Homework 9
of
STAT 3355 Data Analysis for Statisticians & Actuaries

Due: 11:30 am

May 1 (Monday), 2022

Problem 1 (0 point)

Please complete the teamwork peer-review evaluation online: https://forms.gle/pfQDpyDJCoTYMvaX8. Your feedback will be fully credential. Your team project total score will be adjusted based on the feedback from your teammates.

Problem 2 (0 point)

For a simple linear model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$.

- Derive the maximum likelihood estimators for β_0, β_1, and σ^2.
  Answer: The maximum likelihood estimators are the solution of the system of simultaneous equations, each of which is the the derivative of the (log) data likelihood with respect to each of the parameters $(\beta_0, \beta_1, \sigma^2)$. Let's first write down the data likelihood of observing n samples $Y_1, Y_2, \ldots, Y_n$ from a normal distribution, of which p.d.f. is $f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}$:

$$\begin{aligned}
L(\beta_0, \beta_1, \sigma^2 | y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^{n} \frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \\
\log L(\beta_0, \beta_1, \sigma^2 | y_1, \ldots, y_n) &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
&= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 + \beta_0^2 - \beta_1^2 x_i^2 + 2\beta_0\beta_1 x_i - 2\beta_0 y_i - 2\beta_1 x_i y_i) \\
&= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 + n\beta_0^2 - \beta_1^2\sum_{i=1}^{n} x_i^2 + 2\beta_0\beta_1\sum_{i=1}^{n} x_i - 2\beta_0\sum_{i=1}^{n} y_i - 2\beta_1\sum_{i=1}^{n} x_i y_i\right).
\end{aligned}$$

Next we need to solve

$$\begin{cases} \frac{\partial \log L(\beta_0,\beta_1,\sigma^2|y_1,\ldots,y_n)}{\partial \beta_0} &= 0 \\ \frac{\partial \log L(\beta_0,\beta_1,\sigma^2|y_1,\ldots,y_n)}{\partial \beta_1} &= 0 \\ \frac{\partial \log L(\beta_0,\beta_1,\sigma^2|y_1,\ldots,y_n)}{\partial \sigma^2} &= 0 \end{cases} \Rightarrow \begin{cases} -\frac{1}{2\sigma^2}\left(2n\beta_0 + 2\beta_1\sum_{i=1}^n x_i - 2\sum_{i=1}^n y_i\right) &= 0 \\ -\frac{1}{2\sigma^2}\left(2\beta_1\sum_{i=1}^n x_i + 2\beta_0\sum_{i=1}^n x_i - 2\sum_{i=1}^n x_i y_i\right) &= 0 \\ -n\sigma^{-1} - \frac{1}{2}(-\sigma)^{-3}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 &= 0 \end{cases}.$$

The first two equations can be simplified as

$$\begin{cases} \beta_0 n + \beta_1\sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0\sum_{i=1}^n x_i + \beta_1\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \end{cases},$$

where we can get

$$\begin{aligned} \beta_1 &= \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{n\sum_{i=1}^n x_i y_i - n\bar{x}n\bar{y}}{n\sum_{i=1}^n x_i^2 - (n\bar{x})^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ &= \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} \\ &= \frac{S_{xy}}{S_{xx}} \end{aligned}$$

via multiplying $\sum_{i=1}^n x_i$ on both sides of the first equation, multiplying n on both sides of the second equation, and then subtracting the first by the second one. Note that $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $S_{xy} = \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})$, and $S_{xx} = \sum_{i=1}^n (x_i-\bar{x})^2$. Given that $\beta_1 = S_{xy}/S_{xx}$, we can easily get $\beta_0 = \bar{y} - \bar{x}\beta_1$ by plugging the solution of β_1 in the first equation. Given the solution of (β_0, β_1), we can easily get $\sigma_2 = \frac{1}{n}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ by plugging the solution of β_0 and β_1 in the third equation. Therefore, the MLEs are

$$\begin{cases} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2 \end{cases}$$

- Write down the formulae to compute the $(1-\alpha)\times 100\%, \alpha \in (0,1)$ confidence intervals for β_0 and β_1, if σ^2 is known.
  Answer: If $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$ and σ^2 is known, then

$$\begin{cases} \hat{\beta}_1 &\sim \mathrm{N}\left(\beta_1, \sigma^2\frac{1}{S_{xx}}\right) \\ \hat{\beta}_0 &\sim \mathrm{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right) \end{cases} \Rightarrow \begin{cases} \frac{\hat{\beta}_1-\beta_1}{\sigma\sqrt{\frac{1}{S_{xx}}}} &\sim \mathrm{N}(0,1) \\ \frac{\hat{\beta}_0-\beta_0}{\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} &\sim \mathrm{N}(0,1) \end{cases} \Rightarrow \begin{cases} P\left(-z^* \le \frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{\frac{1}{S_{xx}}}} \le z^*\right) &= 1-\alpha \\ P\left(-z^* \le \frac{\hat{\beta}_0-\beta_0}{\sigma/\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} \le z^*\right) &= 1-\alpha \end{cases}.$$

  Therefore, $(1-\alpha)\times 100\%$ confidence interval for β_0 and β_1 are

$$\begin{cases} \beta_1 &\in \left[\hat{\beta}_1 - z^*\sigma\sqrt{\frac{1}{S_{xx}}}, \hat{\beta}_1 + z^*\sigma\sqrt{\frac{1}{S_{xx}}}\right] \\ \beta_0 &\in \left[\hat{\beta}_0 - z^*\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}, \beta_0 + z^*\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right] \end{cases},$$

  where a desirable setting of z^* via `qnorm(1 - α/2, mean = 0, sd = 1)`. For example, $\begin{cases} z^* = 1.65 & \text{if } \alpha = 0.1 \\ z^* = 1.96 & \text{if } \alpha = 0.05 \\ z^* = 2.58 & \text{if } \alpha = 0.01 \end{cases}$

Problem 3 (0 point)

Simulate a sample of $y_1, \ldots, y_{100}$ from a simple linear model $Y = 1 + 2x + \epsilon$, where $\epsilon \sim \mathrm{N}(0, 6^2)$, and x is an arithmetic sequence from 1 to 100, with a step size of 1. Run `set.seed(1)` to set the seed of R's random number generator so that the simulation can be reproduced.

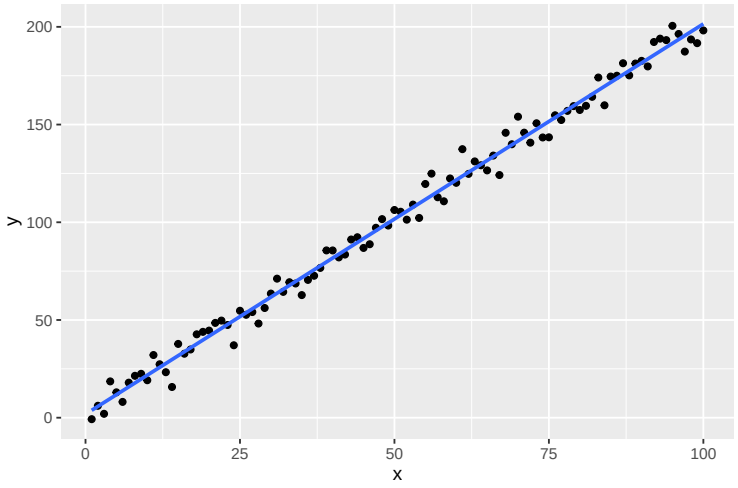
- Make a scatter plot, and fit the data with a regression line.

  Answer:

```
# Load library
library(ggplot2)

# Generate simulated data
set.seed(1)
x <- 1:100
y <- rnorm(100, mean = x, sd = 6)
data <- data.frame(x = x, y = y)

# Make a scatter plot and add a regression line
ggplot(data) + geom_point(mapping = aes(x = x, y = y)) +
   geom_smooth(mapping = aes(x = x, y = y), method = "lm",
   se = FALSE)
```



- Perform a two-sided significance test that $\beta_1 = 2$ versus the alternative that $\beta_1 \neq 2$, following the procedure of seven steps in the lecture note (choose a significance level of 0.05).

  Answer: Let X be a continuous random variable. According to the simulated data, we obtain $n = 100$, $\bar{x} = 50.5$, $\bar{y} = 102.653$, $S_{xy} = 166424.5$, $S_{xx} = 83325$, $\hat{\beta}_1 = S_{xy}/S_{xx} = 1.997$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.790$, and $\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2} = 5.416$.

1. Specify some model for the data: $Y \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$;
2. Identify the null and alternative hypotheses: $H_0 : \beta_1 = 2$ vs. $H_0 : \beta_1 \neq 2$;
3. Specify a test statistic: $t = \frac{\hat{\beta}_1 - 2}{\hat{\sigma}/\sqrt{S_{xx}}}$;
4. Compute the observed value of the test statistic: $z = \frac{\hat{\beta}_1 - 2}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{1.997 - 2}{5.4\hat{1}6/\sqrt{166424.5}} = -0.144$;
5. Compute the p-value under H_0 via `2 * (1 - pt(abs(`t`), df = `n` - 2))`: p-value$= 0.886$;
6. Specify a significance level $\alpha = 0.05$;
7. Compare the p-value and the significance level α: as p-value$= 0.885 > \alpha = 0.05$, H_0 cannot be rejected, which means the data supports $\beta_1 = 2$.

Problem 4 (0 point)

The cost of a house is related to the number of bedrooms it has. Suppose the following table contains data recorded for homes in Dallas.

Price in USD	$300,000	$250,000	$400,000	$550,000	$317,000	$389,000	$425,000	$289,000	$389,000
Number of bedrooms	3	3	4	5	4	3	6	3	4

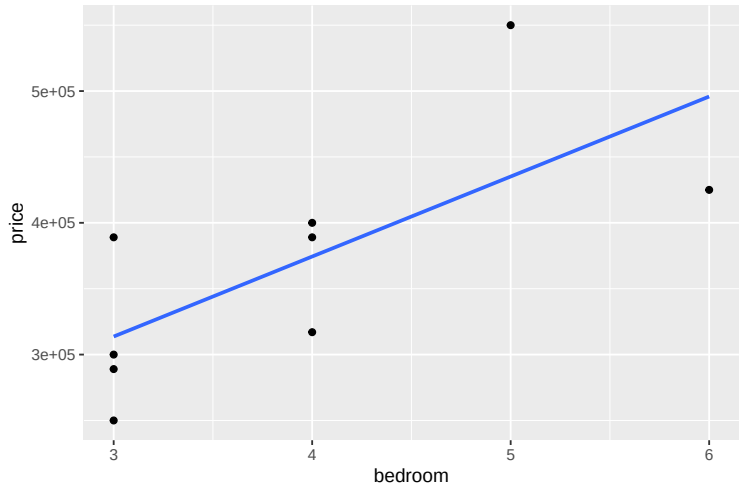
- Make a scatter plot, and fit the data with a regression line.

  Answer:

```
# Load library
library(ggplot2)

# Input data
x <- c(3, 3, 4, 5, 4, 3, 6, 3, 4)
y <- c(300, 250, 400, 550, 317, 389, 425, 289, 389)*1000
house <- data.frame(price = y, bedroom = x)

# Make a scatter plot and add a regression line
ggplot(house) + geom_point(mapping = aes(x = bedroom, y =
   price)) + geom_smooth(mapping = aes(x = bedroom, y =
   price), method = "lm", se = FALSE)
```



- Use the function `prediction()` to find the confidence intervals for the mean price of 2-bedroom house to 8-bedroom houses

  Answer:

```
# Fit a simple linear regression model
```

```
m <- lm(price ~ bedroom, data = house)

# Predict the house price
predict(m, newdata = data.frame(bedroom = 2:8), interval =
   "confidence"))
```

Number of bedrooms	2	3	4	5	6	7	8
Lower bound	$136,928	$241,198	$320,036	$354,066	$368,959	$378,958	$387,276
Fitted value	$252,988	$313,700	$374,413	$435,125	$495,838	$556,550	$617,263
Upper bound	$369,047	$386,202	$428,789	$516,185	$622,716	$734,143	$847,249

Problem 5 (0 point)

The dataset `deflection` in the package `UsingR` contains deflection measurements for various loads. Fit a linear model to the variable `Deflection` as a function of `Load`.

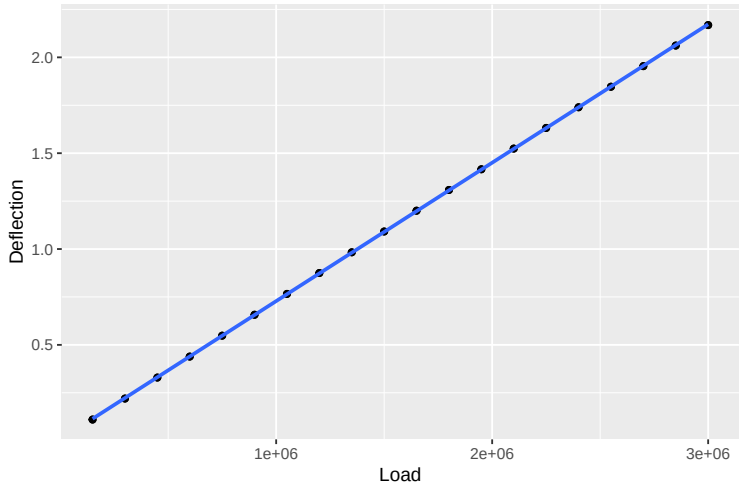
- Make a scatter plot, and fit the data with a regression line.

  Answer:

```
# Load libraries
 library(UsingR)
 library(ggplot2)

 # Load data
 data("deflection")

 # Make a scatter plot and add a regression line
 ggplot(deflection) + geom_point(mapping = aes(x = Load, y =
     Deflection)) + geom_smooth(mapping = aes(x = Load, y =
    Deflection), method = "lm", se = FALSE)
```



- What are the 95% confidence intervals for β_0 and β_1 in the above simple linear regression model?

  Answer: The 95% confidence intervals for β_0 and β_1 can be obtained via the summary function via `confint(lm(Deflection ~ Load, data = deflection))`. For β_0, it is $(4.706 \times 10^{-3}, 7.593 \times 10^{-3})$; For β_1, it is $(7.213 \times 10^{-7}, 7.229 \times 10^{-7})$