Stats101A, Spring 2023 - Homework 4

Luke Villanueva - 206039397

04/27/23

```r
# libraries used
library(tidyverse)
```

Problem 1a

```r
# load in playbill.csv
playbill <- read.csv(file = "playbill.csv", header = TRUE)

glimpse(playbill)
```

```
## Rows: 18
## Columns: 3
## $ Production  <chr> "42nd Street", "Avenue Q", "Beauty and Beast", "Bombay Dre~
## $ CurrentWeek <int> 684966, 502367, 594474, 529298, 570254, 319959, 579126, 13~
## $ LastWeek    <int> 695437, 498969, 598576, 528994, 562964, 282778, 583177, 15~
```

```r
# fit linear model
lin <- lm(CurrentWeek ~ LastWeek, data = playbill)
```

a.

```r
confint(lin, level = 0.95)
```

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

The model could most definitely have a slope of 1 because the lower bound and upper bound for the slope for the linear model is about (0.95,1.01), which places 1 close to either bound.

b.

```
# t stat for intercept = (intercept_estim - 10000) / (se(intercept_estim))
tstat <- (lin$coefficients[1] - 10000) / summary(lin)$coefficients[1,2]
names(tstat) <- NULL

# tstat is left of mean, so lower tail test
tstat
```

```
## [1] -0.3217858
```

```
# get p value from tstat
pt(tstat, df = nrow(playbill) - 1)
```

```
## [1] 0.3757689
```

Regardless of the alpha value (aka the significance level), a p-value of 0.37 says that there is about a 37% chance for 10000 to be the intercept of a sample taken from the list of shows. Therefore, based on usual significance levels, we do **not** have enough information to reject the null hypothesis. So, we fail to reject that $B_0 = 10000$.

c.

```
# estimate using model
predict(lin, data.frame(LastWeek = 400000), interval = "c", level = 0.95)
```

```
##        fit      lwr    upr
## 1 399637.5 388361.9 410913
```

```
# prediction
predict(lin, data.frame(LastWeek = 400000), interval = "p", level = 0.95)
```

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

Based on the estimate and the prediction interval, $450,000 would not be a sensible value for the Current Week's gross office result for the production. This is because the direct estimate based on the model shows the value decreasing. In addition, the prediction interval's upper bound only goes up to $439,442, which means that there is a 95% chance that the true future value would land below or equal to that upper bound. Having a 5% chance of landing above $439k is not sensible.

d.

```
# use r val to get validity of linearity
cor(playbill$LastWeek, playbill$CurrentWeek)
```

```
## [1] 0.998278
```

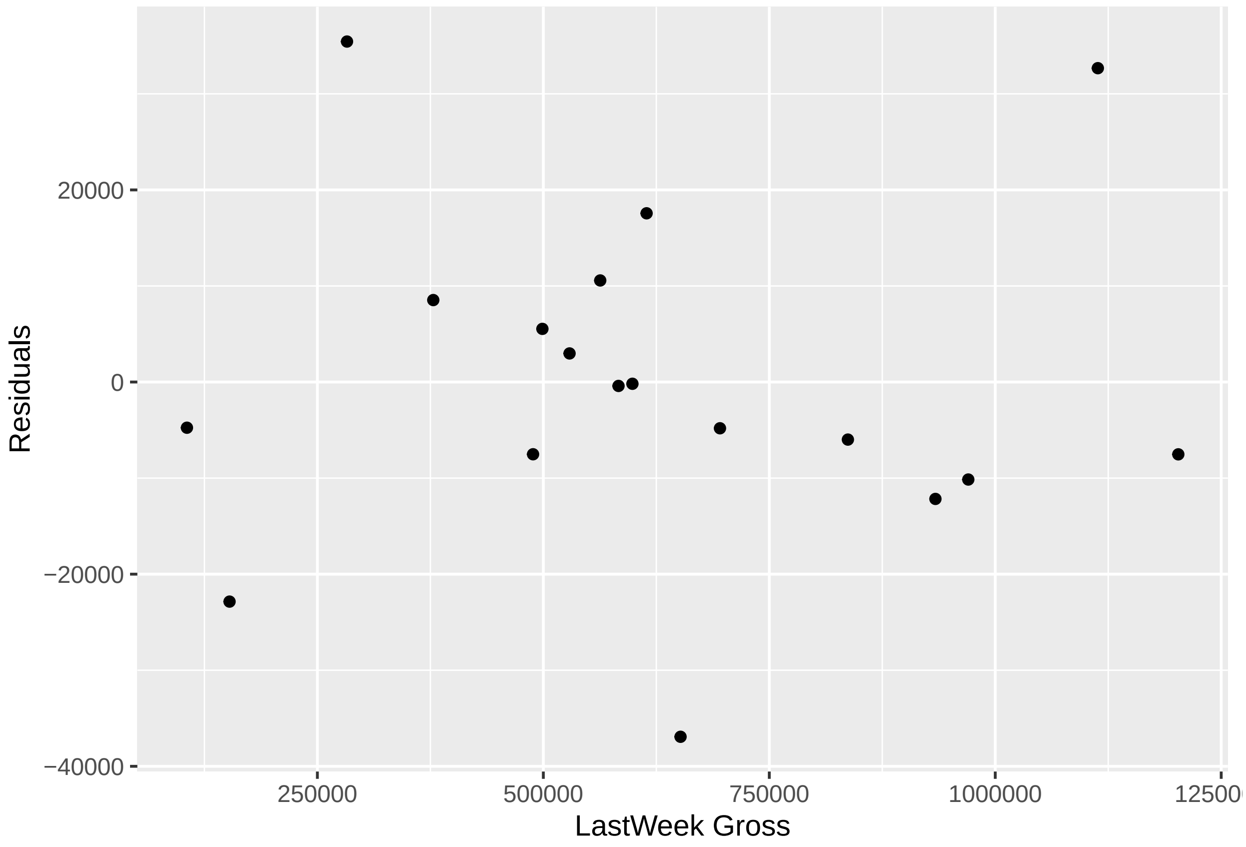
```r
confint(lin, level = 0.95)
```

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

Because there is a strong linear relationship between last week and current week gross results and there is a high likelihood that the true slope of the gross results is close to 1, the prediction rule to assume that next week's gross results would be the same as last week's is decently sensible.

1b

```r
# plot residuals against LastWeek
ggplot(playbill) + geom_point(mapping = aes(x = LastWeek, y = lin$residuals)) + labs(x = "LastWeek Gros
```



The plot suggests that the variance is mostly random. However, there is a slight clumping towards the middle, but that might just be because the sample number is small.

Problem 2a

```r
# load dataset
indicators <- read.delim(file = "indicators.txt", header = TRUE, sep = "\t")

glimpse(indicators)
```

```
## Rows: 18
## Columns: 3
## $ MetroArea           <chr> "Atlanta", "Boston", "Chicago", "Dallas", "Denver"~
## $ PriceChange         <dbl> 1.2, -3.4, -0.9, 0.8, -0.7, -9.7, -6.1, -4.8, -6.4~
## $ LoanPaymentsOverdue <dbl> 4.55, 3.31, 2.99, 4.26, 3.56, 4.71, 4.90, 3.05, 5.~
```

```r
# linear model
lin <- lm(PriceChange ~ LoanPaymentsOverdue, indicators)

# confidence interval
confint(lin, level = 0.95)
```

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

Based on the 95% confidence interval on the slope, the bounds of the slope are both negative, so there is a very likely chance the true slope of the data is also negative.

```r
# using the model for direct estimate
predict(lin, list(LoanPaymentsOverdue = 4))
```

```
##         1
## -4.479585
```

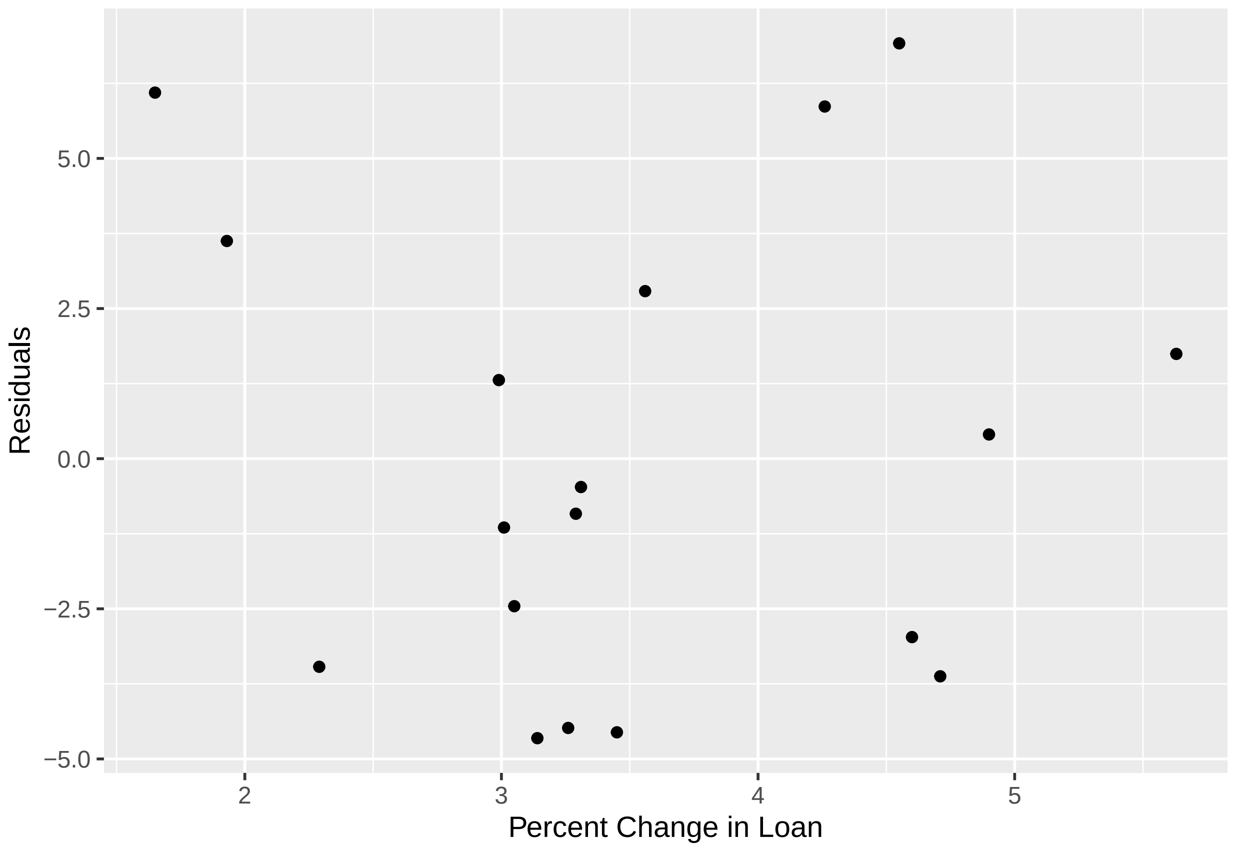
```r
# using estimation
predict(lin, list(LoanPaymentsOverdue = 4), interval = "c", level = 0.95)
```

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

No, 0% for the expected value of price change if overdue loan payment amount is 4% is not a sensible conclusion. This is because the interval bounds are only negative. And with the interval being a 95% confidence level, this implies that the price change being 0% when overdue loan amount is 4% is very unlikely.

2b

```r
ggplot() + geom_point(aes(x = indicators$LoanPaymentsOverdue, y = lin$residuals)) + labs(x = "Percent C
```



The residual plot is random and scattered, which implies that the linear model is a decent fit for the data. However, there is a slight trending downward in the latter half of the plot, but that may disappear if there were more observations in the sample.

Problem 3

a.

```
# load dataset
invoices <- read.delim(file = "invoices.txt", header = TRUE, sep = "\t")

glimpse(invoices)
```

```
## Rows: 30
## Columns: 3
## $ Day      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ Invoices <int> 149, 60, 188, 23, 201, 58, 77, 222, 181, 30, 110, 83, 60, 25,~
## $ Time     <dbl> 2.1, 1.8, 2.3, 0.8, 2.7, 1.0, 1.7, 3.1, 2.8, 1.0, 1.5, 1.2, 0~
```

```
# linear model of time to invoices
lin <- lm(Time ~ Invoices, data = invoices)

lin
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Coefficients:
## (Intercept)     Invoices
##     0.64171      0.01129
```

```
# get summary of linear model
summary(lin)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6417099  0.1222707   5.248 1.41e-05 ***
## Invoices    0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

```
# get confint of intercept
confint(lin, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.391249620 0.89217014
## Invoices    0.009615224 0.01296806
```

The 95% interval of the intercept is (0.39,0.89).

b.

```
# tstat = (slope estimate - 0.01) / SE(slope estimate)
tstat <- (lin$coefficients[2] - 0.01) / summary(lin)$coefficients[1,2]
names(tstat) <- NULL

tstat
```

```
## [1] 0.0105638
```

```r
# tstat is positive, so test right tail
pt(tstat, df = nrow(invoices) - 1, lower.tail = FALSE)
```

```
## [1] 0.4958219
```

Since the p-value is almost 50%, this means that having a slope of 0.01 can occur by chance. Therefore, we fail to reject the null hypothesis.

c.

```r
# point estimate and prediction interval
predict(lin, list(Invoices = 130), interval = "p", level = 0.95)
```

```
##        fit      lwr    upr
## 1 2.109624 1.422947 2.7963
```

The point estimate is 2.1096 hours. The bounds of the 95% prediction interval are (1.4229, 2.7963).

Problem 4

RSS1 < RSS2

SSReg1 > SSReg2

a.

False, because visually, the variance in residuals is greater for model 2.

b.

False, because visually, the variance in the regression for model 1 is expected to be higher than model 2's since the points are closer to the linear model.

c.

False, because visually, the residuals are greater in variance in model 2. And, the data are closer to the model in model 1.

d.

True, because visually, the residuals are smaller for model 1, so RSS for model 1 is less than RSS model 2. In addition, the data points are closer in model 1 than model 2, so the fit is better, implying SSreg for model 1 is higher than model 2.