Contents

Chapter 8

Residual Analysis

In this chapter, we use residulas to detect departure from the model assumptions and suggesting some procedures to deal with these problems.

8.1 Introduction

Recall validity of many of the inferences associated with a regression analysis depends on the error term, ϵ, satisfying certain assumptions.

- $\epsilon \sim N(0, \sigma^2)$
- σ^2 is constant.
- All pairs of error terms are uncorrelated

8.2 Regression Residuals

The error term in a multiple regression model is:

$$\epsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)$$

The exact value of ϵ cannot be calculated. (Why?)

After using the data to obtain least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ of the regression coefficients, we can estimate the value of ϵ associated with each y-value using the corresponding regression residual, that is, the deviation between the observed and the predicted value of ϵ:

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

The **regression residual** is the observed value of the dependent variable minus the predicted value, or:

$$\hat{\epsilon} = y - \hat{y} = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k)$$

We use residuals that estimate the true random error to to check the regression assumptions and refer to **residual analyses**.

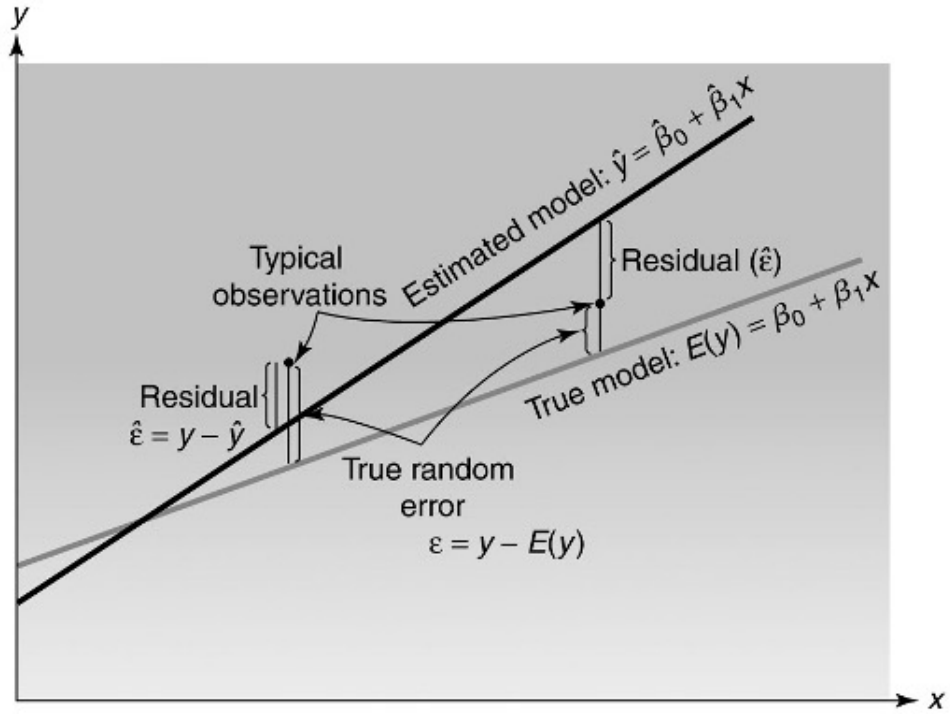


Figure 8.1: Actual random y error ϵ and regression residual $\hat{\epsilon}$

Properties of Regression Residuals

1. The mean of the residuals is equal to 0. This property follows from the fact that the sum of the differences between the observed y-values and their least squares predicted $\hat{y}$ values is equal to 0.

$$\sum \hat{\epsilon}_i = \sum (y_i - \hat{y}_i) = 0$$

2. The standard deviation of the residuals is equal to the standard deviation of the fitted regression model, s. This property follows from the fact that the sum of the squared residuals is equal to SSE, which when divided by the error degrees of freedom is equal to the variance of the fitted regression model, s^2. The square root of the variance is both the standard deviation of the residuals and the standard deviation of the regression model.

$$SSE = \sum \hat{\epsilon}_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$s = \sqrt{\frac{\sum \hat{\epsilon}_i^2}{n - (k+1)}} = \sqrt{\frac{SSE}{n - (k+1)}}$$

Example 8.2.1 *Refer to Example 8.1 from the textbook, data set: OLYMPIC. Consider a regression model relating cholesterol level y to fat intake x. Calculate the regression residuals for*

a. the straight-line (first-order) model

b. the quadratic (second-order) model

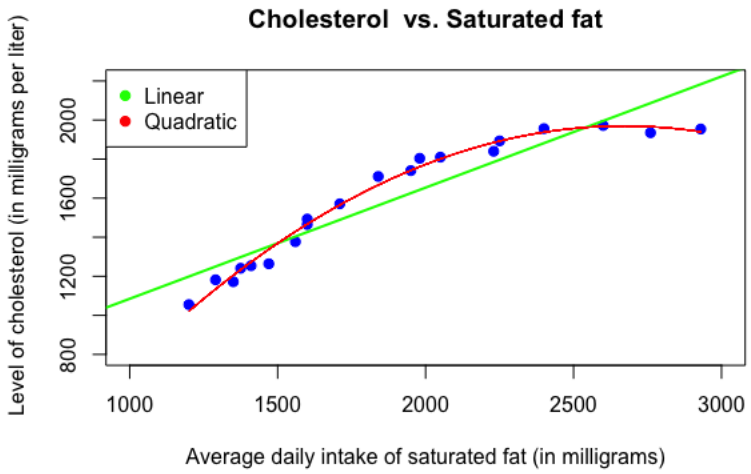
For both models, show that the sum of the residuals is 0.

$y =$ level of cholesterol (in milligrams per liter)

$x =$ average daily intake of saturated fat (in milligrams)

$n = 20$

To increase our information about the sample data, we look at the scatterplot.



a) The RStudio printout for the regression analysis of the first-order model:

$$\hat{y} = 515.70497 + 0.56919\text{FAT}$$

```
> summary(model1)

Call:
lm(formula = CHOLESTEROL ~ FAT, data = OLYMPIC)

Residuals:
     Min       1Q   Median       3Q      Max
-229.429  -73.073    7.498   85.641  161.300

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 515.70497   99.97886   5.158 6.60e-05 ***
FAT           0.56919    0.05145  11.062 1.85e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 115 on 18 degrees of freedom
Multiple R-squared:  0.8718,    Adjusted R-squared:  0.8646
F-statistic: 122.4 on 1 and 18 DF,  p-value: 1.848e-09

> residuals(model1)
        1         2         3         4         5         6         7         8         9
 -67.95904 -112.11040  -88.41310   66.59230   81.98149  147.98689  161.30040   55.00310   74.24094
        10        11        12        13        14        15        16        17        18
-229.42934 -143.73201  -57.34013  -64.26175  -26.64013   38.59230  115.37608  127.45716   96.61932
        19        20
 -23.59690 -151.66718
```

a) The RStudio printout for the regression analysis of the second-order model

```
> summary(residuals(model1))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-229.429  -73.073    7.498    0.000   85.641  161.300
> anova(model1)
Analysis of Variance Table

Response: CHOLESTEROL
          Df  Sum Sq Mean Sq F value    Pr(>F)
FAT        1 1617913 1617913  122.37 1.848e-09 ***
Residuals 18  237980   13221
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\hat{y} = -1.159e + 03 + 2.344e + 00\text{FAT} + -4.390e - 04\text{FAT}^2$$

```
> summary(model2)

Call:
lm(formula = CHOLESTEROL ~ FAT + I(FAT^2), data = OLYMPIC)

Residuals:
   Min     1Q Median     3Q    Max
-73.62 -21.84   5.29  20.34  48.20

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.159e+03  1.306e+02  -8.879 8.59e-08 ***
FAT          2.344e+00  1.354e-01  17.311 3.12e-12 ***
I(FAT^2)    -4.390e-04  3.326e-05 -13.197 2.32e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.28 on 17 degrees of freedom
Multiple R-squared:  0.9886,    Adjusted R-squared:  0.9873
F-statistic: 736.9 on 2 and 17 DF,  p-value: < 2.2e-16

> residuals(model2)
          1           2           3           4           5           6           7           8
 48.1962135 -32.9035274 -73.6207198  25.8696550   5.8740523  43.7577072  43.3802648 -44.5631680
          9          10          11          12          13          14          15          16
 18.4995316  14.3268103  33.7722555   7.4044447 -18.8425051 -51.8616111  -2.1303450  -1.0589664
         17          18          19          20
  9.1447682   0.8919297   4.7059752 -30.8427653
```

```
> summary(residuals(model2))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -73.62  -21.84    5.29    0.00   20.34   48.20
> anova(model2)
Analysis of Variance Table

Response: CHOLESTEROL
          Df  Sum Sq Mean Sq F value    Pr(>F)
FAT        1 1617913 1617913 1299.66 < 2.2e-16 ***
I(FAT^2)   1  216817  216817  174.17 2.318e-10 ***
Residuals 17   21163    1245
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For both models the sum of the residuals is 0.

In order to detect problems with the assumptions, we show how to plot the residuals to reveal model inadequacies. The regression residual can be plotted on the vertical axis against one of the independent variables on the horizontal axis, or against the predicted value $\hat{y}$ (which is a linear function of the independent variables). If the assumptions concerning the error term ϵ are satisfied, we expect to see residual plots that have no trends, no dramatic increases or decreases in variability, and only a few residuals (about 5%) more than 2 estimated standard deviations ($2s$) of ϵ above or below 0.

8.3 Detecting Lack of Fit

Assume that the following general linear model is correctly specified (i.e., that the terms in the model accurately represent the true relationship of y with the independent variables).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

Since $E(\epsilon) = 0 \longrightarrow E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$

If we assumee a **misspecified model** with mean denoted by $E_m(y)$ so that $E(y) \neq E_m(y)$. The hypothesized equation for the misspecified model is $y = E_m(y) + \epsilon$; thus, $\epsilon = y - E_m(y)$. It is easy to see that, for the misspecified model, $E(\epsilon) = E(y) - E_m(y) \neq 0$. That is, for misspecified models, the assumption of $E(\epsilon) = 0$ will be violated.

Detecting Model Lack of Fit with Residuals

1. Plot the residuals, $\hat{\epsilon}$, on the vertical axis against each of the independent variables, $x_1, x_2, ..., x_k$, on the horizontal axis.

2. Plot the residuals, $\hat{\epsilon}$, on the vertical axis against the predicted value, $\hat{y}$, on the horizontal axis.

3. In each plot, look for trends, dramatic changes in variability, and/or more than 5% of residuals that lie outside $2s$ of 0. Any of these patterns indicates a problem with model fit.

Example 8.3.1 *Plot the residuals for model 1 for Example 8.2.1 against fat intake (placing x along the horizontal axis).*

a) What does the plot suggest about a potential lack of fit of the first-order model? How would you modify the model?

b) Construct a residual plot for model 2 similar to the one in part a. What does the plot suggest about the fit of the quadratic model?

a)

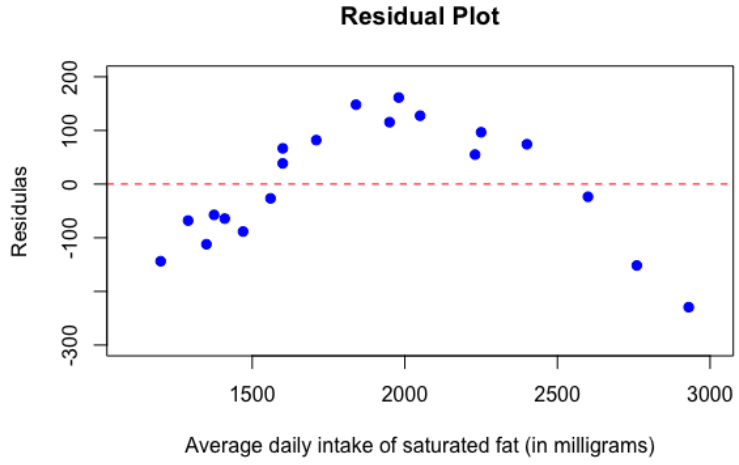


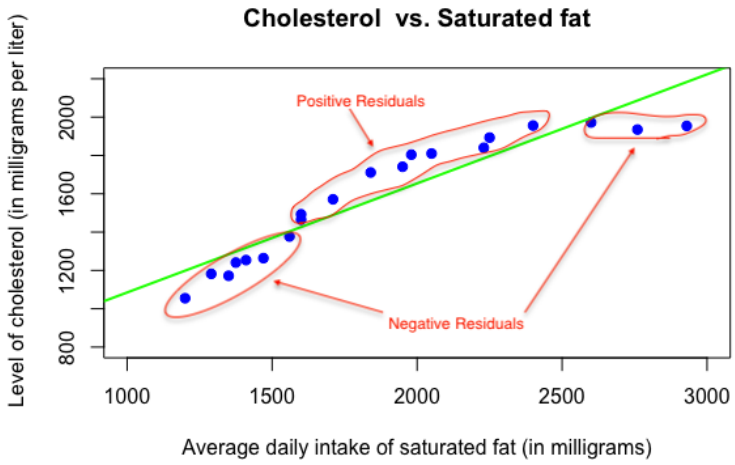
Figure 8.2: RStudio plot of residuals for the first-order model

The residula plot suggests that the relationship is curved (i.e. a quadratic model). We can see the parabolic distribution of the residuals about their mean.

All residuals tend to be positive for athletes with intermediate levels of fat intake and negative for the athletes with either relatively high or low levels of fat intake.

Hence, the assumption $E(\epsilon) = 0$ is violated due to a misspecified model. The parabolic trend suggests that the addition of a second-order (quadratic) term may improve the fit of the model.

Consider model 1 for this data set.



Because of the clear curvilinear trend of the data, the observations with y-values below the predicted values or fitted line for high and low levels of fat intake, x, have negative residuals and the observations with y-values above the fitted line for intermediate levels of x have positive residuals. This trend can be eliminated by fitting a second-order model to the data.

b)

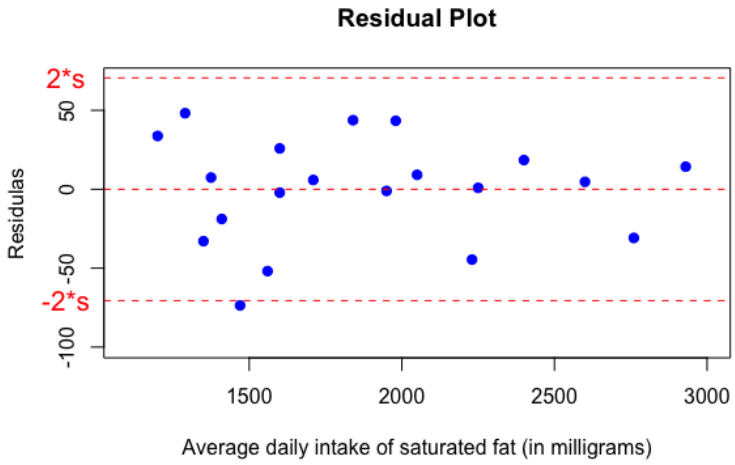


Figure 8.3: RStudio plot of residuals for the quadratic model

The residual plot does not show any distinctive pattern and almost all of the residuals lie within $2s$ of the mean (0), and the

variability around the mean is consistent for both small and large x-values. Also, the the quadratic term (β_2) in model 2 is highly significant. For this model, the assumption of $E(\epsilon) = 0$ is reasonably satisfied.

Partial Residuals

Partial residuals measure the influence of x_j on the dependent variable y after the effects of the other independent variables $(x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_k)$ have been removed or accounted for. If the partial residuals $\hat{\epsilon}^*$ are regressed against x_j in a straight-line model, the resulting least squares slope is equal to $\hat{\beta}_j$, the β estimate obtained from the full model. Therefore, when the partial residuals are plotted against x_j, the points are scattered around a line with slope equal to $\hat{\beta}_j$. Unusual deviations or patterns around this line indicate lack of fit for the variable x_j.

A plot of the partial residuals versus x_j often reveals more information about the relationship between y and x_j than the usual residual plot. A partial residual plot usually indicates more precisely how to modify the model.

The set of **partial regression residuals** for the j th independent variable x_j is calculated as follows:

$$\begin{aligned}\hat{\epsilon}^* &= y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_{j-1} x_{j-1} + \hat{\beta}_{j+1} x_{j+1} + ... + \hat{\beta}_k x_k) \\ &= \hat{\epsilon} + \hat{\beta}_j x_j\end{aligned}$$

where $\hat{\epsilon} = y - \hat{y}$ is the usual regression residual.

Example 8.3.2 *Refer to Example 8.4 from the text book. Data set: COFFEE2*

y = weekly demand for a house brand of coffee at supermarket chain stores (in pounds)

x_1 = price p (in dollars/pound)

$x_2 = \begin{cases} 1 & \text{if advertisment used} \\ 0 & \text{if not} \end{cases}$

$n = 11$

Consider the following model:

$$E(y) = \beta_0 + \beta_1 p + \beta_2 x_2$$

(a) Fit the model to the data. Is the model adequate for predicting weekly demand y?

Using the RStudio printout for the regression analysis:

$$\hat{y} = 2400.18 - 456.30p + 70.18AD$$

The F-value and the corresponding p-value for testing the overall adequacy of the model

$H_0 : \beta_1 = \beta_2 = 0$

H_a : at least one $\beta \neq 0$

```
> Model1=lm(DEMAND~PRICE+AD)
> summary(Model1)

Call:
lm(formula = DEMAND ~ PRICE + AD)

Residuals:
   Min     1Q Median     3Q    Max
-55.52 -37.40 -20.37  31.44  92.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2400.18      68.91   34.83  < 2e-16 ***
PRICE        -456.30      16.81  -27.14  < 2e-16 ***
AD             70.18      21.27    3.30  0.00377 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.88 on 19 degrees of freedom
Multiple R-squared:  0.9752,    Adjusted R-squared:  0.9726
F-statistic: 373.7 on 2 and 19 DF,  p-value: 5.568e-16

> anova(Model1)
Analysis of Variance Table

Response: DEMAND
          Df  Sum Sq Mean Sq F value    Pr(>F)
PRICE      1 1832209 1832209  736.53 < 2.2e-16 ***
AD         1   27090   27090   10.89  0.003765 **
Residuals 19   47265    2488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

are: F-statistic: 373.7 on 2 and 19 DF, p-value: $5.568e-16$ which can be concluded the model contributes information for the prediction of weekly demand, y.

(b) Plot the residuals versus p. Do you detect any trends?

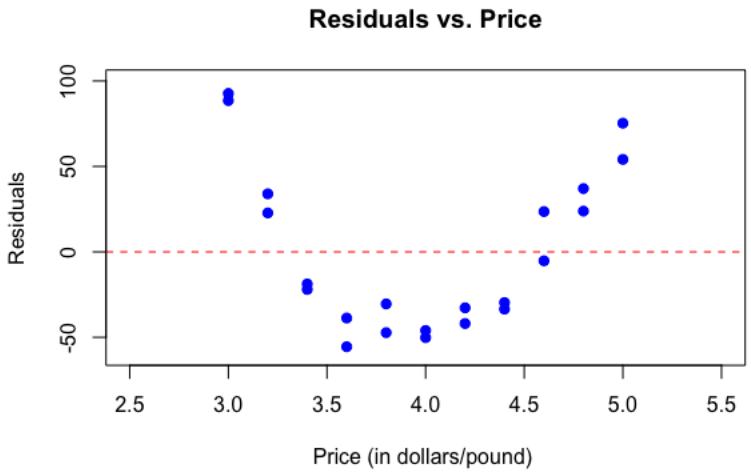


Figure 8.4: RStudio plot of residuals against price for demand model

The scatterplot of the residuals against price for demand model shows a lack of fit because of a clear parabolic trend. Therefore, the weekly demand – price relationship is curvilinear, not linear.

Note that the appropriate transformation on price, $\frac{1}{p}$ is not evident from the plot.

Note: In general, a residual plot will detect curvature if it exists, but may not reveal the appropriate transformation.

(c) Construct a partial residual plot for the independent variable p. What does the plot reveal?

The following plot is a partial residual plot or a component plus residual plot for the two independent variables, p and AD. The focus is on existance of any nonlinear trends.

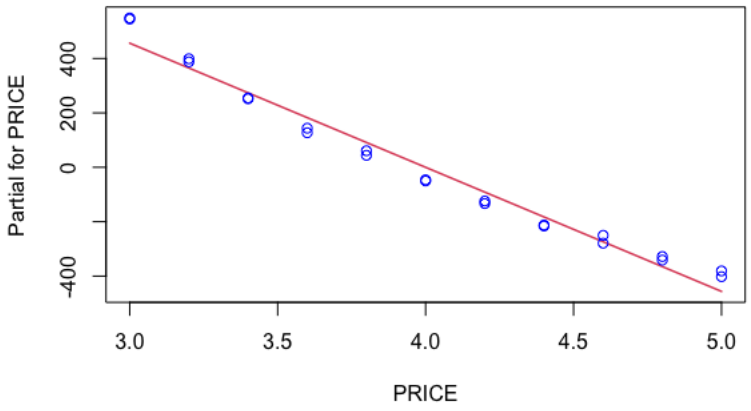


Figure 8.5: RStudio partial residual plot Price in Model 1

The plot suggests the appropriate transformation on price is either $\frac{1}{p}$ or e^{-p}.

(d) Fit the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $x_1 = \frac{1}{p}$. Has the predictive ability of the model improved?

```
Call:
lm(formula = DEMAND ~ x1 + AD)

Residuals:
    Min      1Q  Median      3Q     Max
-21.353  -6.721  -3.707   8.259  23.356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1217.343     14.898  -81.71  < 2e-16 ***
x1           6986.507     56.589  123.46  < 2e-16 ***
AD             70.182      4.732   14.83 6.71e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.1 on 19 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9986
F-statistic:  7731 on 2 and 19 DF,  p-value: < 2.2e-16
```

Figure 8.6: RStudio regression printout for demand model with transformed price

The small p-value of $p - value :< 2.2e - 16$ for testing the overall adequacy of the model indicates that the model is adequate for predicting y.

Question: Compare the values of R^2 and s for Models 1 and 2 and explain your findings.

8.4 Detecting Unequal Variances

Recall that one of the assumptions necessary for the validity of regression inferences is that the error term ϵ has constant constant variance for all levels of the independent variable(s), $var(\epsilon) = \sigma^2$.

Variances that satisfy this property are called **homoscedastic**. Unequal variances for different settings of the independent variable(s) are said to be **heteroscedastic**.

Residuals plots can be used to detect the presence of heteroscedasticity. Often times the reason for the presence of heteroscedasticity is that the variance of the response y is a function of its mean $E(y)$. A residual plot against $\hat{y}$ can be used to check the presence of heteroscedasticity.

- If the response y is a count that has a Poisson distribution, the variance will be equal to the mean $E(y)$.

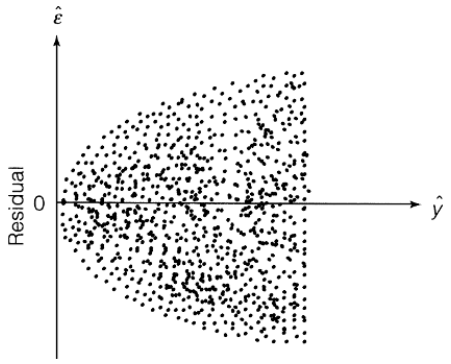


Figure 8.7: A plot of residuals for poisson data

- Many responses are proportions (or percentages) generated by binomial exper- iments.

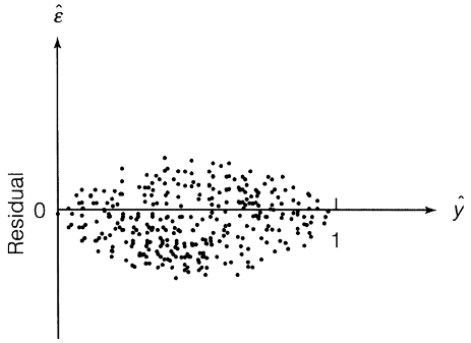


Figure 8.8: A plot of residuals for binomial data (proportions or percentages)

- Multiplicative Models when the response is written as the product of its mean and the random error component; $y = E(y) \times \epsilon$.

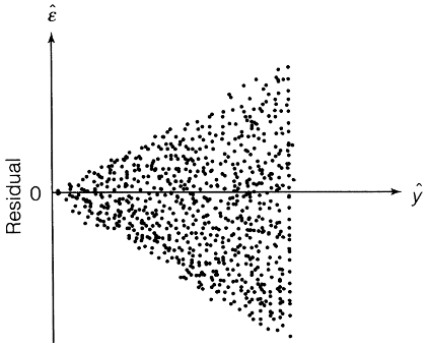


Figure 8.9: A plot of residuals for data subject to multiplicative errors

When the variance of y is a function of its mean, we can often satisfy the least squares assumption of homoscedasticity by transforming the response to some new response that has a constant variance. These are called variance-stabilizing transformations.

Stabilizing transformations for heteroscedastic responses

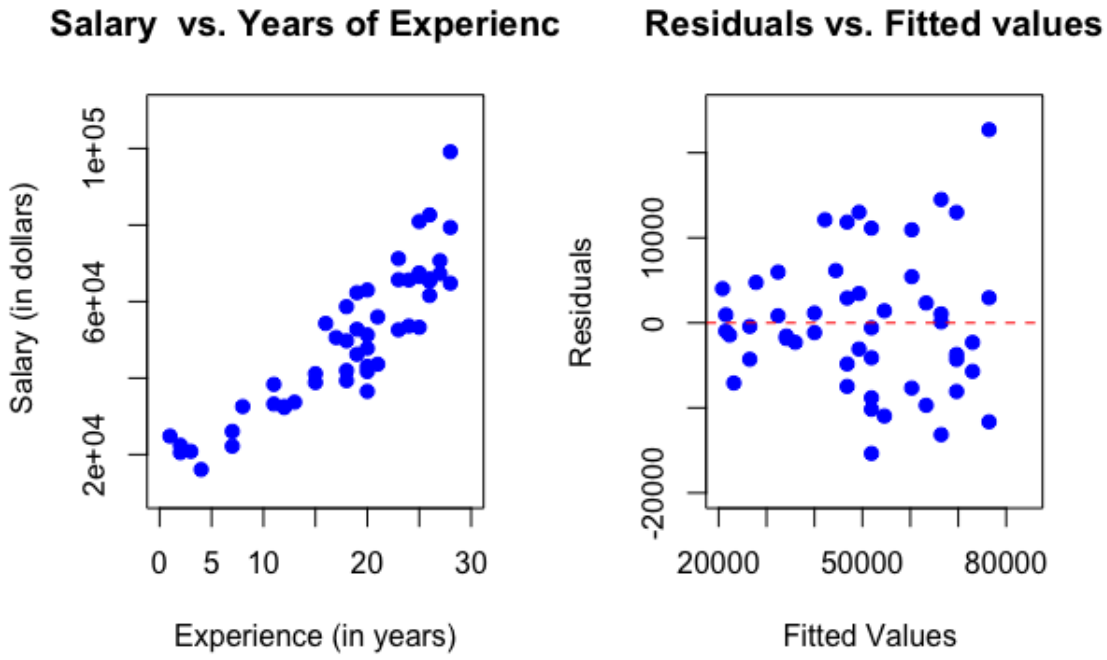
Type of Response	Variance	Stabilizing Transformation
Poisson	$E(y)$	$\sqrt{y}$
Binomial proportion	$\frac{E(y)[1-E(y)]}{n}$	$sin^{-1}\sqrt{y}$
Multiplicative	$[E(y)]^2\sigma^2$	$ln(y)$

Example 8.4.1 *Refer to Example 8.5 from the text book. Data set: SOCWORK*

a) Fit the second-order model to the the data, create a scatterplot of the data points, and a residual plot against $\hat{y}$ and interpret the results.

y = Salary ($)

x = Years of Experience



The residual plot shows the presence of heteroscedasticity. Note that the size of the residuals increases as the estimated mean salary increases. This residual plot indicates that a multiplicative model may be appropriate.

```
> Model4=lm(SALARY~EXP+I(EXP^2))
> summary(Model4)

Call:
lm(formula = SALARY ~ EXP + I(EXP^2))

Residuals:
     Min       1Q   Median       3Q      Max
-15360.3  -4703.4   -783.5   3872.7  22716.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20242.12    4422.59   4.577 3.46e-05 ***
EXP           522.30     616.68   0.847  0.40131
I(EXP^2)       53.01      19.57   2.708  0.00941 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8123 on 47 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.8078
F-statistic:   104 on 2 and 47 DF,  p-value: < 2.2e-16
```

Figure 8.10: RStudio regression printout for second-order model of salary

b) Use the natural log transformation on the dependent variable, and relate $ln(y)$ to years of experience, x, using the second-order model:

$$ln(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Evaluate the adequacy of the model.

```
> y=log(SALARY)
> Model5=lm(y~EXP+I(EXP^2))
> summary(Model5)

Call:
lm(formula = y ~ EXP + I(EXP^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.35492 -0.09022 -0.01778  0.09756  0.26265

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.843e+00  8.479e-02 116.079  < 2e-16 ***
EXP         4.969e-02  1.182e-02   4.203 0.000117 ***
I(EXP^2)    9.415e-06  3.753e-04   0.025 0.980091
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1557 on 47 degrees of freedom
Multiple R-squared:  0.8635,    Adjusted R-squared:  0.8577
F-statistic: 148.7 on 2 and 47 DF,  p-value: < 2.2e-16
```

Figure 8.11: RStudio regression printout for second-order model of natural log of salary

The residual plot, Figure 8.12, indicates that the logarithmic transformation has significantly reduced the heteroscedasticity. Note that the cone shape is gone; there is no apparent tendency of the residual variance to increase as mean salary increases. $R_a^2 = 0.8577$ indicates that about 86% of the variation in $ln(salary)$ is accounted for by the model.

Although the global F -value ($F = 148.7$) and its associated p-value $< 2.2e - 16$ indicate that the model significantly improves upon the sample mean as a predictor of ln(salary), the second-order term does not contribute information for the prediction of ln(salary). (Why?)

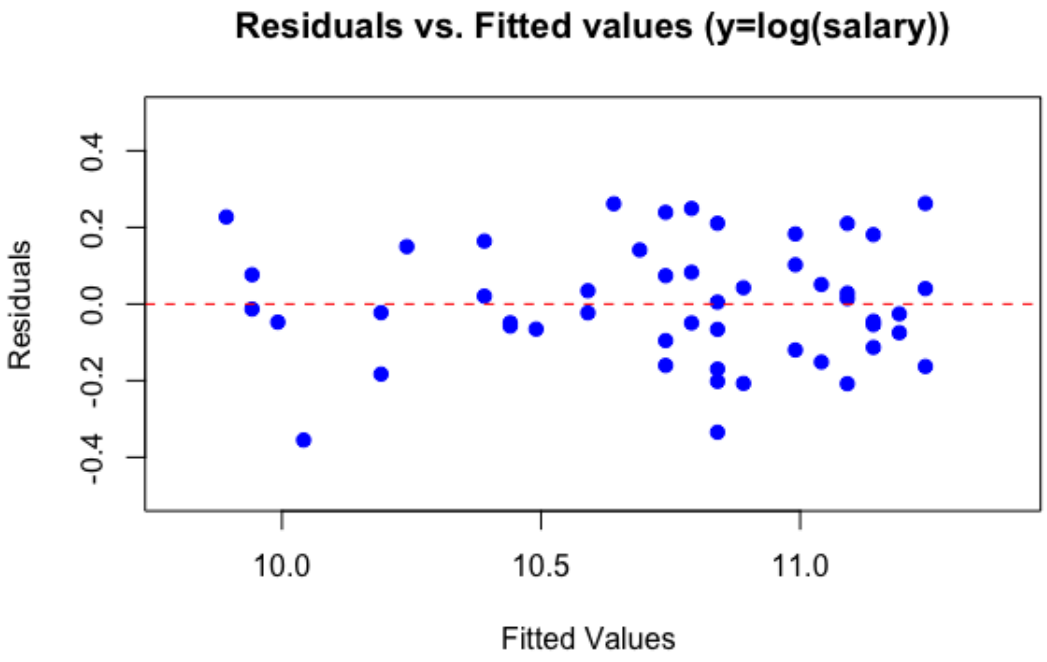


Figure 8.12: RStudio residual plot for second-order model of natural log of salary

```
> Model6=lm(y~EXP)
> summary(Model6)

Call:
lm(formula = y ~ EXP)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35435 -0.09046 -0.01725  0.09739  0.26355

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.841315   0.056356  174.63   <2e-16 ***
EXP         0.049979   0.002868   17.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1541 on 48 degrees of freedom
Multiple R-squared:  0.8635,    Adjusted R-squared:  0.8607
F-statistic: 303.7 on 1 and 48 DF,  p-value: < 2.2e-16
```

Figure 8.13: RStudio residual plot for first-order model of natural log of salary

$$\widehat{ln(y)} = 9.841315 + 0.049979x$$

Note:

- When the transformed model is used to predict the value of $ln(y)$, the predicted value of y is the antilog, $\hat{y} = e^{\widehat{lny}}$.
- The endpoints of the prediction intervals can be transformed back to the original scale without losing their meaning, but confidence intervals for the mean of a transformed response cannot.

8.5 Checking the Normality Assumption

Recall the random errors, ϵ, assumed to be Normally distributed. (i.e. $\epsilon \sim N(0, \sigma^2)$). We use graphical techniques to check this assumption since the available tests tend to have low power when the assumption is violated.
Construct a frequency or relative frequency distribution for the residuals by creating a:

- Histogram
- Stem-and-leaf
- Normal probability plot

Note: The normality assumption is the least restrictive when we apply regression analysis in practice since moderate departures from the assumption of normality have very little effect on Type I error rates associated with the statistical tests and on the confidence coefficients associated with the confidence intervals.

Example 8.5.1 *Consider Example 8.4.1 describe the distribution of the residuals obtained from the model of $ln(salary)$, Model 5.*

The histogram is mound-shaped and reasonably symmetric. The normal probability plot shows the points fall reasonably close to a straight line, indicating that the normality assumption is most likely satisfied.
Note:

- Nonnormality of the distribution of the random error ϵ is often accompanied by heteroscedasticity and the remedy is the variance-stabilizing transformations.
- For a positively skewed distribution of the residuals the square-root transformation on y will stabilize (approximately) the variance and, at the same time, will reduce skewness in the distribution of the residuals.
- If the homoscedasticity assumption is met but the normality assumption is not, normalizing transformations are available such as $\sqrt{y}, log(y), y^2, \frac{1}{\sqrt{y}}, \frac{1}{y}$.

Box and Cox (1964) have developed a procedure for selecting the appropriate transformation to use.

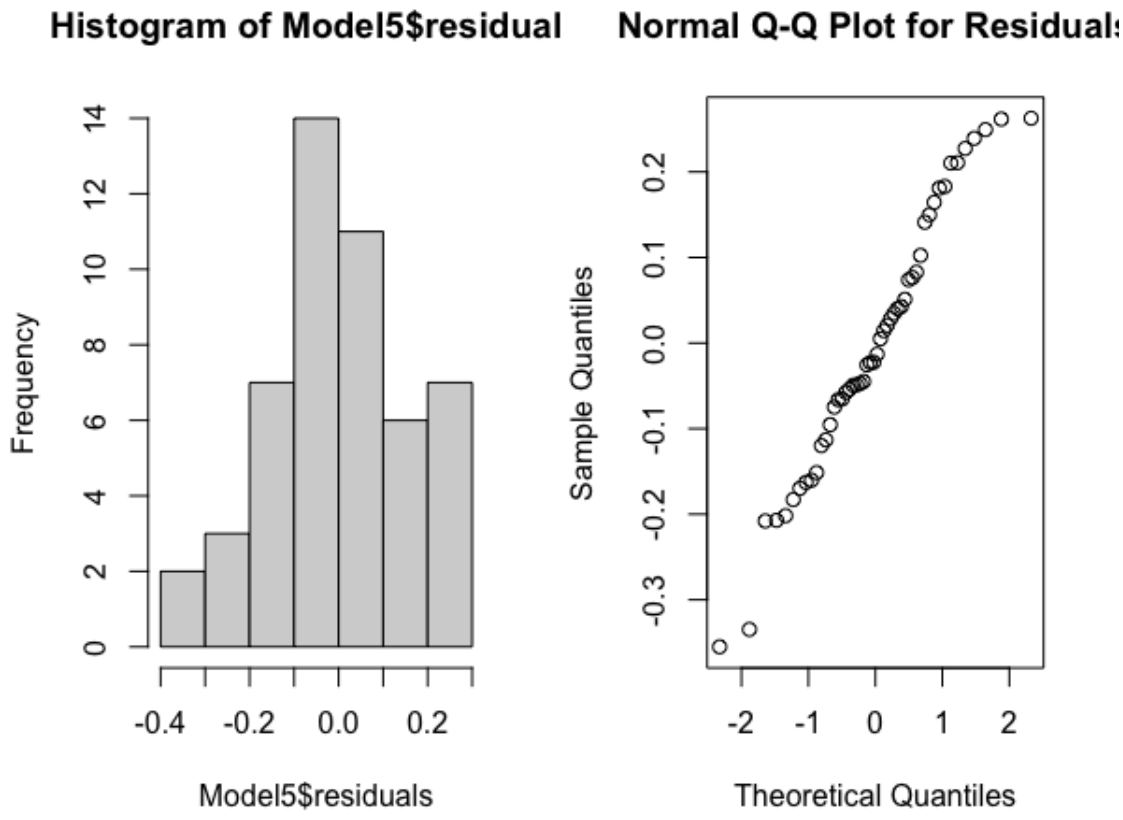


Figure 8.14: RStudio histogram and normal probability plot of residuals from log model of salary

8.6 Detecting Outliners and Identifying Influential Observations

Standardized Residual

The standardized residual, denoted z_i, for the i^{th} observation is the residual for the observation divided by s, that is,

$$z_i = \frac{\hat{\epsilon}_i}{s} = \frac{y_i - \hat{y}_i}{s}$$

Outlier

An observation that does not follow the pattern of the data.

An observation with a residual that is larger than $3s$ (inabsolute value)—or, equivalently, a standardized residual that is larger than 3 (in absolute value)—is considered to be an outlier.

Note: Some text books or software call an observation with a standardized residual that is larger than 2 (in absolute value) an outlier which is more conservative.

Studentized Residuals

As an alternative to standardized residuals, some software packages compute studentized residuals, so named because they follow an approximate Student's t-distribution.

$$z_i^* = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_i}} = \frac{y_i - \hat{y}_i}{s\sqrt{1-h_i}}$$

h_i is called leverage.

Example 8.6.1 *Refer to Example 8.8 from the text book. Data set: FASTFOOD*

We expect a first-order (linear) relationship to exist between mean sales, $E(y)$, and traffic flow. Furthermore, we believe that the level of mean sales will differ from city to city, but that the change in mean sales per unit increase in traffic flow will remain the same for all cities (i.e., that the factors Traffic Flow and City do not interact). The model is therefore:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$y =$ Weekly sales (in thousands of dollars)

$$x_1 = \begin{cases} 1 & \text{if city 1} \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{f city 2} \\ 0 & \text{if not} \end{cases} \quad x_3 = \begin{cases} 1 & \text{f city 3} \\ 0 & \text{if not} \end{cases}$$

$x_4 =$ Traffic flow (in thousands of cars)

(a) Fit the model to the data and evaluate overall model adequacy.

```
> ModelFast=lm(SALES~TRAFFIC+x1 +x2+x3)
> summary(ModelFast)

Call:
lm(formula = SALES ~ TRAFFIC + x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-11.681  -7.331  -1.390   1.719  56.464

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.4592    13.1640  -1.250   0.2264
TRAFFIC       0.3629     0.1679   2.161   0.0437 *
x1            1.1061     8.4226   0.131   0.8969
x2            6.1428    11.6800   0.526   0.6050
x3           14.4896     9.2884   1.560   0.1353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.86 on 19 degrees of freedom
Multiple R-squared:  0.2595,    Adjusted R-squared:  0.1036
F-statistic: 1.665 on 4 and 19 DF,  p-value: 0.1996
```

Figure 8.15: RStudio regression printout for model of fast-food sales

The small value of R^2 and R_a^2 as well as high value of the p-value for the global F-test indicate that the model is not useful for predicting sales.

(b) Plot the residuals from the model to check for any outliers.

The standardized residual plots shows the observation 13 has standardized residual with absolute value greater than 3.

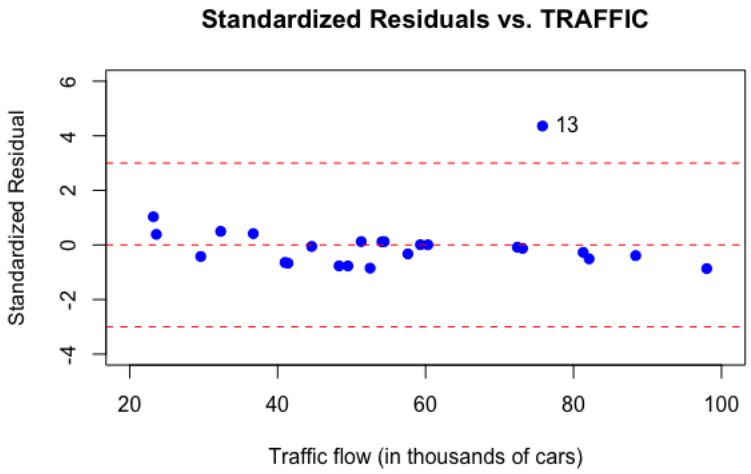


Figure 8.16: RStudio plot of residuals versus traffic flow

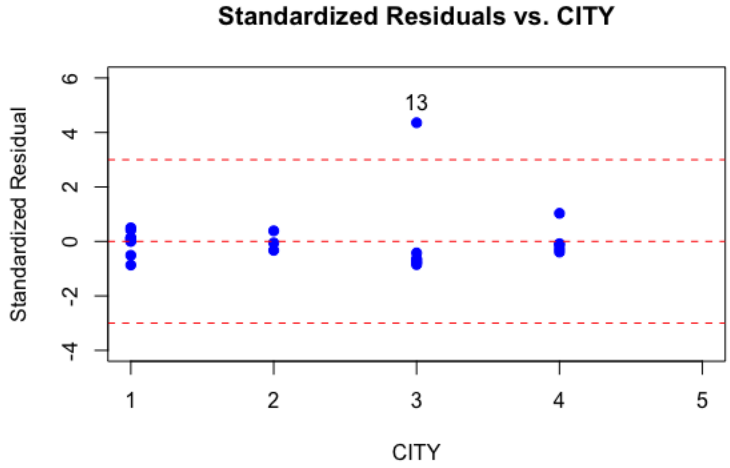


Figure 8.17: RStudioplot of residuals versus city

The following shows the information about observation 13. What is unusual?

```
> identify(y=rstandard(ModelFast), x=FASTFOOD$CITY)
[1] 13
> FASTDUMMY[13,]
   SALES TRAFFIC x1 x2 x3
13    82    75.8  0  0  1
```

(c) Based on the results, part b, make the necessary model modifications and reevaluate model fit. Compare the results with part b results.

```
> FASTDUMMY[13,1]=8.2
> FASTDUMMY[13,]
   SALES TRAFFIC x1 x2 x3
13   8.2    75.8  0  0  1
> ModelFastC=lm(FASTDUMMY$SALES~FASTDUMMY$TRAFFIC+FASTDUMMY$x1 +FASTDUMMY$x2+FASTDUMMY$x3)
> summary(ModelFastC)

Call:
lm(formula = FASTDUMMY$SALES ~ FASTDUMMY$TRAFFIC + FASTDUMMY$x1 +
    FASTDUMMY$x2 + FASTDUMMY$x3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77921 -0.19129 -0.01316  0.26712  0.59254

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.083388   0.321008   3.375 0.003179 **
FASTDUMMY$TRAFFIC 0.103673   0.004094  25.320 4.21e-16 ***
FASTDUMMY$x1      -1.215762   0.205387  -5.919 1.07e-05 ***
FASTDUMMY$x2      -0.530757   0.284819  -1.863 0.077925 .
FASTDUMMY$x3      -1.076525   0.226500  -4.753 0.000138 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3623 on 19 degrees of freedom
Multiple R-squared:  0.9791,    Adjusted R-squared:  0.9747
F-statistic: 222.2 on 4 and 19 DF,  p-value: 1.15e-15
```

Figure 8.18: RStudio regression printout for model of fast-food sales with corrected data point

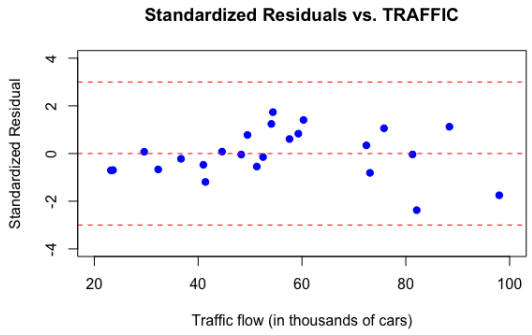


Figure 8.19: RStudio plot of residuals versus traffic flow for model with corrected data point

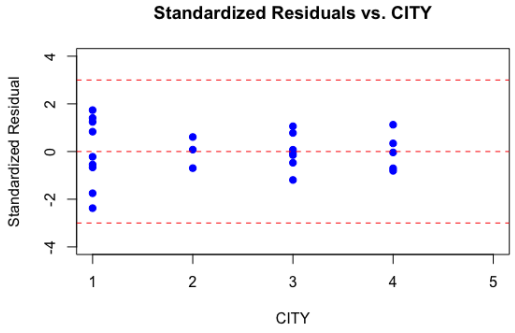


Figure 8.20: RStudioplot of residuals versus city for model with corrected data point

Note: Determine how much influence an an outlier has on the regression analysis before making any decision.

Leverage

The leverage of the i^{th} observation is the weight, h_i, associated with y_i in the equation:

$$\hat{y}_i = h_1y_1 + h_2y_2 + h_3y_3 + ... + h_iy_i + ...h_ny_n$$

where h_1 , h_2 , h_3 . . . h_n are functions of only the values of the independent variables (x's) in the model. The leverage, h_i, measures the influence of y_i on its predicted value $\hat{y}_i$.

Rule of Thumb for Detecting Influence with Leverage

The observed value of y_i is influential if

$$h_i > \frac{2(k+1)}{n}$$

where h_i is the leverage for the ith observation and k = the number of β's in the model (excluding β_0).

The Jackknife

Another technique for identifying influential observations requires that you delete the observations one at a time, each time refitting the regression model based on only the remaining $n-1$ observations.

A deleted residual, denoted d_i, is the difference between the observed response y_i and the predicted value $\hat{y}_{(i)}$ obtained when the data for the i^{th} observation is deleted from the analysis, that is,

$$d_i = y_i - \hat{y}_{(i)}$$

An observation with an unusually large (in absolute value) deleted residual is considered to have large influence on the fitted model.

Cook's Distance:

A measure of the overall influence an outlying observation has on the estimated β coefficients was proposed by R. D. Cook (1979).

Cook's distance, D_i, is calculated for the i^{th} observation as follows:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE}[\frac{h_i}{(1-h_i)^2}]$$

Note that D_i depends on both the residual $(y_i - \hat{y}_i)$ and the leverage h_i for the i^{th} observation.

A large value of D_i indicates that the observed y_i value has strong influence on the estimated β coefficients (since the residual, the leverage, or both will be large).

Values of D_i can be compared to the values of the F distribution with $\nu_1 = k+1$ and $\nu_2 = n-(k+1)$ degrees of freedom. Usually, an observation with a value of D_i that falls at or above the 50^{th} percentile of the F distribution is considered to be an influential observation.

Cook's Distance Cutoff

The observed value of y_i is influential if

$$D_i > \frac{4}{n-(k+1)}$$

Example 8.6.2 *Consider the fast-food sales model.*

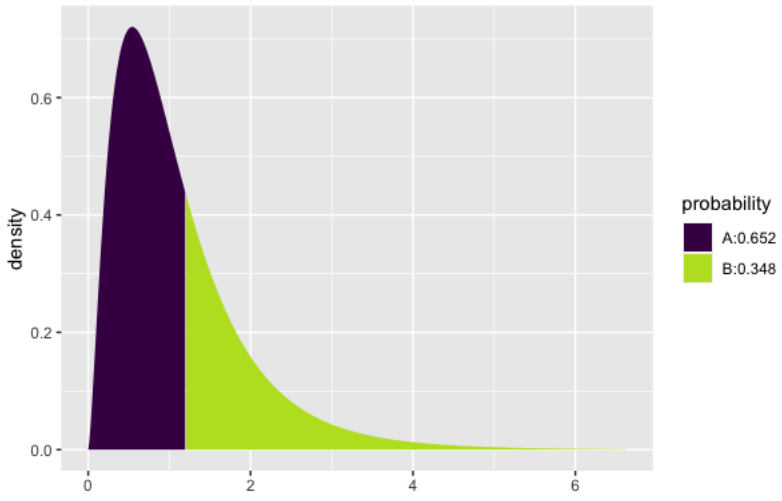
```
> leverage <- round(hatvalues(ModelFast),3)
> StanRes <- round(rstandard(ModelFast),3)
> residual <- round(ModelFast$residuals,3)
> cd <- round(cooks.distance(ModelFast),3)
> Rstudent=round(rstudent(ModelFast),3)
> cbind(SALES,TRAFFIC,leverage,residual,StanRes, cd,Rstudent )
   SALES TRAFFIC leverage residual StanRes    cd Rstudent
1    6.3    59.3    0.111    0.135   0.010 0.000    0.009
2    6.6    60.3    0.111    0.072   0.005 0.000    0.005
3    7.6    82.1    0.181   -6.839  -0.509 0.011   -0.498
4    3.0    32.3    0.200    6.632   0.499 0.012    0.489
5    9.5    98.0    0.308  -10.708  -0.866 0.067   -0.861
6    5.9    54.1    0.114    1.622   0.116 0.000    0.113
7    6.1    54.4    0.113    1.713   0.122 0.000    0.119
8    5.0    51.3    0.118    1.738   0.125 0.000    0.121
9    3.6    36.7    0.173    5.636   0.417 0.007    0.408
10   2.8    23.6    0.376    4.553   0.388 0.018    0.379
11   6.7    57.6    0.365   -3.885  -0.328 0.012   -0.320
12   5.2    44.6    0.334   -0.668  -0.055 0.000   -0.054
13  82.0    75.8    0.239   56.464   4.358 1.196  179.310
14   5.0    48.3    0.143  -10.557  -0.767 0.020   -0.759
15   3.9    41.4    0.149   -9.153  -0.668 0.016   -0.658
16   5.4    52.5    0.145  -11.681  -0.850 0.025   -0.844
17   4.1    41.0    0.150   -8.808  -0.643 0.015   -0.633
18   3.1    29.6    0.188   -5.671  -0.423 0.008   -0.414
19   5.4    49.5    0.143  -10.593  -0.770 0.020   -0.762
20   8.4    73.1    0.204   -1.667  -0.126 0.001   -0.122
21   9.5    81.3    0.224   -3.542  -0.271 0.004   -0.264
22   8.7    72.4    0.203   -1.113  -0.084 0.000   -0.082
23  10.6    88.4    0.255   -5.019  -0.391 0.010   -0.382
24   3.3    23.2    0.453   11.341   1.032 0.176    1.034
```

A statistic related to the deleted residual of the jackknife procedure is the Studentized deleted residual given under the column heading Rstudent.

Studentized deleted residual

The Studentized deleted residual, denoted d_i^*, is calculated by dividing the deleted residual d_i by its standard error s_{d_i} :

$$d_i^* = \frac{d_i}{s_{d_i}}$$



The Studentized deleted residual d_i^* has a sampling distribution that is approximated by a Student's t distribution with $(n-1)-(k+1)$ df.

8.7 Detecting Residual Correlation: The Durbin-Watson test

Acknowledgement

The core content of the slides are from the textbook of this course;

A Second Course in Statistics: Regression Analysis (7th Edition)

by

Mendenhall, William and Sincich, Terry; Pearson Education.

by

Simon J. Sheather