Lecture 2: Simple Linear Regression

Instructor: Jonathan Pipping *Authors: RB, JP, AW*

2.1 Motivation via Normal Random Variables

We'll begin by motivating linear regression through the decomposition of jointly distrubuted random variables. We begin with the case of jointly standard normal variables.

2.1.1 Jointly Standard Normal Random Variables

Definition 2.1 (Jointly Standard Normal Variables)**.** *Let (X, Y) be jointly normal random variables with*

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0, \quad \mathit{Var}(X) = \mathit{Var}(Y) = 1, \quad \mathit{Cov}(X, Y) = \rho$$

Then (X, Y) are jointly standard normal.

Theorem 2.2 (Decomposition of Jointly Standard Normal Variables)**.** *For jointly standard normal variables (X, Y) as defined above, we can write:*

$$Y = \rho X + \sqrt{1 - \rho^2} Z$$

where $Z \sim \mathcal{N}(0, 1)$ and $X \perp\!\!\!\perp Z$ (X is independent of Z).

Proof. Define Z as follows:

$$Z = \frac{Y - \rho X}{\sqrt{1 - \rho^2}} \tag{2.1}$$

We can verify that:

- $\mathbb{E}[Z] = 0$:

$$\begin{aligned} \mathbb{E}[Z] &= \frac{\mathbb{E}[Y] - \rho \mathbb{E}[X]}{\sqrt{1 - \rho^2}} \\ &= \frac{0 - \rho \cdot 0}{\sqrt{1 - \rho^2}} = 0 \end{aligned}$$

- $\mathrm{Var}(Z) = 1$:

$$\begin{aligned} \mathrm{Var}(Z) &= \frac{\mathrm{Var}(Y - \rho X)}{1 - \rho^2} \\ &= \frac{\mathrm{Var}(Y) + \rho^2 \mathrm{Var}(X) - 2\rho \mathrm{Cov}(X, Y)}{1 - \rho^2} \\ &= \frac{1 + \rho^2 - 2\rho^2}{1 - \rho^2} = 1 \end{aligned}$$

- $\text{Cov}(X, Z) = 0$:

$$\begin{aligned}\text{Cov}(X, Z) &= \frac{\text{Cov}(X, Y) - \rho\text{Var}(X)}{\sqrt{1-\rho^2}} \\ &= \frac{\rho - \rho}{\sqrt{1-\rho^2}} = 0\end{aligned}$$

Since Z and X are jointly normal and uncorrelated, they are independent. The result follows by solving for Y:

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

□

Theorem 2.3 (Best Linear Predictor)**.** *For jointly standard normal variables, the best linear predictor (least squares estimate) of Y given X is:*

$$\hat{Y} = \rho X$$

Proof. We aim to minimize the mean squared error:

$$\min_a \mathbb{E}\left[(Y - aX)^2\right]$$

Expanding, we have

$$\begin{aligned}\mathbb{E}\left[(Y - aX)^2\right] &= \mathbb{E}\left[Y^2 - 2aXY + a^2X^2\right] \\ &= \mathbb{E}[Y^2] - 2a\mathbb{E}[XY] + a^2\mathbb{E}[X^2] \\ &= 1 - 2a\rho + a^2\end{aligned}$$

Setting the derivative to zero, we have

$$\begin{aligned}\frac{\partial}{\partial a}\mathbb{E}\left[(Y - aX)^2\right] &= -2\rho + 2a = 0 \\ \implies a &= \rho\end{aligned}$$

So the best linear predictor is

$$\hat{Y} = \rho X$$

□

Theorem 2.4 (Conditional Expectation)**.** *For jointly standard normal variables, the conditional expectation $\mathbb{E}[Y \mid X]$ equals the least squares estimator:*

$$\mathbb{E}[Y \mid X] = \rho X$$

Proof. From the decomposition $Y = \rho X + \sqrt{1-\rho^2}Z$ where $X \perp\!\!\!\perp Z$:

$$\mathbb{E}[Y \mid X] = \rho X + \sqrt{1-\rho^2}\mathbb{E}[Z \mid X]$$

By the independence of X and Z, we have

$$\mathbb{E}[Y \mid X] = \rho X + \sqrt{1-\rho^2}\mathbb{E}[Z]$$

And since $Z \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}[Y \mid X] = \rho X + \sqrt{1-\rho^2} \cdot 0$$

So

$$\mathbb{E}[Y \mid X] = \rho X$$

□

We will now generalize this result to non-standard normal variables.

2.1.2 Generalization to Non-Standard Normals

Suppose now that:

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), \quad \text{Cov}(X,Y) = \rho\sigma_X\sigma_Y$$

Letting $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$, we define standardized versions of X and Y as:

$$X^* = \frac{X - \mu_X}{\sigma_X}, \quad Y^* = \frac{Y - \mu_Y}{\sigma_Y}$$

Then $(X^*, Y^*) \sim \mathcal{N}(0,1)$ with correlation ρ, and from Theorem 2.2 we have that

$$Y^* = \rho X^* + \sqrt{1-\rho^2}Z, \quad Z \sim \mathcal{N}(0,1),\ X^* \perp\!\!\!\perp Z$$

Multiplying both sides by σ_Y and adding μ_Y, we obtain

$$Y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \sigma_Y\sqrt{1-\rho^2}Z$$

And then similar to Theorem 2.4, we have that

$$\mathbb{E}[Y \mid X] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)$$

and the residual is independent Gaussian noise.

2.1.3 Regression When X Is Non-Normal

Suppose X is a standardized random variable with $\mathbb{E}[X] = 0$, $\text{Var}(X) = 1$, but X is not necessarily Gaussian. Let:

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

where $Z \sim \mathcal{N}(0,1)$ and $X \perp\!\!\!\perp Z$.

To compute the least squares estimate of Y given X, we will need the following theorem.

Theorem 2.5.

$$\arg\min_a \mathbb{E}[(Y - aX)^2] = \frac{Cov(X,Y)}{Var(X)}$$

The proof of this Theorem is left as an exercise.

Then we compute the covariance of X and Y as follows:

$$\begin{aligned}\text{Cov}(X, Y) &= \text{Cov}(X, \rho X + \sqrt{1-\rho^2}Z) \\ &= \text{Cov}(X, \rho X) + \text{Cov}(X, \sqrt{1-\rho^2}Z) \\ &= \rho\text{Var}(X) + \sqrt{1-\rho^2}\text{Cov}(X, Z) \\ &= \rho(1) + \sqrt{1-\rho^2} \cdot 0 \\ &= \rho\end{aligned}$$

And since $\text{Var}(X) = 1$, the optimal coefficient is:

$$a = \rho$$

Thus, the best linear predictor of Y given X is still:

$$\hat{Y} = \rho X$$

Furthermore, $\mathbb{E}[Y \mid X] = \rho X$ still holds, because:

$$\begin{aligned}\mathbb{E}[Y \mid X] &= \mathbb{E}[\rho X + \sqrt{1-\rho^2}Z \mid X] \\ &= \rho X + \sqrt{1-\rho^2}\mathbb{E}[Z] \\ &= \rho X + \sqrt{1-\rho^2} \cdot 0 \\ &= \rho X\end{aligned}$$

So even if X is non-Gaussian, the regression relationship remains the same due to the model structure.

2.2 A Model-Based Approach

We will now take a model-based approach to simple linear regression. We will start with a motivating example and then define the model formally.

2.2.1 MLB Batting Averages

Suppose we have access to each MLB player's 2020 and 2021 batting averages from the Lahman Database and no other information. How would we predict a player's 2021 batting average from their 2020 one?

Generally, a good idea is to start with exploratory data analysis. Let's plot the data and see what we learn.

What does the relationship look like?

- Somewhat linear with a positive slope
- Positive relationship: you'd expect that a higher 2020 BA is associated with a higher 2021 BA
- You might imagine drawing a best fit line through the points

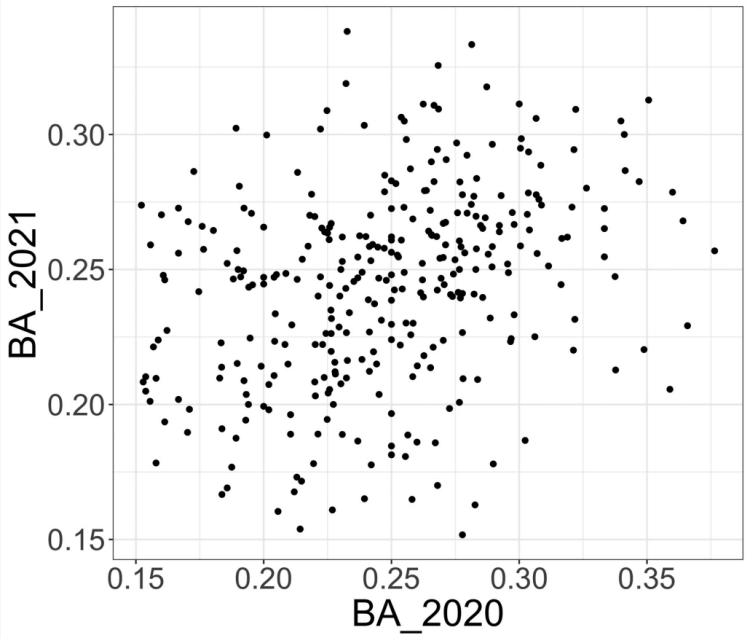

Figure 2.1: Batting Averages for MLB Players in 2020 and 2021

Let's plot that line of best fit!

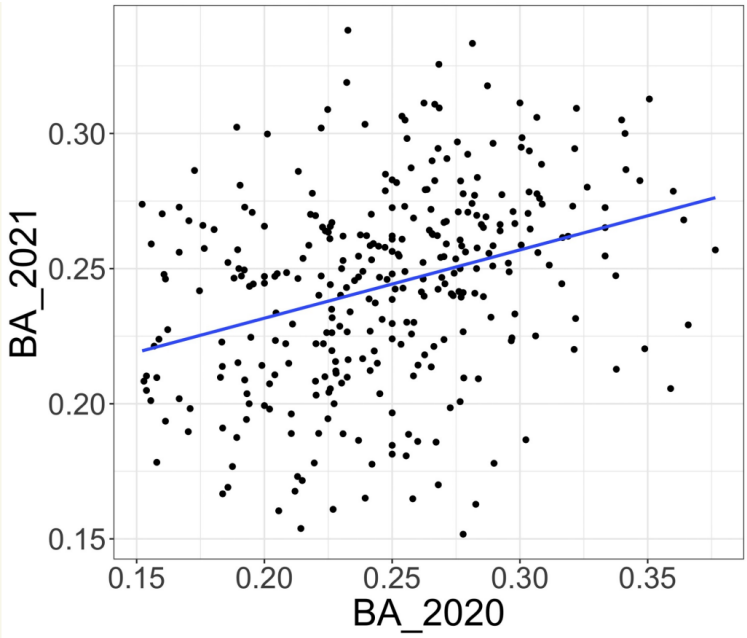

Figure 2.2: Line of Best Fit for MLB Batting Averages

This relationship isn't perfect, there's certainly some correlation but a lot of noise in the data. Still the question remains: how did we get this line? We **set up a model** to define this mathematically.

2.2.2 Setting Up the Model

Index each baseball player by $i = 1, \ldots, n$. Let $X_i = BA_i^{(2020)}$ be our independent (predictor) variable and $Y_i = BA_i^{(2021)}$ be our dependent (response) variable. Assuming a **linear relationship** between each X_i and its corresponding Y_i, we can write:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where β_0 is an unknown constant intercept, β_1 is an unknown constant slope, and ϵ_i is a random, independent, and identically-distributed error (or noise) term with mean 0 and constant variance σ^2. Mathematically, we can write

$$\epsilon_i \text{ i.i.d.}, \quad \mathbb{E}[\epsilon_i] = 0, \quad \text{and} \quad \mathbb{E}[\epsilon_i^2] = \sigma^2$$

We are interested in the conditional expectation of Y_i given X_i, or

$$\mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$$

This represents the "true" underlying line, but we don't know the values of β_0 and β_1. How do we **obtain estimates for these parameters** to obtain the "best fit" line?

2.2.3 Estimating Model Parameters

Definition 2.6 (Ordinary Least Squares)**.** *Find the values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the **Residual Sum of Squares (RSS)**, or the mean squared error.*

Definition 2.7 (Residual Sum of Squares)**.**

$$\begin{aligned} RSS(\beta_0, \beta_1) &= \sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i \mid X_i])^2 \\ &= \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 \end{aligned}$$

Visually, we can think of the RSS as the sum of the squared vertical distances between the observed points and the fitted line (or the squared residuals). In Figure 2.3, we plot the residuals for each point.

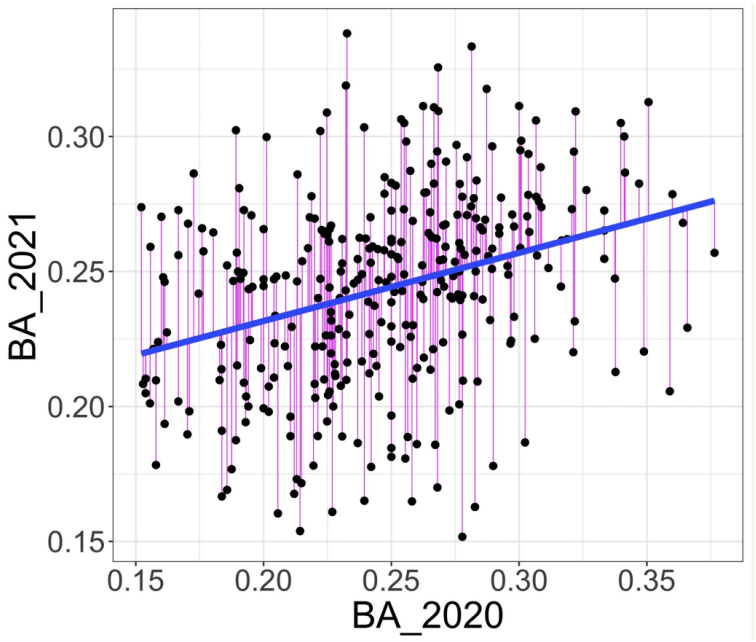

Figure 2.3: Residuals for MLB Batting Averages

Our objective is to find the intercept (β_0) and slope (β_1) which minimize the sum of squares of the lengths of the pink line segments. Mathematically, we write:

$$\begin{aligned}\widehat{\beta}_0, \widehat{\beta}_1 &= \underset{(\beta_0,\beta_1)}{\arg\min} \text{RSS}(\beta_0, \beta_1) \\ &= \underset{(\beta_0,\beta_1)}{\arg\min} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2\end{aligned}$$

We can solve this with calculus, setting the partial derivatives with respect to β_0 and β_1 to zero.

$$\begin{aligned}\frac{\partial}{\partial \beta_0} \text{RSS}(\beta_0, \beta_1) &= \sum_{i=1}^{n} -2(Y_i - (\beta_0 + \beta_1 X_i)) = 0 \\ \implies \frac{1}{n} \sum_{i=1}^{n} \beta_0 &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_1 X_i) \\ \implies \widehat{\beta}_0 &= \bar{Y} - \widehat{\beta}_1 \bar{X}^{\text{a}}\end{aligned} \tag{2.2}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_1}\text{RSS}(\beta_0, \beta_1) &= \sum_{i=1}^{n} -2X_i(Y_i - (\beta_0 + \beta_1 X_i)) = 0 \\
&\implies -\frac{1}{n}\sum_{i=1}^{n} X_i Y_i + \beta_0 \frac{1}{n}\sum_{i=1}^{n} X_i{}^{\text{b}} + \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 0 \\
&\implies \beta_1 \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y} \\
&\implies \widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned} \tag{2.3}$$

So we have closed-form expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. But what do these coefficients mean?

2.3 Interpreting Model Coefficients

2.3.1 Covariance

Recall the definition of Covariance from Lecture 1:

Definition 2.8 (Covariance)**.** *The covariance between two random variables X and Y is defined as*

$$\begin{aligned}
\sigma_{XY} = Cov(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

Note that if $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, then $\sigma_{XY} = \mathbb{E}[XY]$.

- Positive covariance: If when X is positive, Y tends to be positive (and when X is negative, Y tends to be negative), then $\sigma_{XY} > 0$
- Negative covariance: If when X is positive, Y tends to be negative (and when X is negative, Y tends to be positive), then $\sigma_{XY} < 0$

We proceed with some theorems that you will prove in your homework.

Theorem 2.9. *If X and Y are independent, then $\sigma_{XY} = 0$.*

Theorem 2.10. *The **sample covariance** between X and Y, defined as*

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

is an unbiased estimator of σ_{XY}.

Theorem 2.11. *The **sample variance** of X, defined as*

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an unbiased estimator of $\sigma_X^2 = Var(X)$.

We are now ready to define the correlation between two random variables.

[a] $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ are the sample means of the predictor and response variables, respectively.
[b] $\beta_0 \frac{1}{n}\sum_{i=1}^{n} X_i = \beta_0 \bar{X} = (\bar{Y} - \widehat{\beta}_1 \bar{X})\bar{X} = \bar{X}\bar{Y} - \beta_1 \bar{X}^2$, from Equation 2.2

2.3.2 Correlation

Definition 2.12 (Correlation)**.** *The correlation between two random variables X and Y is defined as*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

It is a normalized version of the covariance, and is always between -1 and 1.

Definition 2.13 (Sample Correlation)**.** *The sample correlation between X and Y is defined as*

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

We can use the sample correlation to measure the strength of the linear relationship between X and Y. Consider our estimate $\widehat{\beta}_1$ from Equation 2.3:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \cdot \frac{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \\
&= r_{XY} \cdot \frac{S_Y}{S_X}
\end{aligned}$$

This is the sample version of the formula we derived in Section 2.1.2.

We can also reciprocally express the sample correlation in terms of $\widehat{\beta}_1$:

$$r_{XY} = \widehat{\beta}_1 \frac{S_X}{S_Y}$$

So if X and Y have the same scale (meaning the same sample standard deviations), then the sample correlation is simply the slope of the linear regression line!

2.4 Remarks on Correlation

2.4.1 Correlation Can Be Misleading

Correlation is a measure of **linear association** between two variables. Figure 2.4 shows four plots with equal correlation, but they have **very** different relationships. What does this tell us?

Correlation can be **meaningless** if:

- The relationship is not linear at all
- There are extreme outliers in the data.

It is best to use correlation to describe data whose scatter plots are roughly elliptical (football-shaped). The lesson? **ALWAYS PLOT YOUR DATA!**

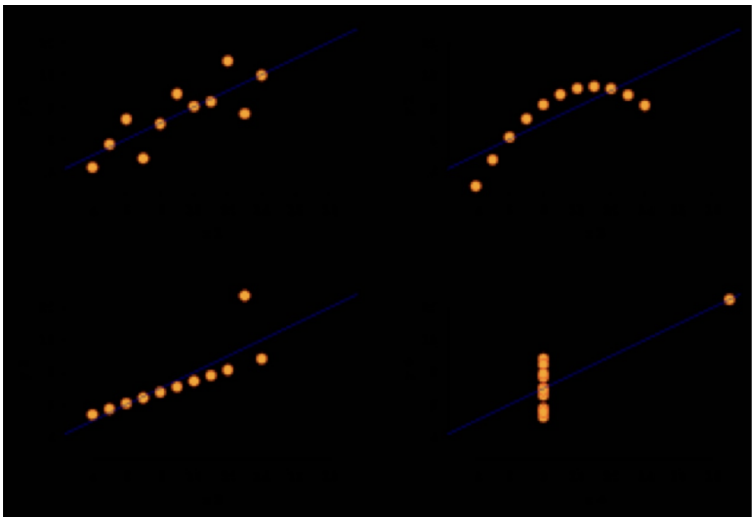

Figure 2.4: Four scatter plots with equal correlation

2.4.2 Case Study: In-Play Batting Average vs Batting Average

In Figure 2.5 below, we plot the MLB season averages for in-play batting average vs overall batting average for each season. It's clear when we plot the data that there is a different relationship between IPBA and BA before 1951 vs after 1951. We should consider splitting the data into two for our analysis.

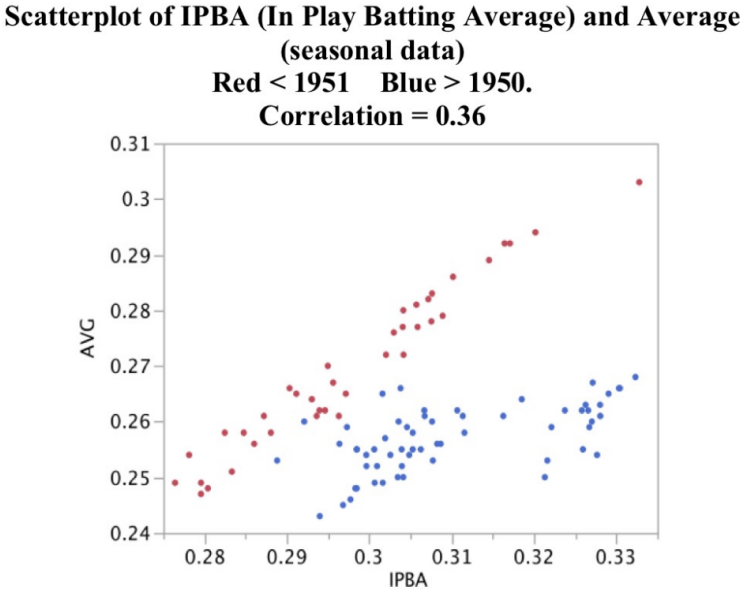

Figure 2.5: The relationship between in-play batting average and overall batting average is clearly different before and after 1951

In Figure 2.6, you can see that the data is much more tightly clustered in an elliptical shape when it is separated. Generally, this shape indicates the best type of data for prediction. Notice that the correlation

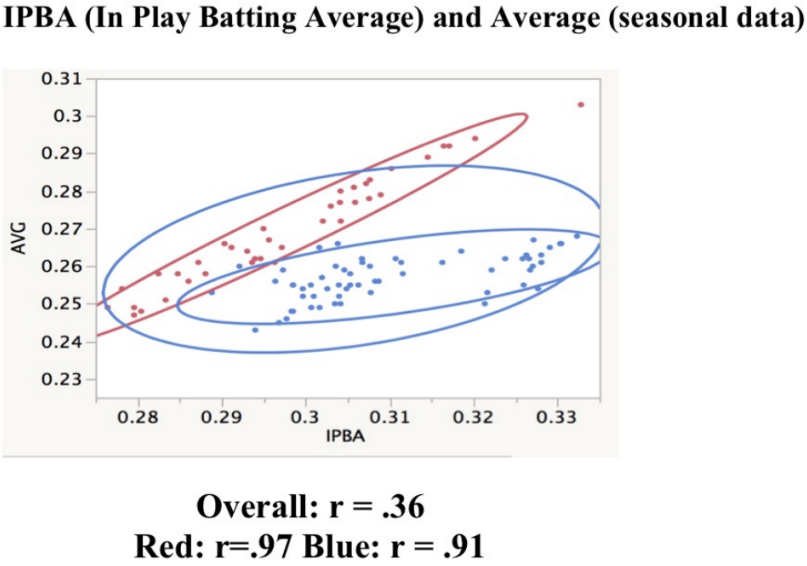

Figure 2.6: The correlation is much higher when the data is split

for either data (red or blue) is almost 3 times larger than the two together (0.97 and 0.91 vs 0.36).

2.5 Back to Our Batting Average Model

In Section 2.2.2, we defined our model as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where Y_i is the dependent variable, X_i is the independent variable, β_0 is the intercept, β_1 is the slope, and ϵ_i is the error term. Letting $X = BA^{(2020)}$ and $Y = BA^{(2021)}$, we now know that $\widehat{\beta}_1$ is a measure of how correlated $BA^{(2020)}$ is with $BA^{(2021)}$.

We use `R` to fit our linear regression model.

```
model = lm(BA_2021 ~ BA_2020, data = ba_data)
summary(model)
```

From this summary output, we can see that $\widehat{\beta}_1 = 0.25$. This means that a batting average increase of 0.020 in 2020 is associated with a predicted batting average increase of 0.005 in 2021.

2.6 Regression to the Mean

If $X_i > \bar{X}$, then $X_i = \bar{X} + \delta$, with $\delta > 0$. Then it follows that:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i \\ &= (\bar{Y} - \widehat{\beta}_1 \bar{X}) + \widehat{\beta}_1 X_i \\ &= \bar{Y} + \widehat{\beta}_1 (X_i - \bar{X}) \\ &= \bar{Y} + \widehat{\beta}_1 \delta\end{aligned}$$

In our example, $\widehat{\beta}_1 = 0.25$, therefore

$$\widehat{Y}_i = \bar{Y} + \frac{\delta}{4}$$

This is an example of **regression to the mean**. $BA_i^{(2020)}$ is δ greater than $\bar{BA}^{(2020)}$, but $\widehat{BA}_i^{(2021)}$ is only $\frac{\delta}{4}$ greater than $\bar{BA}^{(2021)}$. In other words, our prediction of player i's 2021 batting average is somewhere between their 2020 BA and the mean BA in 2021.

References

[Lahman] Friendly, M., Dalzell, C., Monkman, M., Murphy, et al., Lahman: Sean 'Lahman' Baseball Database, 2024. R package version 12.0-0.