Homework 3

Christopher Hainzl

2022-11-10

Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

Question 1

```
baseball <- read.table(file=
"http://jse.amstat.org/datasets/baseball.dat.txt",
header = F, col.names=c("salary", "batting.avg", "OBP", "runs", "hits",
"doubles", "triples", "homeruns", "RBI", "walks", "strike.outs",
"stolen. bases", "errors", "free.agency.elig", "free.agent.91",
"arb.elig", "arb.91", "name"))
head(baseball)
```

```
##   salary batting.avg   OBP runs hits doubles triples homeruns RBI walks
## 1   3300       0.272 0.302   69  153      21       4       31 104    22
## 2   2600       0.269 0.335   58  111      17       2       18  66    39
## 3   2500       0.249 0.337   54  115      15       1       17  73    63
## 4   2475       0.260 0.292   59  128      22       7       12  50    23
## 5   2313       0.273 0.346   87  169      28       5        8  58    70
```

```
## 6   2175       0.291 0.379  104  170      32       2       26 100    87
##   strike.outs stolen..bases errors free.agency.elig free.agent.91 arb.elig
## 1          80             4      3                1             0        0
## 2          69             0      3                1             1        0
## 3         116             6      5                1             0        0
## 4          64            21     21                0             0        1
## 5          53             3      8                0             0        1
## 6          89            22      4                1             0        0
##   arb.91              name
## 1      0 Andre Dawson
## 2      0 Steve Buchele
## 3      0 Kal Daniels
## 4      0 Shawon Dunston
## 5      0 Mark Grace
## 6      0 Ryne Sandberg
```

1a

```
colSums(is.na(baseball))
```

```
##           salary      batting.avg              OBP             runs
##                0                0                0                0
##             hits          doubles          triples         homeruns
##                0                0                0                0
##              RBI            walks      strike.outs    stolen..bases
##                0                0                0                0
##           errors free.agency.elig    free.agent.91         arb.elig
##                0                0                0                0
##           arb.91             name
##                0                0
```

- No, there isn't any missing data.

```
# Determine the variable types of the columns
sapply(baseball, class)
```

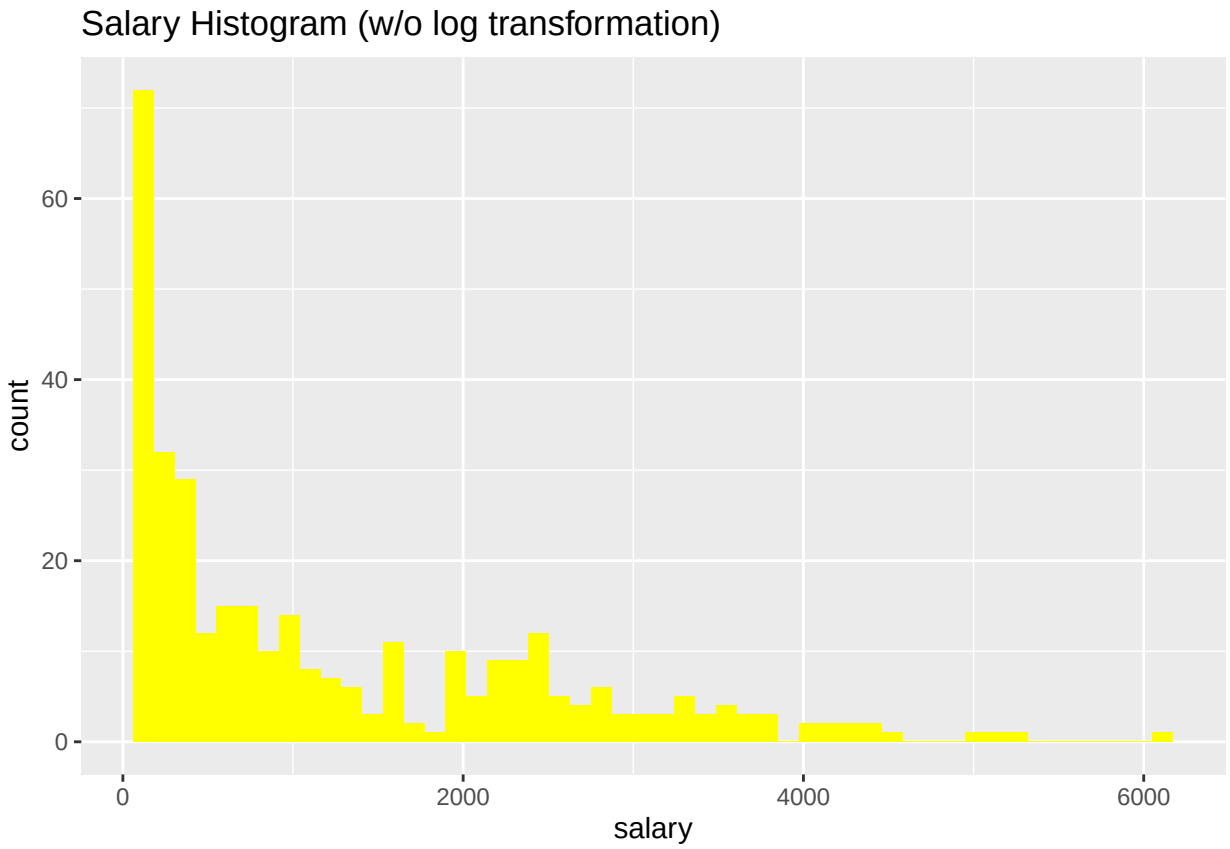
```
##           salary      batting.avg              OBP             runs
##        "integer"        "numeric"        "numeric"        "integer"
##             hits          doubles          triples         homeruns
##        "integer"        "integer"        "integer"        "integer"
##              RBI            walks      strike.outs    stolen..bases
##        "integer"        "integer"        "integer"        "integer"
##           errors free.agency.elig    free.agent.91         arb.elig
##        "integer"        "integer"        "integer"        "integer"
##           arb.91             name
##        "integer"      "character"
```

- 4 predictors are categorical (free.agency.elig, free.agent.91, arb.elig, arb.91), and 10 predictors are continuous, integer counts.

1b

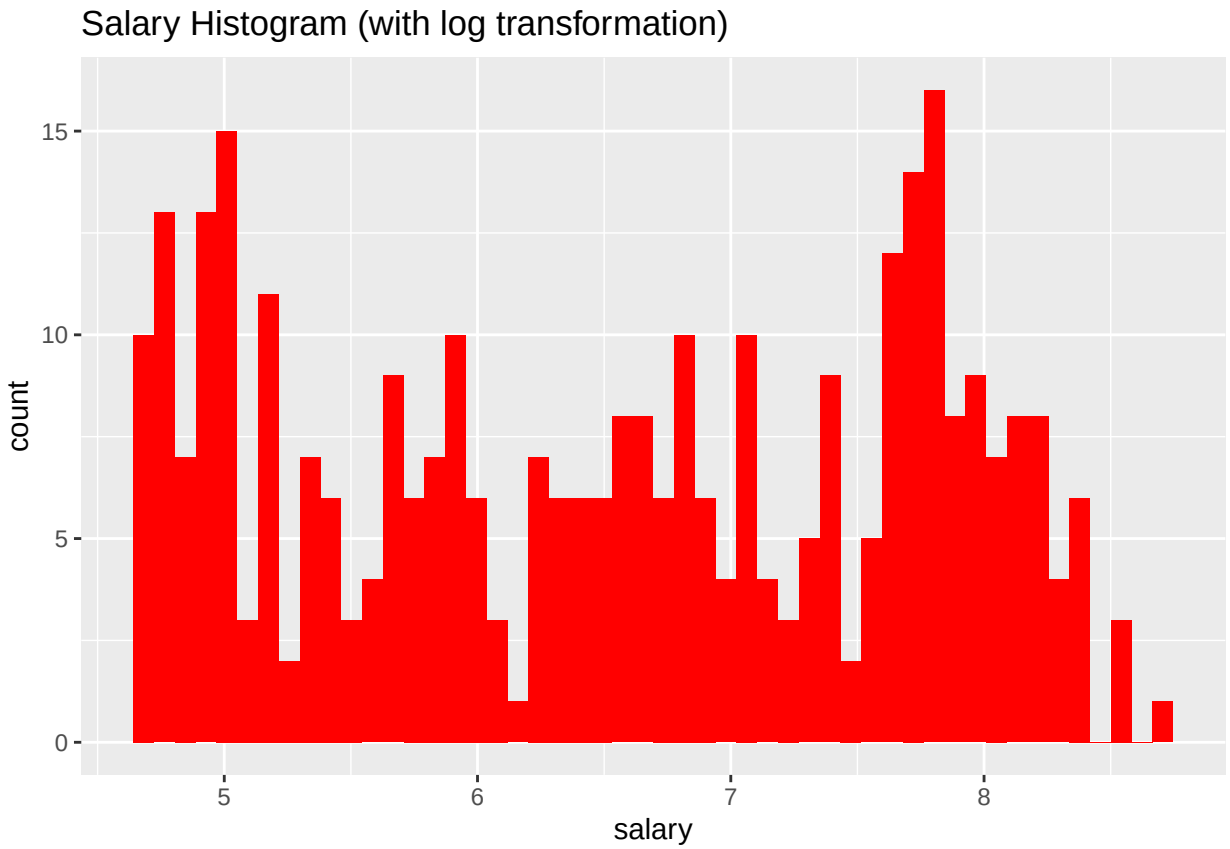
```
# Histogram of salaries without log transformation
ggplot(data = baseball, mapping = aes(x = salary)) +
  geom_histogram(bins = 50, fill = "yellow") +
```

```
  labs(title = "Salary Histogram (w/o log transformation)")
```




```
# perform log transformation on salary
baseball[1] <- log(baseball[1])
```

```
# Histogram of salaries with log transformation
ggplot(data = baseball, mapping = aes(x = salary)) +
  geom_histogram(bins = 50, fill = "red") +
  labs(title = "Salary Histogram (with log transformation)")
```



- The histogram without the log transformation appears to be right-skewed, while the histogram with the log transformation does not appear to have any sort of skewness.

Question 2

2a

```
set.seed(123)
# Sample Indexes
Index = sample(1:nrow(baseball), size = 0.7*nrow(baseball))
# Splitting Data
TrainData = baseball[Index,]
dim(TrainData)
```

```
## [1] 235  18
```

```
TestData = baseball[-Index,]
dim(TestData)
```

```
## [1] 102  18
```

```
head(TrainData)
```

```
##       salary batting.avg   OBP runs hits doubles triples homeruns RBI walks
## 179 4.976734       0.457 0.486    6   16       4       2        0   7     2
## 14  4.744932       0.261 0.370    1    6       0       0        0   2     4
## 195 7.696213       0.279 0.391   94  131      26       2       23  78    90
## 306 6.263398       0.261 0.321   16   53       4       0        1  17    16
```

```
## 118 6.733402       0.287 0.349   59  123      16       1        3  38    37
## 299 7.946264       0.201 0.330   62   97      22       0       22  75    93
##     strike.outs stolen..bases errors free.agency.elig free.agent.91 arb.elig
## 179           2             0      2                0             0        0
## 14            3             0      0                0             0        0
## 195          45             4      4                1             0        0
## 306          28             0      6                1             0        0
## 118          32            12     20                1             0        0
## 299         116             2      4                1             0        0
##     arb.91              name
## 179      0 Scott Cooper
## 14       0 Rey Sanchez
## 195      0 Lou Whitaker
## 306      0 Jamie Quirk
## 118      0 Lenny Harris
## 299      0 Mark McGwire
```

```
head(TestData)
```

```
##      salary batting.avg   OBP runs hits doubles triples homeruns RBI walks
## 2  7.863267       0.269 0.335   58  111      17       2       18  66    39
## 3  7.824046       0.249 0.337   54  115      15       1       17  73    63
## 12 4.941642       0.222 0.307   21   45       9       0        6  22    19
## 15 7.863267       0.300 0.368   69  141      22       3       19  75    53
## 18 6.897705       0.290 0.349   59  141      30       2       16  64    42
## 19 6.829794       0.246 0.323   22   81      14       0        6  26    22
##    strike.outs stolen..bases errors free.agency.elig free.agent.91 arb.elig
## 2           69             0      3                1             1        0
## 3          116             6      5                1             0        0
## 12          56             3      3                0             0        0
## 15          64            31      7                1             0        0
## 18         102            14      6                1             0        0
## 19          26             2      5                1             0        0
##    arb.91              name
## 2       0 Steve Buchele
## 3       0 Kal Daniels
## 12      0 Rick Wilkins
## 15      0 Ivan Calderon
## 18      0 Larry Walker
## 19      0 Gary Carter
```

2b and 2c

```
# 2b
fit.train <- lm(salary ~ batting.avg + OBP +
runs + hits + doubles + triples + homeruns + RBI + walks + strike.outs + stolen..bases + errors,
data = TrainData)

# 2c
summary(fit.train)
```

```
##
## Call:
## lm(formula = salary ~ batting.avg + OBP + runs + hits + doubles +
##     triples + homeruns + RBI + walks + strike.outs + stolen..bases +
```

```
##     errors, data = TrainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90750 -0.60566  0.06554  0.58097  2.08453
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.561171   0.462268  14.193  < 2e-16 ***
## batting.avg   -6.488650   3.835905  -1.692 0.092134 .
## OBP            1.229445   3.431739   0.358 0.720491
## runs          -0.009282   0.008558  -1.085 0.279307
## hits           0.017686   0.004833   3.659 0.000316 ***
## doubles       -0.009300   0.012497  -0.744 0.457523
## triples        0.001948   0.030700   0.063 0.949463
## homeruns       0.044061   0.017837   2.470 0.014259 *
## RBI            0.004447   0.007178   0.620 0.536199
## walks          0.012107   0.006850   1.768 0.078508 .
## strike.outs   -0.013040   0.003009  -4.334 2.22e-05 ***
## stolen..bases -0.001987   0.007026  -0.283 0.777561
## errors        -0.010272   0.010834  -0.948 0.344099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8417 on 222 degrees of freedom
## Multiple R-squared:  0.5152, Adjusted R-squared:  0.489
## F-statistic: 19.66 on 12 and 222 DF,  p-value: < 2.2e-16
```

2d

- When working with a 5% significance level and looking at the individual p-values, salary only appears to have a relationship with hits, home runs, and strikeouts (since the p-values for those three predictors are less than 0.05).

2e

```
# Use the model for the training data to predict the
# test data, and output mean squared error
mean((TestData$salary - predict.lm(fit.train, TestData))^2)
```

```
## [1] 0.6735442
```