Welcome to STA 101!

Statistics is a confrontation with uncertainty.

Statistics confronts uncertainty by quantifying it.

Data analysis

Transforming messy, incomplete, imperfect data into knowledge.

This knowledge usually takes the form of:

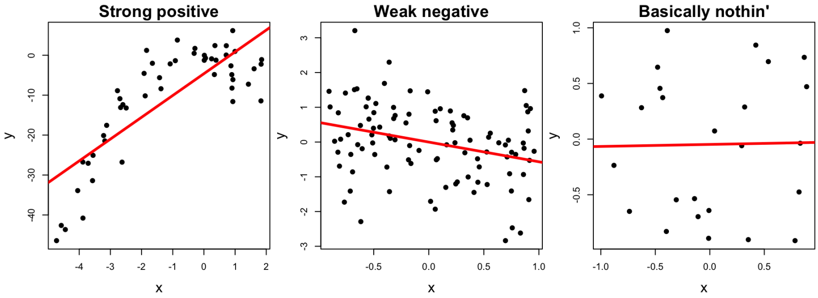
- pictures;
- a concise set of numerical summaries.

Statistical inference

Quantifying our uncertainty about that knowledge:

- **Question**: Given data, what's our best guess at some quantity of interest?
- **Answer**: best-guess $\pm$ margin-of-error

Interpreting the coefficient estimates in simple regression

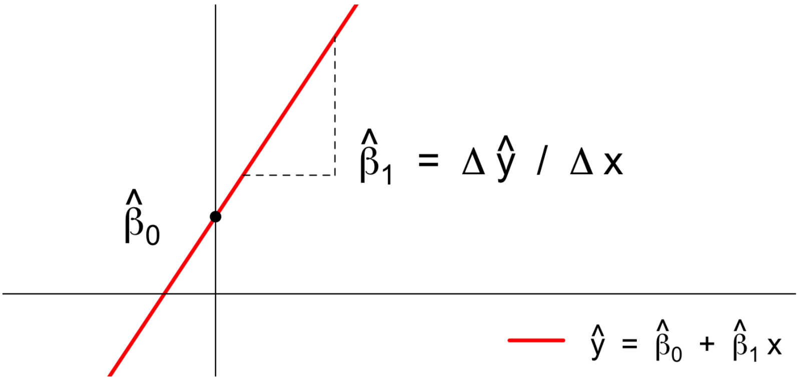


- The sign of $\hat{\beta}_1$ tells you if the variables are positively or negatively correlated;

- The magnitude of $\hat{\beta}_1$ tells you something about the strength of the general association.

Note: The magnitude *does not* tell you about the strength of the correlation per se. I misspoke in a previous class.

Interpreting the coefficient estimates in simple regression



- $\hat{\beta}_0$ is your prediction if $x = 0$ (may not always make sense);
- $\hat{\beta}_1$ is the change in the prediction if x increases by one unit;

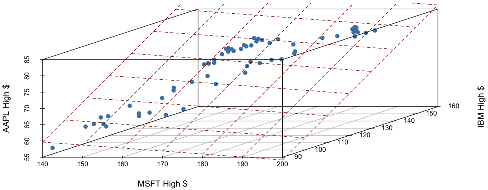
  Note: $\hat{\beta}_1$ is not the "causal impact" of x on y.

Interpreting the coefficients in multiple linear regression

Fitted multiple regression with two predictors:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

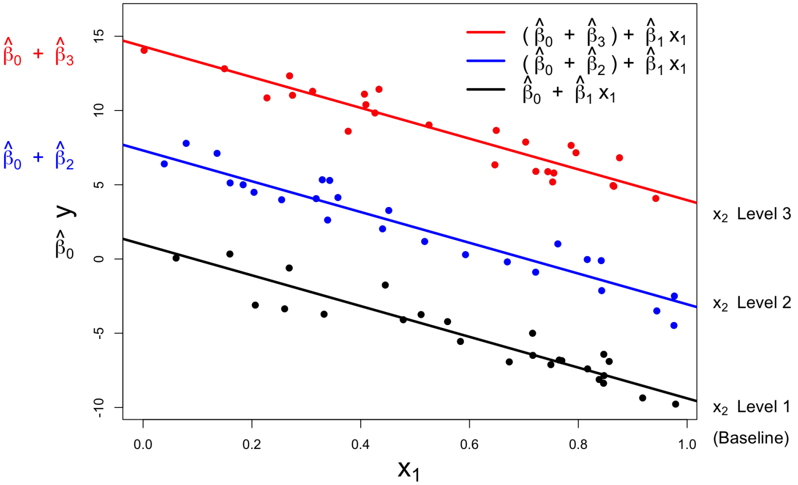
If everything is numerical, the picture looks like this (ick!):



- $\hat{\beta}_0$: if $x_1 = x_2 = 0$, what do we predict for y?
- $\hat{\beta}_1$: if x_2 does not change, and x_1 increases by 1, how does prediction for y change?
- $\hat{\beta}_2$: if x_1 does not change, and x_2 increases by 1, how does prediction for y change?

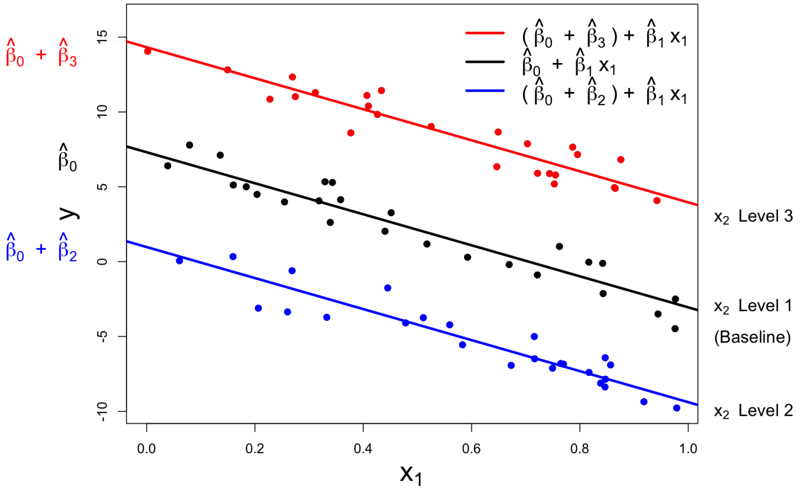
If x_2 is categorical, it's a little different

Parallel best fit lines, one for each level of the variable:



Same slope across the board, but shifting intercept.

If x_2 is categorical, it's a little different



$\hat{\beta}_2$ is *negative*. Level 2 is shifted down from the baseline. Baseline doesn't mean "the lowest one," it means the one we're starting from when we shift.

Variable selection

"Problem": I have a lot of variables (columns) in my data set.

Question: which ones do I include in the model?

Solutions:

- Ehh, just dump 'em all in there;
- None! I hate statistics!
- Only the most important ones! (which are those? important?)

Competing concerns:

- we want a model to predict well;
- we also want it to be "simple" enough that a human can understand why it behaves how it behaves.

Today: Find the set of variables that gives highest *adjusted* R^2.

Variable selection

Today: Find the set of variables that gives highest *adjusted* R^2.

Unfortunately: This is easier said than done. If you have *a lot* of candidate predictors, the list of model gets huge and searching through it becomes hard. You cannot just list everything out.

Solution: stepwise selection methods.

Variable selection: backward elimination

Start with the *full model* (the model that includes all potential predictor variables). Variables are eliminated one-at-a-time from the model until we cannot improve the model any further.

Procedure:

1. Start with a model that has all predictors we consider and compute the adjusted R^2.
2. Next fit every possible model with 1 fewer predictor.
3. Compare adjusted R^2s to select the best model (highest adjusted R^2) with 1 fewer predictor.
4. Repeat steps 2 and 3 until adjusted R^2 no longer increases.

Variable selection: forward stepwise

Forward stepwise regression is the reverse of the backward elimination technique. Instead, of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that improve the model any further.

Procedure:

1. Start with a model that has no predictors.
2. Next fit every possible model with 1 additional predictor and calculate adjusted R^2 of each model.
3. Compare adjusted R^2 values to select the best model (highest adjusted R^2) with 1 additional predictor.
4. Repeat steps 2 and 3 until adjusted R^2 no longer increases.

Review: model fit and R^2

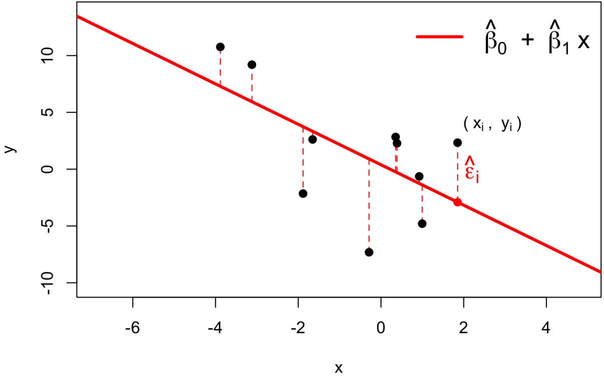
- Quality of fit appears to have something to do with how "all over the place" the residuals are;

- 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

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

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

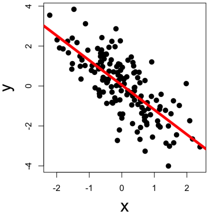


Every data point has one. Some are big, some small, some are positive (data above the line), some negative (data below the line).

how it started vs. how it's going

x	y
2.4	7.8
3.6	-1.1
10.0	4.3
$\vdots$	$\vdots$

$\Longrightarrow$

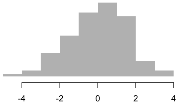


$\Longrightarrow$

$\hat{\varepsilon}$
-2.0
10.0
0.1
$\vdots$

$\downarrow$

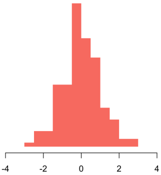
Distribution of y



"The mess the data made."

$\downarrow$

Distribution of $\hat{\varepsilon}$

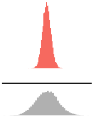


"The leftover after the model tries to clean up."

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

R^2 = proportion of variation explained (number between 0 and 1)

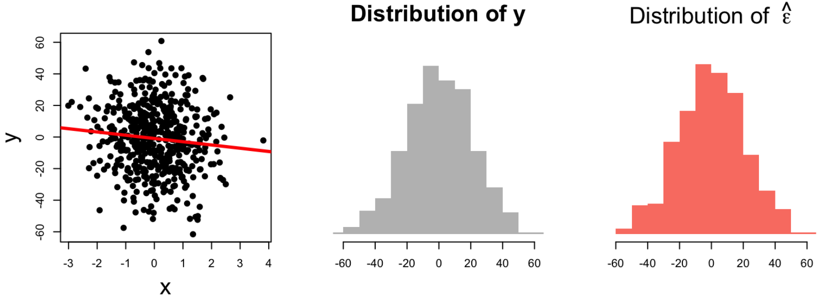
$$= 1 - \frac{\texttt{how it's going}}{\texttt{how it started}}$$

$= 1 -$ 

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

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

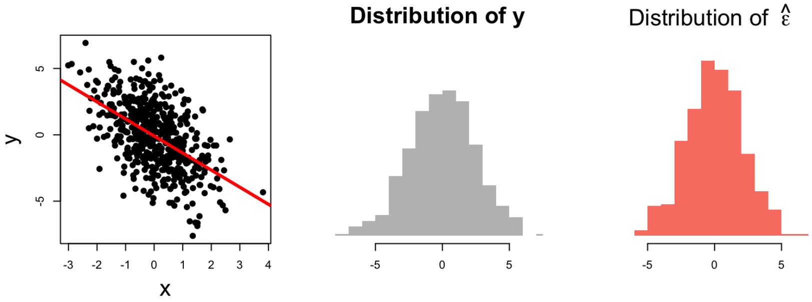
Example: $R^2 \approx 0$



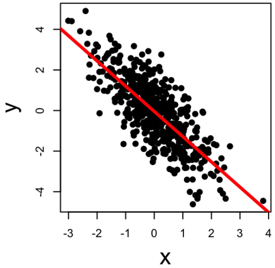
$$\text{var}(\hat{\varepsilon}_i) = \text{var}(y_i) \implies R^2 = 1 - \frac{\text{var}(y_i)}{\text{var}(y_i)} = 1 - 1 = 0.$$

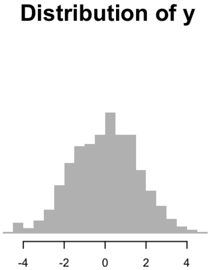
Awful fit. The model didn't explain (clean up) anything.

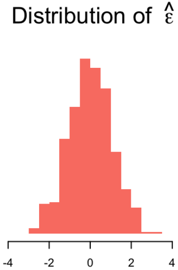
Example: $R^2 \approx 0.25$



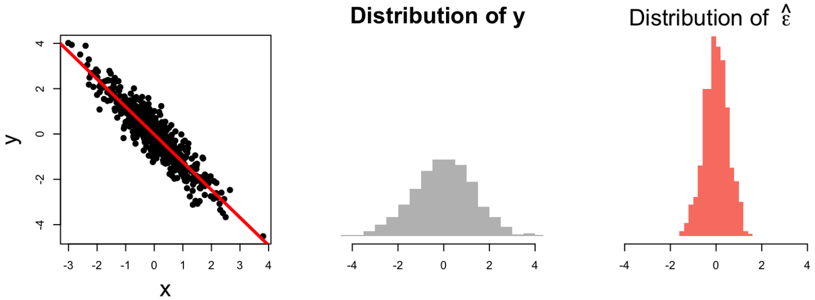
Example: $R^2 \approx 0.5$



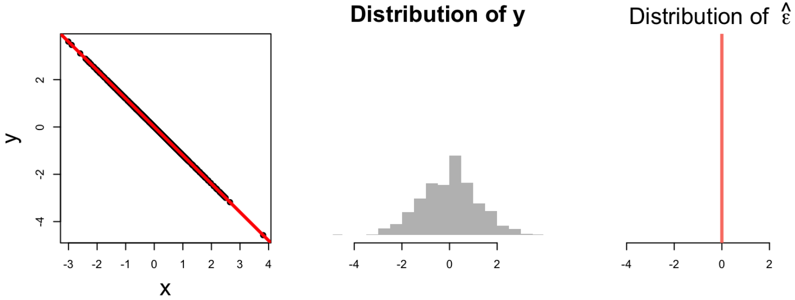




Example: $R^2 \approx 0.85$



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.