Stat 6021: Guided Question Set 6

Tom Lever

10/09/22

For this guided question set, we will use the data set `nfl.txt`, which contains data on NFL team performance from the 1976 season. The variables are:

- y: Games won in the 14-game 1976 season
- x_1: Rushing yards
- x_2: Passing yards
- x_3: Punting average (yards / punt)
- x_4: Field-goal percentage (field goals made / field goals attempted)
- x_5: Turnover differential (turnovers acquired - turnovers lost)
- x_6: Penalty yards
- x_7: Percent rushing (rushing plays / total plays)
- x_8: Opponents' rushing yards
- x_9: Opponents' passing yards

1. Create a scatterplot matrix and find the correlation between all pairs of variables for this data set.

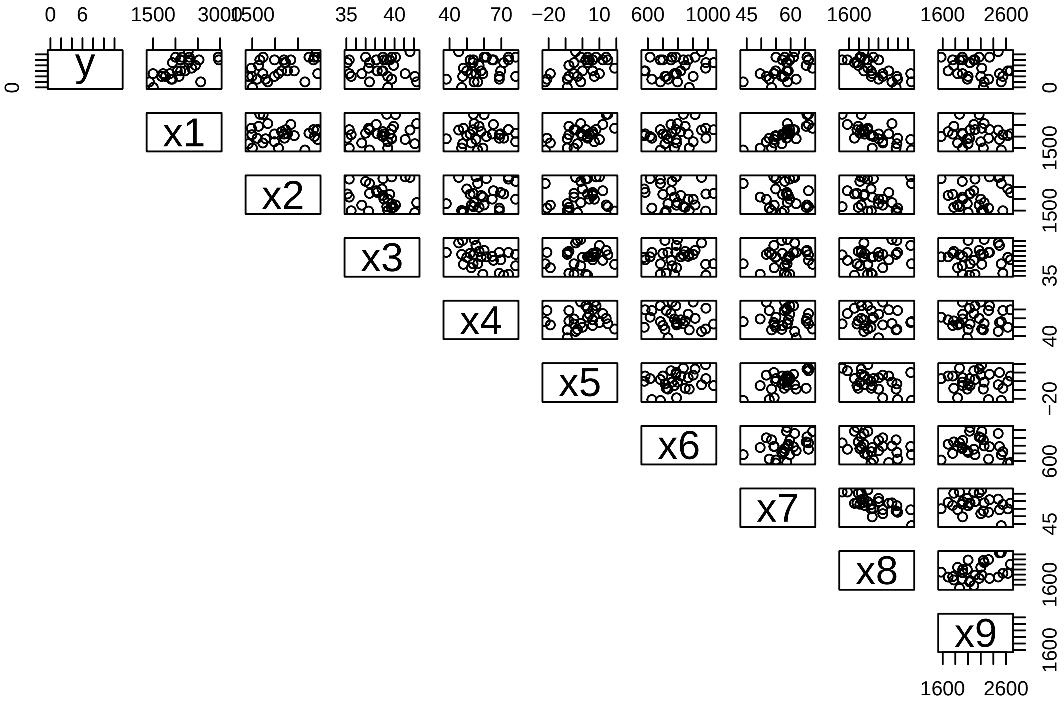
```
library(TomLeversRPackage)
data_set <- read.table("nfl.txt", header = TRUE)
head(data_set, n = 3)
```

```
##    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
```

```
nrow(data_set)
```

```
## [1] 28
```

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



```
correlation_matrix <- round(cor(data_set), 3)
correlation_matrix
```

```
##         y     x1     x2     x3     x4     x5     x6     x7     x8     x9
## y   1.000  0.593  0.483 -0.081  0.258  0.513  0.224  0.545 -0.738 -0.304
## x1  0.593  1.000 -0.037  0.212  0.070  0.600  0.253  0.837 -0.659 -0.111
## x2  0.483 -0.037  1.000 -0.069  0.302  0.135 -0.193 -0.197 -0.051  0.146
## x3 -0.081  0.212 -0.069  1.000 -0.413  0.115 -0.003  0.163  0.290  0.088
## x4  0.258  0.070  0.302 -0.413  1.000  0.149 -0.128 -0.101 -0.164  0.059
## x5  0.513  0.600  0.135  0.115  0.149  1.000  0.259  0.610 -0.470 -0.090
## x6  0.224  0.253 -0.193 -0.003 -0.128  0.259  1.000  0.367 -0.352 -0.173
## x7  0.545  0.837 -0.197  0.163 -0.101  0.610  0.367  1.000 -0.685 -0.203
## x8 -0.738 -0.659 -0.051  0.290 -0.164 -0.470 -0.352 -0.685  1.000  0.417
## x9 -0.304 -0.111  0.146  0.088  0.059 -0.090 -0.173 -0.203  0.417  1.000
```

```
linear_model <- lm(y ~ ., data = data_set)
summarize_linear_model(linear_model)
```

```
##
## Call:
## lm(formula = y ~ ., data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0408 -0.6802 -0.1131  0.9835  2.9785
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.292e+00  1.281e+01  -0.569 0.576312
## x1           8.124e-04  2.006e-03   0.405 0.690329
## x2           3.631e-03  8.410e-04   4.318 0.000414 ***
## x3           1.222e-01  2.590e-01   0.472 0.642750
## x4           3.189e-02  4.160e-02   0.767 0.453289
## x5           1.511e-05  4.684e-02   0.000 0.999746
## x6           1.590e-03  3.248e-03   0.490 0.630338
## x7           1.544e-01  1.521e-01   1.015 0.323547
## x8          -3.895e-03  2.052e-03  -1.898 0.073793 .
## x9          -1.791e-03  1.417e-03  -1.264 0.222490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.83 on 18 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7234
## F-statistic: 8.846 on 9 and 18 DF,  p-value: 5.303e-05
##
## E(y | x) =
##     B_0  +
##     B_x1 * x1 +
##     B_x2 * x2 +
##     B_x3 * x3 +
##     B_x4 * x4 +
##     B_x5 * x5 +
##     B_x6 * x6 +
##     B_x7 * x7 +
##     B_x8 * x8 +
##     B_x9 * x9
## E(y | x) =
##     -7.29193399423624 +
##     0.000812379092565623 * x1 +
##     0.00363114803117619 * x2 +
##     0.122170594156325 * x3 +
##     0.0318914599442878 * x4 +
##     1.51132859736972e-05 * x5 +
##     0.00158997199314889 * x6 +
##     0.154353290365774 * x7 +
##     -0.00389529726882156 * x8 +
##     -0.00179062999422216 * x9
## Number of observations: 28
## Estimated variance of errors: 3.34962287497396
## Multiple R:  0.903104063393996   Adjusted R:  0.850526556891741
```

Answer the following questions based on the output.

(a) Which predictors appear to be linearly related to the number of wins? Which predictors do not appear to have a linear relationship with the number of wins?

Let significance level $\alpha = 0.05$.

Let $t_{\alpha/2,\ n-p}$ be the quantile of a Student's t distribution with $n - p = 28 - 10 = 18$ for which the probability that a test statistic is greater is $\alpha/2 = 0.05/2 = 0.025$.

Since the probability, that the magnitude $|t|$ of a test statistic following the above distribution is greater than $t_{\alpha/2,\ n-p}$, is greater than α for all predictors, all predictors seem to be linearly related to the number of wins.

(b) Do you notice if any of the predictors are highly correlated with one another? If so, which ones?

According to Keith G. Calkins, "correlation coefficients whose magnitudes are [higher than] 0.7… indicate variables which can be considered highly correlated."

```
ifelse(correlation_matrix > 0.7, correlation_matrix, "")
```

```
##    y   x1      x2  x3  x4  x5  x6  x7      x8  x9 
## y  "1" ""      ""  ""  ""  ""  ""  ""      ""  ""
## x1 ""  "1"     ""  ""  ""  ""  ""  "0.837" ""  ""
## x2 ""  ""      "1" ""  ""  ""  ""  ""      ""  ""
## x3 ""  ""      ""  "1" ""  ""  ""  ""      ""  ""
## x4 ""  ""      ""  ""  "1" ""  ""  ""      ""  ""
## x5 ""  ""      ""  ""  ""  "1" ""  ""      ""  ""
## x6 ""  ""      ""  ""  ""  ""  "1" ""      ""  ""
## x7 ""  "0.837" ""  ""  ""  ""  ""  "1"     ""  ""
## x8 ""  ""      ""  ""  ""  ""  ""  ""      "1" ""
## x9 ""  ""      ""  ""  ""  ""  ""  ""      ""  "1"
```

Each predictor is highly correlated with itself. Rushing yards x_1 and percent rushing x_7 are highly correlated.

(c) What predictors would you first consider to use in a multiple linear regression? Briefly explain your choices.

I don't know.

2. Regardless of your answer to the previous question, fit a multiple regression model for the number of games won against the followning three predictors: the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8). Write the estimated regression equation.

```
linear_model <- lm(y ~ x2 + x7 + x8, data = data_set)
summarize_linear_model(linear_model)
```

```
## 
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = data_set)
## 
## 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
## 
## E(y | x) = 
##     B_0  + 
##     B_x2 * x2 + 
##     B_x7 * x7 + 
##     B_x8 * x8
## E(y | x) =
##     -1.8083720587051 +
##     0.0035980702139767 * x2 +
##     0.193960209583223 * x7 +
##     -0.0048154939700504 * x8
## Number of observations: 28
## Estimated variance of errors: 2.91125017423842
## Multiple R:  0.886739490104593   Adjusted R:  0.871547639962855
```

3. Interpret the estimated coefficient for the predictor percentage of rushing plays x_7 in context.

   Holding predictor team's passing yardage x_2 and opponents' yards rushing x_8 constant, for an increase in percentage of rushing plays x_7 of $1/0.194 = 5.156$, expected games won in the 14-game 1976 season increases by 1.

4. What is the estimated number of games a team would win for a team's passing yardage $x_2 = 2000\ yards$, percentage of rushing plays $x_7 = 48$, and opponents' yards rushing $x_8 = 2350\ yards$? Also provide a relevant 95-percent confidence interval for the number of games.

```
linear_model <- lm(y ~ x2 + x7 + x8, data = data_set)
predict(
    linear_model,
    newdata = data.frame(x2 = 2000, x7 = 48, x8 = 2350),
    interval = "confidence"
)
```

```
##        fit      lwr      upr
## 1 3.381448 1.710515 5.052381
```

```
predict(
    linear_model,
    newdata = data.frame(x2 = 2000, x7 = 48, x8 = 2350),
    interval = "predict"
)
```

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

   For a particular predictor $\vec{x_0} = (2000\ yards, 48, 2350\ yards)$ in the three-dimensional space with $x_2^{min} \leq x_2 \leq x_2^{max}$, $x_7^{min} \leq x_7 \leq x_7^{max}$, and $x_8^{min} \leq x_8 \leq x_8^{max}$, the 95-percent confidence interval for the expected / mean number of games a team would win $E\left(y|\vec{x}_0\right)$ is $[1.711,\ 5.052]$. The 95-percent prediction interval for a future number of games a team will win $y_{\vec{x}_0}$ is $[-0.516,\ 7.279]$.

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 with significance level $\alpha = 0.05$.

   We assume that errors are random, are independent, and follow a normal distribution with mean $E\left(\epsilon_i\right) = 0$ and variance $Var\left(\epsilon_i\right) = \sigma^2$. The multiple linear regression model from part 2 is useful in predicting the number of wins during the 1976 season if at least one of the predictor variables in the set $\{x_2, x_7, x_8\}$ contributes significantly to the model. We conduct a test of the null hypothesis $H_0 : \beta_{x_2} = \beta_{x_7} = \beta_{x_8} = 0$. The alternate hypothesis is $H_1 : \beta_{x_2} \neq 0\ or\ \beta_{x_7} \neq 0\ or\ \beta_{x_8} \neq 0$. The alternate hypothesis is also $H_1 : \beta_{x_i} \neq 0\ for\ i \in [2, 7, 8]$. If we reject the null hypothesis, at least one of the predictor variables contributes significantly to the model.

```
analyze_variance(linear_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193  26.172 3.100e-05 ***
## x7         1 139.501 139.501  47.918 3.698e-07 ***
## x8         1  41.400  41.400  14.221 0.0009378 ***
## Residuals 24  69.870   2.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## DFR: 3, SSR: 257.094281532564, MSR: 85.6980938441879
## F0: 29.4368703186446, Fcrit: 3.00878657044736, p: 3.27345828694872e-08
## DFT: 27, SST: 326.964285714286
## R2: 0.786306923310954, Adjusted R2: 0.759595288724823
## Number of observations: 28
```

```
test_null_hypothesis_involving_MLR_coefficients(linear_model, 0.05)
```

```
## Since probability 3.27345828694872e-08 is less than significance level 0.05,
## we reject the null hypothesis.
## We have sufficient evidence to support the alternate hypothesis.
```

Alternatively, since the test statistic $F_0 = 29.437$ is greater than the critical F value $F_{\alpha=0.05,\ k=2,\ n-p=28-3} = 3.009$, we reject the null hypothesis and support the alternate hypothesis.

Since we reject the null hypothesis and support the alternate hypothesis, at least one of the predictor variables contributes significantly to the model. Since at least one of the predictor variables contributes significantly to the model, the model is useful in predicting the number of wins during the 1976 season.