Homework 1, STAT 632

Ken Vu

February 02, 2022

Loading relevant libraries.

```
library(pacman)
```

```
## Warning: package 'pacman' was built under R version 4.1.2
```

```
p_load(ggplot2, alr4)
```

Exercise 0

Here's the link to my GitHub page: \href{https://github.com/Ken-Vu}

Concept Questions

Exercise 1

Part a) The least squares regression line equation is $\boldsymbol{y = -1.1016 + 2.2606x}$.

Part b) For this problem, here are the hypothesis tests.

$$\boldsymbol{H_0 : \beta_1 = 0}$$
$$\boldsymbol{H_A : \beta_1 \neq 0}$$

Given the p-value provided in this problem (i.e. $2 \times 10^{-16} < \alpha = 0.05$), we **reject the null hypothesis**. There's sufficient evidence that the slope is not zero.

Part c) Given the intercept's t-value of -2.699, we get the p-value for the intercept to be **f**, using the *t.test()* function below.

We have degrees of freedom to be $50 - 2 = 48$ since *n=50* and we lose one degree of freedom per mean parameters being estimated (i.e. we have two here).

```
2*pt(q=-2.7,df=48)
```

```
## [1] 0.009548317
```

We have that $\boldsymbol{p \approx 0.0095}$.

Part d) To get the t-statistic, we first know that $\hat{\beta}_1 = 2.2606$, $\beta_1^0 = 0$ (from part a), and $se(\hat{\beta}_1) = 0.0981$.

Plugging into the formula for the t-statistic for the slope gets the following:

$$= \frac{2.2606 - 0}{0.0981}$$

Simplifying the above gives us the following t-statistic for the slope.

$$\boldsymbol{t \approx 23.04383}$$

Part e) When $\alpha = 0.05$ and $df = 48$, $t_{\alpha=0.05, df=48} = 2.010635$ as shown below.

```
qt(1-(0.05/2), df=48)
```

```
## [1] 2.010635
```

We have $\hat{\beta}_1 = 2.2606$ and $se(\hat{\beta}_1) = 0.0981$ so we can find the confidence interval as shown below.

$$(2.2606 - [2.010635 \cdot 0.0981], 2.2606 + [2.010635 \cdot 0.0981])$$

Simplifying the above gives the **95% confidence interval for the slope** below.

$$\boldsymbol{(2.063357, 2.457843)}$$

.

Since zero is not in this confidence interval, this confidence interval **agrees with the hypothesis test results**.

Exercise 2

Part a) First, we derive $R(\hat{\beta})$.

$$\frac{d}{dx} R(\hat{\beta}) = \sum_{i=1}^{n} 2(y_i - \hat{\beta} x_i)^{2-1}(-x_i)$$

$$= -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta} x_i)^1$$

Now, we set $R'(\hat{\beta}) = 0$ and solve for $\hat{\beta}$.

$$\frac{-2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta} x_i)}{-2} = \frac{0}{-2}$$

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta} x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - x_i^2 \hat{\beta} = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i^2 \hat{\beta} = 0$$

Then, we add the term $\sum_{i=1}^{n} x_i^2 \hat{\beta}$ to the right side of the equation and keep solving for $\hat{\beta}$.

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

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

We divide $\sum_{i=1}^{n} x_i y_i$ by $\sum_{i=1}^{n} x_i^2$ to get $\hat{\beta}$.

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

Simplifying the above gets us $\boldsymbol{\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}}$.

Thus, we've **shown** that the least squares estimate for the slope is the result above.

Part b) We have that $Y_i = \beta x_i + e_i$.

Given $E(\hat{\beta})$, we evaluate it knowing from part a that $\boldsymbol{\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}}$.

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

We then substitute y_i in the equation with the definition of Y_i. Then, we simplify.

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

$$E(\hat{\beta}) = E(\frac{\sum_{i=1}^{n} x_i(\beta x_i + e_i)}{\sum_{i=1}^{n} x_i^2})$$

$$E(\hat{\beta}) = E(\frac{\sum_{i=1}^{n} x_i \beta x_i + x_i e_i}{\sum_{i=1}^{n} x_i^2})$$

We move the denominator outside of the expectation as it's a constant.

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n} E(x_i \beta x_i + x_i e_i)}{\sum_{i=1}^{n} x_i^2}$$

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n} E(\beta x_i^2) + E(x_i e_i)}{\sum_{i=1}^{n} x_i^2}$$

The x_is can be moved outside of the expectation as they can be ssumed to be fixed. Thus, we can move them outside of the expectations.

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 E(\beta) + x_i E(e_i)}{\sum_{i=1}^{n} x_i^2}$$

Since $E(e_i) = 0$, the $x_i E(e_i)$ term is gone.

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 \beta}{\sum_{i=1}^{n} x_i^2}$$

$$E(\hat{\beta}) = \frac{\beta \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2}$$

The summations both cancel out so we get the following:

$$E(\hat{\beta}) = E(\beta)$$

$$E(\hat{\beta}) = \beta$$

Thus, we've shown that $E(\hat{\beta}) = \beta$.

Part c) We know that $Var(Y_i) = \sigma^2$ since $e_i \sim N(0, \sigma^2)$.

So $Var(\hat{\beta}) = Var(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2})$.

Since the denominator is a constant, we can pull it out.

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

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

We then substitute y_i in the equation with the definition of Y_i. Then, we simplify.

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} Var(x_i(\beta x_i + e_i))}{(\sum_{i=1}^{n} x_i^2)^2}$$

We can assume x_i to be constant so we can pull out x_i.

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 Var(\beta x_i + e_i)}{(\sum_{i=1}^{n} x_i^2)^2}$$

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 (Var(\beta x_i) + Var(e_i))}{(\sum_{i=1}^{n} x_i^2)^2}$$

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 (x_i^2 Var(\beta) + Var(e_i))}{(\sum_{i=1}^{n} x_i^2)^2}$$

Since β is a constant and $e_i \sim N(0, \sigma^2)$, $Var(\beta) = 0$ and $Var(e_i) = \sigma^2$. Now, we simplify, knowing σ^2's a constant.

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 (x_i^2 \cdot 0 + \sigma^2)}{(\sum_{i=1}^{n} x_i^2)^2}$$

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 (\sigma^2)}{(\sum_{i=1}^{n} x_i^2)^2}$$

$$Var(\hat{\beta}) = \frac{\sigma^2 (\sum_{i=1}^{n} x_i^2)}{(\sum_{i=1}^{n} x_i^2)^2}$$

Simplifying the above by canceling out one of the summations below gives us:

$$Var(\hat{\beta}) = \frac{\sigma^2}{(\sum_{i=1}^{n} x_i^2)^1}$$

Thus, we've shown that $Var(\hat{\beta}) = \frac{\sigma^2}{(\sum_{i=1}^{n} x_i^2)}$.

Data Analysis Questions

Exercise 3

Part a) First, we read in the data set.

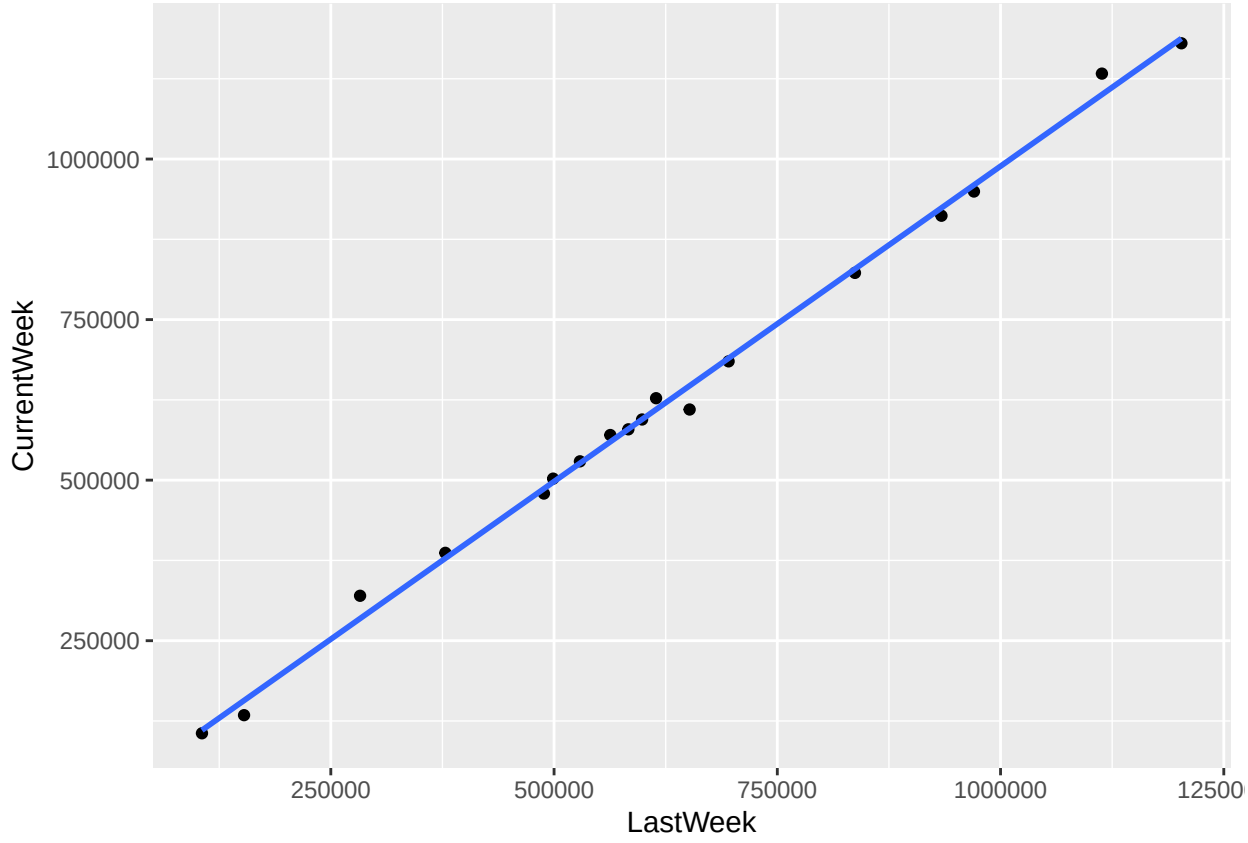
```
playbill_data <- read.csv("playbill.csv")
head(playbill_data)
```

```
##          Production CurrentWeek LastWeek
## 1       42nd Street      684966   695437
## 2          Avenue Q      502367   498969
## 3 Beauty and Beast       594474   598576
## 4     Bombay Dreams      529298   528994
## 5           Chicago      570254   562964
## 6           Dracula      319959   282778
```

Here's a scatterplot of the data.

```
ggplot(playbill_data, aes(x=LastWeek, y=CurrentWeek)) + geom_point() +
  geom_smooth(method=lm, se=F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Part b) First, we generate the linear model.

```
lm1 <- lm(CurrentWeek ~ LastWeek, data=playbill_data)
summary(lm1)
```

```
##
## Call:
## lm(formula = CurrentWeek ~ LastWeek, data = playbill_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36926  -7525  -2581   7782  35443
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.805e+03  9.929e+03   0.685    0.503
## LastWeek    9.821e-01  1.443e-02  68.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18010 on 16 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9963
## F-statistic:  4634 on 1 and 16 DF,  p-value: < 2.2e-16
```

Based on the summary table above, we have the data needed to get confidence intervals for β_0 and β_1.

Let's first get the 95% CI for β_0 and β_1, which is shown below.

```
confint(lm1)
```

```
##                     2.5 %       97.5 %
## (Intercept) -1.424433e+04 27854.099443
## LastWeek     9.514971e-01     1.012666
```

We have that the **95% CI for β_0** to be approximately **(-14244.33, 27854.0994)**.

For the **95% CI for β_1**, it's approximately, **(-0.9514971, 1.012666)**.

Based on the confidence intervals for β_1, it's **plausible** for 1 to be a value for β_1.

Part c) Here's code to get a prediction for the box office for the current week based on the previous week gross of $400,000.

```
predict(lm1, data.frame(LastWeek = 400000), interval = "predict")
```

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

We predict that with LastWeek = 400000, **CurrentWeek $\approx$ $399637.5**

Based on the results above, the **95% prediction interval** is **(359832.8, 439442.2)**. Using the 95% prediction interval, we find that **$450000 is not a plausible value** given the box office of $400000 the previous week.

Part d) While the value of CurrentWeek being equal to PreviousWeek is plausible, it's not a reliable rule to count on as the data we have is based on a limited sample of all of the box office earnings for the previous week and the current week. It's likely, but not guaranteed.

Exercise 4

Part a) Here's a summary of the generated linear model with *Interval* as the response and *Duration* as the predictor

```
lm2 <- lm(Interval ~ Duration, data = oldfaith)
summary(lm2)
```

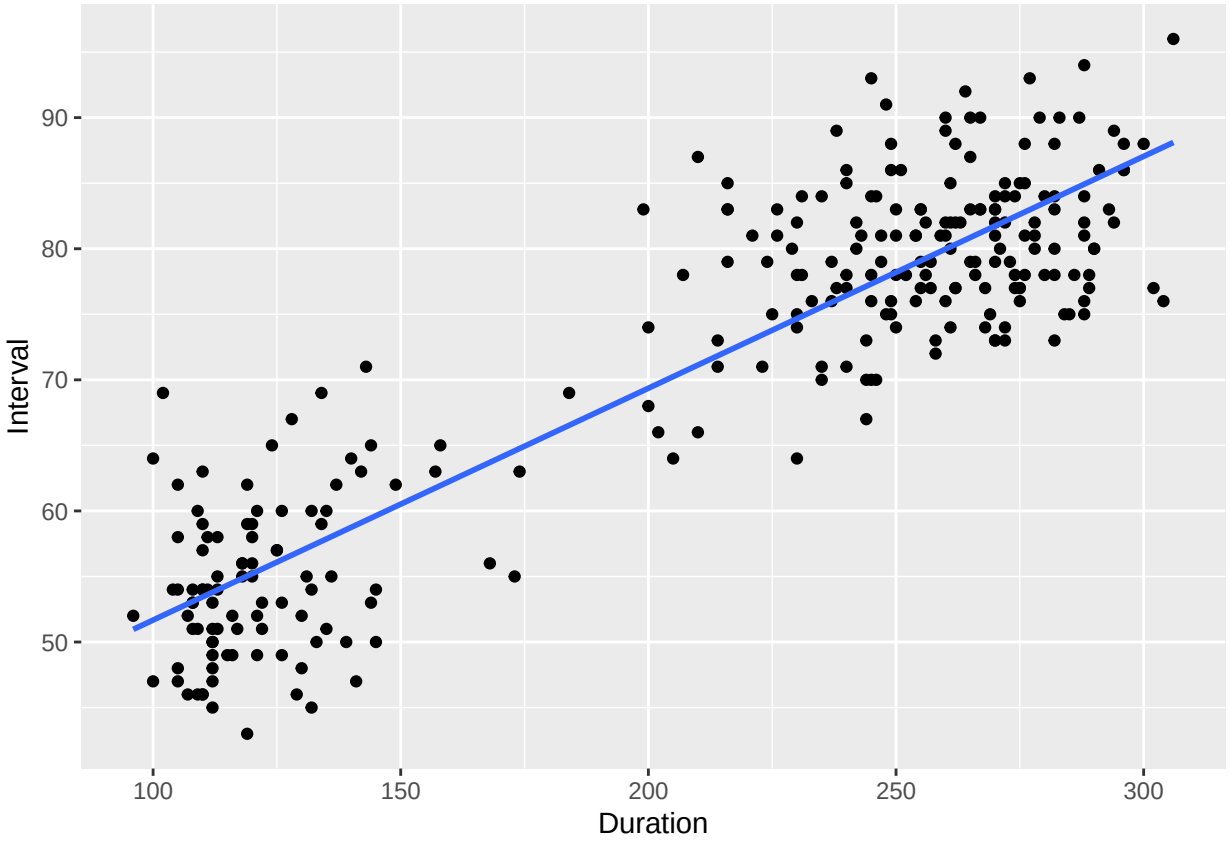
```
##
## Call:
## lm(formula = Interval ~ Duration, data = oldfaith)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3337  -4.5250   0.0612   3.7683  16.9722
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.987808   1.181217   28.77   <2e-16 ***
## Duration     0.176863   0.005352   33.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.004 on 268 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.8022
## F-statistic:  1092 on 1 and 268 DF,  p-value: < 2.2e-16
```

Part b) Here's code for making a scatter plot of *Interval* versus *Duration* with the least squares regression line imposed on it.

```
ggplot(oldfaith, aes(x=Duration, y=Interval)) + geom_point() +
  geom_smooth(method=lm, se=F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Part c) Given *Duration* = 250 seconds, **Interval ≈ 78.204 minutes**. The code is shown below.

```
predict(lm2, data.frame(Duration = 250), interval = "predict")
```

```
##        fit      lwr      upr
## 1 78.20354 66.35401 90.05307
```

Based on the results above, the **95% prediction interval** is **(66.354, 90.053)**.

Part d) Based on the summary table in *part a* of this problem, the coefficient of determination (R^2) indicates

that approximately **80.29% of the variability in the variable *Interval*** can be **explained by the variable *Duration*** .