Lecture 16: model selection

John Zito

Duke University

STA 199 - 3/20/2025

Today's agenda

- **First**: mathy chat about model selection;
- **Second**: finish demo-ing AE 12;
- **Reminder**: Peer eval 2 is due tomorrow;

Model selection

Question: Given many competing models for predicting the same response, which one should we pick? Which model is "best"?

Answer: Find a way to assign a "quality score" to each model, and then pick the model with the highest score:

Model	Score	Verdict
Model A	0.1	$\times$
Model B	0.5	✓
Model C	0.4	$\times$

This is a massive area of statistics. There are "quality scores" for capturing all kinds of objectives: predictive accuracy, goodness-of-fit, simplicity, fairness, etc.

We will study two simple ones: R^2 and *adjusted* R^2.

Model fit and R^2

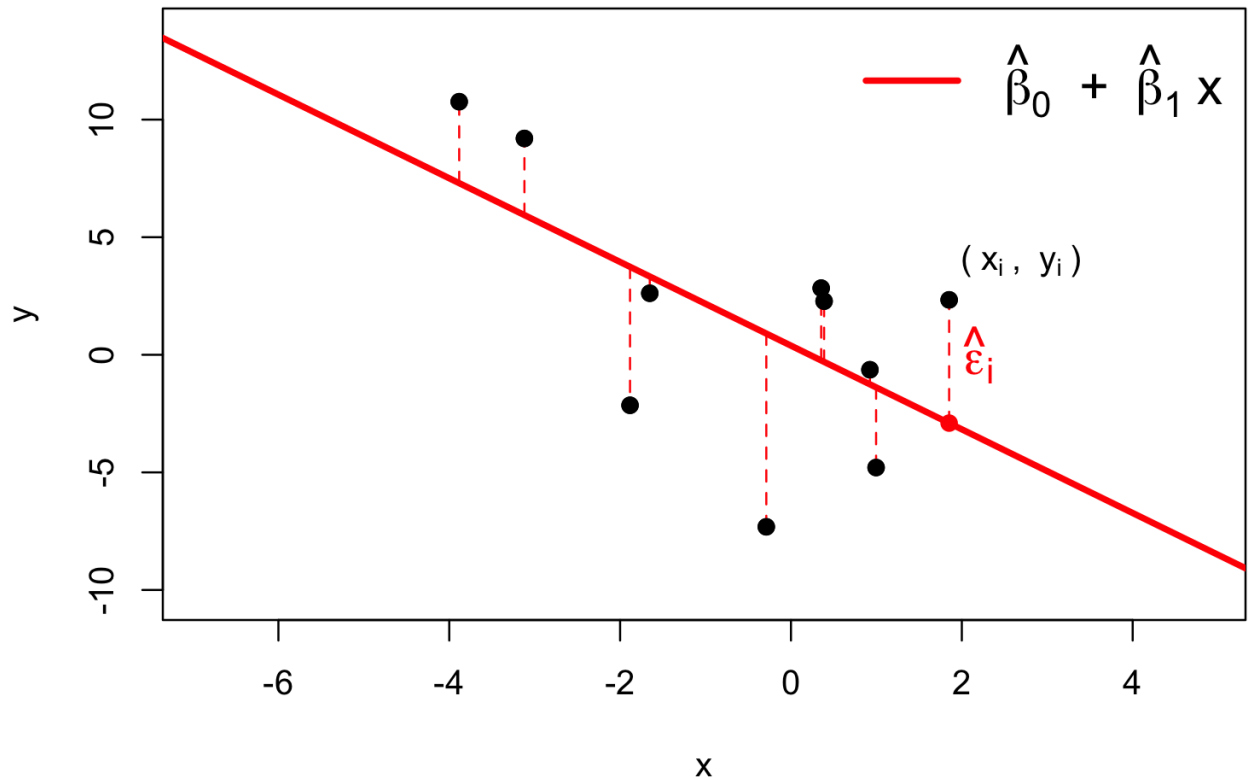
- Quality of fit appears to have something to do with how "all over the place" the residuals are;

- We want to quantify this intuition with a concrete numerical measure that we can use to rank competing models according to goodness-of-fit.

Recall the residuals

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \text{(fitted model)}$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \qquad \text{(residuals)}$$

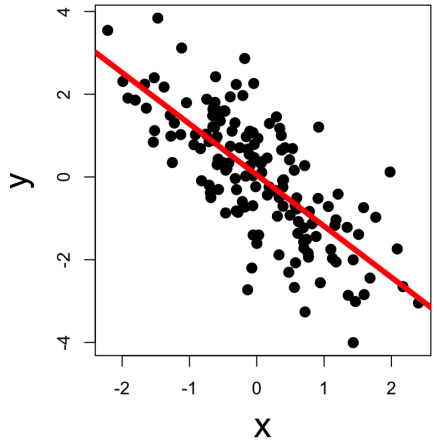


Every data point has one. Some are big, some small, some are positive (data above the line), some negative (data below the line).

how it started vs. how it's going

x	y
2.4	7.8
3.6	−1.1
10.0	4.3
⋮	⋮

$\Longrightarrow$

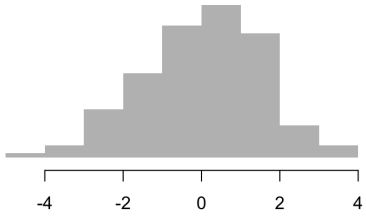


$\Longrightarrow$

$\hat{\varepsilon}$
−2.0
10.0
0.1
⋮

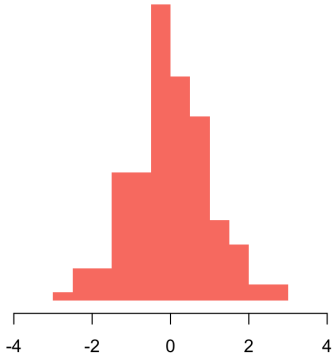
$\downarrow$ $\downarrow$

Distribution of y



"The mess the data made."

Distribution of $\hat{\varepsilon}$



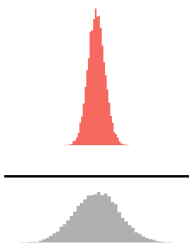
"The leftover after the model tries to clean up."

So, what is R^2? A Rotten Tomatoes score for models

$$R^2 = \begin{array}{c}\text{proportion of}\\ \text{variation explained}\end{array} = 1 - \begin{array}{c}\text{proportion of}\\ \text{variation } \textbf{unexplained}\end{array}$$

$$= 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

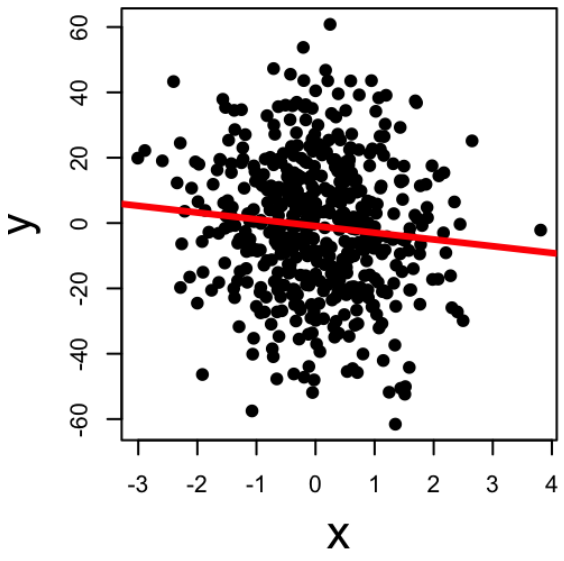
$$= 1 - \frac{}{}$$

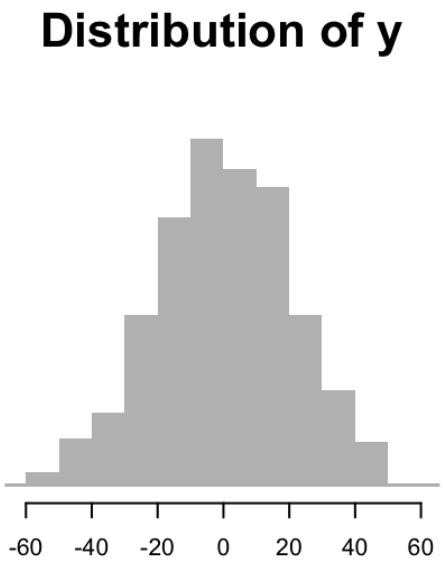


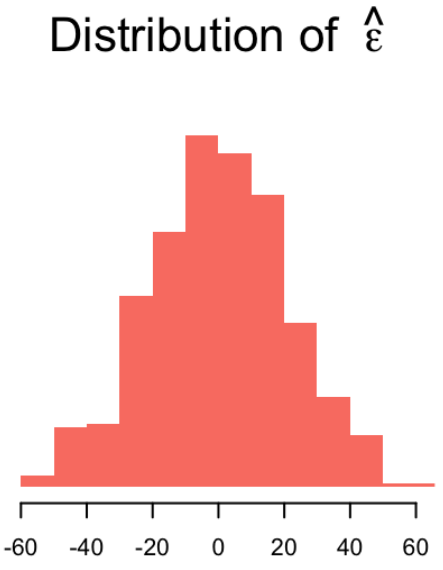
$$= 1 - \frac{\text{spread of leftover}}{\text{spread of original mess}}$$

$$= 1 - \frac{\text{var}(\hat{\varepsilon}_i)}{\text{var}(y_i)}.$$

Example: $R^2 \approx 0$



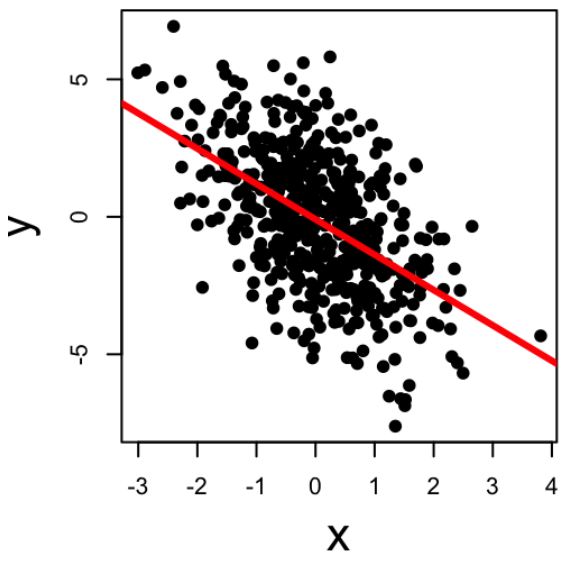


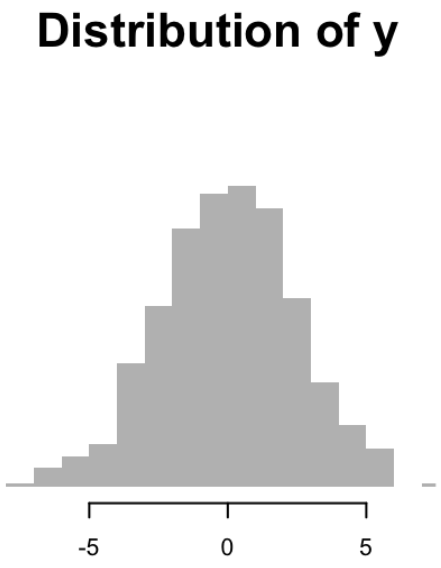


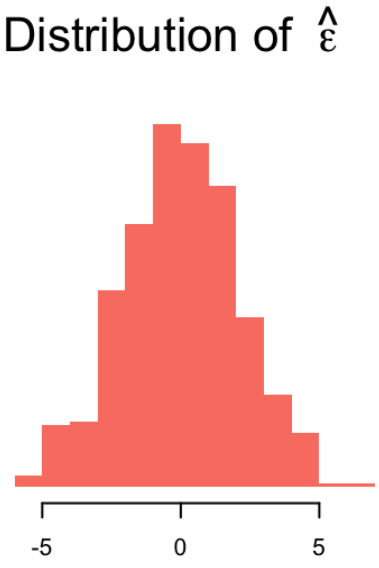
$$\text{var}(\hat{\varepsilon}_i) = \text{var}(y_i) \implies R^2 = 1 - \frac{\text{var}(y_i)}{\text{var}(y_i)} = 1 - 1 = 0.$$

Horrifically awful fit. The model didn't explain (clean up) anything.

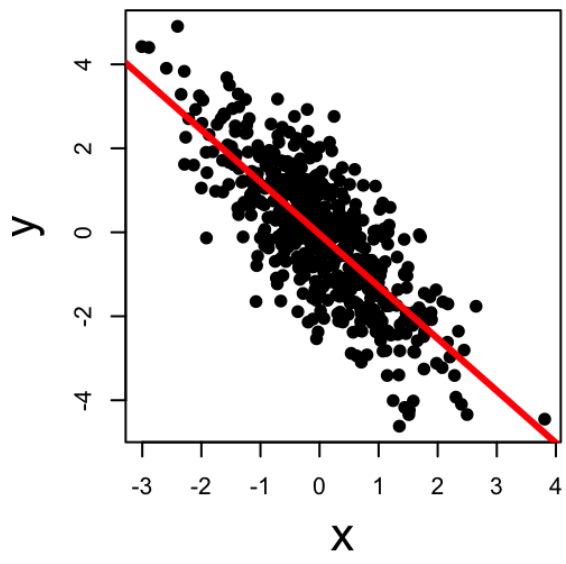
Example: $R^2 \approx 0.25$

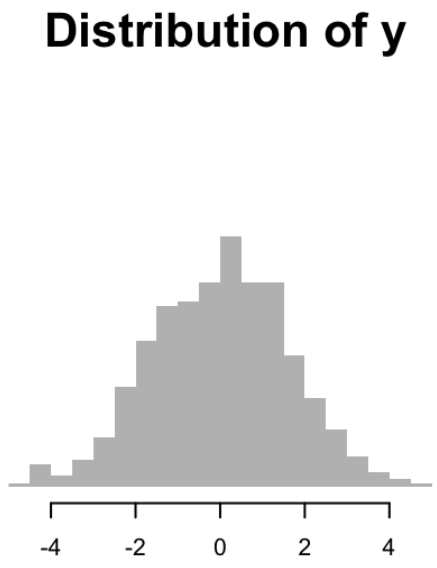


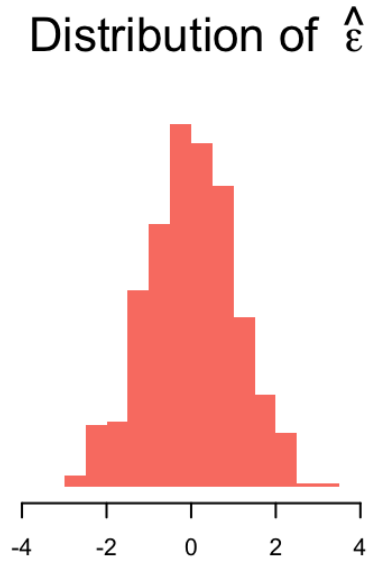




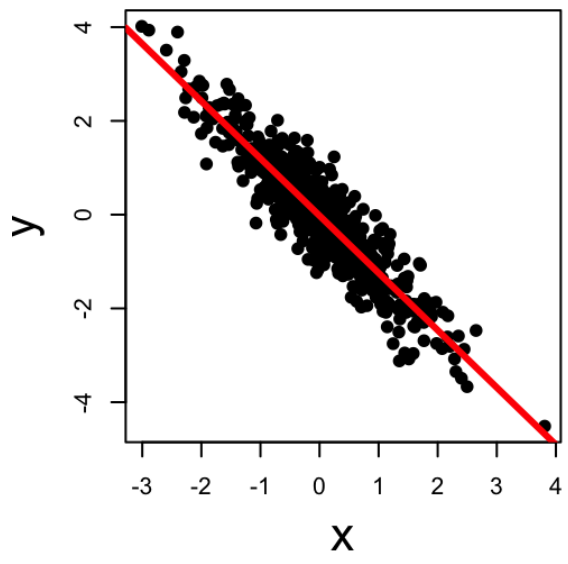
Example: $R^2 \approx 0.5$

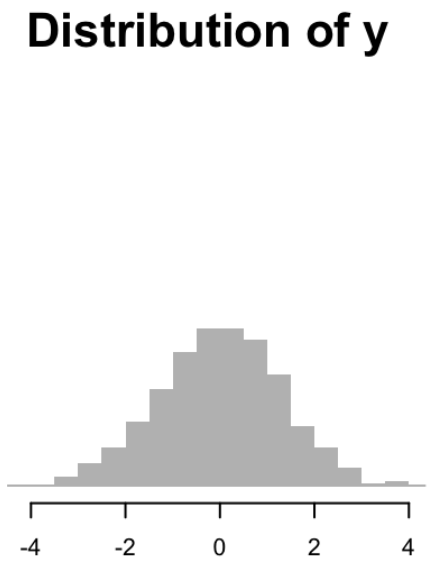


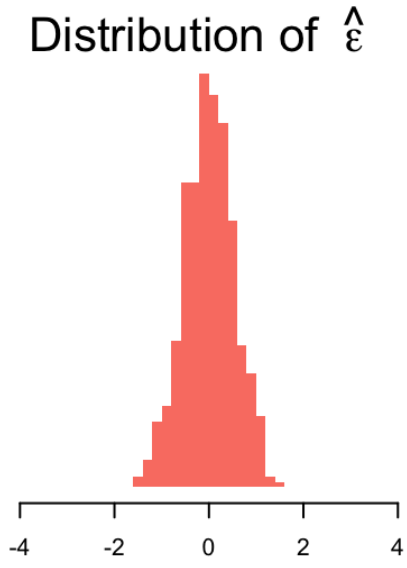




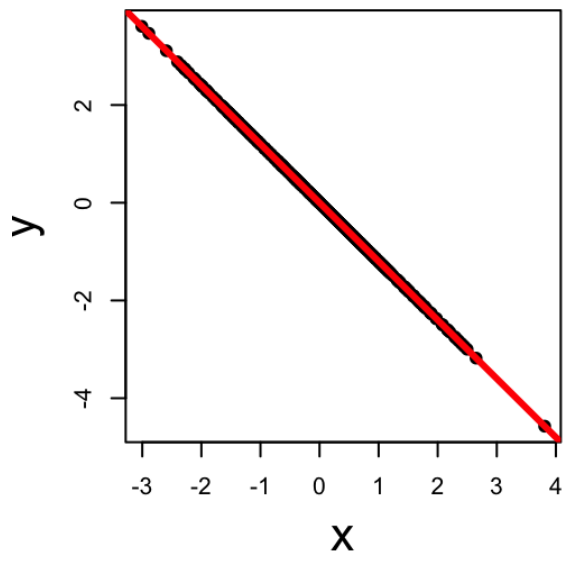
Example: $R^2 \approx 0.85$

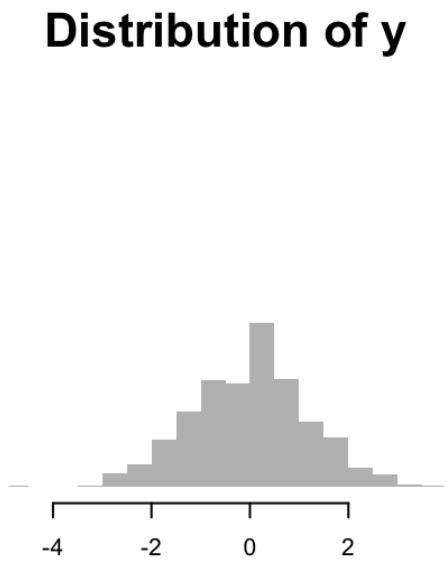


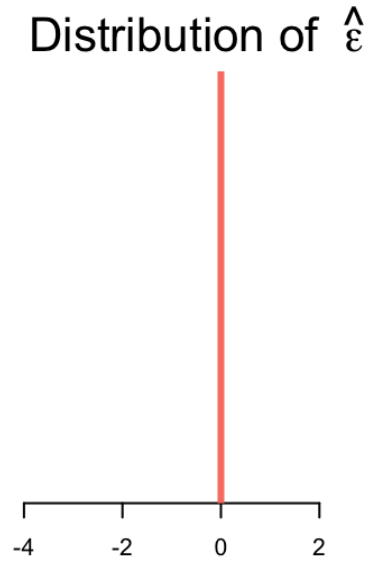




Example: $R^2 = 1$







$$\text{var}(\hat{\varepsilon}_i) = 0 \implies R^2 = 1 - \frac{0}{\text{var}(y_i)} = 1 - 0 = 1.$$

Perfect fit. The model explained (cleaned up) everything.

Adjusted R^2

- R^2 has a nasty mathematical property that it *always goes up* every time you add *any* predictor to the model, even if that predictor is silly and useless;

- Adjusted R^2 is...an adjusted version of R^2 that penalizes the number of predictors in the model.

- Adjusted R^2 is preferable for comparing models.

  Philosophically, we want to select a model that *both*...

  (i) fits/predicts well, *and*...
  (ii) is as simple as possible (so we can understand it).

  Key words: parsimony, Occam's razor, etc.