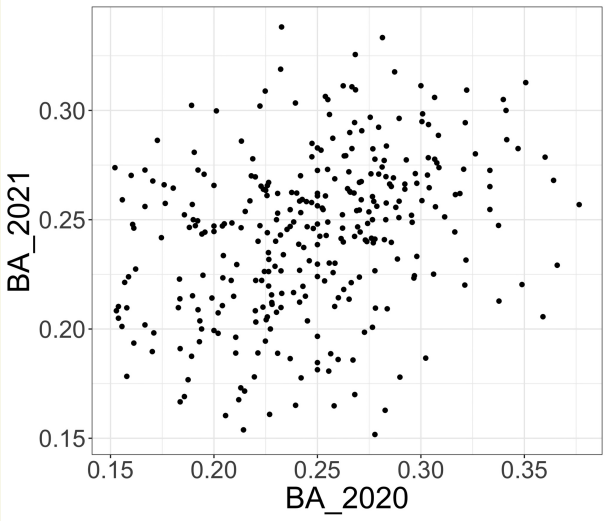
Simple Linear Regression

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

Visualize, if you can.

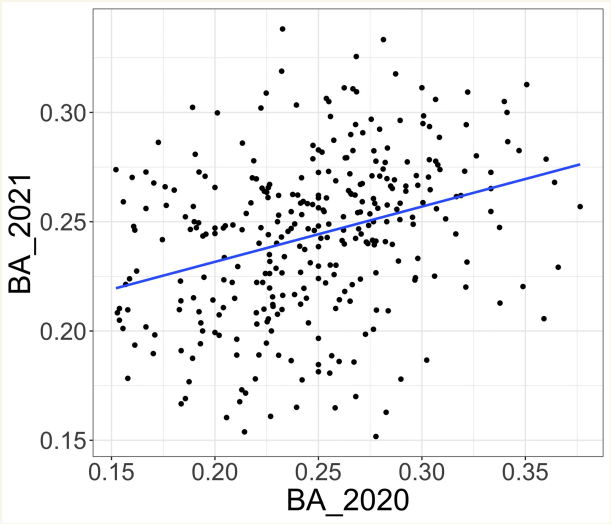


What does the relationship look like?

Looks linear, positive trend, but weak and noisy.

and we expect ↑ BA_{2020} to be associated with a ↑ BA_{2021}.
and we find out of.

We can find the "best fit line":



data generating process

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

$Y_i = \mathbb{E}[Y_i] + \varepsilon_i$

noise
randomness

- What does a best fit line really mean?
- how to find it?

— a representation of a phenomenon

write the picture using math.

Model index the MLB players by $i = 1, \ldots, n$

let X_i = player i's 2020 batting avg BA_i^{2020}

let Y_i = player i's 2021 batting avg BA_i^{2021}

~~assume a "true" linear relationship~~

model the data by a linear relationship,

~~$Y_i = mX_i + b$~~

Model

Statistician's POV:
how was the data generated in some idealized way?

data generating process

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

$Y_i = \mathbb{E}[Y_i] + \varepsilon_i$

noise
randomness

* How can we fit the best fit line to the data?
How can we estimate β_0, β_1?

The β's are parameters

Frequentist statistician: an unknown fixed constant that makes the model work and the ultimately describe how the data was generated in an idealized way.

Our task is to estimate the parameters: if there really was a "true" underlying line that generated the data, what would that line be?

Concretely: estimate β_0, β_1 from the data $\{(x_i, y_i)\}_{i=1}^n$

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

~~find β_0, β_1 that maximize $P(\text{observing } \{(X_i, Y_i)\}_{i=1}^n \mid \beta_0, \beta_1)$~~

↳ MLE ~~maximum likelihood estimation~~

Minimize the "error" between the line and the dots (data).

squared error for player i: $\left(Y_i - (\beta_0 + \beta_1 \cdot X_i)\right)^2$

residual

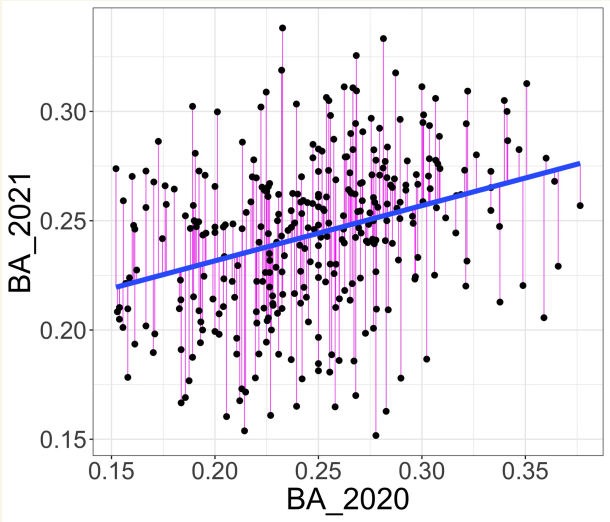
Y_i: actual 2021 BA

$\mathbb{E}(Y_i \mid X_i)$: according to our linear model, predicted 2021 BA

One def. of the best fit line is the line that has least squared error; Minimize the Residual Sum of Squares

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

using calculus ✓



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

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

$$\begin{cases} \beta_0 = \bar{Y} - \beta_1 \bar{X} \end{cases}$$

$$\frac{\partial}{\partial \beta_1} RSS = \sum_{i=1}^{n} 2\left(y_i - (\beta_0 + \beta_1 x_i)\right)(-x_i) = 0$$

$$-\frac{1}{n}\sum x_i y_i + \beta_0 \frac{1}{n}\sum x_i + \beta_1 \frac{1}{n}\sum x_i^2 = 0$$

$\beta_0 \frac{1}{n}\sum x_i \rightarrow \bar{x}$

$\bar{y} - \beta_1 \bar{x}$

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

⋮ algebra

$\hat{\beta}_0$ is our estimate of β_0

hat ^ means estimate

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

What do these formulas mean?

Covariance if X and Y are random variables, then

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

If $\mathbb{E}X=0$ and $\mathbb{E}Y=0$, then

$$\text{cov}(X,Y) = \mathbb{E}(X\cdot Y).$$

Positive covariance means:

On average,

$X>0$ when $Y>0$

$X<0$ when $Y<0$

negative cov means:

on avg,

$X<0 \longleftrightarrow Y>0$

$X>0 \longleftrightarrow Y<0$

Variance $\sigma_X^2 = \text{var}(X) = \mathbb{E}\left[(X-\mathbb{E}X)^2\right] = \text{cov}(X,X)$

on average, how far is X from its own mean

↘ squared error

- In these defs of cov and var, X and Y are random variables.

If we think of our baseball data $\{(x_i,y_i)\}_{i=1}^{n}$ as draws from some random variable,

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

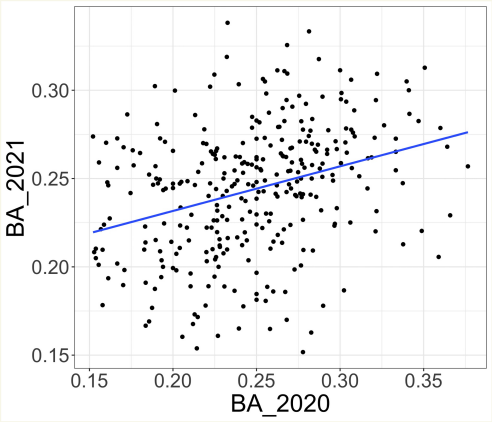
Sample covariance $S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

these are unbiased estimators:

$\mathbb{E}[S_x^2] = \sigma_x^2$ and $\mathbb{E}[S_{xy}] = \sigma_{xy}$

$\mathbb{E}(x \cdot y)$

$$\hat{\beta}_1 = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{sample covariance of } x, y}{\text{sample variance of } x}$$



Correlation is normalized covariance

ranges from -1 to 1, $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$

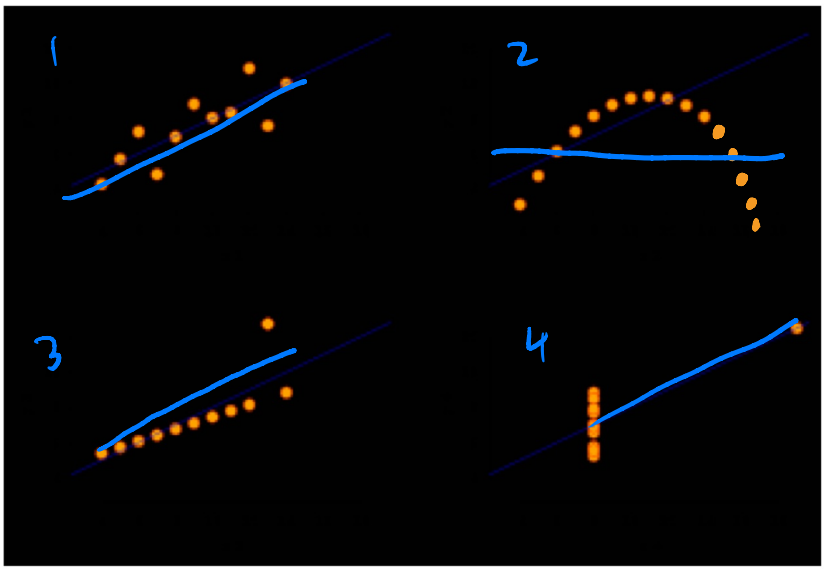
Sample correlation is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \longrightarrow$$

$$\hat{\beta}_1 = r_{xy} \cdot \frac{s_y}{s_x}$$

$$r_{xy} = \hat{\beta}_1 \cdot \frac{s_x}{s_y}$$

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



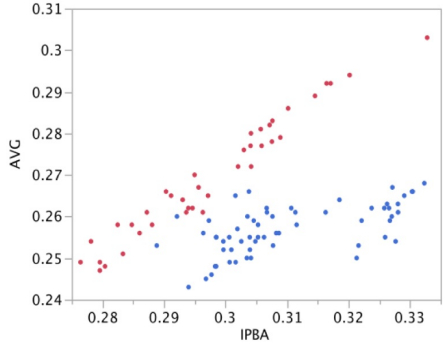
2. no correlation, but there is a relationship!
correlation is fundamentally a measure of linear association

4, 3. outliers can screw you
be careful

1. the ideal case for simple linear regression
   - no crazy outliers
   - linear association

try to plot your data if you can!

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.

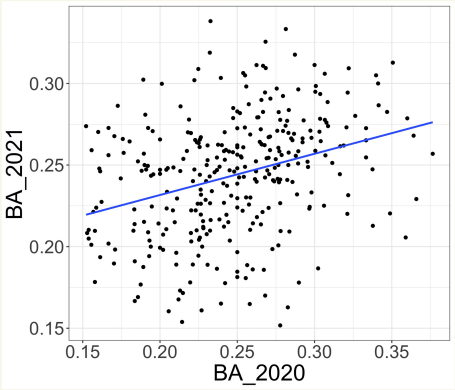
Here we can see that there is quite a different relationship between IPBA and Average before 1951 and after 1951; it may be best for our analysis to split up the data.

Finally, we can estimate the parameters β_0, β_1 of our BA model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ by minimizing the RSS using calculus.

In R: lm

```
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
```



$\hat{\beta}_1 = \frac{1}{4}$ it's not causal, it's observational.

A batting average increase of 0.020 in 2020 is associated with a predicted BA increase of 0.005 in 2021.

avg .245 ↓ .265 | → .245 ↓ .250

Regression to the mean

if $X_i > \bar{X}$, then $X_i = \bar{X} + \delta$ with $\delta > 0$,

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

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

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

$= \bar{y} + \frac{1}{4} \cdot \delta$

ryansbrill.com/lab/lab_2024