Homework 4

Jun Ryu, UID: 605574052

2023-04-28

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

Question 1

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

```
##         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
```

```
model1 <- lm(CurrentWeek ~ LastWeek, data = playbill)
summary(model1)
```

```
##
## 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)

```
confint(model1)
```

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

Using the LastWeek row (β_1), our 95% confidence interval is [0.9514971, 1.012666]. Thus, 1 seems to be a plausible value for β_1 as not only does it fit under our confidence interval, in the context of the problem, it indicates that gross box office results retain a similar value from one week to the next (offset by just β_0, a value that is not large in the scale of this problem), which makes sense.

b)

We have our null is $H_0 : \beta_0 = 10000$ and our alternative is $H_a : \beta_0 \neq 10000$. We will use the t-value as our test statistic. We had 6805 as our observed intercept value and 9929 as our standard error from the summary table.

```
t_stat <- (6805-10000)/9929
# we use n-2 as our degree of freedom
2 * pt(abs(t_stat), nrow(playbill)-2, lower.tail=FALSE)
```

```
## [1] 0.7517816
```

p-value for the intercept is 0.75 (> 0.05), thus we fail to reject the null hypothesis that our intercept is equal to 10000.

c)

```
predict(model1, data.frame(LastWeek = 400000), interval="prediction")
```

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

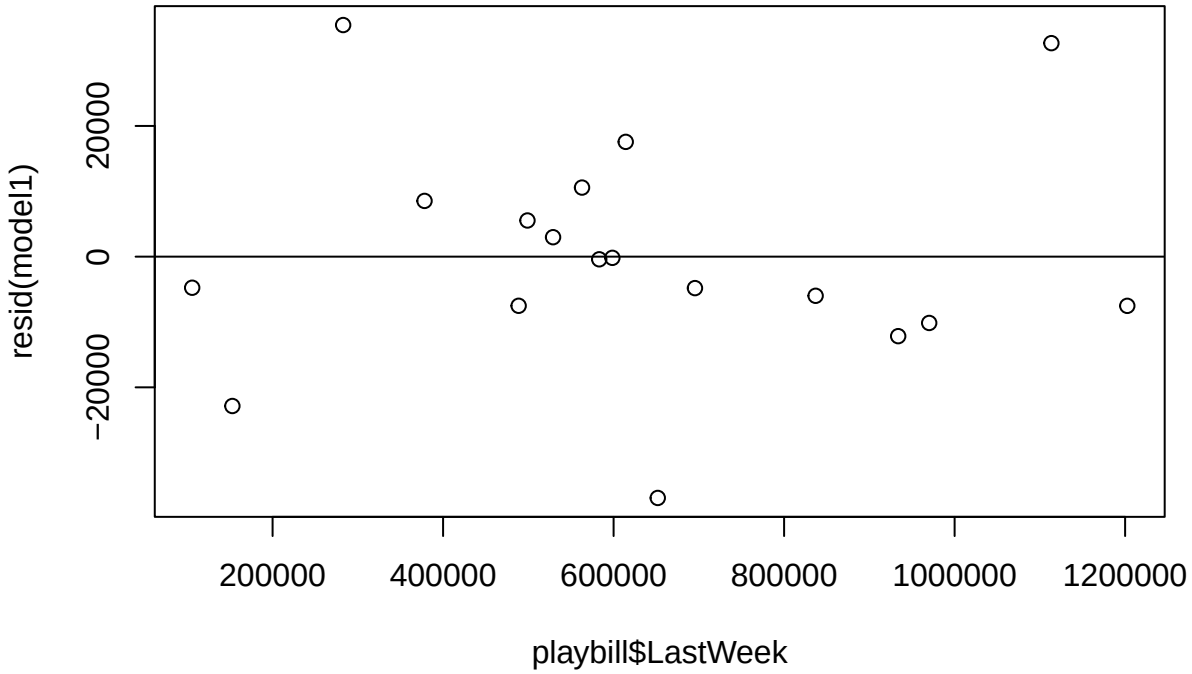
We use a prediction interval here as we are given one specific value of x and trying to predict y. Our prediction interval is [359832.8, 439442.2]. Looking at this interval, a gross box office result of $450000 is not feasible as it is quite above our upper bound for the interval.

d)

Similar to what was stated in part a), as our model gives us a β_1 value of 0.9821, I think the said rule is appropriate since the model's slope is very close to 1.

residual plot)

```
plot(playbill$LastWeek, resid(model1))
abline(0,0)
```



Looking at the residual plot, it seems like the residuals are independent of one another and there seems to be similar amounts of residuals above and below the midline of 0. Also, there appears to be constant variance as well (although harder to tell since there are such few data points). Thus, all this supports the claim that our model is a good linear fit.

Question 2

```
indicator <- read.table("indicators.txt", header = T)
head(indicator)
```

```
##   MetroArea PriceChange LoanPaymentsOverdue
## 1   Atlanta         1.2                4.55
## 2    Boston        -3.4                3.31
## 3   Chicago        -0.9                2.99
## 4    Dallas         0.8                4.26
## 5    Denver        -0.7                3.56
## 6   Detroit        -9.7                4.71
```

```
model2 <- lm(PriceChange ~ LoanPaymentsOverdue, data = indicator)
summary(model2)
```

```
##
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = indicator)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6541 -3.3419 -0.6944  2.5288  6.9163
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.5145     3.3240   1.358   0.1933
## LoanPaymentsOverdue  -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

a)

```
confint(model2)
```

```
##                          2.5 %     97.5 %
## (Intercept)         -2.532112 11.5611000
## LoanPaymentsOverdue -4.163454 -0.3335853
```

Using the LoanPaymentsOverdue row (β_1), our 95% confidence interval is [-4.163454, -0.3335853]. Since this interval is solely comprised of negative values, there surely is evidence of a negative linear association.

b)

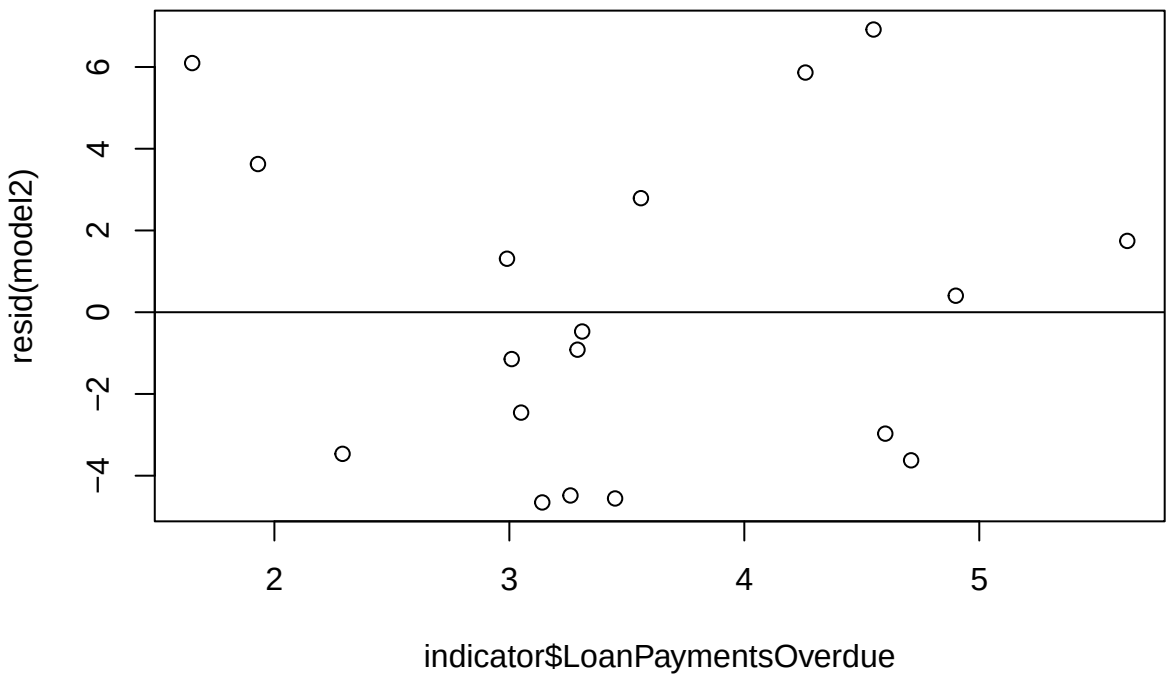
```
predict(model2, data.frame(LoanPaymentsOverdue = 4), interval = "confidence")
```

```
##         fit       lwr       upr
## 1 -4.479585 -6.648849 -2.310322
```

Here, we use a confidence interval instead because we want the mean value of Y instead. Our confidence interval is [-6.648849, -2.310322]. Looking at this interval, 0 is definitely not a feasible value as our confidence interval lies way below 0.

residual plot)

```
plot(indicator$LoanPaymentsOverdue, resid(model2))
abline(0,0)
```



Observing the residual plot, we see independence, normality, and constant variance as similar to Question 1. Thus, this model seems to be a good linear fit.

Question 3

a)

```
# values from the textbook
beta0 <- 0.6417099
standard_error <- 0.1222707

# 95% confidence interval
c(beta0 - 1.96*standard_error, beta0 + 1.96*standard_error)
```

```
## [1] 0.4020593 0.8813605
```

b)

We use t-statistic as we also did in 1b). We have our null is $H_0 : \beta_1 = 0.01$ and our alternative is $H_a : \beta_1 \neq 0.01$.

```
t_stat_2 <- (0.01-0.0112916)/0.0008184
2 * pt(abs(t_stat_2), 28, lower.tail=FALSE) # we use 28 as our degree of freedom
```

```
## [1] 0.1257517
```

p-value is 0.12575 (> 0.05), thus we fail to reject the null hypothesis that the average processing time for an additional invoice is 0.01 hours.

c)

Prediction interval is given by: $\hat{y} \pm t^* SE(\hat{y})$, where $SE(\hat{y}) = \sqrt{\sigma^2(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX})}$. In this case, we have 130 invoices is our mean ($\bar{x}$), so our formula reduces down to $SE(\hat{y}) = \sqrt{\sigma^2(1 + \frac{1}{n})}$ (since $x^* = \bar{x}$). So, here n is our sample size of 30, and we can figure our σ^2 (MSE) by calculating $\frac{RSE^2 * 28}{30}$ since $RSE = \sqrt{\frac{\sum(x - \hat{x})^2}{n-2}}$.

```
# from the textbook
beta0 <- 0.6417099
beta1 <- 0.0112916
rse <- 0.3298
processing_time <- beta0 + beta1 * 130
sigma_squared <- rse^2*28/30
error <- qt(0.975, 28) * sqrt(sigma_squared) * sqrt(1 + 1/30)
c(processing_time - error, processing_time + error) # prediction interval
```

```
## [1] 1.446172 2.773064
```

```
processing_time # point estimate
```

```
## [1] 2.109618
```

Question 4

D is the correct option. Observing both graphs, we see that RSS is smaller for model 1 because the observed values are closer to the regression line in model 1 than model 2. However, for SSreg, the value is greater for model 1 since the regression line stretches farther away from the mean of the sample (in the span of the graph shown) with its steep slope.