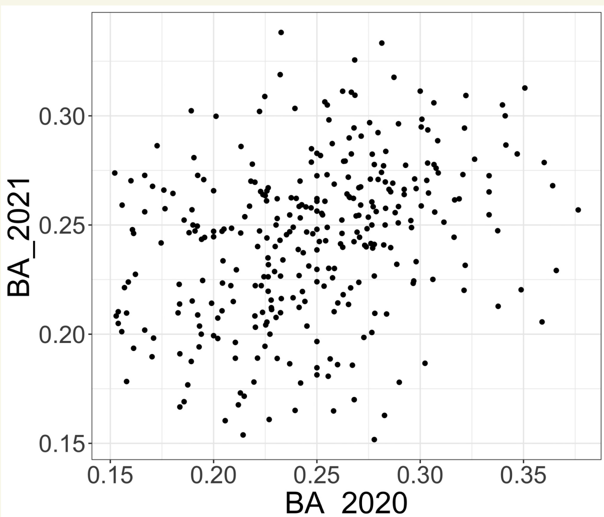
Simple Linear Regression

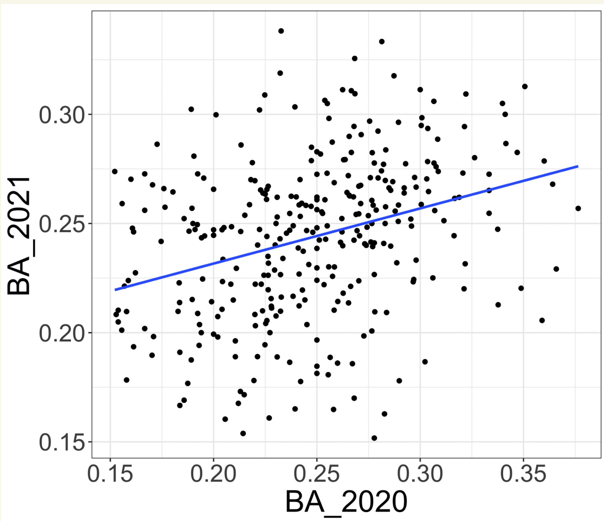
Q Suppose we have access to each MLB player's 2020 batting average and 2021 batting average and no other information.
Predict BA_{2021} from BA_{2020}.

Generally a good idea to begin with exploratory data analysis:



What does the relationship look like?

- Looks linear with a positive slope
  - can imagine drawing a best fit line through the points
  - positive slope (relationship) because, on average, you'd expect that a higher BA_{2020} is associated with a higher BA_{2021}



  - not the most perfect Relationship there is a lot of noise but, still some correlation

- So, how do we get this best fit line?

Model Index each baseball player by i

Let $X_i = BA_i^{(2020)}$ independent predictor variable

Let $Y_i = BA_i^{(2021)}$ dependent response variable

Assume a <u>linear relationship</u>

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

β_0 is an unknown constant intercept

β_1 is an unknown constant slope

ε_i is Random independent and identically distributed noise

$$\boxed{\begin{array}{ll} \mathbb{E}[\varepsilon_i] = 0 & \text{mean zero} \\ \varepsilon_i & \text{iid} \end{array}}$$

with unknown constant variance σ^2

- We are interested in the conditional expectation

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

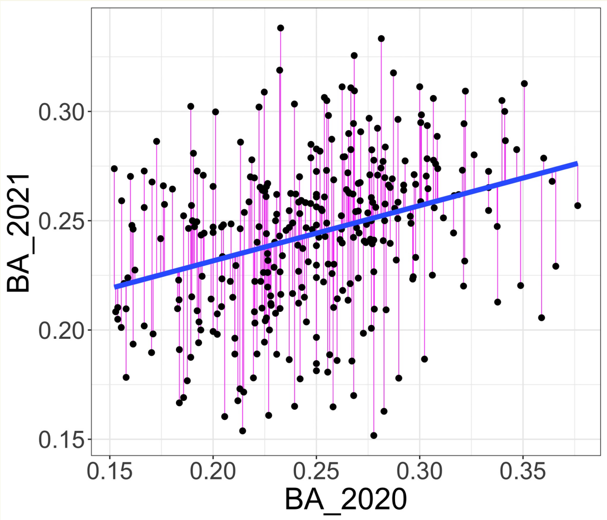
which is the "true" underlying line

- How do we estimate this best fit line? How to obtain estimates $\hat{\beta}_0, \hat{\beta}_1$ of β_0, β_1?

Ordinary Least Squares — find the values β_0, β_1 which minimize the Residual Sum of Squares (RSS) i.e. minimize the mean squared error,

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[Y_i - \mathbb{E}(Y_i | X_i)\right]^2$$

$$= \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2$$



Find the intercept β_0 and slope β_1 (i.e., the blue line) which minimizes the sum of the squares of the lengths of the pink line segments.

$$(\hat\beta_0, \hat\beta_1) = \underset{(\beta_0,\beta_1)}{\operatorname{argmin}} \ RSS(\beta_0, \beta_1)$$

$$= \underset{(\beta_0,\beta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Calculus: set the derivative equal to zero

$$\frac{\partial}{\partial\beta_0} RSS(\beta_0,\beta_1) = \sum_{i=1}^{n} (-2)(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\Longrightarrow \frac{1}{n}\sum_{i=1}^{n} \beta_0 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \beta_1 X_i)$$

$$\Longrightarrow \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\frac{\partial}{\partial\beta_1} RSS(\beta_0,\beta_1) = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

$$\Longrightarrow -\frac{1}{n}\sum Y_i X_i + \underbrace{\beta_0 \frac{1}{n}\sum X_i}_{} + \beta_1 \frac{1}{n}\sum X_i^2 = 0$$

$$= \beta_0 \bar{X}$$

$$= (\bar{Y} - \beta_1 \bar{X})\bar{X}$$

$$= \bar{X}\bar{Y} - \beta_1 \bar{X}^2$$

$$\Longrightarrow \beta_1 \left(\frac{1}{n}\sum X_i^2 - \bar{X}\right)^2 = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$

$\Longrightarrow$ by some algebra...

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Coefficient estimates

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

What do these formulas mean?

Covariance of two random variables X, Y is

$$\sigma_{XY} = \operatorname{cov}(X,Y) := \mathbb{E}\left[(X - \mathbb{E}X)\cdot(Y - \mathbb{E}Y)\right]$$

* If $\mathbb{E}X = 0$, $\mathbb{E}Y = 0$ then $\operatorname{cov}(X,Y) = \mathbb{E}[X \cdot Y]$

- positive covariance:
  if when X is positive Y is positive and when X is negative Y is negative then $\mathbb{E}(XY) > 0$
- negative covariance:
  if when X is negative Y is positive and when X is positive Y is negative then $\mathbb{E}(XY) < 0$

Thm X, Y independant $\Longrightarrow$ $\mathrm{cov}(X,Y) = 0$ (HW)

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

is an unbiased estimate of $\sigma_{xy} = \mathrm{cov}(X,Y)$ (HW)

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

is an unbiased estimate of $\sigma_x^2 = \mathrm{var}(X)$ (HW)

Correlation is normalized covariance,

Ranging from -1 to 1: $\rho_{xy} = \sigma_{xy} / \sigma_x \sigma_y$

Sample correlation $r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$

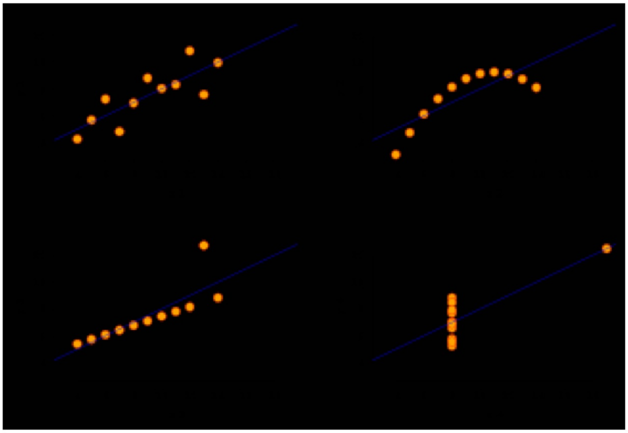
Then $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2} \sqrt{\sum_i (Y_i - \bar{Y})^2}} \cdot \frac{\sqrt{\sum_i (Y_i - \bar{Y})^2}}{\sqrt{\sum_i (X_i - \bar{X})^2}} = r_{xy} \frac{S_y}{S_x}$

$$\boxed{\hat{\beta}_1 = r_{xy} \frac{S_y}{S_x}} \qquad \boxed{r_{xy} = \hat{\beta}_1 \frac{S_x}{S_y}}$$

If X, Y have the same scale (sample variance) then the linear regression slope is the sample correlation!

correlation is a measure of linear association

Which of these is the strongest relationship? Which has the highest correlation?



The correlation would be a misleading summary statistic for the 3 graphs that are not football shaped.

Reminder: Correlation Warnings

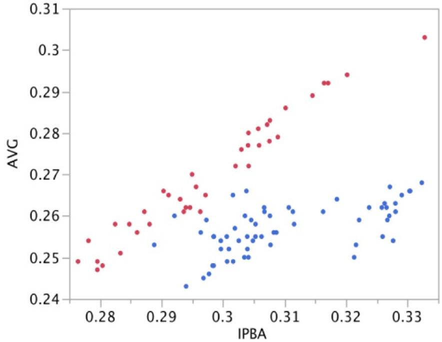
Correlation can be **meaningless** if

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

It is best to use the correlation to describe data whose scatter diagrams are **FOOTBALL** shaped.

ALWAYS PLOT YOUR DATA

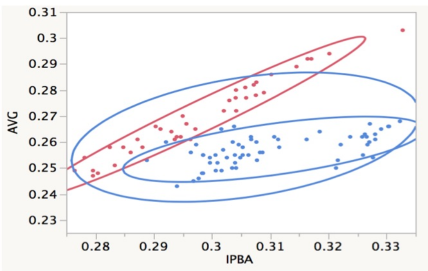
Scatterplot of IPBA (In Play Batting Average) and Average (seasonal data)
Red < 1951   Blue > 1950.
Correlation = 0.36



Each data point is the league average in a single season.

Here we can see that there is quite a different relationship between IPBA and Average before 1951 and after 1951; it may be best for our analysis to split up the data.

Scatterplot of IPBA (In Play Batting Average) and Average (seasonal data)



Overall: r = .36
Red: r=.97 Blue: r = .91

Scatterplot of IPBA (In Play Batting Average) and Average (seasonal data)

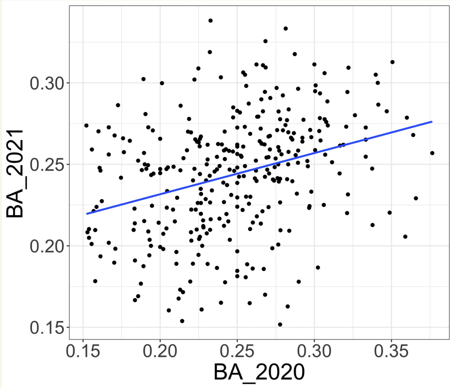
This scatterplot demonstrates how much more tightly the data clusters in the football shape when it is separated; traditionally, the football shape indicates the best type of data for prediction.

The correlation for either data (red or blue) is almost 3 times larger than the two together (.97 and .91 vs .36).

Back to our batting average model:

```r
m1 = lm(data=D1a, BA_2021~BA_2020)

plot_BA_2020_2021_3a = D1a %>%
  mutate(pred = predict(m1, .)) %>%
  ggplot(aes(x = BA_2020, y = BA_2021)) +
  geom_point(size=2) +
  geom_smooth(formula="y~x", method="lm", se=FALSE, linewidth=2)
plot_BA_2020_2021_3a
```



- So, $\hat{\beta}_1$ is a measure of how correlated BA_{2020} is with BA_{2021}

$\boxed{\hat{\beta}_1 = \frac{1}{4}}$

- $\hat{\beta}_1 = \frac{1}{4}$ Reflects that a batting average increase of .020 in 2020 is associated with a predicted batting average increase of .005 in 2021.

Regression to the Mean

If $X_i > \bar{X}$, so $X_i = \bar{X} + \delta$ with $\delta > 0$,

then $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = (\bar{y} - \hat{\beta}_1 \bar{X}) + \hat{\beta} X_i$

$= \bar{y} + \hat{\beta}_1 (X_i - \bar{X}) = \bar{y} + \hat{\beta}_1 \delta = \bar{y} + \frac{\delta}{4}$ $\left(\text{since } \hat{\beta}_1 = \frac{1}{4}\right)$

This is Regression to the Mean:

BA_i^{2020} is δ greater than $\overline{BA}^{2020}$ but $\widehat{BA}_i^{2021}$ is only $\frac{1}{4}\delta$ greater than $\overline{BA}^{2021}$.

In other words, our prediction of player i's batting average in 2021 is somewhere in between the 2020 BA and the mean BA.