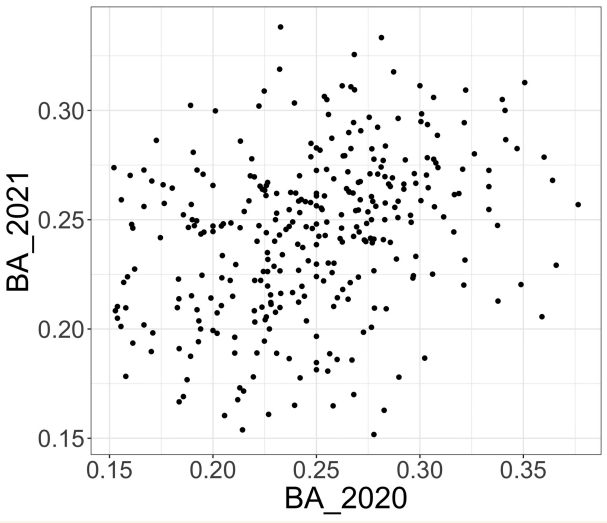
Regression Part 1: Simple Linear Regression

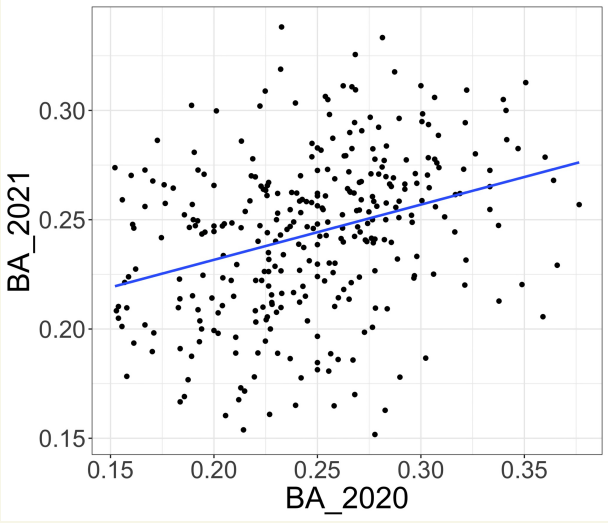
Q Suppose we have access to each MLB player's 2020 batting average and 2021 batting average and no other information.
Predict BA_{2021} from BA_{2020}.

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 linear relationship

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

β_0 is an unknown constant intercept

β_1 is an unknown constant slope

ε_i is Random independent and identically distributed noise

$$\mathbb{E}[\varepsilon_i] = 0 \quad \text{mean zero}$$
$$\varepsilon_i \quad \text{iid}$$

with unknown constant variance σ^2

- We are interested in the conditional expectation

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

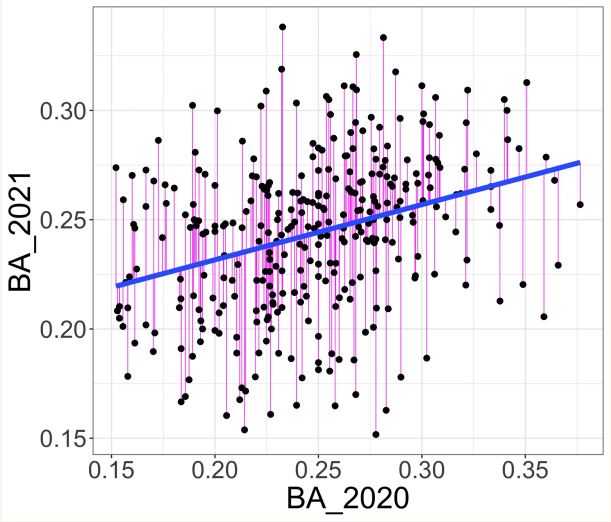
which is the "true" underlying line

- 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

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

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

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

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

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

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

$\beta_0 \frac{1}{n} \sum X_i$

$= \beta_0 \bar{x}$

$= (\bar{y} - \beta_1 \bar{x}) \bar{x}$

$= \bar{x}\bar{y} - \beta_1 \bar{x}^2$

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

$\Longrightarrow$ 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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

What do these formulas mean?

Covariance of two random variables X, Y is

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

Ex Suppose $\mathbb{E}X=0$ and $\mathbb{E}Y=0$

- 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 independent $\Rightarrow$ $\mathrm{Cov}(X,Y) = 0$

Pf $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}X)\cdot(Y-\mathbb{E}Y)]$

$= \mathbb{E}[X-\mathbb{E}X] \cdot \mathbb{E}[Y-\mathbb{E}Y]$ by independence

$= (\mathbb{E}X-\mathbb{E}X)\cdot(\mathbb{E}Y-\mathbb{E}Y)$ by linearity of $\mathbb{E}$

$= 0 \cdot 0 = 0$

Sample covariance $\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

need $n-1$ since we use observed data $\{x_i\}, \{y_i\}$
and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
to get an unbiased estimate $\mathbb{E}(\hat{\sigma}_{XY}) = \sigma_{XY}$

Correlation is normalized covariance, ranging from -1 to 1:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\text{VAR}(X)} = \sqrt{\mathbb{E}X^2 - (\mathbb{E}X)^2}$ is the standard deviation of X

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

sample correlation

$$r_{XY} = \frac{\hat{\sigma}_{XY}}{S_X S_Y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$$

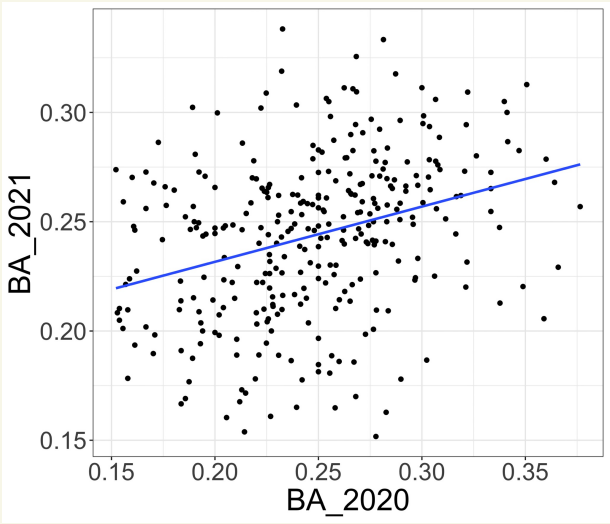
Then $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

$$= \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2} \sqrt{\sum_i (Y_i - \bar{Y})^2}} \cdot \frac{\sqrt{\sum_i (Y_i - \bar{Y})^2}}{\sqrt{\sum_i (X_i - \bar{X})^2}}$$

$$= r_{XY} \frac{S_Y}{S_X}$$

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

Insight: If X, Y are normalized (same variance) then the linear regression slope is simply the sample correlation!



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

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

Regression Line $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i$

$$\hat{Y}_i = \hat{E}(Y_i|X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$= (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 X_i$$

$$= \bar{y} + \hat{\beta}_1 (X_i - \bar{x})$$

If $X_i = \bar{x}$ then $\hat{Y}_i = \bar{y}$

→ $(\bar{x}, \bar{y})$ lies on the Regression Line!!

- Now, suppose $X_i > \bar{x}$

  i.e. $X_i = \bar{x} + \delta$ and $\delta > 0$

  Then $\hat{Y}_i = \bar{y} + \hat{\beta}_1 (X_i - \bar{x})$

  $= \bar{y} + \hat{\beta}_1 \delta$

  $= \bar{y} + \frac{\delta}{4}$ since $\hat{\beta}_1 = \frac{1}{4}$ in our example

This is **Regression to the Mean**:

BA_i^{2020} is δ greater than $\overline{BA}^{2020}$

but $\widehat{BA}_i^{2021}$ is only $\frac{1}{4}$ of 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.

Regression Part 1a: Transformations

Pythagorean Win Percentage

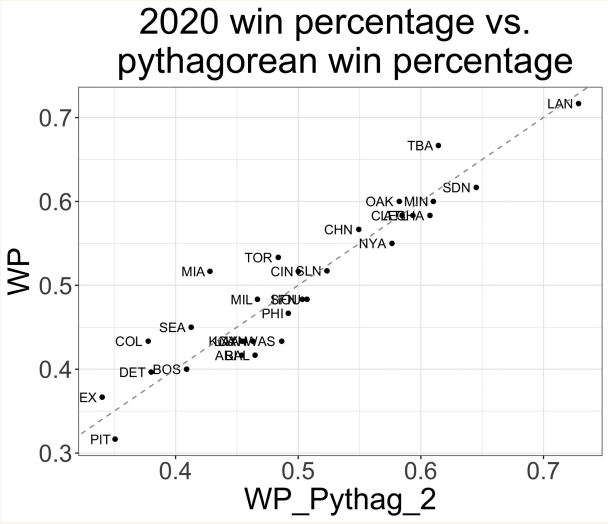
RA = runs allowed by a baseball team in one season.

RS = runs scored.

WP = a team's win percentage.

Bill James' Pythagorean Identity says

$$\widehat{WP} = \frac{RS^2}{RS^2 + RA^2}.$$



Pretty damn good!!

Q The pythagorean exponent is quite good, but is arbitrary. Can we find a better exponent α so that

$$\widehat{WP} = \frac{RS^{\alpha}}{RS^{\alpha} + RA^{\alpha}} \quad ??$$

Using $\{RS, RA, WP\}$ data from the 2020 MLB season, estimate α.

Idea Transform $\mathbb{E}(WP) = \frac{RS^\alpha}{RS^\alpha + RA^\alpha}$ into something that looks linear ($Y = \beta_0 + \beta_1 X$) and then try Linear Regression.

$$\mathbb{E}(WP) = \frac{RS^\alpha}{RS^\alpha + RA^\alpha} = \frac{1}{1 + \left(\frac{RA}{RS}\right)^\alpha}$$

$$\Longrightarrow \quad 1 + \left(\frac{RA}{RS}\right)^\alpha = \frac{1}{\mathbb{E}(WP)}$$

$$\Longrightarrow \quad \left(\frac{RA}{RS}\right)^\alpha = \frac{1}{\mathbb{E}(WP)} - 1 = \frac{1 - \mathbb{E}(WP)}{\mathbb{E}(WP)}$$

$$\Longrightarrow \quad \alpha \log\left(\frac{RA}{RS}\right) = \log\left(\frac{1 - \mathbb{E}(WP)}{\mathbb{E}(WP)}\right)$$

So, let $\mathbb{E}Y = \log\left(\frac{1 - \mathbb{E}(WP)}{\mathbb{E}(WP)}\right)$, $X = \log\left(\frac{RA}{RS}\right)$,

$\beta_1 = \alpha$, and $\beta_0 = 0$.

Then we have $\mathbb{E}Y = \beta_1 \cdot X$

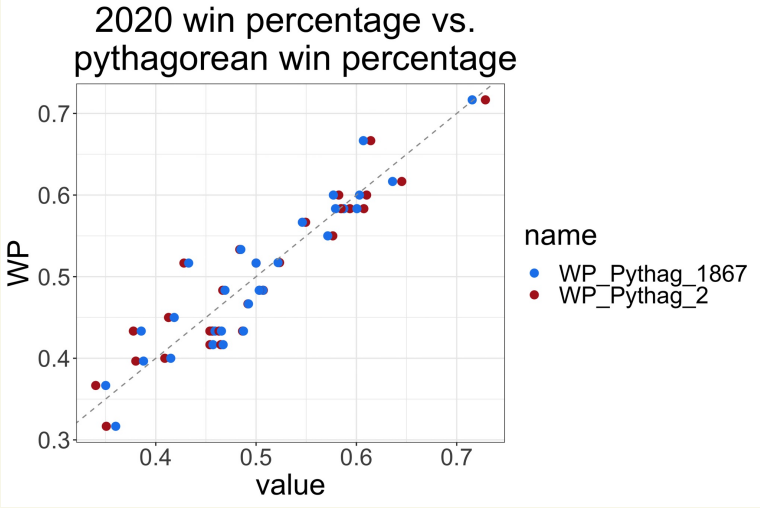
Use linear regression!

```
> m2 = lm(data=data_train, log((1-WP)/WP) ~ log(RA/RS) + 0 )
> m2

Call:
lm(formula = log((1 - WP)/WP) ~ log(RA/RS) + 0, data = data_train)

Coefficients:
log(RA/RS)
     1.867
```

$\alpha = 1.867$

using 2020 data

pretty close to 2!



pretty similar, but blue is slightly better since it is closer to the line $y = x$.

Lesson: Even with a nonlinear relationship like $WP = \frac{RS^{\alpha}}{RS^{\alpha} + RA^{\alpha}}$, we can still use linear Regression by transforming our data!