STAT-285 Homework 9 Solutions

Section 13.4 Question 41 /10

Study Objective: Investigate how eccentricity and axial length in eyes are related to cone cell packing density.

Formulation: Let

- Y_i denote the ith measurement of cone cell packing density (cells/mm^2), for $i = 1, \cdots, n$, with $n = 192$.
- X_{i1} denote the ith measurement of eccentricity (mm), for $i = 1, \cdots, n$.
- X_{i2} denote the ith measurement of axial length (mm), for $i = 1, \cdots, n$.

The relationship between Y_i with X_{i1} and X_{i2} is specified to be

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i,$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and ε_i is independent from X_{i1} and X_{i2}. Here, β_0, β_1, β_2, and σ^2 are unknown parameters.

Part A /2

We are told that $R^2 = 0.834$. This means that 83.4% of the variability in Y can be explained by X_1 and X_2.

To carry out a test of model utility, we consider the following hypothesis test:

Hypothesis Test: $H_0 : \beta_1 = \beta_2 = 0$ vs. $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$.

Test Statistic:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \sim F(k, n - k - 1)$$

under H_0. Here, $k = 2$ is the number of independent variables included in the model. Plugging in the corresponding values results in $F_{obs} = 474.7771$. Since $P_{H_0}(F > F_{obs}) = 1.9962 \times 10^{-74}$, we reject H_0 for essentially any value of α.

Part B /1

We are given $\hat{E}(Y|X_1, X_2) = 35821.792 - 6294.729X_1 - 348.037X_2$. That is, $\hat{\beta}_0 = 35821.792$, $\hat{\beta}_1 = -6294.729$, and $\hat{\beta}_2 = -348.037$. Then

$$\hat{E}(Y|1, 25) = 35821.792 - 6294.729(1) - 348.037(25) = 20826.14$$

Part C /2

Note that

$$\beta_1 = E(Y|X_1 + 1, X_2) - E(Y|X_1, X_2).$$

The model implies by holding axial length fixed, the expected cone cell packing density will decrease by 6,294.729 cells/mm^2 for every 1 mm increase in eccentricity.

Part D /3

A 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{0.025}(189)\widehat{SE}(\hat{\beta}_1) = -6294.729 \pm 1.9726 \times 203.702 \approx [-6696.551, -5892.907]$$

Interpretation: Holding axial length constant, we are 95% confident that the average *decrease* in cone cell packing density by increasing eccentricity by 1 mm is between 5,892.907 and 6,696.551

Part E /2

Hypothesis Test: $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$.

We will conduct the hypothesis test by computing a 95% confidence interval for β_2:

$$\hat{\beta}_2 \pm t_{0.025}(189)\widehat{SE}(\hat{\beta}_2) = -348.037 \pm 1.9726 \times 134.350 \approx [-613.05515, -83.01885].$$

Since $0 \notin [-613.05515, -83.01885]$, we reject H_0 with $\alpha = 0.05$, and conclude that the effect of axial length on cone cell packing density is statistically significant.

Section 13.2 Question 15 /14

Study Objective: Investigate how frying time is related to moisture content in tortilla chips.

Formulation: Let

- Y_i denote the ith measurement of moisture content (%), for $i = 1, \cdots, n$, with $n = 8$.

- X_i denote the ith measurement of frying time (sec), for $i = 1, \cdots, n$.

The relationship between Y_i and X_i is specified as

$$Y_i = f(X_i) + \varepsilon_i,$$

where $f(\cdot)$ is some function, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and ε_i is independent from X_i.

Part A /2

Figure 1 presents a scatter plot of Y_i vs. X_i, for $i = 1, \cdots, 8$. We can see that the relationship between X and Y appears to be a *power* relationship.

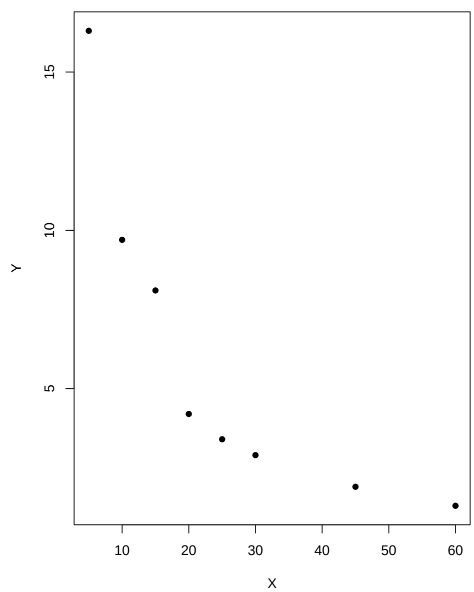

Figure 1: Scatter plot of Y vs. X for data from Section 13.2 Question 15.

Part B /2

Figure 2 presents a scatter plot of $\log Y_i$ vs. $\log X_i$, for $i = 1, \cdots, 8$. We can see that the relationship between $\log X$ and $\log Y$ appears to be linear. That is, the following appears to be appropriate for this data:

$$\log Y_i = \alpha_0 + \alpha_1 \log X_i + \varepsilon_i$$

Part C /2

From the linear model suggested from **Part B**:

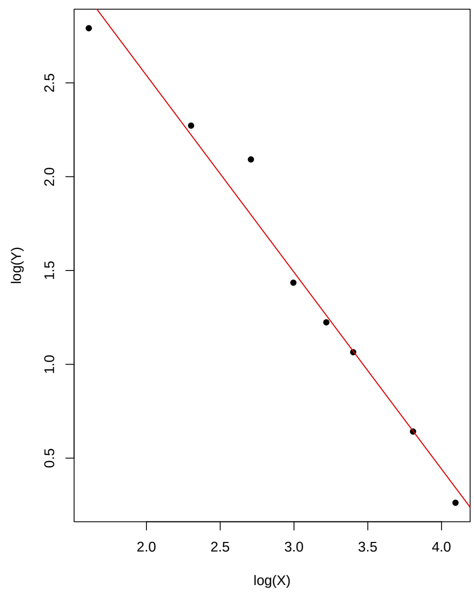

Figure 2: Scatter plot of $\log Y$ vs. $\log X$ for data from Section 13.2 Question 15, and the estimated least squares line.

$$
\begin{aligned}
Y &= \exp\{\alpha_0 + \log X_i^{\alpha_1} + \varepsilon_i\} \\
&= \exp\{\alpha_0\} X_i^{\alpha_1} \exp\{\varepsilon_i\} \\
&= \beta_0 X_i^{\alpha_1} \varepsilon_i^*,
\end{aligned}
$$

where $\beta_0 = \exp\{\alpha_0\}$ and $\varepsilon^* = \exp\{\varepsilon\}$.

Part D /5

Although the wording of the question is kind of confusing, we are asked to provide a (95%) prediction interval of Y given $X = 20$. By fitting a regression line, illustrated in Figure 2, we have

$$\hat{E}(\log Y_i | \log X_i) = 4.638 - 1.049 \log X_i.$$

A point estimate for the predicted value of $\log Y$ given $\log X = \log 20$ is

$$\hat{E}(\log Y_i | \log 20) = 4.638 - 1.049(\log 20) = 1.4953$$

Other quantities we need are

$$\hat{\sigma} = \sqrt{\frac{1}{6}\sum_{i=1}^{8}(\log Y_i - \hat{E}(\log Y_i | \log 20))^2} = 0.1449,$$

$$\overline{\log X} = \sum_{i=1}^{8} \log X_i/8 = 3.0171$$

$$S_{\log X, \log X} = \sum_{i=1}^{8}(\log X_i)^2 - \left(\sum_{i=1}^{8}\log X_i\right)^2 /8 = 2387.5$$

So that a 95% prediction interval for $\log Y$ given $\log X = \log 20$ is

$$\begin{aligned}
&\hat{E}(\log Y_i | \log 20) \pm t_{0.025}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\log 20 - \overline{\log X})^2}{S_{\log X, \log X}}} \\
&= 1.4953 \pm 2.4469 \times 0.1449\sqrt{1 + \frac{1}{8} + \frac{(\log 20 - 3.0171)^2}{2387.5}} \\
&\approx [1.1192, 1.8714].
\end{aligned}$$

Therefore, an approximate 95% prediction interval for Y given $X = 20$ is

$$[\exp\{1.1192\}, \exp\{1.8714\}] = [3.0624, 6.4973].$$

Part E /3

Figure 3 illustrates a scatter plot of the residuals vs. the fitted values, as well as a Normal Q-Q plot of the residuals. We can see that

- The residual corresponding to the third observation is quite large relative to the others.
- There is no apparent trend within in the scatter plot.
- Aside from the residuals pertaining to observations 1 and 3, all of the other points are near the reference line.

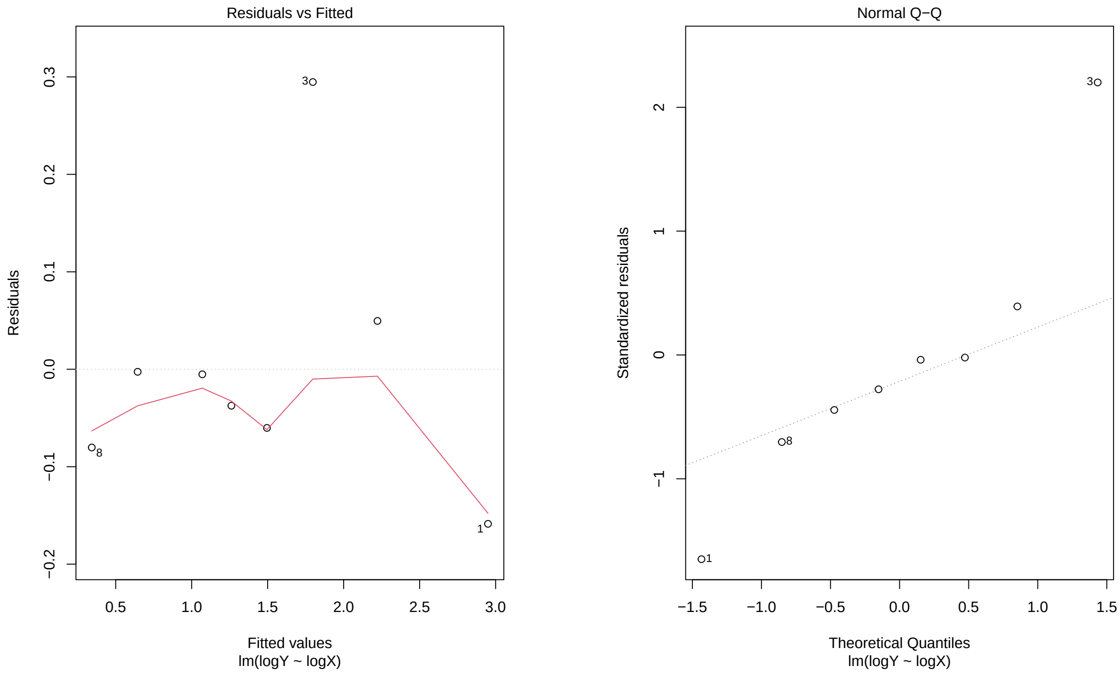

Figure 3:
Left: Scatter plot of the residuals $\hat{e}_i = \log Y_i - \hat{E}(\log Y_i | \log X_i)$ vs. $\hat{E}(\log Y_i | \log X_i)$.
Right: Normal Q-Q plot of $\hat{e}_i$

Section 13.3 Question 29 /8

Study Objective: Investigate how viscosity (MPa · s) is related to free-flow % in high-alumina refractory castables.

Formulation: Let

- Y_i denote the ith measurement of free-flow %, for $i = 1, \cdots, n$, with $n = 7$.
- X_i denote the ith measurement of viscosity (MPa · s), for $i = 1, \cdots, n$.

The relationship between Y_i and X_i is specified to be

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \varepsilon_i,$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and ε_i is independent from X_i. The question gives us $\hat{\beta}_0 = -295.96$, $\hat{\beta}_1 = 2.1885$, and $\hat{\beta}_2 = -0.0031662$

Part A /2

Table 1 displays the data and relevant quantities. The predicted values are $\hat{Y}_i$, the residuals are $\hat{e}_i$, and

$$SSE = \sum_{i=1}^{7} \hat{e}_i^2 = 16.7718$$
$$S^2 = \frac{SSE}{n-3} = 4.1929$$

Table 1: Data and relevant quantities for Section 13.3 Question 29

i	X_i	Y_i	Y_i^2	$\hat{Y}_i$	$\hat{e}_i = Y_i - \hat{Y}_i$	$\hat{e}_i^2$
1	351	81	6,561	82.1342	-1.1342	1.2864
2	367	83	6,889	80.7771	2.2229	4.9414
3	373	79	6,241	79.8502	-0.8502	0.7229
4	400	75	5,625	72.8583	2.1417	4.5870
5	402	70	4,900	72.1567	-2.1567	4.6513
6	456	43	1,849	43.6398	-0.6398	0.4094
7	484	22	484	21.5837	0.4163	0.1733
Total	-	453	32,549	-	-	16.7718

Part B /1

Using information from Table 1,

$$SST = \sum_{i=1}^{7} Y_i^2 - \left(\sum_{i=1}^{7} Y_i \right)^2 /7 = 3233.429$$
$$R^2 = 1 - \frac{SSE}{SST} = 0.9948$$

Approximately 99.48% of the variability in Y can be explained by X and X^2.

Part C /1

Hypothesis Test: $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$.

Test Statistic:

$$T = \frac{\hat{\beta}_2}{\widehat{SE}(\hat{\beta}_2)} \sim t(n-k-1)$$

under H_0. Here, $k = 2$ and plugging in the corresponding values results in $T_{obs} = -6.5483$.

Since $P_{H_0}(|T| > |T_{obs}|) = P_{H_0}(|T| > 6.5483) \approx 0.0028$. We therefore reject H_0 with $\alpha = 0.05$, and conclude that the quadratic term belongs in the regression model.

Part D /2

To have a joint confidence level of at least 95%, we use the Bonferonni procedure and specify α to be

$$\begin{aligned} &100(1-2\alpha)\% \geq 0.95 \\ \Rightarrow\ &\alpha \leq 0.025 \end{aligned}$$

The textbook solution specifies $\alpha = 0.02$, but any value of $\alpha \leq 0.025$ would work too.

A 98% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{0.01}(n-3)\widehat{SE}(\hat{\beta}_1) = 2.1885 \pm 3.7469 \times 0.4050 \approx [0.6708, 3.7062].$$

A 98% confidence interval for β_2 is

$$\hat{\beta}_2 \pm t_{0.01}(n-3)\widehat{SE}(\hat{\beta}_2) = -0.0031662 \pm 3.7469 \times 0.0004835 \approx [-0.0050, -0.0014].$$

Part E /2

Using the estimated regression fit provided by the question, $\hat{E}(Y|X = 400) = 72.8583$

A 95% confidence interval for $E(Y|X = 400)$ is

$$\begin{aligned} &\hat{E}(Y|X = 400) \pm t_{0.025}(n-3)\widehat{SE}(\hat{E}(Y|X = 400)) \\ &= 72.8583 \pm 2.7764 \times 1.198 \\ &\approx [69.532, 76.184] \end{aligned}$$

A 95% prediction interval for a future observation with $X = 400$ is

$$\begin{aligned} &\hat{E}(Y|X = 400) \pm t_{0.025}(n-3)\sqrt{S^2 + \widehat{SE}(\hat{E}(Y|X = 400))^2} \\ &= 72.8583 \pm 2.7764\sqrt{4.1929 + 1.198^2} \\ &\approx [66.2715, 79.4450]. \end{aligned}$$

Due to the extra variability in predicting Y, the prediction interval is wider compared to the confidence interval.

Section 13.4 Question 48 /8

Study Objective: Investigate how three levels of temperature, time of treatment, and tartaric acid concentration are related to weight loss.

Formulation: Let

- Y_i denote the ith measurement of weight loss %, for $i = 1, \cdots, n$, with $n = 15$.
- $X_{i1} \in \{-1, 0, 1\}$ denote the ith level of temperature (in Celsius), for $i = 1, \cdots, n$.
- $X_{i2} \in \{-1, 0, 1\}$ denote the ith level of time of treatment (minutes), for $i = 1, \cdots, n$.
- $X_{i3} \in \{-1, 0, 1\}$ denote the ith level of tartaric acid concentration (g/L), for $i = 1, \cdots, n$.

The relationship between Y_i and X_i is specified to be

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1}^2 + \beta_5 X_{i2}^2 + \beta_6 X_{i3}^2 + \beta_7 X_{i1} X_{i2} + \beta_8 X_{i1} X_{i3} + \beta_9 X_{i2} X_{i3} + \varepsilon_i,$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and ε_i is independent from $\{X_{i1}, X_{i2}, X_{i3}\}$. The question gives us the estimated parameters and relevant quantities to work with. Fitting the regression model in R results in the same estimates provided.

Part A /2

To determine if the specified relationship is meaningful, we conduct the following hypothesis test:
Hypothesis Test: $H_0 : \beta_1 = \cdots = \beta_9 = 0$ vs. H_a : At least one $\beta_j \neq 0$, for $j = 1, \cdots, 9$.

Test Statistic:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \sim F(k, n - k - 1)$$

under H_0. Here, $k = 9$ is the number of independent variables included in the model. Plugging in the corresponding values results in $F_{obs} = 8.3469$. Since $P_{H_0}(F > F_{obs}) = 0.0155$, we fail to reject H_0 with $\alpha = 0.01$.

Part B /2

In terms of notation, let $E(Y|\boldsymbol{X})$ denote the expected value of Y given $\boldsymbol{X} = (X_1, \cdots, X_9)'$, and $E(Y|\mathbf{0})$ denote the expected value of Y given $\boldsymbol{X} = \mathbf{0}$.

With the estimates they provided, $\hat{E}(Y|\mathbf{0}) = 21.9667$, and a 95% confidence interval for $E(Y|\mathbf{0})$ is

$$\begin{aligned}&\hat{E}(Y|\mathbf{0}) \pm t_{0.025}(n - 10)\widehat{SE}(\hat{E}(Y|\mathbf{0}))\\ &= 21.9667 \pm 2.5707 \times 1.248\\ &\approx [18.7586, 25.1748].\end{aligned}$$

Part C /2

With $S^2 = SSE/(n-10) = 4.6758$, a 95% prediction interval for a future observation with $\boldsymbol{X} = \mathbf{0}$ is

$$\begin{aligned} &\hat{E}(Y|\mathbf{0}) \pm t_{0.025}(n-10)\sqrt{\widehat{SE}(\hat{E}(Y|\mathbf{0}))^2 + S^2} \\ &= 21.9667 \pm 2.5707 \times \sqrt{1.248^2 + 4.6758^2} \\ &\approx [15.5488, 28.3846]. \end{aligned}$$

Part D /2

To determine if any of the second order predictors belong in the model, we conduct the following hypothesis test:
Hypothesis Test: $H_0 : \beta_4 = \cdots = \beta_9 = 0$ vs. H_a : At least one $\beta_j \neq 0$, for $j = 4, \cdots, 9$.

Test Statistic: Let SSE_F and SSE_R denote SSE under the full and reduced model, respectively. Since the reduced model has $\ell = 3$ predictors, the test statistic is

$$F = \frac{(SSE_R - SSE_F)/(k-\ell)}{SSE_F/(n-k-1)} \sim F(k-\ell, n-k-1)$$

under H_0. Plugging in the corresponding values results in $F_{obs} = 6.4316$. Since $P_{H_0}(F > F_{obs}) = 0.0296$, we fail to reject H_0 with $\alpha = 0.01$.