Inclass

Hyun Suk (Max) Ryoo (hr2ee)

10/5/2021

Prework

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## 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(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.4.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

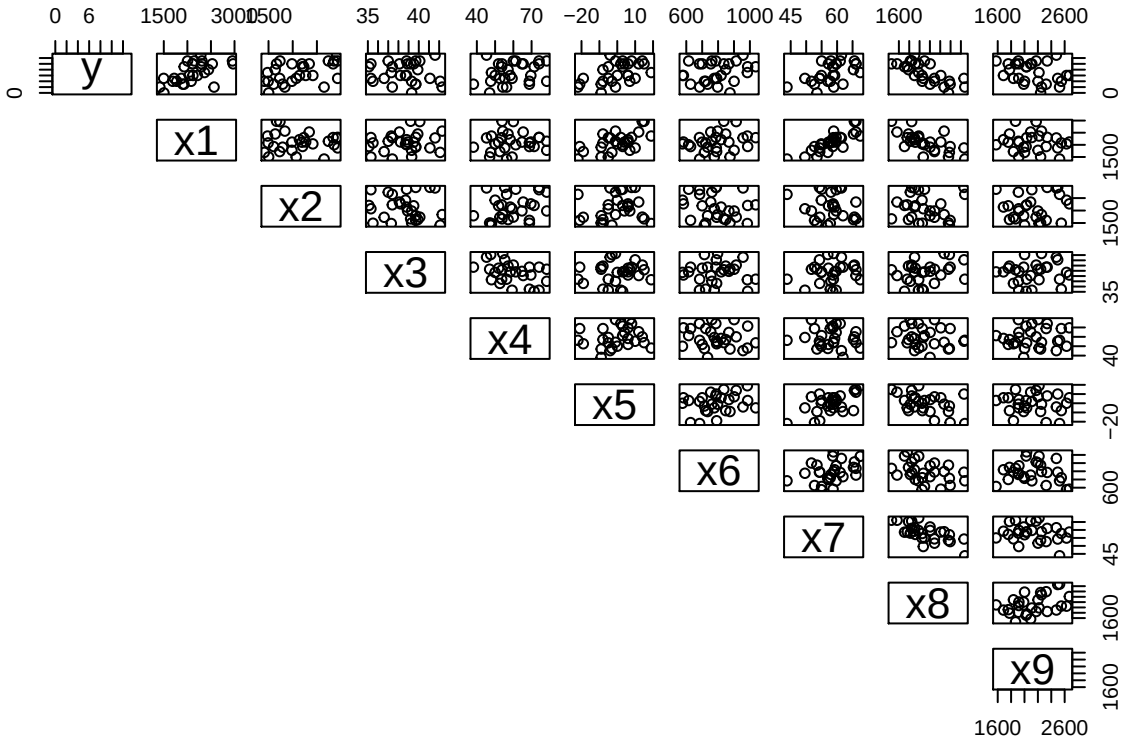
```
## -- Conflicts -------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x MASS::select()  masks dplyr::select()
```

```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/Module6")
data = read.csv("nfl.txt", sep="\t")
head(data)
```

```
##    y   x1   x2   x3   x4 x5  x6   x7   x8   x9
## 1 10 2113 1985 38.9 64.7  4 868 59.7 2205 1917
## 2 11 2003 2855 38.8 61.3  3 615 55.0 2096 1575
## 3 11 2957 1737 40.1 60.0 14 914 65.6 1847 2175
## 4 13 2285 2905 41.6 45.3 -4 957 61.4 1903 2476
## 5 10 2971 1666 39.2 53.8 15 836 66.1 1457 1866
## 6 11 2309 2927 39.7 74.1  8 786 61.0 1848 2339
```

1) Create a scatterplot matrix and find the correlation between all pairs of variables for this data set. Answer the following questions based on the output:

```
pairs(data, lower.panel = NULL)
```




```
cor(data)
```

```
##              y          x1          x2           x3          x4          x5
## y   1.00000000  0.59323604  0.48273470 -0.080812472  0.25847477  0.51320624
## x1  0.59323604  1.00000000 -0.03674736  0.212471227  0.07029904  0.59998017
## x2  0.48273470 -0.03674736  1.00000000 -0.068815157  0.30151583  0.13499515
## x3 -0.08081247  0.21247123 -0.06881516  1.000000000 -0.41309561  0.11509807
```

```
## x4  0.25847477  0.07029904  0.30151583 -0.413095614  1.00000000  0.14902865
## x5  0.51320624  0.59998017  0.13499515  0.115098074  0.14902865  1.00000000
## x6  0.22403447  0.25297272 -0.19283713 -0.003115748 -0.12818435  0.25891534
## x7  0.54534104  0.83728269 -0.19691540  0.162511469 -0.10100316  0.60956318
## x8 -0.73802730 -0.65854627 -0.05104783  0.290438108 -0.16402353 -0.47004608
## x9 -0.30374811 -0.11055739  0.14598149  0.088195595  0.05913611 -0.09028906
##              x6         x7          x8          x9
## y   0.224034472  0.5453410 -0.73802730 -0.30374811
## x1  0.252972716  0.8372827 -0.65854627 -0.11055739
## x2 -0.192837129 -0.1969154 -0.05104783  0.14598149
## x3 -0.003115748  0.1625115  0.29043811  0.08819559
## x4 -0.128184348 -0.1010032 -0.16402353  0.05913611
## x5  0.258915336  0.6095632 -0.47004608 -0.09028906
## x6  1.000000000  0.3670779 -0.35249327 -0.17275608
## x7  0.367077900  1.0000000 -0.68504573 -0.20331784
## x8 -0.352493271 -0.6850457  1.00000000  0.41746519
## x9 -0.172756078 -0.2033178  0.41746519  1.00000000
```

- (A) Which predictors appear to be linearly related to the number of wins? Which predictors do not appear to have a linear relationhsip with the number of wins?
  - Maybe x1, x5, x7, x8 have a linear relationship.
- (B) Do you notice if any of the predictors are highly correlated with one another? If so, which ones?
  - (X1, X7), (X7, X8)
- (C) What predictors would you first consider to use in a multiplie linear regression? Briefly explain your choices.
  - I would start with x1, x8

2) Regardless of your answer to the previous question, fit a multiple regression model for the number of games won against the following three predictors: the team's passing yardage, the percentage of rushing plays, and the opponents' yards rushing. Write the estimated regression equation.

```
result<-lm(y~x2+x7+x8, data=data)
summary(result)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
```

```
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

$y = 0.003598x2 + 0.193960x7 - 0.004815x8 - 1.808372$

3) Interpret the estimated coefficient for the predictor x7 in context.

As the percent rushing increases by one, the games won increases by 0.193960

4) A team with x2 = 2000 yards, x7 = 48 percent, and x8 = 2350 yards would like to estimate the number of games it would win. Also provide a relevant interval for this estimate with 95% confidence.

```
newdata<-data.frame(x2=2000, x7=48, x8=2350)
predict(result, newdata, level=0.95,interval="prediction")
```

```
##        fit        lwr      upr
## 1 3.381448 -0.5163727 7.279268
```

5) Using the output for the multiple linear regression model from part 2, answer the

following question from a client: "Is this regression model useful in predicting the number of wins during the 1976 season?" Be sure to write the null and alternative hypotheses, state the value of the test statistic, state the p-value, and state a relevant conclusion. What is the critical value associated with this hypothesis test? Perform the test at 0.05 significance level.

$H_0 : \beta_1 = ... = \beta_k = 0$
$H_A : \beta_i \neq 0$ For at least one

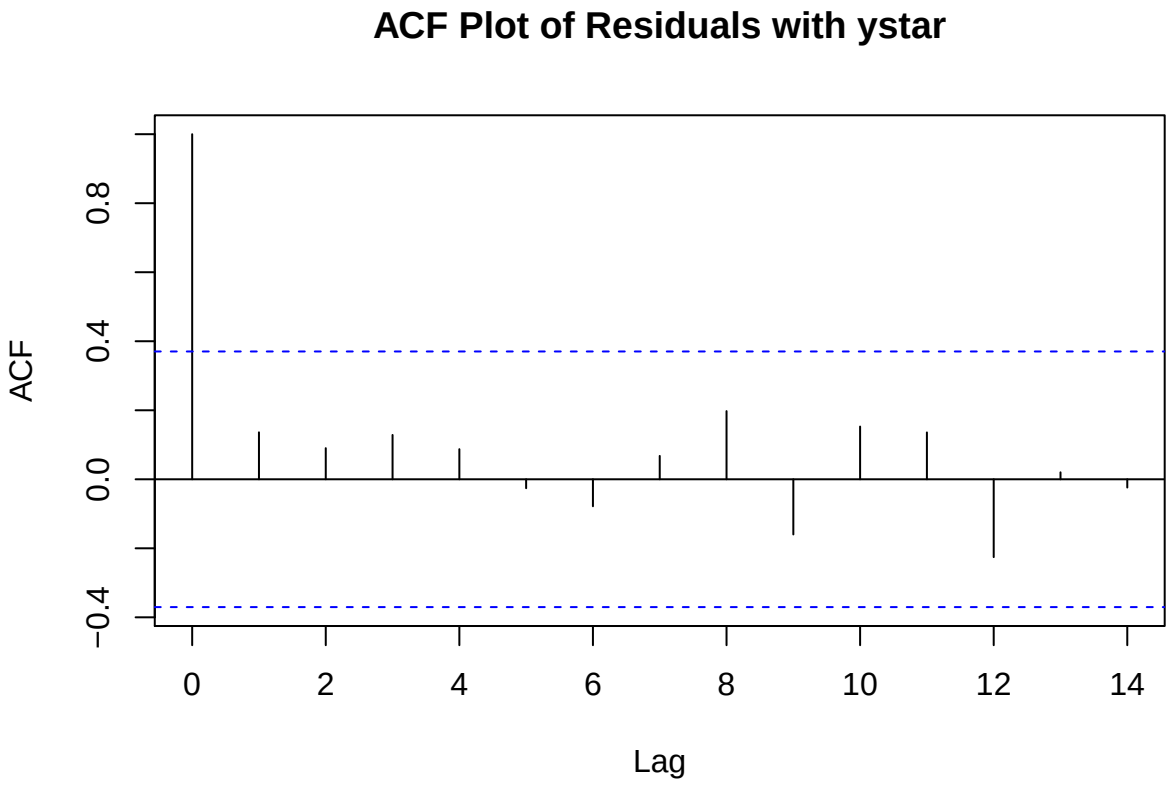
6) Report the value of the t statistic for the predictor x7. What is the relevant conclusion

from this t statistic? Also report the critical value for this hypothesis test. Perform the test at 0.05 significance level.

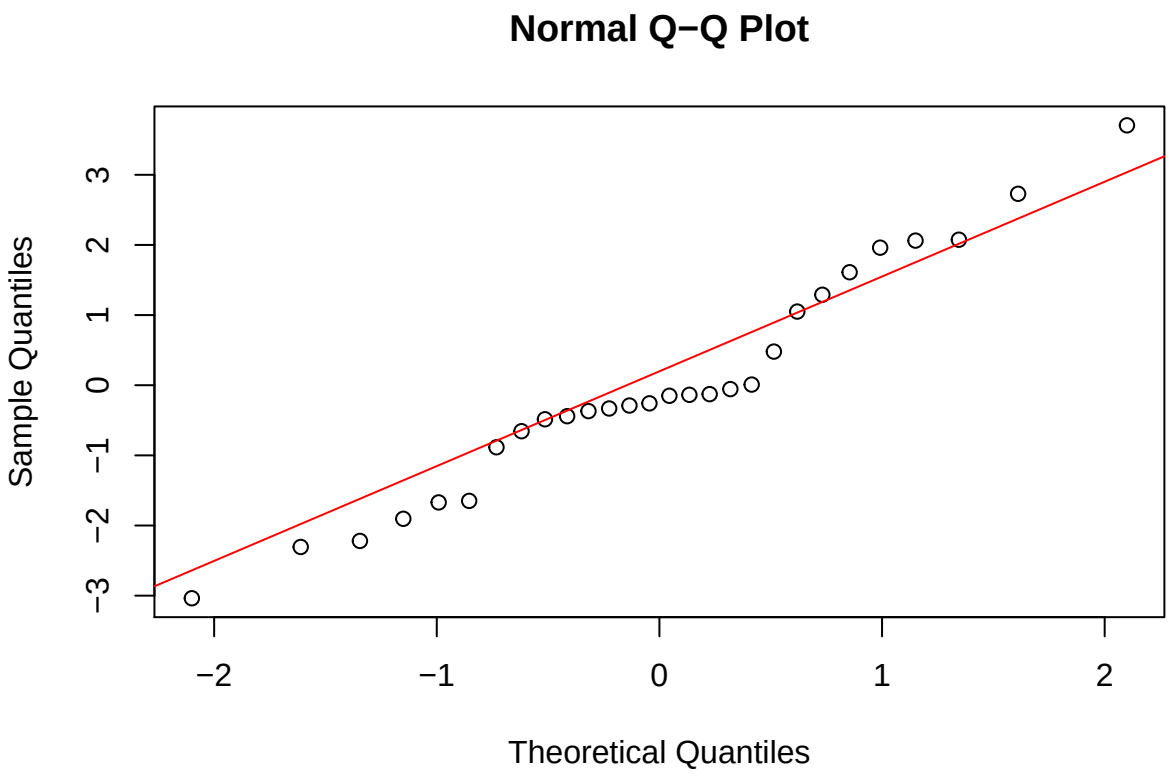
7) Check the regression assumptions by creating a residual plot, an ACF plot of the

residuals, and a QQ plot of the residuals. Comment on these plots.

```r
acf(result$residuals, main="ACF Plot of Residuals with ystar")
```




```r
{
qqnorm(result$residuals)
qqline(result$residuals, col="red")
}
```



8) Consider adding another predictor, x1, the team's rushing yards for the season, to

the model. Interpret the results of the t test for the coefficient of this predictor. A classmate says: "Since the result of the t test is insignificant, the team's rushing yards for the season is not linearly related to the number of wins." Do you agree with your classmate's statement?

```
result<-lm(y~x1+x2+x7+x8, data=data)
summary(result)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x7 + x8, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7456 -0.6801 -0.1941  1.1033  3.7580
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.8791718  8.1955007  -0.107  0.91550
## x1           0.0009045  0.0016489   0.549  0.58862
## x2           0.0035214  0.0007191   4.897 6.02e-05 ***
## x7           0.1437590  0.1280424   1.123  0.27313
## x8          -0.0046994  0.0013131  -3.579  0.00159 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 23 degrees of freedom
```

```
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7524
## F-statistic: 21.51 on 4 and 23 DF,  p-value: 1.702e-07
```