فهرست مطالب

فصل ۱

روش های عیب شناسی در رگرسیون خطی ساده و اعمال تبدیلات خطی

همان طور که می دانیم در رگرسیون خطی ساده فروض اساسی زیر برای مدل همواره برقرار بود.

در بخش اول برقراری فروض اساسی فوق را برسی خواهیم نمود.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E(Y|X = x) = \beta_0 + \beta_1 x[E(\epsilon_i|X = x) = 0$$

$$\mathrm{Var}(Y|X = x) = \sigma^2 \qquad \mathrm{Var}(\epsilon_i) = \sigma^2$$

در بخش دوم خواهیم دید که هرگاه از یک مدل رگرسیون استفاده می کنیم به طور ضمنی یک سری از فرضیات را نیز در نظر میگیریم و در ادامه به منظور بررسی فرضیات خود یک سری ابزارهای مشهود را به عنوان عیب شناسی رگرسیون به کار می بریم بدین منظور ابتدا باید برقراری یا عدم برقراری فرضیات چک شوند و دوم اینکه چنانچه فرضیات برقرار نباشند باید راهی برای غلبه بر این شکل را پیدا کنیم.

در بخش تبدیلات خطی طرز استفاده از تبدیلات را برای حالتی که واریانس خطاها ثابت نباشند و یا اینکه رابطه ی بین متغیر پاسخ و مستقل خطی نباشد را بررسی می کنیم که در این صورت قادر خواهیم شد که یک مدل مناسب به داده ها برازش دهیم.

هدف اصلی این فصل این است که بفهمیم وقتی فرضیات اساسی در مدل رگرسیون نقص می شوند و در واقع چه اتفاقی می افتد و به ازای نقض هر فرضیه ما باید چه رویکرد را در نظر بگیریم.

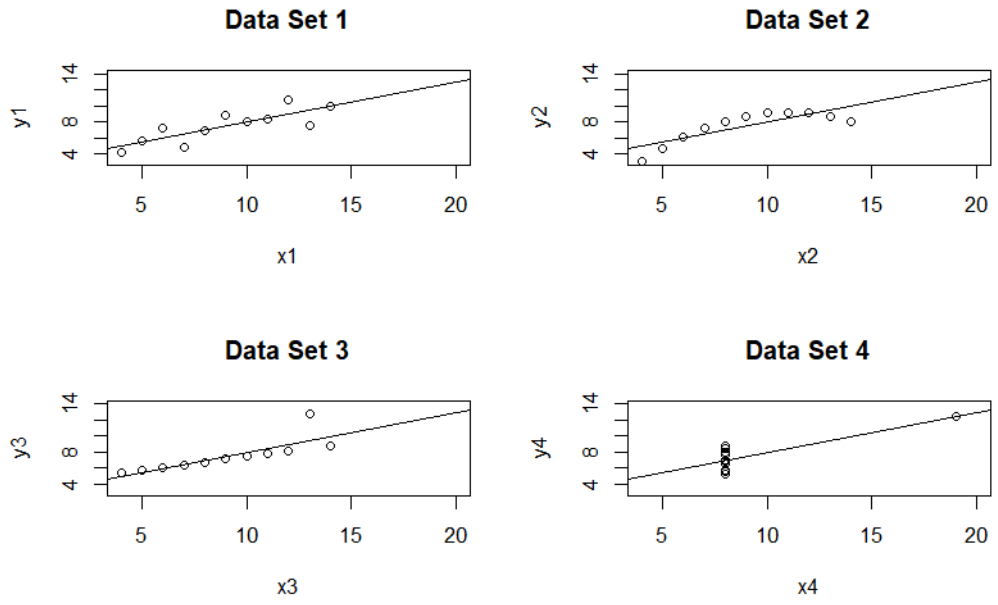
۱.۱ مدل رگرسیون معتبر و نامعتبر:(بررسی ۴ مجموعه داده $An\ Scombe$)

در این بخش چهار مجموعه ساختگی از داده هارا که توسط $An\ Scombe$ ساخته شده اند بررسی می کنیم در این مثال توضیح می دهد که نتایج خروجی در یک مدل رگرسیون می تواند به طور دراماتیک گمراه کننده باشند در یک مدل و منجر به تحلیل نادرست داده ها گردد.

داده ها در جدول زیر آمده است.

جدول ۱.۱: چهار مجموعه داده $Anscombe$

Case	X1	X2	X3	X4	Y1	Y2	Y3	Y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.1	8.84	7.04
7	6	6	6	8	7.14	6.13	6.08	5.25
8	4	4	4	19	4.26	3.1	5.39	12.5
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89



شکل ۱.۱: نمودارهای چهار مجموعه داده *Anscombe*

مدل رگرسیون خطی ساده به صورت زیر آمده است.

$$\hat{Y} = ۳ + ۰⁄۵X$$

همان طور که می بینیم نتیجه مدل برازش شده تا دو رقم اعشار برای همه آن ها یکسان می باشد.

با نگاه به شکل ۱.۱ واضح است که مدل رگرسیون خطی ساده صرفا برای اولین مجموعه از داده ها مناسب است.

زیرا مدل یک تنها مدلی است که برقراری فرضیات اساسی در آن منطقی به نظر می رسد.

$$E(Y|X=x) = \beta_۰ + \beta_۱ X$$

$$\mathrm{Var}(Y|X=x) = \sigma^۲$$

از طرف دیگر به نظر می آید که متغیر پاسخ و مستقل در دومین مجموعه از داده ها دارای رابطه خطی نیستند. سومین مجموعه از داده ها شیب خط رگرسیون تنها توسط یک داده تعیین می گردد. این مثال نشان می دهد که نتایج به دست آمده در خروجی رگرسیون همیشه باید در کنار تحلیل دیداری داده ها چک شود. در این مورد کافی است که به نمودار پراکنش شکل ۱.۱ برای پاسخ به مناسب بودن یا نبودن مدل رگرسیون برازش شده نگاهی بیندازیم.

خروجی رگرسیون در R

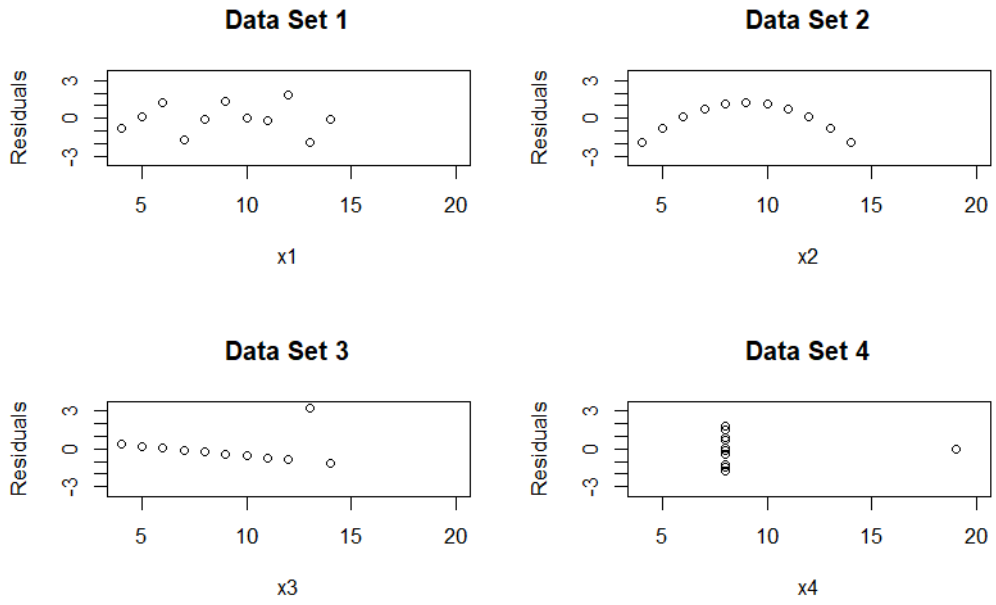
```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,    Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,    Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,    Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,    Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

به هر حال وقتی که چند متغیر مستقل در مدل داریم به ابزار های اضافی برای بررسی مناسب مدل نیازمندیم که در آینده به آنها خواهیم پرداخت.

۱.۱.۱ مانده ها:

یکی از ابزار های مفید برای بررسی مناسب مدل رگرسیون رسم یک یا چند نمودار پراکنش مانده هاست..(یا مانده های استاندارد شده که در ادامه تعریف خواهند شد). این نمودار ها مارا قادر می سازد که به طور شهودی ارزیابی کنیم که آیا مدل مناسب به داده ها برازش شده یا خیر بدون توجه به اینکه چند متغیر مستقل در مدل وجود دارد.



شکل ۲.۱: نمودار مانده ها در مقابل متغیر مستقل

همان طور که در شکل ۲.۱ می بینیم هیچگونه الگویی که نشان دهنده ی غیر تصادفی بودن مانده ها در مقابل X۱ باشد وجود ندارد و بنابر این می توانیم بگوییم که مدل برازش شده به مجموعه داده های یک مناسب است ولی همان طور که می بینیم در سه نمودار (شکل) بعدی هیچ روند تصادفی بین مانده ها و متغیر مستقل وجود ندارد.

۲.۱.۱ استفاده از نمودار های مانده ها برای مشخص نمودن اینکه آیا مدل رگرسیون پیشنهادی معتبر است یاخیر

همان طور که می دانیم یک راه برای چک کردن اینکه مدل رگرسیون معتبر است یا خیر این است که نمودار مانده ها را در مقابل متغیر های مستقل رسم نموده و روند را در آنها بررسی کنیم اگر هیچ روندی وجود نداشته باشد می توان گفت که مدل برازش شده مناسب است.(نمودار پراکنش کاملا

تصادفی است ولی چنانچه طرح یا روندی پیدا شد آنگاه با استفاده از شکل می توان به اطلاعات دست یافت که در مدل برازش شده وجود ندارد.)

مثال ۱.۱.۱. فرض کنید که مدل صحیح به صورت خط زیر باشد.

$$Y_i = E(Y_i|X_i = x_i) + \epsilon_i = \beta_۰ + \beta_۱ X_۱ + \varepsilon_i$$

$$\hat{e}_i = e_i \Rightarrow \hat{Y}_i = \hat{\beta}_۰ + \beta_۱ x_i$$

حال فرض کنید که برآورد های کمترین مربعات $\hat{\beta}_۱$ و $\hat{\beta}_۰$ نزدیک به مقادیر واقعی آن ها یعنی $\beta_۱$ و $\beta_۰$ باشد بنابراین می توان نوشت

$$e_i = Y_i - \hat{Y}_i = (\beta_۰ - \hat{\beta}_۰) + (\beta_۱ - \hat{\beta}_۱)x_i + \varepsilon_i \cong \varepsilon_i$$

بنابراین از رابطه ی اخیر می توان نتیجه گرفت که اگر مدل برازش شده صحیح باشد باید مانده ها (e_i) نیز همانند خطاها (ε_i) رفتاری کاملا تصادفی از خود نشان دهند.

حال اگر مانده ها با متغیرمستقل تغییر کند می توان گفت که یک مدل غلط به داده ها برازش شده است.

مثال ۲.۱.۱. فرض کنید که مدل درجه دوم زیر صحیح باشد

$$y_i = \beta_۰ + \beta_۱ x_i + \beta_۲ x_i^۲ + \varepsilon_i$$

و به اشتباه برآورد کمترین مربعات زیر برای مدل جامعه در نظر گرفته شود.

مدل صحیح $y_i = \beta_۰ + \beta_۱ x_i + \beta_۲ x_i^۲ + \varepsilon_i$

برازش اشتباه $\hat{y}_i = \hat{\beta}_۰ + \hat{\beta}_۱ X_i$

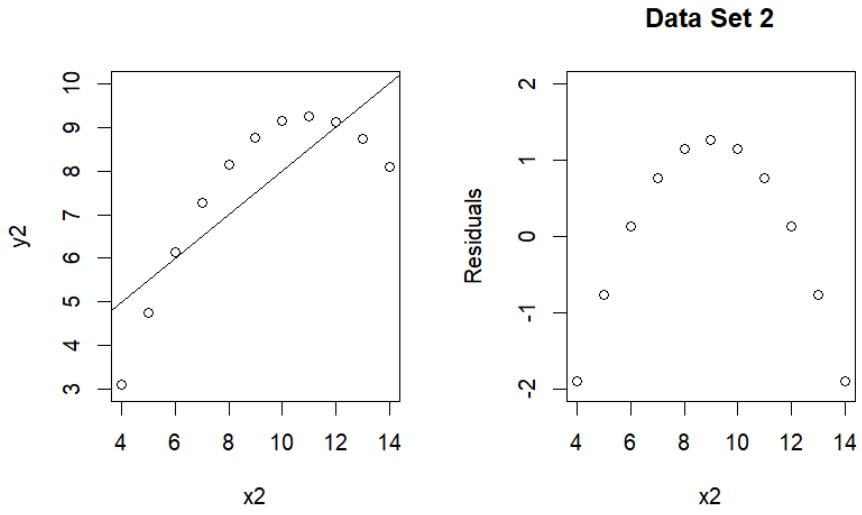
بنابراین به سادگی می توان فرض نمود که برآوردگرهای کمترین مربعات $\hat{\beta}_۰$ و $\hat{\beta}_۱$ به مقادیر واقعی خود نزدیک شوند لذا می توان نوشت:

$$Y_i - \hat{Y}_i = (\beta_۰ - \hat{\beta}_۰) + (\beta_۱ - \hat{\beta}_۱) + \beta_۲ X_i^۲ + \epsilon_i \cong \beta_۲ X_i^۲ + \epsilon_i$$

بنابراین طبق رابطه اخیر چنانچه مدل صحیح به داده ها برازش نشود(به جای مدل رگرسیون درجه دوم از مدل رگرسیون خطی ساده استفاده نماییم.) اگر مانده ها بر حسب X_i ها رسم شوند نشان دهنده یک طرح منحنی شکل خواهند بود که در واقع همان روند می باشد.

۳.۱.۱ مثالی از مدل درجه دوم خطی

فرض کنید که Y یک تابع درجه دو از x های غیر تصادفی باشد حال اگر به اشتباه از مدل رگرسیون خطی ساده استفاده گردد پس از رسم مانده ها در مقابل x یه یک مدل درجه دوم بر می خوریم بنابراین چون شکل مانده ها در مقابل x_i ها یک طرح تصادفی را نشان نمی دهد بنابراین نیاز است که یک مدل درجه دوم به داده ها برازش دهیم. مجموعه داده های دوم $An\ Scomble$ مثالی از این نوع است.



شکل ۳.۱: مجموعه داده های ۲$Anscombe$

شکل ۳.۱ نشان دهنده ی نمودار پراکنش مانده ها در مقابل x_i ها پس از استفاده از مدل رگرسیون خطی ساده (مدل غلط) می باشد همانطور که انتظار می رود یک طرح درجه دوم واضح در این شکل دیده می شود.

۲.۱ ابزار های عیب یابی در رگرسیون: ابزار هایی جهت بررسی اعتبار مدل

در ادامه به بررسی ابزار هایی (که به ابزار های عیب یابی رگرسیون مشهورند) می پردازیم. این ابزار ها برای چک نمودن اعتبار همه ی جنبه های (فرضیات) مدل های رگرسیون به کار می روند هنگامی که یک مدل رگرسیون به داده ها برازش می شود توجه به موارد زیر مهم می باشد:

۱. مشخص نمودن اینکه آیا مدل پیشنهادی یک مدل معتبر است یا خیر؟ ابزار اصلی برای بررسی اعتبار فرضیات رگرسیون همان نمودار های مانده ها می باشد این نمودار ها مارا قادر به ارزیابی

شهودی و فرضیات نقض شده و نکاتی که می توان برای غلبه بر این مشکلات از آن ها استفاده نمود رهنمود می کند.

۲. بررسی اینکه کدام یک از نقاط (اگر وجود داشته باشند) دارای مقادیر متغیر مستقل غیر طبیعی بزرگ هستند که تاثیر زیادی بر روی مدل رگرسیون برازش شده می گذارند .(چنین نقاطی، نقاط اهرمی نامیده می شوند.)

۳. مشخص نمودن اینکه کدام یک از نقاط پرت می باشند یعنی نقاطی که از طرح کلی داده ها تبعیت نمی نمایند.

۴. در صورت وجود داشتن نقاط اهرمی مشخص نمودن اینکه کدام یک از این نقاط اهرمی بد است. اگر یک نقطه اهرمی بد وجود داشته باشد بهتر است که تاثیر آن بر روی مدل برازش شده ارزیابی گردد.

۵. آزمون اینکه آیا فرضیه ثابت بودن واریانس خطاها منطقی است یا خیر؟ اگر نیست ما باید به دنبال راهی برای غلبه براین مشکل باشیم.

۶. اگر داده ها در زمان های متفاوت جمع آوری شده اند آزمون اینکه آیا داده ها بر حسب زمان همبسته هستند یاخیر؟

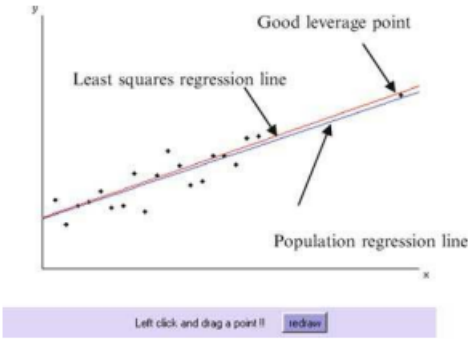
۷. اگر حجم نمونه کوچک است و یا علاقه مند به فواصل پیشگویی هستیم تست اینکه آیا فرضیه نرمال بودن خطاها منطقی است.

موضوع را با بررسی دومین آیتم بالا یعنی نقاط اهرمی شروع می کنیم در بررسی سایر موارد به توضیع مانده های استاندارد نیاز داریم که در ادامه به آن ها خواهیم پرداخت.

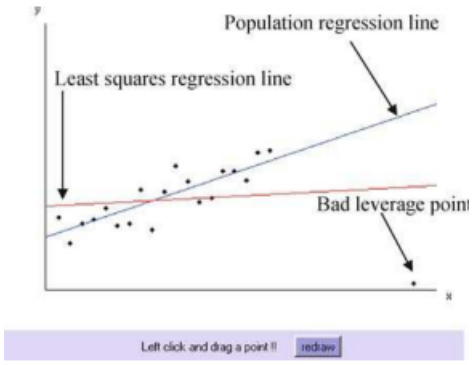
۱.۲.۱ نقاط اهرمی

داده هایی که تاثیر قابل توجهی بر روی مدل برازش شده می گذارند نقاط اهرمی نامیده می شوند این نقاط به دو دسته خوب و بد تفکیک می شوند.

مثالی از نقاط اهرمی خوب و بد:



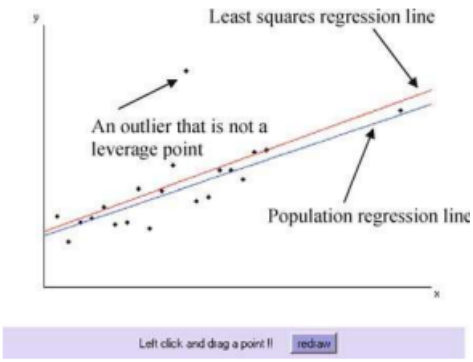
شکل ۴.۱:



شکل ۵.۱:

$Robert\ McCullochs$ از دانشگاه شیکاگو یک برنامک کوچک برای توضیح نقاط اهرمی تولید نمود. این برنامک به طور تصادفی ۲۰ نقطه را که از روی یک خط رگرسیون راست معلوم می آیند را شبیه سازی می کنند. این فرآیند منجر به رسم نقاط در شکل ۳۳.۱ می شود. همان طور که در این شکل دیده می شود یکی از این ۲۰ نقطه دارای مولفه x ای جدا از بقیه است که آن را از توده داده ها جدا می کند. همان طور که خواهیم دید این نقطه که در شکل مشخص شد یک نقطه اهرمی خوب نامیده می شود. این برنامک خط رگرسیون واقعی ($Y = \beta_0 + \beta_1 X$) و خط رگرسیون کمترین مربعات ($\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$) را که بر اساس این ۲۰ نقطه بر داده ها برازش شده مشخص می کند.

حال چنانچه نقاط اهرمی مشخص شده در شکل ۴.۱ به طور قابل توجهی به سمت خط پایین جابه جا کنیم (شکل ۵.۱) و سپس خط کمترین مربعات را مجددا به داده ها ی جدید برازش دهیم در این صورت خواهیم دید که این تک نقطه تاثیر بسیار قابل توجهی بر خط کمترین مربعات داشته و مانند اهرم آن را به سمت خود جابه جا می کند. بنابراین چون این نقطه اهرمی دارای مقدار y ای است که با ۱۹ داده دیگر هم خوانی ندارد (از طرح کلی داده ها تبعیت نمی کند) لذا این نقطه یک نقطه اهرمی بد نامیده می شود.



شکل ۶.۱:

به طور خلاصه می توان گفت که یک نقطه، نقطه اهرمی نامیده می شود اگر مقدار x آن از توده داده ها دارای فاصله قابل توجهی باشد. حال این نقطه یک نقطه اهرمی بد نامیده می شود هرگاه مقدار y آن از طرح سایر نقاط تبعیت نکند به عبارت دیگر یک نقطه اهرمی بد داده ی پرت نیز نامیده می شود. از طرف دیگر با بازگشت به شکل ۴.۱ یک نقطه ی اهرمی،نقطه اهرمی خوب نامیده می شود هرگاه مقدار y آن از طرح کلی سایر داده ها تبعیت کند به عبارت دیگر یک نقطه اهرمی خوب ، نقطه ی اهرمی است که پرت یا دور افتاده نباشد.

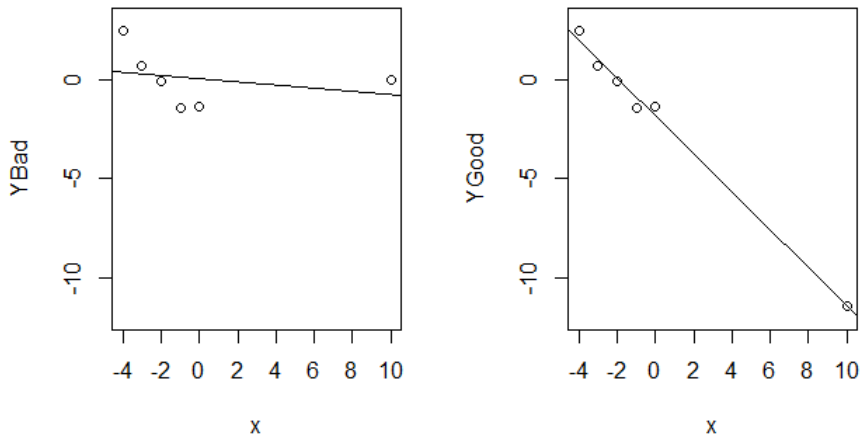
در ادامه بررسی خواهیم نمود که اگر مقدار y یکی از نقاطی را که x آن در وسط قرار گرفته به سمت بالا تغییر دهیم با این تغییر مجموعه داده های جدید در شکل ۶.۱ به همراه خط واقعی رگرسیون و خط کمترین مربعات رسم شده اند همان طور که ملاحظه می کنیم با این جابه جایی تغییر قابل توجهی در خط کمترین مربعات ایجاد نمی شود بنابراین این نقطه، نقطه اهرمی نبوده ولی چون از طرح کلی داده ها تبعیت نمی کند یک نقطه پرت نامیده می شود.

مثال هوبر از نقاط اهرمی خوب و بد :

جدول ۲.۱: مجموعه های داده نقطه اهرم بد و خوب هوبر

x	Y Bad	X	Y Good
-4	2.48	-4	2.48
-3	0.73	-3	0.73
-2	-0.04	-2	-0.04
-1	-1.44	-1	-1.44
0	-1.32	0	-1.32
10	0.00	10	-11.40

همان طور که دیده می شود تنها تفاوت این دو مجموعه داده مقدار Y در $X = ۱۰$ می باشد(در مجموعه داده های بد مقدار Y به ازای $X = ۱۰$ برابر صفر و همین مقدار در داده های خوب برابر با $-۱۱٫۴۰$ می باشد.) این داده ها در شکل رسم شده اند.



شکل ۷.۱: $YGood$و$YBad$ در برابر x با برازش خطوط رگرسیون

همان طور که از شکل بر می آید نقطه $X = ۱۰$ یک نقطه اهرمی در هر دو محموعه از داده ها می باشد. زیرا این نقطه بسیار دور تر از سایر مشاهدات از لحاظ مقادیر X می باشد.

مقایسه ی دو نمودار شکل ۷.۱ به ما این اجازه را می دهد که تاثیر تغییر در متغیر Y به ازای $X = ۱۰$ را بسنجیم، این تغییر در Y منجر به تغییرات قابل توجهی در معادلات برازش خط کمترین مربعات به دو مجموعه از داده ها می شود این تغییرات با دیدن خروجی Rکاملا آشکار است.

خروجی رگرسیون R

```
Call:
lm(formula = YBad ~ x)

Residuals:
      1       2       3       4       5       6
 2.0858  0.4173 -0.2713 -1.5898 -1.3883  0.7463

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06833    0.63279   0.108    0.919
x           -0.08146    0.13595  -0.599    0.581

Residual standard error: 1.55 on 4 degrees of freedom
Multiple R-squared:  0.08237,   Adjusted R-squared:  -0.147
F-statistic: 0.3591 on 1 and 4 DF,  p-value: 0.5813
Call:
lm(formula = YGood ~ x)

Residuals:
      1       2       3       4       5       6
0.47813 -0.31349 -0.12510 -0.56672  0.51167  0.01551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.83167    0.19640  -9.326 0.000736 ***
x           -0.95838    0.04219 -22.714 2.23e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.4811 on 4 degrees of freedom
Multiple R-squared:  0.9923,   Adjusted R-squared:  0.9904
F-statistic: 515.9 on 1 and 4 DF,  p-value: 2.225e-05
```

معادلات کمترین مربعات خطا:

$$Y_{Bad} = ۰/۰۶ - ۰/۰۸x \qquad , R^۲ = ۰/۰۸$$

$$Y_{Good} = -۱/۸۳ - ۰/۹۶x \qquad , R^۲ = ۰/۹۹$$

در ادامه هدف رسیدن به یک قانون عددی در مورد اهرم مربوطه به مشاهده X_i می باشد این قانون می تواند بر اساس دو بند زیر بیان گردد:

الف) فاصله ی X_i از توده سایر X_i ها

ب) مقداری که خط برازش شده رگرسیون توسط آن نقطه جذب (جا به جا) می شود.

دومین بند فوق رابطه دارد با مقداری که $\hat{Y}_i$ (مقدار پیش بینی شده Y در نقطه $X = x_i$) به Y_i بستگی پیدا می کند می توان نوشت:

$$\begin{aligned}
\hat{Y}_i &= \hat{\beta}_۰ + \hat{\beta}_۱ X_i \qquad\qquad = \bar{Y} - \hat{\beta}_۱ \bar{X} + \hat{\beta}_۱ X_i \\
&= \frac{۱}{n} \sum_{j=۱}^{n} Y_j + (X_i - \bar{X}) \sum_{j=۱}^{n} \frac{(X_j - \bar{X}) Y_j}{S_{xx}} \\
&= \sum_{j=۱}^{n} [\frac{۱}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{xx}}] Y_j \\
&= \sum_{j=۱}^{n} h_{ij} Y_j
\end{aligned}$$

$$\begin{cases} \hat{\beta}_۱ = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum_{i=۱}^{n} (X_i - \bar{X}) Y_i}{S_{xx}} \\ \hat{\beta}_۰ = \bar{Y} - \hat{\beta}_۱ \bar{X} \end{cases}$$

باید توجه داشت که مجموعه وزن ها برابر۱ است زیرا:

$$\sum_{j}^{n} h_{ij} = \sum_{j}^{n} [\frac{۱}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{xx}}] = ۱ + ۰ = ۱$$

در برازش متغیر Y در نقطه $X = X_i$ وزن داده شده به متغیر Y_i به صورت زیر محاسبه می گردد:

$$\hat{Y_i} = \sum_{j=۱}^{n} h_{ij} Y_j = h_{ii} Y_i = \sum_{j=۱, i \neq j}^{n} h_{ij} Y_j \qquad (۱.۱)$$

$$h_{ii} = \frac{۱}{n} + \frac{(X_i - \bar{X})^۲}{S_{xx}}$$

جمله ی h_{ii} اهرم مربوط به نقطه i ام نامیده می شود. همان طور که از فرمول مربوط به آن بر می آید مقدار $(X_i - \bar{X})^۲$ در دومین قسمت آن فاصله ی X_i از توده X_i هارا اندازه می گیرد. توجه داشته باشید که h_{ii} تاثیر Y_i را در برازش $\hat{Y_i}$ نشان می دهد.

به عنوان مثال فرض کنید برای مشاهده ای $h_{ii} = ۱$ باشد در این صورت مجموع سایر h_{ij} ها برابر با صفر خواهند شد زیرا کل h_{ij} ها مساوی ۱ است.

$$\hat{Y_i} = h_{ij} y_i + \sum_{j=۱, j \neq i}^{n} h_{ij} Y_j = ۱ \times Y_i + partial\ amount \cong Y_i$$

که در این صورت مقادیر برازش شده و مشاهده شده با هم برابر می شوند باید توجه داشت که h_{ii} تنها به مقدار x_i بستگی دارد نه به مقدار Y_i ها. به راحتی می توان نشان داد که متوسط h_{ii} ها برابر با $\frac{۲}{n}$ می شود:

$$\frac{۱}{n} \sum h_{ii} = \frac{۱}{n} \sum [\frac{۱}{n} + \frac{(X_i - \bar{X})^۲}{S_{xx}}] = \frac{۱}{n} + \frac{\sum (X_i - \bar{X})^۲}{n S_{xx}} = \frac{۱}{n} + \frac{۱}{n} = \frac{۲}{n}$$

یک قانون عددی برای شناسایی نقاط اهرمی:

طبق این قانون نقطه X_i یک نقطه اهرمی یا اهرم بزرگ نامیده می شود هرگاه:

$$h_{ii} > ۲ avemge(h_{ii}) = ۲ \times \frac{۲}{n} = \frac{۴}{n}$$

مجددا به مثال در مجموعه شش تایی هابر ۵.۱ بازگشته و مقادیر h_{ii} ها را برای کلیه مشاهدات محاسبه می نماییم با انجام محاسبات مربوط در می یابیم که :

$$h_{۶۶} = ۰/۹۳۵۹ > \frac{۴}{n} = \frac{۴}{۶} = ۰/۶۷$$

(این مقادیر برای کلیه مشاهدات در جدول زیرمحاسبه شده اند)

جدول ۳.۱: مقادیر اهرمی برای دو مجموعه داده هوبر

i	x_i	Leverage,h_{ii}
1	-4	0.2897
2	-3	0.2359
3	-2	0.1974
4	-1	0.1744
5	0	0.1667
6	10	0.9359

استراتژی برخورد با نقاط اهرمی بد:

۱-حذف نقطه اهرمی بد از مجموعه داده ها:

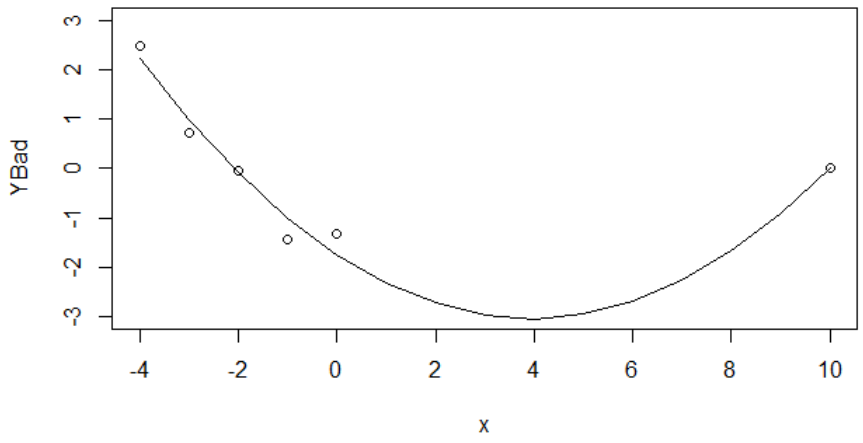
سوال اعتبار مجموعه داده ها مربوط می شود به نقاط اهرمی بد به عبارت دیگر می توان گفت:آیا این نقاط داده ها غیر معمول یا متفاوت از سایر داده ها هستند؟ اگر چنین است می توان این نقاط را حذف نمود و مجددا یک مدل رگرسیون بدون در نظر گرفتن این نقاط به داده ها برازش داد. لذا تنها در صورتی مجاز به حذف داده ها هستیم که با برازش مدل های متفاوت به این داده ها، این داده ها متفاوت از سایرین باشند.

۲-برازش یک مدل متفاوت دیگر به داده ها:

پرسش در مورد اعتبار مدل برازش شده یعنی :"آیا یک مدل رگرسیون غیر صحیح به داده ها برازش شده است؟" اگر چنین است مدل دیگری را با اضافه نمودن متغیر های مستقل به مدل

قبلی برررسی نمایید.(به عنوان مثال مدل چند جمله ای) رویکرد دوم شامل استفاده از اعمال تبدیلات بر روی متغیر های X و یا Y خواهد بود که به طور کامل در همین فصل بررسی خواهندشد.

مثال: در مورد داده های بد هابر چنانچه یک مدل رگرسیون درجه دوم $(Y_{Bad} = \beta_0 + \beta_1 X + \beta_2 X^2)$ به داده ها مجددا برازش دهیم آنگاه خواهیم دید که هیچ نقطه اهرمی در مدل جدید وجود نداشته و همه ی داده ها به خوبی توسط این مدل پوشیده می شوند. شکل ۸.۱ و خروجی این نکته را به خوبی تایید می کند:



شکل ۸.۱: نمودار $YBad$ در مقابل x با برازش مدل درجه دوم اضافه شده

خروجی رگرسیون از R

```
Call:
lm(formula = YBad ~ x + I(x^2))

Residuals:
1          2          3          4          5          6
```

```
0.24695  −0.25918   0.04771  −0.44237   0.42057  −0.01367

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) −1.74057    0.29702   −5.860   0.00991 **
x           −0.65945    0.08627   −7.644   0.00465 **
I(x^2)       0.08349    0.01133    7.369   0.00517 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.4096 on 3 degrees of freedom
Multiple R-squared:  0.952,     Adjusted R-squared:  0.9199
F-statistic: 29.72 on 2 and 3 DF,  p-value: 0.01053
```

۲.۲.۱ مانده های استاندارد شده

گاهی اوقات برای پی بردن به مشکلات به وجود آمده مدل رگرسیون به مانده هایی نیازمندیم که دارای واریانس یکسان یاهم واریانس هستند. زیرا مانده ها بر خلاف خطاها هم واریانس نبوده و در واقع دارای واریانس به صورت زیر هستند:

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$$

$$e_i = Y_i - \hat{Y}_i \Rightarrow e_i = Y_i - \sum_{j=1}^{n} h_{ij}Y_j = (1 - h_{ii})Y_i - \sum_{j=1, j\neq i}^{n} h_{ij}Y_j$$

$$\Rightarrow \mathrm{Var}(e_i) = (1 - h_{ii})^2\sigma^2 + \sigma^2 \sum_{j=1, j\neq i}^{n} h_{ij}^2 = \sigma^2(-2h_{ii} + \sum_{j=1}^{n} h_{ij}^2)$$

از طرفی:

$$\sum_{j=1}^{n} h_{ij}^2 = \sum_{j=1}^{n} \left[\frac{1}{n^2} + \frac{(X_i - \bar{X})^2(X_j - \bar{X})^2}{S_{xx}^2} + \frac{2(X_i - \bar{X})(X_j - \bar{X})}{S_{xx}}\right] = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}} = h_{ii}$$

بنابراین:

$$\mathrm{Var}(e_i) = \sigma^۲(۱ - h_{ii})[*] \quad \forall i = ۱, \ldots, n$$

اگر $h_{ii} \cong ۱$ باشد می توان نتیجه گرفت که $\mathrm{Var}(e_i) \cong ۰$ و این یعنی که e_i در امید ریاضی خود یعنی $E(e_i) = ۰$ چگال یا دارای توزیع تباهیده است. پس اگر i امین نقطه یک نقطه اهرمی باشد می توان نتیجه گرفت که مانده مربوط به آن نقطه دارای واریانس نزدیک به صفر است. همچنین با استفاده از رابطه $\mathrm{Var}(\hat{Y}_i) = \sigma^۲ h_{ii}$ می توان گفت که چون $E(\hat{Y}_i) = Y_i$ و $\mathrm{Var}(\hat{Y}_i) \cong \sigma^۲$ آنگاه $\hat{y}_i \cong Y_i$ است. اما بنابراین رابطه ۲.۲.۱ چون واریانس i ام مانده به اندیس i وابسته است، لذا برای رفع این مشکل و بررسی مانده ها جهت پی بردن به برقراری فرضیات اساسی و عیب یابی رگرسیون، مانده هارا استاندارد می کنیم تا هم واریانس گردند. مانده های استاندارد شده به صورت زیر تعریف می گردند:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{۱ - h_{ii}}} \qquad , \hat{\sigma} = \sqrt{MSE}$$

مزیت مانده های استاندارد شده:

۱- هنگامی که نقاط با اهرم بزرگ (سنگین) وجود دارد، برای عیب یابی رگرسیون بسیار مفید تر این است که به جای رسم نمودار مانده ها، مانده های استاندارد شده را رسم نماییم زیرا حتی اگر خطاها هم واریانس نباشند، مانده های استاندارد شده هم واریانس خواهند بود. (ولی اگر نقاط با اهرم های سنگین وجود نداشته باشد، دو نمودار یعنی مانده ها و مانده های استاندارد شده تفاوت زیادی باهم ندارند.)

۲- نمودار مانده های استاندارد شده سریعا به ما خواهند گفت که کدام نقاط یعنی Y_i دارای اختلاف فاحش با برآوردشان ($\hat{Y_i}$) هستند. به عنوان مثال فرض کنید نقطه ای دارای مانده استاندارد شده ۳.۴ باشد، این بدین معنی است که این نقطه ۳.۴ برابر خطای استاندرد ($\sqrt{MSE}$) از مقدار برآورد شده یعنی ($\hat{Y_i}$) دور تر است، حال اگر خطاها از توزیع نرمال استخراج شده باشند، احتمال مشاهده اینکه نقطه ای ۳.۴ برابر خطای استاندارد دور تر از مقدار برآورد شده باشد تقریبا غیرممکن یا برابر صفر است.

اگر $r_i \in [-۲, ۲]$ باشد می توان گفت که نقطه i ام یک نقطه طبیعی بوده ولی چنانچه $|r_i| > ۲$ باشد، می توان نتیجه گرفت که مشاهده i ام یک مشاهده پرت یا غیر طبیعی است. چنانچه $|r_i| > ۴$ باشد، وضعیت خیلی اضطراری خواهد بود. همچنین می توان گفت که اگر i امین مشاهده یک نقطه اهرمی بد باشد آنگاه $|r_i| > ۲$ باشد، (یک نقطه اهرمی بد، یک یک مشاهده پرت نیز خواهد بود ولی عکس آن صحیح نمی باشد.) و اگر یک نقطه دلخواه ، یک نقطه اهرمی خوب باشد آنگاه $|r_i| \leq ۲$.

می توان نشان داد:

$$\mathrm{Cov}(e_i, e_j) = -h_{ij}\sigma^۲ \quad (i \neq j)$$

زیرا

$$cov(e_i, e_j) = \mathrm{Cov}(Y_i - \sum_{j=۱}^{n} h_{ij}Y_j, Y_j - \sum_{j=۱}^{n} h_{ij}Y_j) = -h_{ii}\sigma^۲ - h_{ij}\sigma^۲ + \sum_{j=۱}^{n} h_{ij}^۲\sigma^۲ = -h_{ij}\sigma^۲$$

$$\Rightarrow corr(e_i, e_j) = \rho(e_i, e_j) = \frac{-h_{ij}}{(۱ - h_{ii})(۱ - h_{jj})} \qquad , i \neq j$$

البته مقدار همبستگی فوق بسیار ناچیز است مخصوصا در مواقعی که مشاهدات بر حسب

زمان جمع آوری شده باشد.

$$\mathrm{Var}(\hat{Y}_i) = \mathrm{Var}(\sum_{j=۱}^{n} h_{ij} Y_j) = \sigma^۲ \sum_{j=۱}^{n} h_{ij}^۲ = \sigma^۲ h_{ii}$$

جدول ۴.۱: تشخیص رگرسیون برای مدل در شکل ۹.۱

case	Coupon rate	Bid price	Leavrage	Residuals	Std.Residualse
1	7.000	92.94	0.049	–3.309	–0.812
2	9.000	101.44	0.029	–0.941	–0.229
3	7.000	92.66	0.049	–3.589	–0.881
4	4.125	94.50	0.153	7.066	1.838
5	13.125	118.94	0.124	3.911	1.001
6	8.000	96.75	0.033	–2.565	–0.625
7	8.750	100.88	0.029	–0.735	–0.179
8	12.625	117.25	0.103	3.754	0.949
9	9.500	103.34	0.030	–0.575	–0.140
10	10.125	106.25	0.036	0.419	0.102
11	11.625	113.19	0.068	2.760	0.685
12	8.625	99.44	0.029	–1.792	–0.435
13	3.000	94.50	0.218	10.515	2.848
14	10.500	108.31	0.042	1.329	0.325
15	11.250	111.69	0.058	2.410	0.595
16	8.375	98.09	0.030	–2.375	–0.578
17	10.375	107.91	0.040	1.313	0.321
18	11.250	111.97	0.058	2.690	0.664
19	12.625	119.06	0.103	5.564	1.407
20	8.875	100.38	0.029	–1.618	–0.393
21	10.500	108.50	0.042	1.519	0.372
22	8.625	99.25	0.029	–1.982	–0.482
23	9.500	103.63	0.030	–0.285	–0.069
24	11.500	114.03	0.064	3.983	0.986
25	8.875	100.38	0.029	–1.618	–0.393
26	7.375	92.06	0.041	–5.339	–1.306
27	7.250	90.88	0.044	–6.136	–1.503
28	8.625	98.41	0.029	–2.822	–0.686
29	8.500	97.75	0.030	–3.098	–0.753
30	8.875	99.88	0.029	–2.118	–0.515
31	8.125	95.16	0.032	–4.539	–1.105
32	9.000	100.66	0.029	–1.721	–0.418
33	9.250	102.31	0.029	–0.838	–0.204
34	7.000	88.00	0.049	–8.249	–2.025
35	3.500	94.53	0.187	9.012	2.394

مثال ۱.۲.۱. مثال قیمت قراردادهای خزانه داری آمریکا

در این مثال خواهیم دید که تعداد کمی از داده های پرت می توانند تاثیر قابل توجهی بر مدل برازش شده بگذارند. همچنین خواهیم دید که حذف این نقاط، به طرز قابل توجهی بر بهبود مدل برازش شده و فواصل اطمینان خواهند گزارد. مدل مورد بررسی به قرار زیر است:

X: قیمت کوپن قرارداد که سالی دوبار تعیین می گردد.

Y: قیمت پیشنهادی جاری

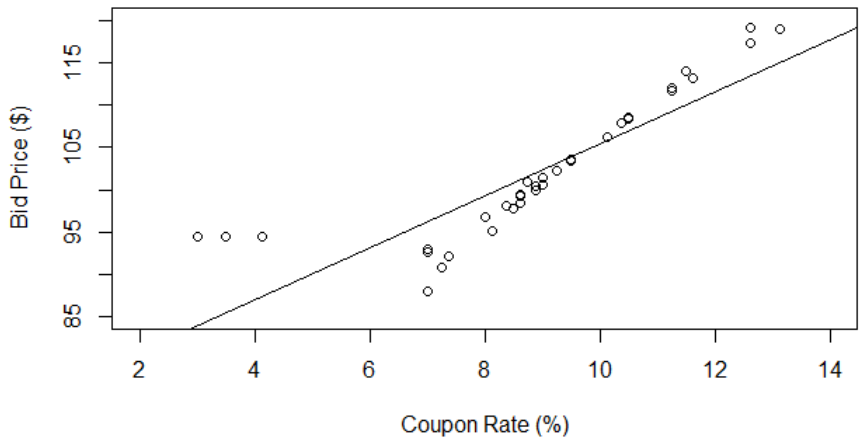
مدل برازش شده به تمام داده ها:

$$\hat{Y}_i = ۷۴٫۷۸ + ۳٫۰۶X_i \qquad (R^۲ = ۰٫۷۵۱۶)$$

$$\hat{\beta}_۰ \in (۶۹٫۰۳, ۸۰٫۵۲) \Rightarrow (٪۲٫۵, ٪۹۷٫۵)$$

$$\hat{\beta}_۱ \in (۲٫۴۴, ۳٫۶۹) \qquad ٪ ۹۵$$

یعنی اگر یک واحد متغیر X افزایش یابد می توان گفت که به طور میانگین متغیر پاسخ یعنی Y حداقل۲٫۴۴ و حداکثر۳٫۶۹واحد افزایش خواهد یافت.(با احتمال ۹۵٪)



شکل ۹.۱: نموداری از داده های اوراق قرضه با حداقل مربعات خط گنجانده شده

خروجی رگرسیون R

```
Call:
lm(formula = BidPrice ~ CouponRate)

Residuals:
Min      1Q  Median      3Q     Max
-8.249  -2.470  -0.838   2.550  10.515

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
  2.5 % 97.5 %
(Intercept) 69.036 80.537
CouponRate   2.442  3.690
```

همان طور که دیدیم:

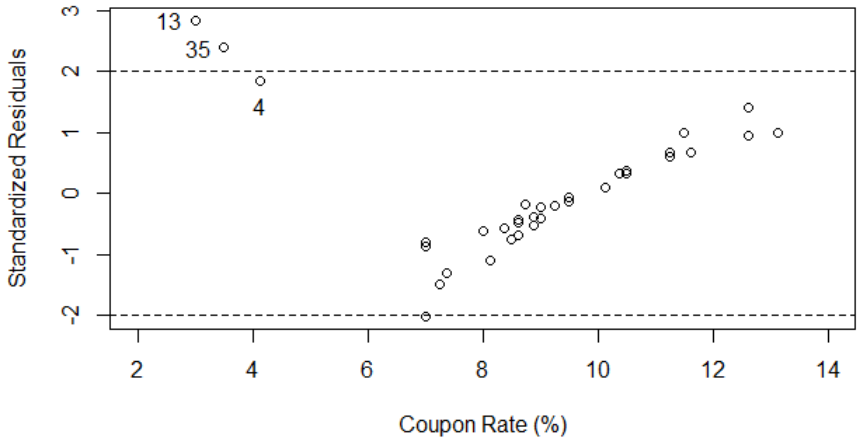
$$mean(h_{ii}) = \frac{۲}{n}$$

هرگاه برای داده ای $h_{ii} > \frac{۴}{n} = ۲mean(h_{ii})$ ،آنگاه می توان گفت که نقطه i ام یک نقطه اهرمی است.

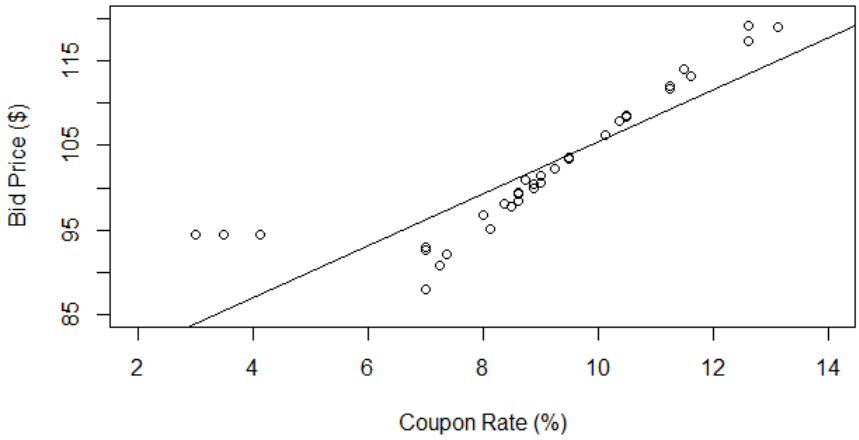
برای این مثال $\frac{۴}{n} = \frac{۴}{۳۵} = ۰/۱۱$ و بنابراین نقاط ۴و۵و۱۳و۳۵اهرمی می باشند.

اما برای پیدا کردن مشاهدات پرت چون برای نقاط ۱۳و۳۴و۳۵ چون $|r_i| > ۲$ است، لذا این نقاط، نقاط پرت خواهند بود. بنابراین نقاط ۱۳و۳۵ نقاط اهرمی بد و نقاط ۵و۴ نقاط اهرمی خوب هستند. در شکل ۱۰.۱ نمودار متغیر مستقل در مقابل مانده های استاندارد شده رسم شده است. از شکل این نمودار می توان گفت که مدل برازش شده مناسب نمی باشد زیرا نمودار پراکندگی شکلی غیر تصادفی را نشان می دهد. با حذف نقاط پرت، برآورد مدل به صورت زیر خواهد شد:

$$\hat{Y_i} = ۵۷/۲۹ + ۴/۸۳X_i \quad , R^۲ = ۰/۹۸۵۲$$



شکل ۱۰.۱: طرحی از باقیمانده های استاندارد شده با تعدادی از موارد نمایش داده شده



شکل ۱۱.۱: نموداری از داده های اوراق قرضه با پیوندهای "گل" حذف شده

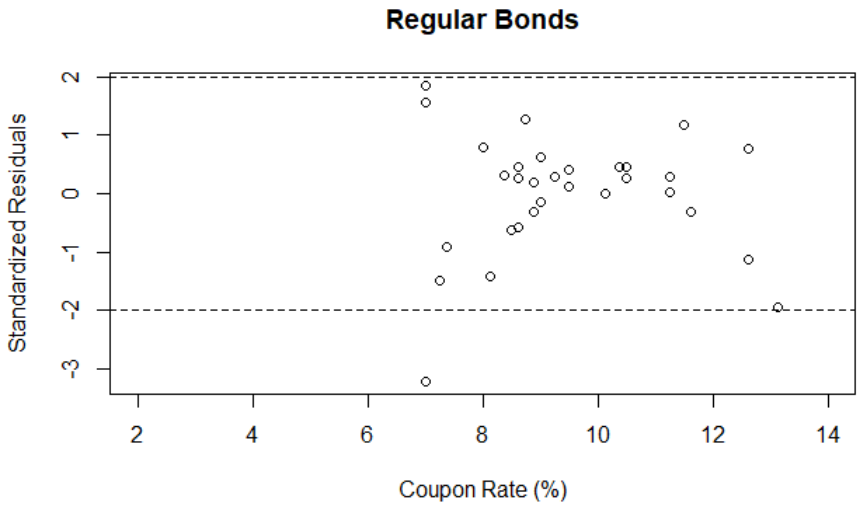
خروجی رگرسیون R

```
Call:
lm(formula = BidPrice ~ CouponRate, subset = (1:35)[-c(4, 13,
35)])

Residuals:
Min       1Q  Median      3Q     Max
-3.1301 -0.3789  0.2240  0.4576  1.8099

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.2932     1.0358   55.31   <2e-16 ***
CouponRate    4.8338     0.1082   44.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 1.024 on 30 degrees of freedom
Multiple R-squared:  0.9852,    Adjusted R-squared:  0.9847
F-statistic:  1996 on 1 and 30 DF,  p-value: < 2.2e-16
```

همان طور که می بینیم $\hat{\beta}_1$ در مدل اخیر حتی در فاصله اطمینان بدست آمده برای $\hat{\beta}_1$ بر اساس کلیه داده ها قرار نمی گیرد که نشان دهنده تحلیل خام داده ها براساس کلیه داده ها می باشد و لذا بدون حذف داده های پرت، نتایج بسیار گمراه کننده خواهند بود.

این مثال نشان دهنده اهمیت برآورد ها و فواصل اطمینان بر اساس یک مدل معتبر است.

چنانچه مجددا مانده های استاندارد را در مقابل X_i ها رسم کنیم(برای مدل صحیح) می توان به مناسب بودن مدل پس از حذف داده های پرت پی برد. (۱۲.۱شکل)



شکل ۱۲.۱: مانده های استاندارد در مقابل X_i

۳.۲.۱ توصیه ای برای برخورد با مشاهدات پرت و نقاط اهرمی:

(آ) نقاط نباید به طوری عادی از مجموعه داده ها حذف گردند و آنالیز بدون آن ها انجام گردد فقط به این دلیل که آن ها به مدل برازش شده سازگار نیستند. نقاط پرت و اهرمی بد در واقع سیگنال هایی برای نشان دادن بدی مدل برازش شده هستند.

(ب) مشاهدات پرت می توانند نشان دهنده مشکلاتی باشند که تاکنون پیش نیامده و در آینده ممکن است رخ دهد .ممکن است آن ها اشاره کننده به برازش مدل دیگر باشند که در آن دیگر مشاهدات پرت، پرت نباشند.

۴.۲.۱ ارزیابی تاثیر نقاط متفاوت بر مدل برازش شده:

در این قسمت تاثیر هر مشاهده به تنهایی را بر مدل رگرسیون برازش شده بررسی خواهیم نمود. در مثال قبل دیدیم که سه نقطه پرت توانستند با تاثیر قابل توجهی که بر مدل می گذارند خط رگرسیون برازش شده را به طور شگفت آوری جابه جا نمایید. در ادامه آماره ای را معرفی خواهیم نمود که به کمک آن می توان اندازه تاثیر هر مشاهده دلخواه بر مدل برازش شده به تنهایی را محاسبه نمود.

این آماره، فاصله کوک ($cook\ distance$) نامیده می شود که نخستین بار توسط کوک ۱۹۷۷ پیشنهاد گردید و به شکل زیر محاسبه می شود:

$$D_i = \frac{\sum_{j=۱}^{n}(\hat{Y_{j(i)}} - \hat{Y_j})^۲}{۲MSE}$$

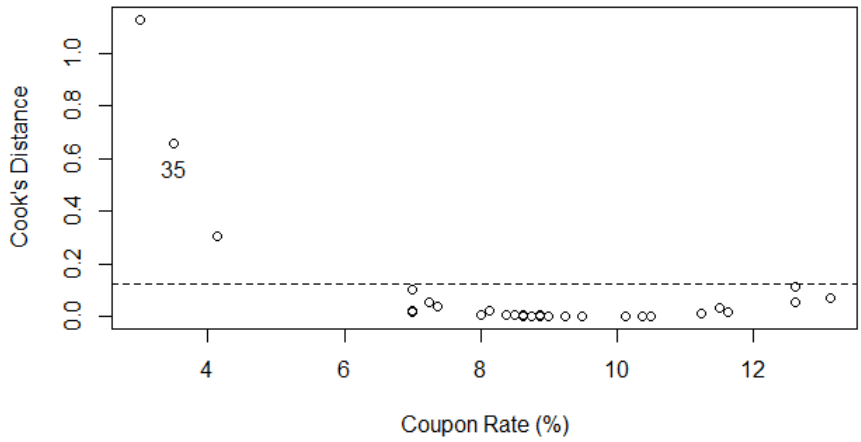
که در آن $\hat{Y}_{j(i)}$ نشان دهنده برازش مدل مدل بدون استفاده از مشاهده i ام به $(n-۱)$ داده باقیمانده است.

در عمل استفاده از فرمول فوق برای محاسبه D_i ها سخت است زیرا برای هر i، باید i امین مشاهده را کنار گذاشته و مدل را بدون برازش آن برازش نموده و سپس مجموع توان دوم مقادیر برازش شده بدون مشاهده i ام از مقادیر برازش شده به کلیه داده ها را محاسبه نمود. لذا برای حل این مشکل می توان از فرمول محاسباتی D_i ها استفاده نمود:

$$D_i = \frac{r_i^۲}{۲} \cdot \frac{h_{ii}}{۱ - h_{ii}}$$

بنابراین می توان گفت که D_i به طور مستقیم با $r_i^۲$ سرو کار داشته و یک تابع صعودی از h_{ii} است. به عبارت دیگر اگر h_{ii} بزرگ و $r_i^۲$ نیز بزرگ شود می توان نتیجه گرفت که مشاهده i ام تاثیر زیادی بر برازش مدل خواهد داشت. در حالت کلی اگر D_i کمتر از ۱ باشد، مشاهده i ام پرت نبوده و حذف آن تاثیر معنی داری بر مدل برازش شده نخواهد گذاشت. فاکس (Fox) در سال ۲۰۰۲، مقدار $\frac{۴}{n-۲}$ را به عنوان یک مقدار بحرانی برای مدل رگرسیون خطی ساده معرفی نمود و چنانچه $D_i > \frac{۴}{n-۲}$، آنگاه می توان گفت که مشاهده i ام تاثیر گزار بر مدل برازش شده است.

مثال ۲.۲.۱. در مثال ۱.۲.۱ آماره D_i مربوط به سه مشاهده ۴و۳۵و۱۳ از حد $\frac{۴}{n-۲} =$ ۰٫۱۲ تجاوز نموده و $D_{۱۳} > ۱$ شده است.



شکل ۱۳.۱: طرحی از فاصله کوک در برابر نرخ کوپن

۵.۲.۱ بررسی نرمال بودن توزیع خطاها:

همان طور که می دانیم، اگر حجم نمونه تصادفی کم باشد، برای استفاده از آماره t مربوط به آزمون فرضیات و همچنین بدست آوردن فواصل اطمینان برای پارامتر ها، فرضیه نرمال بودن خطاها مورد نیاز است. این فرضیه می تواند با نگاه کردن به توزیع مانده ها یا مانده های استاندارد شده بررسی گردد.

می توان نوشت:

$$e_i = Y_i - \hat{Y}_i = Y_i - \sum_{j=1}^{n} h_{ij} Y_j = \beta_0 + \beta_1 X_i + \epsilon_i - \sum_{j=1}^{n} h_{ij}(\beta_0 + \beta_1 X_j + \epsilon_j)$$
$$= \beta_0 + \beta_1 X_i + \epsilon_i - \beta_0 \sum_{j=1}^{n} h_{ij} - \beta_1 \sum_{j} h_{ij} X_j - \sum_{j=1}^{n} h_{ij}\epsilon_j$$

اما

$$\sum_{j=1}^{n} h_{ij} X_j = \sum_{j=1}^{n} \frac{X_j}{n} + \sum_{j=1}^{n} \frac{X_j (X_i - \bar{X})(X_j - \bar{X})}{S_{xx}} = \bar{X} + \frac{(X_i - \bar{X}) S_{xx}}{S_{xx}} = X_i$$

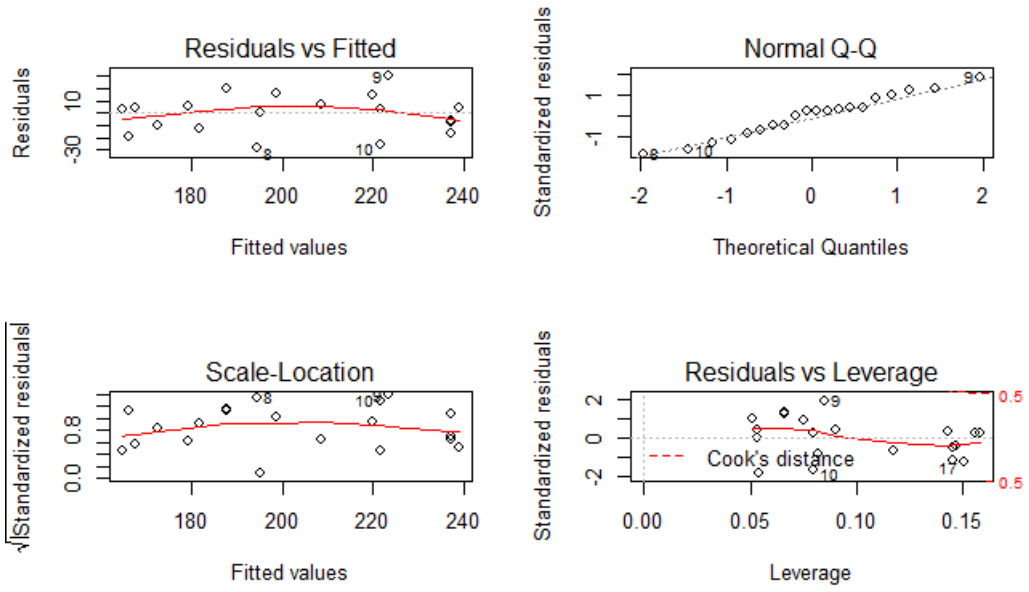
بنابراین:

$$e_i = \beta_0 + \beta_1 X_i + \epsilon_i - \beta_0 - \beta_1 X_i - \sum_{j=1}^{n} h_{ij}\epsilon_j = \epsilon_i - \sum_{j=1}^{n} h_{ij}\epsilon_j$$

حال اگر حجم نمونه مورد بررسی کوچک باشد، ممکن است جمله دوم رابطه بالا بزرگ تر از اولی شود و لذا حتی اگر خطاها نرمال نباشند باز هم e_i ها نرمال به نظر می آیند. در صورتی که اگر حجم نمونه بزرگ باشد واریانس جمله اول بسیار بزرگ تر از واریانس جمله دوم شده

و لذا اگر خطاها نرمال نباشند در نمودار پراکنش مانده ها واضح خواهد بود.

اما برای بررسی دقیق تر این موضوع کافی است نمودار $Q-Q$ (چندک-چندک) یا $p-p$ (نمودار احتمال) توزیع نرمال را برای مانده های استاندارد شده رسم نماییم. در این نمودار مانده های استاندارد شده به ترتیب صعودی (چندک های تجربی) را در مقابل چندک های توزیع نرمال استاندارد (چندک های ریاضی مورد انتظار) به ترتیب بر روی محور Y ها و X ها رسم می نماییم.



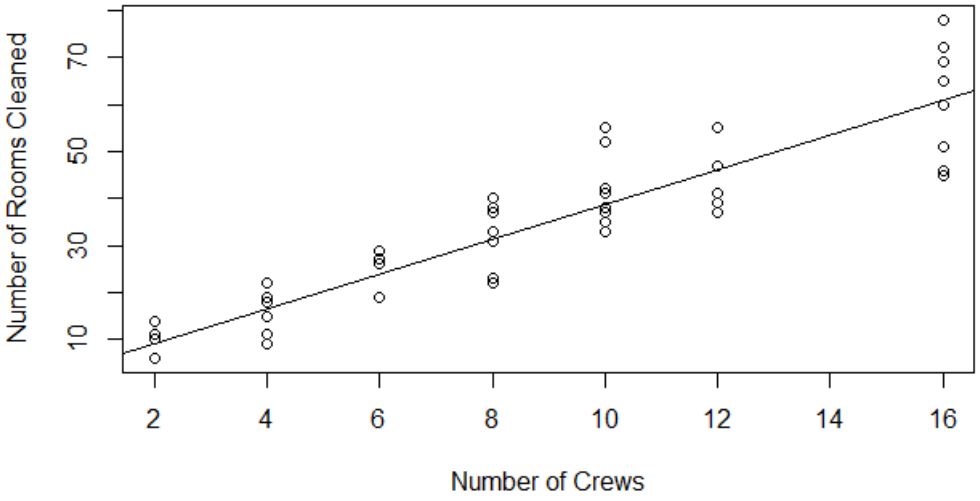
شکل ۱۴.۱: یک نمودار $Q-Q$ معمولی و سایر نمودارها

بنابر این :

$$\begin{cases} y & r_{(1)} < r_{(2)} < \cdots < r_{(i)} < \cdots < r_{(n)} \\ x & \phi^{-1}\dfrac{(i-1/2)}{n} \end{cases}$$

برای بررسی نرمال بودن، اگر نمودار پراکنش داده ها بر روی خط راست واقع شد آنگاه توزیع

خطاها نرمال و در غیر این صورت اگر منحنی ظاهر شد، نرمال نخواهد بود. همچنین اگر نمودار پراکنش بر روی خط $y = x$ واقع شد، توزیع خطاها نرمال استاندارد است.



شکل ۱۵.۱: نمودار داده های تمیز کردن اتاق با حداقل مربعات خطا

مثال ۳.۲.۱. بر اساس مثال ۱.۲.۱ نرمال بودن توزیع خطاها در داده های مربوط به به قیمت های قرار داد های خزانه داری بررسی گردد.

با استفاده از دستور $plot(Lm(Y\ x))$ می توان نمودار مانده ها و مانده های استاندارد در مقابل مقادیر برازش شده، نمودار $Q-Q$ و مانده های استاندارد در مقابل قدرت نفوذ داده ها (لوریج ها) را رسم نمود.

از رسم مانده های استاندارد در مقابل لوریج ها می اوان نقاط اهرمی بد را مشخص نمود و از رسم نمودار مانده ها در مقابل مقادیر برازش شده می توان به یک سویی مدل برازش شده و

ثبات واریانس پی برد. اگر این نمودار طرحی را نشان دهد می توان گفت که مدل برازش شده مناسب نیست(طرح منحنی) و یا واریانس ثابت نیست(طرح قیفی شکل)

چرا e_i هارا در مقابل $\hat{Y}_i$ رسم می کنیم در مقابل Y_i ها؟؟

زیرا e_i با $\hat{Y}_i$ ناهمبسته است ولی با Y_i همبسته است. ($\underline{e}'\hat{Y}_i = ۰$)

$$\rho(e_i, Y_i) = \sqrt{۱ - R^۲}$$

$$\rho(e_i, Y_i) = \frac{\underline{e}'\underline{Y}}{\sqrt{e'e.\sum(Y_i - \bar{Y})^۲}} = \frac{\sum e_i(Y_i - \bar{Y})}{\sqrt{\sum e_i^۲.\sum(Y_i - \bar{Y})^۲}}\sqrt{۱ - R^۲}$$

در چه صورت رسم نمودار فوق منطقی است؟

وقتی که مدل برازش کامل داشته باشد یعنی $R^۲ = ۱$

۶.۲.۱ بررسی ثبات واریانس:

همان طور که می دانیم یکی از فروض اساسی رگرسیون، ثابت بودن واریانس خطاها می باشد. هرگاه واریانس جملات خطا ثابت نباشند، دو رویکرد برای برخورد با آن داریم:

۱- استفاده از تبدیلات

۲- استفاده از رگرسیون وزنی

جدول ۵.۱: داده های مربوط به تمیز کردن اتاق ها

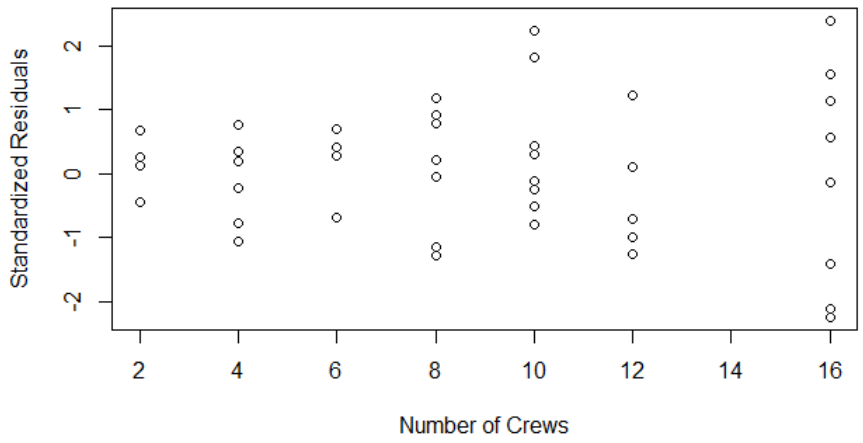
case	Number of crews	Rooms cleaned	case	Number of crews	Rooms cleaned
1	16	51	28	4	18
2	10	37	29	16	72
3	12	37	30	8	22
4	16	46	31	10	55
5	16	45	32	16	65
6	4	11	33	6	26
7	2	6	34	10	52
8	4	19	35	12	55
9	6	29	36	8	33
10	2	14	37	10	38
11	12	47	38	8	23
12	8	37	39	8	38
13	16	60	40	2	10
14	2	6	41	16	65
15	2	11	42	8	31
16	2	10	43	8	33
17	6	19	44	12	47
18	10	33	45	10	42
19	16	46	46	16	78
20	16	69	47	2	6
21	10	41	48	2	6
22	6	19	49	8	40
23	2	6	50	12	39
24	6	27	51	4	9
25	10	35	52	4	22
26	12	55	53	12	41
27	4	15			

مثال ۴.۲.۱. در این مثال X به عنوان تعداد خدمه و Y به عنوان تعداد اتاق های نظافت شده در نظر گرفته شده اند. با توجه به داده ها هدف برآورد یک مدل رگرسیونی و نهایتا پیش بینی تعداد اتاق های است که می تواند توسط ۴و۱۶ خدمه تمیز گردد.

رگرسیون برازش شده به کل داده ها:

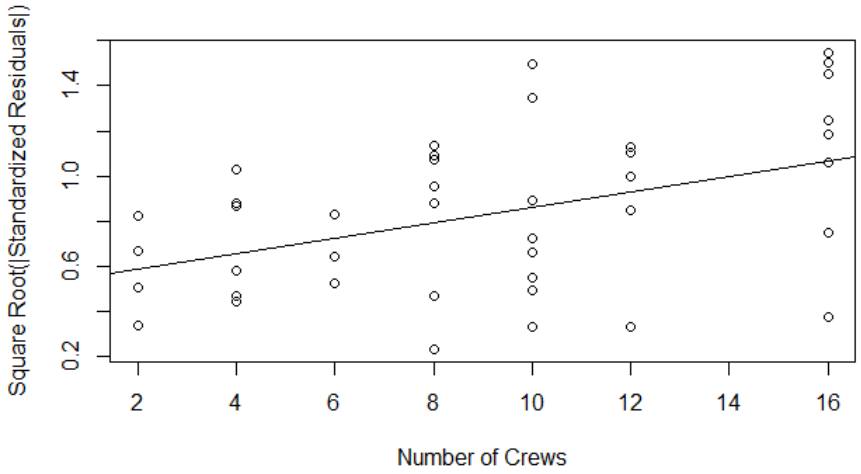
خروجی رگرسیون R

```
Call:
lm(formula = Rooms ~ Crews)

Residuals:
Min        1Q    Median        3Q       Max
-15.9990  -4.9901   0.8046   4.0010  17.0010

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7847     2.0965   0.851    0.399
Crews         3.7009     0.2118  17.472   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 7.336 on 51 degrees of freedom
Multiple R-squared:  0.8569,    Adjusted R-squared:  0.854
F-statistic: 305.3 on 1 and 51 DF,  p-value: < 2.2e-16
predict(m1,newdata=data.frame(Crews=c(4,16)),level=0.95)
       fit       lwr       upr
1 16.58827  1.58941 31.58713
2 60.99899 45.81025 76.18773
```



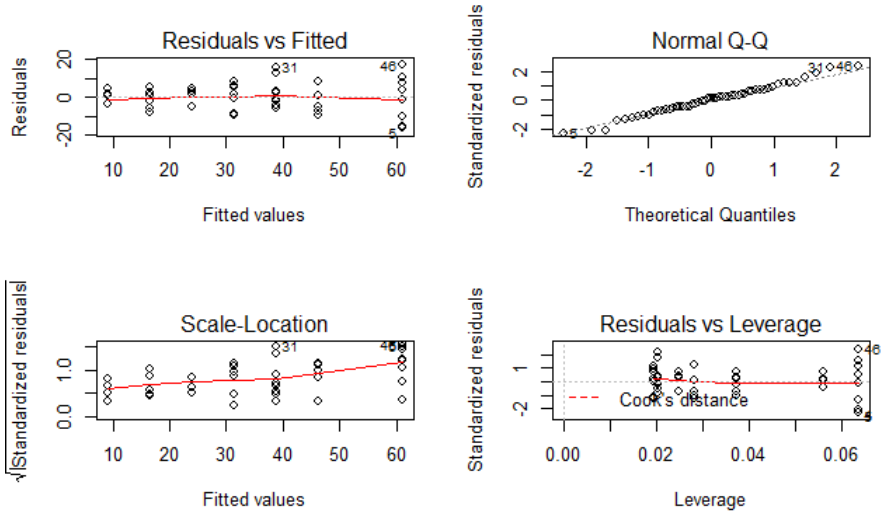
شکل ۱۶.۱: باقی مانده های استاندارد شده در مقابل X

همان طور که در شکل ۱۶.۱ بر می آید، تغییرات مانده ها ی استاندارد شده با افزایش متغیر X رو به افزایش است. بنابراین می توان نتیجه گرفت که واریانس خطاها ثابت نبوده و می توان گفت که این نمودار شکل قیفی شکل دارد. کوک در سال ۱۹۹۹ گفت که یک راه تشخیص موثر برای ثابت نبودن واریانس خطاها، رسم |مانده های استاندارد|$^{۱/۲}$ در مقابل X_i ها است. توان ۰٫۵ در واقع به خاطر کاهش مقدار مطلق ضریب چولگی داده هاست.

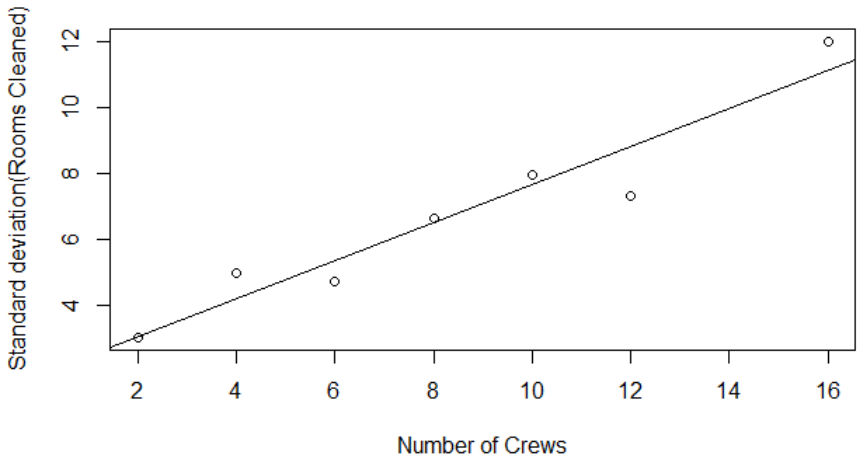


شکل ۱۷.۱: طرح تشخیصی با هدف تشخیص واریانس خطای غیر ثابت

در شکل ۱۷.۱ نمودار |مانده های استاندارد|$^{۱/۲}$ در مقابل X_i ها رسم شده است. همچنین خط کمترین مربعات برازش شده نیز برای این نمودار رسم شده است. یک روند افزایشی در این نمودار بدیهی است، که نشان دهنده افزایش واریانس نسبت به X_i هاست.



شکل ۱۸.۱: نمودار رگرسیون تشخیصی



شکل ۱۹.۱: X_iها در مقابل انحراف معیار

جدول ۶.۱: انحراف معیار Y برای هر مقدار x

Crews	N	StDev(Rooms cleaned)
2	9	3.00
4	6	4.97
6	5	4.69
8	8	6.64
10	8	7.93
12	7	7.29
16	10	12.00

بر اساس شکل ۱۸.۱،۱۹.۱ که این مقادیر در جدول ۶.۱ محاسبه شده اند ثابت نبودن واریانس و افزایش آن بر حسب X_i ها آشکار است.

۳.۱ تبدیلات

در حالت کلی استفاده از تبدیلات می تواند در موارد زیر صورت گیرد:

۱- غلبه بر مشکل ثابت نبودن واریانس جملات خطا

۲- برآورد درصد تاثیرات

۳- غلبه بر مشکل غیر خطی بودن

۱.۳.۱ استفاده از تبدیلات برای ثابت نمودن واریانس خطاها

در حالتی که واریانس جملات خطا ثابت نباشد می توان برروی متغیر پاسخ یا مستقل یا هر دو تبدیلاتی را در جهت ثابت نمودن واریانس اعمال کرد.

مثال ۱.۳.۱. برای استفاده از یک تبدیل مناسب به مثال ۴.۲.۱ که ثابت نبودن واریانس در آن آشکار شده باز می گردیم. ابتدا با بررسی متغیر وابسته که به صورت تعداد بیان شده بود

در می بابیم که یک متغیر گسسته شمارش پذیر است. معمولا متغیر های تصادفی شمارش پذیر از توزیع پواسن تبعیت می کنند.

فرض کنیم متغیر Y دارای توزیع پواسن با پارامتر λ باشد، در این صورت واریانس آن نیز برابر λ خواهد بود، در چنین مواردی مناسب ترین تبدیل برای ثابت واریانس تبدیل رادیکالی است، می توان نوشت:

$$f(Y) = f(E(Y)) + f'(E(Y))(Y - E(Y)) + \dots$$

$$\mathrm{Var}(f(y)) = [f'(E(Y))]^{۲}\mathrm{Var}(Y) = [f'(\lambda)]^{۲}\lambda = C$$

$$f'(\lambda) = \frac{c' + \sqrt{c}}{\sqrt{\lambda}} \Rightarrow f(\lambda) = \int c'\lambda^{-۱/۲}d\lambda = c''\lambda^{۱/۲} \Rightarrow \sqrt{\lambda}$$

اکنون با در نظر گرفتن $f(Y) = \sqrt{Y} = Y^{۱/۲}$ می توان نوشت:

$$\mathrm{Var}(Y^{۱/۲}) = [\frac{۱}{۲}E^{-۱/۲}(Y)]^{۲}\mathrm{Var}(Y) = (\frac{۱}{۲}\lambda^{-۱/۲})^{۲}\lambda = cte$$

حال با توجه به اینکه داده ها در هر دو بعد قابل شمارش می باشند، لذا برای هر دو متغیر X و Y از تبدیل یکسان رادیکالی استفاده می نماییم. اگر ازاین تبدیل استفاده نماییم، معادله رگرسیون برازش شده و برآورد ها در دونقطه ۴و۱۶به صورت زیر خواهد بود:

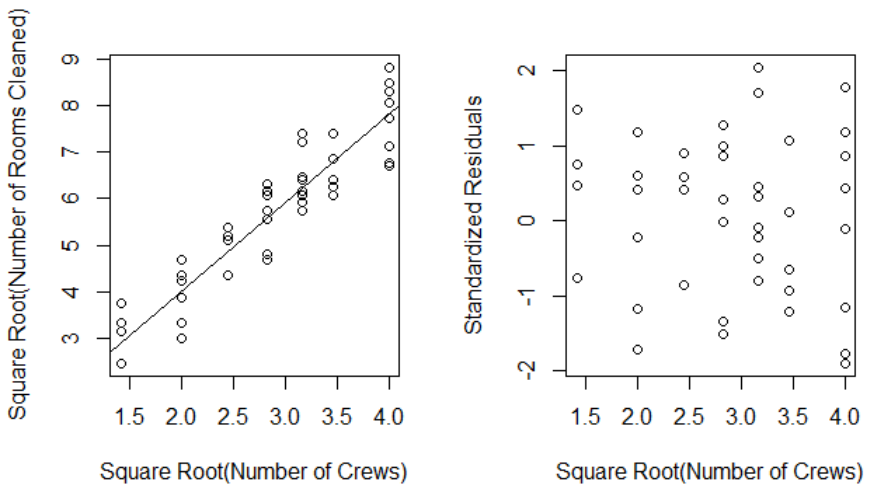
$$\sqrt{Y} = \hat{\beta}_{۰} + \hat{\beta}_{۱}\sqrt{X}$$

جدول ۷.۱: پیش بینی ها و۹۵٪ فواصل پیش بینی برای تعداد اتاق ها

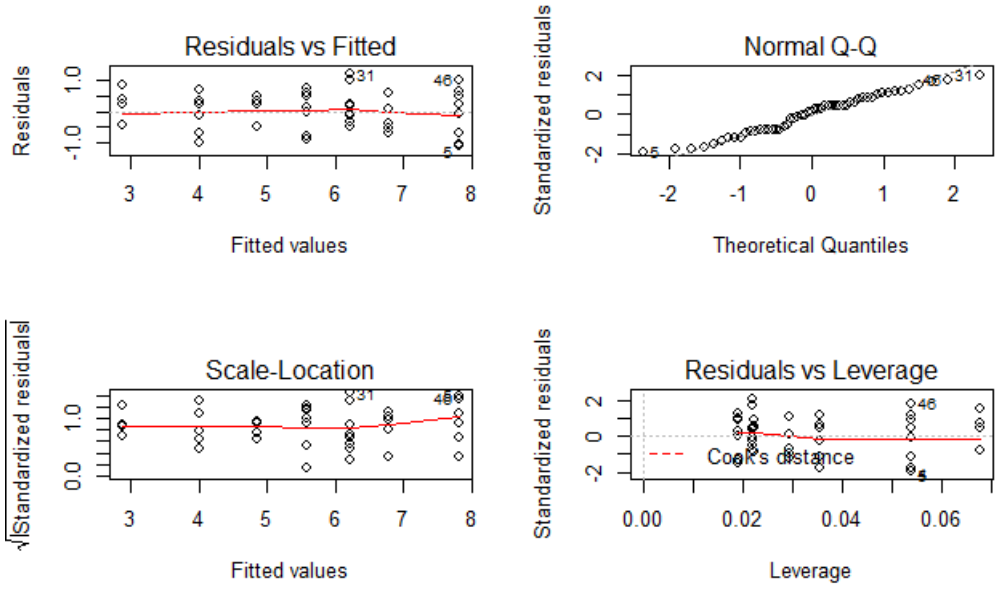
x , Crews	Prediction	Lower limit	Upper limit
4 (transformed data)	16 =(4.003 2)	8 =(2.790 2)	27=(5.217 2)
4 (raw data)	17	2	32
16 (transformed data)	61 =(7.806 2)	43 =(6.582 2)	82 =(9.031 2)
16 (raw data)	61	46	76

خروجی رگرسیون R

```
Call:
lm(formula = sqrtrooms ~ sqrtcrews)

Residuals:
Min        1Q    Median        3Q       Max
-1.09825 -0.43988  0.06826  0.42726  1.20275

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2001     0.2757   0.726    0.471
sqrtcrews     1.9016     0.0936  20.316   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.594 on 51 degrees of freedom
Multiple R-squared:   0.89,     Adjusted R-squared:  0.8879
F-statistic: 412.7 on 1 and 51 DF,  p-value: < 2.2e-16

fit       lwr       upr
1 4.003286 2.789926 5.216646
2 7.806449 6.582320 9.030578
```



شکل ۲۰.۱: مانده های استاندارد در مقابل $\sqrt{X_i}$



شکل ۲۱.۱: رگرسیون تشخیصی

در شکل ۲۰.۱ (سمت راست) تقریبا می توان گفت که واریانس ثابت شده است و شکل قیفی خود را از دست داده اند. در شکل ۲۱.۱ در قسمت پایین سمت چپ یعنی نمودار $\sqrt{|r_i|}$ در مقابل $\hat{Y}_i$ جدید ($\sqrt{\hat{Y}_i}$) دیگر روند افزایشی شدیدی بر خلاف شکل بدون تبدیل نمی توان دید. بنابراین استفاده از تبدیل رادیکالی مشکل ثابت نبودن واریانس را تا حد زیادی حل نمود.

همان طور که در شکل ۷.۱ دیده می شود، فواصل اطمینان مقادیر برازش شده ،تبدیل یافته برای X_i های کوچک پهنای کمتر و برای X_i های بزرگ پهنای بیشتری دراد،نسبت به داده های تبدیل نیافته.زیرا واریانس خطاها با افزایش متغیر X افزایش می یابد و بنابراین انتظار می رود که برای X_i های بزرگ تر فاصله اطمینان پهن تری نسبت به X_i های کوچک تر به دست آید که این موضوع بر اساس داده های تبدیل نیافته قابل مشاهده نبود و لذا فواصل اطمینان

تبدیل نیافته معتبر نمی باشد.

$$E(Y(\hat{x^*})) = \hat{Y}^* \pm t_{\frac{\alpha}{۲}}(n-۲)S\sqrt{\frac{۱}{n} + \frac{(x^* - \bar{X})^۲}{S_{xx}}}$$
$$Y(x^*) = \hat{Y}^* \pm t_{\frac{\alpha}{۲}}(n-۲)S\sqrt{۱ + \frac{۱}{n} + \frac{(x^* - \bar{X})^۲}{S_{xx}}}$$

۲.۳.۱ استفاده از تبدیلات لگاریتمی برای برآورد درصد تاثیرات

در این بخش نشان خواهیم داد که چگونه می توان با استفاده از تبدیل لگاریتمی، درصد تاثیرات مثبت متغیر ها را برآورد نمود.

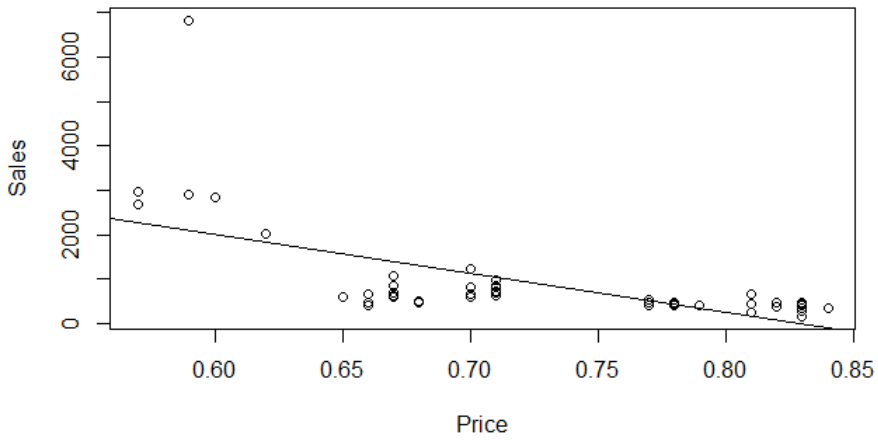
مدل رگرسیون زیر را در نظر بگیرید:

$$\log(Y) = \beta_۰ + \beta_۱\log(X) + \epsilon \qquad \log(Y) = \ln(Y)$$

$$\hat{\beta}_۱ = \frac{\Delta\log(Y)}{\Delta\log(X)} = \frac{\log(Y_۲) - \log(Y_۱)}{\log(X_۲) - \log(X_۱)} = \frac{\log(Y_۲/Y_۱)}{\log(X_۲/X_۱)}$$
$$\cong \frac{\frac{Y_۲}{Y_۱} - ۱}{\frac{X_۲}{X_۱} - ۱} = \frac{۱۰۰(\frac{Y_۲}{Y_۱} - ۱)}{۱۰۰(\frac{X_۲}{X_۱} - ۱)} = \frac{۱۰۰(\frac{Y_۲ - Y_۱}{Y_۱})}{۱۰۰(\frac{X_۲ - X_۱}{X_۱})} = \frac{٪\Delta Y}{٪\Delta X}$$

به عبارت دیگر می توان گفت که یک درصد افزایش در X منجر به $\hat{\beta}_۱٪$ کاهش یا افزایش در Y خواهد شد.

$$\Rightarrow ٪\Delta Y = \hat{\beta}_۱ \times ٪\Delta X$$

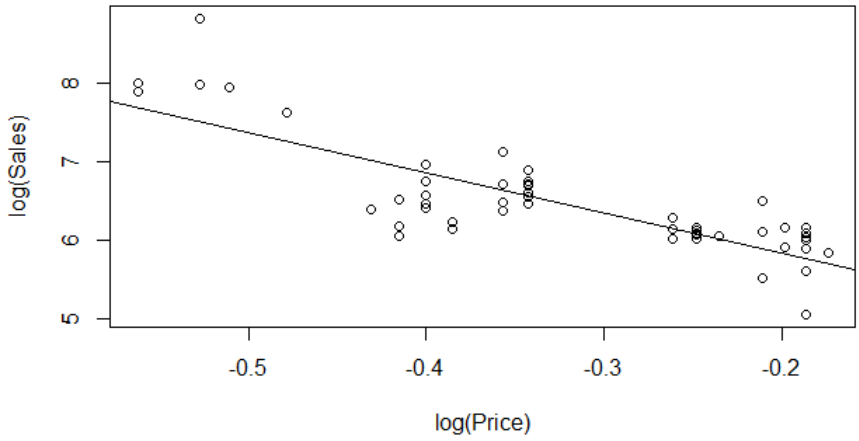


شکل ۲۲.۱: نمودار فروش در برابر قیمت

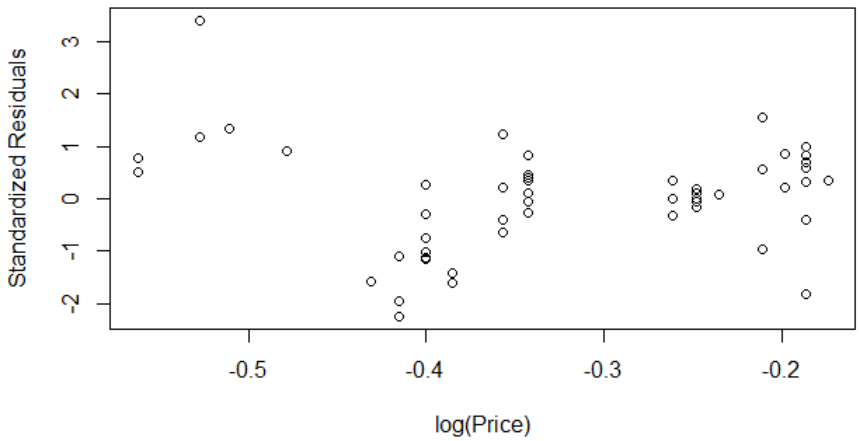
مثال ۲.۳.۱. در این مثال هدف بررسی رابطه قیمت(p) یک محصول با تعداد فروش (Q) بر حسب هزار عدد آن است. می خواهیم بدانیم (برآورد کنیم) درصد تاثیر(تغییر) روی فروش به ازای یک درصد افزایش در قیمت. نمودار پراکنش داده ها در شکل ۲۲.۱ رسم شده است. بدیهی است که براساس این شکل، یک خط راست به خوبی به داده ها برازش نخواهد شدو بر اساس برازش یک خط راست، تعداد زیادی داده پرت خواهیم داشت. اگر مدل را به داده ها برازش دهیم:

$$\log(Q) = \beta_0 + \beta_1 \log(p)$$

مجددا اگر $\log(P)$ را در مقابل $\log(Q)$ رسم کنیم در شکل ۲۳.۱ حاصل می شود.



شکل ۲۳.۱: نمودار فروش در برابرقیمت



شکل ۲۴.۱: نمودار باقیمانده های استاندارد شده در برابر قیمت

مدل رگرسیون فوق به صورت زیر به داده ها برازش می شود:

خروجی رگرسیون R

```
Call:
lm(formula = log(Sales) ~ log(Price))

Residuals:
Min        1Q    Median        3Q       Max
−0.88973 −0.18188  0.04025  0.22087  1.31026

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8029     0.1744   27.53  < 2e−16 ***
log(Price)   −5.1477     0.5098  −10.10 1.16e−13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 0.4013 on 50 degrees of freedom
Multiple R-squared:  0.671,     Adjusted R-squared:  0.6644
F-statistic:   102 on 1 and 50 DF,  p-value: 1.159e−13
```

براساس آنچه دیدیم، $\hat{\beta}_1$ در واقع برابر است با درصد تغییرات در Q و P .یعنی به ازای ۱٪ افزایش در قیمت، ۵/۱ درصد کاهش در فروش حاصل تقاضا خواهد شد.از آنجا که درآمد برابر است با قیمت ضرب در تعداد فروش، بنابراین افزایش قیمت بسیار به ضرر فروشنده خواهد بود. زیرا تغییرات در فروش تقریبا ۵ برابر تغییرات در قیمت است.

۳.۳.۱ استفاده از تبدیلات برای غلبه بر غیر خطی بودن

در حالت کلی، دوراه برای غلبه بر مشکل غیر خطی بودن وجود دارد:

۱- نمودار های پاسخ معکوس

۲- تبدیلات باکس-کاکس

در هر دو صورت ممکن است یکی از سه حالت زیر اتفاق بیفتد:

(آ) فقط متغیر وابسته نیاز به تبدیل داشته باشد

(ب) فقط متغیر مستقل نیاز به تبدیل داشته باشد

(ج) هردو متغیر نیاز به تبدیل داشته باشند.

استفاده از تبدیل رگرسیون معکوس بر روی متغیر وابسته

فرض کنید رابطه بین X و Y به صورت زیر است:

$$Y = g(\beta_۰ + \beta_۱ X + \varepsilon)$$

به طوری که $g(۰)$ یک تابع نامعلوم باشد. می توان نوشت:

$$g^{-۱}(Y) = \beta_۰ + \beta_۱ X + \varepsilon$$

مثلا:

$$Y = (\beta_۰ + \beta_۱ X + \varepsilon)^۳ \Rightarrow \sqrt[۳]{Y} = \beta_۰ + \beta_۱ X + \varepsilon \Rightarrow g^{-۱}(Y) = \sqrt[۳]{Y}$$

$$Y = \exp\beta_۰ + \beta_۱ X + \varepsilon \Rightarrow \log(Y) = \beta_۰ + \beta_۱ X + \varepsilon \Rightarrow g^{-۱}(Y) = \log(Y)$$

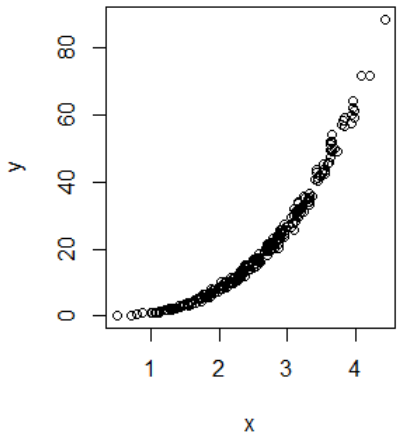
مثال شبیه سازی شده

فرض کنید ۲۵۰ داده از مدل زیر شبیه سازی شده اند:

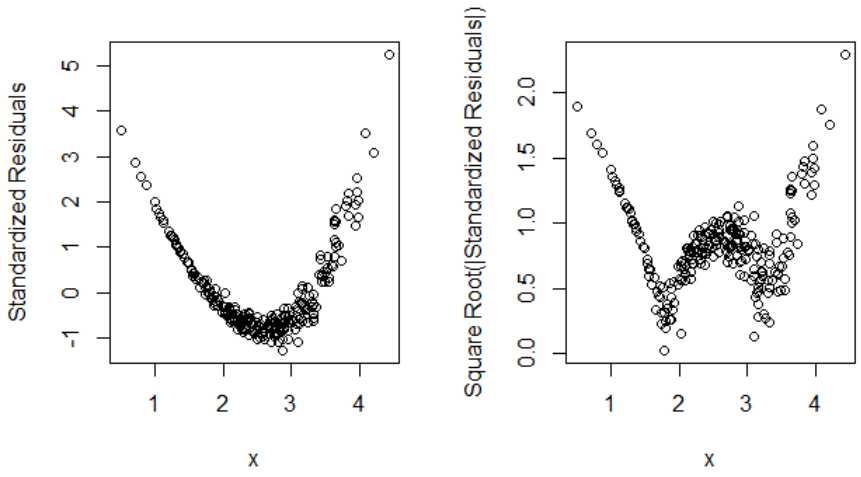
$$Y = (\beta_۰ + \beta_۱ X + \varepsilon)^۳ \qquad , X \sim^{i.i.d} N \qquad \varepsilon \sim^{i.i.d} N$$

هدف برآورد تابع $g^{-1}(Y) = Y^{1/3}$ است. نمودار پراکنش Y و X در شکل ۲۵.۱ رسم شده اند. ابتدا بابرازش خط راست شروع می نماییم:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (۲.۱)$$



شکل ۲۵.۱: نمودار Y در مقابل x برای داده های تولید شده



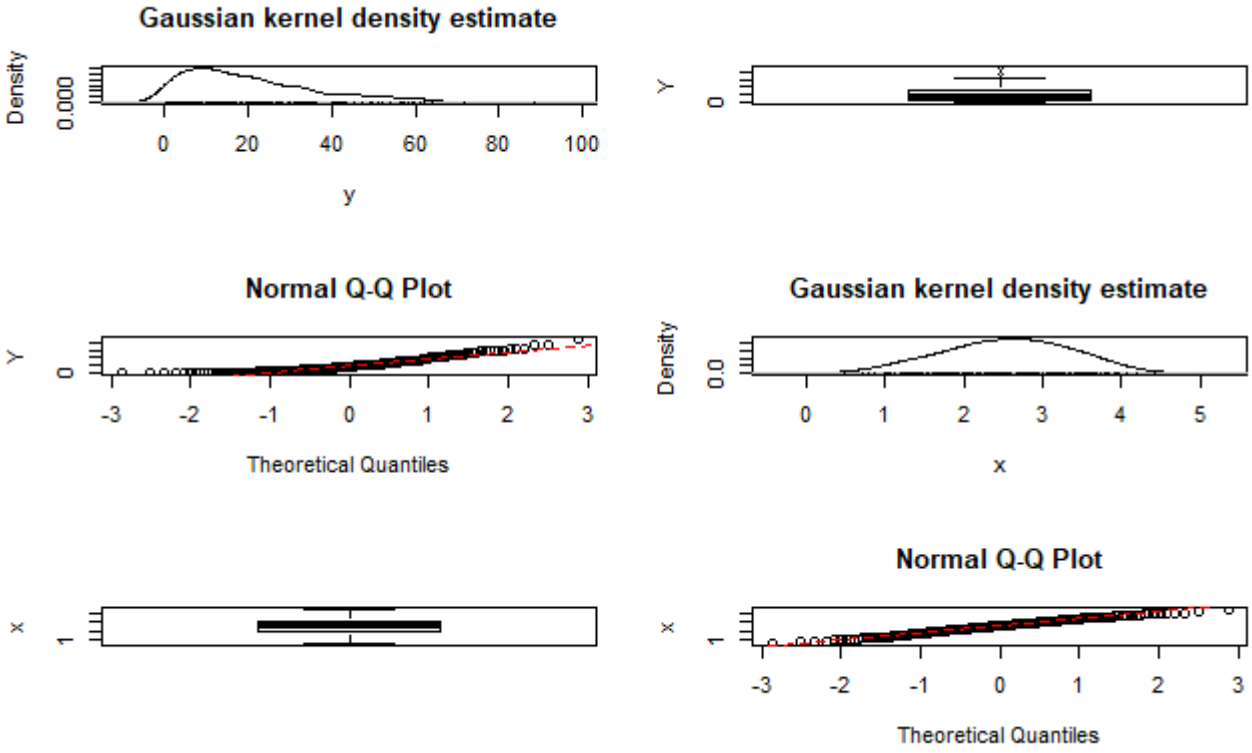
شکل ۲۶.۱: نمودارهای تشخیصی

در شکل ۲۶.۱ نمودار پراکنش مانده های استاندارد شده در مقابل X و جزر قدر مطلق مانده های استاندارد شده در مقابل X رسم شده است. به طوری که اولی نشان دهنده رابطه غیر خطی بین متغیر وابسته و مستقل بوده (عدم کفایت مدل) و دمی نشان دهنده این است که واریانس مانده های استاندارد شده ثابت نمی باشد لذا واریانس جملات خطا در مدل خخطی ثابت نیست. در چنین مواردی طبیعی است که باید تبدیلات را روی Y یا X یا هر دو بررسی نماییم. بدین منظور می توان شکل توزیع Y و X را بررسی نمود. لذا نمودار های جعبه ای و $Q-Q$ برآورد های توابع چگالی را برای دو متغیر رسم می نماییم. در شکل ۲۷.۱ این نمودار ها رسم شده اند و بر اساس آن ها در می یابیم که متغیر Y دارای توزیعی چاوله به راست است و بهتر است که تبدیل را روی آن به تنهایی به کار بریم زیرا توزیع X متقارن است. البته نمودار برآورد توابع چگالی بر اساس روش رگرسیون ناپارامتری کرتل انجام و رسم شده است.اکنون

مدل را به شکل زیر در نظر بگیرید:

$$Y = g(\beta_0 + \beta_1 X + \varepsilon) \Rightarrow g^{-1}(Y) = \beta_0 + \beta_1 X + \varepsilon$$

اگر β_0 و β_1 معلوم باشند، می توان شکل تابع $g^{-1}(\circ)$ را بر اساس نمودار پراکنش Y در مقابل $\beta_0 + \beta_1 X$ مشخص نمود.



شکل ۲۷.۱: نمودارهای جعبه، نمودارهای $Q-Q$ معمولی و تخمین تراکم هسته Yو x

در شکل ۲۸،نمودار پراکنش $\hat{Y}$ (مدل برازش شده در رگرسیون خطی ساده) در مقابل Y رسم شده و برای مقادیر $\lambda = ۱, ۱/۳, ۱$ منحنی $\hat{Y} = Y^{\lambda}$ برازش شده است. همانطور که از این شکل بر می آید مدل $\hat{Y} = \hat{\beta}_۰ + \hat{\beta}_۱ X = Y^{۱/۳}$ به بهترین نحو به داده ها برازش پیدا می کند که البته این موضوع طبیعی است زیرا داده ها از مدل $Y = (\hat{\beta}_۰ + \hat{\beta}_۱ X + \varepsilon)^۳$ شبیه سازی شده اند. به این روش نمودار پاسخ معکوس نیز گفته می شود زیرا برای رسم نمودار پراکنش، متغیر Y روی محور x ها و متغیر $\hat{Y}$ روی محور y ها در نظر گرفته می شود. [۱]

انتخاب تبدیلات توانی

به منظور برآورد تابع $g^{-۱}(۰)$ در حالت کلی خانواده تبدیلات توانی مقیاسی زیر را که روی متغیر مطلقا مثبت Y تغریف می شوند، مورد بررسی قرار می دهیم:

$$\Psi_s(Y, \lambda) = \begin{cases} \dfrac{Y^{\lambda} - ۱}{\lambda} & \lambda \neq ۰ \\ \log(Y) & \lambda = ۰ \end{cases}$$

این خانواده دارای خواص زیر است:

۱- $\psi_s(Y, \lambda)$ تابعی پیوسته بر حسب λ است.

۲- تبدیل لگاریتمی نیز عضوی از این خانواده است، زیرا:

$$\lim_{\lambda \to ۰} \psi_s(Y, \lambda) = \lim_{\lambda \to ۰} \frac{Y^{\lambda} - ۱}{\lambda} = \lim_{\lambda \to ۰} \frac{e^{\lambda \log(Y)} - ۱}{\lambda} = \lim_{\lambda \to ۰} \frac{\log(Y) e^{\lambda \log(Y)}}{۱}$$

۳- این خانواده از تبدیلات حفظ کننده جهت تغییرات بین X و Y است یعنی اگر X و Y رابطه مثبت(منفی) باهم داشته باشند آنگاه X و $\psi_s(Y, \lambda)$ هم دارای رابطه مثبت(منفی) خواهند بود.

[1] Inverse response plot

بنابراین به منظور برآورد تابع $g^{-1}(\circ)$ ، مدل های شکل زیر را مورد بررسی قرار می دهیم:

$$E(\hat{Y}|Y=y) = \alpha_\circ + \alpha_1 \psi_s(Y,\lambda) \qquad (۳.۱)$$

برای مقدار معلوم λ ، مدل فوق یک مدل رگرسیون خطی ساده ، با متغیر مستقل $\psi_s(Y,\lambda)$ است و متغیر وابسته نیز $\hat{Y}$ (مقدار برازش شده به مدل رگرسیون خطی ساده با روش کمترین مربعات) خواهد بود.

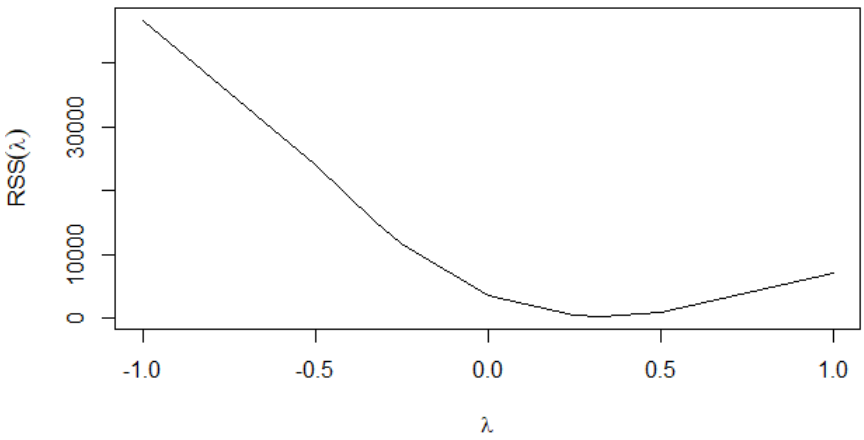
بنابراین باید مدل ۳.۱ را به کمک روش کمترین مربعات برازش داده و مقادیری از λ بهینه خواهند بود که SSE را مینیمم نمایند. معمولا فقط مقادیری از λ جهت انتخاب λ ی بهینه آزمایش می شوند که متعلق به مجموعه زیر باشند:

$$\lambda \in \{-1, -\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1\}$$

ادامه مثال شبیه سازی شده

در شکل۲۹ نمودار $SSE(\lambda)$ بر حسب λ رسم شده است. همان طور که از این نمودار بر می آید، λ ی بهینه بین صفر و ۰٫۵ قرار گرفته است و به طور دقیق تر می توان گفت که $\hat{\lambda}_{op+} = ۰٫۳۳۲$. بنابراین $g^{-1}(Y) = Y^{۰٫۳۳۲} \cong Y^{\frac{1}{3}}$.

یک روش جانشین دیگر این است که به طور هم زمان پارامتر های $\alpha_\circ$ و α_1 و λ در مدل ۳.۱ به روش کمترین مربعات برآورد شده و مقدار بهینه λ مشخص گردد. این کار می تواند توسط تابع $inverse.response.plot$ در کتابخانه $alr۳$ نرم افزار R انجام گردد.



شکل ۲۸.۱: نمودار $RSS(l)$ در برابر l برای مجموعه داده های تولید شده

تبدیلات باکس-کاکس روی متغیر وابسته Y

باکس و کاکس در سال ۱۹۶۴ یک روش کلی از تبدیلات را بر روی متغیر وابسته اکیدا مثبت مورد بررسی قرار دادند. همان طور که می بینیم، این خانواده از تبدیلات می توانند به طور همزمان بر روی متغیر یا متغیر های مستقل نیز به کار برده شوند. روش باکس-کاکس در واقع نزدیک کننده توزیع متغیر تبدیل یافته به توزیع نرمال است. قبل از معرفی این تبدیلات، حالتی را بررسی می نماییم که X و Y دارای توزیع نرمال دو متغیره باشند.

رگرسیون خطی ساده وقتی X و Y هر دو دارای توزیع نرمالند:

فرض کنید:

$$X_i, Y_i \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

طبق خخواص توزیع نرمال دو متغیره داریم:

$$Y_i|X_i = x_i \sim N(\mu_y - \rho_{xy}\frac{\sigma_y}{\sigma_x}\mu_x + \rho_{xy}\frac{\sigma_y}{\sigma_x}x_i, \sigma_y^{۲}(۱ - \rho_{xy}^{۲}))$$

$$\Rightarrow Y_i|X_i = x_i \sim N(\beta_۰ + \beta_۱ x_i, \sigma^{۲})$$

به طوریکه:

$$\beta_۰ = \mu_y - \beta_۱\mu_x \qquad , \beta_۱ = \rho_{xy}\frac{\sigma_y}{\sigma_x} \qquad , \sigma^{۲} = \sigma_y^{۲}(۱ - \rho_{xy}^{۲})$$

$$\Rightarrow E(Y|X = x_i) = \beta_۰ + \beta_۱ x_i$$

تبدیلات باکس-کاکس در واقع تبدیلی را به دست می آورند که توزیع X و Y رابه سمت توزیع نرمال دو متغیره تبدیل نموده و نتیجتا مدل رگرسیون Y روی X خطی گردد. این روش براساس رویکرد درستنمایی بوده و لذا در ادامه مروری بر اصل درستنمایی می اندازیم.

روش درستنمایی ماکزیمم در برآورد مدل رگرسیون خطی هرگاه X و Y نرمال باشند

همان طور که در قسمت قبل دیدیم هرگاه X و Y دارای توزیع نرمال توام باشند آنگاه:

$$Y_i|X_i = x_i \sim N(\beta_۰ + \beta_۱ x_i, \sigma^{۲}) \Rightarrow f(y_i|x_i) = \frac{۱}{\sigma\sqrt{۲\pi}}\exp\{-\frac{(y_i - \beta_۰ - \beta_۱ x_i)^{۲}}{۲\sigma^{۲}}\}$$

$$\Rightarrow L(\beta_۰, \beta_۱, \sigma^{۲}|\underline{Y}) = \prod_{i=۱}^{n} f(y_i|x_i) = (\frac{۱}{\sigma\sqrt{۲\pi}})^{n}\exp\{-\frac{۱}{۲\sigma^{۲}}\sum_{i=۱}^{n}(y_i - \beta_۰ - \beta_۱ x_i)^{۲}\}$$

$$\Rightarrow L(\beta_۰, \beta_۱, \sigma^{۲}|\underline{Y}) = \log L = -\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\sigma^{۲}) - \frac{۱}{۲\sigma^{۲}}\sum_{i=۱}^{n}(y_i - \beta_۰ - \beta_۱ x_i)^{۲} \quad (۴.۱)$$

برآورد های ML:

$$\begin{cases} \frac{\partial l}{\partial \beta_۰} = ۰ & \Rightarrow \\ \frac{\partial l}{\partial \beta_۱} = ۰ & \Rightarrow \\ \frac{\partial l}{\partial \sigma^۲} = ۰ & \Rightarrow \end{cases} \quad \begin{cases} \hat{\beta}_۰ = \bar{Y} - \hat{\beta}_۱ \bar{X} \\ \hat{\beta}_۱ = \frac{S_{xy}}{S_{xx}} \\ \hat{\sigma^۲} = \frac{۱}{n} \sum_{i=۱}^{n} (Y_i - \hat{\beta}_۰ - \hat{\beta}_۱ X_i) = \frac{SSE}{n} \end{cases}$$

و بنابراین فقط برآورد $\sigma^۲$ با روش کمترین مربعات فرق دارد.

$$\hat{\sigma^۲}_{ls} = \frac{SSE}{n - ۲}$$

تبدیلات باکس-کاکس بر روی متغیر وابسته

باکس و کاکس در سال ۱۹۶۴ خانواده بهبود یافته از تبدیلات توانی زیر را مورد بررسی قرار دادند:

$$\Psi_M(Y, \lambda) = \Psi_M(Y, \lambda) \times gm(Y)^{۱-\lambda} = \begin{cases} gm(Y)^{۱-\lambda} \times \frac{Y^\lambda - ۱}{\lambda} & \lambda \neq ۰ \\ gm(Y) \times \log(Y) & \lambda = ۰ \end{cases}$$

به طوری که:

$$gm(Y) = \sqrt[n]{\prod_{i=۱}^{n} Y_i} = \exp\{\frac{۱}{n} \sum_{i=۱}^{n} \log(Y_i)\}$$

روش باکس-کاکس در واقع تبدیل را روی متغیر Y اعمال می کند که توزیع آن را به توزیع نرمال نزدیک نماید، یعنی $\Psi_M(Y, \lambda)$ دارای توزیع نرمال است. بنابراین برآورد به روش درستنمایی ماکزیمم به صورت زیر خواهمد بود:

$$\log\{l(\beta_۰, \beta_۱, \sigma^۲, \lambda | \Psi_M(Y, \lambda))\} = -\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\sigma^۲) - \frac{۱}{۲\sigma^۲} \sum_{i=۱}^{n} (\psi(Y_i, \lambda) - \beta_۰ - \beta_۱ x_i)^۲$$

حال باتوجه به اینکه لگاریتم تابع درستنمایی جدید همان تابع ۴.۱ در قسمت قبل است و فقط فرق آن این است که Y_i ها با $\Psi_M(y_i, \lambda)$ جایگزین شده اند و چون ژاکوبین تبدیل برابر است با یک، لذا برآورد گرهای $\sigma^۲, \beta_۱, \beta_۰$ همان قبلی بوده و باید فقط λ برآور شود. داریم:

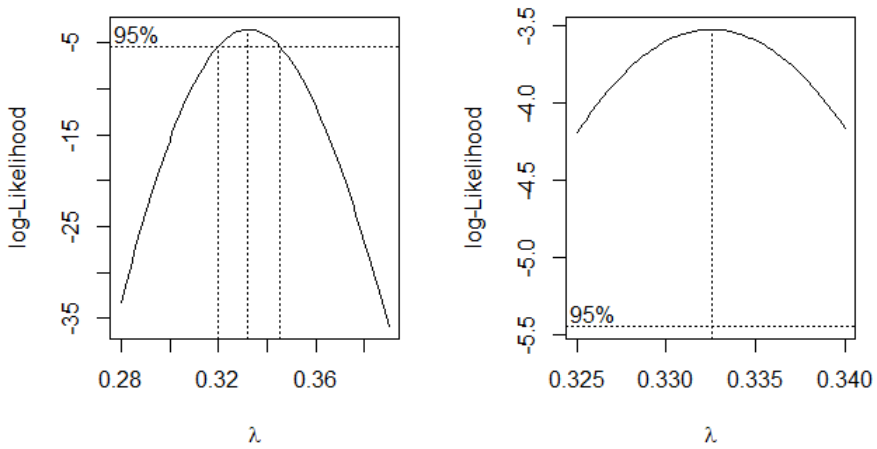
$$\log\{l(\beta_۰, \beta_۱, \sigma^۲, \lambda|\Psi_M(Y, \lambda))\} = -\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\frac{SSE(\lambda)}{n}) - \frac{۱}{۲SSE(\lambda)/n}SSE(\lambda)$$
$$= -\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\frac{SSE(\lambda)}{n}) - \frac{n}{۲}$$

به طوری که

$$SSE(\lambda) = \sum_{i=۱}^{n}(\psi(y_i, \lambda) - \hat{\beta}_۰ - \hat{\beta}_۱ x_i)^۲$$

حال چون ماکزیمم سازی تابع درستنمایی معادل است با مینیمم سازی $SSE(\lambda)$ ، لذا مقدار برآوردگر درستنمایی ماکزیمم بهینه λ با مینیمم سازی $SSE(\lambda)$ حاصل خواهد شد.

پیاده سازی روش باکس کاکس در مثال شبیه سازی شده



شکل ۲۹.۱: احتمال ورود به سیستم برای روش تبدیل $Box - Cox$

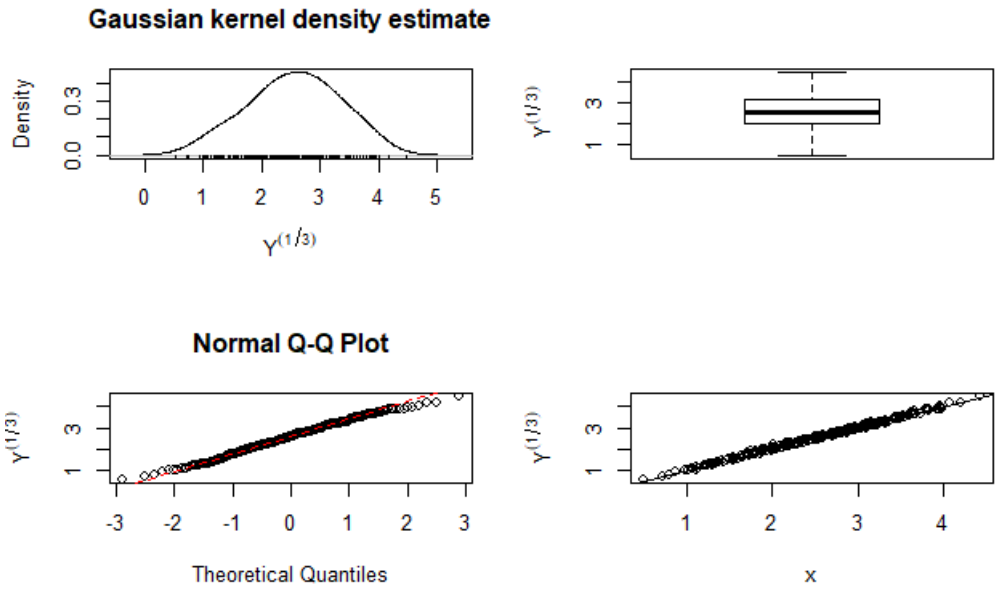
در شکل ۲۹.۱ لگاریتم تابع درستنمایی در مقابل λ رسم شده است. مقدار بهینه λ و همچنین فاصله اطمینان ۹۵٪ برای λ در این شکل رسم شده است. با رسم این شکل در مقیاس بزرگتر در می یابیم که مقدار بهینه λ برابر است با ۰٫۳۳۳ یعنی تبدیل مناسب برابر است با:

$$g^{-1}(Y) = Y^{1/3}$$

خروجی رگرسیون R

```
Call:
lm(formula = ty ~ x)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.008947 0.011152 0.802 0.423
x 0.996451 0.004186 238.058 <2e-16 ***
---
```

```
Residual standard error: 0.05168 on 248 degrees of freedom
Multiple R-Squared: 0.9956, Adjusted R-squared: 0.9956
F-statistic: 5.667e+04 on 1 and 248 DF, p-value: < 2.2e-16
```



شکل ۳۰.۱: نمودارهای جعبه، نمودارهای $Q-Q$ معمولی و تخمین تراکم هسته $Y^{۱/۳}$

از شکل ۳۰.۱ گواه بر نرمال بودن توزیع $Y^{۱/۳}$ می باشند.همچنین در نمودار پراکنش $Y^{۱/۳}$ در مقابل X به وضوح رابطه خطی بین متغیر های مستقل و وابسته دیده میشود.

اعمال تبدیل صرفا بر روی متغیر مستقل

همانند قسمت های قبل، می توان خانواده تبدیلات مقیاسی را در مورد متغیر اکیدا مثبت X به شکل زیر تعریف نمود:

$$\psi_s(X,\lambda)=\begin{cases}\dfrac{X^{\lambda}-۱}{\lambda} & \lambda\neq ۰\\ \log(X) & \lambda=۰\end{cases}$$

بنابراین مدل مورد نظر به صورت زیر خواهد بود:

$$E(Y|X=x) = \alpha_۰ + \alpha_۱\psi_s(x,\lambda) \qquad (۵.۱)$$

مدل فوق برای مقدار معلوم λ ،یک مدل رگرسیون خطی ساده خواهد بود به طوریکه متغیر وابسته Y ولی متغیر مستقل $\psi_s(x,\lambda)$ است. همانند گذشته، می توان به کمک روش کمترین مربعات و مینیمم سازی $SSE(\lambda) = \sum_{i=۱}^{n}(Y_i - X_۰ - \alpha_۱\psi(x_i,\lambda))^۲$ مقدار بهینه λ را برآورد نموده و تبدیل مناسب را پیشنهاد نمود.

همچنین، می توان از تبدیل باکس -کاکس استفاده نمود و توزیع متغیر X را نرمال نمود. در این روش نیازی به انجام رگرسیون خطی نبوده و این روش به طور مستقیم بر روی X اعمال خواهد شد . البته قابل توجه است که ممکن است با استفاده از هیچ تبدیلی نتوان یک مدل رگرسیون خطی ساده معنی دار برای Y روی X برازش داد (چه تبدیل روی X انجام گردد یاY یا هردو) این حالت زمانی اتفاق می افتد که متغیر یا متغیر های مستقلی که باید در مدل رگرسیون مورد نظر قرار گیرند، به مدل وارد نشده باشند. همچنین ممکن است که تبدیل باکس-کاکس نتواند توزیع متغیر تبدیل یافته را به نرمال تبدیل نماید.

اعمال تبدیل بر روی متغیر پاسخ و متغیر(های) مستقل

هرگاه توزیع هر دو متغیر X و Y بسیار چاوله باشد و هر دو نیاز به تبدیل داشته باشند، می توان از یکی از دو رویکرد زیر استفاده نمود:

رویکرد۱

(آ) تبدیل را روی متغیر X اعمال نماییم به طوریکه متغیر تبدیل یافته یعنی $\psi_s(x,\lambda_x)$

دارای توزیعی تقریبا نرمال شود. این کار با استفاده از تبدیل یک متغیره امکان پذیر است.

(ب) با در نظر گرفتن متغیر تبدیل یافته به عنوان متغیر جدید، مدل رگرسیون خطی به شکل $Y = g(\beta_۰ + \beta_۱ \psi_s(x, \lambda_x) + \varepsilon)$ در نظر گرفته شده و با استفاده از نمودار پاسخ معکوس تبدیل مناسب برای برآورد $g^{-۱}(۰)$ استخراج گردد.

رویکرد ۲

با استفاده از تبدیل تعمیم یافته چند متغیره باکس-کاکس، به طور همزمان توزیع توام (X, Y) به توزیع نرمال دو متغیره تبدیل گردد. این رویکرد در ادامه بررسی خواهد شد.

تعمیم چند متغیره روش تبدیلات باکس-کاکس

ولیلا $(velilla)$ در سال ۱۹۹۳ با پیشنهاد این روش که توسیعی برای روش یک متغیره باکس-کاکس بود، به طور هم زمان توزیع دو یا چند متغیر را به توزیع نرمال چند متغیره تبدیل نمود. در این بخش، تنها دو متغیر X و Y را مورد بررسی قرار داده و خانواده تبدیلات بهبود یافته دو متغیره را به صورت زیر تعریف می کنیم:

$$(\psi_M(X, \lambda_x), \psi_M(Y, \lambda_Y))$$

$$\psi_M(Y, \lambda_Y) = \psi_s(Y, \lambda_Y) gm(Y)^{۱-\lambda_Y} = \begin{cases} gm(Y)^{۱-\lambda_Y} \dfrac{Y_Y^{\lambda} - ۱}{\lambda_Y} & \lambda_Y \neq ۰ \\ gm(Y)\log(Y) & \lambda_Y = ۰ \end{cases}$$

به طور مشابه $\psi_M(X, \lambda_x)$ قابل تعریف است. در ادامه به منظور برآورد λ_Y و λ_x، تابع درستنمایی $(\psi_M(Y, \lambda_Y), \psi_M(X, \lambda_x))$ را تشکیل داده و با ماکزیمم نمودن آن مقادیر λ_Y و λ_x برآورد خواهند شد.

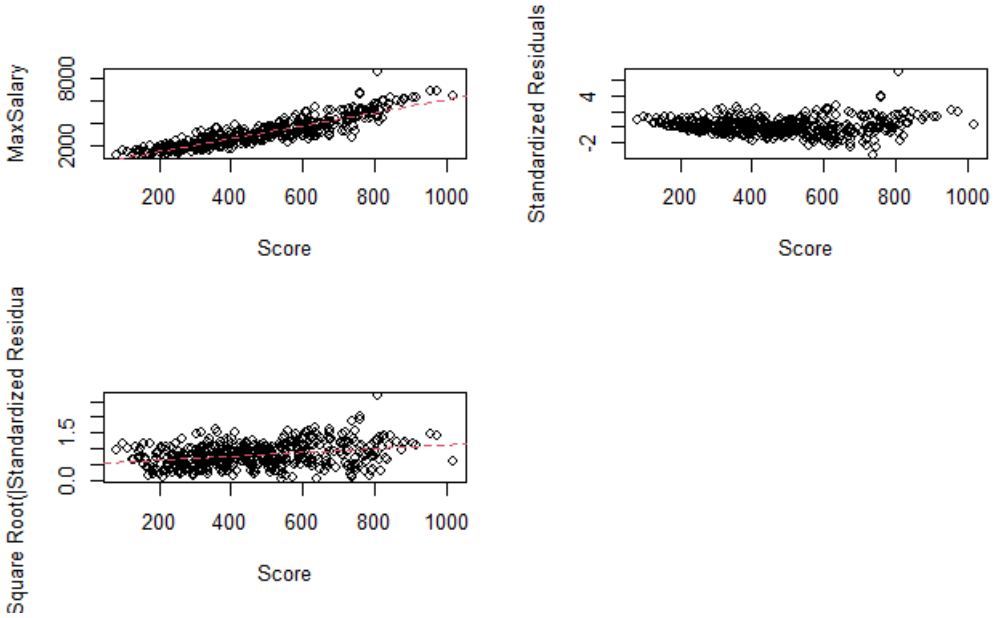
مثال: داده های مربوط به حقوق دولت

در این مثال ۴۹۵ داده مربوط به ماکزیمم حقوق شغل های غیر اتحادیه ای در بخش دولتی نیمه غربی آورده شده است. داده ها در پکیج $alr۴$ به نام $Salarygov.txt$ ذخیره شده اند. در ادامه مدل رگرسیون را جهت پیش بینی حقوق به داده ها برازش می دهیم.

برازش را با مدل $Maxsalary = \beta_۰ + \beta_۱ Score + \varepsilon$ شروع می نماییم به طوریکه

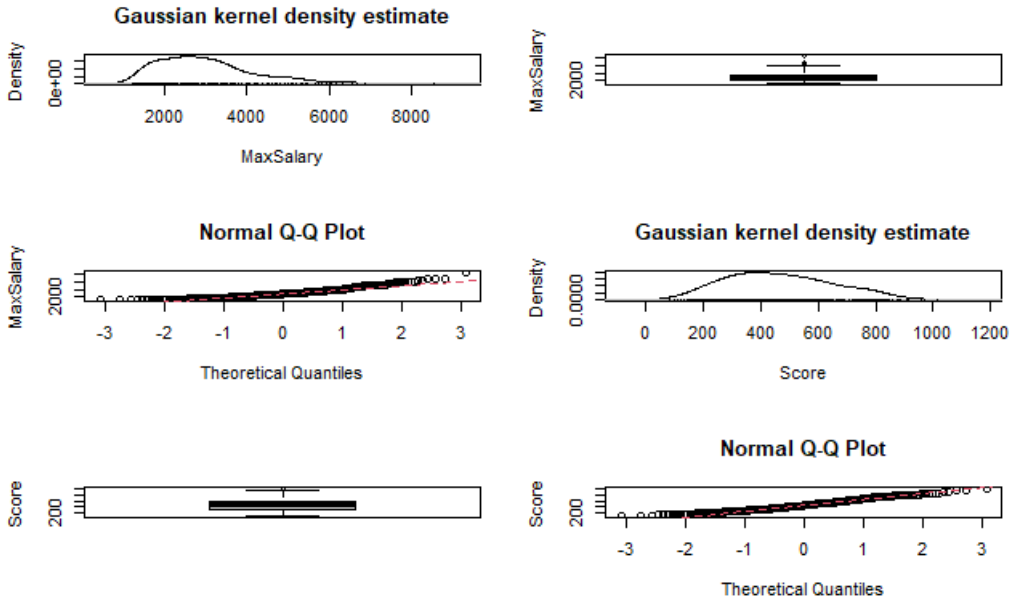
$Score$: متغیر مستقل بوده و نمره کلاس شغلی بر اساس سختی-سطح مهارت-آموزش های مورد نیاز و میزان مسئولیت پذیری که با یک مشاوره بخش دولتی اداره می شود.

$Maxsalary$: ماکزیمم حقوق داده شده به کارمندان یک کلاس شغلی



شکل ۳۱.۱: نمودارهای مرتبط با یک مدل خط مستقیم به داده های تبدیل نشده

در شکل ۳۱.۱ ،نمودار پراکنش داده ها و مدل برازش شده، مانده های استاندارد در مقابل $Score$ ، $\sqrt{|r_i|}$ در مقابل $Score$ رسم شده است. همان طور که از این شکل بر می آید، نامناسب بودن مدل و نا ثابت بودن واریانس، بدیهی است. بنابراین از تبدیل روی متغیر (ها) استفاده می نماییم.



شکل ۳۲.۱: نمودارهای داده های تبدیل نشده

برای به کار بردن تبدیل، ابتدا به شکل توزیع دو متغیر وابسته و مستقل نگاه می کنیم. (۳۲.۱) در این شکل نمودار های جعبه ای، برآورد تابع چگالی و $Q-Q$ برای هر دو متغیر رسم شده است. همان طور که از این شکل بر می آید، توزیع هیچ یک از متغیر ها نرمال نبوده و چوله استو لذا باید از تبدیل روی هر دو استفاده نماییم.

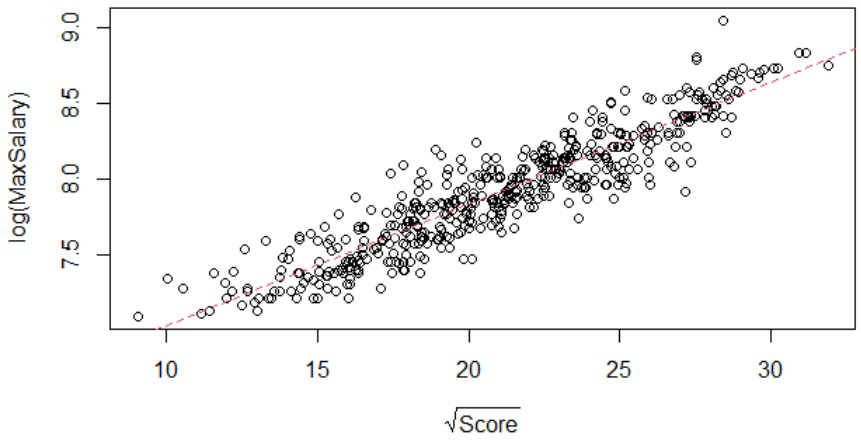
استفاده از رویکرد۲:با استفاده از دستور $bctrans$ در کتابخانه $alr۴$ می توان تبدیل باکس-

کاکس دو متغیره را بر روی X و Y اعمال نمود.

خروجی رگرسیون R

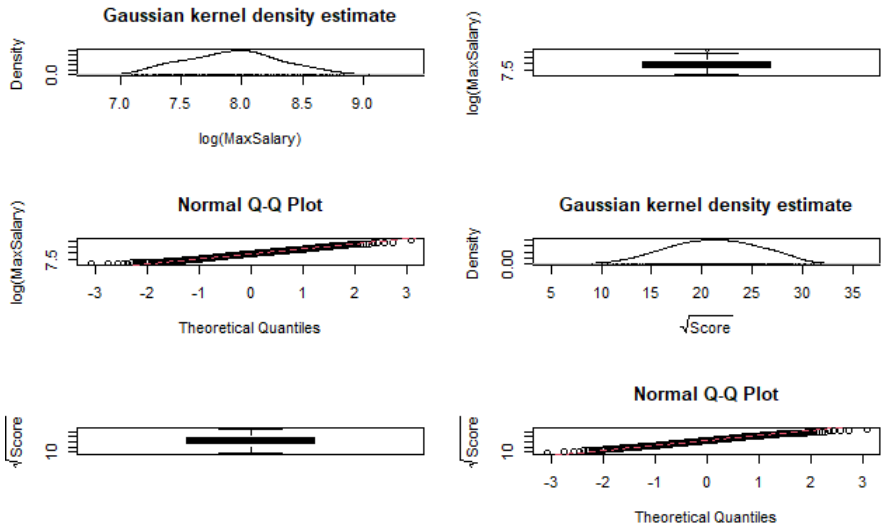
```
box.cox Transformations to Multinormality
Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
MaxSalary −0.0973 0.0770 −1.2627 −14.2428
Score 0.5974 0.0691 8.6405 −5.8240
LRT df p.value
LR test, all lambda equal 0 125.0901 2 0
LR test, all lambda equal 1 211.0704 2 0
```

مقادیر بهینه برای λ_Y و λ_x به ترتیب برابر صفر و ۰٫۵ به دست آمده اند. بنابراین باید $\log(Maxsalary)$ بر روی $\sqrt{Score}$ رگرسیون گردد.



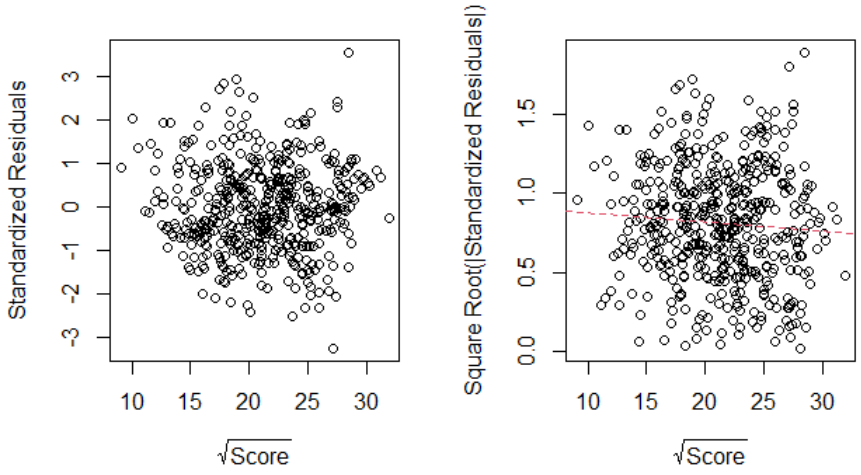
شکل ۳۳.۱: نمودار $(MaxSalary)$ و امتیاز با خط کمترین مربع اضافه شده

در شکل ۳۳.۱ نمودار پراکنش $\log(Maxsalary)$ بر روی $\sqrt{Score}$ رسم شده و خط کمترین مربعات بر آن برازش گردیده است. همان طور که از این شکل بر می آید، شواهد نشان دهنده رابطه خطی بهتری برای متغیر های تبدیل یافته است.



شکل ۳۴.۱: نمودارهای داده های تبدیل شده

در شکل ۳۴.۱ مجددا نمودار های جعبه ای، برآورد توابع چگالی و $Q-Q$ برای هر دو متغیر رسم شده که می توان فهمید توزیع متغیر های مربوط به نحو معنی داری به نرمال نزدیک شده است و چاولگی آن ها از بین رفته است.



شکل ۳۵.۱: نمودارهای تشخیصی از مدل بر اساس داده های تبدیل شده

در شکل ۳۵.۱ نموداری مربوط به عیب یابی رگرسیون رسم شده که نشان دهنده استاندارد بودن رفتار مانده ها می باشند.

رویکرد ۱: در این قسمت رویکرد ۱ را مورد بررسی قرار می دهیم، یعنی ابتدا متغیر $Score$ را تحت تبدیل باکس-کاکس معرفی شده به سمت متغیر نرمال نزدیک می نماییم و سپس با ماکزیمم نمودن لگاریتم تابع درستنمایی مقدار بهینه λ_x را برآورد می نماییم. پس از آن با تبدیل X به $\psi_M(x,\lambda_x)$ مدل رگرسیون $Y = g(\beta_0 + \beta_1\psi_M(x,\lambda_x) + \varepsilon)$ را برازش نموده و تابع $g^{-1}(\circ)$ را برآورد می نماییم.(به روش پاسخ معکوس).

خروجی رگرسیون R

```
bcPower Transformation to Normality
Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1     0.5481            0.5       0.3606       0.7357
```

```
Likelihood ratio test that transformation parameter is equal to 0
(log transformation)
LRT df       pval
LR test, lambda = (0) 35.16895  1 3.023e-09

Likelihood ratio test that no transformation is needed
LRT df       pval
LR test, lambda = (1) 21.09339  1 4.3743e-06
```

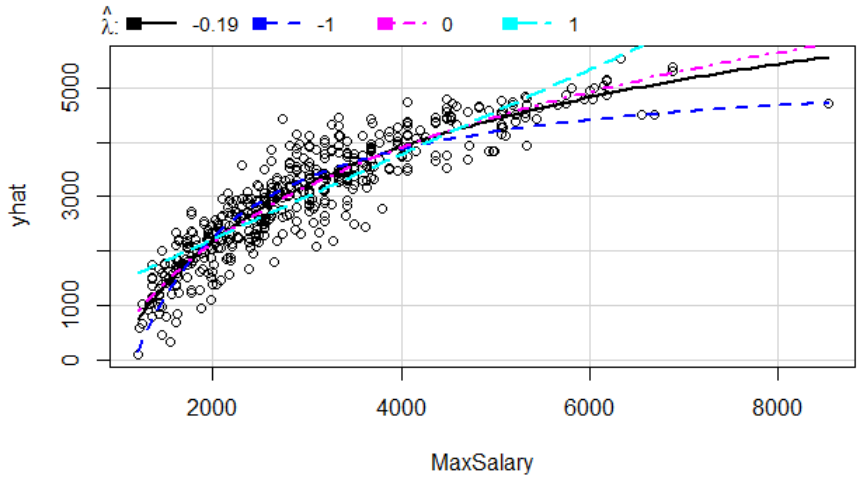
مقدار بهینه برای λ_x طبق خروجی بالا برابر ۰٫۵۴ گزارش شده و این یعنی تبدیل مناسب برای X، یک تبدیل رادیکالی است. بنابراین:

$$\max Salary = g(\beta_0 + \beta_1\sqrt{Score} + \varepsilon) \qquad (۶.۱)$$

و یا

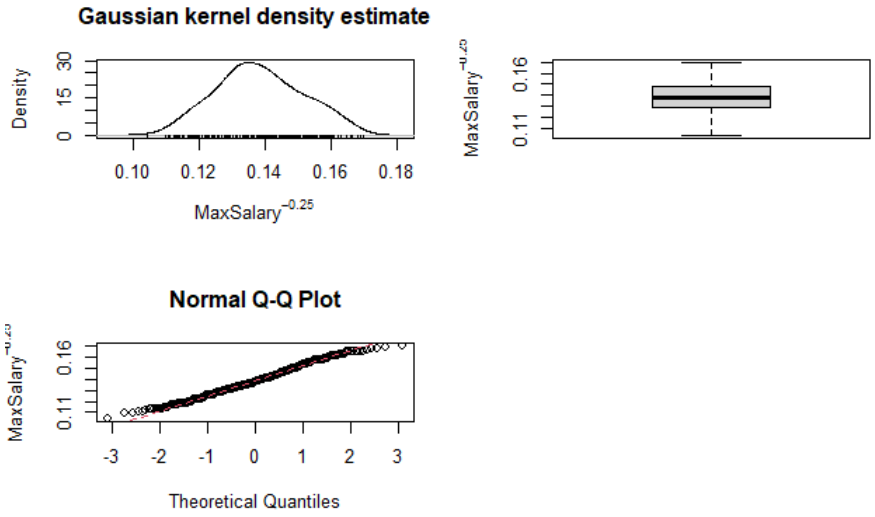
$$g^{-1} = (\max Salary)\beta_0 + \beta_1\sqrt{Score} + \varepsilon$$

برای استفاده از روش نمودار پاسخ معکوس، نمودار $\hat{Y}$ $(\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X)$ در مقابل Y رسم شده است. (باید توجه داشت که چون متغیر X نرمال شده است این روش قابل استفاده است.)



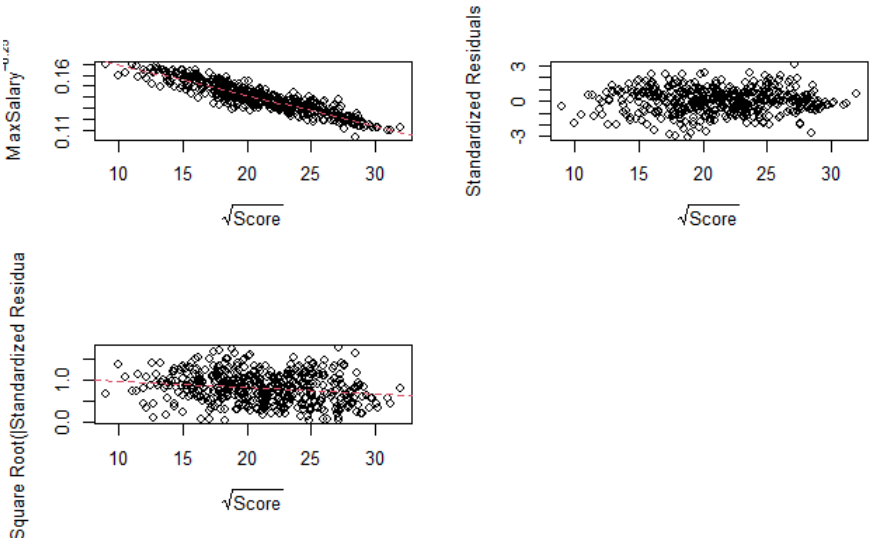
شکل ۳۶.۱: نمودار پاسخ معکوس بر اساس مدل ۶.۱

بر اساس شکل ۳۶.۱ بهترین منحنی برازش یافته به نمودار پراکنش داده ها مربوط به $\lambda = -۰/۱۹$ است و لذا مدل مورد نظر به صورت $\max Salary^{-۰/۱۹} = \beta_۰ + \beta_۱\sqrt{Score} + \varepsilon$ باید در نظر گرفته شود.



شکل ۳۷.۱: نمودارهای متغیر $MaxSalary$ تبدیل شده

در شکل ۳۷.۱ نمودار های جعبه ای - برآورد تابع چگالی و $Q-Q$ برای متغیر Y تبدیل یافته رسم شده اند که نسبتا نشان دهنده نرمال شدن توزیع پاسخ تبدیل یافته هستند.



شکل ۳۸.۱: نمودارهای متغیر $MaxSalary$ تبدیل شده

همچنین در شکل ۳۸.۱ نمودار پراکنش متغیر های تبدیل یافته در مقابل هم به همراه برازش خط کمترین مربعات رسم شده که می توان گفت داده های تبدیل یافته به طور معنی دار تری نسبت به داده های خام باهم رابطه خطی دارند. همچنین در این شکل نمودار مانده های استاندارد در مقابل $\sqrt{Score}$ و جزر قدر مطلق مانده های استاندارد در مقابل $\sqrt{Score}$ رسم شده است. در مقایسه این شکل با شکل ۳۵.۱ (نمودار های رگرسیون تشخیصی مربوط به تبدیل باکس-کاکس) می توان گفت که مدل باکس-کاکس مقداری بهتر از روش ۱ به داده ها برازش پیدا نموده است. (واریانس مانده ها در روش باکس-کاکس ثابت تر به نظر می رسد.) در شکل ۳۸.۱ واریانس مانده ها با افزایش متغیر مستقل جدید یعنی $\sqrt{Score}$ مقداری کاهش می یابد. بنابراین در مقایسه این دو رویکرد با هم، مدل اولی یعنی $\log(MaxSalary) = \beta_0 + \beta_1\sqrt{Score} + \varepsilon$ به دومی ترجیه داده می شود.

البته لازم به گفتن است که کوک و ویزبرگ در سال ۱۹۹۹، اشاره به نتایج یکسان دو روش در بسیاری از موارد نمودند ولی همان طور که دیدیم در مثال اخیر، دو روش منجر به نتایج یکسان نشدند. آنها همچنین رویکرد ۱ را استوار تر از رویکرد ۲ می دانند یعنی اگر در مجموعه داده ها، داده یا داده های پرت وجود داشته باشد، رویکرد ۱ منجر به نتایج بهتری خواهد شد و رویکرد ۲ نسبت به داده های پرت حساسیت بیشتری نشان می دهد.

فصل ۲

کمترین مربعات وزنی

در این فصل رویکرد کمترین مربعات وزنی برای حل مشکل ثابت نبودن واریانس جملات خطا بررسی خواهد شد.

۱.۲ برازش مدل رگرسیون خطی ساده به روش کمترین مربعات وزنی

مدل رگرسیون خطی ساده $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ را با فرض $E(\varepsilon_i) = 0, \mathrm{Var}(\varepsilon_i) = \frac{\sigma^2}{W_i}$ در نظر بگیرید. اگر W_i دارای مقداری بزرگ و در نتیجه ε_i کوچک باشد، آنگاه برآورد $\hat{Y}_i$ باید بسیار به مقدار Y_i نزدیک باشد و برعکس اگر W_i کوچک شود، ε_i بزرگ شده و لذا زوج (X_i, Y_i) باید از داده ها حذف شود و برازش کمترین مربعات بدون در نظر گرفتن این زوج انجام گردد.

برآورد های کمترین مربعات وزنی پارامتر های $\beta_۰, \beta_۱$:

قبل از برآورد پارامتر ها، ابتدا باید واریانس جملات خطا را ثابت نمود، لذا داریم:

$$Y_i = \beta_۰ + \beta_۱ X_i + \varepsilon_i$$

$$\Rightarrow \sqrt{W_i} Y_i = \sqrt{W_i}\beta_۰ + \sqrt{W_i}\beta_۱ X_i + \sqrt{W_i}\varepsilon_i \to \mathrm{Var}(\sqrt{W_i}\varepsilon_i) = \frac{\sigma^۲}{W_i} \times W_i = \sigma^۲$$

بنابراین برآوردگرهای مورد نظر با مینیمم سازی عبارت زیر که مجموع مربعات خطای وزنی نامیده می شود، حاصل می شود:

$$WRSS = \sum_{i=۱}^{n} (\sqrt{W_i} Y_i - \sqrt{W_i}\hat{\beta}_۰ - \sqrt{W_i}\hat{\beta}_۱ X_i)^۲$$

$$= \sum W_i (Y_i - \hat{\beta}_۰ - \hat{\beta}_۱ X_i)^۲$$

همان طور که در رابطه اخیر دیده می شود هرچه $\mathrm{Var}(\varepsilon_i)$ کوچک تر باشد و W_i بزرگ تر باشد، وزن داده شده به مشاهده i ام یعنی (X_i, Y_i) بزرگ تر شده و لذا تاثیر بیشتری در برآورد پارامتر ها خواهد داشت و بر عکس اگر $\mathrm{Var}(\varepsilon_i)$ بزرگ باشد و W_i کوچک شده و مشاهده i ام تاثیر کمتری بر برازش ما خواهد گذاشت که این موضوع به لحاظ منطقی نیز درست به نظر می رسد. زیرا اگر $\mathrm{Var}(\varepsilon_i)$ کوچک باشد یعنی دقت مشاهده i ام زیاد بوده و باید وزن بیشتری به آن دهیم و اگر $\mathrm{Var}(\varepsilon_i)$ بزرگ باشد یعنی مشاهده i ام دارای دقت کمتری است و لذا باید ارزش کمتری برای آن قائل شویم تا جایی که اگر $W_i \to ۰$ یعنی $\mathrm{Var}(\varepsilon_i) = \frac{\sigma^۲}{W_i} \to \infty$ آنگاه وزن داده شده به مشاهده i ام در برازش مورد نظر صفر خواهد بود.

جهت به دست آوردن برآوردگرهای کمترین مربعات وزنی (WLS) باید $WRSS$ را مینیمم

نماییم.

با مشتق گیری از $WRSS$ نسبت به $\hat{\beta}_۰$ و $\hat{\beta}_۱$ داریم:

$$\begin{cases} \frac{\partial WRSS}{\partial \hat{\beta}_۰} = -۲\sum_{i=۱}^{n} W_i(Y_i - \hat{\beta}_۰ - \hat{\beta}_۱ X_i) = ۰ \qquad or \qquad \sum_{i=۱}^{n} W_i e_i = ۰ \\ \frac{\partial WRSS}{\partial \hat{\beta}_۱} = -۲\sum_{i=۱}^{n} W_i X_i(Y_i - \hat{\beta}_۰ - \hat{\beta}_۱ X_i) = ۰ \qquad or \qquad \sum_{i=۱}^{n} W_i X_i e_i = ۰ \end{cases}$$

$$\Rightarrow \begin{cases} \sum W_i Y_i - \hat{\beta}_۰ \sum W_i - \hat{\beta}_۱ \sum W_i X_i = ۰ \Rightarrow \hat{\beta}_۰ = \frac{\sum W_i Y_i - \hat{\beta}_۱ \sum W_i X_i}{\sum W_i} \; [*] \\ \sum W_i X_i Y_i - \hat{\beta}_۰ \sum W_i X_i - \hat{\beta}_۱ \sum W_i X_i^۲ = ۰ \end{cases} \tag{۱.۲}$$

با جایگذاری [*] در رابطه ۱.۲ داریم:

$$\hat{\beta}_{۱W} = \frac{(\sum W_i)(\sum W_i X_i Y_i) - (\sum W_i X_i)(\sum W_i Y_i)}{(\sum W_i)(\sum W_i X_i^۲) - (\sum W_i X_i)^۲}$$

$$= \frac{\frac{\sum W_i X_i Y_i}{\sum W_i} - \frac{(\sum W_i X_i)(\sum W_i Y_i)}{(\sum W_i)^۲}}{\frac{\sum W_i X_i^۲}{\sum W_i} - \frac{(\sum W_i X_i)^۲}{(\sum W_i)^۲}} = \frac{\sum W_i (X_i - \bar{X}_W)(Y_i - \bar{Y}_W)}{\sum W_i (X_i - \bar{X}_W)^۲} \tag{۲.۲}$$

به طوریکه:

$$\bar{X}_W = \frac{\sum W_i X_i}{\sum W_i}, \qquad \bar{Y}_W = \frac{\sum W_i Y_i}{\sum W_i}$$

با جایگذاری رابطه ۲.۲ در [*] می توان نوشت:

$$\hat{\beta}_{۰W} = \frac{\sum W_i Y_i}{\sum W_i} - \hat{\beta}_۱ \frac{\sum W_i X_i}{\sum W_i} = \bar{Y}_W - \hat{\beta}_{۱W} \bar{X}_W$$

$$\hat{\beta}_{۱W} = \frac{WS_{xy}}{WS_{xx}}$$

مثال ۱.۱.۲. مجددا به مثال ۵.۱ قبل بر می گردیم. همان طور که دیدیم واریانس خطاها ثابت نبوده و با افزایش X افزایش می یافت. حال با استفاده از رگرسیون وزنی مدل را مجددا

برآورد نموده و پیشگویی مورد نظر را انجام خواهیم داد. همان طور که در رگرسیون وزنی دیدیم برای ثابت نمودن جملات خطا باید وزن ها را برابر $W_i = \frac{۱}{\mathrm{Var}(Y_i)} = \frac{۱}{\mathrm{Var}(\varepsilon_i)}$ در نظر می گیریم.

خروجی رگرسیون R

```
Call:
lm(formula = Rooms ~ Crews, weights = 1/StdDev^2)

Weighted Residuals:
Min        1Q    Median        3Q       Max
-1.43184 -0.82013  0.03909  0.69029  2.01030

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8095     1.1158   0.725    0.471
Crews         3.8255     0.1788  21.400   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.9648 on 51 degrees of freedom
Multiple R-squared:  0.8998,    Adjusted R-squared:  0.8978
F-statistic:   458 on 1 and 51 DF,  p-value: < 2.2e-16

fit       lwr       upr
1 16.11133 13.71210 18.51056
2 62.01687 57.38601 66.64773
```

فواصل اطمینان برای کمترین مربعات وزنی

برای یافتن فاصله اطمینان در مدل رگرسیون وزنی شده، انحراف از معیار مقادیر پیشگویی شده در نقطه $X = x$ به صورت زیر محاسبه می شود:

$$\hat{\sigma}\sqrt{۱ + \frac{۱}{n} + \frac{(x^* - \bar{X})^۲}{WS_{xx}}} \qquad \sigma_{\hat{W}} = \sqrt{MSE} = \frac{WRSS}{n - ۲}$$

بنابراین:

$$Y(x^*) = Y|x^* : (\hat{\beta_{\circ w}} + \hat{\beta_{\backslash w}} x^*) \pm t_{\frac{\alpha}{۲}}(n-۲)\hat{\sigma_w}\sqrt{۱ + \frac{۱}{n} + \frac{(x^* - \bar{X})^۲}{WS_{xx}}}$$

لوریج در کمترین مربعات وزنی

همان طور که می دانیم برآورد کمترین مربعات وزنی Y در نقطه $X = X_i$ برابر است با:

$$\hat{Y_{wi}} = (\hat{\beta_{\circ w}} + \hat{\beta_{\backslash w}} x_i) \qquad , \beta_{\backslash w} = \frac{WS_{xy}}{WS_{xx}} \qquad , \hat{\beta_{\circ w}} = \bar{Y}_w - \beta_{\backslash w}\bar{X}_w$$

بنابراین:

$$\hat{Y_{wi}} = \bar{Y}_w - \hat{\beta_{\backslash w}}(x_i - \bar{X}_w) = \frac{\sum_{j=\backslash}^{n} W_j Y_j}{\sum_{k=\backslash}^{n} W_k} + \sum_{j=\backslash}^{n} \frac{W_j(X_j - \bar{X}_w)}{WS_{xx}} Y_j (X_i - \bar{X}_w)$$

$$= \sum_{j=\backslash}^{n} \big[W_j^s + \frac{W_j - (X_i - \bar{X}_w)(X_j - \bar{X}_w)}{WS_{xx}}\big] Y_j = \sum_{j=\backslash}^{n} h_{wij} Y_j$$

$$W_j^s = \frac{W_j}{\sum_{k=\backslash}^{n} W_k} \qquad , h_{wij} = W_j^s + \frac{W_j(X_i - \bar{X}_w)(X_j - \bar{X}_w)}{WS_{xx}}$$

$$\hat{Y}_w = h_{wii} Y_i + \sum_{j \neq i} h_{wij} Y_j \qquad , h_{wii} = W_i^s + \frac{W_i(X_i - \bar{X}_w)^۲}{WS_{xx}}$$

بنابراین می توان گفت که هرگاه همه وزن ها برابر باشند یعنی $W_\backslash = \cdots = W_n \to W_i = \frac{۱}{n}$ انگاه روش کمترین مربعات وزنی و کمترین مربعات معادل خواهند شد.

برازش مدل کمترین مربعات وزنی با استفاده از روش کمترین مربعات

مدل رگرسیون زیر را در نظر بگیرید:

$$Y_i = \beta_\circ + \beta_\backslash X_i + \varepsilon_i \qquad , \varepsilon_i \sim (\circ, \frac{\sigma^۲}{W_i})$$

$$\Rightarrow \sqrt{W_i} Y_i = \beta_\circ \sqrt{W_i} + \beta_\backslash \sqrt{W_i} X_i + \sqrt{W_i}\varepsilon_i$$

همان طور که در ابتدای فصل دیدیم، در مدل جدید واریانس جملات خطا ثابت بوده و لذا می توان از کمترین مربعات معمولی برای برآورد پارامتر ها استفاده نمود، زیرا:

$$var(\sqrt{W_i}\varepsilon_i) = W_i\frac{\sigma^۲}{W_i} = \sigma^۲$$

بنابراین، تعریف متغیر های جدید به شکل زیر می توان مدل جدید را به صورت زیر تعریف نمود:

$$Y_{Newi} = \beta_۰ X_{۱Newi} + \beta_۱ X_{۲Newi} + \varepsilon_{Newi}$$

$$Y_{Newi} = \sqrt{W_i}Y_i \quad , X_{۱Newi} = \sqrt{W_i} \quad , X_{۲Newi} = \sqrt{W_i}X_i \quad , \varepsilon_{Newi} = \sqrt{W_i}\varepsilon_i$$

مثال برآورد پارامتر ها در مثال قبل به روش جدید

باتعریف متغیر های جدید در مساله ۵.۱ خروجی زیر به دست می آید.

خروجی رگرسیون R

```
Call:
lm(formula = ynew ~ x1new + x2new - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43184 -0.82013  0.03909  0.69029  2.01030

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
x1new    0.8095     1.1158   0.725    0.471
x2new    3.8255     0.1788  21.400   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.9648 on 51 degrees of freedom
Multiple R-squared:  0.9617,    Adjusted R-squared:  0.9602
```

```
F-statistic: 639.6 on 2 and 51 DF,  p-value: < 2.2e-16

fit       lwr       upr
1 3.243965 1.286166 5.201763
2 5.167873 3.199481 7.136265
```

$$Y = \hat{\beta}_{۰} X_{۱New}(\frac{۱}{\sqrt{\text{Var}(Y|X=۴)}}) + \hat{\beta}_{۱} X_{۲New}(۴ \times \frac{۱}{\sqrt{\text{Var}(Y|X=۴)}})$$
$$= (۰/۸۰۹۵)(\frac{۱}{۴/۹۷}) + ۳/۸۲۵۸ \times \frac{۴}{۴/۹۷} = ۳/۲۴$$

در فاصله اطمینان بدست آمده نیز کافیست جای $X_{۱New}$ مقدار $\frac{۱}{۴/۹۷}$ و به جای $X_{۲}$ مقدار $\frac{۴}{۴/۹۷}$ را قرار می دهیم.

مانده های کمترین مربعات وزنی

$$e_{wi} = \sqrt{Wi}(Y_i - \hat{Y_{wi}}) \Rightarrow \sum e_{wi}^{۲} = \sum W_i(Y_i - \hat{Y_{wi}})^{۲} = \sum_{i=۱}^{n} W_i(Y_i - \hat{\beta_{۰w}} - \hat{\beta_{۱w}} X_i)^{۲}$$

کاربرد کمترین مربعات وزنی

در بسیاری از موارد متغیر وابسته Y_i ،میانگین یا میانه متغیر های پاسخ در n_i مشاهده مستقل است و لذا $\text{Var}(Y_i) = \frac{۱}{n_i}$ و بنابراین با استفاده از رگرسیون وزنی باید $W_i = n_i$ در نظر گرفته شود.

فصل ۳

عیب شناسی در رگرسیون چند گانه و انجام تبدیلات بر روی آن

در این فصل روش هایی را برای آزمون فرضیات اساسی در مدل رگرسیون چندگانه و نهایتا عیب شناسی در این مدل ارائه خواهیم نمود. همچنین در صورت امکان روش هایی را برای غلبه بر این مشکلات (از قبیل ثابت نبودن واریانس خطاها و غیر خطی بودن رابطه بین متغیر های وابسته و مستقل) پیشنهاد خواهیم نمود.

۱.۳ عیب شناسی در رگرسیون چندگانه

در هنگام برازش یک مدل رگرسیون چندگانه برای بررسی برقراری فرضیات اساسی توجه به نکات زیر ضروری است:

۱- مشخص کنیم که آیا مدل برازش شده به داده ها معتبر است یا خیر(برازش کافی به داده ها دارد یا خیر) ابزار اصلی جهت رسیدن به این هدف و بررسی برقراری فرضیات رسم نمودار مانده های استاندارد و مقادیر برازش شده است. همچنین رسم نمودار های

پراکنش حاشیه ای نیز در انجام این امر بسیار مهم است.

۲- مشخص نمودن اینکه کدام یک از نقاط دارای تاثیر زیادی بر برازش مدل به دست آمده هستند (این کار می تواند به کمک مثلا لوریج ها انجام گردد).

۳- مشخص نمودن داده های پرت در صورت وجود. یعنی داده هایی که از طرح کلی مدل برازش شده و سایر داده ها تبعیت نمی کنند.

۴- ارزیابی تاثیر هر یک از متغیر های پیشگو به تنهایی برروی متغیر پاسخ با استفاده از نمودار های متغیر اضافه شده.

۵- ارزیابی مقدار همبستگی بین متغیر های پیشگو با استفاده از معیار های معرفی شده به عنوان مثال عوامل تورم واریانس.

۶- آزمون اینکه آیا فرضیه ثابت بودن واریانس خطاها برقرار است یا خیر. اگر خیر چگونه می توان براین مشکل غلبه نمود.

۷- اگر داده های زمانی داریم یعنی داده ها در طول زمان جمع آوری شده اند بررسی اینکه آیا این داده ها بر روی زمان همبستگی دارند یا خیر.

یادآوری رگرسیون چندگانه در فرمت ماتریسی

فرض کنید $Y_{n\times ۱}$ بردار متغیر پاسخ و $X_{n\times(p+۱)}$ ماتریس طرح باشد. یعنی

$$Y=\begin{bmatrix}Y_۱\\Y_۲\\\vdots\\Y_n\end{bmatrix} \qquad ,X=\begin{bmatrix}۱ & X_{۱۱} & \dots & X_{۱p}\\۱ & X_{۲۱} & \dots & X_{۲p}\\\vdots & \vdots & \ddots & \vdots\\۱ & X_{n۱} & \dots & X_{np}\end{bmatrix}$$

بر اساس ماتریس های فوق، بردار پارامتر ها و خطاها به صورت زیر قابل تعریفند:

$$\beta = (\beta_0 \beta_1 \dots \beta_p)', \qquad \varepsilon = (\varepsilon_1 \varepsilon_2 \dots \varepsilon_n)'$$

بنابراین می توان نوشت:

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim (0, \sigma^2 I)$$

$$\hat{\beta} = (X'X)^{-1}X'Y \Rightarrow \hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY, \quad H = X(X'X)^{-1}X'$$

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

۱.۱.۳ لوریج ها در رگرسیون چندگانه

همان طور که دیدیم در مدل رگرسیون خطی ساده با لوریج ها معیاری برای نشان دادن تاثیر مشاهده i ام بر روی مدل برازش شده بودند. در مدل رگرسیون چندگانه نیز این مقادیر همین نقش را را داشته و به صورت زیر تعریف می شوند:

$$\hat{Y} = HY, \qquad \hat{Y}_i = (Diameter\, i\, matrix\, H)XY = \sum_{j=1}^{n} h_{ij}Y_j = h_{ii}Y_i + \sum_{j=1, j\neq i}^{n} h_{ij}Y_j$$

بنابراین می توان نوشت:

$$h_{ii} = x_i(X'X)^{-1}x_i', \qquad h_{ij} = x_i(X'X)^{-1}x_j'$$

مشاهده i ام یک مشاهده تاثیر گذار است. $\rightarrow h_{ii} > 2\frac{p+1}{n}$

به طوریکه x_i سطر i ماتریس طرح یعنی $(x_i | x_{i1}\ x_{i2}\ \dots\ x_{ip})$ و x_j سطر j از آن سطر است.

۲.۱.۳ بررسی خواص مانده ها در مدل رگرسیون چندگانه

بردار مانده ها در مدل رگرسیون چندگانه به صورت زیر تعریف می شود:

$$e = Y - \hat{Y} = (I - H)Y \Rightarrow E(e) = E(Y - \hat{Y}) = E\big[(I - H)Y\big]$$

$$\Rightarrow E(e) = (I - H)E(Y) = (I - H)X\beta = X\beta - HX\beta = ۰$$

$$\mathrm{Var}(e) = \mathrm{Var}\big[(I - H)Y\big] = \sigma^۲(I - H)I(I - H)' = \sigma^۲(I - H)$$

زیرا H یک ماتریس متقارن و $(I - H)$ خود توان است. $H = H'$

$$(I - H)(I - H) = I - ۲H + H^۲ = I - ۲H + H = I - H$$

تمرین:

$$cov(e, \hat{Y}) = ۰, cov(e, Y) = \sigma^۲(I - H)$$

بنابراین می توان مانده های استاندارد را به صورت زیر تعریف نمود:

$$\mathrm{Var}(e_i) = \sigma^۲(۱ - h_{ii}) \Rightarrow r_i = \frac{e_i}{s\sqrt{۱ - h_{ii}}}, \qquad s = \sqrt{MSE} = \sqrt{\frac{SSE}{n - (p + ۱)}}$$

چنانچه $|r_i| > ۲$ آنگاه می توان گفت که مشاهده i ام ی داده پرت است و در غیر این صورت پرت نیست. باید توجه داشت در صورتی یک مشاهده پرت خواهد بود که مدل برازش شده معتبر باشد و در غیر این صورت نباید برچسب داده پرت به آن زد. همچنین ممکن است برای یک مجموعه از داده ها $r_i = -۱/۸$ شود (به ویژه وقتی n کوچک است) ولی در مقایسه با طرح کلی داده ها آن داده یک داده پرت به نظر آید. بنابراین قبل از حذف یک داده به عنوان

داده پرت، بهترین کار نگاه به نمودار پراکنش داده ها و بررسی چند معیار به طور همزمان برای پرت بودن داده هاست که در ادامه معرفی خواهند شد.

استفاده از مانده ها و مانده های استاندارد برای بررسی مدل

به زبان ساده، مدل رگرسیون چندگانه یک مدل معتبر برای داده هاست هرگاه میانگین شرط Y با شرط X یک تابع خطی از X باشد و واریانس شرطی $\mathrm{Var}(Y|X=x)=\sigma^2$ ثابت باشد.

$$E(Y|X=x)=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p$$

بنابراین یک مدل در صورتی معتبر است که نمودار مانده های استاندارد در مقابل هریک از متغیر های پیشگو یا هر ترکیب خطی از آن ها دارای دو خاصیت زیر باشد:

۱- چون در صورت برازش یک مدل صحیح $E(e_i)=0$ است، لذا باید نمودار پراکنش فوق یک طرح تصادفی حول محور x ها را نشان دهد.

۲- یک تغییر پذیری ثابت حول محور x ها وجود داشته باشد. (چون واریانس مدل صحیح ثابت است) بنابراین می توان گفت که اگر در نمودار های پراکنش فوق هر گونه طرحی وجود داشته باشد غیر معتبر بودن مدل برازش شده آشکار است.

نمودار مربوط به متغیر افزوده

به کمک این نمودار ها به طور شهودی می توان به تاثیر هریک از متغیر های پیشگو در توجیه متغیر پاسخ بدون در نظر گرفتن تاثیر سایر متغیر های پیشگو پی برد. ابتدا مدل

$$Y=X\beta+\varepsilon,\ \mathrm{Var}(\varepsilon)=\sigma^2 I \qquad (۱.۳)$$

در نظر بگیرید. فرض کنید می خواهیم مقدمات ورود یک متغیر پیشگوی جدید را مانند Z به مدل فوق فراهم نماییم، یعنی مدل زیر را در نظر بگیریم:

$$Y = X\beta + Z\alpha + \varepsilon, \qquad Z = (Z_1, Z_2 \ldots, Z_n)', \alpha \in \Re \qquad (۲.۳)$$

بنابراین باید ابتدا تاثیر جزئی متغیر جدید Z در Y را بدون در نظر گرفتن تاثیر سایر p متغیر قبلی X بر Y بررسی نماییم. برای بررسی چنین موضوعی کافی است مانده های حاصل از رگرسیون $Y = X\beta + \varepsilon$ را برروی مانده های حاصل از مدل $Z = X\delta + \varepsilon$ رگرسیون نماییم. به عبارت دیگر باید $e_{YX} = Y - \hat{Y} = (I - H_X)Y$ را به عنوان متغیر وابسته جدید روی متغیر مستقل جدید $e_{ZX} = Z - \hat{Z} = (I - H_X)Z$ رگرسیون نماییم به طوریکه

$$H_X = X(X'X)^{-1}X'$$

بنابراین می توان گفت که روش متغیر افزوده در واقع نشان دهنده تغییرات Y است که توسط X قابل توجیه نبوده و پس از حذف اثر آن توسط Z توجیه پذیر است.

می توان نوشت:

$$e_{YX} = Y - \hat{Y} = (I - H_X)Y$$

$$e_{ZX} = Z - \hat{Z} = (I - H_X)Z$$

$$Y = X\beta + Z\alpha + \varepsilon \overset{\times(I-H_X)}{\rightarrow} (I - H_X)Y = (I - H_x)X\beta + (I - H_X)Z\alpha + (I - H_X)\varepsilon$$

$$\Rightarrow e_{YX} = (X - X(X'X)^{-1}X'X)\beta + e_{ZX}\alpha + \varepsilon^* \quad (\varepsilon^* = (I - H_X)\varepsilon)$$

$$\Rightarrow e_{YX} = 0 + e_{ZX}\alpha + \varepsilon^*$$

$$\Rightarrow e_{YX} = e_{ZX}\alpha + \varepsilon^*$$

تمرین

اگر $\hat{\alpha}_{AVP}$ مقدار برآورد α در مدل فوق و $\hat{\alpha}_{LS}$ برآورد α در مدل $Y = X\beta + Z\alpha + \varepsilon$ باشد، می توان ثابت نمود که $\hat{\alpha}_{AVP} = \hat{\alpha}_{LS}$. برای حالت سه متغیره به عنوان تمرین ثابت شود یعنی مدل $Y = \beta_0 + \beta_1 X + \varepsilon$ را در نظر گرفته و با اضافه نمودن متغیر Z به مدل مقدار α را از دو روش برآورد و درستی رابطه فوق را تصدیق نمایید.

$$Y = \beta_0 + \beta_1 X + \alpha Z + \varepsilon$$

حال اگر مدل ۲.۳ صحیح باشد آنگاه نمودار متغیر افزوده باید نقاطی را تولید نماید که به طور تصادفی حول خطی با شیب $\hat{\alpha}_{LS}$ که از مبدا میگذرد پراکنده شوند.

مثال: بررسی رابطه در منوی قیمت در یک رستوران جدید در نیویورک

در این مطالعه متغیر ها به صورت زیر تعریف می شوند:

Y :قیمت یک شام بر حسب دلار

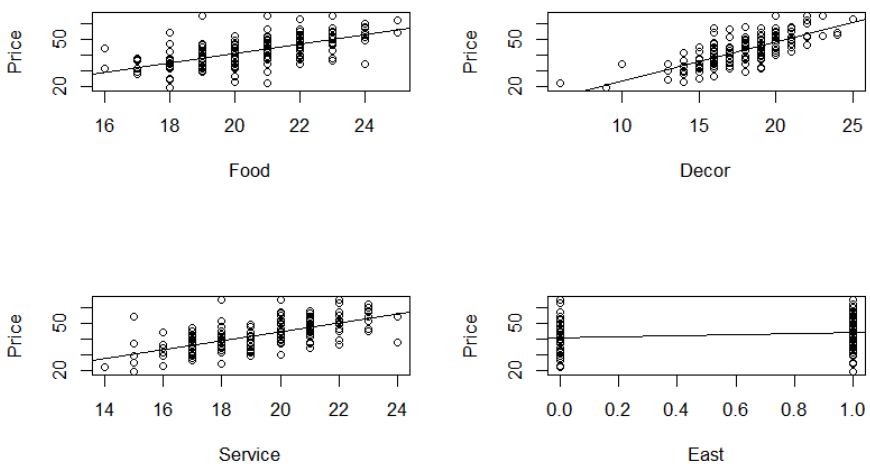
X_1: نمره غذا توسط مشتری

X_2: نمره دکوراسیون توسط مشتری

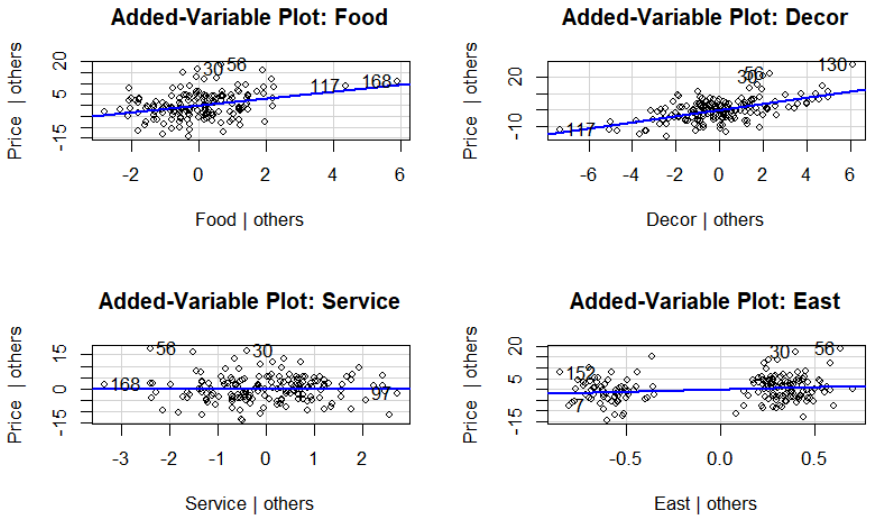
X_3: نمره سرویس توسط مشتری

X_4: متغیر توضیحی ۱ برای شرق و ۰ برای غرب

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$



شکل ۱.۳: نمودار پراکندگی Y، قیمت در برابر هر پیش بینی کننده



شکل ۲.۳: نمودار متغیر های اضافه شده

در شکل ۱.۳ تاثیر هر یک از متغیر های مستقل روی Y نمایش داده شده است. برای بررسی اثر هر یک از متغیر های مستقل به تنهایی بر روی Y ، نمودار های افزوده برای هر یک از

متغیر های پیشگو در شکل ۲.۳ رسم شده است. همان طور که دیده می شود به غیر از متغیر مستقل سرویس سایر متغیر ها دارای شیب خطی مخالف با صفر هستند. بنابراین می توان گفت که متغیر سرویس دارای تاثیر اندکی در متغیر وابسته است.

۲.۳ تبدیلات

در این بخش تبدیلات را بررسی خواهیم نمود که به دو منظور زیر بر روی متغیر (ها) اعمال می گردند:

۱- غلبه بر مشکل غیر خطی بودن

۲- غلبه بر مشکل ثابت نبودن واریانس

۱.۲.۳ استفاده از تبدیلات برای غلبه بر غیر خطی بودن

مانند فصل سوم کتاب ، در این بخش نیز دو روش کلی برای یافتن تبدیل مناسب معرفی خواهند شد که عبارتند از:

۱- نمودار پاسخ معکوس($Inverse\ Response\ Plot$)

۲- روش باکس-کاکس

در سه موقعیت می توانیم از تبدیلات استفاده نماییم که عبارتند از:

a- اعمال تبدیل تنها بر روی متغیر پاسخ

b- اعمال تبدیل بر روی متغیر یا متغیر های مستقل

c- اعمال تبدیل بر روی متغیر های پاسخ و مستقل

استفاده از تبدیل به روش رگرسیون معکوس تنها روی متغیر پاسخ

فرض کنید مدل صحیح رگرسیون به شکل زیر باشد:

$$Y = g(\beta_۰ + \beta_۱ X_۱ + \cdots + \beta_p X_p)$$

که در آن $g(.)$ یک تابع نامعلوم است. مدل فوق با استفاده از تابع معکوس $g(.)$ می تواند به راحتی به مدل رگرسیون خطی زیر تبدیل گردد:

$$g^{-۱}(Y) = \beta_۰ + \beta_۱ X_۱ + \cdots + \beta_p X_p$$

به عنوان مثال:

$$Y = \exp\{\beta_۰ + \beta_۱ X_۱ + \cdots + \beta_p X_p\} \Rightarrow \log(Y) = \beta_۰ + \beta_۱ X_۱ + \cdots + \beta_p X_p$$

مثال

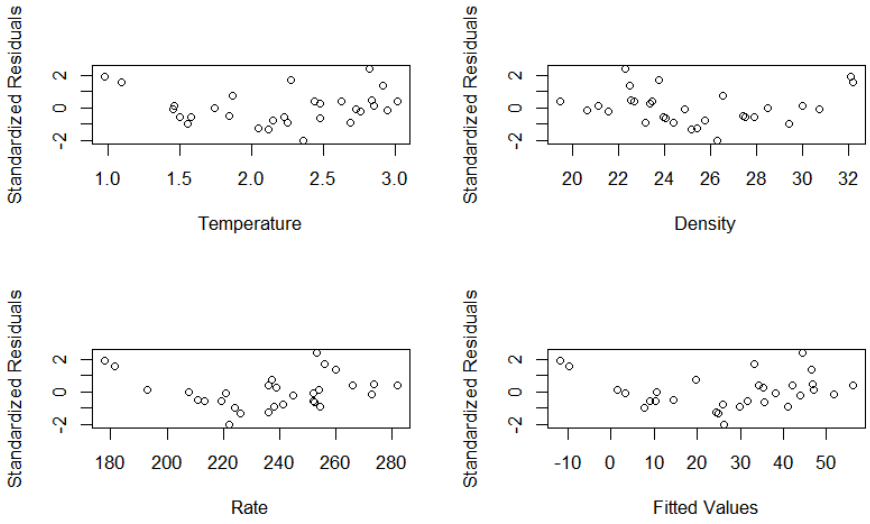
در این مثال می خواهیم متغیر Y (نرخ خرابی$Defective$) را بر روی $X_۱$ ($Temp$)و $X_۲$ ($density$) و $X_۳$ ($Rate$) رگرسیون نماییم.

ابتدا با مدل زیر شروع می نماییم:

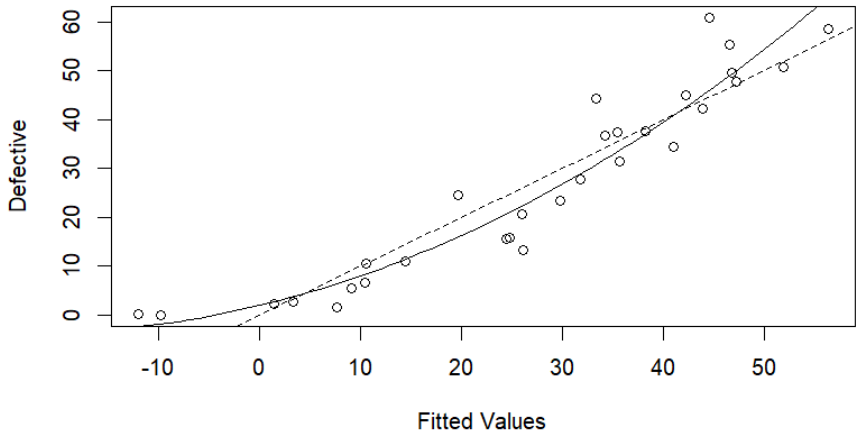
$$Y = \beta_۰ + \beta_۱ X_۱ + \beta_۲ X_۲ + \beta_۳ X_۳ + \varepsilon$$

جدول ۱.۳: داده‌های مربوط به نرخ‌های معیوب

Temperature	Density	Rate	Defective	Temperature	Density	Rate	Defective
0.97	32.08	177.7	0.2	2.76	21.58	244.7	42.2
2.85	21.14	254.1	47.9	2.36	26.3	222.1	13.4
2.95	20.65	272.6	50.9	1.09	32.19	181.4	0.1
2.84	22.53	273.4	49.7	2.15	25.73	241	20.6
1.84	27.43	210.8	11	2.12	25.18	226	15.9
2.05	25.42	236.1	15.6	2.27	23.74	256	44.4
1.5	27.89	219.1	5.5	2.73	24.85	251.9	37.6
2.48	23.34	238.9	37.4	1.46	30.01	192.8	2.2
2.23	23.97	251.9	27.8	1.55	29.42	223.9	1.5
3.02	19.45	281.9	58.7	2.92	22.5	260	55.4
2.69	23.17	254.5	34.5	2.44	23.47	236	36.7
2.63	22.7	265.7	45	1.87	26.51	237.3	24.5
1.58	27.49	213.3	6.6	1.45	30.7	221	2.8
2.48	24.07	252.2	31.5	2.82	22.3	253.2	60.8
2.25	24.38	238.1	23.4	1.74	28.47	207.9	10.5



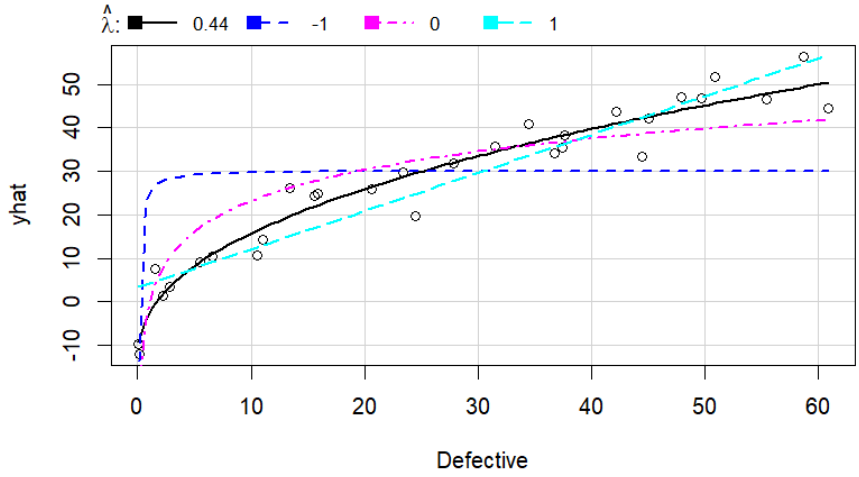
شکل ۳.۳: نمودار باقی مانده های استاندارد شده



شکل ۴.۳: نمودار پراکنش

شکل ۳.۳ اشاره به یک طرح غیر تصادفی دارد. در شکل ۴.۳ نمودار پراکنش $\hat{Y}$ در مقابل Y رسم شده که با استفاده از آن به ضعف یک خط راست به این نقاط پی برد. بنابراین مدل خطی برای داده ها مناسب نبوده و باید به دنبال یک تبدیل مناسب باشیم. بدین منظور از روش رسم پاسخ معکوس استفاده می نماییم. فرض می کنیم مدل مناسب به صورت زیر باشد:

$$Y = g(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon) \quad or \quad g^{-1}(Y) = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon$$



شکل ۵.۳: نمودار پاسخ معکوس

با رسم مقادیر $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$ در مقابل Y در شکل ۵.۳ در می یابیم که تبدیل مناسب $g^{-1}(Y) = Y^{0.44}$. بنابراین مدل مناسب برای این داده ها عبارت است از:

$$Y^{0.5} or Y^{0.44} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

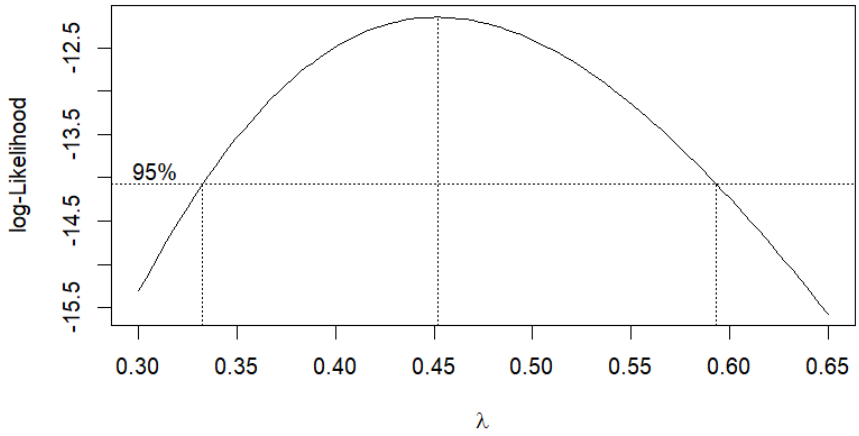
اعمال تبدیلات برروی متغیر پاسخ با استفاده از روش باکس-کاکس

همانند گذشته، خانواده بهبود یافته تبدیلات توانی بر روی Y به صورت زیر تعریف خواهند شد:

$$\psi_M(Y,\lambda) = \psi_s(Y,\lambda) gm(Y)^{1-\lambda} = \begin{cases} gm(Y)^{1-\lambda}\dfrac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ gm(Y)\log(Y) & \lambda = 0 \end{cases}$$

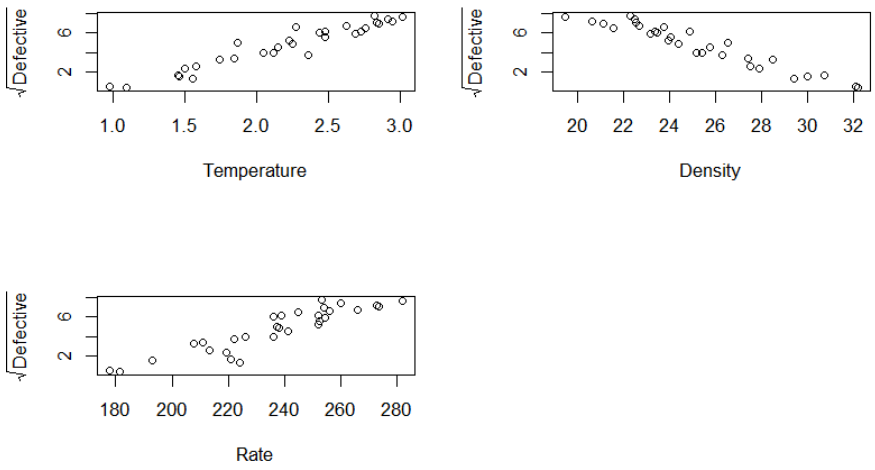
این روش بر این اساس است که در واقع برای برخی مقادیر λ ، $\psi_M(Y,\lambda)$ دارای توزیع نرمال تقریبی است. این روش مقدار بهینه λ را با ماکزیمم سازی تابع درستنمایی برآورد می نماید.

مثال: بررسی مثال گذشته با رویکرد باکس-کاکس

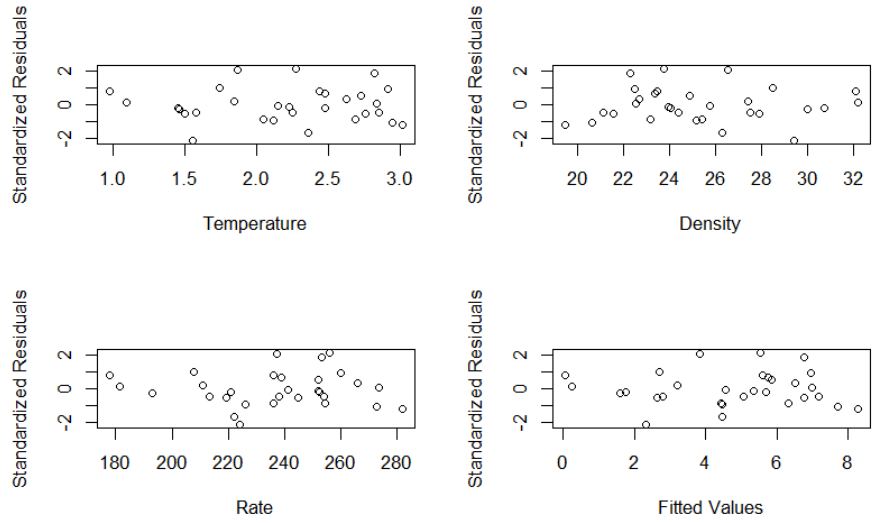


شکل ۶.۳: نمودار تابع درستنمایی

در شکل ۶.۳ ، تابع درستنمایی $\psi_M(Y,\lambda)$ در مقابل Y رسم شده است. بر این اساس، مقدار ماکزیمم کننده λ برابر با ۰٫۴۵ خواهد بود که به جواب روش نمودار معکوس بسیار نزدیک است. در ادامه، نمودار پراکنش $\sqrt{Y}$ در مقابل سه متغیر مستقل رسم شده است و مشاهده می شود که رابطه آن ها به یک رابطه خطی نزدیک تر شده است.

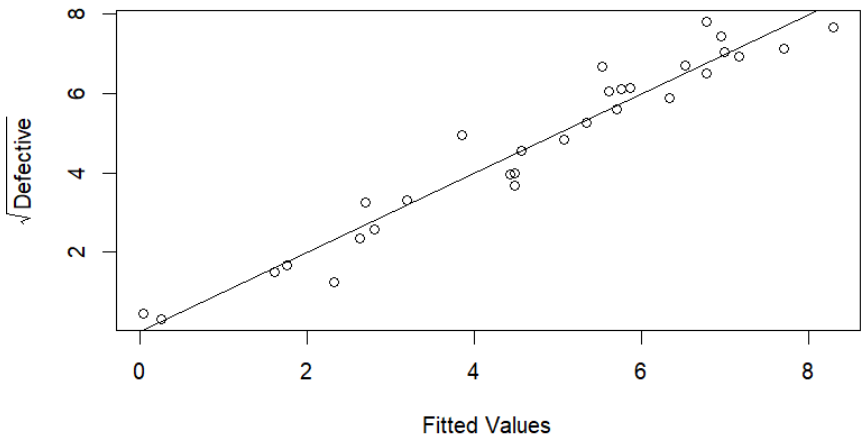


شکل ۷.۳: نمودار $Y^{0.5}$



شکل ۸.۳: نمودار مانده های استاندارد شده

در شکل ۸.۳ نمودار پراکنش مانده های استاندارد در مقابل متغیر های مستقل و مقدار برازش شده برای مدل تبدیل یافته رسم شده که همه آن ها بر مناسب بودن مدل جدید تاکید دارند.



شکل ۹.۳: نمودار $\sqrt{Y}$

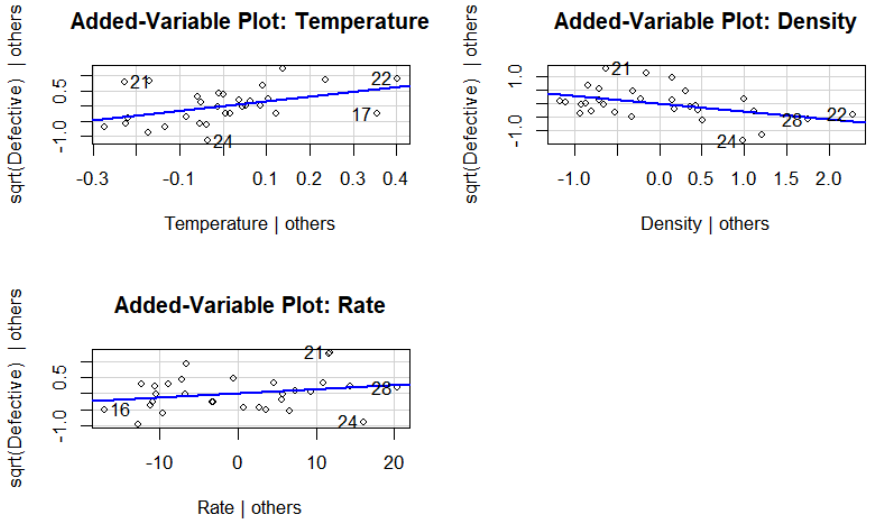
باتوجه به اینکه خط برازش شده به این نقاط به نیم ساز ربع اول نزدیک است، لذا می توان به نیکویی مدل برازش شده پی برد.

خروجی رگرسیون R

```
Call:
lm(formula = sqrt(Defective) ~ Temperature + Density + Rate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10147 -0.28502 -0.07716  0.34139  1.13951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.59297    5.26401   1.062   0.2978
Temperature   1.56516    0.66226   2.363   0.0259 *
Density      -0.29166    0.11954  -2.440   0.0218 *
Rate          0.01290    0.01043   1.237   0.2273
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

```
Residual standard error: 0.5677 on 26 degrees of freedom
Multiple R-squared: 0.943, Adjusted R-squared: 0.9365
F-statistic: 143.5 on 3 and 26 DF, p-value: 2.713e-16
```

با استفاده از این خروجی می توان فهمید که متغیر مستقل $X_۳$ تاثیر معنی داری بر تغییرات $\sqrt{Y}$ نداشته و ضریب آن برابر صفر است.



شکل ۳.۱۰: نمودار های متغیر افزوده

این مطلب در شکل ۳.۱۰ به وضوح فهمیده می شود(برای اطمینان بیشتر زیرا$X_۳$ با سایر متغیر های مستقل بر اساس نمودار پراکنش اولیه رابطه خطی دارد و لذا بهتر است اثر جزئی آن بر متغیر وابسته به تنهایی سنجیده گردد. همچنین در صورت وجود هم خطی بین متغیر های مستقل ، آماره هایt استودنت مربوط به معنی داری ضرایب آن ها دیگر قابل اعتماد نمی باشد.)

انجام تبدیلات به طور توام بر روی متغیر های وابسته و مستقل

در حالتی که توزیع متغیر های مستقل و پاسخ همگی دارای چولگی بوده و نیاز به تبدیل برای نرمال چند متغیره شدن داشته باشند، می توانیم از یکی از دو رویکرد زیر استفاده نماییم.

رویکرد۱:

این روش ترکیبی از روش باکس-کاکس چند متغیره و روش نمودار پاسخ معکوس است. در این روش ابتدا با استفاده از روش باکس-کاکس چند متغیره، تبدیل مناسب را برای متغیر های مستقل یافته و سپس به کمک روش نمودار پاسخ معکوس تبدیل مناسب را برای متغیر پاسخ پیشنهاد می دهیم.

$$(X_1, X_2, \ldots, X_p) \to \psi_M(X_1, \lambda_{x_1}), \psi_M(X_2, \lambda_{x_2}), \ldots, \psi_M(X_p, \lambda_{x_p})$$

$$Y = g(\beta_0 + \beta_1 \psi_M(X_1, \lambda_{x_1}) + \ldots, \psi_M(X_p, \lambda_{x_p}))$$

$$plot(Y, \hat{Y}) \to (\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots, \hat{\beta}_p X_p)$$

$$Y^{\lambda} = \beta_0 + \beta_1 \psi_M(X_1, \lambda_{x_1}) + \ldots, \psi_M(X_p, \lambda_{x_p}) + \varepsilon$$

رویکرد ۲:

در این روش با استفاده از متد باکس-کاکس چند متغیره تبدیلات مناسب برای Y و $X_p, \ldots, X_1$ را به دست می آوریم.

مثال: سود مجله

یک تحلیلگر علاقه مند به فهم ارتباط بین سود حاصل از فروش یک مجله و آگهی های آن است. او داده هایش را بر اساس بررسی ۳۰۰ مجله در ایالات متحده جمع و متغیر های وابسته

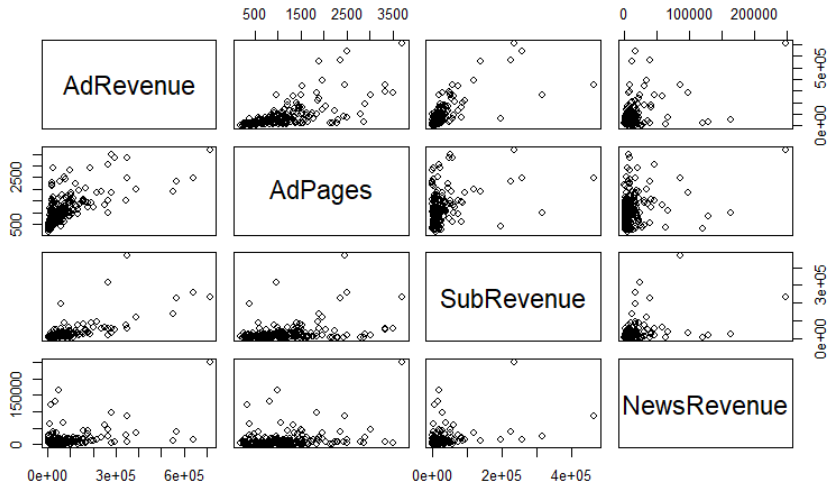
و مستقل را به شرح زیر تعریف نموده است:

Y: سود حاصل از آگهی

$X_۱$: تعداد صفحاتی که در آن ها آگهی(تبلیغات) می شود.

$X_۲$: درآمد حاصل از مشترکین

$X_۳$: درآمد حاصل از دکه ها



شکل ۱۱.۳: نمودار پراکنش

در شکل ۱۱.۳ نشان دهنده وجود چولگی در کلیه متغیر هاست. همچنین به نظر نمی رسد که بین متغیر های مستقل رابطه ی خطی وجود داشته باشد بنابراین می توان به اعمال تبدیل روی متغیر های مستقل و وابسته روی آورد.

اعمال تبدیلات روی متغیر ها با استفاده از رویکرد۱:

بر این اساس ابتدا تبدیل مناسب برای متغیر های مستقل را با استفاده از باکس-کاکس چند

متغیره به دست می آوریم.

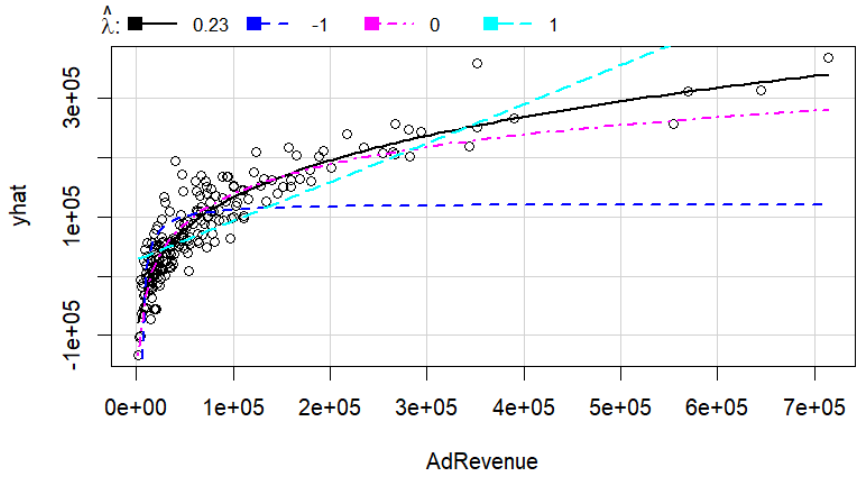
خروجی رگرسیون R

```
box.cox Transformations to Multinormality
Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
AdPages 0.1119 0.1014 1.1030 −8.7560
SubRevenue −0.0084 0.0453 −0.1864 −22.2493
NewsRevenue 0.0759 0.0333 2.2769 −27.7249
LRT df p.value
LR test, all lambda equal 0 6.615636 3 0.08521198
LR test, all lambda equal 1 1100.018626 3 0.00000000
```

طبق این خروجی می توان گفت که λ مربوط به هر سه متغیر مستقل تقریبا برابر صفر است.

پس، تبدیل مناسب یک تبدیل لگاریتمی است. بنابراین:

$$Y = g(\beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \beta_3 \log(X_3) + \varepsilon) \qquad (۳.۳)$$



شکل ۱۲.۳: نمودار معکوس

با توجه به شکل ۱۲.۳ در می یابیم که تبدیل مناسب برای Y به ازای $\lambda_y = ۰/۲۳$ حاصل می شود. بنابراین مدل نهایی به صورت زیر خواهد بود:

$$Y^{۰/۲۳} = \beta_۰ + \beta_۱\log(X_۱) + \beta_۲\log(X_۲) + \beta_۳\log(X_۳) + \varepsilon \qquad (۴.۳)$$

اما با توجه به شکل $\lambda = ۰$ نیز برای مقادیر کوچک و متوسط متغیر Y ، برازش خوبی به داده ها دارد و چون مقیاس متغیر های مستقل با وابسته یکی است، لذا تبدیل جانشین دیگر می تواند $\log(Y)$ باشد و مدل به صورت زیر در نظر گرفته شود:

$$\log(Y) = \beta_۰ + \beta_۱\log(X_۱) + \beta_۲\log(X_۲) + \beta_۳\log(X_۳) + \varepsilon \qquad (۵.۳)$$

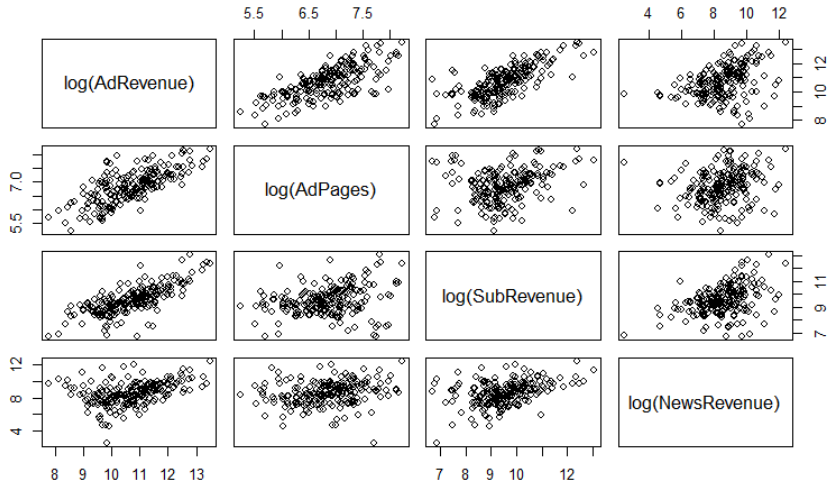
اما به هرحال برای هر کدام از دو تبدیل فوق بهتر است آن را که جدول آنالیز واریانس قابل قبول تری را می دهد، به عنوان مدل نهایی در نظر بگیریم.

اعمال تبدیل بر روی متغیر ها با استفاده از رویکرد ۲:

خروجی رگرسیون R

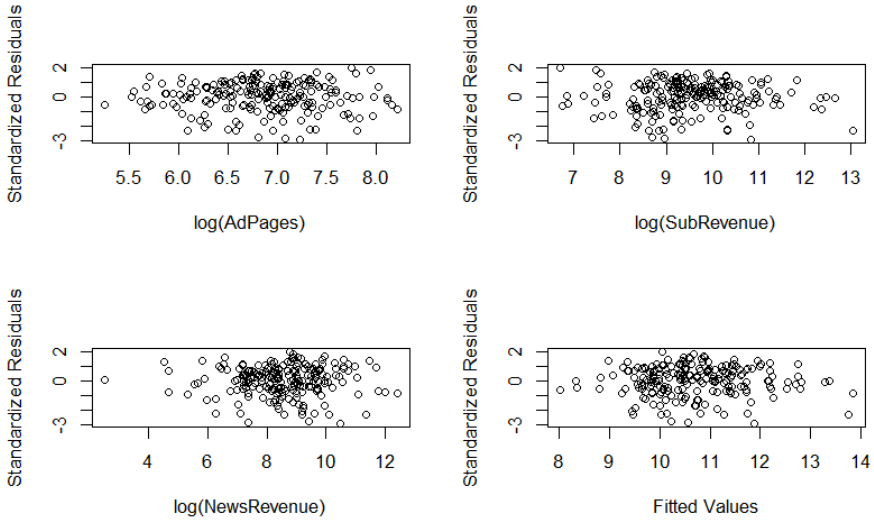
```
box.cox Transformations to Multinormality
             Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
AdRevenue       0.1071  0.0394       2.7182      -22.6719
AdPages         0.0883  0.0836       1.0566      -10.9068
SubRevenue     -0.0153  0.0362      -0.4217      -28.0413
NewsRevenue     0.0763  0.0330       2.3087      -27.9682
LRT df p.value
LR test, all lambda equal 0 13.87021 4 0.007721018
LR test, all lambda equal 1 1540.50928 4 0.000000000
```

بر اساس خروجی های بالا می توان نتیجه گرفت که روش باکس-کاکس برای کلیه متغیر های تبدیل لگاریتمی را به عنوان یک تبدیل مناسب معرفی می نماید.



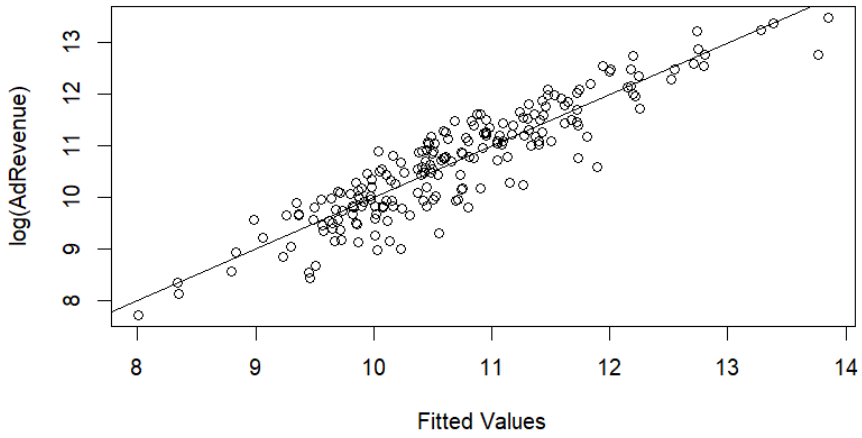
شکل ۱۳.۳: نمودار پراکنش ماتریسی

همان طور که از شکل ۱۳.۳ بر می آید، متغیر های تبدیل یافته رابطه خطی معنی دار تری باهم دارند.



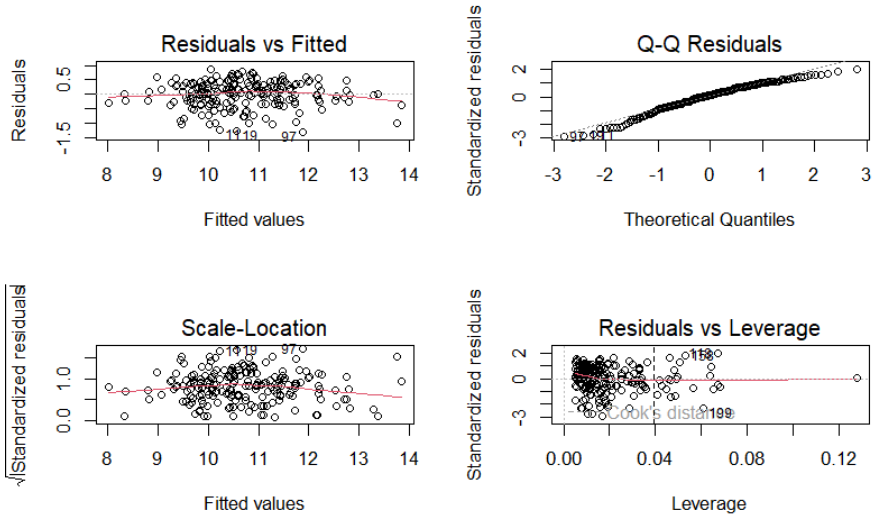
شکل ۱۴.۳: نمودار مانده های استاندارد در مقابل متغیر های جدیدو $\hat{Y}$

شکل ۱۴.۳ همگی نشان دهنده طرح تصادفی هستند. بنابراین مدل فوق مدلی مناسب برای داده ها می باشد.



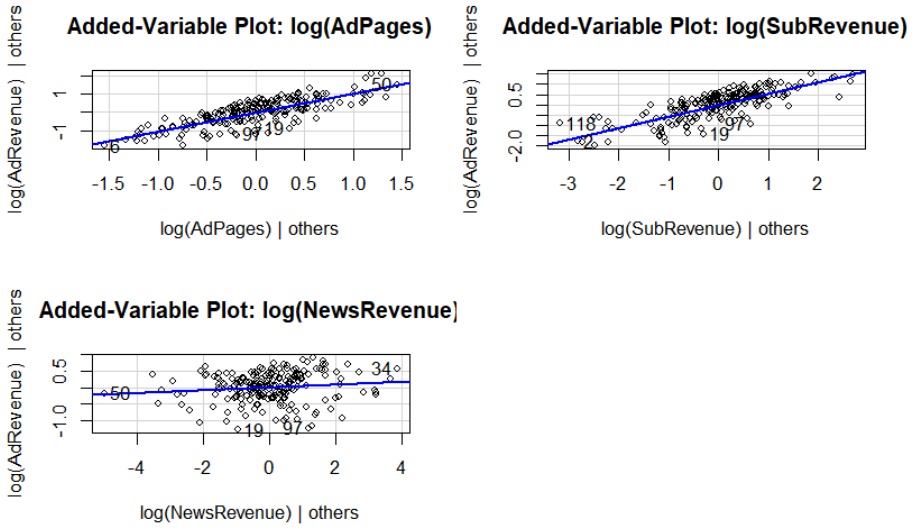
شکل ۱۵.۳: نمودار Y در مقابل $\hat{Y}$

همچنین در شکل ۱۵.۳ نمودار Y در مقابل $\hat{Y}$ برای متغیر های تبدیل یافته رسم شده و چون نقاط تقریبا بر روی نیمساز ربع اول و سوم واقعند، لذا مدل برازش کافی به داده ها دارد.



شکل ۱۶.۳: نمودار های عیب شناسی

شکل ۱۶.۳ همگی دلالت بر مناسب بودن مدل مورد بررسی دارند. البته نمودار پایین سمت چپ این شکل نشان می دهد که واریانس خطاها افزایش و سپس کاهش دارد. همچنین در نمودار پایین سمت راست خط نقطه چین عمودی نشان دهنده مرز لوریج ها یعنی $\frac{۲(p+۱)}{n} = ۰/۰۳۹$ و خط افقی تشان دهنده مرکزیت مانده های استاندارد است.



شکل ۱۷.۳: نمودار های متغیر افزوده

بر اساس شکل ۱۷.۳ می توان فهمید که $\log(X_۳)$ تاثیر معنی داری بر روی مدل ۵.۳ ندارد و بهتر است حذف گردد.بااستفاده از خروجی زیر نیز این مطلب کاملا آشکار است.

خروجی رگرسیون R

```
Call:
lm(formula = log(AdRevenue) ~ log(AdPages) + log(SubRevenue) +
log(NewsRevenue))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) −2.02894 0.41407 −4.900 1.98e−06 ***
log(AdPages) 1.02918 0.05564 18.497 < 2e−16 ***
log(SubRevenue) 0.55849 0.03159 17.677 < 2e−16 ***
log(NewsRevenue) 0.04109 0.02414 1.702 0.0903 .
---
Residual standard error: 0.4483 on 200 degrees of freedom
Multiple R−Squared: 0.8326, Adjusted R−squared: 0.8301
F−statistic: 331.6 on 3 and 200 DF, p−value: < 2.2e−16
```

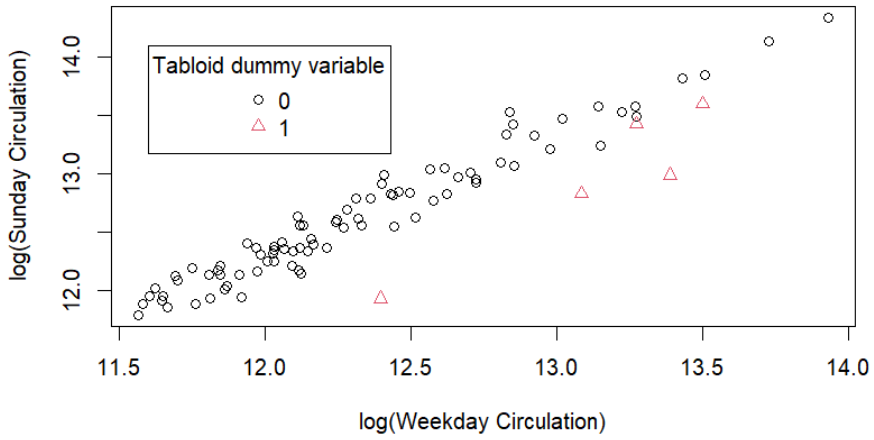
مثال: برآورد تیراژ یک روزنامه در یکشنبه

در این مثال بر اساس ۸۹ مشاهده در ایالات متحده آمریکا متغیر های زیر تعریف شده اند:

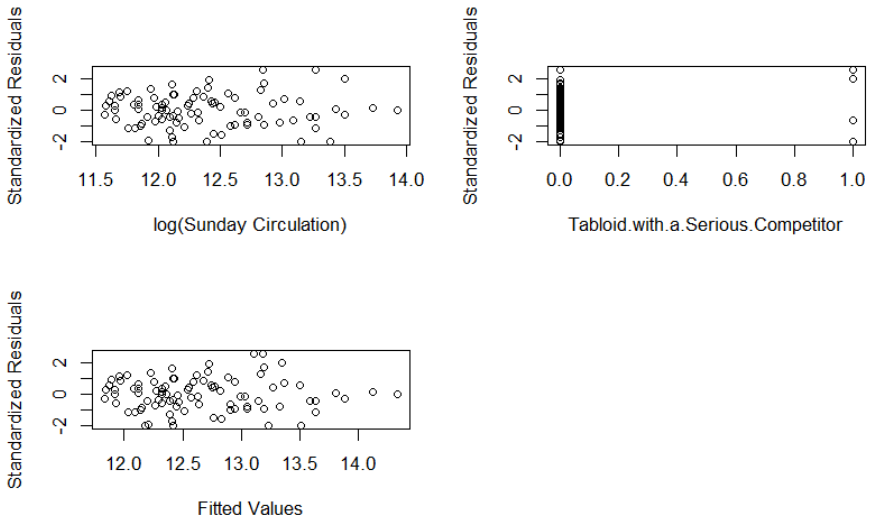
Y: لگاریتم تیراژ در یکشنبه

$X_۱$: لگاریتم تیراژ در سایر روز های هفته

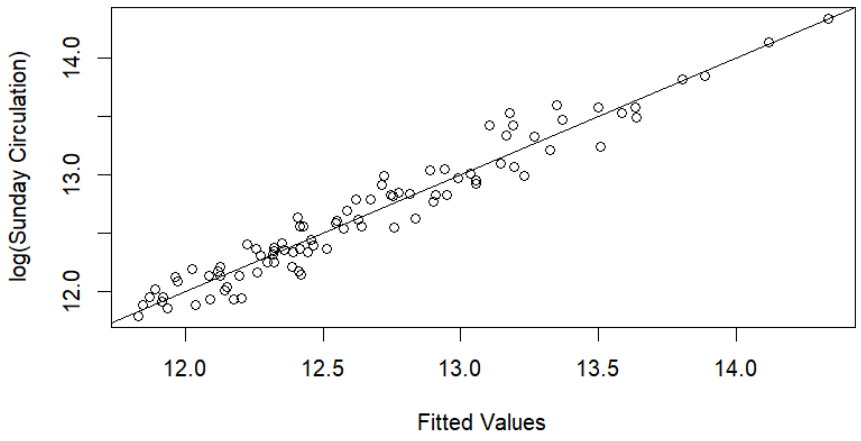
$X_۲$: متغیر دو حالتی ۰و۱ به طوریکه صفر برای حالتی است که در یک شهر تنها یک روزنامه چاپ می شود و ۱ برای حالتی است که در آن شهر روزنامه دیگری که حالتی خلاصه دارد نیز علاوه بر روزنامه اصلی چاپ می شود.



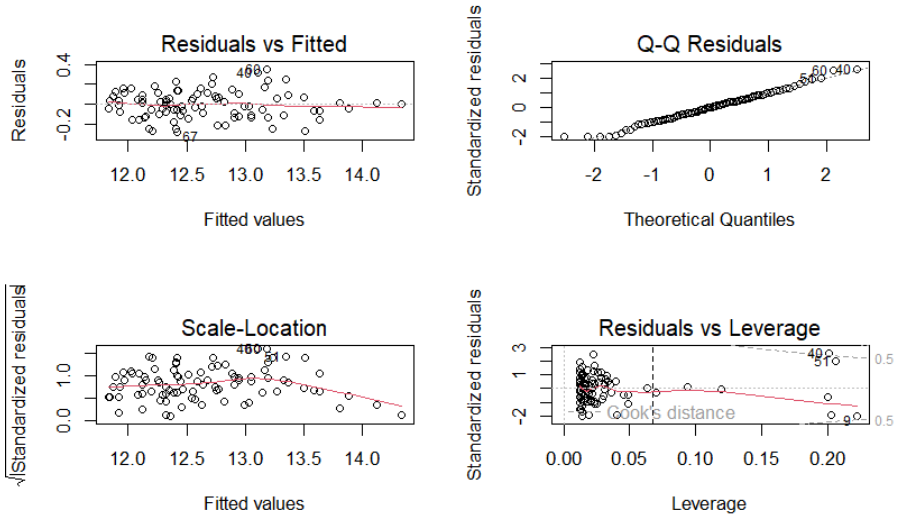
شکل ۱۸.۳: نمودار پراکنش Y در مقابل $X_۱$



شکل ۱۹.۳: نمودار مانده های استاندارد در مقابل متغیر های مستقل



شکل ۲۰.۳: نمودار پراکنش Y در مقابل $\hat{Y}$



شکل ۲۱.۳: نمودار عیب یابی رگرسیون

شکل ۱۸.۳،۱۹.۳ نشانگر طرح تصادفی هستند. با توجه به شکل ۲۰.۳ می توان به نیکویی برازش پی برد. و شکل ۲۱.۳ اعتبار مدل لگاریتمی را تایید می نمایند.

خروجی رگرسیون R

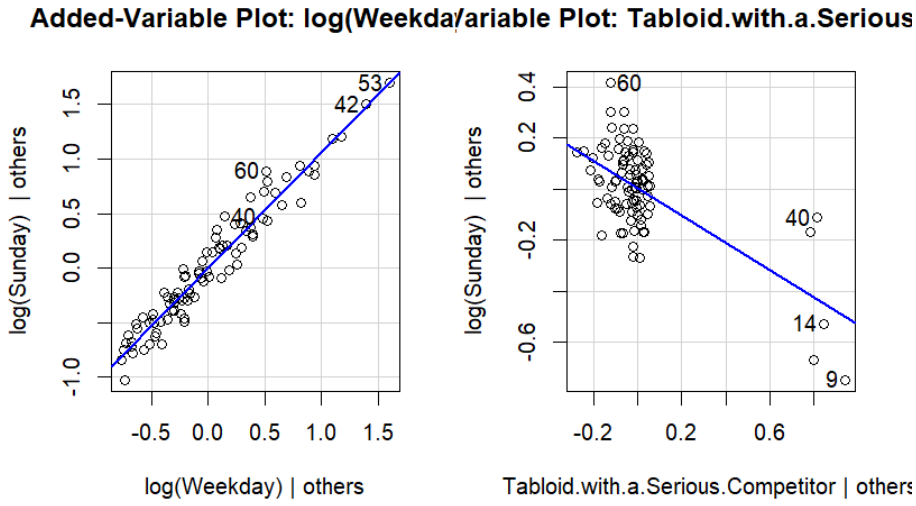
```
Call:
lm(formula = log(Sunday) ~ log(Weekday) + Tabloid.with.a.Serious.
Competitor)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.44730
0.35138 -1.273
log(Weekday) 1.06133
0.206
0.02848 37.270 < 2e-16 ***
Tabloid.with.
a.Serious.
Competitor -0.53137 ---
0.06800 -7.814 1.26e-11 ***
Residual standard error: 0.1392 on 86 degrees of freedom
Multiple R-Squared: 0.9427,
```

```
Adjusted R-squared: 0.9413
F-statistic: 706.8 on 2 and 86 DF, p-value: < 2.2e-16 ---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

براساس خروجی بالا می توان به معنی داری هر دو متغیر مستقل در مدل پی بردو بر اساس آن می توان گفت:

(آ) با ثابت بودن متغیر X_2 ، درصد افزایش متغیر پاسخ به ازای یک واحد افزایش X_1 برابر است با ۱٫۶٪

(ب) با ثابت بودن متغیر X_1 ، متغیر پاسخ به ازای یک واحد افزایش X_2 مقدار ۵۳٫۱٪ کاهش می باید.



شکل ۲۲.۳: نمودار های متغیر افزوده

با توجه به شکل ۲۲.۳ می توان به معنی داری اثر هر یک از متغیر های مستقل بر این مدل

پی برد.

در انتها ما قادر خواهیم بود که مقدار متغیر Y را به ازای تعداد تیراژ هفتگی ۲۱۰۰۰۰۰۰۰ به صورت زیر برآورد و فاصله اطمینان ۹۵٪ به دست آوریم:

خروجی رگرسیون R

```
Tabloid.with.a.Serious.Competitor=1
fit  lwr  upr
[1,] 12.02778 11.72066 12.33489
Tabloid.with.a.Serious.Competitor=0
fit  lwr  upr
[1,] 12.55915 12.28077 12.83753
```

$$X_۲ = ۱ \to \hat{Y} = ۱۲٫۰۲۸ \qquad (۱۱٫۷۲۱, ۱۲٫۳۳۵)$$

$$X_۲ = ۰ \to \hat{Y} = ۱۲٫۵۵۹ \qquad (۱۲٫۲۸, ۱۲٫۸۳۸)$$

$$X_۲ = ۱ \to \text{تعداد تیراژ یکشنبه} = \exp(۱۲٫۰۲۸)$$

$$X_۲ = ۰ \to \text{تعداد تیراژ یکشنبه} = \exp(۱۲٫۵۵۹)$$

۳.۳ هم خطی چندگانه

اگر بین دو یا چند متغیر مستقل در مدل رگرسیون چندگانه رابطه خطی وجود داشته باشد، گوییم مدل رگرسیون دارای هم خطی است. هم خطی کامل زمانی رخ می دهد که یکی از متغیر های مستقل یک تابع دقیق از یک یا چند متغیر مستقل دیگر باشد و هم خطی ناقص زمانی رخ می دهد که این تابع تقریبی باشد. در عمل هم خطی کامل رخ نخواهد داد زیرا اگر یکی از از متغیر های مستقل تابعی دقیق از سایر متغیر های دیگر باشد، خطای اندازه گیری سبب می شود که این رابطه دقیق به یک رابطه تقریبی تبدیل گردد. حال چنانچه هم خطی

تقریبی اتفاق افتد، آنگاه $|X'X| \to ۰$ و لذا ماتریس $(X'X)^{-۱}$ وجود نخواهد داشت.

منابع اصلی هم خطی

۱- انتخاب روش گردآوری داده ها

۲- گذاشتن قید هایی روی مدل یا جامعه

۳- شناسایی مدل

۴- انتخاب مدلی که تعداد متغیر های مستقل آن بیشتر از مشاهدات باشد.

در حالت کلی با زیاد شدن متغیر های پیشگو، احتمال به وجود آمدن هم خطی بین متغیر های مستقل افزایش می یابد.

اثرات هم خطی

همان طور که گفته شد، اگر هم خطی ناقص وجود داشته باشد، $|X'X| \cong ۰$ شده و لذا درایه های ماتریس $(X'X)^{-۱}$ خیلی بزرگ خواهند شد و چون $\hat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma^۲}(X'X)^{-۱}$، لذا می توان نتیجه گرفت که دقت برآورد کمترین مربعات کاهش یافته و همچنین فواصل اطمینان به دست آمده برای پارامتر ها پهن خواهند شد.گاهی اوقات ممکن است حتی علامت پارامتر های $\beta_۰, \ldots, \beta_p$ به اشتباه برآورد گردد. همچنین مقادیر آماره های t مربوط به ضرایب رگرسیون بسیار کوچک شده و غیر معنی دار به نظر می رسند زیرا

$$j = ۰, ۱, \ldots, p \qquad t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma^۲} c_{j+۱,j+۱}}}$$

که در آن $c_{j+۱,j+۱}$ $(j+۱)$ امین درایه روی قطر اصلی ماتریس $(X'X)^{-۱}$ است.

چگونه می توان به وجود هم خطی پی برد؟

گاهی اوقات آماره F جدول آنالیز واریانس معنی دار است ولی تعداد زیادی یا همه آماره های t مربوط به ضرایب متغیر های رگرسیون معنی دار نیستند. در چنین مواردی می توان به وجود هم خطی پی برد ولی بسیاری از مجموعه داده هایی که همخطی معنی داری دارند این رفتار را نشان نمی دهند و لذا این معیار چندان سودمند نمی باشد. در چنین مواردی دو معیار زیرمی توانند مورد استفاده قرار گیرند:

۱- عدد شرطی:

این آماره بر اساس مقادیر ویژه ماتریس $(X'X)$ به صورت زیر تعریف می شود:

$$K(X'X) = \sqrt{\frac{\max \lambda_i}{\min \lambda_i}}, \qquad i = ۱, \ldots, p$$

که در آن $\lambda_p, \ldots, \lambda_۱$ مقادیر ویژه ماتریس $(X'X)$ هستند.

۲- عامل تورم واریانس [۱]

عامل تورم واریانس مربوط به متغیر مستقل j ام را با VIF_j نشان داده و به صورت زیر تعریف می کنیم:

$$VIF_j = \frac{۱}{۱ - R_j^۲}$$

که در آن $R^۲$ ضریب تعیین در مدل رگرسیون X_j روی سایر متغیر های پیشگو یا مستقل می باشد.

$$R_j^۲ \to ۱ \Rightarrow VIF_j \to \infty$$

[1] Variance Inflation Factor

هرگاه $VIF_j > ۵$ $(R^۲ > ۰/۸)$ آنگاه می توان به وجود هم خطی پی برد.

همانگونه که در رگرسیون ۱ دیدیم، در مدل خطی دو متغیره $Y = \beta_۰ + \beta_۱ X_۱ + \beta_۲ X_۲ + \varepsilon$

$$\mathrm{Var}(\hat{\beta}_j) = \frac{۱}{۱ - r_{۱۲}^۲} \frac{\sigma^۲}{S_{X_j X_j}}, \qquad S_{X_j X_j} = \sum_{j=۱}^{n} (X_{ij} - \bar{X}_j)^۲ \quad j = ۱, ۲$$

بنابراین می توان نوشت:

$$\mathrm{Var}(\hat{\beta}_j) = VIF_j \frac{\sigma^۲}{S_{X_j X_j}} \Rightarrow VIF_j \to \infty \Rightarrow \mathrm{Var}(\hat{\beta}_j) \to \infty$$

دقت برآوردگر کمترین مربعات کم می شود.

رابطه اخیر در حالت کلی نیز درست بوده و می توان نشان داد:

$$Y = \beta_۰ + \beta_۱ X_۱ + \cdots + \beta_p X_p + \varepsilon$$

$$\mathrm{Var}(\hat{\beta}_j) = \frac{۱}{۱ - r_{۱۲}^۲} \frac{\sigma^۲}{S_{X_j X_j}}, \qquad j = ۱, \ldots, p$$

$$\Rightarrow \mathrm{Var}(\hat{\beta}_j) = VIF_j \frac{\sigma^۲}{S_{X_j X_j}}$$

مثال: ساخت پل

در مسئله ساخت یک پل، متغیر های زیر قابل بررسی می باشند:

Y: زمان ساخت بر حسب روز $(time)$

$X_۱$: مساحت پل برحسب ۱۰۰۰فوت مربع $(DArea)$

$X_۲$: هزینه ساخت بر حسب ۱۰۰۰ دلار $(Coust)$

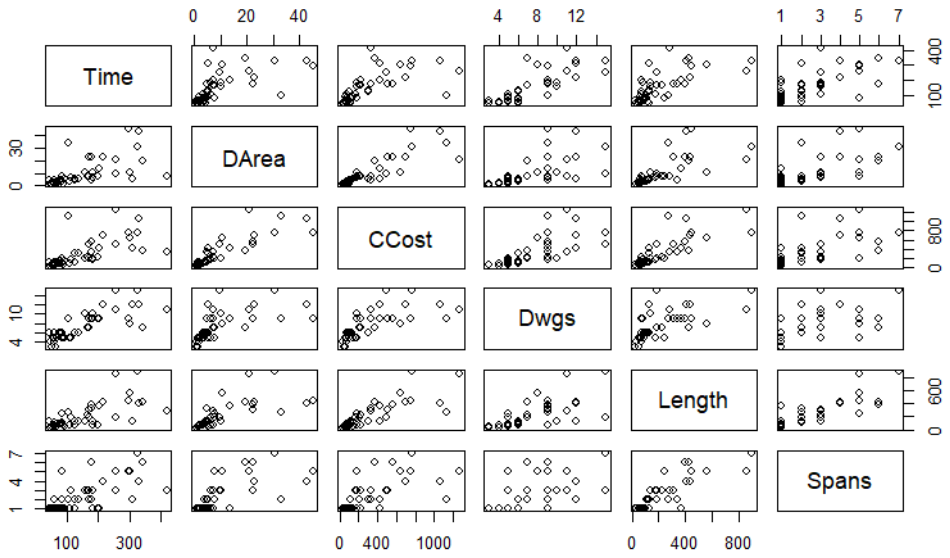
$X_۳$: تعداد طراحان پل $(DWgs)$

$X_۴$: طول پل بر حسب فوت $(Length)$

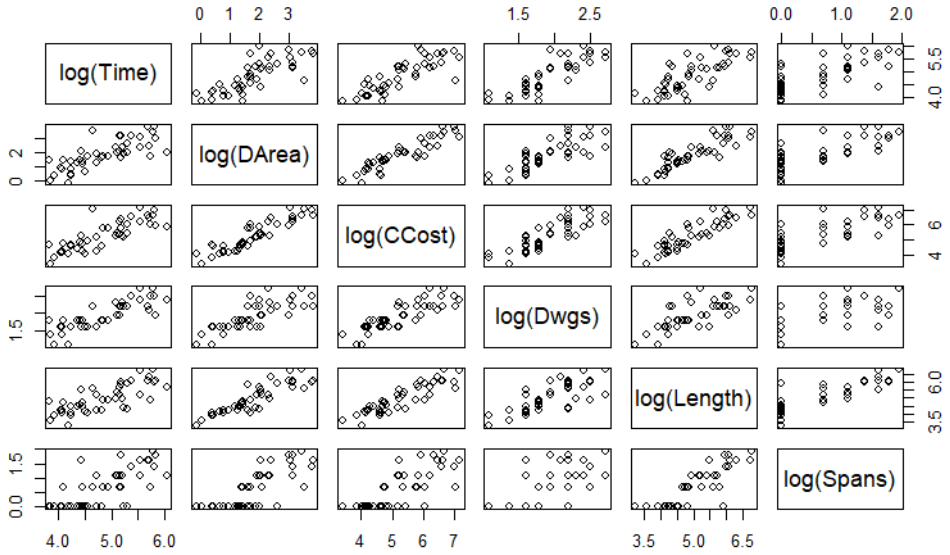
$X_۵$: تعداد طاق ها $(Spans)$

ابتدا به کمک روش باکس-کاکس چند متغیره، تبدیل مناسب جهت تبدیل داده ها به یک توزیع نرمال چند متغیره و ایجاد رابطه خطی بین آن ها را به دست می آوریم. می توان λ ی مناسب را برای کلیه متغیر ها برابر صفر در نظر گرفت و در نتیجه مدل زیر حاصل می شود:

$$\log(Y) = \beta_۰ + \beta_۱\log(X_۱) + \cdots + \beta_۵\log(X_۵) + \varepsilon$$

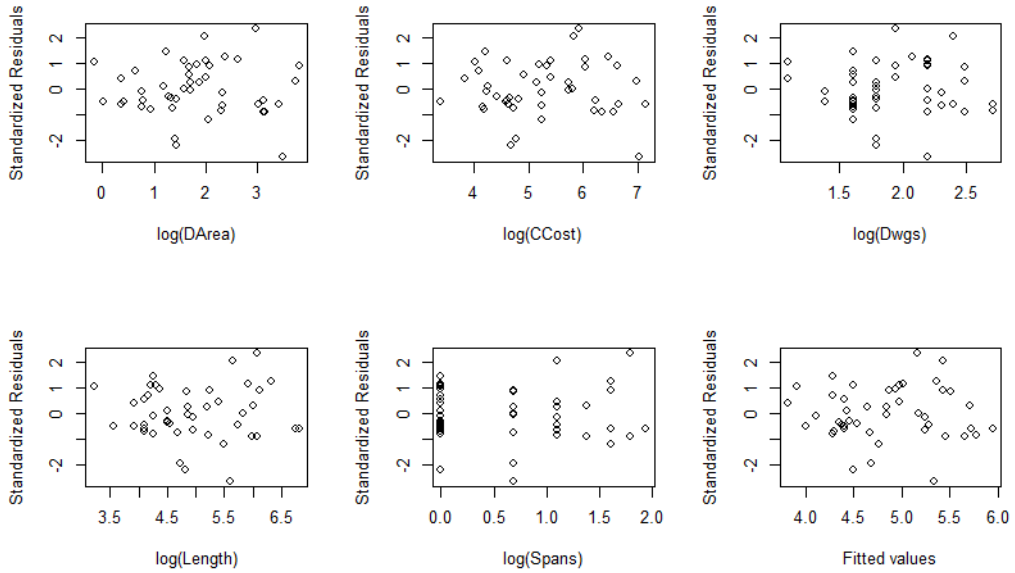


شکل ۲۳.۳: ماتریس نمودار پراکنش قبل از تبدیل



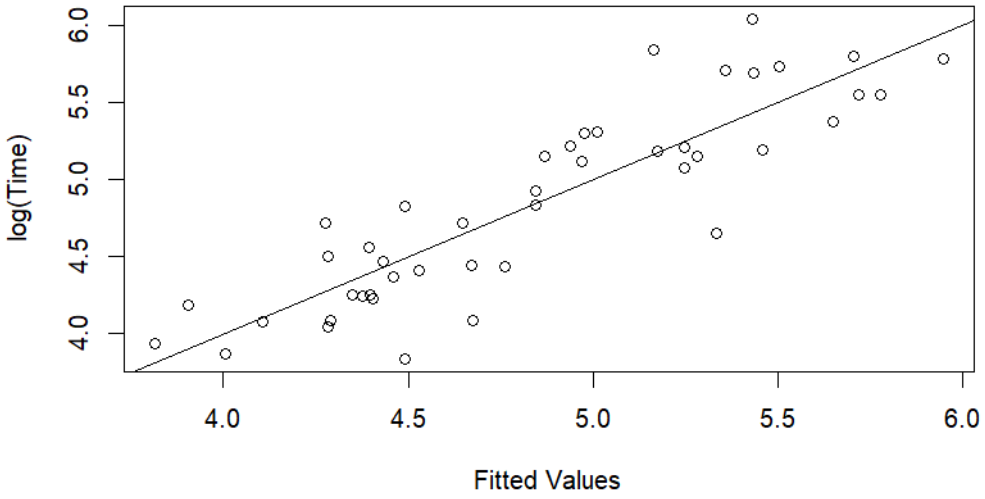
شکل ۲۴.۳: ماتریس نمودار پراکنش بعد از تبدیل

در شکل های ۲۳.۳،۲۴.۳ به ترتیب ماتریس نمودار پراکنش قبل و بعد از تبدیل را نشان می دهند.

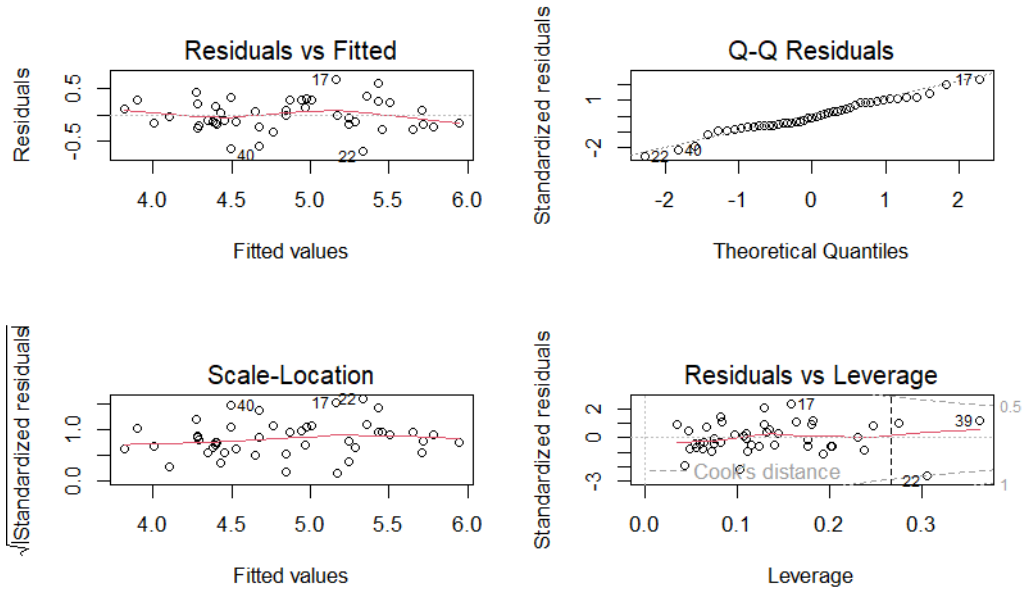


شکل ۲۵.۳: نمودار های تشخیصی

شکل ۲۵.۳ نمودار مانده های استاندارد در مقابل متغیر های وابسته و مستقل برای مدل تبدیل یافته را نشان می دهد.(همه نمودار ها نشان دهنده طرح تصادفی هستند.)



شکل ۲۶.۳: نمودار $Y^{\log(Y)}$ در مقابل $\log(\hat{Y})$



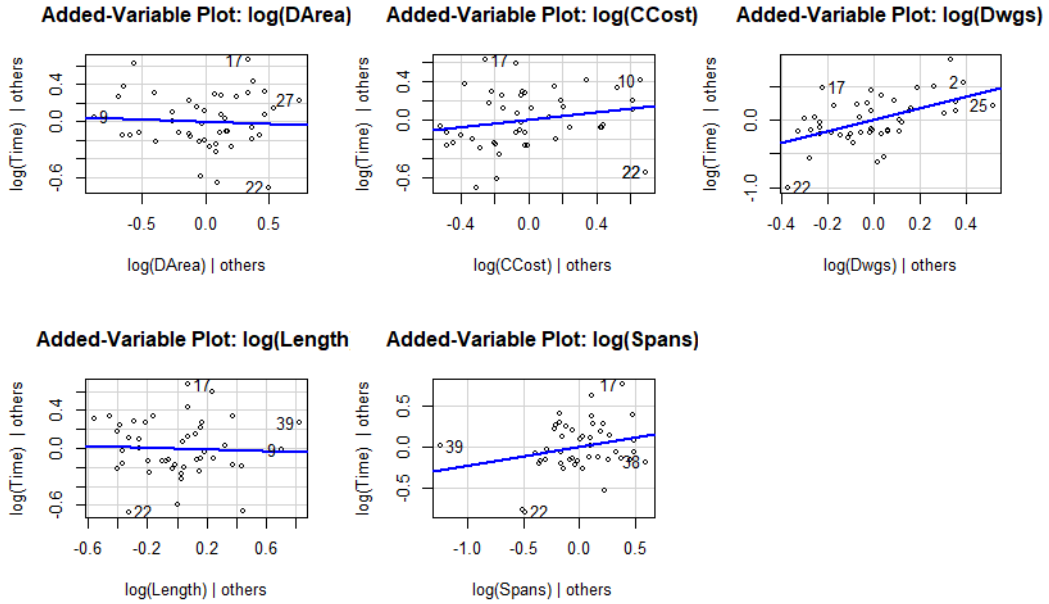
شکل ۲۷.۳: نمودار عیب یابی

نمودار های بالا همگی دلالت بر معتبر بودن مدل فوق دارند.

خروجی نرم افزار R

```
Call:
lm(formula = log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) +
log(Length) + log(Spans))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.28590 0.61926 3.691 0.00068 ***
log(DArea) −0.04564 0.12675 −0.360 0.72071
log(CCost) 0.19609 0.14445 1.358 0.18243
log(Dwgs) 0.85879 0.22362 3.840 0.00044 ***
log(Length) −0.03844 0.15487 −0.248 0.80530
log(Spans) 0.23119 0.14068 1.643 0.10835
---
Residual standard error: 0.3139 on 39 degrees of freedom
Multiple R-Squared: 0.7762, Adjusted R-squared: 0.7475
F-statistic: 27.05 on 5 and 39 DF, p-value: 1.043e-11
```

اما هنگامی که به جدول آنالیز واریانس خروجی R بر می خوریم می بینیم که با وجود اینکه آماره F در سطح بالایی معنی دار است، تنها آماره t مربوط به متغیر $X_۳$ معنی دار بوده و سایر متغیر ها دارای اثر معنی داری بر مدل رگرسیون نمی باشند. علاوه بر این علامت مربوط به ضرایب برآورد شده متغیر های مستقل $X_۱$ (مساحت پل) و $X_۴$ (طول پل) منفی است یعنی هرچه سطح (طول) یک پل افزایش یابد زمان ساخت آن کمتر می شود که این جمله مطلقا غلط است.



شکل ۲۸.۳: نمودار های متغیر افزوده

همچنین بر اساس شکل ۲۸.۳ تنها متغیر مستقل $X_۳$ دارای اثر معنی دار بر مدل بوده و سایر متغیر ها فاقد اثر معنی داری بر مدل رگرسیون نمی باشند.هرگاه دو یا چند متغیر مستقل با هم بستگی بالا وارد یک مدل رگرسیون شوند، آن ها به طور موثری اطلاعات مشابهی را در مورد متغیر وابسته حمل کرده و این باعث می شود که روش کمترین مربعات نتواند اثر جزئی هر یک از آن ها را روی متغیر پاسخ تشخیص دهد. در چنین مواقعی F جدول آنالیز واریانس بزرگ شده و سطح معنی داری آن بسیار افزایش می یابد در حالیکه ممکن است تعداد کمی از ضرایب رگرسیون معنی دار شوند. مشکل دیگر ایجاد شده در چنین مواقعی، برآورد غلط علامت برخی ضرایب رگرسیون می باشد.

خروجی نرم افزار R

```
           logDArea logCCost logDwgs logLength logSpans
logDArea      1.000    0.909   0.801     0.884    0.782
logCCost      0.909    1.000   0.831     0.890    0.775
logDwgs       0.801    0.831   1.000     0.752    0.630
logLength     0.884    0.890   0.752     1.000    0.858
logSpans      0.782    0.775   0.630     0.858    1.000
```

همان طور که دیده می شود بین متغیر های مستقل همبستگی خطی بالایی وجود دارد.

همچنین مقادیر VIF برای متغیر های مستقل عبارتند از:

$X_۱$	$X_۲$	$X_۳$	$X_۴$	$X_۵$
۷٫۱۶	۸٫۴۸	۳٫۴۱	۸٫۰۱	۳٫۸۸

فصل ۴

نحوه انتخاب متغیر ها در مدل

در این بخش روش هایی برای انتخاب بهترین مدل ممکن از میان تمام مدل های قابل انتخاب، معرفی خواهند شد. مدل رگرسیون چندگانه زیر را در نظر بگیرید:

$$Y = \beta_۰ + \beta_۱ X_۱ + \cdots + \beta_p X_p + \varepsilon$$

روش های انتخاب متغیر ها به انتخاب بهترین زیر مجموعه ممکن از متغیر های مستقل منجر می شود.

ارزیابی زیر مجموعه های نهایی متغیر های مستقل

در ادامه معیار هایی جهت ارزیابی زیر مجموعه های از متغیر های مستقل ارائه خواهد شد:

۱-معیار $R۲_{adj}$

همان طور که می دانیم اضافه نمودن متغیر های مستقل بی ربط اغلب باعث افزایش $R^۲$ می شوند ولی این اتفاق در مورد $R^۲$ تصحیح شده که به صورت زیر تعریف می شود نمی افتد:

$$R^۲_{adj} = ۱ - \frac{SSE/(n-p-۱)}{S_{yy}/(n-۱)}$$

می توان ثابت نمود که در صورتی افزایش $R^۲_{adj}$ با افزودن یک متغیر به مدل اتفاق خواهد افتادهرگاه آماره F جزئی مربوط به آن متغیر مستقل از یک بیشتر باشد. در عمل تعداد متغیر های مستقل طوری انتخاب می شود که منجر به بیشترین آماره $R^۲_{adj}$ شوند. می توان نشان داد ماکزیمم $R^۲_{adj}$ برای متغیر های مستقلی اتفاق خواهد افتاد که دارای کمترین $\frac{SSE}{n-p-۱}$ باشند.

البته چنانچه افزایش یک متغیر مستقل جدید به مدل باعث افزایش جزئی در $R^۲_{adj}$ شود برای تفسیر بهتر نتایج و سادگی بیشتر، بهتر است آن متغیر را در مدل نهایی در نظر نگیریم. به عنوان مثال فرض کنید:

$$p = ۱۰ \rightarrow R^۲_{adj} = ۰/۶۹۲ \qquad p = ۹ \rightarrow R^۲_{adj} = ۰/۶۹۱$$

$$p = ۸ \rightarrow R^۲_{adj} = ۰/۵۴۱$$

در مدل های فوق هر چند $p = ۱۰$ متغیر مستقل دارای بیشترین مقدار $R^۲_{adj}$ است ولی بهتر است که $p = ۹$ متغیر در نظر گرفته شود زیرا تغییر چندانی در $R^۲_{adj}$ ایجاد نشده است.

سه معیار زیر تنها در صورتی قابل استفاده می باشند که توزیع متغیر های پیشگو و وابسته نرمال باشد.

در صورت نرمال بودن توزیع آن ها لگاریتم تابع درستنمایی برابر است با:

$$\log L(\beta_۰, \ldots, \beta_p, \sigma^۲|Y) = -\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\sigma^۲) - \frac{۱}{۲\sigma^۲}\sum_{i=۱}^{n}(Y_i - \beta_۰ - \beta_۱X_{۱i} - \cdots - \beta_pX_{pi})^۲$$

$$\Rightarrow \log L(\hat{\beta}_۰, \ldots, \hat{\beta}_p, \hat{\sigma^۲}|Y) = -\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\sigma^۲) - \frac{۱}{۲\sigma^۲}SSE$$

$$\Rightarrow \hat{\sigma^۲}_{MLE} = \frac{SSE}{n}, \qquad \hat{\sigma^۲}_{LS} = S^۲ = \frac{SSE}{n-p-۱}$$

$$\Rightarrow \log L(\hat{\beta}_۰, \ldots, \hat{\beta}_p, \hat{\sigma^۲}|Y) = -\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\frac{SSE}{n}) - \frac{n}{۲}$$

معیار AIC [۱]

این معیار بر اساس لگاریتم تابع درستنمایی تعریف شده و نیکویی برازش را اندازه می گیرد.

AIC معیاری جهت اندازه گیری اطلاعات بوده و به صورت زیر تعریف می شود:

$$AIC = -۲\big[\log L(\hat{\beta}_۰, \ldots, \hat{\beta}_p, \hat{\sigma^۲}|Y) - (p+۲)\big]$$

$$-۲\big[-\frac{n}{۲}\log(۲\pi) - \frac{n}{۲}\log(\frac{SSE}{n}) - \frac{n}{۲} - (p+۲)\big]$$

$$= n\log(\frac{SSE}{n}) + ۲p + other\ terms$$

به طوریکه سایر جملات به نحوه برازش مدل بستگی نداشته و برای همه مدل های ممکن مشارکند .

بنابراین در نرم افزار R، AIC به صورت زیر محاسبه می شود:

$$AIC = n\log(\frac{SSE}{n}) + ۲p$$

لازم به توضیح است که بهترین مدل، مدلی است که دارای کمترین AIC باشد.

معیار AIC **تصحیح شده**(AIC_C)

اگر حجم نمونه کوچک و یا تعداد پارامتر های مدل نسبتا زیاد باشند، بهتر است از AIC_c استفاده شود. اگر $\frac{n}{p+۲} \leq ۴۰$ باشد، AIC_c باید به کار برده شود. این معیار به صورت زیر

[1] Akaikes Information criterian

محاسبه می شود:

$$AIC_c = AIC + \frac{۲(p+۲)(p+۳)}{n-p-۱}, \qquad AIC_c \overset{n\to\infty}{\to} AIC$$

معیار BIC [۲]

معیار اطلاع نیز به صورت محاسبه می شود:

$$BIC = -۲\log L(\hat{\beta}_۰, \ldots, \hat{\beta}_p, \hat{\sigma^۲}|Y) + (p+۲)\log(n)$$

مشابه معیار AIC هرچه BIC کوچک تر باشد مدل برازش بهتری به داده ها دارد.

معیار C_p **مالوس**

این معیار نخستین بار توسط مالوس پیشنهاد و به صورت زیر تعریف شد:

$$C_p^* = \frac{SSE^*}{S^۲} - (n - ۲p^*)$$

که در آن SSE_p^* مجموع مربعات خطای مدلی با p^* پارامتر از جمله $\beta_۰$ و $s^۲$ برآورد $\sigma^۲$ یا MSE با بیشترین تعداد متغیر های مستقل ممکن است. حال اگر مدلی با p^* پارامتر کفایت کند در این صورت $E(SSE_{p^*}) \cong (n-p^*)\sigma^۲$ و چون بنا به فرض $E(S^۲) = \sigma^۲$ است لذا می توان نوشت:

$$E(C_p^*) \cong \frac{(n-p^*)\sigma^۲}{\sigma^۲} - (n - ۲p^*) = p^*$$

بنابراین می توان نتیجه گرفت هرگاه مقدار C_p^* به p^* نزدیک شود، مدل مناسب خواهد بود.

[2] Bayesian Information criterion

تصمیم گیری در مورد کالکسیون همه زیر مجموعه های ممکن از متغیر های پیشگو:

برای انجام تصمیم گیری در مورد متغیر های مفید در مدل دو راه کلی زیر که کاملا متفاوتند وجود دارد:

۱- همه رگرسیون های ممکن در این روش بر اساس برازش همه ۲^p مدل رگرسیون ممکن به متغیر ها و شناسایی زیر مجموعه ای از متغیر های مستقل است که معیار $R^۲_{adj}$ برازش را ماکزیمم و یا معیار های اطلاع را مینیمم نماید. البته باید توجه داشت که معیار های متفاوت ممکن است منجر به نتایج متفاوتی گردد و هیچ معیاری به تنهایی همیشه بهترین نخواهد بود. لذا باید بیشتر از یک معیار را در نظر بگیریم.

مثال ساخت پل:

ابتدا مدلی شامل تمام متغیر ها را در نظر می گیریم:

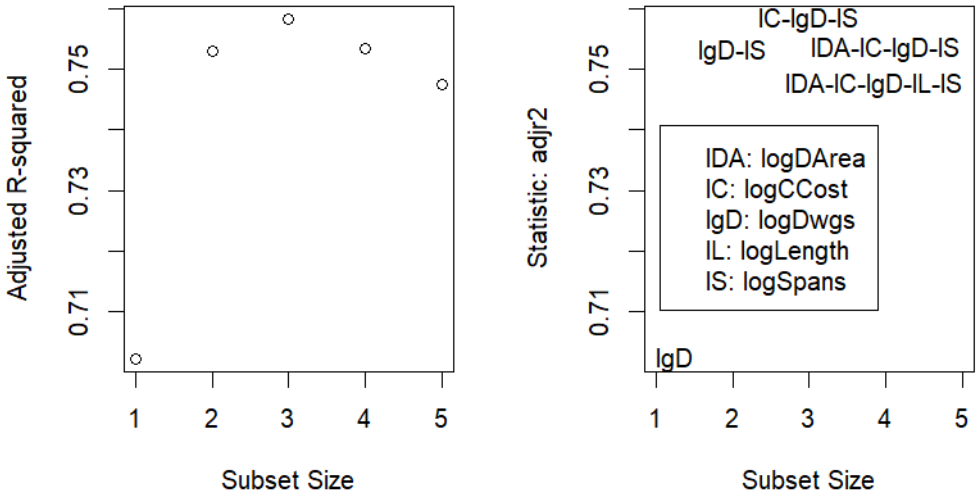
$$\log(Y) = \beta_۰ + \beta_۱\log(X_۱) + \cdots + \beta_۵\log(X_۵) + \varepsilon$$

خروجی نرم افزار R

```
Call:
lm(formula = log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) +
log(Length) + log(Spans))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.28590 0.61926 3.691 0.00068 ***
log(DArea) −0.04564 0.12675 −0.360 0.72071
log(CCost) 0.19609 0.14445 1.358 0.18243
log(Dwgs) 0.85879 0.22362 3.840 0.00044 ***
log(Length) −0.03844 0.15487 −0.248 0.80530
log(Spans) 0.23119 0.14068 1.643 0.10835
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '. 0.1 ' 1
Residual standard error: 0.3139 on 39 degrees of freedom
```

```
Multiple R-Squared: 0.7762, Adjusted R-squared: 0.7475
F-statistic: 27.05 on 5 and 39 DF, p-value: 1.043e-11
```

همان طور که مشاهده می شود با وجود معنی دار بودن آماره F جدول آنالیز واریانس، غالب آماره های t جزئی به غیر از ضریب مربوط به متغیر $X_۳$ معنی دار نمی باشند. بنابراین تحقیق در مورد انتخاب متغیر های پیشگو را با ماکزیمم سازی $R^۲_{adj}$ شروع می کنیم.



شکل ۱.۴: نمودار $R^۲_{adj}$ در مقابل متغیر های مستقل

به عنوان مثال، زیر مجموعه بهینه از متغیر های مستقل به حجم ۲ شامل $X_۵, X_۳$ و زیر مجموعه بهینه از متغیر های مستقل به حجم ۳ شامل $X_۲, X_۵, X_۳$ است.

طبق جدول دو معیار $R^۲_{adj}, AIC$ بهترین مجموعه ممکن را شامل $X_۲, X_۵, X_۳$ پیشنهاد می شود ولی، در معیار AIC_c, BIC بهترین مجموعه ممکن را شامل $X_۵, X_۳$ در نظر می گیرند.

جدول ۱.۴: مقادیر $R^2_{adj}, AIC, AIC_C, BIC$ برای زیر مجموعه های بهینه

size	Predictors	R^2_{adj}	AIC	AIC_C	BIC
1	log(Dwgs)	0.702	–94.90	–94.31	–91.28
2	log(Dwgs), log(Spans)	0.753	–102.37	–101.37	–96.95
3	log(Dwgs), log(Spans), log(CCost)	0.758	–102.41	–100.87	–95.19
4	log(Dwgs), log(Spans), log(CCost), log(DArea)	0.753	–100.64	–98.43	–91.61
5	log(Dwgs), log(Spans), log(CCost), log(DArea), log(Length)	0.748	–98.71	–95.68	–87.87

با مقایسه آماره های این جدول و تفاوت بسیار جزئی آماره های R^2_{adj}, AIC برای دو و سه متغیر مستقل بهترین انتخاب ممکن مدلی شامل X_5, X_3 است که خروجی آن در زیر آمده است.

خروجی نرم افزار R

```
Call:
lm(formula = log(Time) ~ log(Dwgs) + log(Spans))
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.66173 0.26871 9.905 1.49e-12 ***
log(Dwgs)    1.04163 0.15420 6.755 3.26e-08 ***
log(Spans)    0.28530 0.09095 3.137 0.00312 *
---
Residual standard error: 0.3105 on 42 degrees of freedom
Multiple R-Squared: 0.7642, Adjusted R-squared: 0.753
F-statistic: 68.08 on 2 and 42 DF, p-value: 6.632e-14
Call:
lm(formula = log(Time) ~ log(Dwgs) + log(Spans) + log(CCost))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.3317 0.3577 6.519 7.9e-08 ***
log(Dwgs) 0.8356 0.2135 3.914 0.000336 ***
log(Spans) 0.1963 0.1107 1.773 0.083710 .
log(CCost) 0.1483 0.1075 1.380 0.175212
---
Residual standard error: 0.3072 on 41 degrees of freedom
Multiple R-Squared: 0.7747, Adjusted R-squared: 0.7582
F-statistic: 46.99 on 3 and 41 DF, p-value: 2.484e-13
```

$$\log(Y) = \beta_۰ + \beta_۳\log(X_۳) + \beta_۵\log(X_۵)$$

۲-گزینش به روم گام به گام [۳]

این روش در واقع جنبه اصلاح شده روش رگرسیونی پیشرو است که به ما اجازه می دهد در هر مرحله متغیر های لحاظ شده در مراحل قبلی را دوباره امتحان کنیم. متغیری که در مرحله قبلی در مدل وارد شده ممکن است در مرحله بعدی بخاطر همبستگی اش با سایر متغیر های مستقل در الگو زائد به نظر برسد.برای بررسی این موضوع در هر مرحله یک آزمون F جزئی برای هر متغیر که در مدل است انجام می شود ولو اینکه این متغیر جدید ترین متغیر وارد شده به مدل رگرسیونی باشد و صرف نظر از اینکه چه وقت به مدل وارد شده است. در هر مرحله متغیر با کوچکترین F جزئی بی معنی (در صورت وجود) حذف شده و مدل با متغیر های باقی مانده دوباره برازش می شود و F های جزئی محاسبه می گردند و به طور مشابه امتحان می گردند و این کار ادامه پیدا می کند. این فرآیند تازمانی ادامه می یابد که متغیر های بیشتری نتوانند به مدل رگرسیون وارد شوند (به دلیل بی معنی بودن آماره ی F جزئی آن ها) و یا از مدل حذف نگردند. بنابراین روش گام به گام حداکثر $p+(p-۱)+\cdots+p+۱ = \frac{p(p+۱)}{p}$ مدل از میان تمام ۲^p مدل ممکن را بررسی می کند. لذا می توان گفت که روش گام به گام لزوما مدلی را ارائه می کند که بهترین مدل از لحاظ کمترین معیار های اطلاع گفته شده در قسمت قبل است. به عبارت دیگر نتایج این روش روی داده های مثال قبل و همچنین روش های پیش رو و پس رو همه بهترین مدل ممکن را مدلی شامل متغیر های $X_۲, X_۵, X_۳$ معرفی

[3] Stepwise Selection

خواهند نمود.