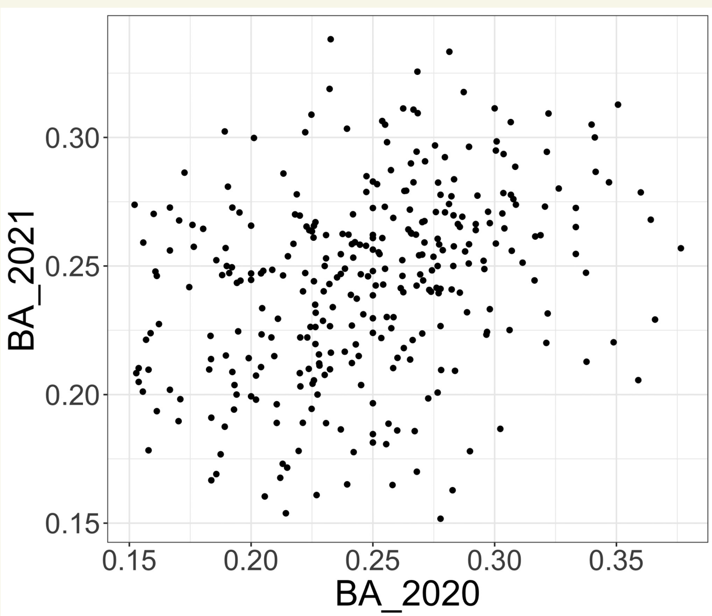
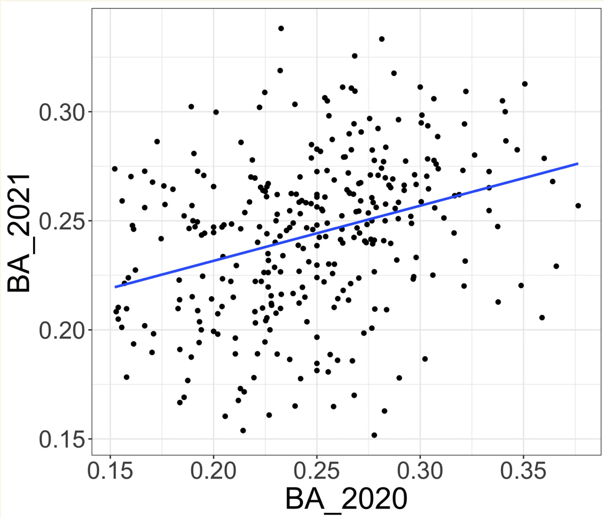
Simple Linear Regression

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

Always good to begin with plotting visualization,



Looks linear-ish
Positive slope
can imagine drawing a best fit line



How do we get a best fit line?

Model index each baseball player in the dataset by i

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

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

Assume a linear Relationship

$$\boxed{Y_i = \underbrace{\beta_0 + \beta_1 \cdot X_i}_{b+mx} + \underbrace{\varepsilon_i}_{\text{noise}}}$$

ε_i is a random variable (noise)
with $\mathbb{E}\,\varepsilon_i = 0$

β_0, β_1 are unknown constants

goal is to estimate β_0, β_1
to get the best fit line

$$\hat{Y}_i := \mathbb{E}(Y_i \mid X_i) = \mathbb{E}(\beta_0 + \beta_1 X_i + \varepsilon_i)$$

$$= \beta_0 + \beta_1 X_i + \cancelto{0}{\mathbb{E}(\varepsilon_i)}$$

$$\approx \hat{\beta}_0 + \hat{\beta}_1 X_i$$

How to estimate β_0 and β_1 ?

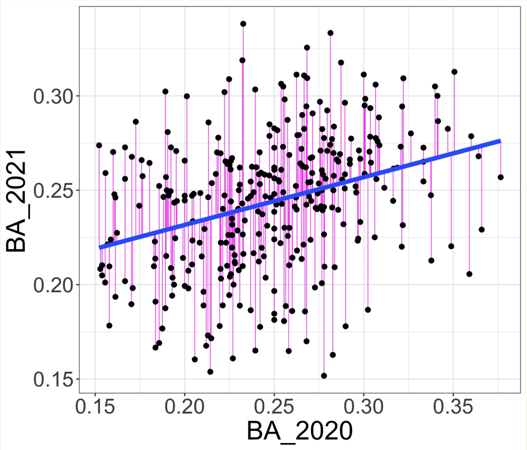
Ordinary least squares (OLS) — choose the values of β_0, β_1 which minimize the Residual Sum of Squares

i.e. minimize mean square error (MSE),

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2$$

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

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{(\beta_0, \beta_1)}{\arg\min}\ RSS(\beta_0, \beta_1)$$



$$\hat{\beta}_0, \hat{\beta}_1 = \underset{(\beta_0, \beta_1)}{\arg\min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Calculus: to minimise a function, set the derivative equal to 0 and solve.

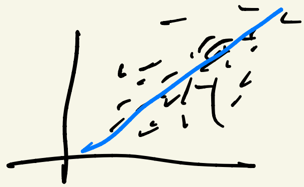
$$\begin{cases} \dfrac{d}{d\beta_0} RSS = \displaystyle\sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)\cdot(-1) = 0 \\ \dfrac{d}{d\beta_1} RSS = \displaystyle\sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)\cdot(-X_i) = 0 \end{cases}$$

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_1 X_i) = \frac{1}{n}\sum_{i=1}^{n}\beta_0 \implies \bar{Y} - \beta_1 \bar{X} = \beta_0$$

$$-\frac{1}{n}\sum X_i Y_i + \underbrace{\underbrace{\beta_0 \frac{1}{n}\sum X_i}_{\beta_0 \bar{X}}}_{(\bar{Y} - \beta_1 \bar{X})\bar{X}} + \beta_1 \frac{1}{n}\sum X_i^2 = 0$$

$$\beta_1\left(\frac{1}{n}\sum X_i^2 - (\bar{X})^2\right) = \frac{1}{n}\sum X_i Y_i - \bar{X}\bar{Y}$$

$$\begin{cases} \beta_1 = \dfrac{\frac{1}{n}\sum x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum x_i^2 - (\overline{x})^2} \\[2em] \beta_0 = \overline{y} - \beta_1 \overline{x} \end{cases} \approx \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}$$



$$\operatorname{var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$$

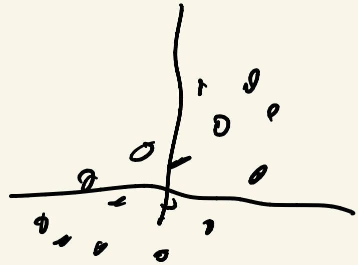
$$\begin{aligned}
\operatorname{cov}(X,Y) &= \mathbb{E}\big[(X - \mathbb{E}X)\cdot(Y - \mathbb{E}Y)\big] \\
&= \mathbb{E}\big[X\cdot Y - X\cdot\mathbb{E}Y - \mathbb{E}X\cdot Y + \mathbb{E}X\cdot\mathbb{E}Y\big] \\
&= \mathbb{E}[X\cdot Y] - \mathbb{E}[X\cdot\mathbb{E}Y] - \mathbb{E}[\mathbb{E}X\cdot Y] + \mathbb{E}[\mathbb{E}X\cdot\mathbb{E}Y] \\
&= \mathbb{E}[X\cdot Y] - \mathbb{E}Y\,\mathbb{E}X - \cancel{\mathbb{E}X\,\mathbb{E}Y} + \cancel{\mathbb{E}X\,\mathbb{E}Y} \\
&\approx \sum_{i=1}^{n} (x_i - \overline{x})\cdot(y_i - \overline{y})
\end{aligned}$$

$$\boxed{\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}X\,\mathbb{E}Y}$$

$$\approx \frac{1}{n}\sum x_i y_i - \overline{x}\,\overline{y}$$

pretend $\mathbb{E}X=0$, $\mathbb{E}Y=0$

$\mathrm{Cov}(X,Y) = \mathbb{E}(X\cdot Y)$



X,Y positive covariance:

$$\mathbb{E}(X\cdot Y) > 0$$

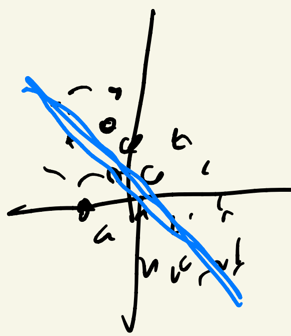
$X\cdot Y>0$ on average

then on average,

when $X>0$ $Y>0 \rightarrow XY>0$

when $X<0$ $Y<0 \rightarrow XY>0$

X,Y negative covariance:



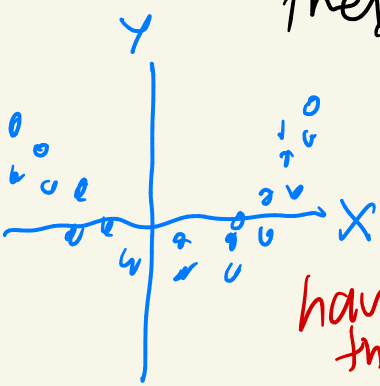
$$\mathbb{E}(X\cdot Y) < 0$$

$X\cdot Y<0$ on average

then on average,

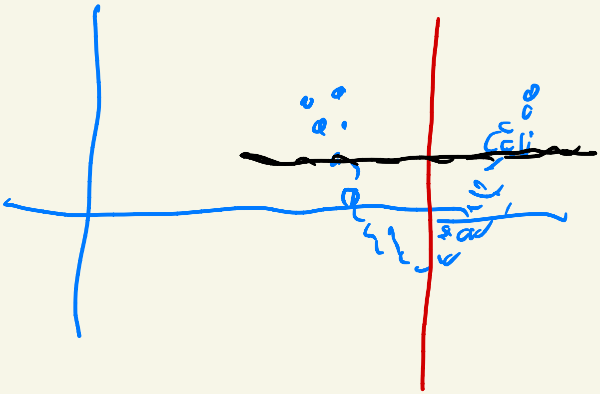
when $X<0$ $Y>0$

when $X>0$ $Y<0$



having 0 covariance does not mean there's no relationship.

means can't glean much info abt the relationship of the signs (assuming mean 0)



Covariance is a measure of linear relationship

slope

$$\hat{\beta}_1 \approx \frac{\mathrm{Cov}(X,Y)}{\mathrm{VAR}(X)}$$

$$\begin{cases} \beta_1 = \dfrac{\frac{1}{n}\sum X_i Y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum X_i^2 - (\bar{X})^2} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2} \\[2em] \beta_0 = \bar{Y} - \beta_1\bar{X} \end{cases}$$

Showed $\mathrm{var}(X) := \mathbb{E}(X-\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$

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

$$= \mathbb{E}(X\cdot Y) - \mathbb{E}X\,\mathbb{E}Y$$

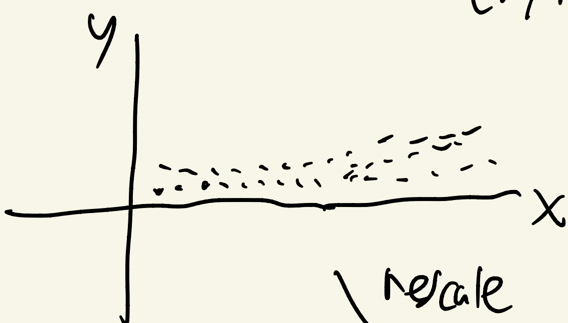
<u>Sample covariance</u> $\quad \sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$

<u>Sample variance</u> $\quad \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$

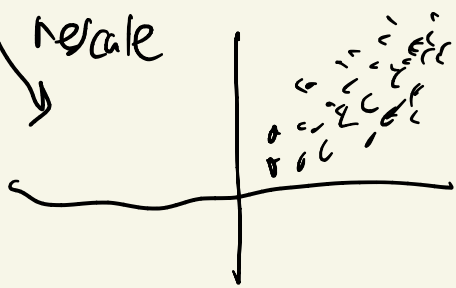
<u>CoRRelation</u>

$$\mathrm{COR}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Sd}(X)\cdot\mathrm{Sd}(Y)}$$

$$-1 \le \mathrm{COR}(X,Y) \le 1$$



rescale



Sample correlation $\quad r_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

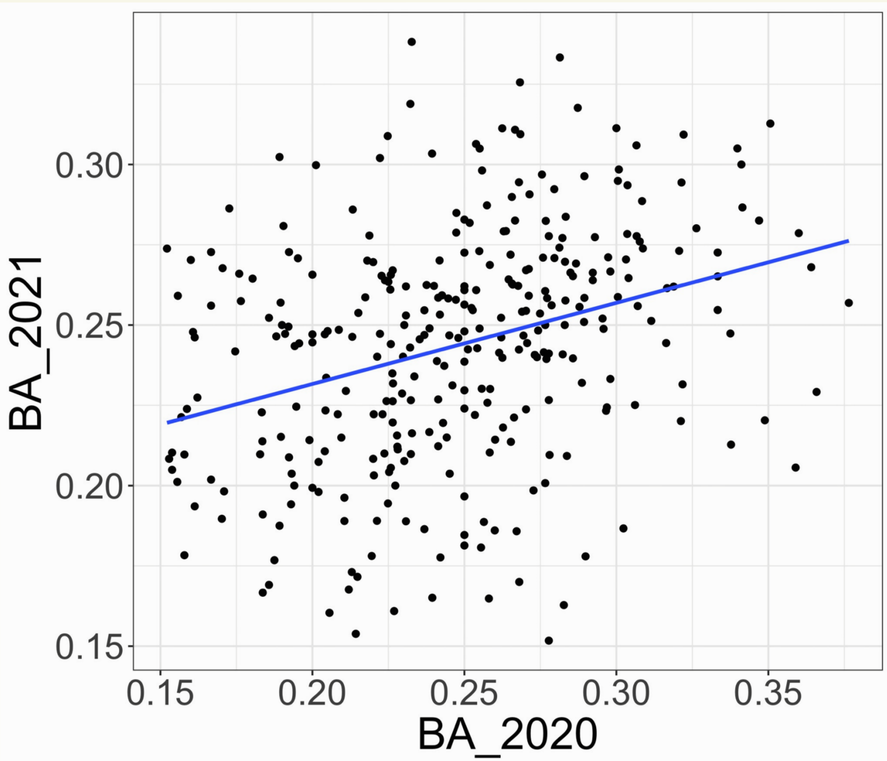
$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \cdot \frac{\sqrt{\sum_i (y_i - \bar{y})^2}}{\sqrt{\sum_i (y_i - \bar{y})^2}}$$

$$= \left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - x)^2}\sqrt{\sum (y_i - \bar{y})^2}} \right) \cdot \left(\frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{\sum (x_i - \bar{x})^2}} \right)$$

$$= \boxed{r_{xy} \cdot \frac{\sigma_y}{\sigma_x} = \hat{\beta}_1}$$

$$\boxed{r_{xy} = \hat{\beta}_1 \frac{\sigma_x}{\sigma_y}}$$

If x, y are normalized (have same variance) then linear regression slope is the sample correlation.



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

$[-1, 1]$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{1}{4}$

$\bar{X} \approx .25 = \frac{1}{4}, \quad \bar{Y} \approx \frac{1}{4}$

$\hat{\beta}_1 = \frac{1}{4}, \quad \hat{\beta}_0 = 3/16$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x \approx \frac{3}{16} + \frac{1}{4} x$$

<u>regression to the mean</u>

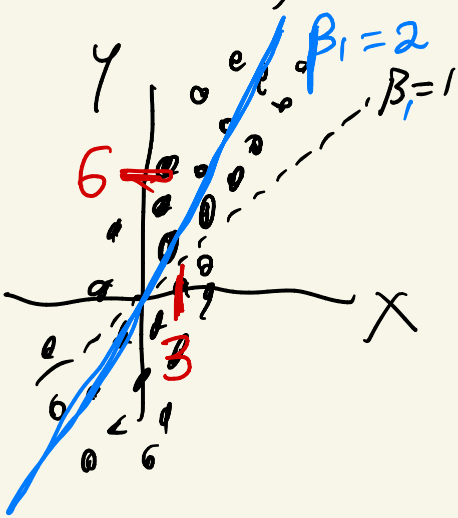
if $x = \bar{x} = \frac{1}{4}$, then $\hat{y} = \frac{1}{4}$

if $x > \bar{x}$, $x = .3$, $\hat{y} = \frac{13}{48} > \frac{12}{48}$

$< .3$

if $x < \bar{x}$, $x = .2$, $\hat{y} = \boxed{\frac{19}{80}} < \frac{20}{80}$

$> .2$



- regression code
- 2nd example → Pythagorean win percentage

Pythagorean Win Percentage

RA = runs allowed by a (baseball) team in one season

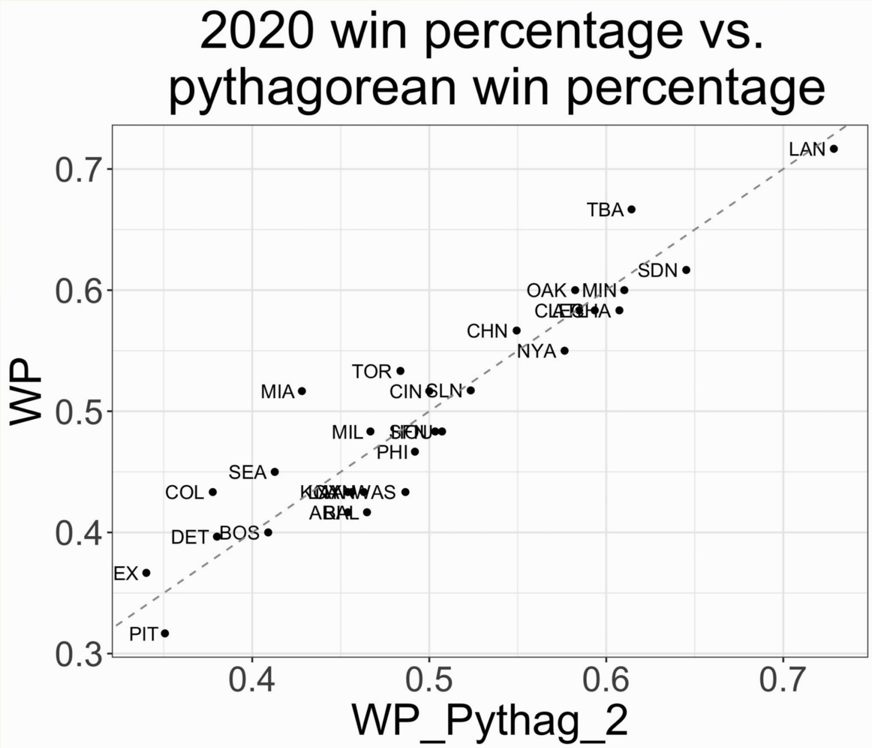
RS = runs scored in season

WP = team's win percentage that season

Bill James

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

Pythagorean



$$\widehat{WP} = \frac{RS^\alpha}{RS^\alpha + RA^\alpha} \qquad \text{find } \alpha?$$

<u>Model</u> $$\mathbb{E}\, WP = \frac{RS^\alpha}{RS^\alpha + RA^\alpha} = \frac{1}{1 + \left(\frac{RA}{RS}\right)^\alpha}$$

$$\Rightarrow \quad 1 + \left(\frac{RA}{RS}\right)^{\alpha} = \frac{1}{EWP}$$

$$\Rightarrow \quad \left(\frac{RA}{RS}\right)^{\alpha} = \frac{1}{EWP} - 1$$

$$= \frac{1 - EWP}{EWP}$$

$$\Rightarrow \quad \log \left(\frac{RA}{RS}\right)^{\alpha} = \log(\,\cdot\,)$$

$$\Rightarrow$$

$$\underbrace{0}_{\beta_0} + \boxed{\underbrace{\alpha}_{\beta_1} \cdot \underbrace{\log\left(\frac{RA_i}{RS_i}\right)}_{X_i} = \underbrace{\log\left(\frac{1 - wp_i}{wp_i}\right)}_{Y_i}}$$

Simple linear regression!

close to 2

$$\boxed{\hat{\alpha} = 1.867}$$

using 2020 data
pretty close to 2!