STAT 632: Homework 1

Ashish Ashish

Due: Feb 7th, 2025 at 11:59pm

Excercise 0 : link to Github. https://github.com/ashishfreaksout/Stat632

Concept Questions

Excercise 1 :

(a) Least Squares Regression Line

The equation for the least squares regression line is given by:

$$\hat{y} = b_0 + b_1 x$$

From the regression summary:

$$b_0 = -1.1016$$
$$b_1 = 2.2606$$

Thus, the regression equation is:

$$\hat{y} = -1.1016 + 2.2606x$$

(b) Hypothesis Test for the Slope

The hypotheses for testing whether the slope is significantly different from zero are:

$$H_0 : \beta_1 = 0 \quad \text{(No relationship between } x \text{ and } y\text{)}$$
$$H_A : \beta_1 \neq 0 \quad \text{(There is a relationship between } x \text{ and } y\text{)}$$

The p-value for the slope is $< 2e-16$, which is extremely small. Since this is far below the typical significance level ($\alpha = 0.05$), we **reject the null hypothesis** and conclude that the slope is significantly different from zero.

(c) Missing p-value for the Intercept

The p-value is calculated using the t-statistic formula:

```
t_statistic <- -2.699  # t-value for the intercept
p_value <- 2 * pt(t_statistic, df = 50 - 2)# Compute the two-tailed p-value
p_value
```

```
## [1] 0.009573193
```

missing p-value is 0.0095

(d) Missing t-statistic for the Slope

for the slope:

$$t = \frac{2.2606}{0.0981} = 23.048$$

(e) 95% Confidence Interval for the Slope

A confidence interval for the slope is given by:

$$b_1 \pm t^* \cdot SE(b_1)$$

where:

- t^* is the critical value from the t-distribution with $df = 50 - 2 = 48$. For a 95% confidence level, t^* is:

```
tcrit <- qt(0.975, df=50-2) # value of tcritical
conf1 <- 2.2606 - 0.0981*tcrit #first conf interval
conf2 <- 2.2606 + 0.0981*tcrit #second conf interval
conf1
```

```
## [1] 2.063357
```

```
conf2
```

```
## [1] 2.457843
```

Since the confidence interval $(2.0633, 2.4579)$ **does not include 0**, it agrees with the hypothesis test's conclusion that the slope is significantly different from zero.

Excercise 2:

Consider the linear regression model through the origin:

$$Y_i = \beta x_i + e_i, \quad i = 1, \ldots, n \tag{1}$$

where the errors are independent and normally distributed:

$$e_i \sim N(0, \sigma^2). \tag{2}$$

(a) Finding the Least Squares Estimate of β

To find the least squares estimate of β, we minimize the residual sum of squares:

$$R(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2. \tag{3}$$

Taking the derivative with respect to β and setting it to zero:

$$\frac{d}{d\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2 = \sum_{i=1}^{n} 2(y_i - \beta x_i)(-x_i) = 0. \tag{4}$$

Expanding and solving for β:

$$\sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 = 0. \tag{5}$$

Thus, the least squares estimate of β is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{6}$$

(b) Expectation of $\hat{\beta}$

Taking the expectation:

$$E(\hat{\beta}) = E\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right). \tag{7}$$

Substituting $Y_i = \beta x_i + e_i$:

$$E(\hat{\beta}) = E\left(\frac{\sum_{i=1}^{n} x_i(\beta x_i + e_i)}{\sum_{i=1}^{n} x_i^2}\right). \tag{8}$$

Expanding the summation:

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i \beta x_i + \sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}. \tag{9}$$

Since $E(e_i) = 0$, the second summation vanishes:

$$E(\hat{\beta}) = \frac{\beta \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta. \tag{10}$$

Thus, $\hat{\beta}$ is an unbiased estimator of β.

(c) Variance of $\hat{\beta}$

Using the variance property:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right). \tag{11}$$

Substituting $Y_i = \beta x_i + e_i$:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_i(\beta x_i + e_i)}{\sum_{i=1}^{n} x_i^2}\right). \tag{12}$$

Since variance only affects the error term:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right). \tag{13}$$

Using the property that $e_i \sim N(0, \sigma^2)$ and are independent:

$$\mathrm{Var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 \sigma^2}{(\sum_{i=1}^{n} x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}. \tag{14}$$

Thus, the variance of $\hat{\beta}$ is:

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}. \tag{15}$$

Data Analysis Questions

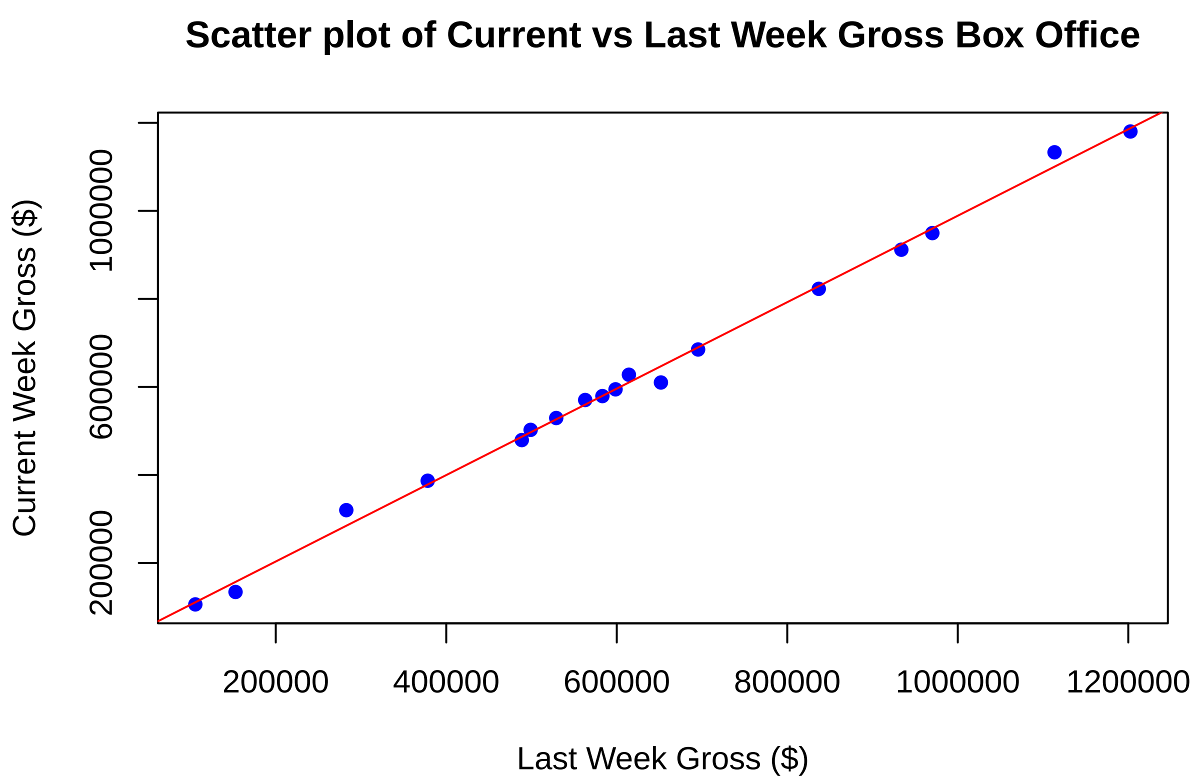
Excercise 3:

```
library(readr)
playbill <- read_csv("~/Downloads/playbill.csv")
head(playbill)
```

```
## # A tibble: 6 x 3
##   Production         CurrentWeek LastWeek
##   <chr>                    <dbl>    <dbl>
## 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
```

(a) Load the data, make a scatter plot, and fit the regression model

```r
lm1 <- lm(CurrentWeek ~ LastWeek, data = playbill)
# Scatter plot with regression line
plot(CurrentWeek ~ LastWeek,data = playbill,
     main = "Scatter plot of Current vs Last Week Gross Box Office",
     xlab = "Last Week Gross ($)",
     ylab = "Current Week Gross ($)",
     pch = 16, col = "blue")
abline(lm1, col = "red")
```



(b) Compute 95% confidence intervals for the intercept and slope

```r
confint(lm1)
```

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

The 95% confidence interval for the slope β_1 is:

$$0.9515 \leq \beta_1 \leq 1.0127$$

Since the value 1 falls within this confidence interval, it suggests that $\beta_1 = 1$ is a plausible value. This means that next week's gross box office revenue could reasonably be predicted using this week's revenue.

\subsection*{(c) Predict the gross box office for a show with $400,000 in the previous week}

```r
new_data <- data.frame(LastWeek = 400000)

# Estimate expected gross box office revenue
predict(lm1, new_data)
```

```
##        1
## 399637.5
```

```r
# Compute 95% prediction interval
predict(lm1, new_data, interval = "prediction", level = 0.95)
```

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

Using the fitted regression model, the estimated gross box office result for the current week is:

$$\hat{Y} = 399637.5$$

The 95% prediction interval for the gross box office results in the current week is:

$$(359832.8, 439442.2)$$

Since $450,000 is outside this prediction interval, it is unlikely (but not impossible) that a production with $400,000 in the previous week's gross box office will achieve $450,000 in the current week.

(d) Evaluating the Prediction Rule: "Next Week's Gross Box Office Equals This Week's Gross Box Office"

```r
#summary of the linear model
summary(lm1)
```

```
##
## Call:
## lm(formula = CurrentWeek ~ LastWeek, data = playbill)
##
## 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
```

According to the summary, the R-squared value is 0.9966 that means the model is quite efficient for prediction. That means that promotors can use this model to predict the next week's earnings.

Excercise 4:

(a) Perform Simple Linear Regression

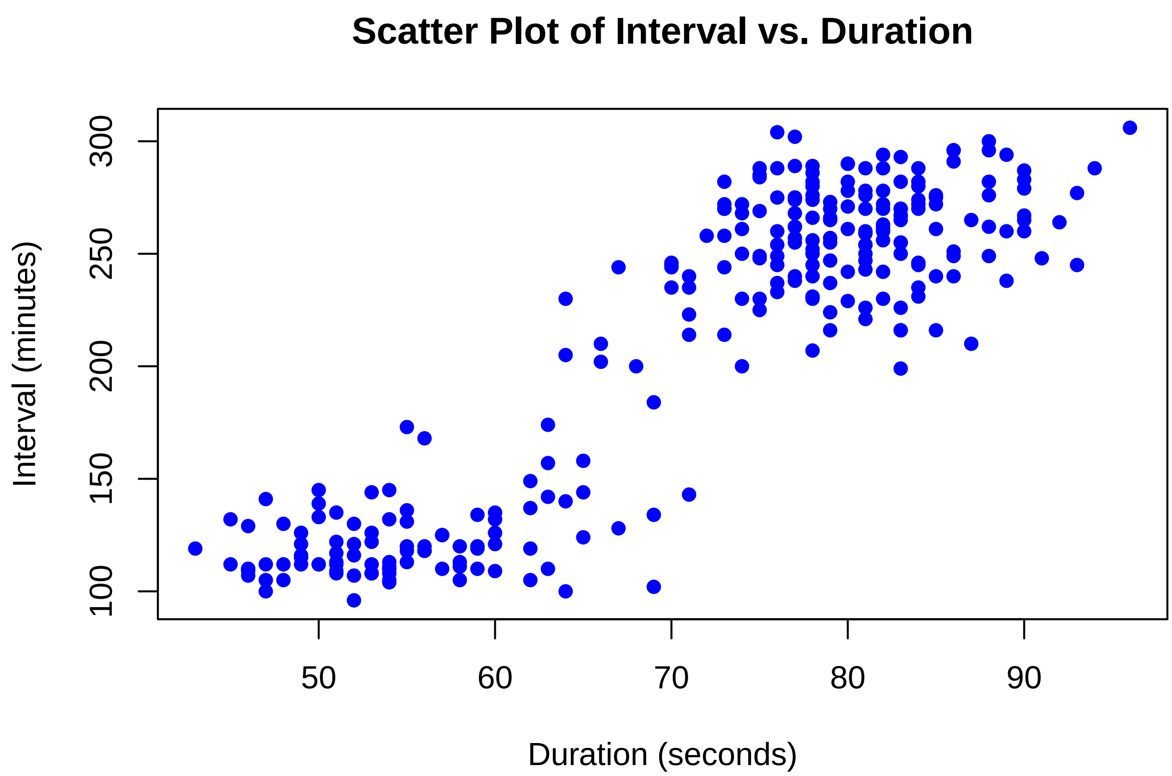
```r
library(alr4)

# Fit the linear model
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
```

(b) Scatter Plot with Regression Line

```r
# scatter plot of duration and interval
plot(Duration ~ Interval, data = oldfaith,
     main = "Scatter Plot of Interval vs. Duration",
     xlab = "Duration (seconds)",
     ylab = "Interval (minutes)",
     pch = 16, col = "blue")
abline(lm2, col = "red")
```



(c) Compute 95% confidence intervals

```r
# new data frame with the given Duration
new_data <- data.frame(Duration = 250)

# Predict Interval with confidence and prediction intervals
prediction <- predict(lm2, newdata = new_data, interval = "prediction", level = 0.95)
prediction
```

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

this indicates that if an eruption lasts 250 seconds, the predicted waiting time until the next eruption is approximately 78.2 minutes.

Additionally, the 95% prediction interval for the waiting time is [66.35, 90.05] minutes. This means that, based on the regression model, we expect the actual waiting time to fall within this range 95% of the time for a new observation.

(d) R^2 Interpretation

Multiple R-squared is 0.8029 , this means the model provides a good fit to the data, meaning that eruption duration is a strong predictor of waiting time. However, there is still some unexplained variation, suggesting that additional factors might influence the waiting time.