Simple Linear Regression

Grinnell College

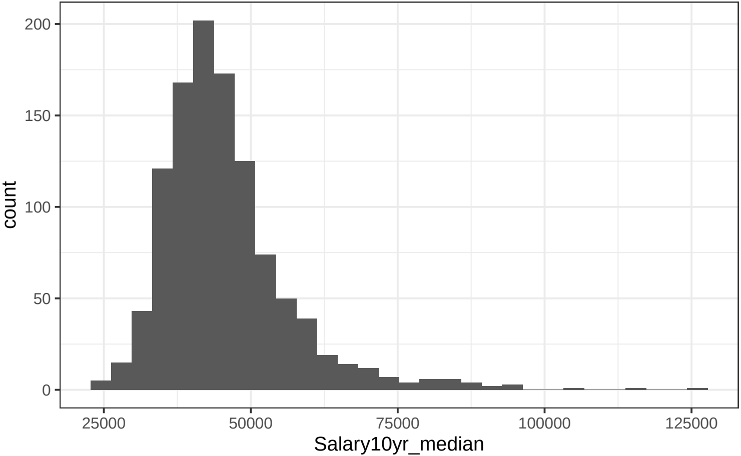
September 26, 2025

Review

- Scatterplot descriptions
  - form, strength, direction, outliers
- Pearson's correlation (r)
  - strength and direction of linear relationship for 2 quant. variables
- Spearman's correlation (ρ)
  - strength and direction of *monotone* relationship
  - more robust to outliers

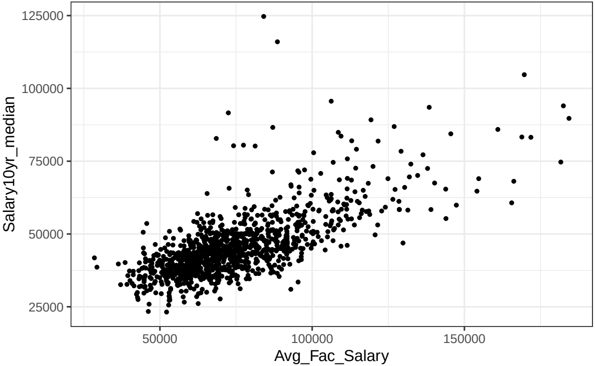
Motivation

If I asked you to guess your income after ten years, how would you guess?



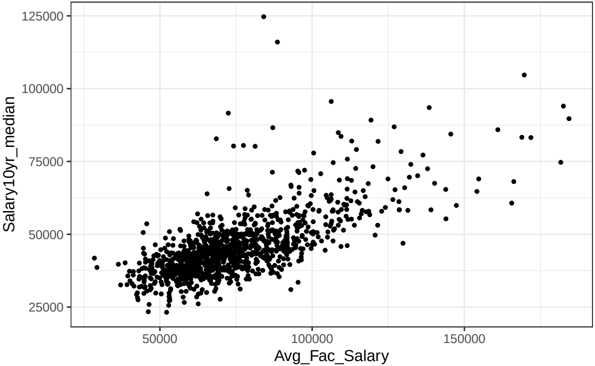
Motivation

If I told you my salary, how would you guess your (future) income?



Motivation

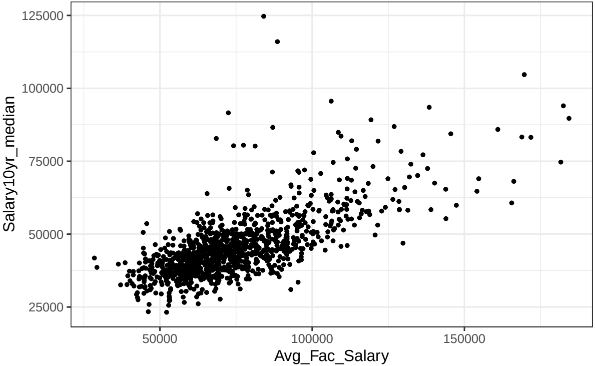
If I told you my salary, how would you guess your (future) income?



Linear Regression allows us to do this formally

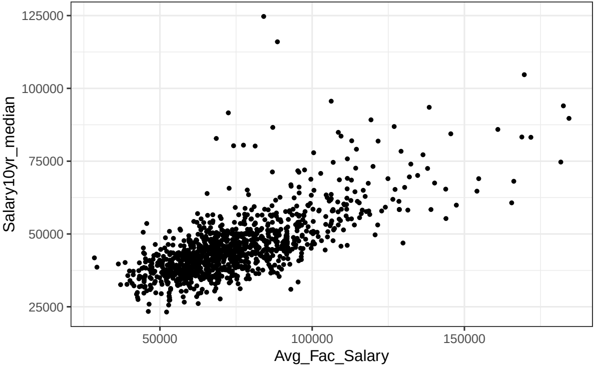
Correlation, Causation Review

Should you all tell the administration to raise my salary so your future income increases?



Correlation, Causation Review

Should you all tell the administration to raise my salary so your future income increases?



Yes!! But it won't actually increase your income. They are correlated but one doesn't cause the other (at least directly).

Basic Idea

Regression is how we model data; for us it's the "best fit line"

Two Main Goals:

- Use the regression/our best fit line(s) to describe the relationship between the explanatory variable(s) and the response variable
  - Science!
- Use the explanatory variable(s) to predict the response variable
  - Machine Learning/AI stuff
  - Business/finance investments
  - Planning around weather

Notation

- The variable being predicted is the *response* (aka “variable of interest”)
  - Usually denoted as y
- the variable we are using to do the prediction/explanation is the *explanatory variable* (aka “covariate” or occasionally “predictor”)
  - Usually denoted as x or X
- The estimates themselves are usually denoted with a “hat”
  - $\hat{y}$ is our predicted response
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are our estimated intercept and slope of the regression line (more in a second)

Notation Comparison

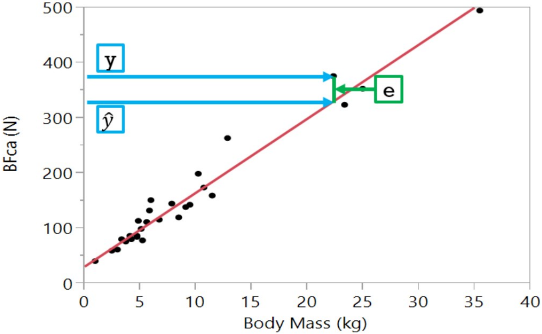
Statisticians use different symbols to write out a line than what you probably saw in HS algebra

Algebra
y = mx + b
m = slope: change in y over the change in x (rise / run)
b = intercept: value where the line cross the y-axis
All points fall exactly on the line

Statistics
$\hat{y} = \beta_0 + \beta_1 X$
β_1 = slope
β_0 = intercept
Not all of our data points will exactly on the line → variability

How it works

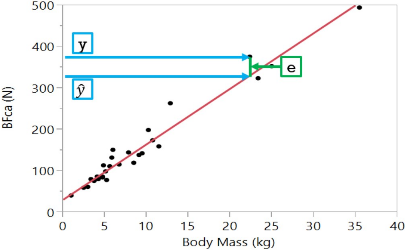
A regression line for the canidae data set predicting bite force (response) using body mass (explanatory)



- y's denote the values of the datapoints for the response variable
- points on the line are predicted values for the y's, denoted as $\hat{y}$
  - $\hat{y}$ are ALWAYS on our best-fit-line
- residual: difference between data and predictions (e = y - $\hat{y}$)

How it works

The **regression line** is the line that best fits through the data



- Need to define "best"
- Optimality critera: minimizes sum of squared residuals $\sum e_i^2$
- *Least Squares Regression* is another, more explicit name for this

Some Formulas

- $\hat{y} = \beta_0 + \beta_1 X$ **(regression equation)**
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ **(estimated regression equation)**
- $\hat{\beta}_1 = (\frac{s_x}{s_y})r$ (estimated slope)
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ (estimated intercept)
- e = y - $\hat{y}$ (**residual**)

Regression Line vs Estimated Regression Line

What is the difference between these two? Why do we have two?

- $\hat{y} = \beta_0 + \beta_1 X$ **(regression equation)**
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ **(estimated regression equation)**

Regression Line vs Estimated Regression Line

What is the difference between these two? Why do we have two?

- $\hat{y} = \beta_0 + \beta_1 X$ **(regression equation)**
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ **(estimated regression equation)**

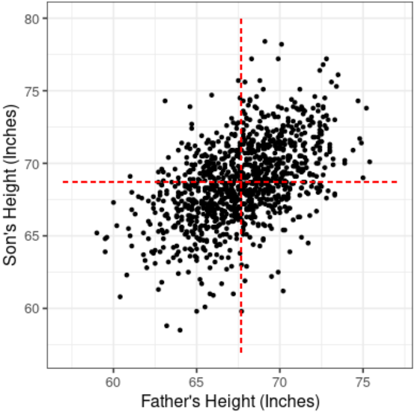
β_0 and β_1 are population parameters which means we almost never know them. Instead we have to estimate them using our sample.

Again, ^ (called hat) means estimated

Pearson's Height Data

	Mean	Std.Dev.	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

Father	Son
65.0	59.8
63.3	63.2
65.0	63.3
65.8	62.8
61.1	64.3
63.0	64.2
⋮	⋮



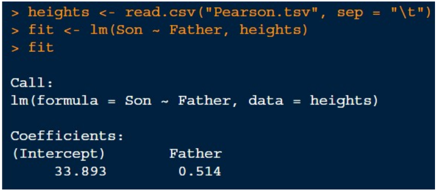
Pearson's Height Data

We could calculate our regression line using info from this table.

	Mean	Std.Dev.	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

Regression equation:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$$\hat{\beta}_1 = (\frac{s_x}{s_y})r$$
$$= (\frac{2.81}{2.74})0.501 = 0.514$$

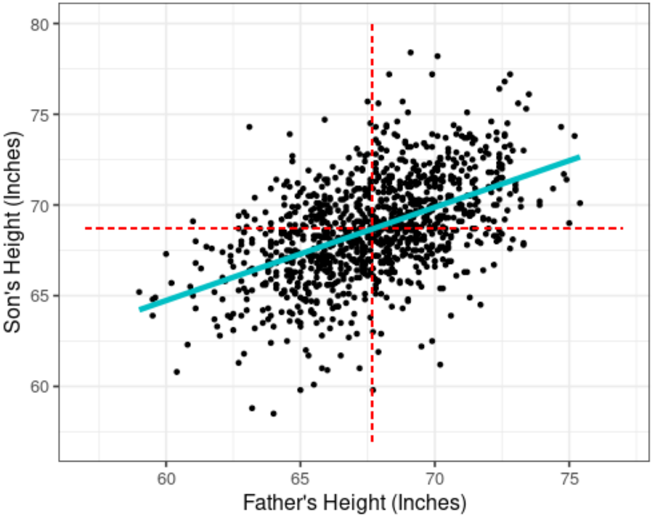


```
> heights <- read.csv("Pearson.tsv", sep = "\t")
> fit <- lm(Son ~ Father, heights)
> fit

Call:
lm(formula = Son ~ Father, data = heights)

Coefficients:
(Intercept)       Father
     33.893        0.514
```

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$
$$= 68.68 - 0.514 * 67.68 = 33.893$$

Pearson's Height Data – Plot Line

We can make R graph the line on our scatterplot.



Pearson's Height Data – Prediction

The formula for the regression line

$$\hat{y} = \beta_0 + X\beta_1$$

can be expressed in context of our original data and our estimated values

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

Given the Father's height, we can predict the son's height using this equation by plugging in a value for the father's height

Example: Predict the height of the son for a father with a height of 65in.

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times 65.0 = ?$$

Pearson's Height Data – Prediction

The formula for the regression line

$$\hat{y} = \beta_0 + X\beta_1$$

can be expressed in context of our original data and our estimated values

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

Given the Father's height, we can predict the son's height using this equation by plugging in a value for the father's height

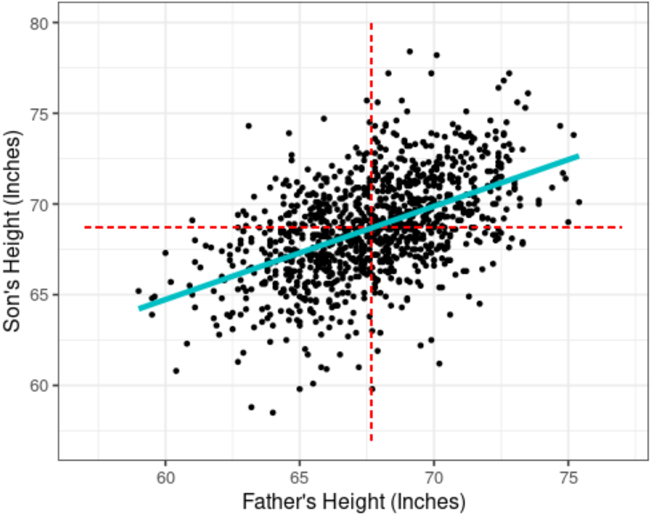
Example: Predict the height of the son for a father with a height of 65in.

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times 65.0 = 67.30 in.$$

Pearson's Height Data – Prediction

Predicted Son's Height = 67.30 inches for a father with height = 65in

- Check to see if our prediction makes sense on the graph



Residual

A **Residual** is the difference between an observed value and a prediction

- often labeled as **e** (e for "error", occasionally ϵ)
- e = y - $\hat{y}$

Interpretation: the residual tells us whether we have over- or under-predicted the values for the response variable in our data (and by how much)

- positive value → under-predicted
- negative value → over-predicted
- hard truth → I always forget which is which

Pearson's Height Data – Residual

In our data set, the first father had a height of 65 inches. We can calculate the residual for this father. We predicted the son's height to be 67.30 inches.

$$\begin{aligned} e &= y - \hat{y} \\ &= \text{observed value - predicted value} \\ &= 59.8 in. - 67.30 in. = -7.5 in. \end{aligned}$$

Father	Son
65.0	59.8
63.3	63.2
65.0	63.3
65.8	62.8
61.1	64.3
63.0	64.2
⋮	⋮

Interpretation: We overpredicted the height of this particular son by 7.5 inches

Next Time

At this point everything we've done in linear regression has only been a mathematical result

- The Best-fit-line is a geometric minimization problem
- We have yet to make assumptions

Next time, we will introduce the assumptions for SLR and then interpretations for the slope and intercept

- Assumptions are wrong -> best fit line is wrong
- ALWAYS check the assumptions before you worry about your interpretations/model's results
- NO INTERPRETATION for $\hat{\beta}_0$ or $\hat{\beta}_1$ is valid if the assumptions are broken (in a meaningful way)

Assumptions

Need to check all of the following (in any order)

1. X and y have a linear relationship
   - Want a straight line
   - Curvy lines, expoenential growth, etc... won't work
2. The errors are normally distributed with mean 0
3. The errors have a constant st. dev. (homoskedasticity)
4. The errors are not correlated (independent)

The last of these can be abbreviated to

$$e_i \overset{\text{iid}}{\sim} N(0, \sigma) \tag{1}$$

Checking Assumptions: Independence

This is the odd man out as it's hard to check visually

- Has to be check via critical reflection
- Verrrry common assumtion to mess up ("psuedo-replication")
    - $psuedo - replication$ is when a an observational unit is measured twice (or more) and treated as two (or more) units
    - Ways to deal with this (mixed models)
- Things to ask yourself:
    - Is there a reason one observation might influence another observation?
    - If I told you the errors of the observations around a given observation would you have any information?

Checking Assumptions

The majority of your assumptions can and will be check visually using graphics created from the residuals. Two super common types

- Residual by Predicted
  - x-axis is predicted value
  - y-axis is the residual
  - want a cloud of points centered around the residual = 0 line.
  - Super important graph

- QQ-plot (Q = Quantile)
  - x-axis is the predicted values
  - y-axis is the observed values
  - want a straight 45 degree line

COLOR and SHAPE your residuals by other variables!!!!

- Odd man out is independent errors
- Homoskedasticity
    - Does a scatterplot of your predictions vs your residuals create a spread out pattern?
    - Is the spread of the errors roughly consistent across the graph?
- Normally distributed
    - Is the scattering of residuals roughly equal above and below the line at 0?
    - Are most of the errors concentrated around the line at 0?
- X and y have a linear relationship
    - Scatterplot of X and y