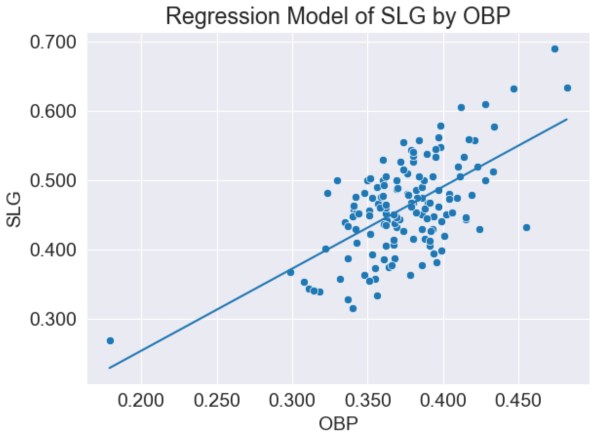
Problem 1

Using the `hofbatting.csv` file that you've come to know and love, conduct a regression analysis to determine if On-Base Percentage (OBP) can be used to predict Slugging Percentage (SLG).

(a) Test whether this model is providing useful information. In other words, is the slope non-zero? (If not, the model $y = \bar{x}$ could not be dismissed.)

- BLUF: There is enough evidence to determine that model is providing useful information. The OBP can be used to predict the SLG.
- Hypothesized model: $\hat{y} = 0.0175 + 1.182x$
- Scatterplot



- Parameter estimates: Intercept: $\hat{\beta}_0 = 0.0175$. Slope: $\hat{\beta}_1 = 1.182$.
- Coefficient of determination: $r^2 = 0.379$
- Hypothesis test
  - Hypotheses: $H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$
  - Significance: $\alpha = 0.05$
  - Test Statistic: $t = 9.399$
  - P-value: $p < 0.0001$
  - Technical Conclusion: We reject the null hypothesis and conclude the slope parameter, β_1, is not equal to zero.

(b) What is the expected slugging percentage for players with an OBP of 0.32? Give a confidence interval for $\alpha = 0.05$.

For OBP = .320, the expected SLG is .396, with a 95% confidence interval of [.340, .412].

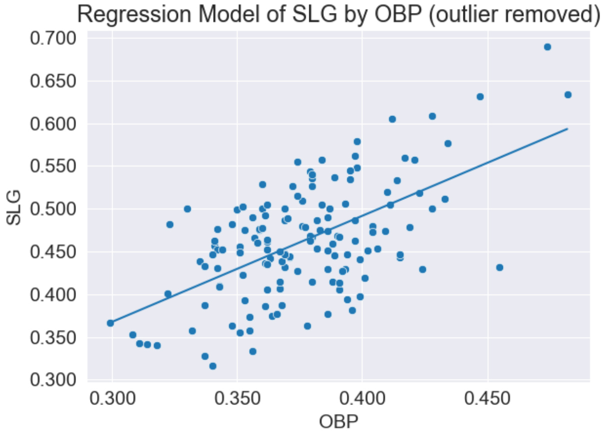
(c) What would you expect the slugging percentage of a new inductee with an OBP of 0.32 to be? Give a confidence interval for $\alpha = 0.05$.

For a new inductee with an OBP of .320, we expect the SLG to be .396, with a 95% prediction interval of [.290, .502].

(d) We previously identified Willard Brown as an outlier. Redo the analysis from the previous problems excluding his data. Did it make a difference?

- BLUF: There is enough evidence to determine that model is providing useful information. The OBP can be used to predict the SLG.

- Hypothesized model: $\hat{y} = -0.0041 + 1.239x$
- Scatterplot



- Parameter estimates: Intercept: $\hat{\beta}_0 = -0.0041$. Slope: $\hat{\beta}_1 = 1.239$.
- Coefficient of determination: $r^2 = 0.345$
- Hypothesis test
  - Hypotheses: $H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$
  - Significance: $\alpha = 0.05$
  - Test Statistic: $t = 8.704$
  - P-value: $p < 0.0001$
  - Technical Conclusion: We reject the null hypothesis and conclude the slope parameter, β_1, is not equal to zero.
- Estimation: For OBP = .320, the expected SLG is .392, with a 95% confidence interval of [.337, .410].
- Prediction: For a new inductee with an OBP of .320, we expect the SLG to be .396, with a 95% prediction interval of [.286, .499].

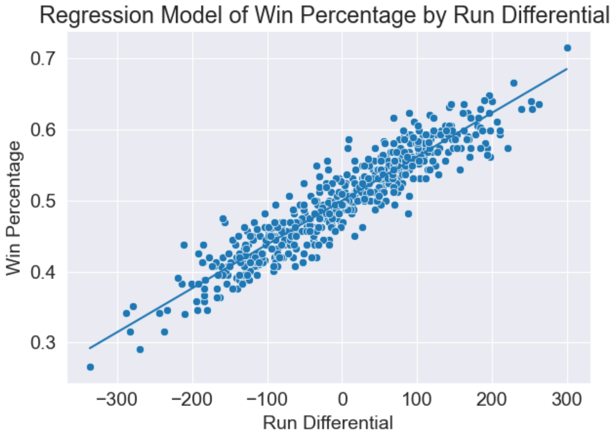
Problem 2

In baseball, it is hypothesized that we can use the run differential to predict the number of wins a team will have by the end of the season. Use the file `Teamdata.csv` to test this concept.

(a) Create a column of data for Run Differential (runs - runs allowed or $R - RA$) and a column for Win Percentage ($W/(W + L)$). Use these values to determine if the Run Differential can be used to predict the percentage of wins a team will end up with.

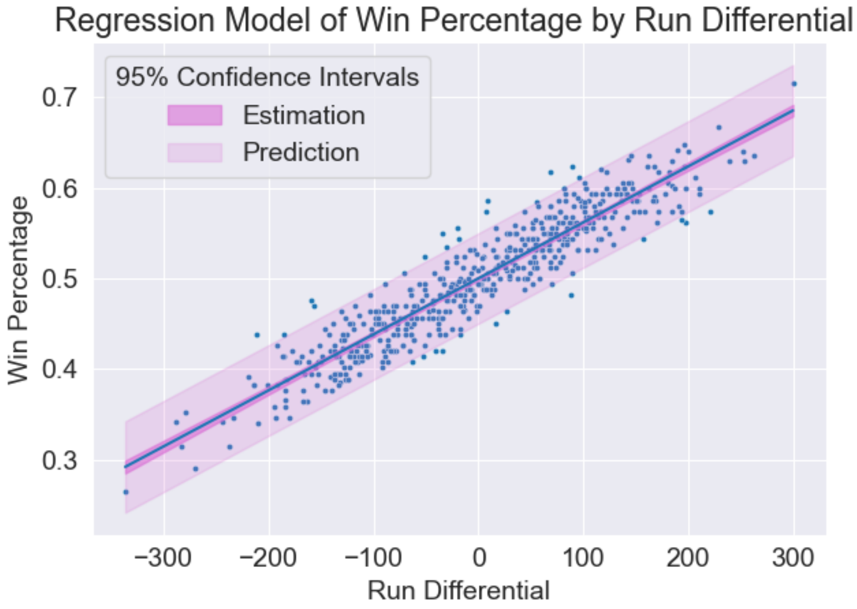
- BLUF: There is sufficient evidence to determine that the run differential can be used to predict the win percentage for a team.
- Hypothesized model: $\hat{y} = 0.5 + 0.0006x$

- Scatterplot



- Parameter estimates: Intercept: $\hat{\beta}_0 = 0.5$. Slope: $\hat{\beta}_1 = 0.0006$.
- Coefficient of determination: $r^2 = 0.875$
- Hypothesis test
  - Hypotheses: $H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$
  - Significance: $\alpha = 0.05$
  - Test Statistic: $t = 61.5$
  - P-value: $p < 0.0001$
  - Technical Conclusion: We reject the null hypothesis and conclude the slope parameter, β_1, is not equal to zero.

(b) Create a plot showing the confidence intervals for mean estimation and prediction overlaid on the original scatterplot. Reduce the dot size so that you can see the lines/zones.

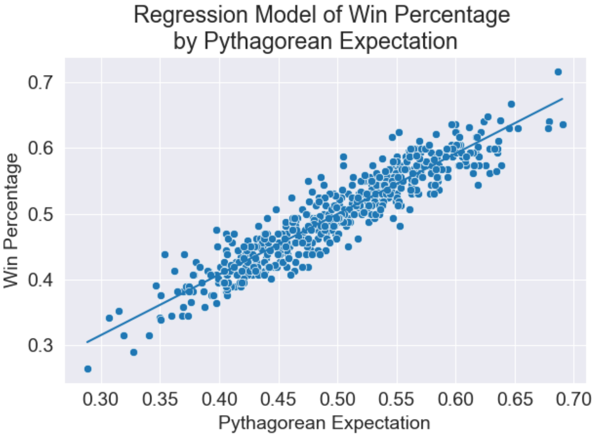


(c) Bill James, the godfather of sabermetrics, empirically derived a non-linear formula to estimate winning percentage called the Pythagorean Expectation.

$$W_{pct} = \frac{R^2}{R^2 + RA^2}$$

Create a new variable representing the Pythagorean Expectation. Now use this new column to replace the Run Differential and re-run your analysis.

- BLUF: There is sufficient evidence to determine that the Pythagorean expectation can be used to predict the win percentage for a team.
- Hypothesized model: $\hat{y} = 0.0394 + 0.9201x$
- Scatterplot



- Parameter estimates: Intercept: $\hat{\beta}_0 = 0.0394$. Slope: $\hat{\beta}_1 = 0.9201$.
- Coefficient of determination: $r^2 = 0.874$
- Hypothesis test
  - Hypotheses: $H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$
  - Test Statistic: $t = 61.13$
  - P-value: $p < 0.0001$
  - Technical Conclusion: We reject the null hypothesis and conclude the slope parameter, β_1, is not equal to zero.

(d) The 2001 Seattle Mariners had 116 wins and 46 losses with a +300 Run Differential in the data. Find this row in your data and pretend it's a new team with identical stats.

Use both the Run Differential and Pythagorean Expectation models to create confidence intervals for the Win Percentage expected for another team with identical performance to the Mariners.

Since we are pretending this is a new team, we want the prediction intervals. This new team has a run differential of 300 and a Pythagorean expectation of 0.686. Those values are used to predict the win percentage using their respective models. The results are shown in the table below with 95% prediction intervals.

Model	Expected Win Percentage	Lower Prediction Interval	Upper Prediction Interval
Run Differential	0.685	0.635	0.735
Pythagorean Expectation	0.671	0.620	0.721

(e) What are the pros and cons of each of these models (PE and RD)? Which would you rather use? Why? Think about practical usage as well as statistical accuracy.

A benefit of the Pythagorean Expectation (PE) model is that the PE value is very similar to the win percentage. So if all you want is a rough estimate (within a few percentage points), the PE value itself could be used without having to run through the model. A benefit of the Run Differential (RD) model is that RD is a commonly tracked and reported baseball stat, unlike PE. It is likely RD would be readily available so the RD model would be easier to use. If someone wanted to use the PE model, they would probably have to calculate the PE value first, and then use the model prediction (if they aren't using the PE value as a rough estimate). As far as statistical accuracy, the models perform very similarly. The coefficient of determination is almost identical. Figure 1 shows the residuals for both models next to each other. Both models produce similar residuals in both distribution and range.

Since the accuracy of both models is similar, I would use the Run Differential model because I think it is more practical.

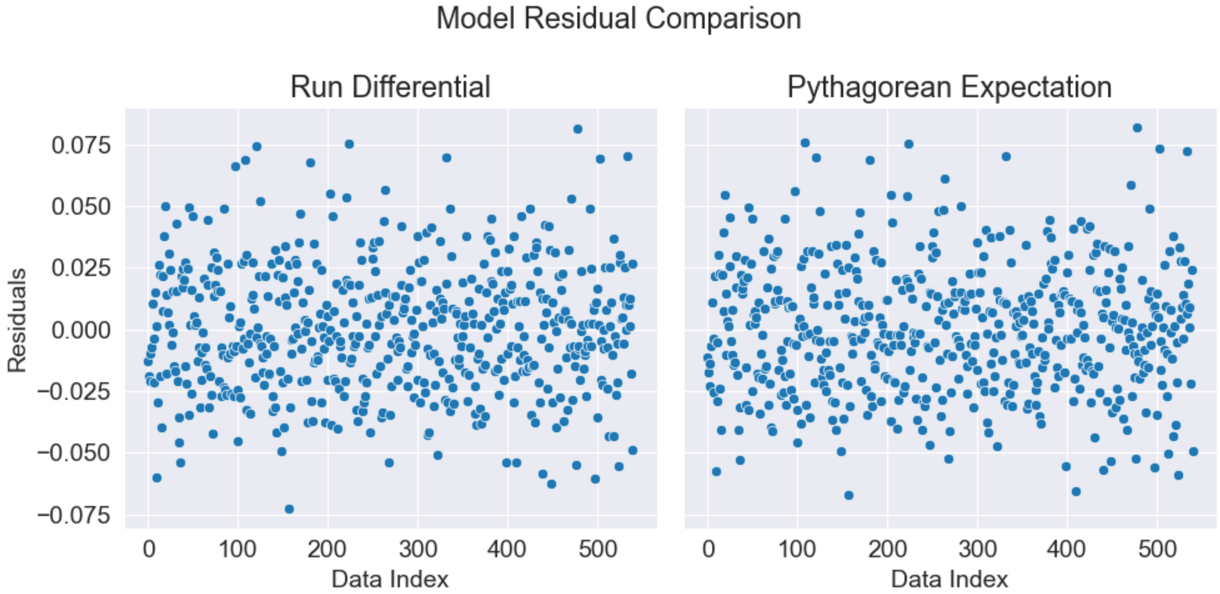


Figure 1