Project 3 158: Multiple Linear Regression

Tesfa Asmara and Kevin Loun

4/09/2022

Contents

1 Introduction **1**

2 Hypothesis **2**

3 Feature Engineering **2**

4 Interaction Variables **3**

5 Computational Model **3**

6 Statistical Model **4**

7 R^2 and $R^2_{a,p}$ **4**

8 Intepretation of the Coefficients **4**

9 Analysis of Residuals and Leverage **5**

10 Interpretation of the Model **7**

11 Confidence and Prediction Intervals **7**

12 Summary **8**

Bibliography **8**

1 Introduction

League of Legends (abbreviated LOL) is a multiplayer online battle arena video game developed and published by Riot Games. It has been one of the most popular video games since its release in 2009. In LOL, there are two teams of 5 players battle against each other in the Summoner's Rift, which is the main and most classical map. The goal is to be the first to destroy the opposing team's "Nexus", a structure located in the heart of each teams' base and protected by defensive towers. There are hundreds of champions with unique abilities waiting for players to choose from and to use forming various team compositions based on their strategies. Commonly, players collect gold by killing enemies and minions or destroy turrets. Players use the gold earned to purchase more powerful items and, thus, gain advantages in the following team fights.

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`.

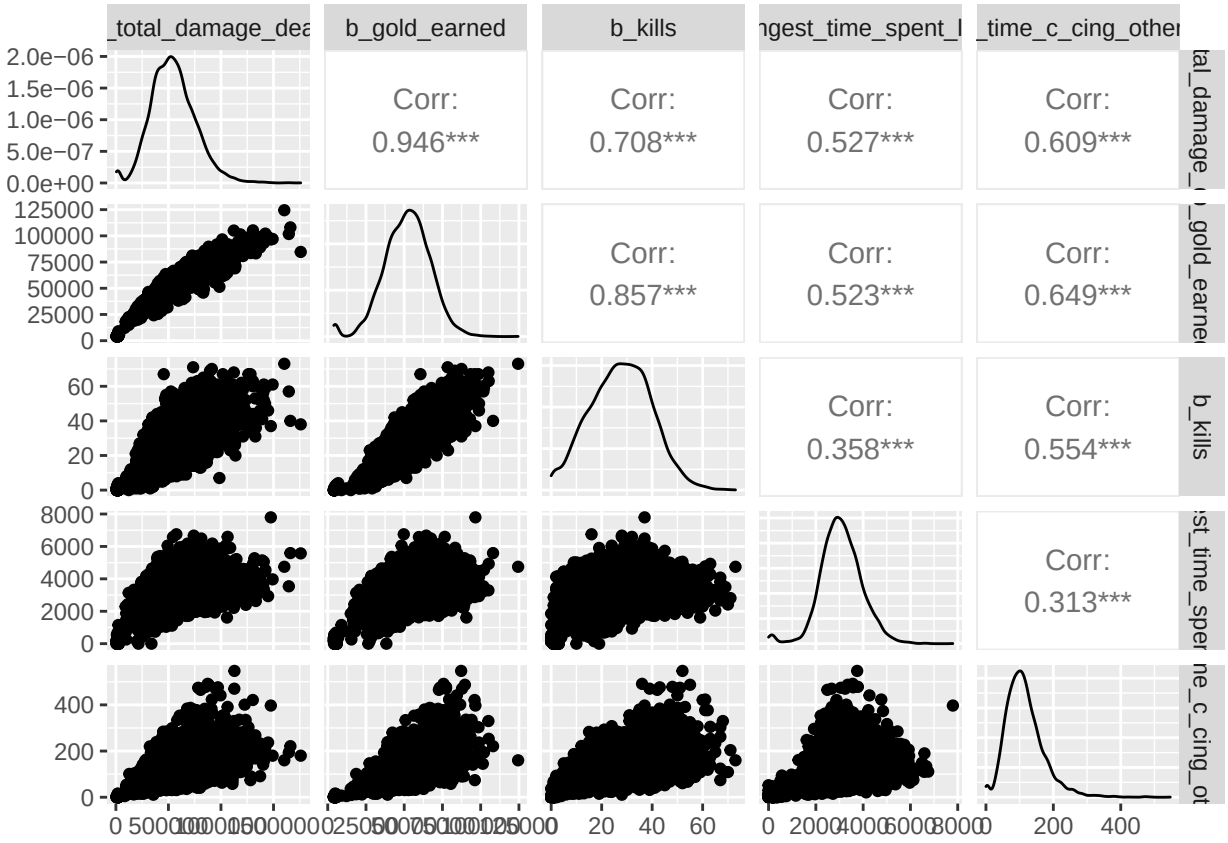
For this project, the following variables are of interest: time spent crowd controlling others, map side, longest time spent living, kills, gold earned, and total damage dealt. A figure including all the relevant variables and their description is attached at the end.

2 Hypothesis

By page 214 in ALSM, a single predictor variable in the model would have provided an inadequate description since a number of key variables affect the response variable in important and distinctive ways. Furthermore, in situations of this type, we frequently find that predictions ofthe response variable based on a model containing only a single predictor variable are too imprecise to be useful. A more complex model, containing additional predictor variables, typically is more helpful in providing sufficiently precise predictions of the response variable. The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We consider the following research question: Can we better understand the total damage dealt for the average Gold-ranked player on the blue team?

3 Feature Engineering



For $n = 7509$ 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. The test alternatives are: $H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$ and H_a : not all βs in H_0 are zero. 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,` and `b_time_c_cing_others` have been taken into consideration.

5 Computational Model

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

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

and ,

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

.

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 adjusted coefficient of multiple determination, $R^2_{a,p}$, criterion identifies several subsets of variables for which $R^2_{a,p}$ is high. When using $R^2_{a,p}$ as the decision criterion, we seek to eliminate or add variables depending on whether they lead to the largest improvement in $R^2_{a,p}$ and we stop when adding or elimination of another variable does not lead to further improvement in $R^2_{a,p}$. By this process, the four-parameter model ,

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

is identified as best; it has $\texttt{max}(R^2_{a,p}) = 0.9343514$ and will serve as the selected model.

7 R^2 and $R^2_{a,p}$

```
## [1] 45147232304
```

The coefficient of multiple determination, denoted by R^2, of a linear model measures the proportionate reduction of total variation in `b_total_damage_dealt` associated with the use of the set of variables `b_gold_earned, b_kills, b_longest_time_spent_living,` and `b_time_c_cing_others`. $R^2_{a,p}$ adjusts R^2 for the number of variables in the model by dividing each sum of squares by its associated degrees of freedom. R^2 is a biased estimate of the amount of variability explained by the model when applied to model with more than one predictor. To get a better estimate, we use $R^2_{a,p}$. $R^2_{a,p}$ describes the strength of a model fit, and it is a useful tool for evaluating which predictors are adding value to the model, where adding value means they are (likely) improving the accuracy in predicting future outcomes.

Calculated from the test data, the linear model we selected has $R^2 = 0.9321065$ and $R^2_{a,p} = 0.9319978$. This means that 93.2106472% 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`.

8 Intepretation of the Coefficients

The parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are sometimes called partial regression coefficients because they reflect the partial effect of one predictor variable when the other predictor variable is included in the model and is held constant. Let us now consider the meaning of the regression coefficients in the selected model.

The parameter $\beta_0 = -2.0687578 \times 10^5$ is the Y intercept of the regression plane. If the scope of the model includes $X_1 = X_2 = X_3 = X_4 = 0$, then $\beta_0 = -2.0687578 \times 10^5$ represents the mean response $\mathbb{E}\{Y\}$ at $X_1 = X_2 = X_3 = X_4 = 0$. Otherwise, β_0 does not have any particular meaning as a separate term in the regression model. The parameter β_1 indicates the change in the mean response $\mathbb{E}\{Y\}$ per unit increase in X_1 when X_2, X_3, X_4 is held constant. Likewise, β_2, β_3, β_4 indicates the change in the mean response per

unit increase in X_2 when X_1, X_3, X_4 is held constant, X_3 when X_1, X_2, X_4 is held constant, and X_4 when X_1, X_2, X_3 is held constant, respectively.

We can evaluate our coefficients based on their p-value to determine if they are significant. Upon closer inspection it appears that `b_longest_time_spent_living` and `b_time_c_cing_others` are not significant in our final model. They have p-values of 0.208 and .608 respectively. We set our α to be 0.05 and as such these coefficients are not significant, however, `b_gold_earned` and `b_kills` are significant as they have p-values lower than our threshold.

```
## # A tibble: 5 x 5
##   term                          estimate std.error statistic   p.value
##   <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                  -206876.    4744.      -43.6   2.17e-309
## 2 b_gold_earned                     18.2      0.171   107.    0
## 3 b_kills                        -6751.     178.      -37.9   2.20e-248
## 4 b_longest_time_spent_living       -1.89     1.50     -1.26  2.08e-  1
## 5 b_time_c_cing_others             -13.6     25.9      -0.525 6.00e-  1
```

9 Analysis of Residuals and Leverage

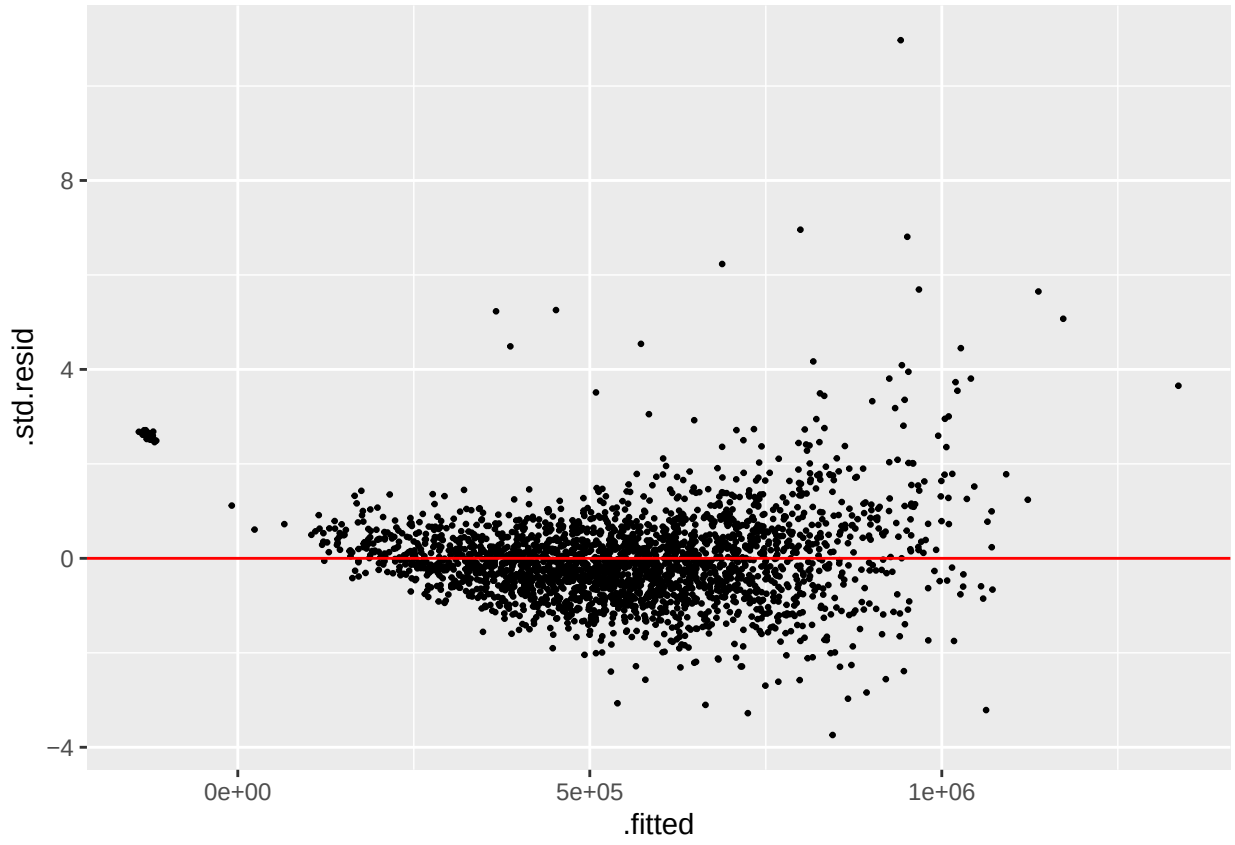


Figure 1: Studentized Residual Plot

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

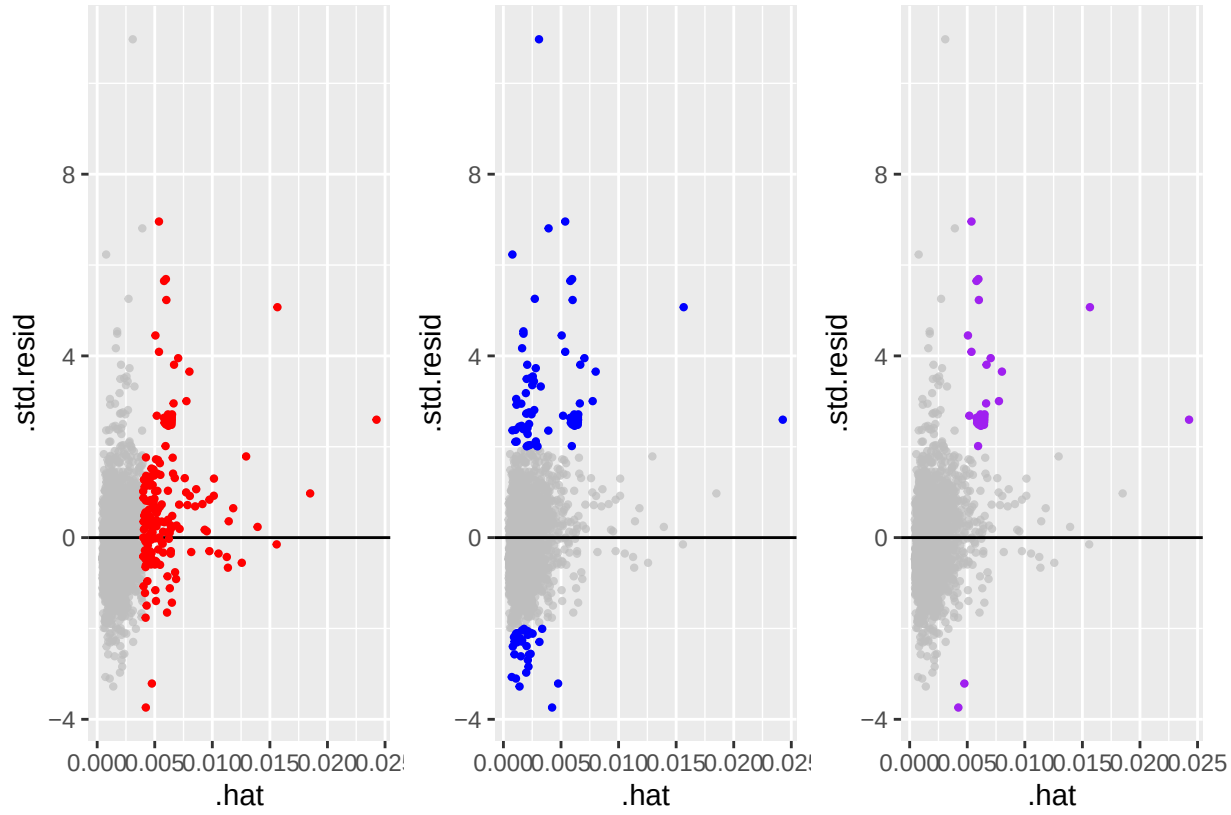


Figure 2: Left: Leverage Scatter Plot; Right: Outlier Scatter Plot

The diagonal elements, h_{ii}, of the hat matrix are a measure of the distance between the `b_gold_earned`, `b_kills`, `b_longest_time_spent_living`, and `b_time_c_cing_others` values for the ith case and the means of the `b_gold_earned`, `b_kills`, `b_longest_time_spent_living`, and `b_time_c_cing_others` values for all n cases. A leverage value h_{ii} is usually considered to be large if it is more than twice as large as the mean leverage value, denoted by $\bar{h} = \frac{p}{n}$. Approximately 7.5079872% of the cases have leverage values above the cut-off leverage of 0.0039936. Additionally, about 4.8322684% of the cases have semi-studentized residuals $e_i^* \notin (-2, 2)$ and none of the cases have semi-studentized residuals $e_i^* > 10$. We identify that about 1.956869% of the cases are outlying with respect to their Y values and their X values.

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

```
## Warning: Could not calculate the predicate for layer 2; ignored
```

```
## Warning: Could not calculate the predicate for layer 2; ignored

## Warning: Could not calculate the predicate for layer 2; ignored
```

After identifying cases that are outlying with respect to their Y values and their X values, the next step is to ascertain whether or not these outlying cases are influential. A useful measure of the influence that case i has on the fitted value Y_i is given by $(DFFITS)_i$.

The letters DF stand for the difference between the fitted value Y_i for the ith case when all n cases are used in fitting the regression function and the predicted value $Y_{i(i)}$ for the ith case obtained when the ith case is omitted in fitting the regression function. Of those cases that are outlying with respect to their Y values and their X values, 100% of their $(DFFITS)_i$ values exceed our guideline of $2\sqrt{\frac{p}{n}}$ for a medium-size data set.

Cook's distance measure, D_i, considers the influence of the ith case on all n fitted values. We consider 0% of the cases influential on the fit of the regression function because their D_i values exceed or equal 1.

The DFBETAS value by its sign indicates whether inclusion of a case leads to an increase or a decrease in the estimated regression coefficient, and its absolute magnitude shows the size of the difference relative to the estimated standard deviation of the regression coefficient. A large absolute value of $(DFBETAS)_{k(i)}$ is indicative of a large impact of the ith case on the kth regression coefficient. We consider 95.9183673% of the cases influential on β_0 because their $(DFBETAS)_{0(i)}$ values exceed $\frac{2}{\sqrt{n}}$. Likewise, we consider 95.9183673%, 36.7346939%, 95.9183673%, and 24.4897959% of the cases influential on $\beta_1, \beta_2, \beta_3, \beta_4$ because their $(DFBETAS)_{1(i)}, (DFBETAS)_{2(i)}, (DFBETAS)_{3(i)}, (DFBETAS)_{4(i)}$ values exceed $\frac{2}{\sqrt{n}}$, respectively.

All three influence measures (DFFITS, Cook's distance, and DFBETAS) did not identify a particular case, seeing as the size of the collection of cases that exceed the threshold for all of these tests is 0. Hence, the extent of the influence may not be large enough to call for consideration of remedial measures.

10 Interpretation of the Model

Our model is of type $p-1$ variables, where $p = 5$. We say $p-1$ instead of p because including the intercept there are p parameters in need of estimation. There were four independent variables, `b_gold_earned, b_kills, b_longest_time_spent_living,` and `b_time_c_cing_others`.

Initially we began with a model that included interactions between these variables but they were insignificant and were dropped from the model. These were most likely non-significant because changing one variable did not have a strong affect on any other variable, that is an increase or decrease in one variable did not result in an increase or decrease in any other variabilty. From the correlation matrix earlier, there seems to be a high correlation between `b_gold_earned` and `b_kills` which can indicate multicollinearity or that one of these variables can take the place of each other. To assess whether there really is multicollinearity we use the variance inflation factor for each variable. The variance inflation factor for each variable did not exceed our cutoff of 10, which is taken as an indication that multicollinearity may not be unduly influencing the least squares estimates.

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

11 Confidence and Prediction Intervals

We can now choose to look at the confidence intervals for the mean and individual response for a game with 34 `b_kills`, 368654 `b_gold_earned`, 216 `b_time_c_cing_others`, and 4392 `b_longest_time_spent_living`.

We found that the confidence interval for the mean response and individual response is and , respectively. This means that we are 95% confident that our true `b_total_damage_dealt` lies within this interval.

We can now choose to look at the confidence intervals for the mean and individual response at gold earned = 10.8676537, which is the median of our possible values for log(gold earned). We found that the confidence interval for the mean response and individual response is $(6.1694433 \times 10^6, 6.3712386 \times 10^6)$ and $(6.121801 \times 10^6, 6.4188809 \times 10^6)$, respectively. This means that we are 95% confident that our true `b_total_damage_dealt` lies within this interval.

12 Summary

Our final linear model is ,

$$\begin{aligned} \widehat{\text{b\_total\_damage\_dealt}} = -206875.78 + 18.22(\text{b\_gold\_earned}) - \\ 6750.84(\text{b\_kills}) - 1.89(\text{b\_longest\_time\_spent\_living}) - \\ 13.61(\text{b\_time\_c\_cing\_others}) \end{aligned} \tag{6}$$

,

.

We initially began by considering the following research question: Are one or more of the independent variables, `b_gold_earned`, `b_kills`, `b_longest_time_spent_living`, or `b_time_c_cing_others` in the model useful in predicting the future values of `b_total_damage_dealt`? We first began asnwering this question by assesing whether any feature engineering or interaction variables were neccessary in our model. We found that based on Table B.6 in ALSM that no feature engineering was needed. Through an F-test we found that non of the interaction variables were significant enough to add to our model. We then chose to approach the question through the use of a computational and statistical model. After comparing both models we concluded that the statistical model was the best choice for answering our question. We then proceeded to asses our data for any outliers that could affect our model; we accomplished this by using metrics such as: studentizeed residuals, cook's distance, dfbetas, and leverage. The outliers identified were not a result of data errors or misformulated regression so they were dropped from the data which in the end made the data more normal. Finally, a test for multicollinearity was applied due to a high correlation between variables. It was found that based on our cutoff value, no multicollinearity was present in our data.

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_total_damage_dealt	The total damage dealt by the blue side team.
b_gold_earned	The gold obtained by the blue side team.
b_kills	The kills obtained by the blue side team.
b_longest_time_spent_living	The longest time spent living obtained by the blue side team.
b_time_c_cing_others	The time spent crowd controlling others by the blue side team.

Figure 3: Variables and their descriptions