STAT632 HW1

Brandon Keck

2025-02-07

Exercise 0

Link to github page: https://github.com/branistician87

Exercise 1

(a)

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

(b)

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Based on the p-value of **2e-16** the conclusion is that reject the null hypothesis in favor of the alternative. We have enough evidence to conclued that the slope is significantly different than 0.

(c)

p = 2 X P(T > |t|)

```
2 * pt(q = -2.699, df = 48)
```

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

The missing p-value for the intercept is $\boxed{0.009573193}$

(d)

$$T = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$$

$$T = \frac{2.2606 - 0}{0.0981} = \boxed{23.044}$$

(e)

```
# Calculate 95% Confidence Interval
n = 50
tcrit <- qt(0.975, df = n -2)
2.2606 * tcrit - 0.0981 # lower bound
```

```
## [1] 4.447141
```

```
2.2606 * tcrit + 0.0981 # upper bound
```

```
## [1] 4.643341
```

$\boxed{Thus\ our\ 95\ interval\ is\ (4.447141, 4.643341)}$

Since the 95% confidence interval does not include 0 this agrees with the hypothesis test to reject the H_0

Exercise 2

(a)

Show that the least squares estimate of the slope is given by:

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

Consider the linear regression model through the origin given by $Y_i = \beta x_i + e_i$

Where Y_i is the response variable and x_i is the predictor variable, and β is the slope coefficient we are wanting to estimate.

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

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

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

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

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

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

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

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

(b)

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

Recall the formula for the least squares estimate of β

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

Using Linearity of expectation:

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

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

Recall that $y_i = \beta x_i + \epsilon_i$

$E(y_i) = \beta x_i\ because\ E(\epsilon_i) = 0$

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

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

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

Therefore, we have shown that $E(\hat{\beta}) = \beta$

(c)

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

Recall that:

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

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

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

$$Var(\sum_{i=1}^n x_i y_i) = Var \sum_{i=1}^n (x_i y_i)$$

$$Var(x_i y_i) = x_i^2 Var(y_i)$$

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

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

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

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

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

Exercise 3

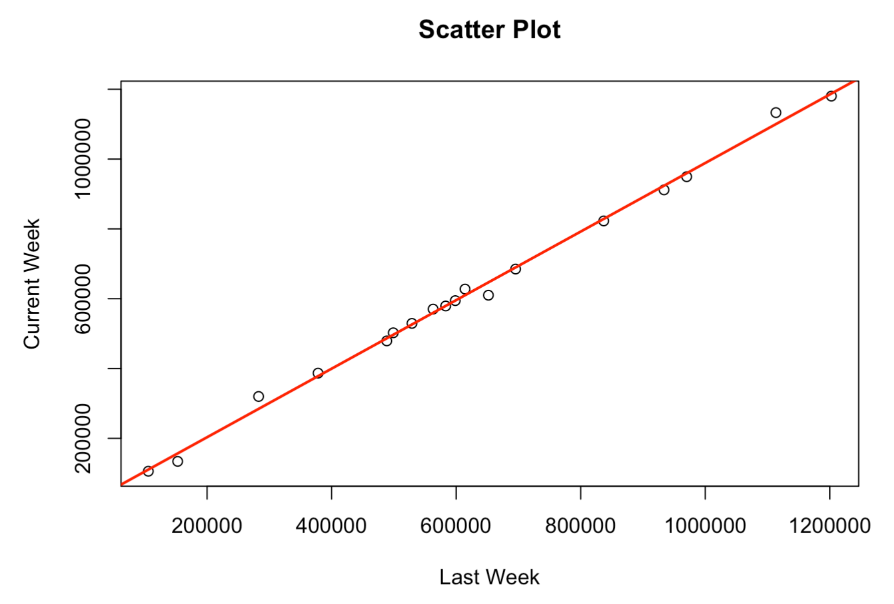
(a)

```
# Load in the dataset
library(readr)
playbill <- read_csv("~/Documents/EastBay/Sping 2025/Stat632Regression/playbill.csv")
```

```
## Rows: 18 Columns: 3
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Production
## dbl (2): CurrentWeek, LastWeek
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Then we fit a linear model, $Y = \beta_0 + \beta_1 + \epsilon$

```
# Fit the linear model
lm1 <- lm(CurrentWeek ~ LastWeek, data = playbill)

# Creates a scatter plot
plot(CurrentWeek ~ LastWeek, data = playbill,
     xlab = "Last Week",
     ylab = "Current Week",
     main = "Scatter Plot")

# Regression Line
abline(lm1, col = "red", lwd = 2)
```




```
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
```

(b)

```
# function to compute 95% confidence intervals
confint(lm1)
```

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

The confidence interval for β_1 : (9.514971e-01, 1.012666). The value 1 lies within this interval. Therefore, 1 is a plausible value for β_1 since it falls withing the 95% confidence interval.

(c)

```
df <- data.frame(LastWeek = 400000)

predict(lm1, newdata = df, interval = "prediction")
```

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

Based on the prediction interval 450,000 is **not** a feasible value for gross box office production with 400,000 in gross box office the previous week.

(d)

It seems reasonable to use the regression model to estimate the amount of gross box office results. However, the rule that next week's gross box office results will be equal to this week's box office results does not appear reasonable. There is such a large interval in our prediction that it might suggest variability within the data.

Exercise 4

(a)

```
# Load the library alr4
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

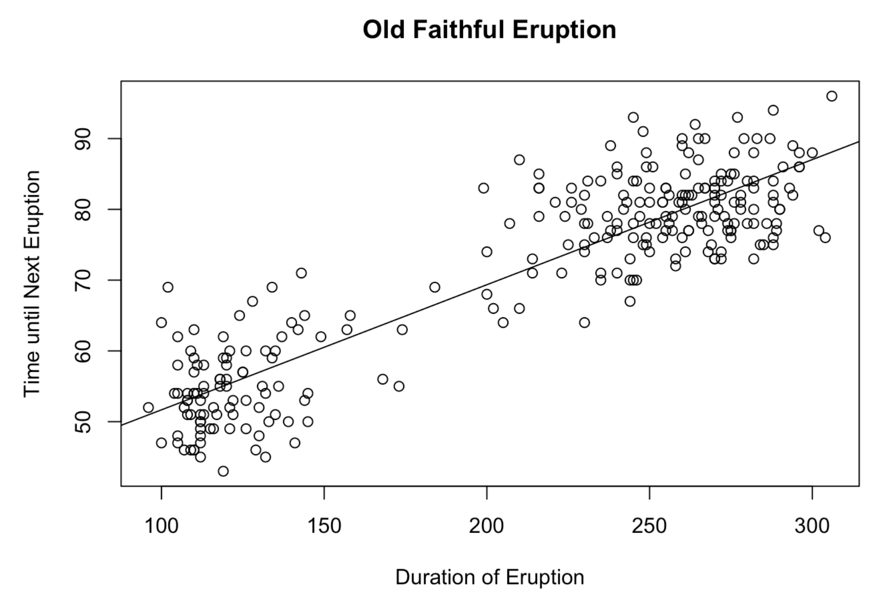
```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
# Create the linear model
lm2 <- lm(Interval ~ Duration, data = oldfaith)

# Summary statistics
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)

```
# Creates a scatter plot
plot(Interval ~ Duration, data = oldfaith,
     xlab = "Duration of Eruption",
     ylab = "Time until Next Eruption",
     main = "Old Faithful Eruption")

# Regression Line
abline(lm2)
```



(c)

```
# Create a 95% Prediction
df2 <- data.frame(Duration = 250)

predict(lm2, newdata = df2, interval = "prediction")
```

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

From the predicted value an eruption that lasts 250 seconds the predicted time until the next eruption is approximately 78.20354 minutes. While the actual time that someone would have to wait if they arrived at the end of a 250 second eruption is between 66.35 and 90.05 minutes.

(d)

From earlier our coefficient of determination was 0.8029 which means that 80.29% of the variability in "Interval" or (the time in minutes) until the next eruption is explained by "Duration" in seconds.