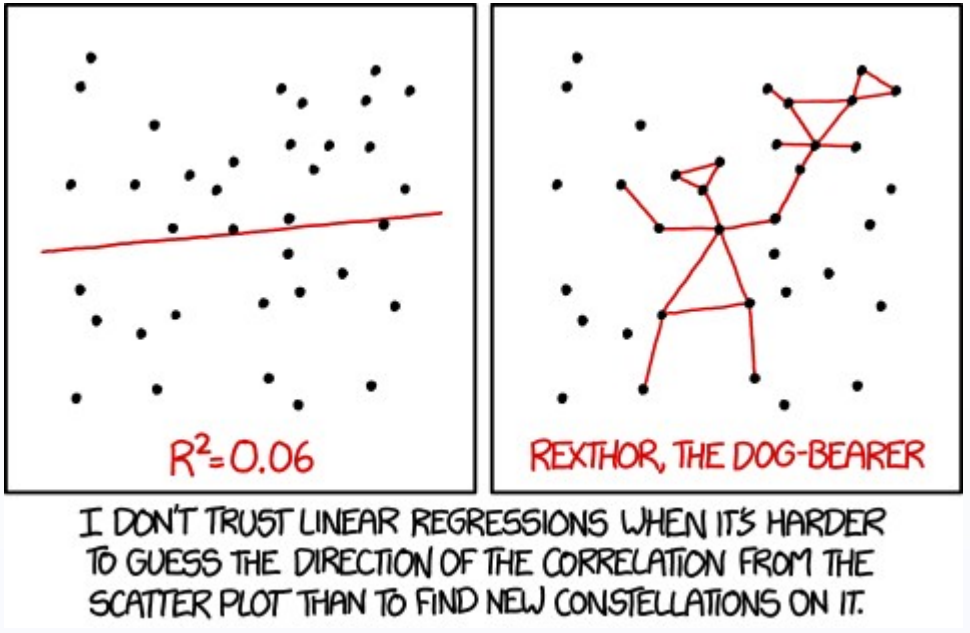
EGM101 – Skills Toolbox

Week 6, Part 4: The Coefficient of Determination

How good is my model?

- We've seen how to:
  - Assess linear relationship between variables (correlation)
  - Fit a linear model to observations (regression)
- Need some way to assess how well model fits to the data

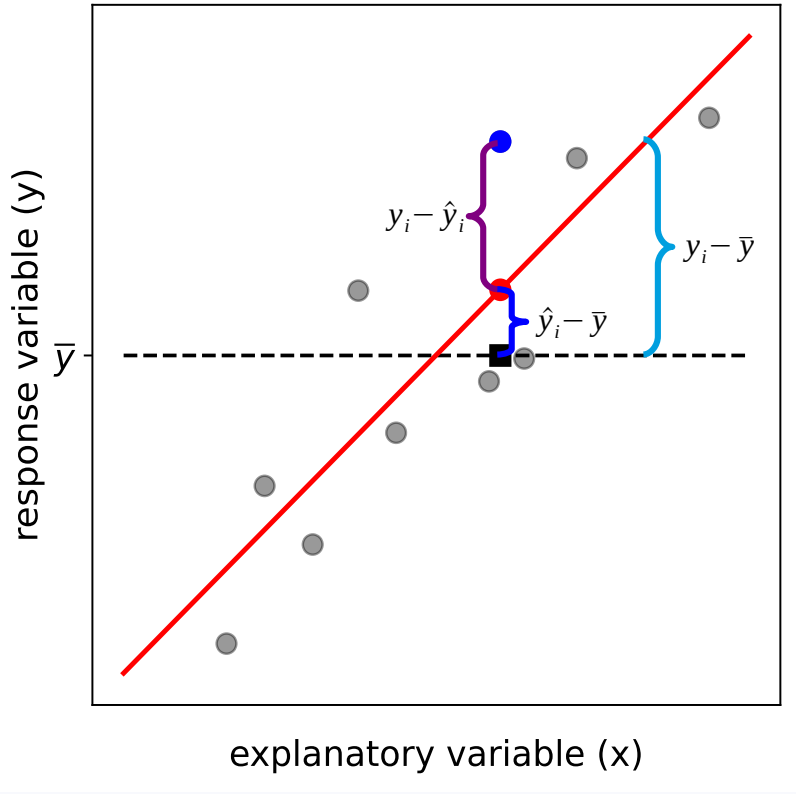


xkcd.com/1725

Variability

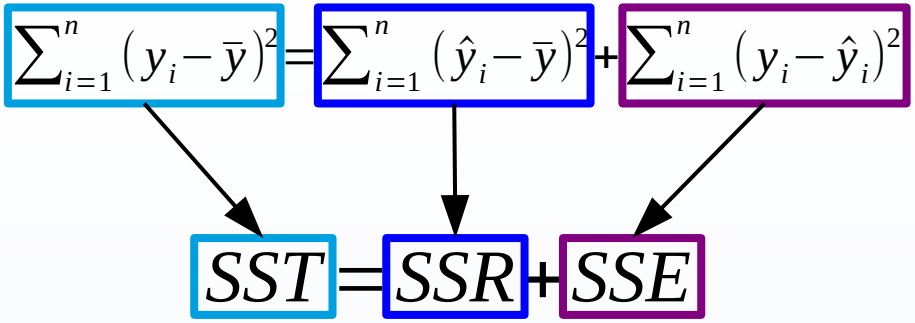
- Variability of y_i: $y_i - \bar{y}$
- Can think of variability of y as having two components:
  - Part explained by model
  - Part unexplained by model
- Remember: residuals, deviation from mean sum to zero
- So, use sum of squares:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$



Coefficient of Determination (general)

- Coefficient of determination (R^2): proportion of total variability accounted for by the model
- Often expressed as percent
  - Best case (perfect fit): $R^2 = 1$ (100%)
  - "Baseline" ($\hat{y} = \bar{y}$): $R^2 = 0$ (0%)
- Note: it is possible to have $R^2 < 0$
  - But, we won't be dealing with those cases.



$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Coefficient of Determination (simple case)

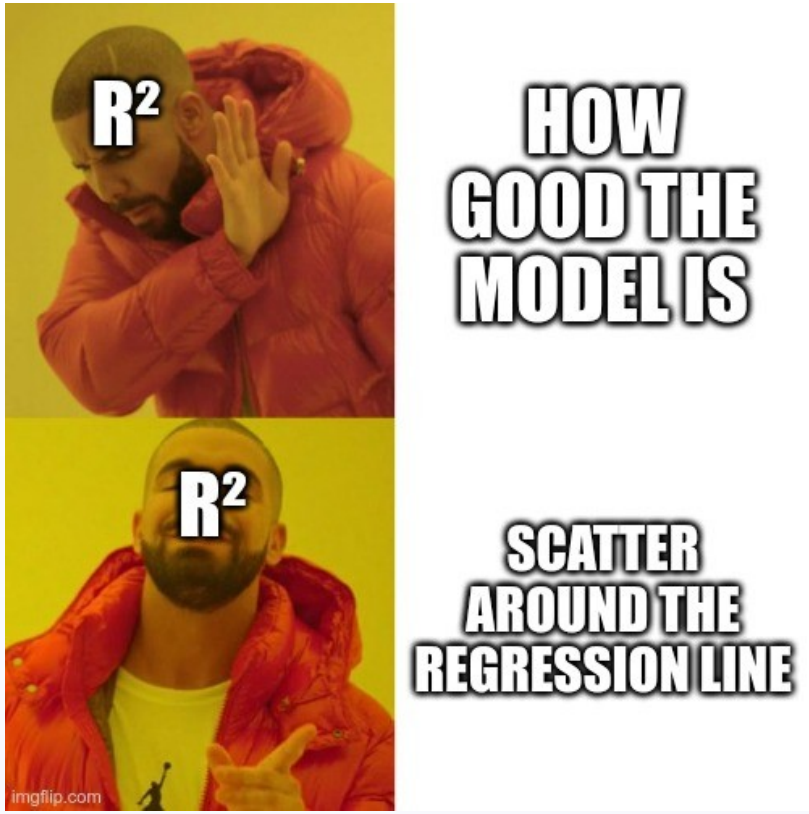
- Recall: in simple linear regression, only one explanatory variable (x)
- In this case, R^2 is the square of Pearson's correlation coefficient (r):

$$R^2 = r^2 = \left(\frac{\mathrm{cov}(x, y)}{s_x s_y} \right)^2$$

- This is likely the form that you will most commonly encounter (and many textbooks pretend it's the only one)

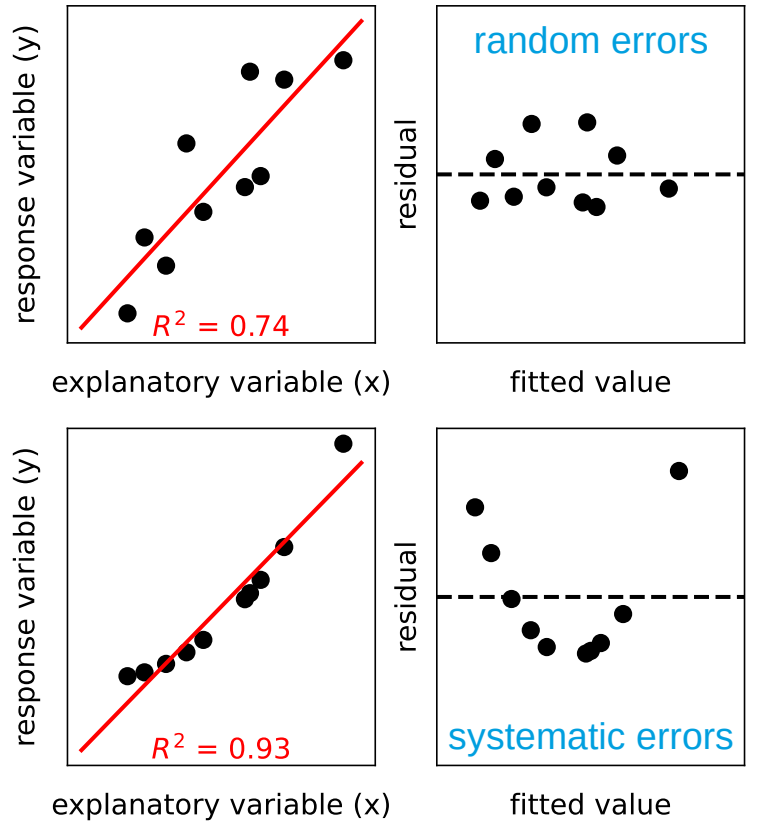
Interpreting the R^2 value

- R^2 tells us:
  - Scatter of data points around the best-fit line
  - Proportion of variability of dependent variable explained by the independent variable
- R^2 does not tell us:
  - How good the model is



High R^2 does not mean it's the right model!

- One way to check: look at the residual plot
  - Random errors: no apparent pattern in the errors
- Systematic bias: non-random errors (i.e., a pattern)
  - Indicates variability not accounted for in the model (bad fit)
- Solution: use a different model
  - Additional explanatory variables
  - Non-linear terms



So what's a good R^2, anyway?

- It depends on context, goal:
  - Understanding relationship between variables
  - Predicting unknown values
- Other questions:
  - How much of variability can be explained?
  - Is the relationship statistically significant? (more on this later...)
  - How precise are the predictions of the model?

Summary

- Coefficient of Determination (R^2):
  - Measures scatter of data around regression line
  - Measures how much of the variability in the response variable is explained by the explanatory variable(s)
- A high R^2 value does not tell us whether we're using the correct model for our data
- Interpreting/evaluating R^2 value depends on context, goal

Additional Resources

- Illowsky and Dean, Chapter 12.3
- Weiss, Chapter 4.3
- R-squared or coefficient of determination [Khan Academy]
- How To Interpret R-squared in Regression Analysis [Jim Frost]
- How High Does R-squared Need to Be? [Jim Frost]
- Five Reasons Why Your R-squared can be Too High [Jim Frost]