STAT 4310 Final Project

Duy Truong / David Wiley

April 17, 2019

Counter-Strike: Global Offensive - Player Titles

Introduction:

Counter Strike: Global Offensive is a first-person shooter action game. The most popular competitive game mode involves ten people split into two teams of five of either counter-terrorists or terrorists. The games are played per map and per round consisting of a best of 30 round.

Problem:

Since its release in 2012 and consequent first Major Championship in 2013, the game has finished 14 Major Championships. We've gathered data on the top 157 players as of April 2019 in order to see if individual player statistics are a major contributing factor in the number of Major Championship titles are held by a player. Our null hypothesis being that we believe that the relationship between the player performance is related to the number of Major Championship Titles they have. Our regressors are gathered from a website called HLTV.org that keeps track of player's performance throughout their entire career.

Purpose:

We decided to use all the statistics reported including Total Kills, the total amount of kills they have accumulated throughout their career. Headshot percentage, their career percentage of headshot kills over total kills. Total deaths, the career total of deaths. K/D ratio, a ratio of their career total kills to total kills. Damage per round, the career average of how much damage they inflict per round. Grenade damage per round, the career average of how much grenade damage they inflict per round. Maps and Rounds played; the career total times played. Kills, Deaths, and Assists per round, a career average of these based on round performance. Saved by and Saving Teammates, a career average of how many times a round either the player was saved or saving another teammate. Rating 1.0, HLTV's personal rating of each individual player; and finally Major Titles, the amount of times a player has won a Major Championship.

Multiple Linear Regression

From our coefficients we have:
x1 = 3.198596e-04, representing the total amount of kills in the player's career
x2 = 1.739149e-01, a proportion of headshot kills to kills
x3 = -8.698547e-04, number of total deaths in the player's career
x4 = -4.916218e+00, the ratio of total kills to total deaths
x5 = -6.188700e-02, the amount of damage done per round
x6 = 4.799665e-02, the amount of damage done with grenades only per round
x7 = 3.472583e-02, the total amount of maps played in the player's career
x8 = -8.801416e-04, the total amount of rounds played in a the player's career
x9 = 4.953407e+00, the proportion of kills per round
x10 = 1.188841e+01, the proportion of assist kills per round
x11 = 7.454971e+00, the proportion of deaths per round
x12 = -1.953565e+01, the proportion of times saved by a teammate per round
x13 = -1.070168e+01, the proportion of times saving a teammate per round
x14 = 2.114881e+00, HLTV's ranking system

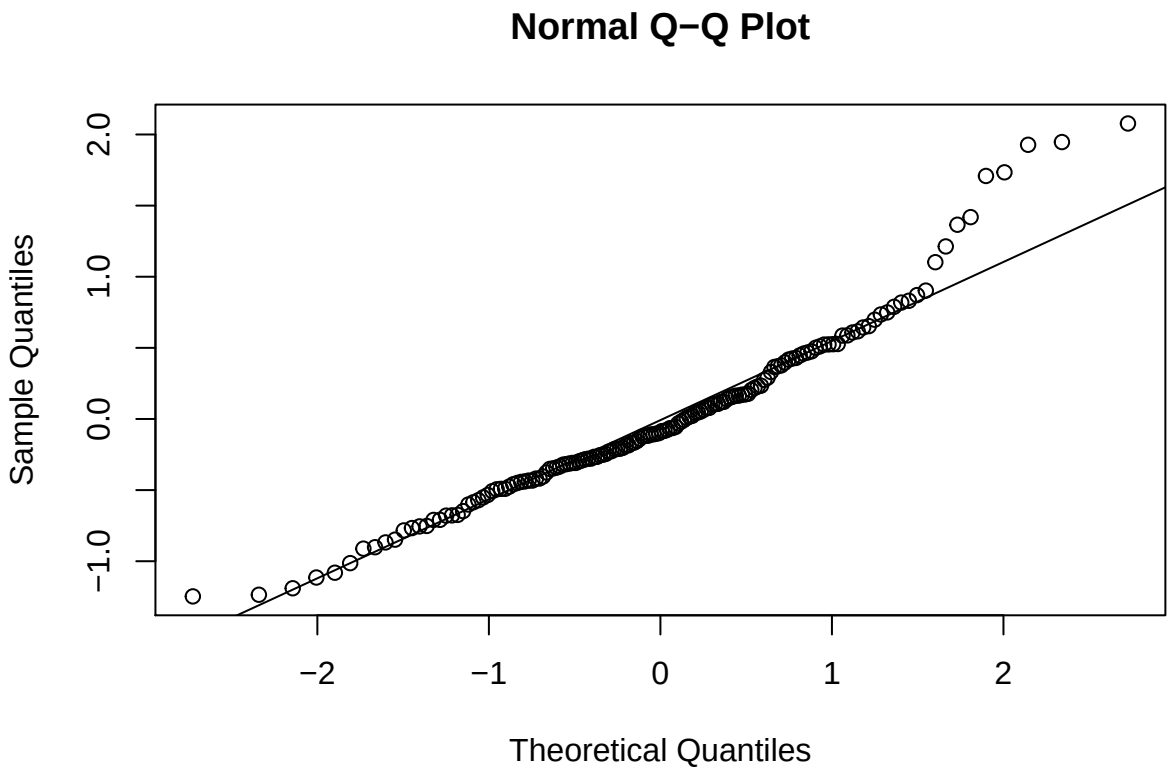
These represent an increase or decrease in relationship to the number of Major Titles won, holding the other variables constant.

Thus our multiple linear regression model is

$\hat{y} = (-0.702)+(0)x_1+(0.159)x_2+(-0.001)x_3+(-4.946)x_4+(-0.062)x_5+(0.048)x_6+(0.035)x_7+(-0.001)x_8+(4.941)x_9+(11.87$

Plotting Data

```
qqnorm(resid(fit))
qqline(resid(fit))
```




```
#new_dat01 = data.frame(dat)
#new_dat02 = new_dat01[,-1]
#ok = lm(Major.Titles~.,new_dat02)
#ok_sum = summary(ok)
#std_res= ok_sum$residuals/ok_sum$sigma
#qqnorm(std_res)
#qqline(std_res)
#shapiro.test(ok_sum$residuals)
#plot(ok$fitted.values, std_res)
#abline(h=0, col ='red')
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.95397, p-value = 5.037e-05
```

Looking at the QQ Plot we can see a light-tail distribution and with the Shapiro test giving a p-value of

4.982e-05 we reject the null hypothesis which means our data is not normally distributed.

Testing p-value

```
runs.test(fit$residuals, alternative = "two.sided")
```

```
##
##  Runs Test - Two sided
##
## data:  fit$residuals
## Standardized Runs Statistic = 0.48194, p-value = 0.6298
```

```
bartels.test(fit$residuals, alternative = "two.sided")
```

```
##
##  Bartels Test - Two sided
##
## data:  fit$residuals
## Standardized Bartels Statistic = 0.21237, RVN Ratio = 2.034,
## p-value = 0.8318
```

Both our runs and bartels test give a p-value larger than an $\alpha = 0.01$ which means we fail to reject the null at 1% significance level. We interpret this to mean that the autocorrelation is 0 mean our residuals are unrelated.

Testing for significance of Regression

$H_0 : \hat{\beta}_0 = H_0 : \hat{\beta}_1 = ... = H_0 : \hat{\beta}_j = 0;$

$H_1 : \hat{\beta}_j \neq 0$

```
# Test for significance of Regression.
sumfit
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
##     x10 + x11 + x12 + x13 + x14, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24713 -0.38314 -0.09235  0.36680  2.07730
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.020e-01  1.004e+01  -0.070  0.94435
## x1           3.209e-04  3.513e-04   0.914  0.36248
## x2           1.586e-01  8.365e-01   0.190  0.84989
## x3          -8.697e-04  5.146e-04  -1.690  0.09324 .
## x4          -4.946e+00  7.291e+00  -0.678  0.49865
## x5          -6.154e-02  6.128e-02  -1.004  0.31698
## x6           4.780e-02  6.486e-02   0.737  0.46237
## x7           3.474e-02  1.166e-02   2.978  0.00342 **
## x8          -8.815e-04  6.419e-04  -1.373  0.17184
## x9           4.941e+00  1.145e+01   0.431  0.66677
## x10          1.188e+01  7.832e+00   1.516  0.13168
## x11          7.416e+00  1.224e+01   0.606  0.54556
```

```
## x12         -1.953e+01  8.024e+00  -2.434  0.01617 *
## x13         -1.067e+01  6.508e+00  -1.639  0.10339
## x14          2.118e+00  2.925e+00   0.724  0.47007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6479 on 141 degrees of freedom
## Multiple R-squared:  0.4055, Adjusted R-squared:  0.3464
## F-statistic: 6.868 on 14 and 141 DF,  p-value: 1.383e-10
```

```
# P Value from summary(fit)
pval = 1.378e-10
pval
```

```
## [1] 1.378e-10
```

Since the p-value is 1.378e-10 which is less than α, we reject H_0 and conclude that the regression model is significant.
Since the regression model is significant we can say that there is a linear relationship between the response y and the regressors $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_10, x_11, x_12, x_13, x_14$

Contribution of Regressors

```
sumfit
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
##     x10 + x11 + x12 + x13 + x14, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24713 -0.38314 -0.09235  0.36680  2.07730
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.020e-01  1.004e+01  -0.070  0.94435
## x1           3.209e-04  3.513e-04   0.914  0.36248
## x2           1.586e-01  8.365e-01   0.190  0.84989
## x3          -8.697e-04  5.146e-04  -1.690  0.09324 .
## x4          -4.946e+00  7.291e+00  -0.678  0.49865
## x5          -6.154e-02  6.128e-02  -1.004  0.31698
## x6           4.780e-02  6.486e-02   0.737  0.46237
## x7           3.474e-02  1.166e-02   2.978  0.00342 **
## x8          -8.815e-04  6.419e-04  -1.373  0.17184
## x9           4.941e+00  1.145e+01   0.431  0.66677
## x10          1.188e+01  7.832e+00   1.516  0.13168
## x11          7.416e+00  1.224e+01   0.606  0.54556
## x12         -1.953e+01  8.024e+00  -2.434  0.01617 *
## x13         -1.067e+01  6.508e+00  -1.639  0.10339
## x14          2.118e+00  2.925e+00   0.724  0.47007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6479 on 141 degrees of freedom
```

```
## Multiple R-squared:  0.4055, Adjusted R-squared:  0.3464
## F-statistic: 6.868 on 14 and 141 DF,  p-value: 1.383e-10
```

From the regression table we can see that the p-values of the intercept is significantly greater than α of 0.05. Therefore at a 5% significance level we can interpret that the intercept is equal to 0.

From the regression table we can see that the p-values of $x_1 : x_6, x_8 : x_{11}, x_{13} : x_{14}$ are greater than α of 0.05. Therefore at a 5% significance level we can interpret that the regressor x_1 is not contributing significantly to the model of y.

We can also see that the p-values of x_7, x_{12} is less than α of 0.05. Therefore at a 5% significance level we can interpret that the regressor x_7 is contributing significantly to the model of y, given that all our other regressons are also in the model.

From this we can conclude that the regressors x_7 and x_{12} are contributing significantly to our model of y, while the other regressors have little to not contribution to the model.

```
sumfit$r.squared
```

```
## [1] 0.4054516
```

```
sumfit$adj.r.squared
```

```
## [1] 0.3464184
```

Our $R^2 = 0.41$ which means that only 41% of our variance in total Major titles is explained by all regressors. Adding regressors that do not have an impact decreased our adjusted R^2 to 0.35.

```
newfit = lm(y~x7 + x12,dat)
sumnewfit = summary(newfit)
sumnewfit$r.squared
```

```
## [1] 0.2868018
```

```
sumnewfit$adj.r.squared
```

```
## [1] 0.277479
```

With only x_7 and x_{12} our $R^2 = 0.29$ and the adjusted $R^2 = 0.28$ which are values approximately 12% different than before which shows still that these regressors alone have the most impact on our original model.

Confidence Interval

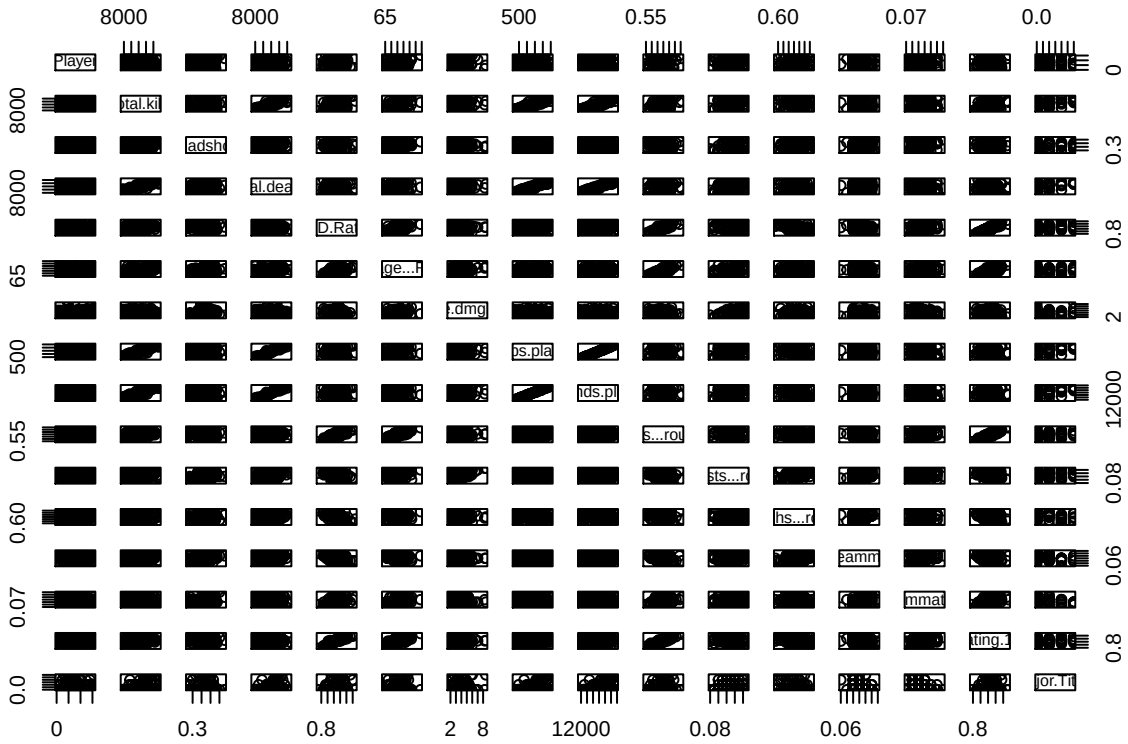
Looking at the two most significant regressors alone, we can say with 95% confidence that our coefficient x_7 lies in between 1.17e-02 and 0.06 while x_{12} lies between -3.5e+01 and -3.7 with all regressors.

While having only these two regressors in the model we can say with 95% confidence that our coefficient x_7 lies in between 0.002 and 0.003 while x_{12} lies between -32.6 and -11.9

Eigensystem and Measure of Multicollinearity

After testing for multicollinearity, we see there are high VIFs associated with x_1, x_3, x_4, x_7, x_8, and x_9, with x_8, total amount of rounds played, having the strongest level of multicollinearity at vif of 1724.

```
plot(dat)
```



From both our plots and table we can see a very strong positive relatioship among our regressors. It is not surprising to see a strong positive relationship between x_7 and x_8 because in this game, the Maps consisnt of rounds, so as the number of Maps increase we can undoubtedly expect an increase in the number of rounds played.

We calculated our kappa value as 11497, which is indicative of strong multicollinearity in our data set.

Condition Indices

exx Values for x_{11}, x_{12}, x_{13}, and x_{14} are above 1000 leading us to believe that these regressors are involved with multicollinearity. Further testing is required.

Influential Observations

From calculating our table of influences, we can see there are 14 possible influential observations, these infuential observations are related to players Xantares, HEN1, Calyx, dupreeh, ropz, flusha, JW, gla1ve, pashaBiceps, Relyks, Dima, Karrigan, ngiN, and MSL. Interesting note is that 4 of these players, dupreeh, flusha, JW, and gla1ve have the max amount of possible Major Titles currently while PashaBiceps has 1 title and the rest have 0.

After removing an individual player, we saw no difference to the R^2 value or the y-intercept value across all removals.

Best Subset

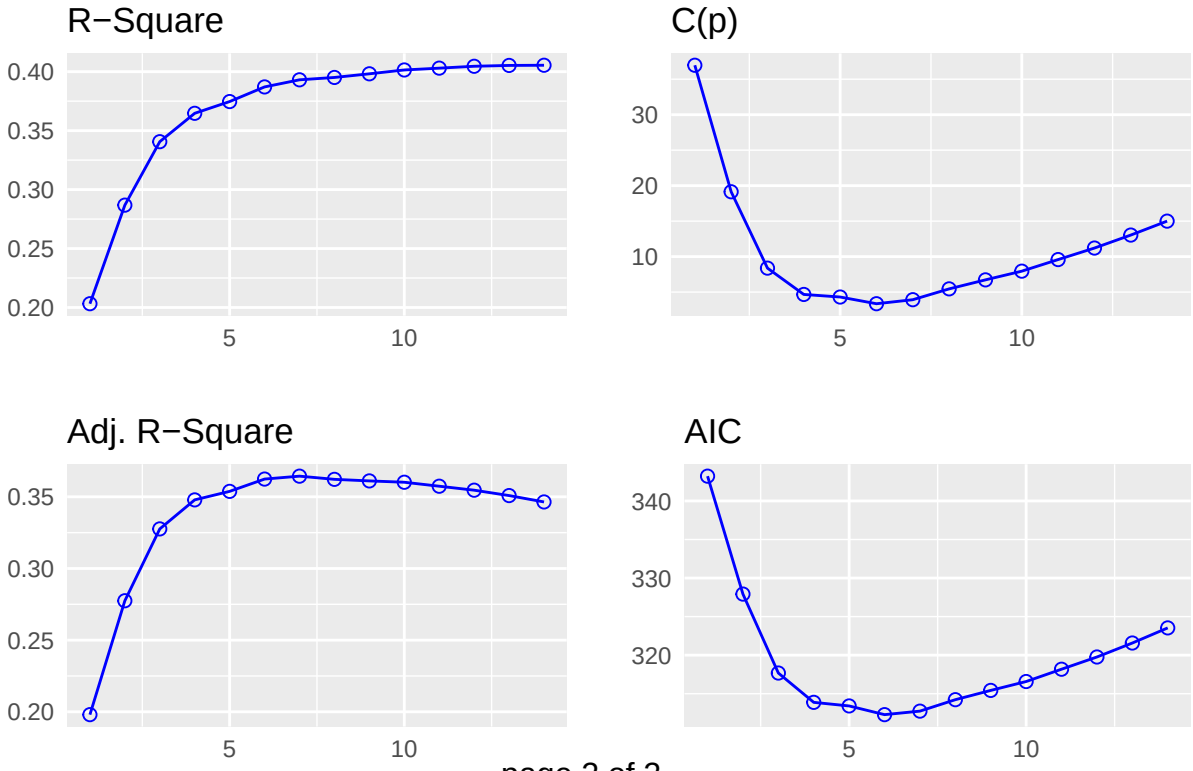
```
suppressMessages(library(olsrr))
best = ols_step_best_subset(lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14,dat))
best
```

```
##                    Best Subsets Regression
## -------------------------------------------------------------
## Model Index    Predictors
```

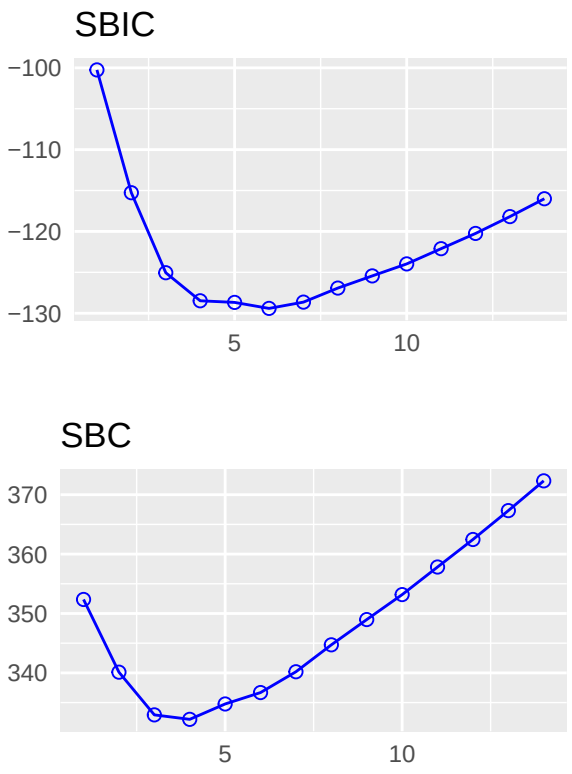
```
## ------------------------------------------------------------
##      1         x7
##      2         x7 x12
##      3         x7 x8 x12
##      4         x6 x7 x8 x12
##      5         x7 x8 x10 x12 x13
##      6         x3 x7 x10 x11 x12 x13
##      7         x3 x7 x8 x10 x11 x12 x13
##      8         x1 x3 x4 x7 x8 x10 x12 x13
##      9         x1 x3 x5 x7 x8 x10 x11 x12 x13
##     10         x1 x3 x5 x6 x7 x8 x10 x11 x12 x13
##     11         x1 x3 x5 x6 x7 x8 x10 x11 x12 x13 x14
##     12         x1 x3 x4 x5 x6 x7 x8 x10 x11 x12 x13 x14
##     13         x1 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14
##     14         x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14
## ------------------------------------------------------------
##
##                                                  Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC         SBC         MSEP
## ----------------------------------------------------------------------------------------------------
##   1        0.2032      0.1980      0.1784    36.9666    343.2120    -100.2555    352.3616    0.5218
##   2        0.2868      0.2775      0.2546    19.1384    327.9189    -115.2680    340.1184    0.4732
##   3        0.3406      0.3276      0.3032     8.3752    317.6793    -125.0487    332.9286    0.4433
##   4        0.3647      0.3479      0.3115     4.6678    313.8792    -128.4783    332.1783    0.4328
##   5        0.3746      0.3538      0.3212     4.3054    313.4138    -128.6778    334.7628    0.4317
##   6        0.3871      0.3624      0.3235     3.3611    312.2856    -129.4093    336.6844    0.4289
##   7        0.3931      0.3644      0.3185     3.9282    312.7402    -128.6408    340.1889    0.4304
##   8        0.3951      0.3622      0.3094     5.4542    314.2256    -126.9243    344.7241    0.4349
##   9        0.3982      0.3611      0.2996     6.7231    315.4285    -125.4371    348.9769    0.4386
##  10        0.4014      0.3602      0.2892     7.9498    316.5809    -123.9699    353.1792    0.4423
##  11        0.4030      0.3574      0.2825     9.5822    318.1765    -122.1091    357.8246    0.4474
##  12        0.4046      0.3546      0.2777    11.2061    319.7615    -120.2453    362.4595    0.4524
##  13        0.4053      0.3509       0.269    13.0360    321.5734    -118.1871    367.3212    0.4583
##  14        0.4055      0.3464       0.261    15.0000    323.5336    -116.0063    372.3313    0.4647
## ----------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
plot(best)
```

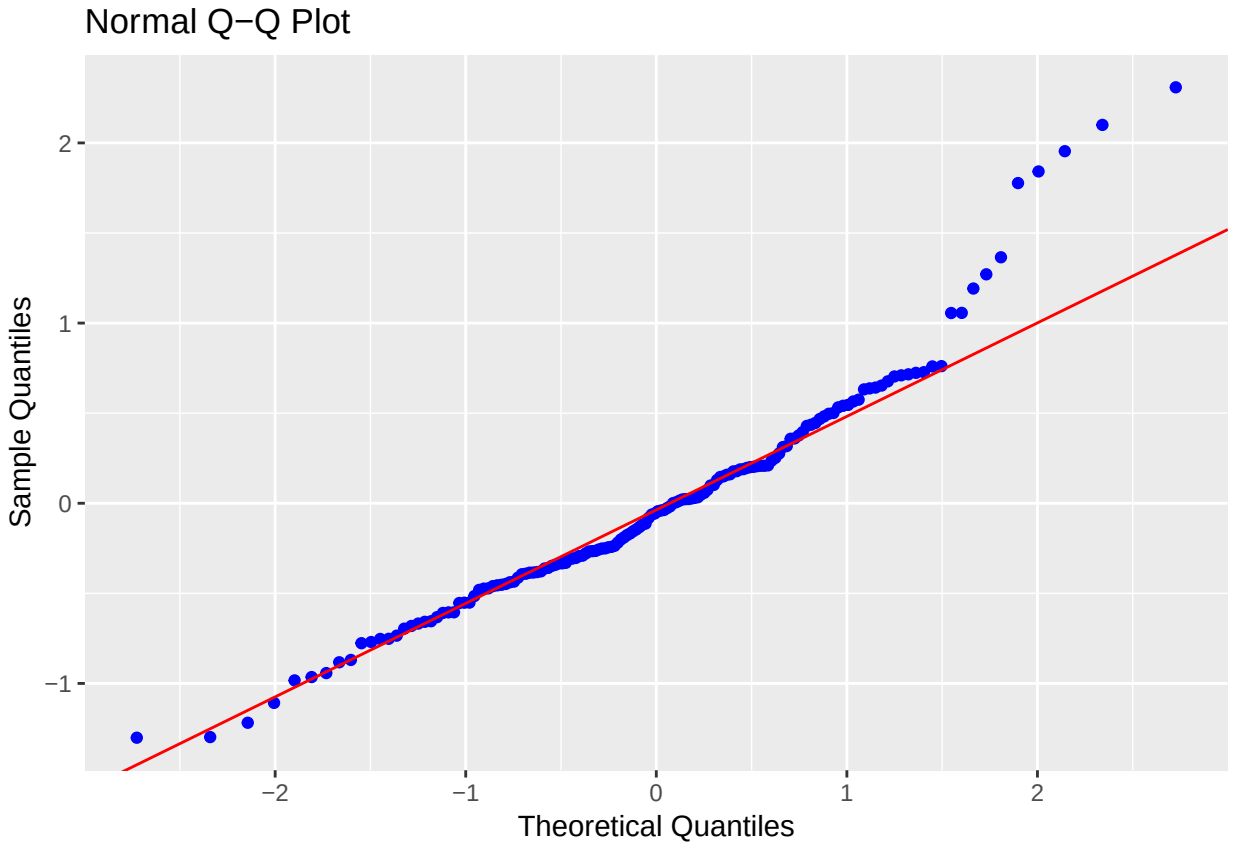
page 1 of 2



page 2 of 2




```
ols_plot_resid_qq(lm(y~x3+x7+x10+x11+x12+x13))
```



After reviewing our best fit models, it looks like the sixth model, $x_3, x_7, x_{10}, x_{11}, x_{12}, x_{13}$, would be our best fit model based from our data. This would then make our model to be:

$$\hat{y} = -11.1497 - 0.0013x_3 + 0.0241x_7 + 8.0514x_{10} + 17.7394x_{11} - 18.9007x_{12} - 9.9036x_{13}$$

Conclusion:

Based on the data and our statistical analysis I believe it is safe to say that there is some association between player statistics and their amount of Major Titles they have. Most interestingly was that the two most contributing regressors were Maps played and the proportion of times saved by a teammate per round, the latter being a factor we hadn't thought would ever be a major contributing regressor. Even more surprising was that albeit small, the proportion of teammates saved, had a negative impact in relationship to titles won. Maps played playing a role though is not as big as a surprise, those who have more play time should see marked increase in their gameplay. The R^2 value is not very high, but this is to be expected. Counter Strike: Global Offensive (CSGO) is a highly skill-based game, but also heavily relies on teamwork. Individuals can create opportunities, but in the end its your teammate that must follow up these plays. When a team as a whole performs well, we can see the raw data showing more wins in the upper echelon of players, especially those who have remained on a team for several majors.