Math 158 Project Part 2: Simple Linear Regression

Anna Choi and Riley Elliott

Introduction

We scraped all of the data used in the following analysis from baseballsavant.com, the online depot for all publicly available advanced metrics on Major League Baseball players. Our sample is the 132 MLB hitters who had at least 502 plate appearances during the 2021 season. Our data set is a collection of these players' 2021 statistics. The population we will generalize our findings to is all MLB players with similar plate appearance quantities, regardless of year.

Our variables of interest in this study are average launch angle (LA) and isolated power (ISO). Launch angle is the vertical angle at which the ball leaves a hitter's bat, where 0° is parallel to the ground and 90° is straight up in the air. This average expresses the mean angle for all of a hitter's batted balls throughout the course of the 2021 season. Isolated power is meant to quantify how much power a hitter demonstrates during games. It calculates the rate at which they hit for "extra" total bases. Where a single is 1 total base, a double is 2, a triple is 3, and a home run is 4, ISO is calculated by $ISO = \frac{\text{(Total Bases) - (Singles)}}{\text{(At Bats)}}$. LA will be used as the explanatory variable, while ISO will be used as the response.

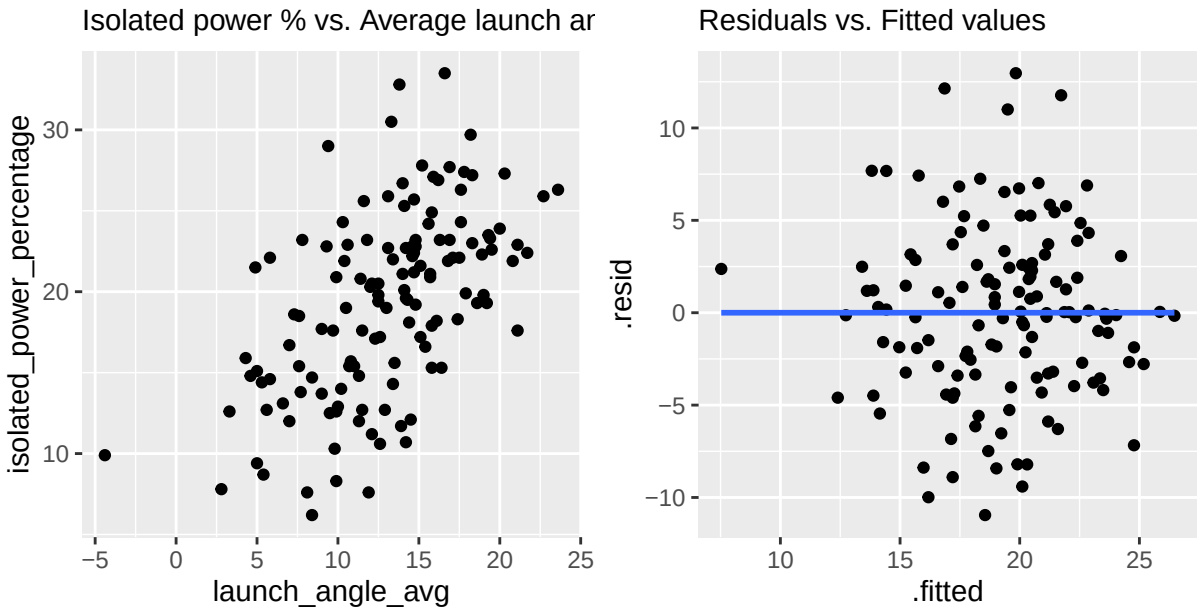
We have multiplied isolated power by a factor of 100 to express the rate as a percentage rather than a decimal. This was a cosmetic choice. We did this to make our graphs and slopes more presentable. We will refer to it as isolated power percentage.

Hypotheses

We hypothesize that average launch angle and isolated power have a positive linear relationship. This hypothesis is based on two related ideas. A high average launch angle means that a player is hitting more fly balls, which will generally travel further than ground balls because of ground friction. It is easier to hit a home run (worth 4 total bases) when hitting the ball further. Balls that travel far enough also end up in the outfield, where there are fewer fielders who are farther away from the bases. Hitting the ball to the outfield should make it easier for players to hit doubles (2 total bases) and triples (3 total bases). In running a linear regression test on these variables, we are essentially testing the degree to which launch angle can predict how much power a MLB batter demonstrates during games.

Checking the assumptions for linear regression

```
## geom_smooth: na.rm = FALSE, orientation = NA, se = FALSE
## stat_smooth: na.rm = FALSE, orientation = NA, se = FALSE, method = lm
## position_identity
```



Linearity

The data seems very linear. The fitted regression line cuts pretty much right through the middle of the data the whole way through.

Independent variables

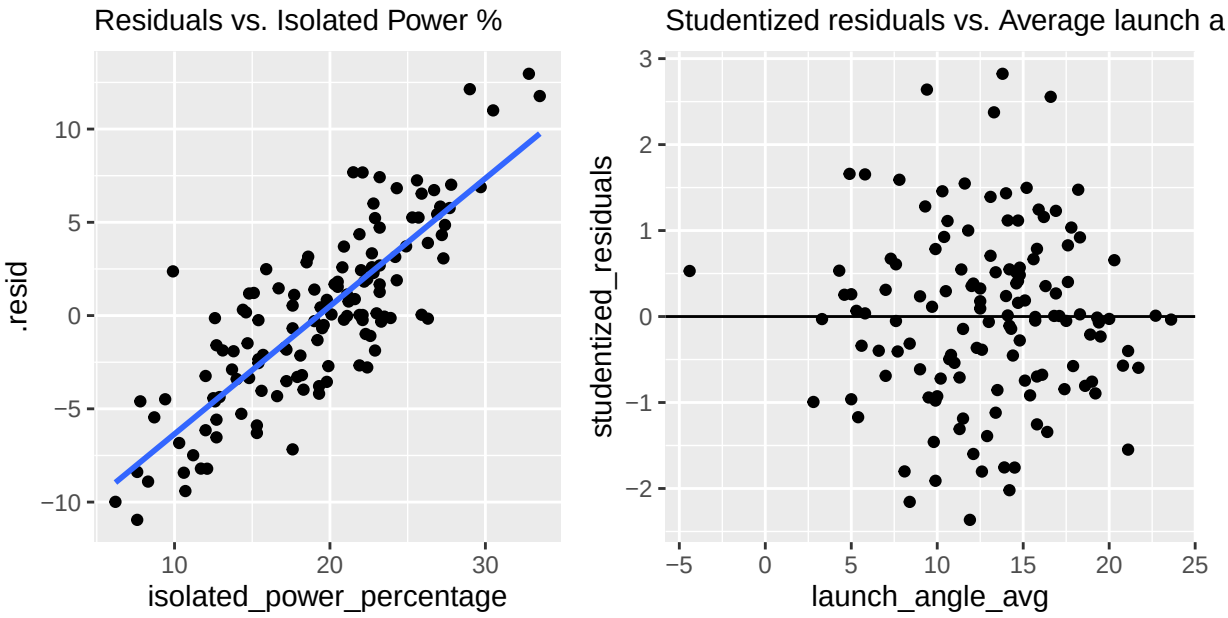
We know from the synthesis/gathering of this data that no data points (players) are repeated, and each player's statistics are necessarily independent from every other.

Normally distributed errors

The residuals appear to be distributed normally around 0. The concentration of points increases as it gets closer to 0, though the distribution is pretty wide (implying imperfect correlation). The concentration may be a bit thinner when we get to the larger-magnitude positive residuals compared to the larger-magnitude negative residuals, perhaps implying some skew. This issue appears to be minor, and might disappear with a larger sample size.

Equal Variance

The standard deviation of data points about the fitted regression line seems to be consistent no matter what the fitted value is. Homoscedasticity appears to be sound.



The stronger the correlation between average launch angle and isolated power percentage, the weaker the correlation between isolated power percentage and the residuals of the regression model. There will be some correlation in the latter regardless of explanatory/response correlation. There appears to be fairly strong correlation in the residual plot above, implying weak correlation between average launch angle and isolated power percentage.

Outliers

From plotting studentized residuals against our original explanatory variable, we can see that the former is limited to an absolute value of 3 for all data points, thus there are no clear outliers in the dataset.

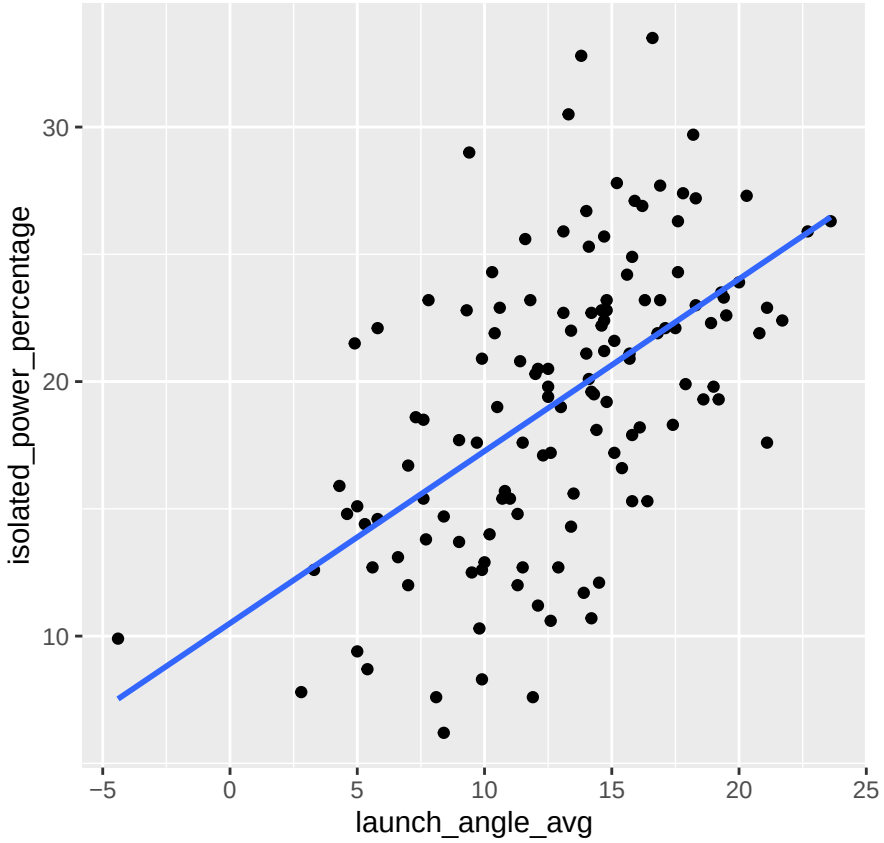
T-test for β_1

$H_0 : \beta_1 = 0$
$H_a : \beta_1 > 0$

```
## # A tibble: 2 x 7
##   term             estimate std.error statistic  p.value conf.low conf.high
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)        10.5      1.21        8.71 1.19e-14    8.12     12.9
## 2 launch_angle_avg    0.676    0.0875      7.73 2.62e-12    0.503     0.849
```

Since the p-value of our one-tailed test is $\frac{2.62\times10^{-12}}{2} = 1.31 \times 10^{-12}$ which is very small, we reject the null hypothesis that $\beta_1 = 0$, hence the slope coefficient must be greater than zero, i.e. there is a positive linear relationship between the isolated power percentage and average launch angle. A 95% confidence interval for the slope coefficient β_1 rounded to three significant figures is $(0.503, 0.849)$.

Plotting the variables




```
## # A tibble: 1 x 4
##   launch_angle_avg .fitted .lower .upper
##              <dbl>   <dbl>  <dbl>  <dbl>
## 1             13.0    19.3   18.4   20.1
```

We can say with 95% confidence that the mean isolated power percentage in our greater population for players with an average launch angle of 12.95° is between 18.448% and 20.077%. This range centers on 19.263%, our model-fitted value. We chose a 12.95° LA because that is the sample average.

```
## # A tibble: 1 x 4
##   launch_angle_avg .fitted .lower .upper
##              <dbl>   <dbl>  <dbl>  <dbl>
## 1             13.0    19.3   9.87   28.7
```

We can estimate that 95% of players in our population with an average launch angle of 12.95° will have an isolated power percentage between 9.867% and 28.658%. This is a massive range. The fitted value and the lower and upper bound are both roughly a factor of 2 away from each other. In practice, an ISO% of 9.867% is terrible and an ISO of 28.658% is fantastic. This indicates that the correlation between the two variables is not strong, and that not much of the variation in ISO% can be explained by average launch angle. We will test the correlation strength in the next section.

Assessing the fit of the model

```
## # A tibble: 2 x 6
##   term                df sumsq meansq statistic   p.value
##   <chr>            <int> <dbl>  <dbl>     <dbl>     <dbl>
```

```
## 1 launch_angle_avg     1 1337. 1337.       59.7  2.62e-12
## 2 Residuals          130 2910.   22.4      NA   NA
```

$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{2909.912}{4246.505} \approx 0.315$

This R^2 value indicates that only 31.5% of the variation in isolated power percentage can be explained by our linear model on average launch angle. This matches up with what we found when performing the prediction intervals and what we noticed while assessing the LINE conditions (the standard deviation is large).

Conclusion

Since we rejected the null hypothesis in favor of our alternative hypothesis (that there is a positive linear relationship between average launch angle and isolated power percentage), we expect that future MLB players recording 502+ plate appearances who have higher average launch angles will have a higher ISO. We also found, however, that the correlation between the variables was not strong, hence we would not feel comfortable predicting such a player's ISO based on their average launch angle alone.

A contributing factor for weak correlation may have been outliers, especially the observation on the far left of our original plot, which can affect R^2 greatly. Yet this argument is refuted by our plot of studentized residuals, which showed we did not have significant outliers. Nevertheless, limiting the range of the average launch angle to positive values to exclude the data point on the left may be a future consideration.

We feel confident in our assessment that the LINE conditions for this two-variable relationship hold, and that the relative lack of correlation can be best explained by the multitude of other variables that could affect a player's ISO. Some include average exit velocity, pull percentage, and outside zone swing percent. If we plotted ISO against a combination of these variables in a multiple linear regression analysis, we believe that we could create a better model.