HOMEWORK 2

Yiqiao Yin [YY2502]

Contents

PROBLEM 1 (Q1)

We want to investigate the relationship between Sales of Product (Y) and population (X). According to model (2.1) from textbook, we have

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

where

- β_0 and β_1 are parameters
- X_i are known covariate
- ϵ_i are independent $\mathcal{N}(0, 1)$

We are given model (the upper / lower bound is under 95% confidence interval):

Parameter	Estimated Value	Lower Bound	Upper Bound
Intercept	7.43119	-1.18518	16.0476
Slope	0.755048	0.452886	1.05721

Answer:

1) We have the model $Y_i = 7.43119 + 0.755048 \cdot X_i$ from analysis. We observe from model that there is a positive linear coefficient, i.e. $\beta_1 = 0.755048$ from the results. We can set up hypothesis:

- H_0: $\beta_1 = 0$ (null hypothesis states that there is no relationship)
- H_1: $\beta_1 \neq 0$ (alternative hypothesis states that there is a relationship)

Next, we compute test statistics. Notice that the upper/lower is 95% so this implies that level of significance is $\alpha = 5\%$ for this test. This means that under null hypothesis, we have critical value to 1.96. In other words, we may find the standard error for β_1 to be

$$\hat{\beta}_1 \pm 1.96 \cdot \text{SE}(\hat{\beta}_1) = [0.452886, 1.05721] \Rightarrow \text{SE}(\hat{\beta}_1) = \frac{1.05721 - 0.452886}{2} \frac{1}{1.96} = 0.1541643 \approx 0.15$$

Hence, we compute test statistic

$$\text{test-statistic} = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{0.755048}{0.1551643} = 4.866119 \approx 4.87$$

Since test-statistic $\approx 4.57 > 1.96$, we reject H_0. Hence, we conclude that there is sufficient evidence to reject null hypothesis and that there is a statistically significant relationship between X and Y.

Remark: p-value is

```
dnorm(4.87)
```

```
## [1] 2.823909e-06
```

2) The worst case scenario is indeed "no sales" which means the sales has minimum value 0 in real world. However, the 95% confidence interval is a theoretical value taking variation into consideration. There is no real world meaning behind the negative value. Theoretically speaking, we model Y by $Y = \beta_0 + \beta_1 X + \epsilon_i$. If we run a regression of $Y \sim \beta_0 + \beta_1 X$, we get a decent model but the randomness is still there in the data, ϵ_i. Technically speaking, taking dollar sales to zero, we are still left with $Y \sim \beta_0 + \epsilon_i$. Since it is randomness, we can never fully capture exactly what that is in real data. This is why we use $\text{SE}(\hat{\beta}_0)$ to help us derive a confidence interval using some α level (usually $\alpha = 5\%$) to understand how much the estimated value can vary.

PROBLEM 2 (Q4)

From Q1.19, we have the following

Parameter	Estimated Value	Std Error
Intercept	2.11405	0.32089
Slope	0.03883	0.01277

1) Compute 99% confidence. In other words, we are using $\alpha = 0.01$.

```
# critical = qnorm(1-0.01/2); critical
critical = qt(1-0.01/2, 118); critical
```

```
## [1] 2.618137
```

We can use

$$\hat{\beta}_1 \pm 2.618 \cdot \text{SE}(\beta_1) = 0.03883 \pm 2.618 \cdot 0.01277 \Rightarrow [0.00539, 0.07226]$$

The confidence interval does not contain zero.

2) Hypothesis Testing

Let us state the hypothesis:

- H_0: $\beta_1 = 0$
- H_1: $\beta_1 \neq 0$

Compute test statistic

$$\text{test-statistic} = \frac{0.03883}{0.01277} \approx 3.04072$$

The critical value, under significance level of 0.01, is

```
# critical = qnorm(1-0.01/2); critical
critical = qt(1-0.01/2, 118); critical
```

```
## [1] 2.618137
```

We observe that test-statistic is 3.04 which is greater than critical value of 2.61. Thus, we reject H_0 and conclude that we have sufficient evidence to say that there is a linear association between ACT scores and GPA end of freshman year.

3) Compute p-value

```
p_Value = dnorm(0.03883/0.01277); p_Value
```

```
## [1] 0.00391896
```

```
# p_Value = dt(0.03883/0.012777, 118); p_Value
```

The p-value is 0.0039 which is less than 1% significance level. It is consistent with the conclusion made from computing test-statistics.

PROBLEM 3 (Q7)

Let us load the data and run linear regression model.

```
setwd("C:/Users/eagle/OneDrive/Course/CU Stats/STATS GR6101 - Applied Statistics I/Data")
data = read.csv("CH01PR22.csv", header = FALSE)
colnames(data) = c("Hardness", "ElapseTime")

print("Presents data: ")
```

```
## [1] "Presents data: "
```

```
data
```

```
##    Hardness ElapseTime
## 1       199         16
## 2       205         16
## 3       196         16
## 4       200         16
## 5       218         24
## 6       220         24
## 7       215         24
## 8       223         24
## 9       237         32
## 10      234         32
## 11      235         32
## 12      230         32
## 13      250         40
## 14      248         40
## 15      253         40
## 16      246         40
```

```
LM1 <- lm(data$Hardness~data$ElapseTime)
```

The data has two columns: - let the covarite to be elapsed time, this is X_i - let the response to be hardness, this is Y_i

Let us run linear regression

```
summary(LM1)
```

```
##
## Call:
## lm(formula = data$Hardness ~ data$ElapseTime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1500 -2.2188  0.1625  2.6875  5.5750
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     168.60000    2.65702   63.45  < 2e-16 ***
## data$ElapseTime   2.03438    0.09039   22.51 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.234 on 14 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9712
## F-statistic: 506.5 on 1 and 14 DF,  p-value: 2.159e-12
```

This is the results from Q1.22 which Q7 in Chapter 2 refers to.

1) From the regression results above we have $\hat{\beta}_0 = 168.6$ and $\hat{\beta}_1 = 2.03438$. Thus, we may write linear model to be

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X = 168.6 + 2.03438X$$

The results imply that a single unit change of X increases Y by 2.03438. Recall critical value at 99%:

```r
# critical = qnorm(1-0.01/2); critical
critical = qt(1-0.01/2, 14); critical
```

```
## [1] 2.976843
```

and thus we have

$$99\% \text{ confidence interval} = [2.03438 - 2.977 \cdot 0.09039, 2.03438 + 2.977 \cdot 0.09039] = [1.765289, 2.303471]$$

and we observe that zero does not fall in this range.

2) Hypothesis Testing

Let us state the hypothesis:

- $H_0 : \hat{\beta}_1 = 2$
- $H_1 : \hat{\beta}_1 \neq 2$

Compute test statistics:

$$\text{test-statistics} = \frac{\hat{\beta}_1 - 2}{\text{SE}(\hat{\beta}_1)} = \frac{2.03438 - 2}{0.09039} = 0.3803518 \approx 0.38$$

Compare test-statistic with critical value

$$\text{test-stat} = 0.38 < 2.977 = \text{critical value}$$

and also notice that p-value is

```r
dt(0.38, 14)
```

```
## [1] 0.3628588
```

which is greater than 1%.

Hence, we fail to reject H_0 and conclude that we do not have enough evidence to say that the linear association is not 2.

3) Compute power of the test

$$|.3|/0.1 = 3 \rightarrow \text{Power} = 0.5$$

PROBLEM 4 (Q63)

From Homework 1, we have obtained the following:

Risk	Results	Formula	$\text{SE}(\hat{\beta}_0)$	$\text{SE}(\hat{\beta}_1)$
MSE1	7192580	Y~f(X|Geography=1(NE)), Y = 9223.82 + 522.16 X	851.77	37.13
MSE2	4329650	Y~f(X|Geography=2(NC)), Y = 13581.41 + 238.67 X	575.14	27.23
MSE3	7376003	Y~f(X|Geography=3(S)), Y = 10529.79 + 330.61 X	612.48	27.13
MSE4	8000959	Y~f(X|Geography=4(W)), Y = 8615.05 + 440.32 X	1052.20	45.37

With the above information, we can answer the following questions:

1) We want to compute confidence interval of β_1 for each of the linear models above.

Model 1:

$$99\% \text{ confidence interval for } \beta_1 = [522.16 - 1.64 \cdot 37.13, 522.16 + 1.64 \cdot 37.13] = [461.2668, 583.0532]$$

Model 2:

$$99\% \text{ confidence interval for } \beta_1 = [238.67 - 1.64 \cdot 27.23, 238.67 + 1.64 \cdot 27.23] = [194.0128, 283.3272]$$

Model 3:

$$99\% \text{ confidence interval for } \beta_1 = [330.61 - 1.64 \cdot 27.13, 330.61 + 1.64 \cdot 27.13] = [286.1168, 375.1032]$$

Model 4:

$$99\% \text{ confidence interval for } \beta_1 = [440.32 - 1.64 \cdot 45.37, 440.32 + 1.64 \cdot 45.37] = [365.9132, 514.7268]$$

All 4 linear models have positive slopes. That means, we are expecting similar linear association between Bachelor Degree X and Income per Capita Y for each of the regions. Though the numbers vary, they stay around 200 to 530 range most of the time.

2) The estimated regression models have different $\hat{\beta}_1$s, but they are all positive and stay in the range from 200 to 530.

PROBLEM 5

This problem is about Wilcoxin Rank-sum Test. Consider $X_1, ..., X_n$ and $Y_1, ..., Y_m$. We want to test whether they have equal mean. Let us state the hypothesis

- $H_0 : \mu_X = \mu_Y$
- $H_1 : \mu_X \neq \mu_Y$

1) Let R_j while $j = 1, 2, ..., n$ be the ranks of X_i among the stacked data of $[X_1, X_2, ..., X_n, Y_1, ..., Y_m]$. Then we can compute test statistic

$$\text{W-statistic} = \sum_{j=1}^{n} R_j$$

Under H_0, we have $\mathbb{E}W = n(n+1)/2$ and $\text{var}(W) = (mn(n+a)/12)$. We also have the asymptotic distribution

$$\frac{\text{W-stat} - \mathbb{E}W_{H_0}}{\sigma(W)_{H_0}} = \frac{\text{W-stat} - \frac{n(n+1)}{2}}{\sqrt{\text{var}(\frac{mn(n+1)}{12})}} \to \mathcal{N}(0,1)$$

This is consistent with t test or z test. We reject H_0 if W-stat is greater than $\mathbb{E}W + z_\alpha \sqrt{\text{var}(W)}$

2) If we have data with sufficient sample size, we can find p-value by using standard normal distribution and search for probability that the test statistic is greater than critical value. If we have data lack of sufficient sample size, we can find p-value by using t-distribution adjusting for degrees of freedoms and search for proabability that the test statistic is greater than critical value.

3) From part a), we discussed that the asymptotic distribution Wilcoxin test statistic under null hypothesis follow standard normal distribution. This means that we can say, for certain significance level, the probability of test statistic within certain range of critical value to be the following

$$\mathbb{P}(-z_{\alpha/2} < \frac{\text{W-stat} - \frac{n(n+1)}{2}}{\sqrt{\text{var}(\frac{mn(n+1)}{12})}} < z_{\alpha/2}) = 1 - \alpha$$

and by moving equations around we get the following confidence interval

$$[\text{LB}, \text{UB}] = \left[\frac{n(n+1)}{2} - z_{\alpha/2} \cdot \sqrt{\text{var}\frac{mn(n+1)}{12}}, \frac{n(n+1)}{2} + z_{\alpha/2} \cdot \sqrt{\text{var}\frac{mn(n+1)}{12}}\right]$$

of which the lower bound, LB, and the uppwer bound, UB, are defined as above.