Simple Linear Regression

Grinnell College

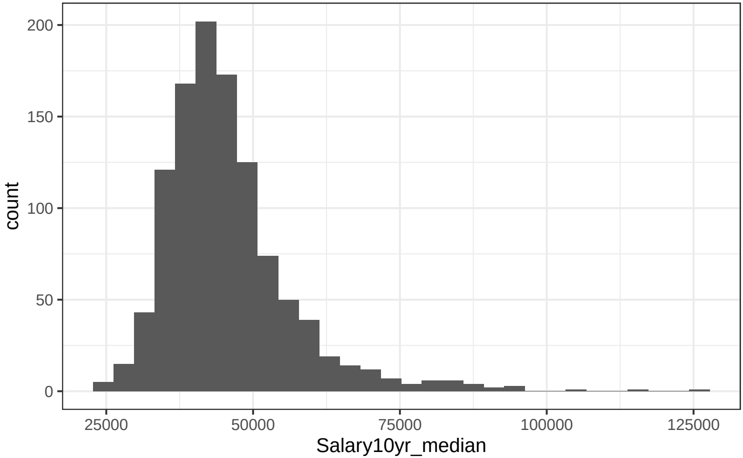
September 26, 2025

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

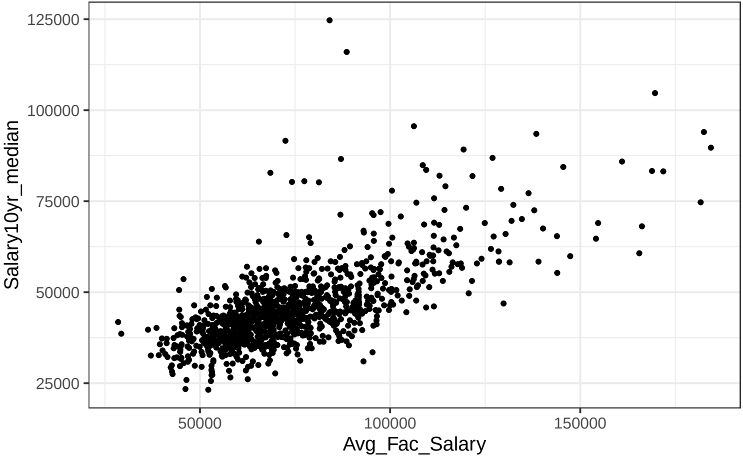
Motivation

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



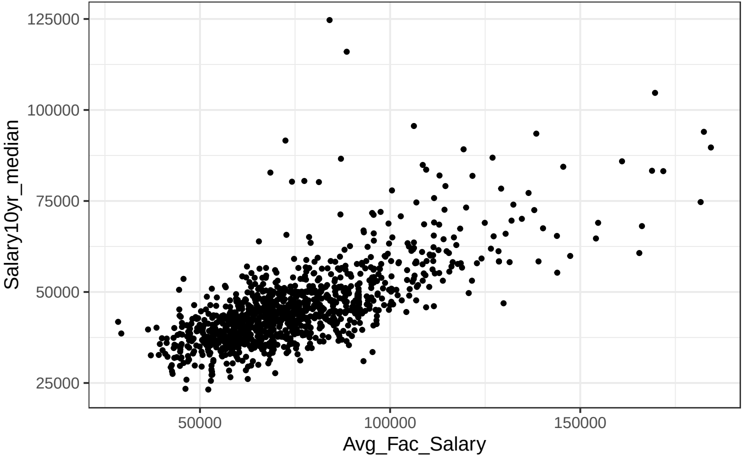
Motivation

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



Motivation

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



Linear Regression allows us to do this formally

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
  - Prediction and AI stuff

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

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

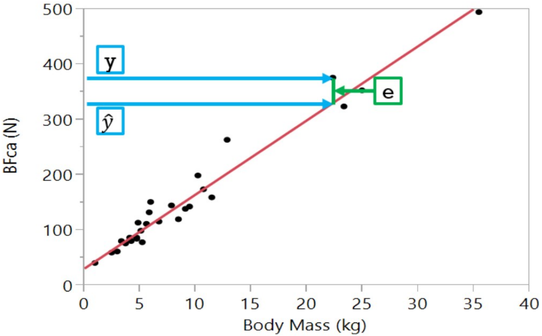
β_1 = slope

β_0 = intercept

Not all of our data points will exactly on the line $\rightarrow$ variability

How it works

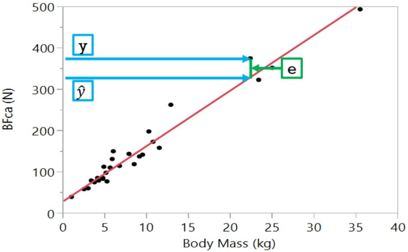
A regression line for the canidae data set predicting bite force (response) using body mass (explanatory)



- 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 ($\mathbf{e}$ = 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

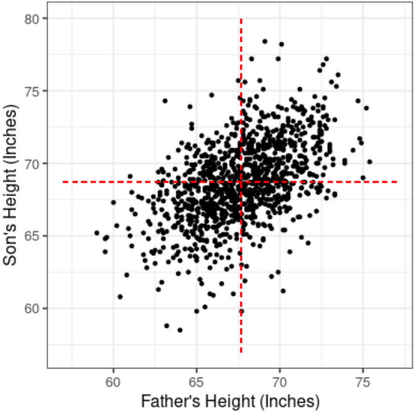
Some Formula's

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

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:
$\hat{y} = b_0 + b_1 X$

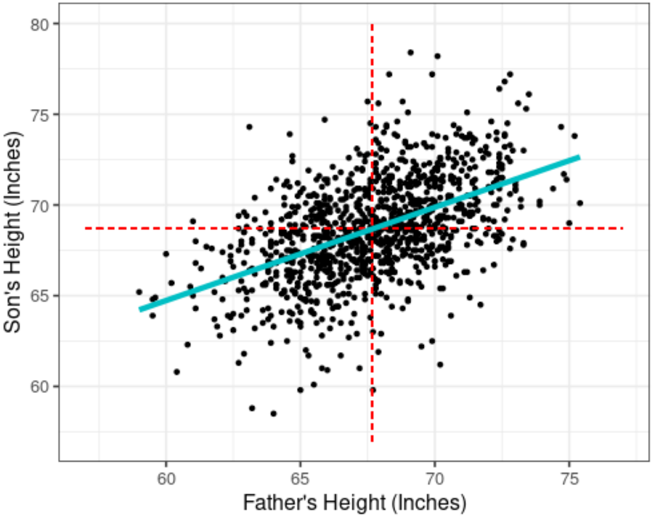
$$b_0 = (\frac{s_x}{s_y})r$$
$$= (\frac{2.81}{2.74})0.501 = 0.514$$

```
> 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
```

$$b_1 = \overline{y} - b_1\overline{x}$$
$$= 68.68 - 0.514 * 67.68 = 33.893$$

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 terms our our original variables and what we wish to predict

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

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

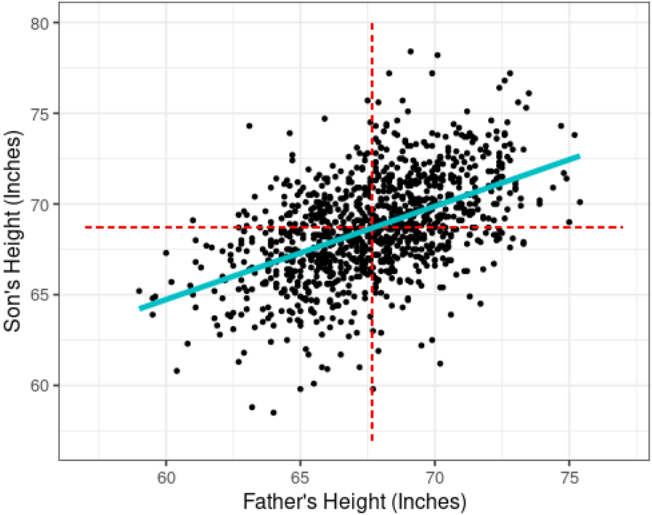
Example: Predict the height of the son for a father with a height of 65in.

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

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 $\rightarrow$ under-predicted
- negative value $\rightarrow$ over-predicted
- hard truth $\rightarrow$ 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.

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

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

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
⋮	⋮

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
- NO INTERPRETATION for $\hat{\beta}_0$ or $\hat{\beta}_1$ is valid if the assumptions are broken (in a meaningful way)