STAT 608, Spring 2022 - Assignment 3
Solutions

1. Question 1, Chapter 3, p. 103.

   (a) While this model does seem to work with great precision, it appears that the true underlying relationship between distance and fare may not be a straight line. If, for example, the underlying model is actually a parabola, our model will systematically overestimate or underestimate the price of the airfare, depending on the flight distance. This problem will be exacerbated the more we attempt to extrapolate, if we do.

   (b) As above, it is likely that we need to fit something other than a straight line relationship. If we used a transformation or parabola, it appears we may still fail to fit the point with maximum airfare; that point may need to be investigated for other explanations, and perhaps an indicator variable can be added to prevent it from influencing the fit of the rest of the model.

2. Explain in words why when we create confidence intervals and prediction intervals using a transformed response variable Y, we can't simply take the inverse transformation of the endpoints to get a confidence or prediction interval in the original units of Y.

   When we transform the data the interval may change, and when we try to back-transform the interval, we will need to add or multiply by an additional term in order for the interval to agree with the original data. This is known as the correction factor and it changes for a given transformation.

3. Recall the model with two indicator variables from question 3 of the previous homework. Calculate the hat matrix (use software if you like; it might be faster by hand). Explain what that projection matrix does and why it makes sense, as if to someone who has taken one semester of statistics.

   In this dummy variable case, take $m = 3$, $n = 4$, for example. Then the design matrix and the hat matrix are

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \; \mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \begin{bmatrix} 1/3 & 1/3 & 1/3 & & & & \\ 1/3 & 1/3 & 1/3 & & & & \\ 1/3 & 1/3 & 1/3 & & & & \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \\ & & & 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

   Recall that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. The hat matrix projects $\mathbf{y}$ onto its average. The design matrix tells us which group an individual is in: the calcium or the placebo group, say. So the hat matrix will take us from the vector $\mathbf{y}$, where every individual has a different blood pressure, down into the vector space defined by the columns of $\mathbf{X}$, where there are only two groups. So there will only be two blood pressures in that sub-space: an average for the calcium group and an average for the placebo group.

Also note that the two parameters of interest in this model are the population means for the two groups, and our linear model here estimates those two population means with their sample means, as we might have expected.

4. For the simple linear regression model in the case that our assumption is met that the errors are independent and identically distributed with variance σ^2:

   (a) Show that the formula for the vector of residuals $\hat{\mathbf{e}}$ can be expressed compactly using the hat matrix:
   $$(\mathbf{I}-\mathbf{H})\,\mathbf{y}$$
   In this problem, the model $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}$ can be written as
   $$\mathbf{y}=\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\mathbf{e}} \implies \hat{\mathbf{e}}=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{y}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}=\mathbf{y}-\mathbf{H}\mathbf{y}=(\mathbf{I}-\mathbf{H})\mathbf{y}.$$

   (b) Show that the covariance matrix of the residuals is therefore equal to
   $$(\mathbf{I}-\mathbf{H})\,\boldsymbol{\Sigma}\,(\mathbf{I}-\mathbf{H})',$$
   where $\boldsymbol{\Sigma}$ is the covariance matrix of the errors. Show that the covariance matrix of the residuals reduces to $(\mathbf{I}-\mathbf{H})\,\sigma^2$. (Please show that $\mathbf{H}$ is idempotent; that is, that $\mathbf{HH}=\mathbf{H}$.)
   Idempotent:
   $$\mathbf{HH}=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{I}\mathbf{X}'=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'=\mathbf{H}$$

   Symmetric:
   $$\mathbf{H}'=\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)'=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'=\mathbf{H}$$

   Note also that
   $$(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})=\mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{HH}=\mathbf{I}-\mathbf{H}$$
   Therefore
   $$\begin{aligned}\mathrm{Cov}\left(\hat{\mathbf{e}}|\mathbf{X}\right)&=\mathrm{Cov}\left((\mathbf{I}-\mathbf{H})\mathbf{y}|\mathbf{X}\right)=(\mathbf{I}-\mathbf{H})\mathrm{Cov}\left(\mathbf{y}|\mathbf{X}\right)(\mathbf{I}-\mathbf{H})'\\&=(\mathbf{I}-\mathbf{H})\sigma^2(\mathbf{I}-\mathbf{H})'=\sigma^2(\mathbf{I}-\mathbf{H})\end{aligned}$$

(c) Conclude that $\operatorname{Cov}(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2$, $i \neq j$.
From (b), when $i \neq j$, $\operatorname{Cov}(\hat{e}_i, \hat{e}_j) = (0 - h_{ij})\,\sigma^2 = -h_{ij}\sigma^2$

5. For the simple linear regression model, show that the hat matrix $\mathbf{H}$ has the following properties:

(a) $\mathbf{H}$ is symmetric.

$$\mathbf{H}' = \left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)' = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \mathbf{H}$$

(b) $0 \leq h_{ii} \leq 1$, where h_{ii} is the i^{th} diagonal entry of the hat matrix. (**HINT**: First show that $h_{ii} \geq h_{ii}^2$ and note that $h_{ii} = \sum_j h_{ij}^2$.)

$$h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2 \geq h_{ii}^2 \implies h_{ii}(1 - h_{ii}) \geq 0 \implies 0 \leq h_{ii} \leq 1$$

(c) The off-diagonals of the hat matrix are found by the formula

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}.$$

First reference textbook section 6.1.1, where we are told that

$$\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \frac{1}{SXX} \begin{bmatrix} \frac{1}{n}\sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}$$

If you work out the algebra, you find that

$$\begin{aligned} h_{ij} &= \frac{1}{SXX}\left[\frac{1}{n}\sum_i x_i^2 + x_i x_j - x_i\bar{x} - x_j\bar{x}\right] \\ &= \frac{1}{SXX}\left[\frac{1}{n}\left(\sum_i x_i^2 - n\bar{x}^2 + n\bar{x}^2\right) + x_i x_j - x_i\bar{x} - x_j\bar{x}\right] \\ &= \frac{1}{SXX}\left[\frac{SXX}{n} + \bar{x}^2 + x_i x_j - x_i\bar{x} - x_j\bar{x}\right] \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX} \end{aligned}$$

(d) Finally, the text states that "There is a small amount of correlation present in standardized residuals, even if the errors are independent." Comment on when the covariances of the residuals are close to zero, for a fixed sample size. Why does it make sense that the covariances are close to zero in those situations?

When $n \to \infty$, the denominators of h_{ij} will be very large and $h_{ij} \to 0$. For a fixed sample size, it makes sense that the covariances are small numbers.

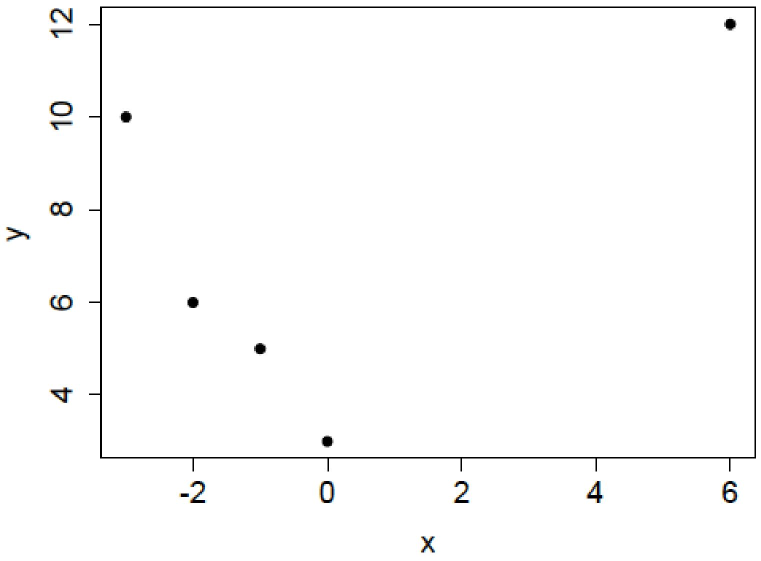

Figure 1: Scatterplot of the data.

6. Under the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + e_i$, suppose the following data are collected and recorded: $x = [-3, -2, -1, 0, 6]$, $y = [10, 6, 5, 3, 12]$. Show your work, and do the following calculations by hand. (You may double-check with a computer.)

   (a) First, create a quick sketch of the scatterplot of the data. Label any potential outliers or leverage points as such. Write out your design matrix.

   Figure 1 shows the scatterplot. The design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 6 \end{bmatrix}$$

   (b) Using $\hat{y} = 7.2 + 0.5x$, calculate the residuals $\hat{e}_i$.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = 7.2 + 0.5 \begin{bmatrix} -3 \\ -2 \\ -1 \\ 0 \\ 6 \end{bmatrix} = \begin{bmatrix} 5.7 \\ 6.2 \\ 6.7 \\ 7.2 \\ 10.2 \end{bmatrix}$$

   And

$$\hat{\mathbf{e}} = (\mathbf{y} - \hat{\mathbf{y}}) = \begin{bmatrix} 4.3 \\ -0.2 \\ -1.7 \\ -4.2 \\ 1.8 \end{bmatrix}$$

(c) Compute the leverage for each observation. Use the rule $h_{ii} > 4/n$ to identify potential leverage points. Are any points of high leverage "good" or "bad"?

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 0 \\ 0 & 50 \end{bmatrix}, \ \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \frac{1}{50}\begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \frac{1}{50}\begin{bmatrix} 19 & 16 & 13 & 10 & -8 \\ 16 & 14 & 13 & 10 & -2 \\ 13 & 12 & 11 & 10 & 4 \\ 10 & 10 & 10 & 10 & 10 \\ -8 & -2 & 4 & 10 & 46 \end{bmatrix}$$

Using the rule $h_{ii} > 4/n = 4/5 = 0.8$, we can see $h_{55} = 46/50 > 0.8$, the point (6, 12) is a potential leverage point, and it is "bad" because it does not fall near the line that would fit the other points.

(d) Compute the variances of the residuals. (Assume the variance of the errors to simply be σ^2.)

The variances of the residuals are

$$\text{Var}\left(\hat{e}_i\right) = \sigma^2(1 - h_{ii}) = \frac{\sigma^2}{50}\begin{bmatrix} 31 \\ 36 \\ 39 \\ 40 \\ 4 \end{bmatrix} = \sigma^2\begin{bmatrix} 0.62 \\ 0.72 \\ 0.78 \\ 0.80 \\ 0.08 \end{bmatrix}$$

The residual e_4 has the highest variance.

(e) Compute the standardized residuals ($s = 3.755$). Comment no why this answer seems to conflict a bit with the answer to part (b) above.

The standardized residuals are

$$\frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}\begin{bmatrix} 1.454 \\ -0.063 \\ -0.513 \\ -1.251 \\ 1.695 \end{bmatrix}$$

Point 5 has the highest standardized residual in absolute value, even though point 1 has the most extreme residual. Because the standardized residuals are obtained by dividing the residuals by the variance of the residuals, points with low variance will tend to have high standardized residuals.

(f) Comment on why the point with the highest leverage in this dataset had the smallest variance.

Since the variance is given by $\sigma^2(\mathbf{I} - \mathbf{H})$, points with high leverages will tend to have low variances. The bad leverage point is pulling the line to be very close to it, reducing the variance of the residual there.

7. When Y has both mean and variance equal to μ, we showed in the notes that the appropriate transformation of Y for stabilizing the variance is the square root transformation. Now, suppose that Y has mean equal to μ and variance equal to μ^2. Show that the appropriate transformation of Y for stabilizing variance is the log transformation. (Question 7, Chapter 3, page 112 of the textbook.)

   Suppose $f(Y) = \log(Y)$, $\mathrm{E}(Y) = \mu$, and $\mathrm{Var}(Y) = \mu^2$. Using Taylor series expansion,

   $$f(Y) = f(\mu) + f'(\mu)(Y - \mu) + \cdots$$

   Hence,

   $$\mathrm{Var}(f(Y)) \approx \left(f'(\mu)\right)^2 \mathrm{Var}(Y) = \left(\frac{1}{\mu}\right)^2 \times \mu^2 = 1$$

   Because the variance of $f(Y)$ is constant in μ (that is, the variance is no longer a function of the mean), the variance has been stabilized and the log transformation is appropriate.

8. Download the dataset called `company.csv` from Canvas. The dataset contains a systematic sample (every tenth company; we'll take these as randomly selected) for the Forbes 500 list. The variables of interest are `Sales` and `Assets` of the companies (both in millions of U.S. dollars). As with many financial datasets, many of these variables are skewed. Your job is to choose appropriate power transformations such that the relationship between `Assets` (response variable) and `Sales` (explanatory) are approximately linear.

   (a) Begin by creating a scatterplot of `Sales` and `Assets` and fit a simple linear regression line. What transformations does your scatterplot suggest? Create diagnostic plots for this model (Model 1). Discuss any weaknesses of this model.

   Figure 2 shows the scatterplot of `Sales` and `Assets`. In Figure 3, the Q-Q plot shows that the residuals are not normally distributed. The first and third plots show that most fitted values are on the left-hand side of the graph. The Scale-Location plot shows that observations 16, 48, and 54 are outliers; point 43 has a standardized residual that also exceeds our rule of thumb of absolute standardized residual exceeding 2. The residual vs. leverage plot shows that observation 40 is a high leverage point (all of points 33, 40, and 43 have leverage values that exceed our rule of thumb of $4/n$). The Cook's distance plot in Figure 4 shows that observation 16 is the point with the worst Cook's distance; points 40, 43, 48, and 54 are also above our cutoff for bad outliers / high influence.

   (b) Choose an appropriate transformation for `Sales`. Explain how you made your choice. Include plots if applicable.

   I would like to use the Box-Cox transformation to see which power is the most appropriate value for doing the transformation. (We know that when $\lambda = 0$, it is the log transformation.)

   By using the R code, we find that $\lambda = -0.0675$ is the most appropriate power transformation, which is very close to 0. Also note that the Wald

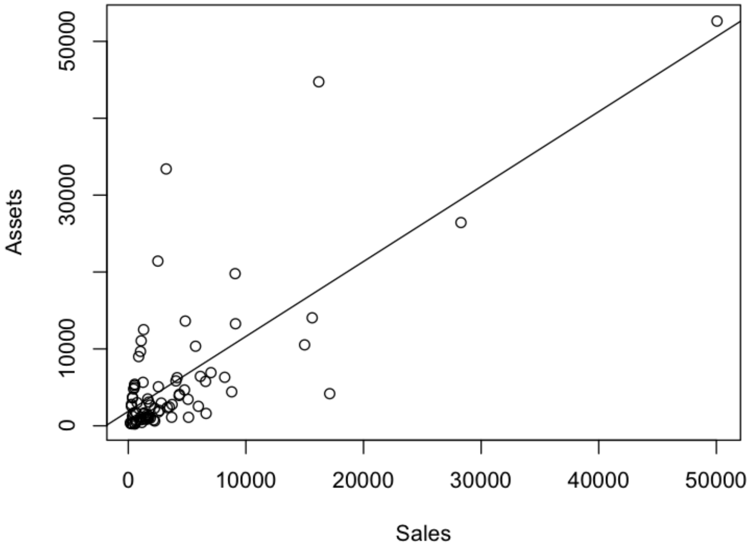

Figure 2: Scatterplot of the data

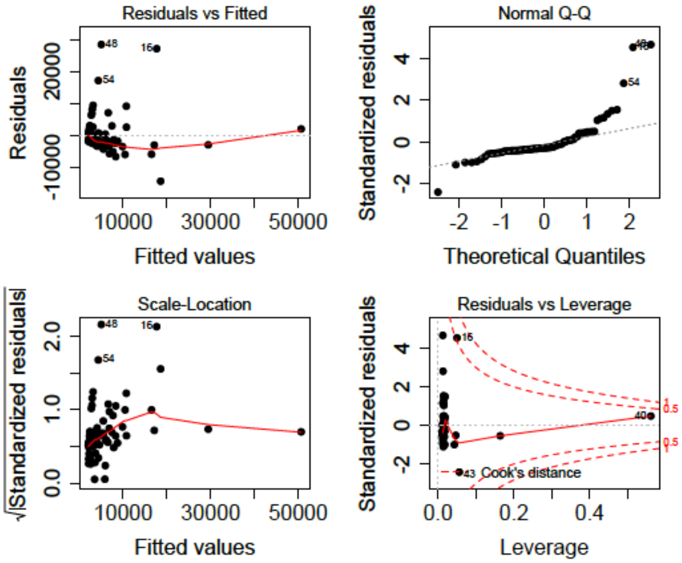

Figure 3: Diagnostic plot

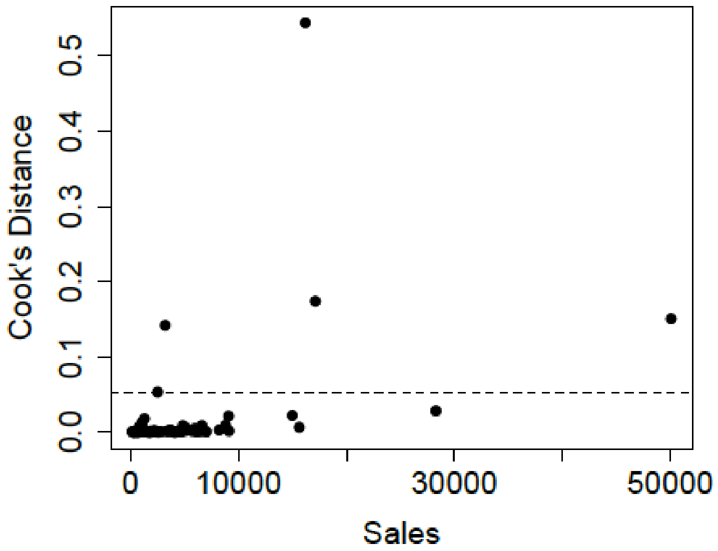

Figure 4: Cook's distance plot

confidence interval includes the value 0, and the likelihood ratio test for testing that $\lambda = 0$ has a large p-value, so we fail to reject the hypothesis that a log transformation is appropriate. Therefore, the natural log transformation for sales is an appropriate transformation.

```
bcPower Transformation to Normality
      Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Sales   -0.0675           0      -0.2329       0.0979

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                             LRT df    pval
LR test, lambda = (0) 0.6481734  1 0.42077

Likelihood ratio test that no transformation is needed
                            LRT df       pval
LR test, lambda = (1) 170.5833  1 < 2.22e-16
```

(c) Choose an appropriate transformation for `Assets`, and again explain how you made your choice. Because using an inverse response plot in this example is messy, you can just (1) fit a regression model of `Assets` vs. the transformed version of `Sales` that you chose in part (b), then (2) pass the fitted model into the `powerTransform` function. No plots required.

This time, I run a regression model regressing `Assets` on log(`Sales`). I use that model in the powerTransform (Box-Cox) function in R so that I transform residuals to be as close to normal as possible (rather than the

y-variable). We find that the best transformation for y is now -0.0166, and 0 is in the Wald confidence interval; we also fail to reject a null hypothesis that $\lambda = 0$ because of the large p-value. Therefore the natural log transformation is an acceptable transformation for y.

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.0166           0      -0.1688       0.1357

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                             LRT df    pval
LR test, lambda = (0) 0.04564128  1 0.83083

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 159.6628  1 < 2.22e-16
```

(d) Call the model with both variables transformed Model 2. Create diagnostic plots for this model, and discuss any weaknesses of this model.

After log transformations of both `Sales` and `Assets`, the new x and y variables are linearly related, as shown in Figure 5. The diagnostic plots improve a lot compared with Model 1, as shown in Figure 6. The residual plot is less concentrated than Model 1, and the Q-Q plot shows the residuals are closer to a normal distribution. In Model 2, all assumptions are met: linear relationship, independent observations, normality of residuals and constant variance.

(e) Compare Model 1 and Model 2. Which model is preferable?

As shown in (d), Model 2 met all the assumptions. Therefore, I prefer Model 2.

(f) Using the model $\log(\texttt{Assets}) = \beta_0 + \beta_1 \log(\texttt{Sales})$, interpret the slope in the context of the problem.

The slope is 0.587, which corresponds to the percentage change in assets for every percentage change in sales. According to our model, for every 1% increase sales, there is approximately a 0.587% increase in assets, on average.

(g) Again using the model $\log(\texttt{Assets}) = \beta_0 + \beta_1 \log(\texttt{Sales})$, find a 95% confidence interval for the average assets of a company with $6,571$ million in sales, as Hewlett-Packard did. Interpret your confidence interval in context.

To transform this interval into the original units, add one half of the mean square error to each endpoint and apply the exponential function, we have a 95% confidence interval of $(6,870, 12,889)$. I am 95% confident that companies with 6,571 million in sales would have assets between 6,870 and 12,889 million, on average.

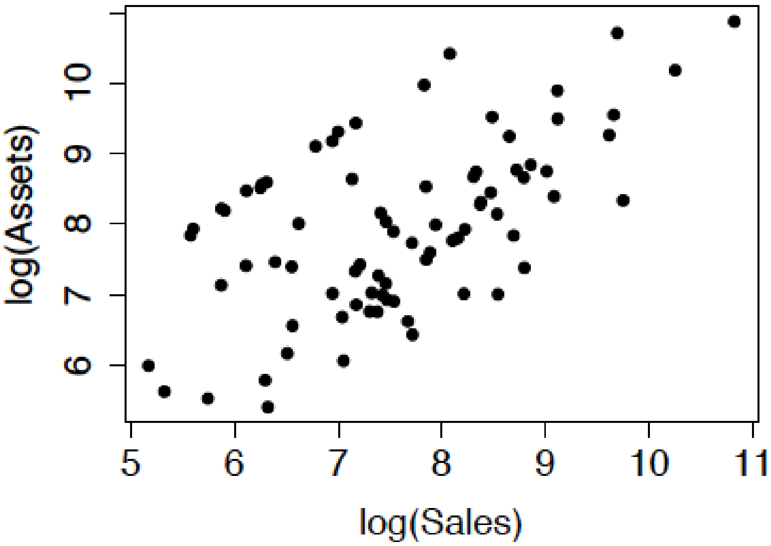

Figure 5: The relationship between `Sales` and `Assets` after transformation

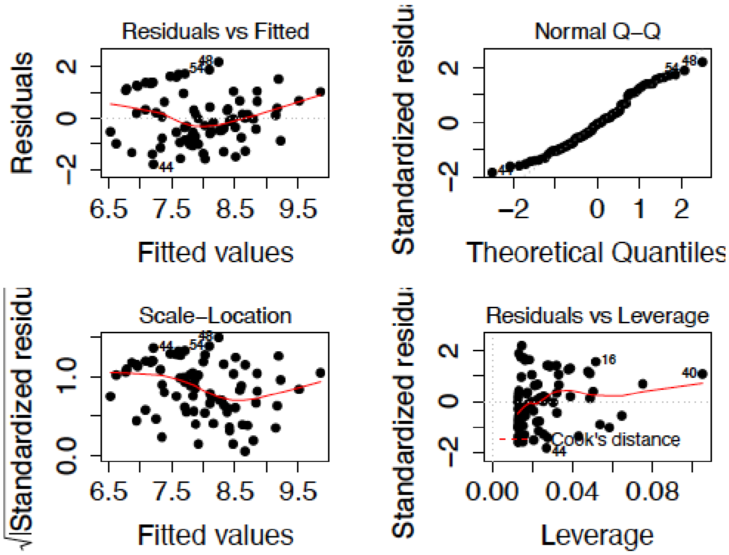

Figure 6: Diagnostic plot after transformation