Lecture V: Intro to Linear regression

STA9750

Fall 2018

Baruch COLLEGE CUNY

Logistics

- Today: Lecture on correlation and simple linear regression
- Later in the week, I will upload a handout and datasets that cover how to do simple linear regression with SAS
- Please look at it and try to go through it before next lecture
- Next lecture, I'll go through the handout and answer your questions, and then I'll talk about multiple linear regression

Today

- Correlation
- Simple linear regression
- Transformations

Correlation

- The correlation between 2 quantitative random variables measures the ***linear*** association between 2 quantitative variables
- It can be computed in different equivalent ways (see textbook). For example, if our data are pairs:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- We can compute standardized values:

$$z_{x_i} = \frac{x_i - \overline{x}}{s_x} \qquad z_{y_i} = \frac{y_i - \overline{y}}{s_y}$$

- And, finally, compute the ***correlation coefficient***:

$$\boxed{r = \frac{1}{n-1}(z_{x_1} z_{y_1} + z_{x_2} z_{y_2} + \cdots + z_{x_n} z_{y_n})}$$

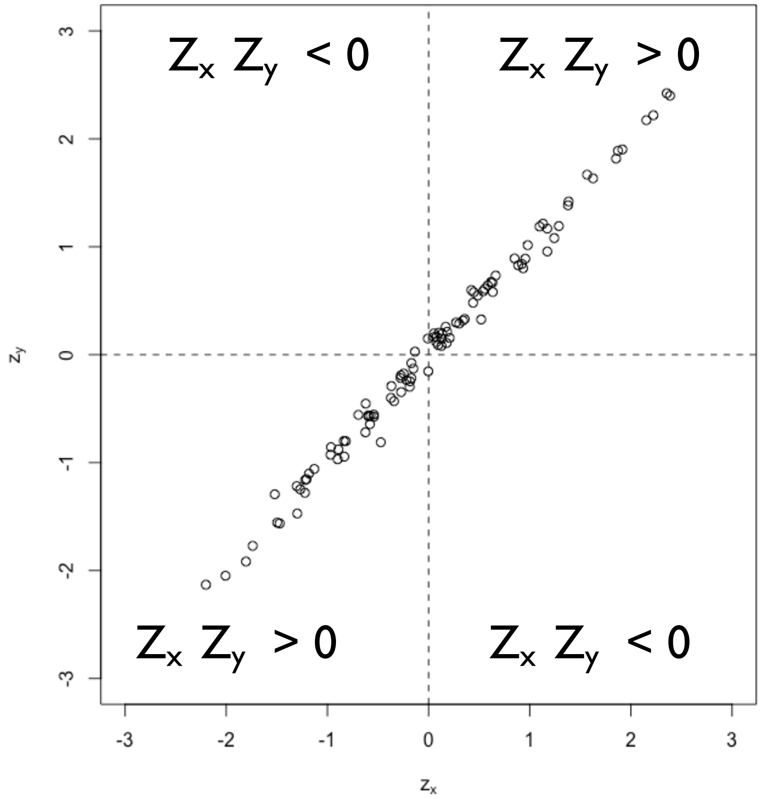
Interpreting correlation formula

- Standardized data have 0 mean
- That is, the scatterplot of z_y against z_x is centered at (0,0)
- Keep in mind:

$$r = \frac{1}{n-1}(z_{x_1}z_{y_1} + z_{x_2}z_{y_2} + \cdots + z_{x_n}z_{y_n})$$

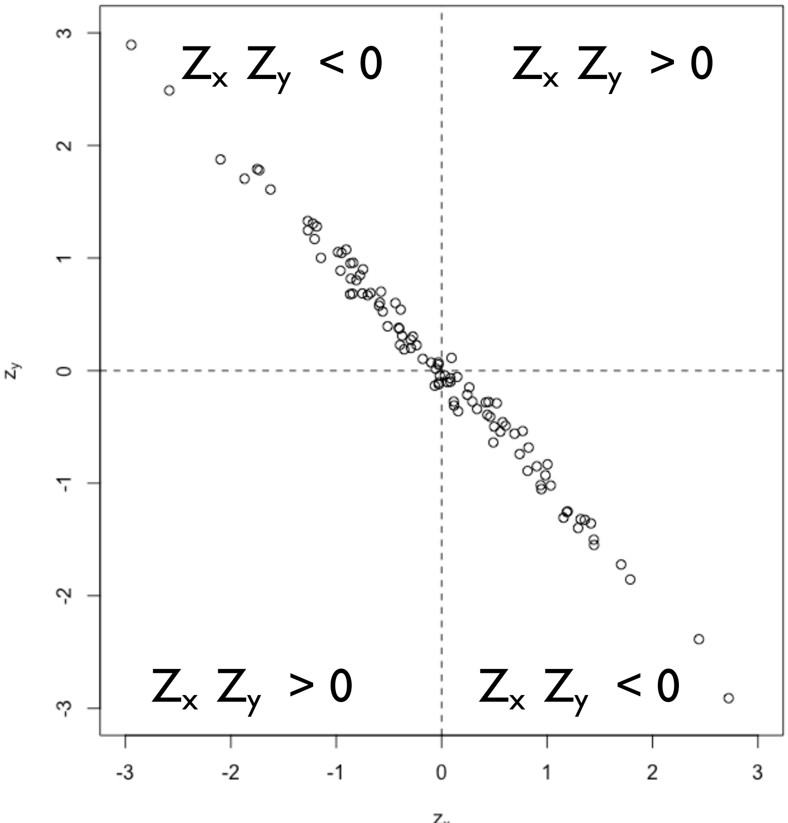
- *r* is always between -1 and 1. The extremes are attained when there are perfect linear relationships (with negative and positive slope, respectively)

Positively correlated ($r > 0$)



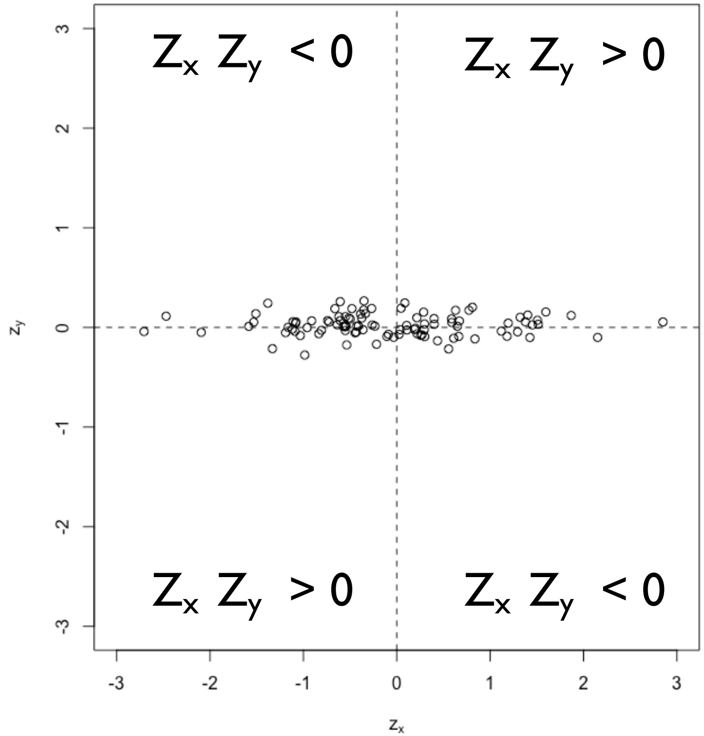
$$r = \frac{1}{n-1}\overbrace{(z_{x_1} z_{y_1} + z_{x_2} z_{y_2} + \cdots + z_{x_n} z_{y_n})}^{\text{mostly positive}}$$

Negatively correlated *(r < 0)*



$$r = \frac{1}{n-1}\underbrace{\left(z_{x_1}z_{y_1} + z_{x_2}z_{y_2} + \cdots + z_{x_n}z_{y_n}\right)}_{\text{mostly negative}}$$

No association

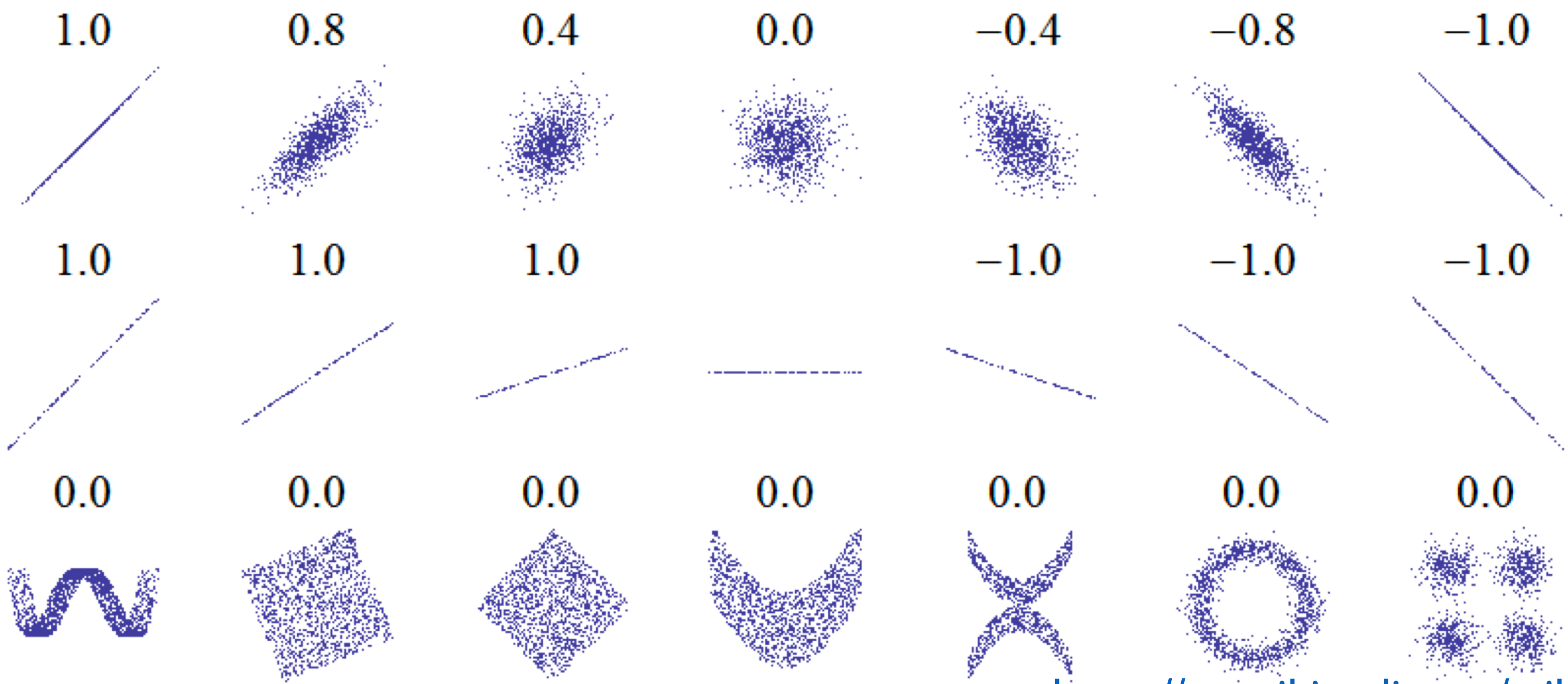


roughly the same positive & negative… will cancel out & r ~ 0

$$r = \frac{1}{n-1}\overbrace{(z_{x_1}z_{y_1} + z_{x_2}z_{y_2} + \cdots + z_{x_n}z_{y_n})}$$

r measures the strength and direction of **linear dependence**:

- *If there is a clear pattern, but it isn't linear… r is inadequate!*



https://en.wikipedia.org/wiki/Correlation_and_dependence

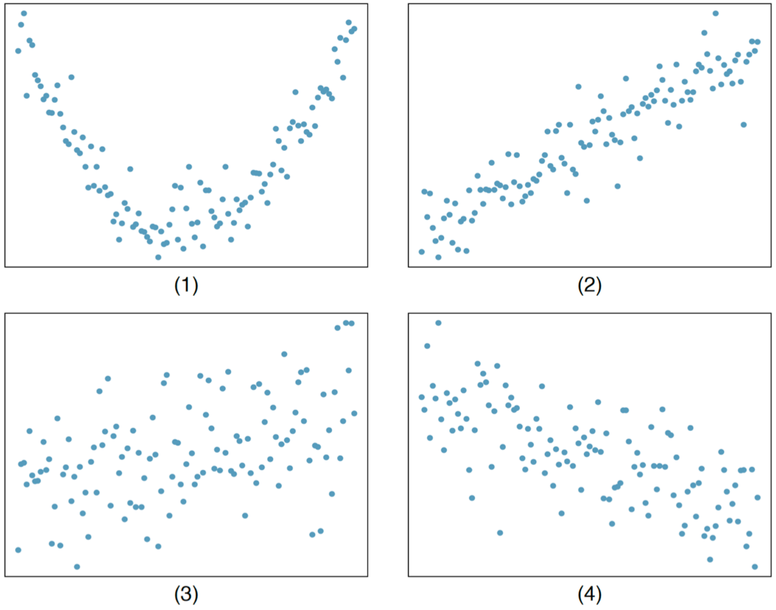
7.7 Match the correlation, Part I. Match the calculated correlations to the corresponding scatterplot.

(a) $r = -0.7$

(b) $r = 0.45$

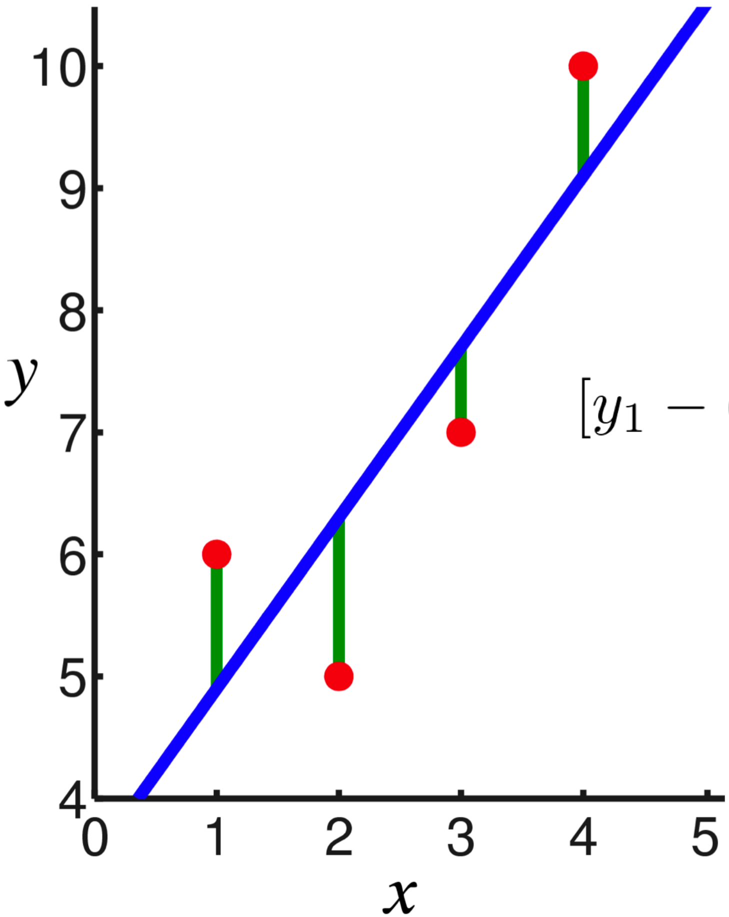
(c) $r = 0.06$

(d) $r = 0.92$



https://www.openintro.org/

Finding the best line: least squares



Goal

Find the best line $b_0 + b_1 x$

Least squares

Find b_0 and b_1 that **minimize**

$$[y_1 - (b_0 - b_1 x_1)]^2 + [y_2 - (b_0 - b_1 x_2)]^2 + \cdots$$
$$+ [y_n - (b_0 + b_1 x_n)]^2$$

Solution

$$b_0 = \overline{y} - b_1 \overline{x} \qquad b_1 = r \, \frac{s_y}{s_x}$$

Solution

- The least squares line is given by

$$\boxed{b_0 = \overline{y} - b_1\overline{x}}$$

$$\boxed{b_1 = r\,\frac{s_y}{s_x}}$$

- Predicted/fitted values:

$$\hat{y}_i = b_0 + b_1 x_i$$

- Residuals (errors): "observed minus predicted:"

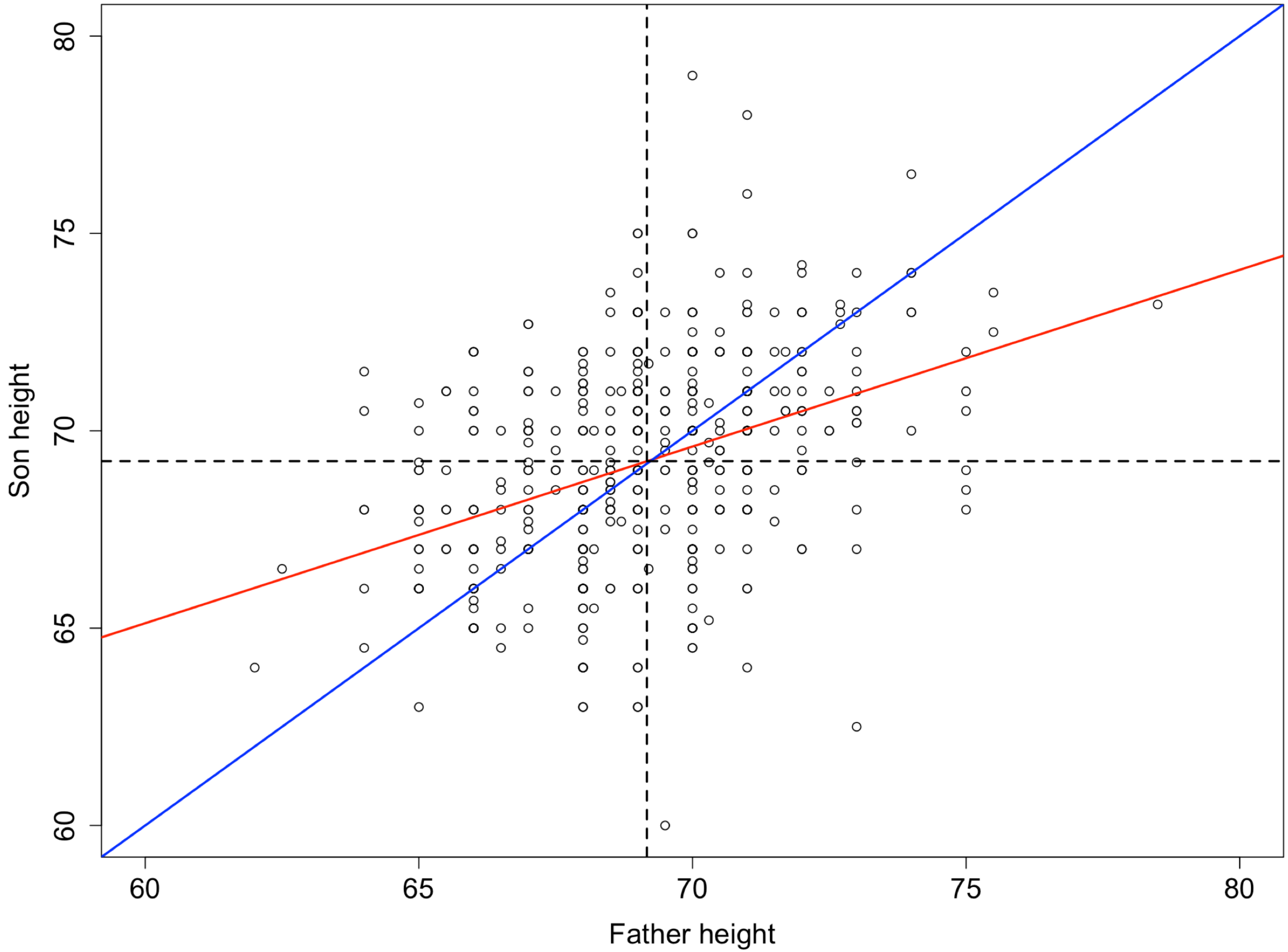
$$e_i = (y_i - \hat{y}_i)$$

Galton’s example

- In 1886, Galton published a study where he compared the statures of fathers and sons

Red line: least squares line

Blue line: $y = x$ [Son height = Father height]

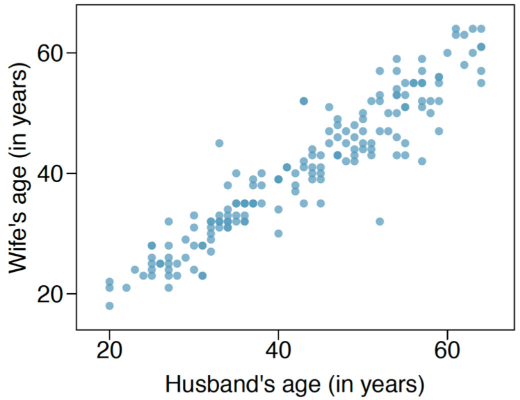


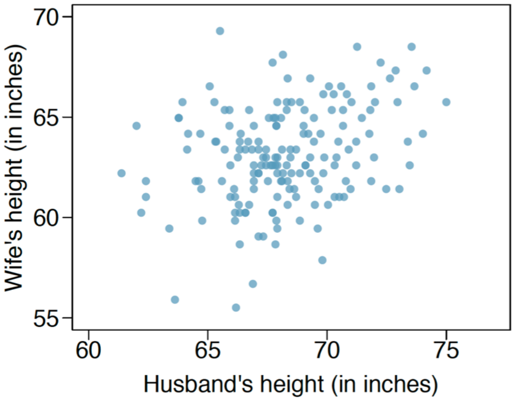
“Regression” to the mean

- If your father is tall, you’re likely to be tall, but shorter than he is
- If your father is short, you’re likely to be short, but taller than he is

That is, if your father is at the extremes, you’re likely to “regress” to the overall population mean

7.6 Husbands and wives, Part I. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives.[16] The scatterplot on the left shows the wife's age plotted against her husband's age, and the plot on the right shows wife's height plotted against husband's height.





(a) Describe the relationship between husbands' and wives' ages.

(b) Describe the relationship between husbands' and wives' heights.

(c) Which plot shows a stronger correlation? Explain your reasoning.

(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between husbands' and wives' heights?

https://www.openintro.org/

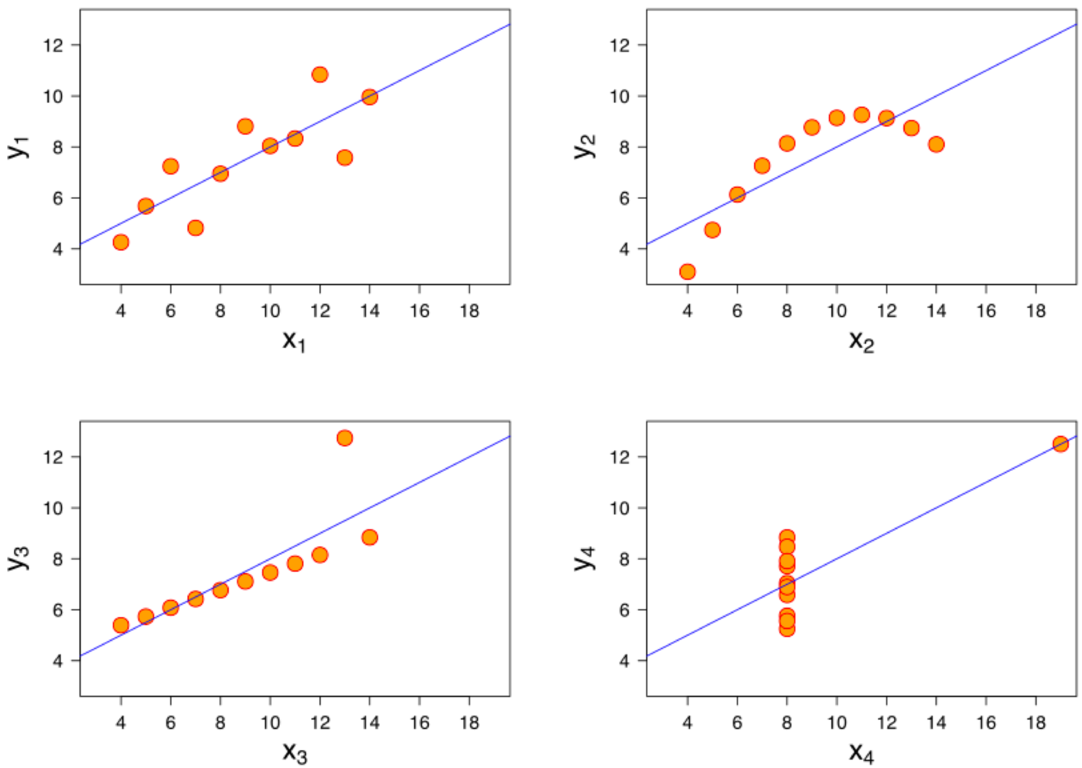
Coefficient of determination: R^2

- R^2 is very widely used measure for quantifying how "good" the least squares line and it is simply

$$R^2 = r^2$$

- It can be interpreted as the fraction of the total variability that is explained by the regression line
- Be careful: it doesn't tell us if the line is "adequate"

Anscombe's quartet



All datasets have $R^2 = 0.67$

… But vastly different stories!

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Inference?

- So far, we haven't made any distributional assumptions
- We just found the "best" line
- If we make some assumptions, we'll be able to find CIs and do hypothesis tests
- *Normal linear model*

$$\boxed{y_i \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)}$$

Assumptions:

- Independence of outcomes y_i for i in 1:n (given the x_i).
- Normality
- Homoscedasticity (equal variance across observations, which doesn't depend on x_i)
- Linearity

If the assumptions hold…

$$\mathrm{CI}_{1-\alpha}(\beta_0) = b_0 \pm t_{\alpha/2,n-2}\, s_{b_0}$$

$$\mathrm{CI}_{1-\alpha}(\beta_1) = b_1 \pm t_{\alpha/2,n-2}\, s_{b_1}$$

$t_{\alpha/2,n-2}$ is the $100(1-\alpha/2)\%$ quantile of a Student-*t* with *n*-2 degrees of freedom

The std. errors are $s_{b_1} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{(n-2)\sum_{i=1}^n (x_i - \overline{x})^2}}$ $\quad s_{b_0} = s_{b_1}\sqrt{\sum_{i=1}^n x_i^2/n}$

From here, we can do hypothesis tests by checking whether the intervals contain certain values (for example, if the interval for the slope contains 0)

How do we check assumptions?

- Since

$$y_i \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \Rightarrow y_i - (\beta_0 + \beta_1 x_i) \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

… then, if the assumptions are satisfied:

$$e_i = y_i - (b_0 + b_1 x_i) \overset{\text{iid}}{\approx} N(0, s^2)$$

Assumptions:

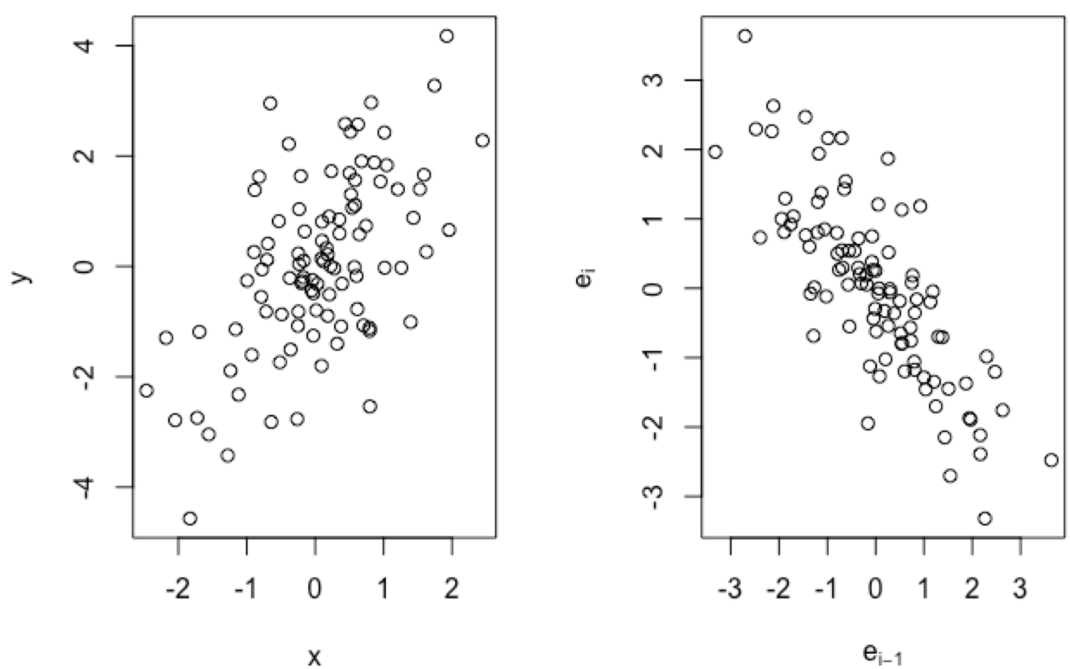
1. Independence of outcomes y_i for i in 1:n (given the x_i).
2. Normality
3. Homoscedasticity (equal variance across observations, which doesn't depend on x_i)
4. Of course, linearity

How to check them:

1. Check if e_i are *strongly* correlated (e.g. serial correlation, if observations are taken over time)
2. Q-Q plot of e_i
3. Scatterplot of e_i vs $b_0+b_1x_i$
4. Scatterplot of e_i vs $b_0+b_1x_i$

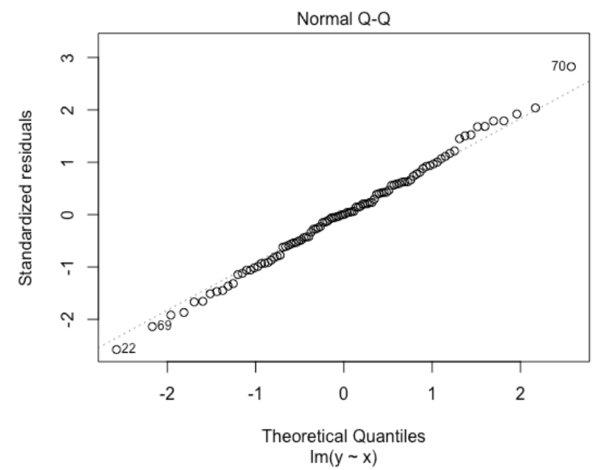
Independence?

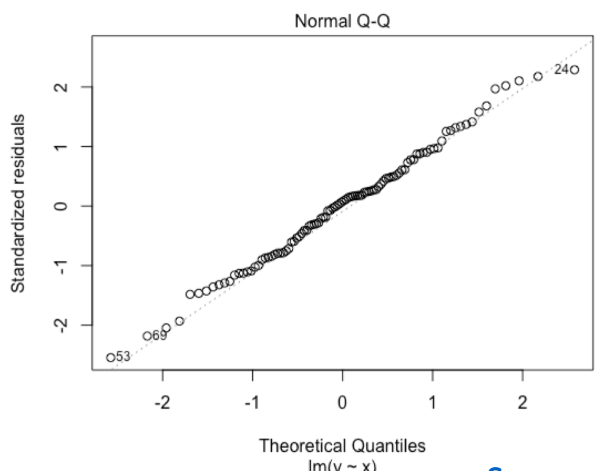
- Hard to check unless data are collected over time or there are clear “groups” or variables that were not included in the regression



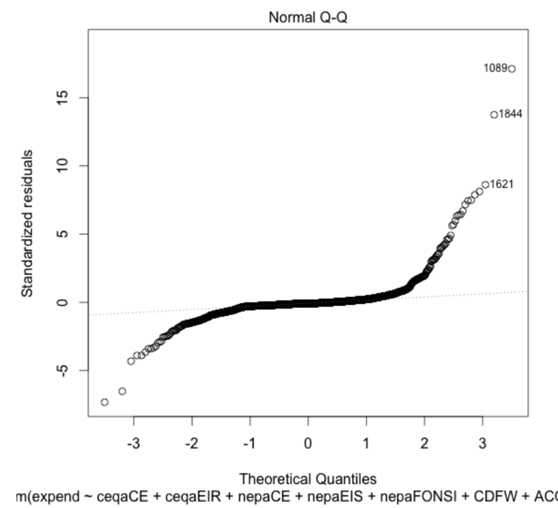
Normality? Q-Q plot: see if it is roughly linear

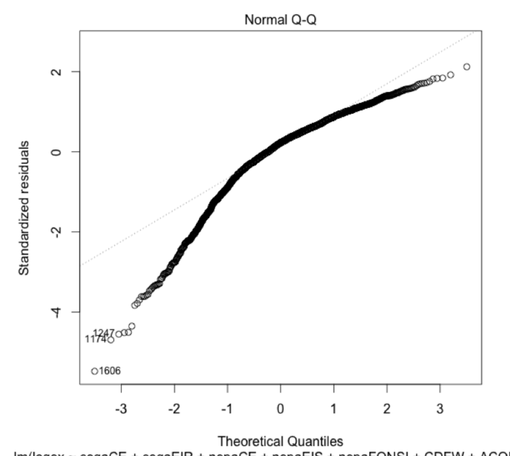
OK





Bad



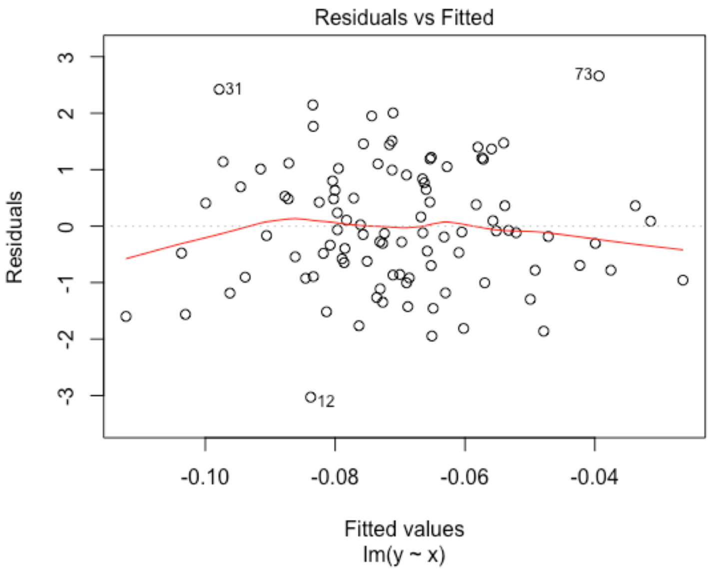


[Source of bad QQ-plots: https://stats.stackexchange.com/questions/160562/what-to-do-if-residual-plot-looks-good-but-qq-plot-doesnt-after-transforming-t](https://stats.stackexchange.com/questions/160562/what-to-do-if-residual-plot-looks-good-but-qq-plot-doesnt-after-transforming-t)

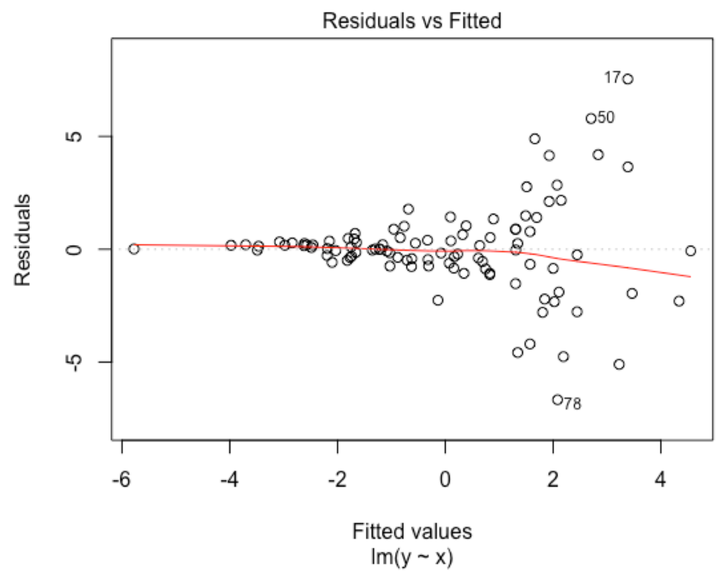
Homoscedasticity?

Constant spread in scatterplot of e_i vs $b_0+b_1x_i$

OK

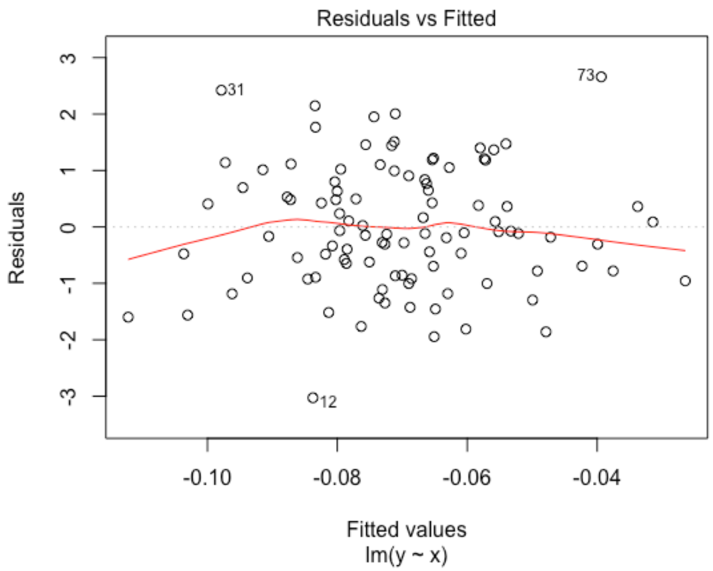


Bad

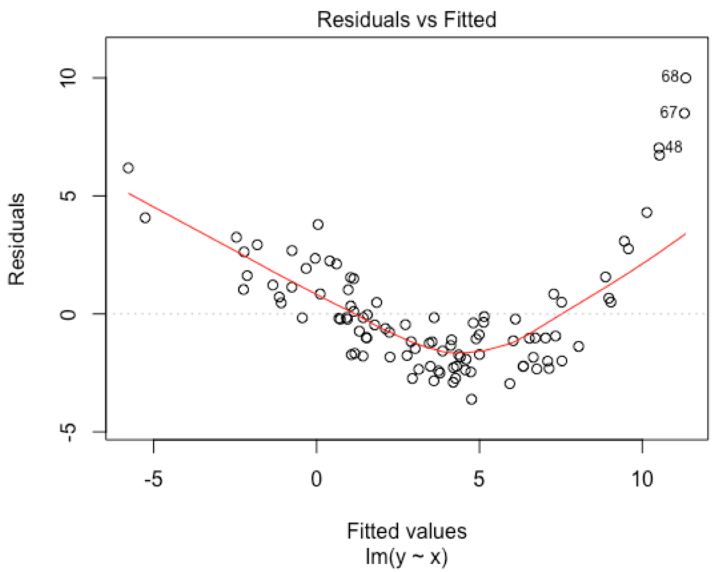


Linearity? No obvious patterns in scatterplot of e_i vs $b_0+b_1x_i$

OK



Bad



Next time…

- Simple linear regression with SAS
- Intro to multiple linear regression