Yuhan Zhu
A54482257

STT 863 Homework 2

2.1. A student working on a summer internship in the economic research department of a large corporation studied the relation between sales of a product (Y, in million dollars) and population (X, in million persons) in the firm's 50 marketing districts. The normal error regression model (2.1) was employed. The student first wished to test whether or not a linear association between Y and X existed. The student accessed a simple linear regression program and obtained the following information on the regression coefficients:

Parameter	Estimated Value	95 Percent Confidence Limits	
Intercept	7.43119	−1.18518	16.0476
Slope	.755048	.452886	1.05721

a. The student concluded from these results that there is a linear association between Y and X. Is the conclusion warranted? What is the implied level of significance?

b. Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.

a. Let's say H_0: β_1 (slope) $= 0$ no linear relationship
H_a: $\beta_1 \neq 0$ there are linear relationship

From the normal error regression model we get the 95% confidence interval $(0.452886, 1.05271)$ which does not contain 0 in it.

$\therefore$ The 95% CI reject the H_0, the conclusion is warranted. X and Y has linear association.

Confidence level is 95% means there are a 5% chance of incorrectly rejecting the null hypothesis.

b. It is theoretically possible for the intercept to be negative in the regression model.

It may not have meaniful interpretation in real-world. like this case, sales etc.

So, it is important to understand the limition of a model and model should be interpreted with caution.

*2.5. Refer to **Copier maintenance** Problem 1.20.

a. Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

b. Conduct a t test to determine whether or not there is a linear association between X and Y here; control the α risk at .10. State the alternatives, decision rule, and conclusion. What is the P-value of your test?

c. Are your results in parts (a) and (b) consistent? Explain.

d. The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e. Does b_0 give any relevant information here about the "start-up" time on calls—i.e., about the time required before service work is begun on the copiers at a customer location?

a.
```
```{r 2.5a}
# Load the data
data <- read.table("CH01PR20.txt", header = FALSE)
colnames(data) <- c("Y", "X")

# Fit a linear regression model
model <- lm(Y ~ X, data = data)

# Get the slope coefficient and its 90% confidence interval
b1 <- coef(model)["X"]
conf_int_90 <- confint(model, level = 0.9)["X", ]

print(paste0("Slope:",b1))
print(conf_int_90)
```
```

```
[1] "Slope:15.0352480417755"
      5 %     95 %
14.22314 15.84735
```

b. $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

If p-value < α 0.1 then we will reject the H_0

```{r 2.5b}
#p-value
summary(model)$coefficients["X", "Pr(>|t|)"]
```

```
[1] 4.009032e-31
```

p-value: $4.009 \times 10^{-31} \ll 0.1$

$H_0: \beta_1 = 0$ rejected.

There is linear association between X and Y

c.
```{r 2.5c}
summary(model)$coefficients
```

```
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -0.5801567  2.8039411 -0.2069076 8.370587e-01
X           15.0352480  0.4830872 31.1232581 4.009032e-31
```

The results in part (a) and (b) are consistent.

CI, slope, and t-test show evidence of linear association between X and Y.

d. $H_0: \beta_1 \leq 14$ $H_a: \beta_1 > 14$

If p-value < 0.05, we will reject the null hypothesis

```{r 2.5d}
beta <- 14
#t-test
t_statistic <- (b1 - beta) / summary(model)$coefficients["X", "Std. Error"]
#p-value for the one-tailed test
p_value_one_tailed <- 1 - pt(t_statistic, df = df.residual(model))4
print(paste(t_statistic,p_value_one_tailed))
```

```
[1] "2.14298373611535 0.0189076589843856"
```

P-value $0.01891 < 0.05$. Reject the null hypothesis $H_0: \beta_1 \leq 14$

$\therefore$ This standard is not being satisfied by Tri-city.

e. $b_0 = -0.5801567$

I think this b_0 is not meaningful in reality due to it is less than 0. It doesn't necesary mean the "start-up". time. It might indicate the model is not perfectly describe the relationship between X and Y.

*2.14. Refer to **Copier maintenance** Problem 1.20.

a. Obtain a 90 percent confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your confidence interval.

b. Obtain a 90 percent prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be?

c. Management wishes to estimate the expected service time *per copier* on calls in which six copiers are serviced. Obtain an appropriate 90 percent confidence interval by converting the interval obtained in part (a). Interpret the converted confidence interval.

d. Determine the boundary values of the 90 percent confidence band for the regression line when $X_h = 6$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

a. $\hat{Y}_h \pm t_{\alpha/2}\sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$

```{r 2.14a}
X_h <- 6
Y_hat_h <- predict(model, newdata = data.frame(X = X_h))
#90% confidence level
t_critical_90 <- qt(1 - 0.05, df = df.residual(model))
#MSE
MSE <- deviance(model) / df.residual(model)
#observations
n <- nrow(data)
#X_bar
X_bar <- mean(data$X)
#standard error for the mean prediction
SE_mean_prediction <- sqrt(MSE * ((1/n) + ((X_h - X_bar)^2 / sum((data$X - X_bar)^2))))
#90% confidence interval for the mean service time
CI_mean_90 <- c(Y_hat_h - t_critical_90 * SE_mean_prediction, Y_hat_h + t_critical_90 * SE_mean_prediction)
CI_mean_90
```

```
       1        1 
87.28387 91.97880 
```

$\therefore$ (87.28387, 91.97880) is the 90% CI b.

```{r 2.14b}
SE_prediction <- sqrt(MSE * (1 + (1/n) + ((X_h - X_bar)^2 / sum((data$X - X_bar)^2))))

PI_90 <- c(Y_hat_h - t_critical_90 * SE_prediction, Y_hat_h + t_critical_90 * SE_prediction)
PI_90
```

```
        1         1 
 74.46433 104.79833 
```

90% CI for service time on the next call
(74.46433 , 104.79833)

It is wider than (a). Because the prediction interval accounts for both variability in mean service time and individual observations mean.

c.

```{r 2.14c}
CI_mean_per_copier_90 <- CI_mean_90 / X_h
CI_mean_per_copier_90
```

```
       1        1 
14.54731 15.32980 
```

We are 90% confident mean service time pre copier for calls with 6 copiers is between 14.54731 mins to 15.3298 mins.

d. In (d), the interval's behavior across all X and in (a) the interval is specific to X=6

(a) gives the Confidence interval for the mean response at X=6, (d) shows the behavior of this intervals across the range of X.

The interval for part (d) is same as (a)
(87.28 , 91.98)

2.51. Show that b_0 as defined in (2.21) is an unbiased estimator of β_0.

Given $E(b_0) = \beta_0$

$b_0 = \bar{Y} - b_1 \bar{X}$ (2.21)

we know b_1 is unbiased estimator of β_1

$E(b_1) = \beta_1$

$E(\bar{Y}) = \mu Y$ $E(\bar{X}) = \mu(X)$

$\therefore$ $b_0 = \bar{Y} - b_1 \bar{X}$

$E(b_0) = E(\bar{Y} - b_1 \bar{X})$

$E(b_0) = E(\bar{Y}) - E(b_1) E(\bar{X})$

$E(b_0) = \mu Y - \beta_1 \mu X$

$\beta_0 = \mu Y - \beta_1 \mu X$

$\mu Y = \beta_0 + \beta_1 \mu X$

$\Rightarrow$ $E(b_0) = \beta_0 + \beta_1 \mu X - \beta_1 \mu X$

$\Rightarrow$ $E(b_0) = \beta_0$

$E(b_0) = \beta_0$ shows b_0 is an unbiased estimator of β_0