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.

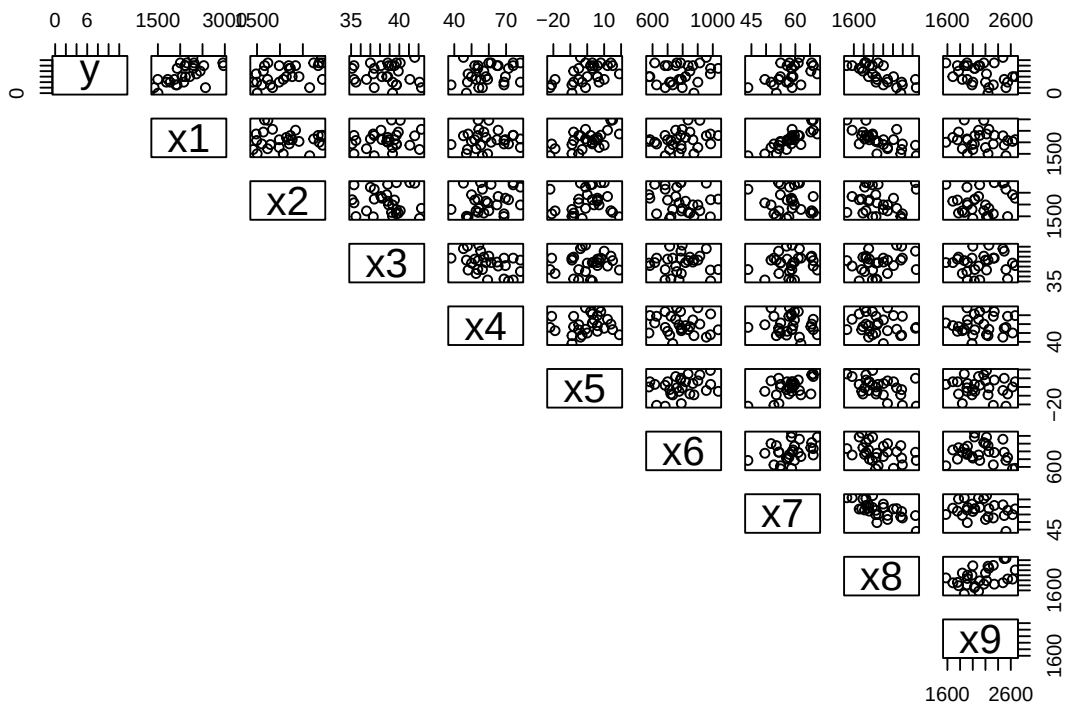
```
library(TomLeversRPackage)
data_set <- read.table("nfl.txt", header = TRUE)
head(data_set, n = 3)
```

```
##    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 ~ x1, data = data_set)
#summarize_linear_model(linear_model)
```

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?

We note that predictors in the set $\{x_1, x_2, x_5, x_7, x_8\}$ have moderate to high correlations with the number of wins. The last predictor is the only one moderate to highly negatively associated with the number of wins.

The predictors in the set $\{x_3, x_4, x_6, x_9\}$ do not have a strong linear relationship with the number of wins.

Considering statistics for each of nine linear models of number of games won y versus one predictor

x,

predictor x	test statistic t_0	probability p
x_1	3.758	0.00877
x_2	2.811	0.00927
x_3	-0.413	0.68300
x_4	1.364	0.18400
x_5	3.049	0.00522
x_6	1.172	0.25200
x_7	3.317	0.00269
x_8	-5.577	7.38e-06
x_9	-1.626	0.11609

Let significance level $\alpha = 0.05$.

Let $t_{\alpha/2,\ n-p} = t_{0.05/2,\ 28-2}$ be the quantile / critical value of a Student's t distribution with $n - p = 28 - 2 = 26$ degrees of freedom for which the probability that a test statistic is greater is $\alpha/2 = 0.05/2 = 0.025$.

Since the probability p, that the magnitude $|t|$ of a random test statistic following the above distribution is greater than t_0, is less than α for each predictor in the set $\{x_1, x_2, x_5, x_7, \text{and } x_8\}$, we have sufficient evidence to reject a null hypothesis that the predictor is not linearly related to the number of wins. We have sufficient evidence to support an alternate hypothesis that the predictor is linearly related to the number of wins. Since the probability is greater than α for each predictor in the set $\{x_3, x_4, x_6, \text{and } x_9\}$, we have insufficient evidence to reject a null hypothesis that the predictor is not linearly related to the number of wins. The predictor may not 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 (https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm), "Correlation coefficients whose magnitude[s] are between 0.9 and 1.0 indicate variables which can be considered very highly correlated. Correlation coefficients whose magnitude[s] are between 0.7 and 0.9 indicate variables which can be considered highly correlated. Correlation coefficients whose magnitude[s] are between 0.5 and 0.7 indicate variables which can be considered moderately correlated. Correlation coefficients whose magnitude[s] are between 0.3 and 0.5 indicate variables which have... low correlation[s]. Correlation coefficients whose magnitude[s] are less than 0.3 have little if any (linear) correlation."

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

```
##    y         x1       x2  x3  x4  x5  x6  x7       x8        x9
## y  "1"       ""       ""  ""  ""  ""  ""  ""       "-0.738"  ""
## x1 ""        "1"      ""  ""  ""  ""  ""  "0.837"  ""        ""
## x2 ""        ""       "1" ""  ""  ""  ""  ""       ""        ""
## x3 ""        ""       ""  "1" ""  ""  ""  ""       ""        ""
## x4 ""        ""       ""  ""  "1" ""  ""  ""       ""        ""
## x5 ""        ""       ""  ""  ""  "1" ""  ""       ""        ""
## x6 ""        ""       ""  ""  ""  ""  "1" ""       ""        ""
## x7 ""        "0.837"  ""  ""  ""  ""  ""  "1"      ""        ""
## x8 "-0.738"  ""       ""  ""  ""  ""  ""  ""       "1"       ""
## x9 ""        ""       ""  ""  ""  ""  ""  ""       ""        "1"
```

```
ifelse(abs(correlation_matrix) > 0.5, correlation_matrix, "")
```

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

```
## y  "1"      "0.593"  ""  ""  ""  "0.513" ""  "0.545"  "-0.738" ""
## x1 "0.593"  "1"      ""  ""  ""  "0.6"   ""  "0.837"  "-0.659" ""
## x2 ""       ""       "1" ""  ""  ""      ""  ""       ""       ""
## x3 ""       ""       ""  "1" ""  ""      ""  ""       ""       ""
## x4 ""       ""       ""  ""  "1" ""      ""  ""       ""       ""
## x5 "0.513"  "0.6"    ""  ""  ""  "1"     ""  "0.61"   ""       ""
## x6 ""       ""       ""  ""  ""  ""      "1" ""       ""       ""
## x7 "0.545"  "0.837"  ""  ""  ""  "0.61"  ""  "1"      "-0.685" ""
## x8 "-0.738" "-0.659" ""  ""  ""  ""      ""  "-0.685" "1"      ""
## x9 ""       ""       ""  ""  ""  ""      ""  ""       ""       "1"
```

Each predictor is highly correlated with itself. Rushing yards x_1 and percent rushing x_7 are highly correlated. x_1 is moderately correlated with x_5 and x_8. x_5 is moderately correlated with x_7. x_7 is moderately correlated with x_8.

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

I would consider using x_1, x_2, x_5, x_7, and x_8 in a multiple linear regression as these predictors have high correlation with the number of wins.

I would consider using x_2 as it has a high correlation with the number of wins, and a subset of $\{x_1, x_5, x_7, x_8\}$ as these are correlated with the number of wins but are also correlated among themselves.

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.

```r
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
## Critical value t(alpha/2 = 0.05/2, DFRes = 24): 2.06389856162803
```

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 estimated number of a games a team would win is 3.

   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 $[-0.516,\ 7.279]$. We are asked for the prediction interval.

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 $H_0 : \beta_{x_2} = \beta_{x_7} = \beta_{x_8} = 0$ that all coefficients in the set $\{\beta_{x_2}, \beta_{x_7}, \beta x_8\}$ are 0. The alternate hypothesis

is $H_1 : \beta_{x_2} \neq 0$ *or* $\beta_{x_7} \neq 0$ *or* $\beta_{x_8} \neq 0$ that at least one of the coefficients in the set $\{\beta_{x_2}, \beta_{x_7}, \beta x_8\}$ is not 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, F(alpha = 0.05, DFR = 3, DFRes = 24): 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 value $F_{\alpha=0.05,\ df_R=k=3,\ df_{Res}=n-p=28-4} = 3.009$, 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.

6. Report the value of the t statistic for the predictor x_7. What is the relevant conclusion from this t statistic? Also report the critical value for this hypothesis test. Perform the test at significance level 0.05.

   The test statistic for the predictor percent rushing x_7 $t_0 = 2.198$. The critical value $t_{\alpha/2=0.05/2,\ n-p=28-4} = 2.064$.

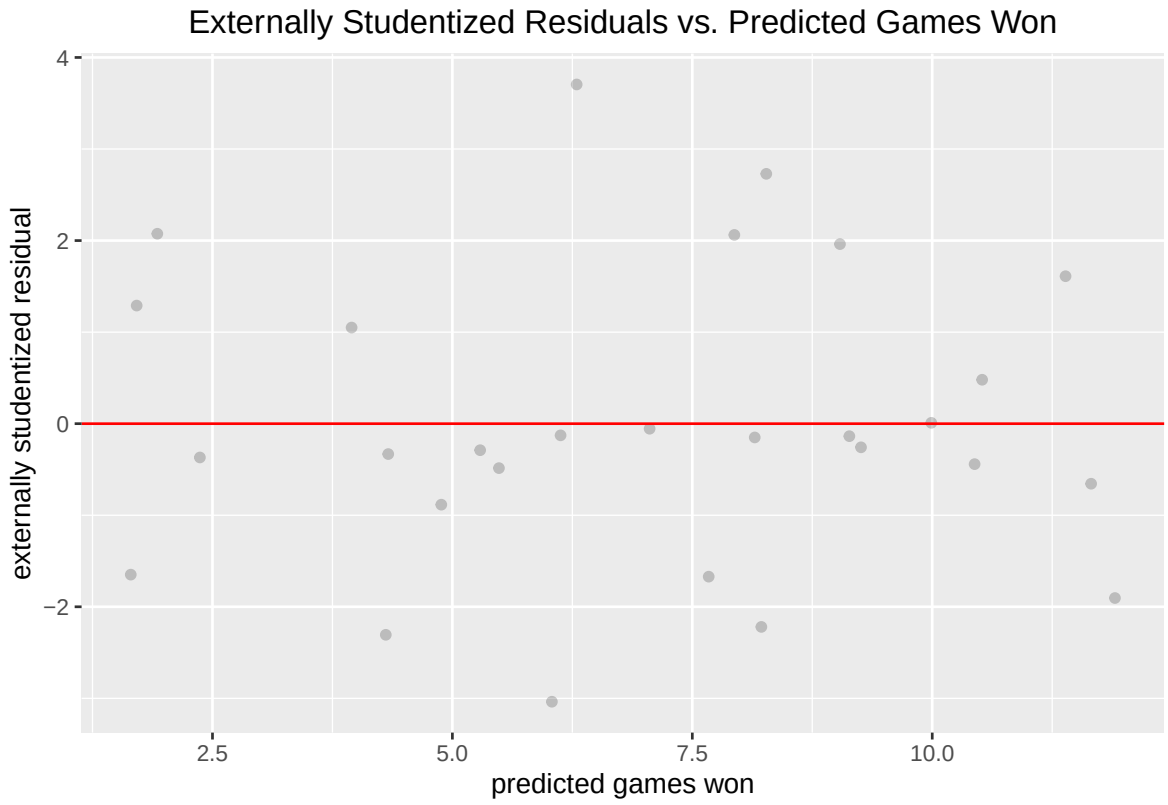
   In parallel, the probability, that the magnitude $|t|$ of a random test statistic is greater than the magnitude $|t_0|$ of our test statistic, $p = 0.0378$.

   Since the test statistic for the predictor x_7 is greater than the critical value, and the probability is less than the significance level $\alpha = 0.05$, we reject a null hypothesis that the individual predictor x_7 is insignificant in predicting the response y and can be removed from the model. We have sufficient evidence to support the alternate hypothesis that the individual predictor percent rushing x_7 is significant in predicting the response number of games won y and cannot be removed from the model.

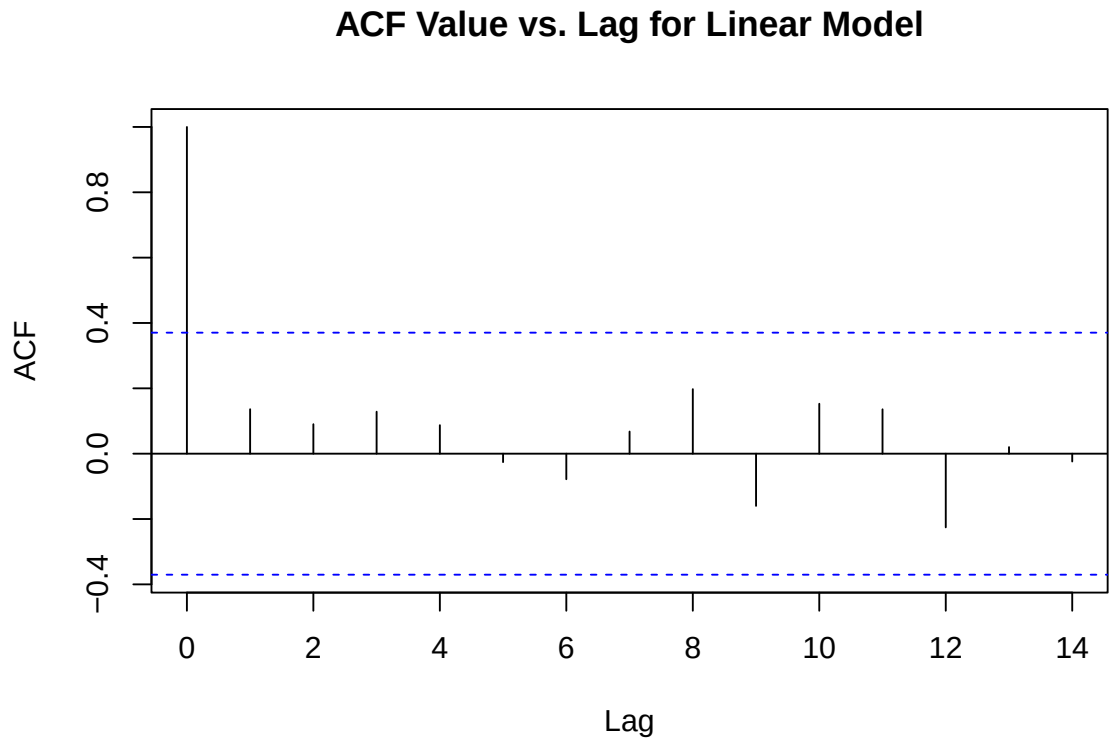
7. Check the regression assumptions by creating a residual plot, an ACF plot of the residuals, and a QQ plot of the residuals. Comment on these plots.

```
library(ggplot2)
ggplot(
    data.frame(
        externally_studentized_residual = linear_model$residuals,
```

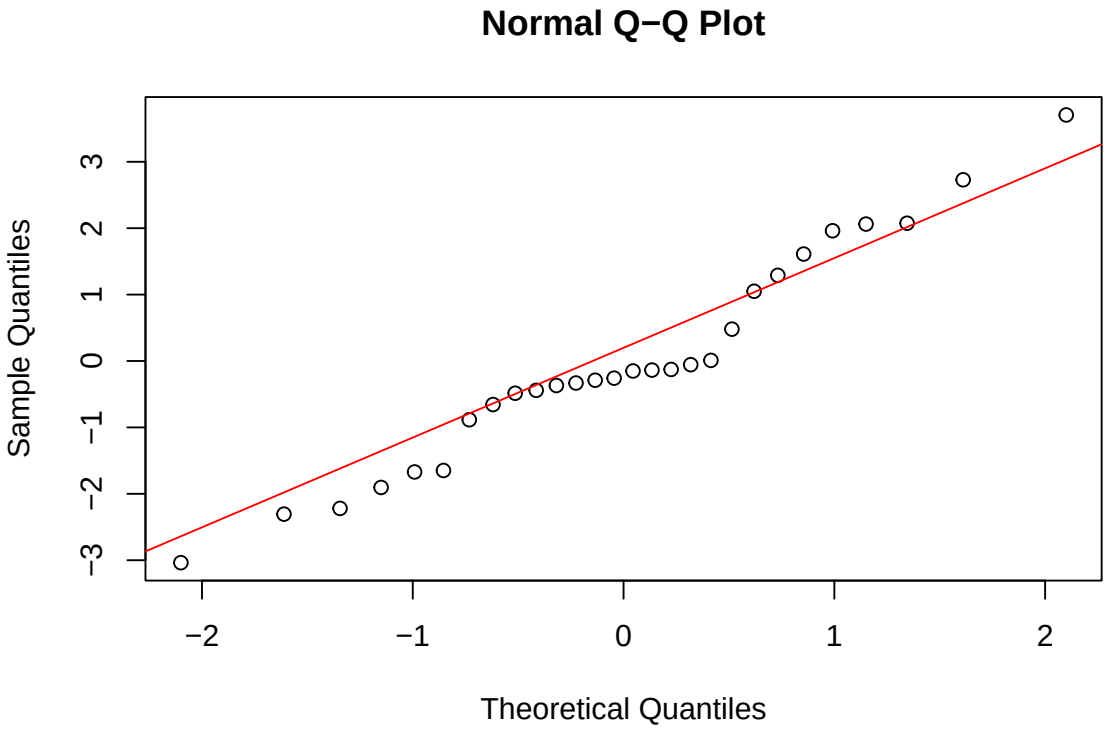
```
        predicted_games_won = linear_model$fitted.values
    ),
    aes(x = predicted_games_won, y = externally_studentized_residual)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted games won",
        y = "externally studentized residual",
        title = "Externally Studentized Residuals vs. Predicted Games Won"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```




```
acf(linear_model$residuals, main = "ACF Value vs. Lag for Linear Model")
```




```
qqnorm(linear_model$residuals)
qqline(linear_model$residuals, col = "red")
```



1. The assumption that the relationship between response / games won and predictors is linear, at least approximately, is met cannot be addressed.
2. The assumption that the residuals of the linear model of games won versus predictors have mean 0 is met. Residuals are evenly scattered around $e = 0$ at random.
3. The assumption that the distributions of residuals of the linear model for different weights have constant variance is met. Residuals are evenly scattered around $e = 0$ with constant vertical variance.
4. The assumption that the residuals of the linear model are uncorrelated is met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since all ACF value are insignificant, we have insufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We have insufficient evidence to conclude that the residuals of the linear model are correlated. We have insufficient evidence to conclude that the assumption that the residuals of the linear model are uncorrelated is not met.
5. The assumption that the residuals of the linear model are normally distributed is met. A linear model is robust to these assumptions. Considering a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model, since observations lie near the line of best fit / their theoretical values, a probability vs. externally studentized residuals plot / distribution is normal.

8. Consider adding another predictor x_1, the team's rushing yards for the season, to the linear model defined in part 2. 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".

   The test statistic for the predictor percent rushing x_1 $t_0 = 0.549$. The critical value $t_{\alpha/2=0.05/2,\ n-p=28-5} = 2.069$.

```
linear_model <- lm(y ~ x1 + x2 + x7 + x8, data = data_set)
summarize_linear_model(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x7 + x8, data = data_set)
##
## 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
##
## E(y | x) =
##     B_0 +
##     B_x1 * x1 +
##     B_x2 * x2 +
##     B_x7 * x7 +
##     B_x8 * x8
## E(y | x) =
##     -0.879171806754579 +
##     0.000904461554456168 * x1 +
##     0.00352135032121328 * x2 +
##     0.143759004067199 * x7 +
##     -0.00469941159827706 * x8
## Number of observations: 28
## Estimated variance of errors: 2.99860008211101
## Multiple R:  0.888294012756423   Adjusted R:  0.867399632927676
## Critical value t(alpha/2 = 0.05/2, DFRes = 23): 2.06865761041905
```

In parallel, the probability, that the magnitude $|t|$ of a random test statistic is greater than the magnitude $|t_0|$ of our test statistic, $p = 0.589$.

Since the test statistic for the predictor x_1 is less than the critical value, and the probability is greater than the significance level $\alpha = 0.05$, we have insufficient evidence to reject a null hypothesis that the individual predictor x_1 is insignificant in predicting the response number of games won y in the presence of the other predictors. We have insufficient evidence to support the alternate hypothesis that the individual predictor rushing yards x_1 is significant in predicting the response number of games won y in the presence of the other predictors. I agree with my classmate in that the result of the t test is insignificant. I would change my classmate's wording to "we have insufficient evidence to support the alternate hypothesis that the individual predictor rushing yards x_1 is significant in predicting the response number of games won y in the presence of the other predictors, and may be removed from the model".

To address the classmate's statement, we need to fit a simple linear regression model with x_1, the team's rushing yards, as the only predictor. The multiple linear regression model is not meant to address the

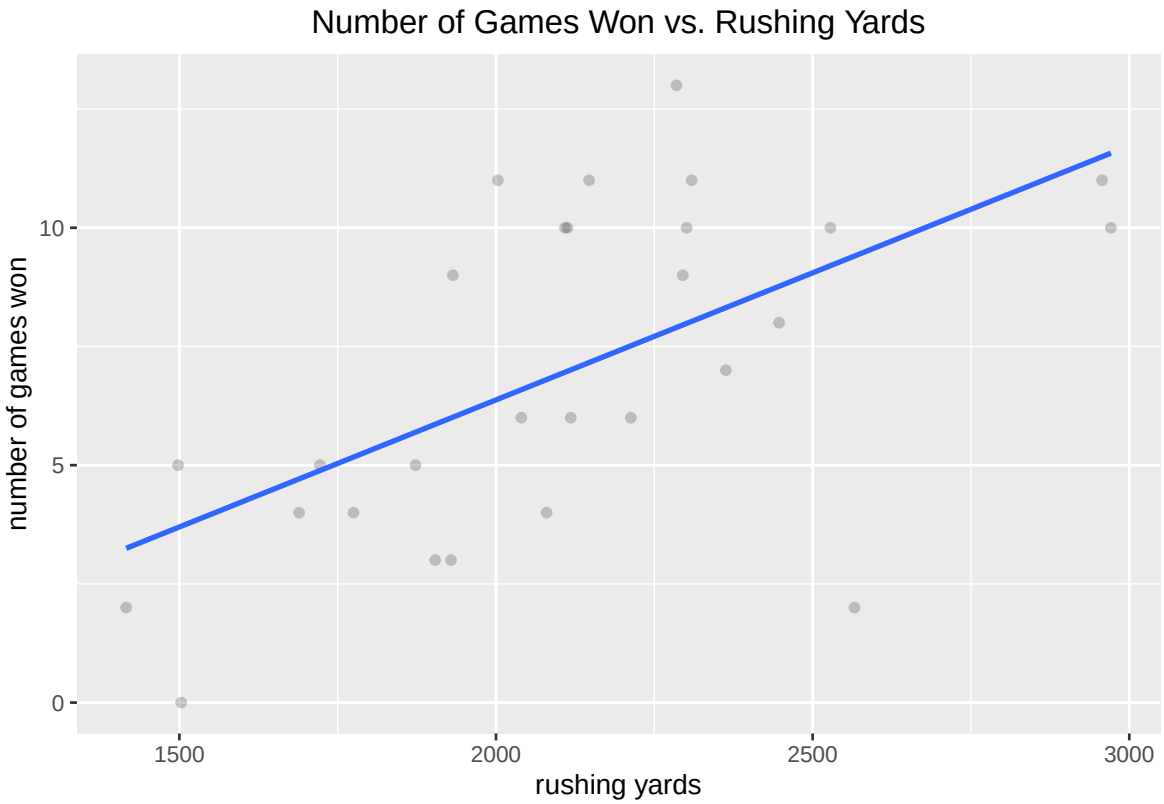
classmate's statement.

Fitting a simple linear regression model with x_1 as the only predictor and plotting number of games won y and versus the team's rushing yards x_1, we see that x_1 is a significant predictor and is linearly related to the number of wins.

```
linear_model <- lm(y ~ x1, data = data_set)
summarize_linear_model(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x1, data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4037 -1.3665 -0.6416  2.2539  5.1002
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.330015   3.053809  -1.418 0.168093
## x1           0.005352   0.001424   3.758 0.000877 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.855 on 26 degrees of freedom
## Multiple R-squared:  0.3519, Adjusted R-squared:  0.327
## F-statistic: 14.12 on 1 and 26 DF,  p-value: 0.000877
##
## E(y | x) =
##     B_0 +
##     B_x1 * x1
## E(y | x) =
##     -4.33001501055134 +
##     0.00535220560361909 * x1
## Number of observations: 28
## Estimated variance of errors: 8.14984886695685
## Multiple R:  0.593236043460868   Adjusted R:  0.571841932437823
## Critical value t(alpha/2 = 0.05/2, DFRes = 26): 2.05552943864287
```

```
ggplot(data_set, aes(x = x1, y = y)) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "rushing yards",
        y = "number of games won",
        title = "Number of Games Won vs. Rushing Yards"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```



x_1 is linearly related to the response on its own. That being said, it is not needed in a model with x_2, x_7, and x_8 as it doesn't improve the predictions significantly as it is moderately correlated with x_8 and highly correlated with x_7 / doesn't provide much addition insight to the prediction when the other predictors are already in the model.

```
library(dplyr)
cor(data_set %>% select(y, x1, x2, x7, x8))
```

```
##             y          x1          x2         x7          x8
## y   1.0000000  0.59323604  0.48273470  0.5453410 -0.73802730
## x1  0.5932360  1.00000000 -0.03674736  0.8372827 -0.65854627
## x2  0.4827347 -0.03674736  1.00000000 -0.1969154 -0.05104783
## x7  0.5453410  0.83728269 -0.19691540  1.0000000 -0.68504573
## x8 -0.7380273 -0.65854627 -0.05104783 -0.6850457  1.00000000
```