STAT 608 HW 1

Jack Cunningham (jgavc@tamu.edu)

9/6/24

1)

```r
playbill <- read.csv("playbill.csv")
```

```r
playbill_slr <- lm(data = playbill, formula = CurrentWeek ~ LastWeek)
slr_data <- summary(playbill_slr)
slr_data
```

```

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
```

a)

```r
B_1 <- slr_data$coefficients[2,"Estimate"]
B_1_se <- slr_data$coefficients[2,"Std. Error"]
n = length(playbill[,"CurrentWeek"])
lower <- B_1 - B_1_se*qt(1-0.05/2, df = n - 2)
upper <- B_1 + B_1_se*qt(1-0.05/2, df = n - 2)
c(lower,upper)
```

```
[1] 0.9514971 1.0126658
```

The 95% confidence interval for β_1 is (0.951, 1.013) The suggested value of 1 for β_1 is reasonable because it is within our confidence interval.

b)

We are testing the hypothesis: $H_0 : \beta_0 = 10000, H_1 : \beta_0 \neq 10000$. To do this we have the T statistic:

$T = \frac{\hat{\beta_0} - 10000}{SE(\beta_0)}$ and its absolute value needs to be greater than $t_{n-2}(\alpha/2)$ to reject the null hypothesis.

```r
t_quantile_rejection <- qt(1 - 0.05/2,df = n - 2)
B_0_hat <- slr_data$coefficients[1,"Estimate"]
B_0_hat_se <- slr_data$coefficients[1,"Std. Error"]
T_statistic_B_0 <- (B_0_hat - 10000)/B_0_hat_se
T_statistic_B_0
```

```
[1] -0.3217858
```

In this case our T statistic equals to -0.322 whose absolute value is far lower than 2.12.

```r
lower_B_0 <- B_0_hat - B_0_hat_se*qt(1-0.05/2, df = n - 2)
upper_B_0 <- B_0_hat + B_0_hat_se*qt(1 - 0.05/2, df = n -2)
c(lower_B_0,upper_B_0)
```

```
[1] -14244.33  27854.10
```

In fact the 95% confidence interval for β_0 is extremely wide, it is $(-1.4244327 \times 10^4, 2.7854099 \times 10^4)$. Therefore I wouldn't use the currently fitted model to predict current week sales when previous week sales are closer to zero. This is mainly due to the fact that our data only covers a range of previous week sales values from $200,000 to $1,200,000.

c)

To construct a 95% prediction interval for $x^\star = 400000$ we use the below formula:

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm t(\alpha/2, n-2) S \sqrt{\frac{1}{n} + 1 + \frac{(x^\star - \bar{x})}{SXX}}$$

We can do this manually or use the predict function.

```r
prediction <- B_0_hat + B_1*400000
S <- sqrt(1/(n-2)*sum((playbill[,"CurrentWeek"]-B_0_hat - B_1*playbill[, "LastWeek"])^2))
SXX <- sum((playbill[,"LastWeek"] - mean(playbill[,"LastWeek"]))^2)

upper <- prediction + qt(1 - 0.05/2,df = n -2)*S*sqrt(1/n + 1 + (400000 - mean(playbill[,
lower <- prediction - qt(1 - 0.05/2,df = n -2)*S*sqrt(1/n + 1 + (400000 - mean(playbill[,

c(lower,upper)
```

```
[1] 359832.8 439442.2
```

```r
current_predicted_value <- predict(playbill_slr,
                                   newdata = data.frame(LastWeek = c(400000)),
                                  interval = "prediction",
                                  confidence = 0.95)
current_predicted_value
```

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

The value \$450,000 is not a reasonable value to expect for the current week given a previous week of \$400,000. This is because it falls outside of our 95% prediction interval at $x^\star = 400000$: $(3.5983275 \times 10^5, 4.394422 \times 10^5)$.

d)

The model that the promoters of Broadway plays are using can be represented mathematically as $\hat{Y} = \beta_0 + \beta_1 x$ where $\beta_0 = 0$ and $\beta_1 = 1$. Thus to see if this is an appropriate rule we can test to see if the predictions generated by this simplified model are within the prediction interval for various values of last week data. Since we only have values of x from \$200,000 to \$1,200,000 in our dataset, I use evenly spaced values of x in that range and see if the predictions are within the prediction intervals generated from our simple linear regression model.

```
x_to_test <- seq(200000, 1200000,10000)
conf_intervals_for_x <- predict(playbill_slr,
                                newdata = data.frame(LastWeek = x_to_test),
                                interval = "prediction",
                                confidence = 0.95)
x_in_range <- x_to_test < conf_intervals_for_x[,"upr"]& x_to_test > conf_intervals_for_x[,

x_percent_in_range <- sum(x_in_range)/length(x_in_range)
x_percent_in_range
```

```
[1] 1
```

From this test we can see that the promoters prediction rule of using last week's gross box office result to predict the current week's box office result is a good rule of thumb (for last week gross box office amounts between $200,000 and $1,200,000). However it is worth noting that the promoters should expect the actual value of current week sales to not exactly match their prediction and rather randomly deviate around it.

2)

$Var(Y_i|X_i = x_i) = Var(\beta_0 + \beta_1 x_i + e_i) = Var(e_i)$.

We assume that Y is related to X through the linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$.

We assume that x_i is manually chosen by the practitioner or "known fixed constants".

3)

To create a linear regression model we need a method to estimate the parameters. The least squares criterion estimates the parameters by finding the values that minimize the sum of squared differences between our predicted values of the response variable and the actual values of the response variable in the data.

4)

i)

Since there is no intercept the design matrix is:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \dots \\ X_n \end{bmatrix}$$

ii)

$$y = X\beta + e_i$$

$$e_i^2 = (y - X\beta)^T(y - X\beta)$$

$$e_i^2 = y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta$$

$$e_i = y^Ty - 2y^TX\beta + \beta^TX^TX\beta$$

We minimize e_i^2. So we take the derivative of this with respect to β.

$$\frac{dRSS(\beta)}{d\beta} = -2y^TX + (X^TX + (X^TX)^T)\beta$$

We set this equal to zero:

$$-2y^TX + (2X^TX)\beta = 0$$

Recalling that X is a vector of length n consisting of all x values we have $y^TX = \sum_{i=1}^n y_ix_i$ and $X^TX = \sum_{i=1}^n x_i^2$. We have:

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

b)

i)

Using that $E[y_i] = \beta x_i$.

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

Using that $E(y_i) = \beta x_i$:

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

ii)

##Possibly Used Later##

The design matrix X is:

$$\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ ... & ... \\ 1 & X_n \end{bmatrix}$$

ii)

The principle of least squares has us minimizing the following function:

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n} e^2 = e^T e$$

We substitute in using the fact that $e = y - X\beta$.

$$(y - X\beta)^T(y - X\beta) = (y^T - \beta^T X^T)(y - X\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta$$

$$= y^T y - (X^T y)^T \beta - (X^T y)^T \beta + \beta^T (X^T X)\beta$$

$$= y^T y - 2(X^T y)^T \beta + \beta^T (X^T X)\beta$$

We now take the derivative relative to β and set it equal to zero.

$$\frac{RSS(\beta)}{d\beta} = -2(X^T y)^T + (X^T X + (X^T X)^T)\beta = 0$$

$$2(X^T X)\beta = 2(X^T y)^T$$

$$\beta = (X^T X)^{-1}(X^T y)^T$$

Where:

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

And thus:

$$(X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

We can also simplify the first value:

$$\frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2}$$

That is $\frac{1}{(n)SXX}$

And:

$$X^T y = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

So:

$$(X^T X)^{-1}(X^T y) = \frac{1}{n}(\frac{1}{SXX}$$

##End Possibly Used Later ##

6.

We know the below fact about covariance:

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Then:

$$Cov(aX, bY) = E(aXbY) - E(aX)E(bY) = abE(XY) - abE(X)E(Y) = ab(E(XY) - E(X)E(Y)) = ab(Cov(X, Y$$

7.

When a confidence interval for a particular $x^\star$ is created we are talking about the confidence interval of the statistic $E(Y|x^\star)$, the mean value of Y at a particular $x^\star$.

The confusion comes from thinking that the confidence interval of $E(Y|x^\star)$ is the same as a prediction interval for a single observation of Y at a given $x^\star$. The prediction interval will always be wider than the confidence interval as it also incorporates the variance of the random error e_i.