Project 3 158: Multiple Linear Regression

Tesfa Asmara and Kevin Loun

4/09/2022

Contents

1 Introduction

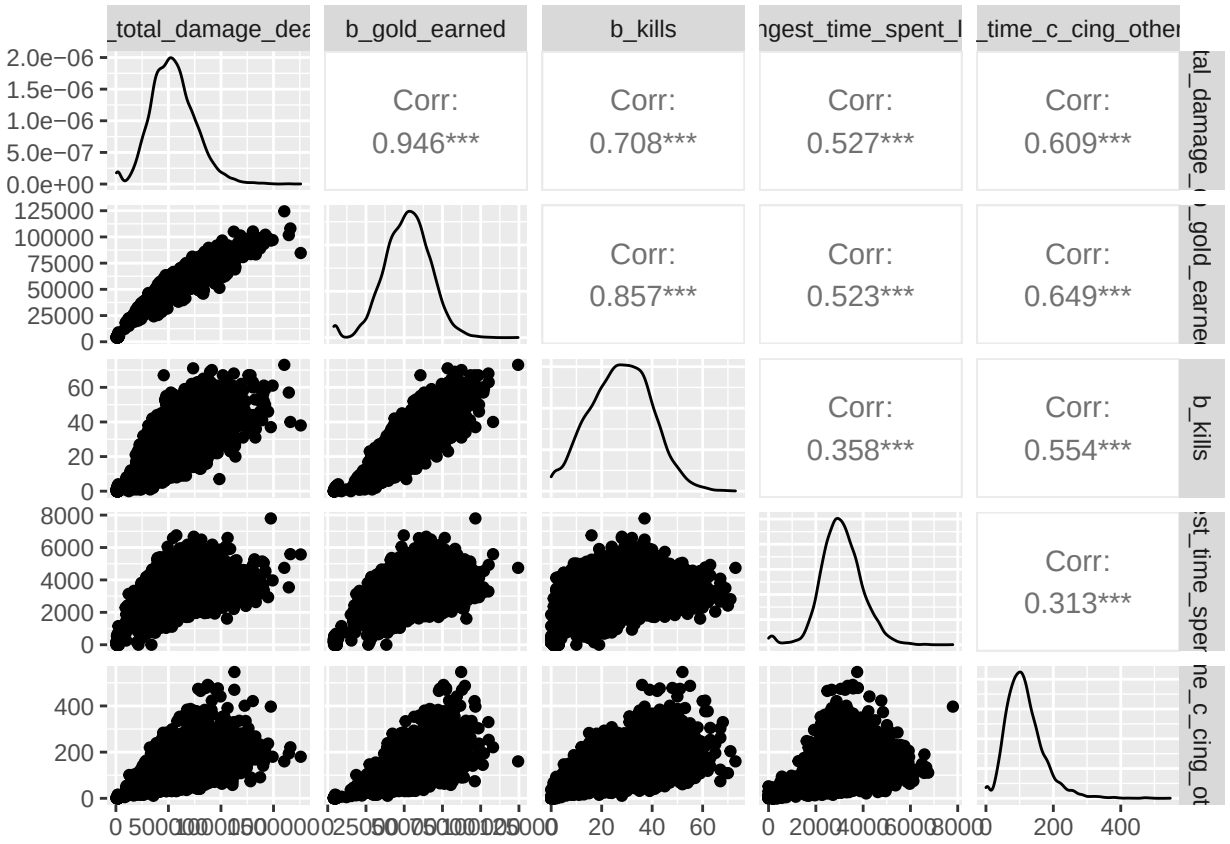
The dataset for this project contains 10,000 League of Legends ranked matches from the North American region with 775 variables offered through the Riot Games API, provided on Kaggle (Games 2021a)(James 2020). Each match is pulled from players who rank Gold in the League system, a ranking system that matches players of a similar skill level to play with and against each other. Amongst North American players, the Gold skill level was the second most common tier, achieved by 27.7 percent of players, or approximately 49.86 million players when considered against Riot Games' player base of 180 million ("Distribution of League of Legends (Lol) Summoners in North America as of October 2021, by Tier" 2021)(Games 2021b). This dataset will be referred to as `lol10`.

For this project, the following variables are of interest: time spent crowd controlling others, map side, longest time spent living, kills, gold earned, and total damage dealt. A figure including all the relevant variables and their description is attached at the end.

2 Hypothesis

We consider the following research question:

3 Feature Engineering



For $n = 5$ observations, Table B.6 in ALSM is employed to assess whether or not the magnitude of the correlation coefficient supports the reasonableness of the normality assumption. The feature engineering we conducted was minimal.

4 Interaction Variables

We wish to test formally in the `lol10` dataset whether interaction terms between the four explanatory variables should be included in the regression model. We therefore need to consider the following regression model: ,

$$\begin{aligned} \text{b\_total\_damage\_dealt} = \beta_0 &+ \beta_1(\text{b\_gold\_earned}) + \\ &\beta_2(\text{b\_kills}) + \beta_3(\text{b\_longest\_time\_spent\_living}) + \\ &\beta_4(\text{b\_time\_c\_cing\_others}) + \beta_5(\text{b\_gold\_earned} \times \text{b\_kills}) + \\ &\beta_6(\text{b\_gold\_earned} \times \text{b\_longest\_time\_spent\_living}) + \beta_7(\text{b\_gold\_earned} \times \text{b\_time\_c\_cing\_ot} \\ &\beta_8(\text{b\_kills} \times \text{b\_longest\_time\_spent\_living}) + \beta_9(\text{b\_kills} \times \text{b\_time\_c\_cing\_others}) + \\ &\beta_{10}(\text{b\_longest\_time\_spent\_living} \times \text{b\_time\_c\_cing\_others}) + \epsilon \end{aligned} \tag{1}$$

,

.

We wish to test whether any interaction terms are needed. We do so by performing a partial F-test by fitting both the reduced and full models separately and thereafter comparing them using the `anova()` function.

Since $F \approx 297.3748455$ (p-value ≈ 0), we reject the null hypothesis $H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$ at the $\alpha = 0.05$ level of significance. This means that the interaction terms do not contribute significant information to the `b_total_damage_dealt` once the explanatory variables `b_gold_earned, b_kills, b_longest_time_spent_living, b_time_c_cing_others` have been taken into consideration.

5 Computational Model

From our domain experience, we consider the following parsiminous models: ,

$$, \begin{array}{c} \text{b\_total\_damage\_dealt} = \beta_0 + \beta_1(\text{b\_gold\_earned}) + \\ \beta_2(\text{b\_kills}) + \epsilon \end{array}, \tag{2}$$

and ,

$$, \begin{array}{l} \text{b\_total\_damage\_dealt} = \beta_0 + \beta_1(\text{b\_gold\_earned}) + \\ \beta_2(\text{b\_kills}) + \beta_3(\text{b\_longest\_time\_spent\_living}) +, \\ \epsilon \end{array} \tag{3}$$

.

We compare the two models using cross-validation prediction error.

In comparing which model is better, the CV RMSE provides information on how well the model did predicting each $1/v$, where $v =$ the number of folds, hold out sample. We can compare the model RMSE to the original variability seen in the `b_total_damage_dealt` variable. The original variability (measured by standard deviation) of `b_total_damage_dealt` was 2.1223789×10^5. After running Model 1, the remaining variability (measured by RMSE averaged over the folds) is 2625.6822105; after running Model 2, the remaining variability (measured by RMSE averaged over the folds) is 2638.9578378.

Hence, the better computational model is ,

$$, \begin{array}{c} \text{b\_total\_damage\_dealt} = \beta_0 + \beta_1(\text{b\_gold\_earned}) + \\ \beta_2(\text{b\_kills}) + \epsilon \end{array}, \tag{4}$$

.

6 Statistical Model

For the four predictors in the `lol10` data, we know there are $2^4 = 16$ possible models. The four-parameter model ,

$$, \begin{array}{l} \text{b\_total\_damage\_dealt} = \beta_0 + \beta_1(\text{b\_gold\_earned}) + \\ \beta_2(\text{b\_kills}) + \beta_3(\text{b\_longest\_time\_spent\_living}) +, \\ \beta_4(\text{b\_time\_c\_cing\_others}) + \epsilon \end{array} \tag{5}$$

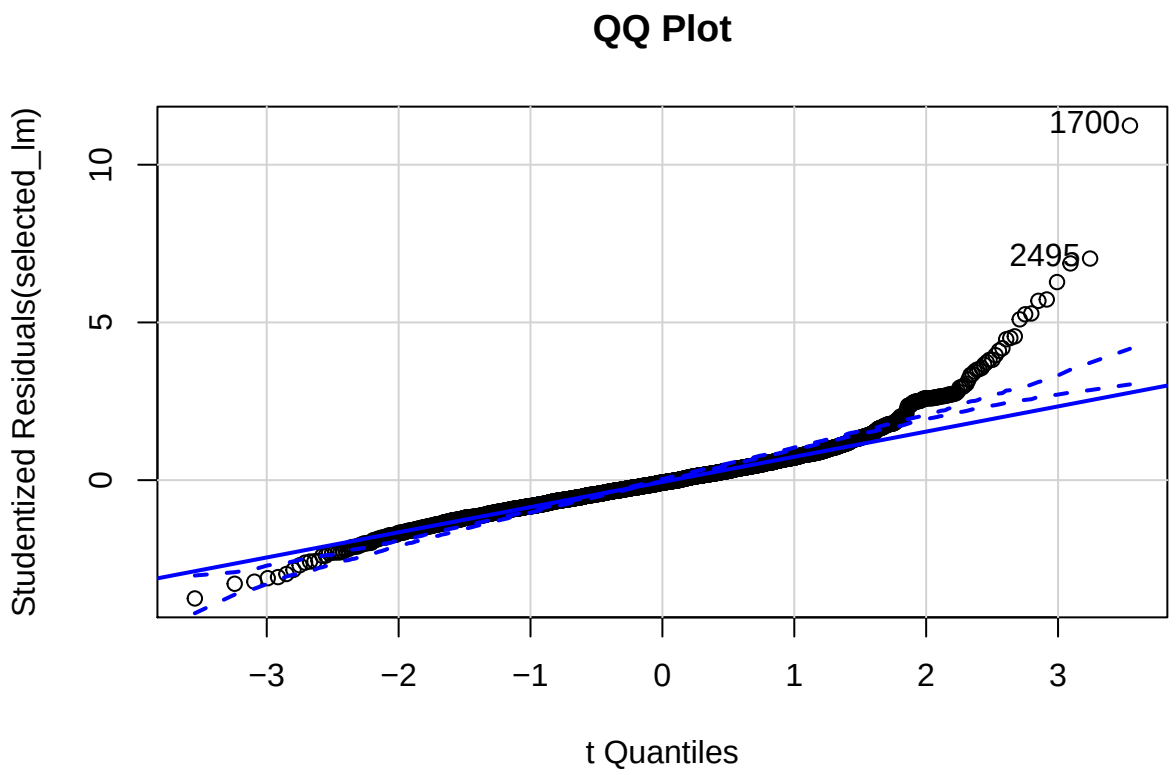
is identified as best by the $R_{a,p}$ criterion; it has $\texttt{max}(R_{a,p}) = 4$ and will serve as the selected model.

For the test data, the linear model we selected has $R^2 = 0.9321065$ and a $R^2_{adj} = 0.9319978$. Therefore, 0.0093211% of the variability in `b_total_damage_dealt` for players who rank Gold in the North American region is explained by the variables `b_gold_earned, b_kills, b_longest_time_spent_living, b_time_c_cing_others`.

```
# Assessing Outliers
outlierTest(selected_lm) # Bonferonni p-value for most extreme obs
```

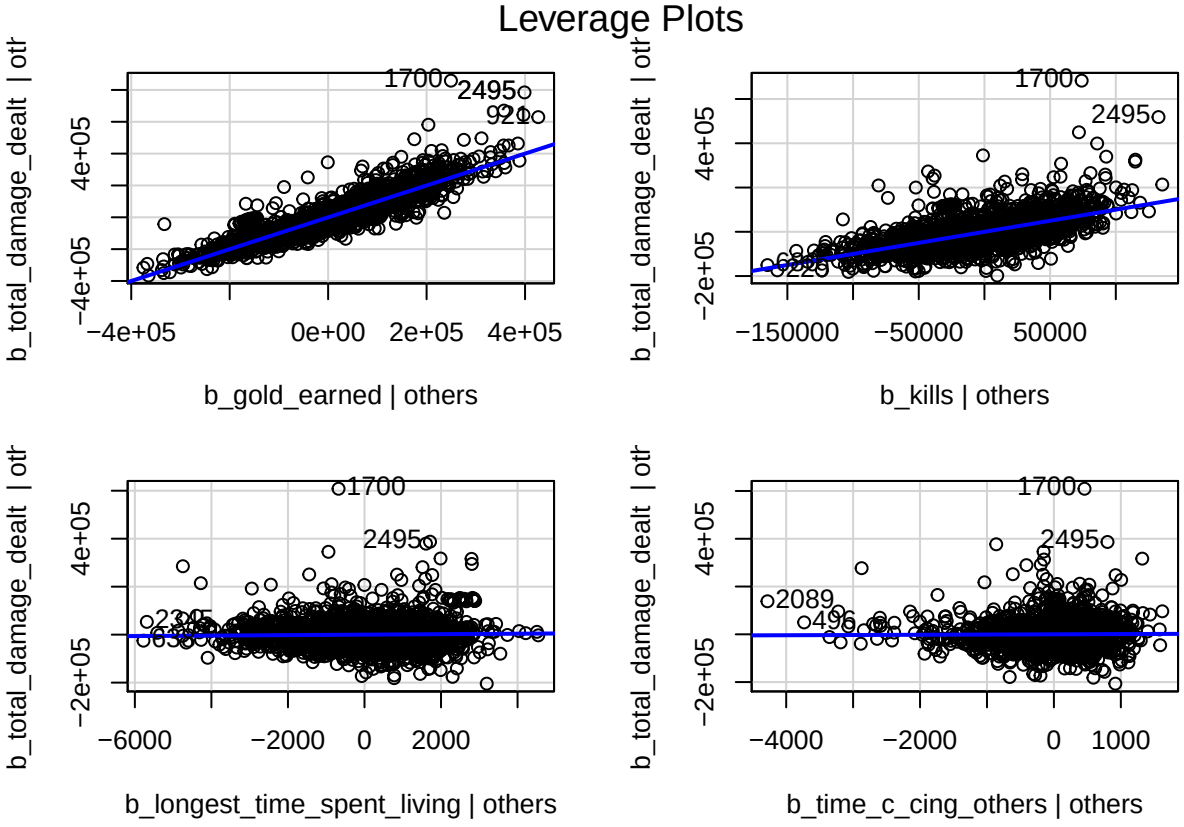
```
##       rstudent unadjusted p-value Bonferroni p
## 1700 11.241725         1.2239e-28   3.0646e-25
## 2495  7.024868         2.7520e-12   6.8910e-09
## 1259  6.870322         8.0580e-12   2.0177e-08
## 1972  6.279619         3.9899e-10   9.9908e-07
## 413   5.727456         1.1417e-08   2.8588e-05
## 335   5.684848         1.4615e-08   3.6596e-05
## 1857  5.284712         1.3679e-07   3.4253e-04
## 1063  5.258432         1.5761e-07   3.9465e-04
## 1666  5.097518         3.6990e-07   9.2624e-04
## 1656  4.559145         5.3838e-06   1.3481e-02
```

```
qqPlot(selected_lm, main="QQ Plot") #qq plot for studentized resid
```

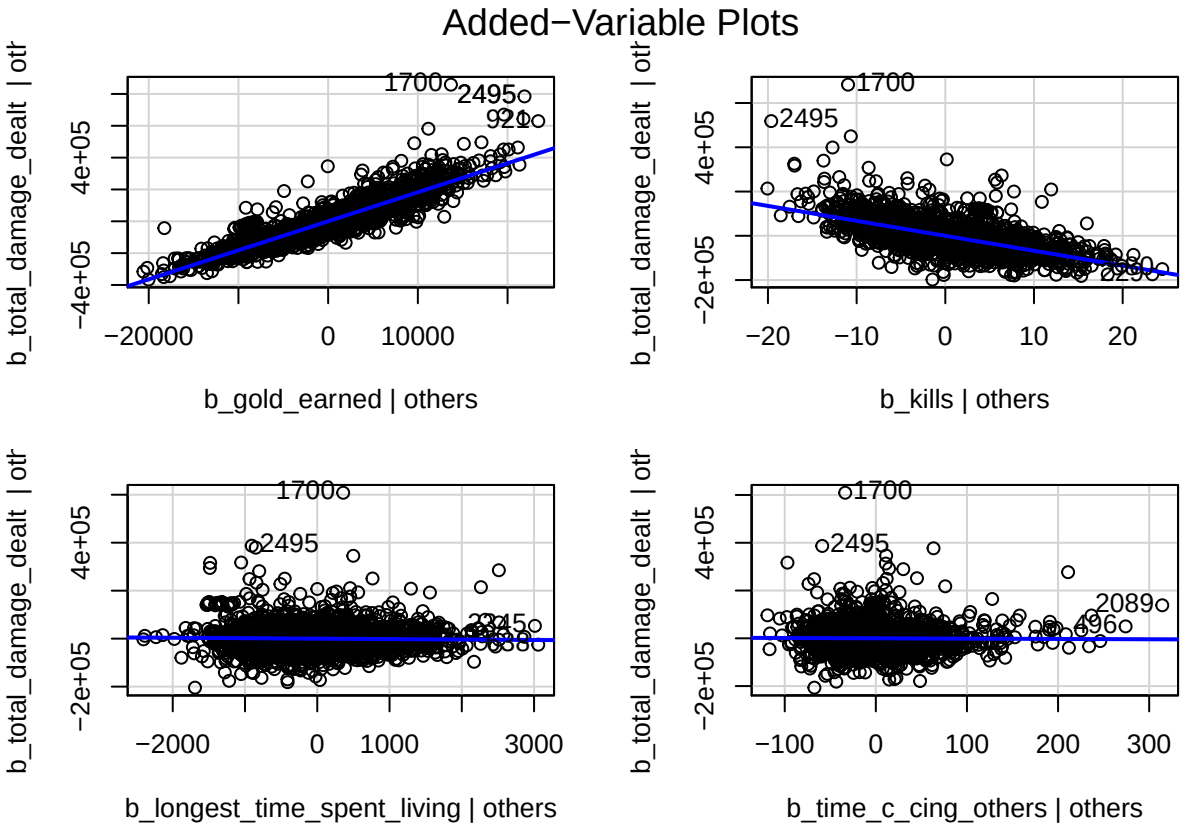



```
## [1] 1700 2495
```

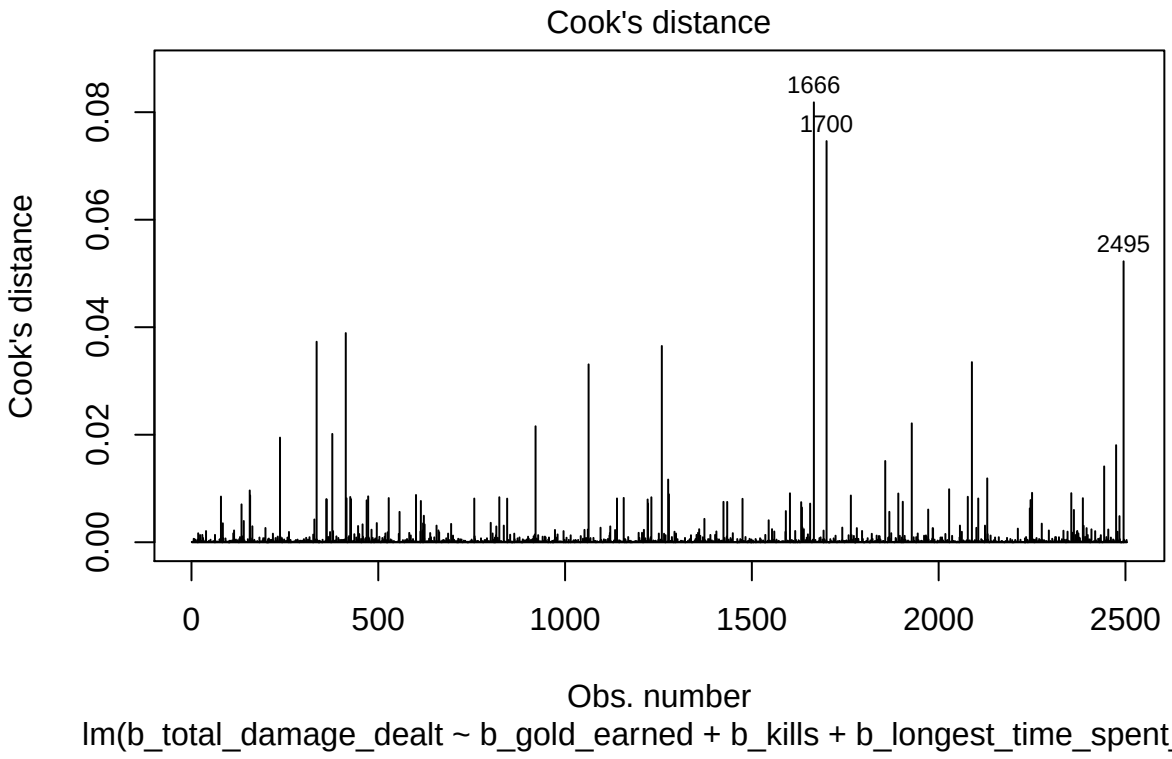
```
leveragePlots(selected_lm) # leverage plots
```



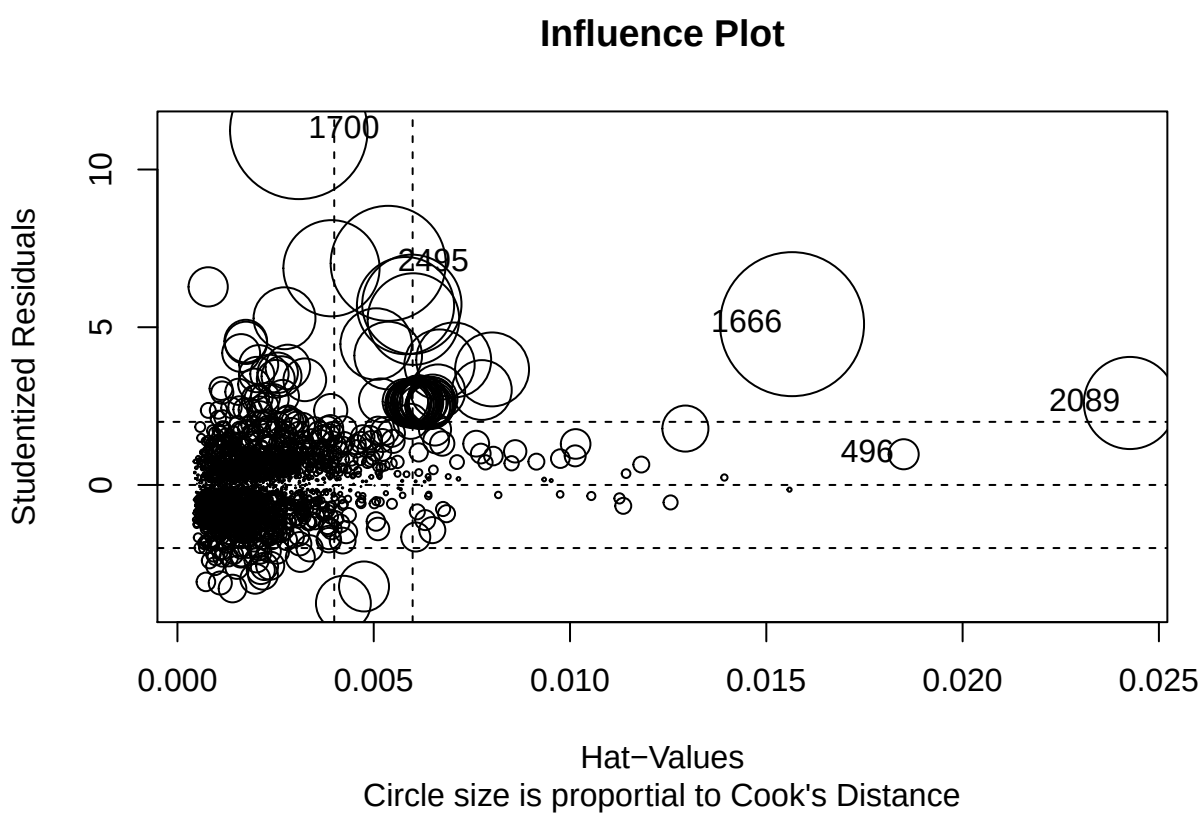

```
# Influential Observations
# added variable plots
avPlots(selected_lm)
```




```
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(lol10_test)-length(selected_lm$coefficients)-2))
plot(selected_lm, which=4, cook.levels=cutoff)
```

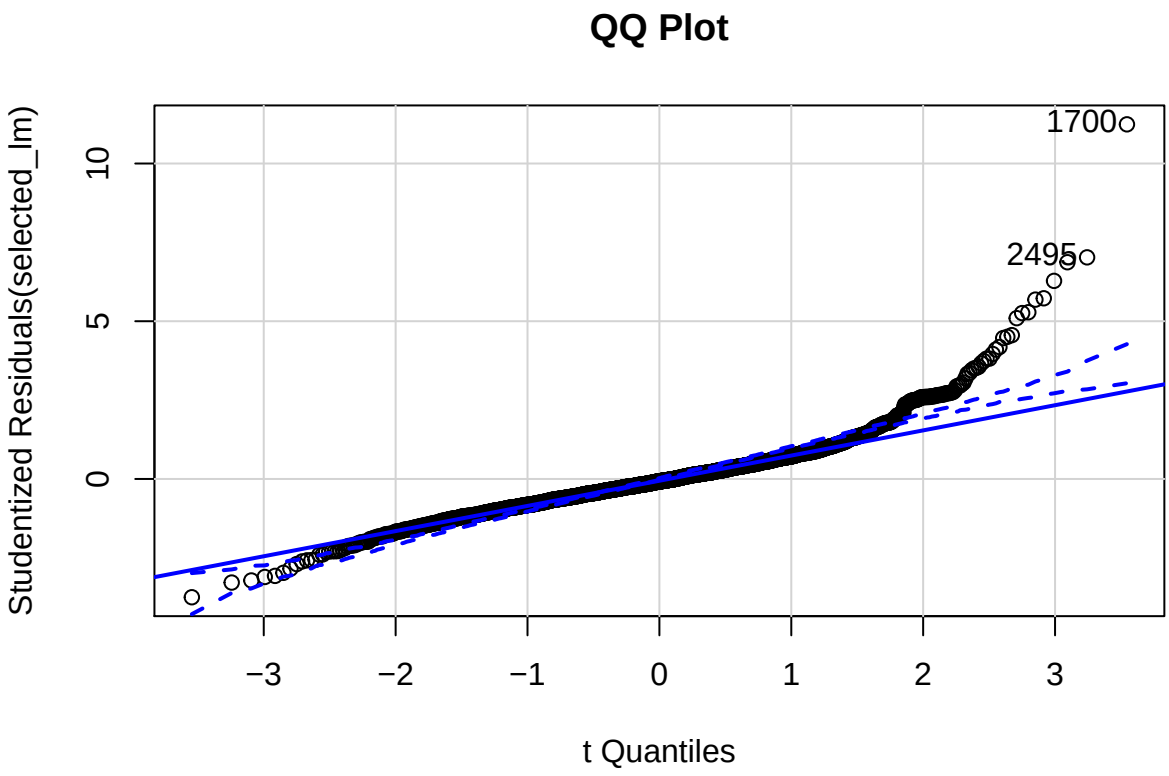



```
# Influence Plot
influencePlot(selected_lm, main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )
```



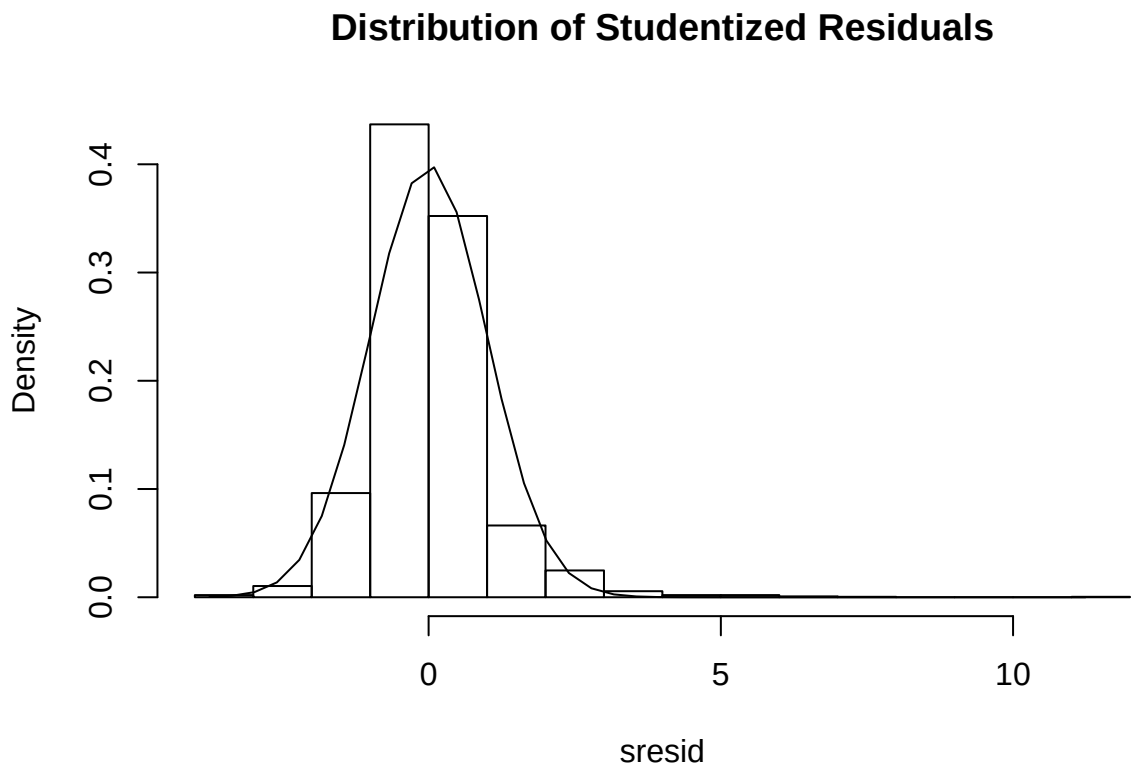

```
##          StudRes         Hat       CookD
## 496    0.9728662 0.018498547 0.003567732
## 1666   5.0975175 0.015654895 0.081833242
## 1700  11.2417253 0.003090876 0.074621180
## 2089   2.5985059 0.024262284 0.033502519
## 2495   7.0248676 0.005368642 0.052262053
```

```
# Normality of Residuals
# qq plot for studentized resid
qqPlot(selected_lm, main="QQ Plot")
```



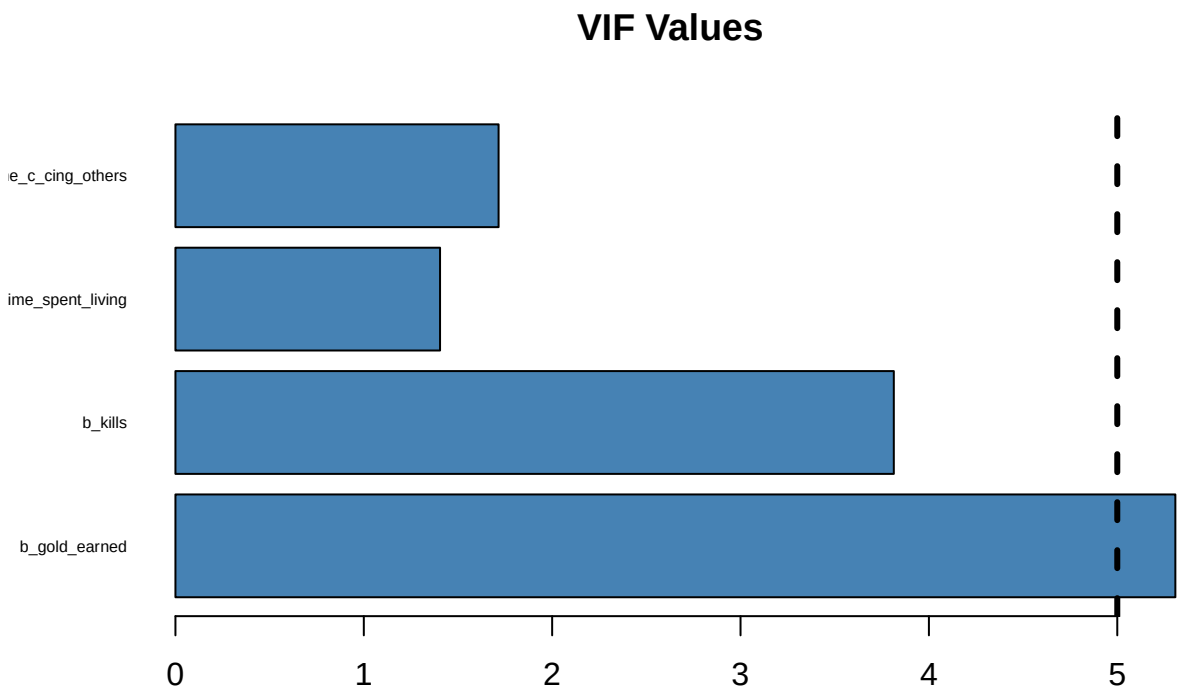

```
## [1] 1700 2495
```

```
# distribution of studentized residuals
sresid <- studres(selected_lm)
hist(sresid, freq=FALSE,
   main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```




```
# Evaluate Collinearity
vif(selected_lm) # variance inflation factors
```

```
##               b_gold_earned                     b_kills
##                    5.308366                    3.813736
## b_longest_time_spent_living       b_time_c_cing_others
##                    1.404775                    1.715489
```



-Interpret your β coefficients to the best of your ability.

-Report the R^2 and Adjusted-R^2 values on the test data.

-A complete analysis of the residuals and influence points.

-Try to give an interpretation of the model that makes sense.

-Give CIs for a mean predicted value and a future predicted value for at least one combination of X's (from your final linear model).

-Summarize your report.

-Add pretty language.

-Fix Variables and their descriptions table for MLR.

-Report on any relationships between the explanatory variables.

Bibliography

"Distribution of League of Legends (Lol) Summoners in North America as of October 2021, by Tier." 2021. *Statista*. https://www.statista.com/statistics/807298/league-of-legends-player-tier/.

Games, Riot. 2021a. "Riot Games Api." *Riot Developer Portal*. https://developer.riotgames.com/apis.

———. 2021b. *Twitter*. Twitter. https://twitter.com/riotgames/status/1455172784938651649?s=20&t=AQmQGrTa1ijf6u3cEDPZcg.

James. 2020. "League of Legends Ranked Match Data from Na." *Kaggle*. https://www.kaggle.com/jamesbting/league-of-legends-ranked-match-data-from-na.

Variable	Description
b _kills	The number of kills obtained by summoners on the blue side of the map.
b _gold_earned	The gold obtained by summoner 1 on the blue side of the map.
b _longest_time_spent_living	Sum of the longest time spent alive by summoners on the blue side of the map.
b_time_c_cing_others	The total time spend crowd controlling enemy players by summoners of the blue side of the map.
b_total_damage_dealt	Total damage done by summoners on the blue side of the map.

Figure 1: Variables and their descriptions