Homework 1

Jacob Thielemier

2 February 2024

Question 1

- Observing data at least as extreme as the observed data, given that the null hypothesis is true.

Question 2

(a)

- Linearity: The relationship between the independent variables and the dependent variable is linear.
- Independence: The residuals are independent of each other.
- Constant Variance: The residuals have constant variance at every level of the independent variables.
- Normality: The residuals are normally distributed.

(b)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

(c)

- If y is correlated with x_1 or x_2 then the assumption of independence would not be satisfied.

(d)

- One example would be a data set of incomes. An income distribution would be right-skewed, meaning that a small amount of people have very high incomes compared to the rest. This skewness violates the normality assumption.

(e)

- A data set containing daily temperatures of a city over a year would violate the independence assumption. For example, the temperature on successive days correlate with each other because today's temperature is a predictor of tomorrow's temperature.

(f)

- The stock markets data violate the constant variance assumption. In times of higher market volatility, the range of stock price movements is much larger compared to periods of low movement with a smaller range.

Question 3

(a)

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

(b) Use the likelihood function above

$$\ell(\mu) = \log L(\mu, \sigma^2; y_1, \ldots, y_n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$\frac{\partial}{\partial\mu}\ell(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)$$

$$\sum_{i=1}^{n} y_i - n\mu = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\hat{\mu} = \bar{y}$$

(c)

$$\frac{\partial}{\partial\sigma^2}\ell(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$0 = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

(d)

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$\frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\frac{\partial}{\partial \sigma^2}\ell(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

(e) Bias of the MLE μ

$$\text{Bias}(\hat{\mu}) = E[\hat{\mu}] - \mu$$

$$\text{Bias}(\hat{\mu}) = \mu - \mu = 0$$

Bias of the MLE σ^2

$$\text{Bias}(\hat{\sigma^2}) = E[\hat{\sigma^2}] - \sigma^2$$

$$\text{Bias}(\hat{\sigma^2}) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Question 4

(a)

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 SSW = \sum_{i=1}^{j}\sum_{i=1}^{n}(y_i - \overline{y}_i)^2 SSB = \sum_{i=j}^{n}(\overline{y}_j - \overline{y})^2$$

- Can use the Pythagorean Theorem with $y_i - \overline{y}$ being the hypotenuse and the two sides being $(\overline{y}_j - \overline{y})^2$ and $(y_{ij} - \overline{y}_j)^2$
- Proof:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right)^2 \\
&= \sum_{i=1}^{n}\left((y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\right) \\
&= \sum_{i=1}^{n}\left((y_i - \hat{y}_i)^2\right) + \sum_{i=1}^{n}\left(\epsilon_i^2\right) + \sum_{i=1}^{n}\left(2\epsilon_i(\hat{y}_i - \bar{y})\right) \\
&= \sum_{i=1}^{n}\left((\hat{y}_i - \bar{y})^2\right) + \sum_{i=1}^{n}\left(\epsilon_i^2\right) + \sum_{i=1}^{n}\left(2\epsilon_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} - \bar{y})\right) \\
&= \sum_{i=1}^{n}\left((\hat{y}_i - \bar{y})^2\right) + \sum_{i=1}^{n}\left(\epsilon_i^2\right) + 2\left(\beta_0 - \bar{y}\right)\sum_{i=1}^{n}\left(\epsilon_i\right) + 2\beta_1\sum_{i=1}^{n}\left(\epsilon_i x_{i1}\right) + \ldots + 2\beta_p\sum_{i=1}^{n}\left(\epsilon_i x_{ip}\right) \\
&= \sum_{i=1}^{n}\left((\hat{y}_i - \bar{y})^2\right) + \sum_{i=1}^{n}\left(\epsilon_i^2\right) \\
&= SSW + SSB
\end{aligned}$$

(b)

$$H_o : \mu_1 = \mu_2 = \mu_3$$

(c)

- SST (Total Sum of Squares) measures the total variance in the data. It quantifies how much the data points vary from the overall mean of the data.
- SSW (Sum of Squares Within) measures the variance within each group. It quantifies how much the data points within each group vary from their respective group means.
- SSB (Sum of Squares Between) measures the variance between the groups. It quantifies how much the group means vary from the overall mean of the data.

Question 5

(a)

- I would use a typical hypothesis test to compare the means between the 25mg group and the 50mg group. A two-sample t-test would be the best test to compare just those two groups against each other.

(b)

- I would use a one-way Analysis of Variance (ANOVA) test to compare the average response values between the three groups.

(c)

- $y = \beta_0 + \beta_1 x + \varepsilon$ where y is the response, x is the predictor, β_o is the y-intercept, β_1 is the slope, and ε is the error

(d)

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ where y is the response, x is the predictor, β_o is the y-intercept, β_1 is the coefficient, β_2 is the coefficient for the quadratic (dosage squared),and ε is the error

Question 6

(a)

- Likelihood for SLR is: $L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}\right)$

(b)

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right]$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left[\frac{(y_i - \beta_0 - \beta_1 x_i) x_i}{\sigma^2} \right]$$

$$\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Question 7

- The formula $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$
- SS_{res} is the residual sum of squares which is the difference between the observed and predicted values. When we add an irrelevant covariate the SS_{res} will decrease and SS_{tot} will stay the same

Question 8

(a)

- We can use the model of $y = Q_0 e^{kt}$ then transform the data to $z = log(Y)$ which gives us $z = c + kt$ and we get the confidence interval for k

(b)

- If we use the same model for a total of five cultures to calculate the average growth rate $k_1, , , , , , k_5$ we would have a major flaw having to split up our data while sharing the same σ^2
- A better formula would be $y_i t = Q_0 i e^{k_t + \epsilon_{it}}$ to allow us to run one calculation. A new logarithm would be $z_{it} = kt + \epsilon_{it}$

Question 9

(a)

- The two-sample t-test is $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ and the ANOVA F-statistic is $F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$
- The ANOVA $F - statistic$ is the square of the $t - statistic$ from the $t - test$ since there is only one degree of freedom for the between-group variability for the ANOVA test.

(b)

- In the model this $I(j == 1)$ means that the model will estimate the mean of group 1 when $j = 1$ and the mean of group 2 when $j = 0$ with the difference of the means being the coefficient β_1
- If you test that β_1 is significantly different than 0, then $t - value$ from the test would be the same as the $t - value$ from a two-sample $t - test$

Question 10

(a)

```
##     2.5 %    97.5 %
## 0.9514971 1.0126658
```

- Our 95% confidence interval is 0.951 to 1.013. 1 is a plausible value for β_1 this is because returns are likely to be similar from week to week.

(b)

$$H_0 : \beta_0 = 10000 H_1 : \beta_0 \neq 10000$$

```
## [1] 0.7517807
```

- Conduct a $t - test$ and the $p - value$ is 0.75. This means we fail to reject the null hypothesis.

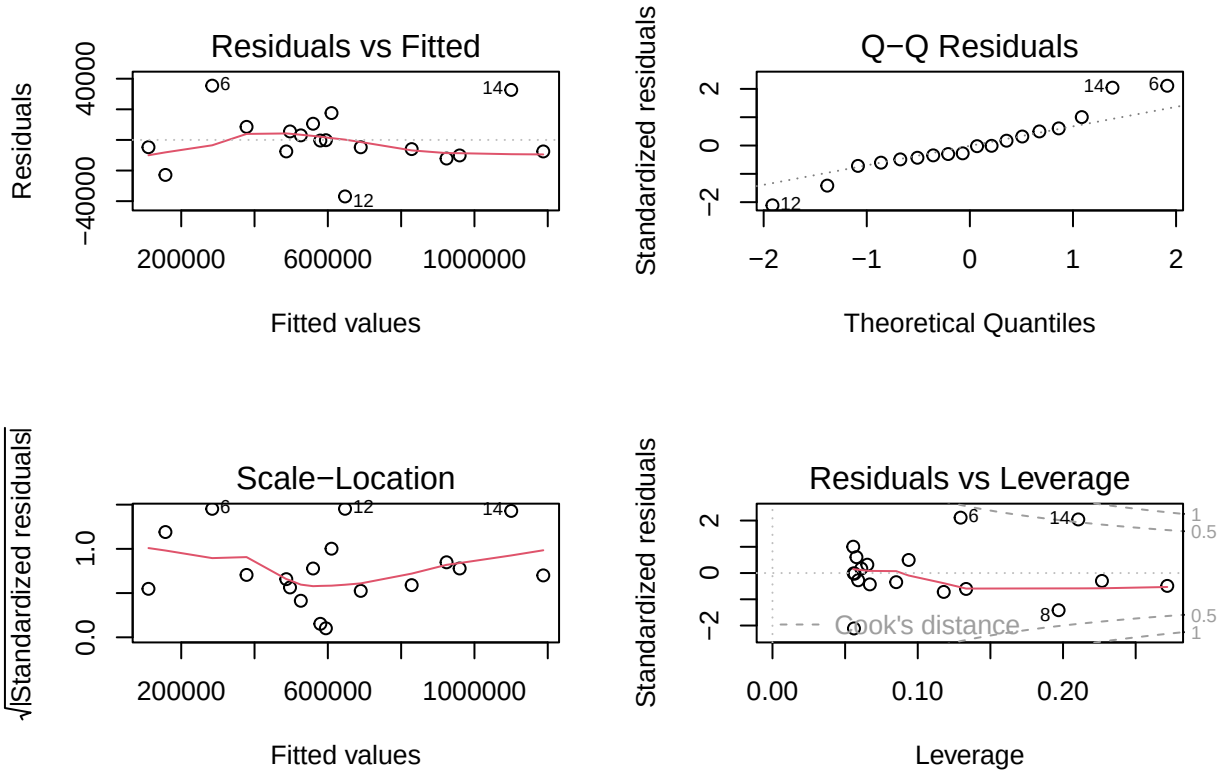
(c)

```
##        fit      lwr      upr
## 1 399637.5 359832.8 439442.2
```

- A $450,000 gross box office results is not feasible. This is because our upper limits on a 95% prediction interval is $439,442.20.

(d)

- This rule is acceptable because of the nearly perfect correlation from one week to the next. But if we look at the residuals the we can see that there are three values that do not fit.



Question 11

(a)

```
## [1] 0.4012042 0.8822156
```

- The 95% confidence interval is 0.401 to 0.882

(b)

$$H_0 : \beta_1 = 0.01 H_1 : \beta_1 \neq 0.01.$$

```
## [1] 0.1253666
```

- We fail to reject the null hypothesis becuase the $t - value$ is -1.58 and the $p - value$ is 0.12. We cannot say that the true average processing time is significantly different from 0.01 hours.

(c)

- The point estimate is 2.11 and the 95% CI is 1.45 to 2.77