Inference in Regression

Administrivia

- **Homework 6** posted. Due Friday after Break.
- Good **Milestones**:
  - Problems 1-3 this week
  - Problems 4-6 the week after break

Simple Linear Regression

Given data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, fit a simple linear regression of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad \text{where} \qquad \epsilon_i \sim N(0, \sigma^2)$$

Estimates of the intercept and slope parameters are estimated by minimizing

$$SSE = \sum_{I=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

The least-squares estimate of the parameters are

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \text{and} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

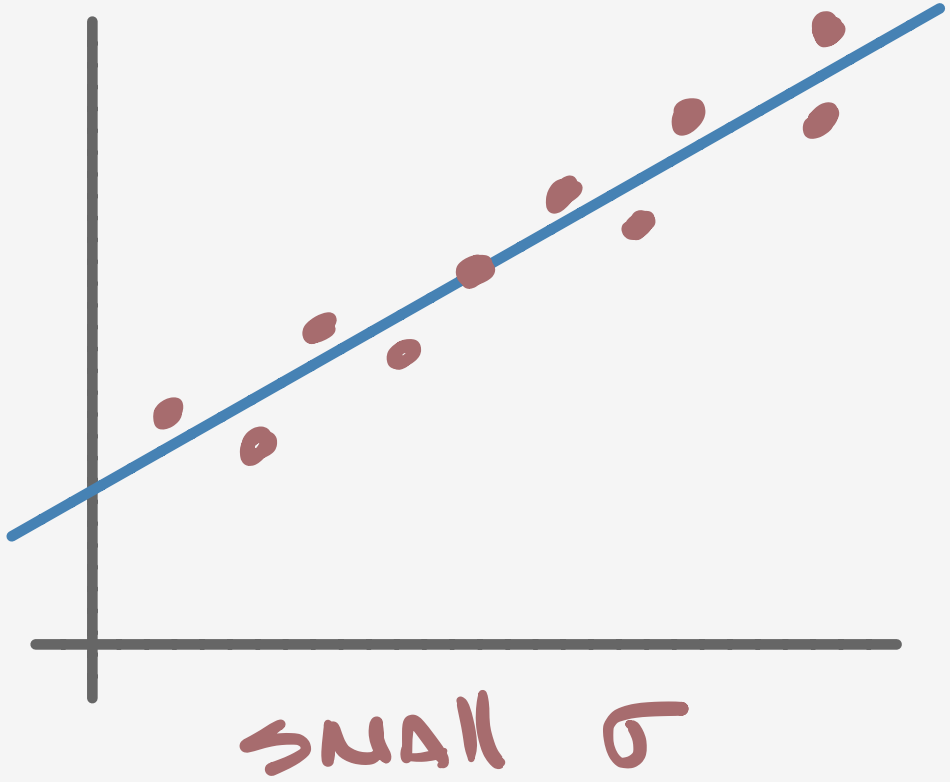
RoadMap

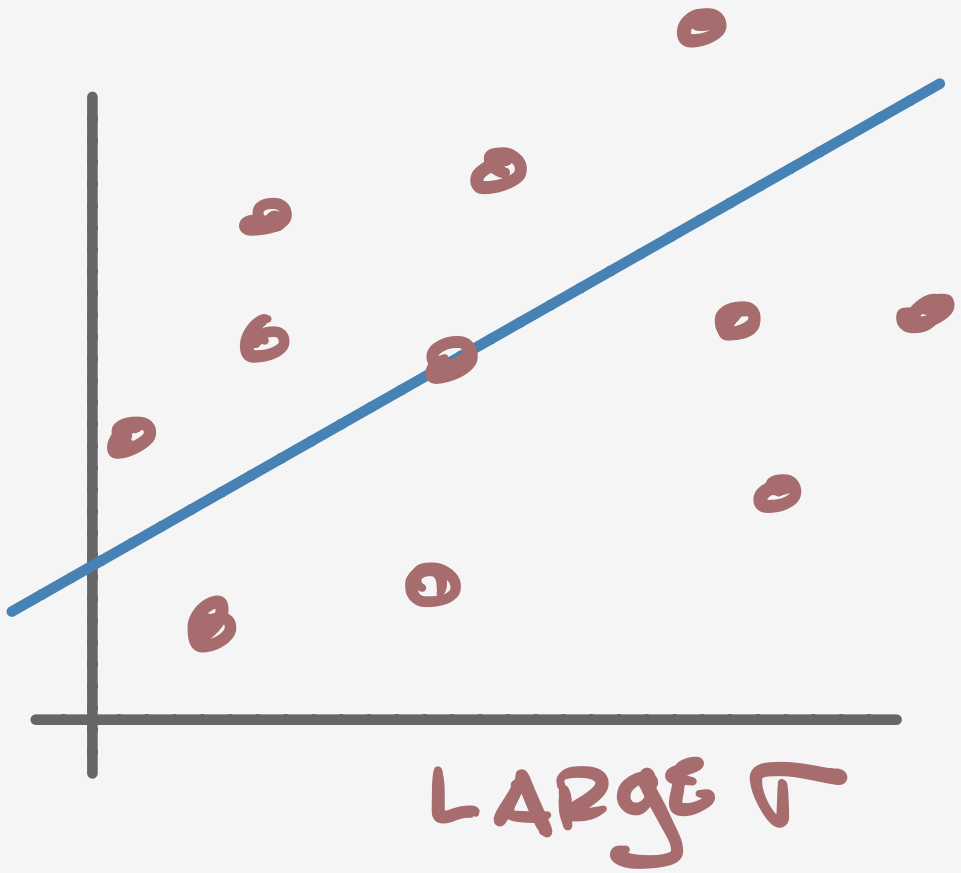
Today we'll see how we can:

- Estimate the variance in the data about the true regression line
- Quantify the goodness-of-fit in our simple linear regression model
- Perform inference on the regression parameters

Estimating the Variance

The parameter σ^2 determines the spread of the data about the true regression line





Estimating the Variance

An estimate of σ^2 will be used in computing confidence intervals and doing hypothesis testing on the estimated regression parameters

Recall that the sum of squared errors is given by

$$SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Our estimate of the variance, $\hat{\sigma}^2$, is given by

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

Estimating the Variance

- The divisor (n-2) in the estimate $\hat{\sigma}^2$ is the number of degrees of freedom (df) associated with the estimate of SSE

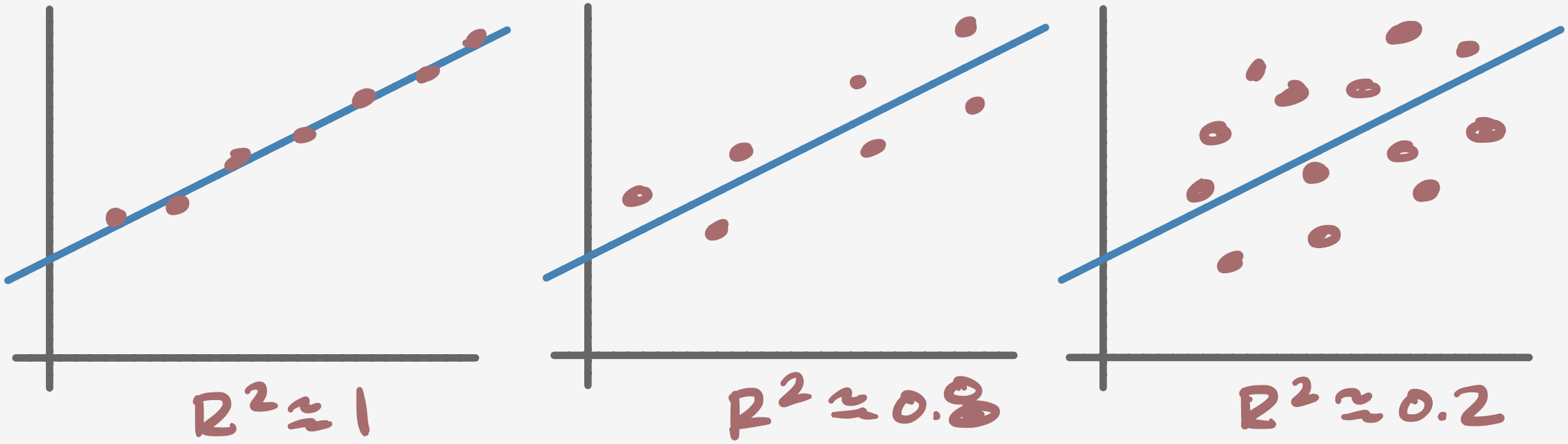
- This is because to obtain $\hat{\sigma}^2$, the two parameters $\hat{\alpha}$ and $\hat{\beta}$ must first be estimated, which results in a loss of 2 degrees of freedom

MEAN: $\bar{x} = \frac{1}{n}\sum_i x_i$

VAR: $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$

The Coefficient of Determination

- The coefficient of determination, R^2, quantifies how well the model explains the data



- R^2 is a value between 0 and 1.

The Coefficient of Determination

The sum of squares error $SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$

can be interpreted as a measure of how much variation in y is left unexplained by the model - that is, how much cannot be attributed to a linear relationship.

The regression sum of squares is given by $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

A quantitative measure of the total amount of variation in observed y values is given by the so-called total sum of squares

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ (DOES NOT DEPEND on x)

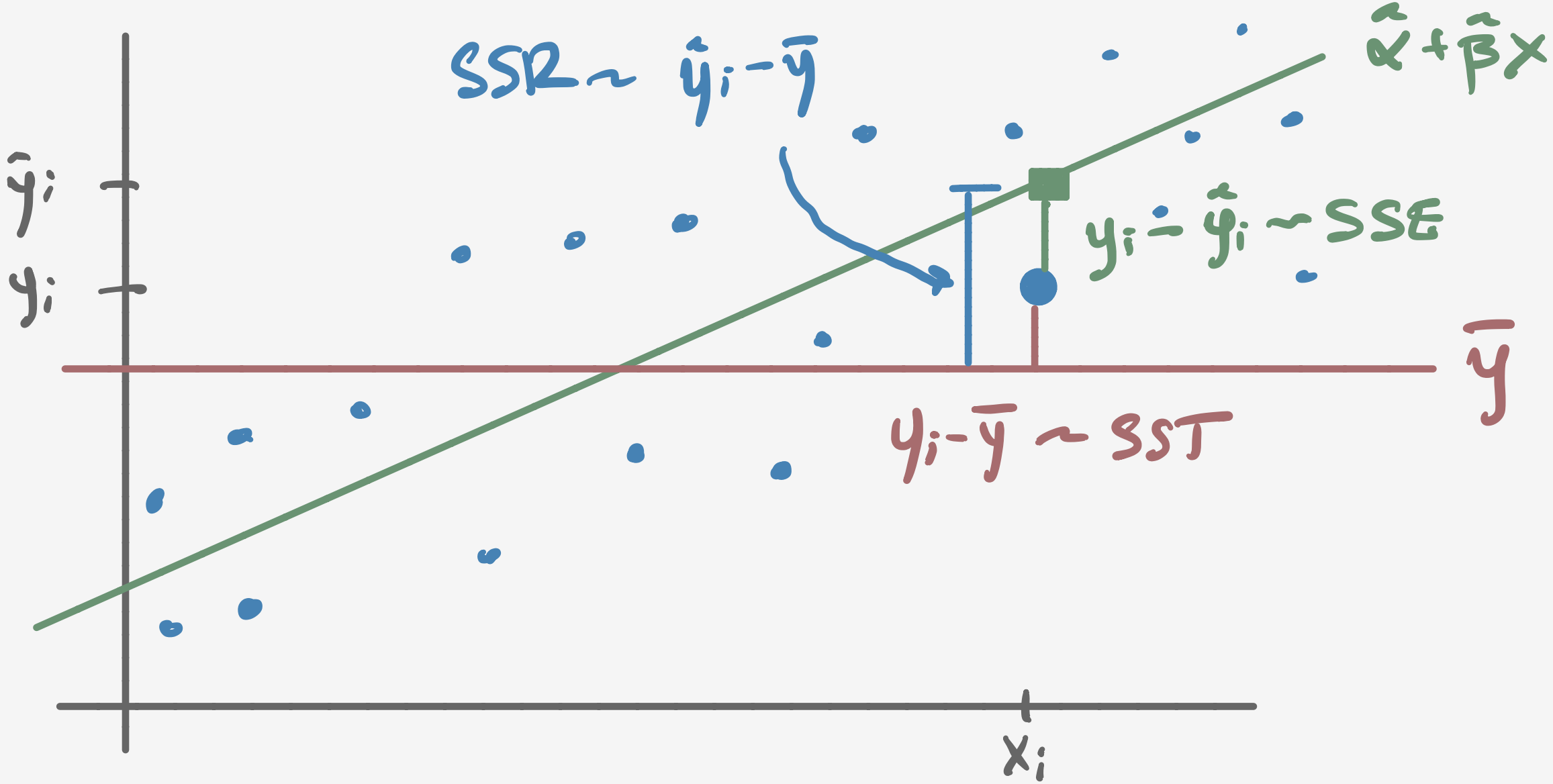
The Coefficient of Determination

The sum of squared deviations about the least-squares line is smaller than the sum of squared deviations about any other line, i.e. SSE < SST unless the horizontal line itself is the least-squares line

The ratio SSE/SST is the proportion of total variation in the data that cannot be explained by the simple linear regression model, and the coefficient of determination is

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

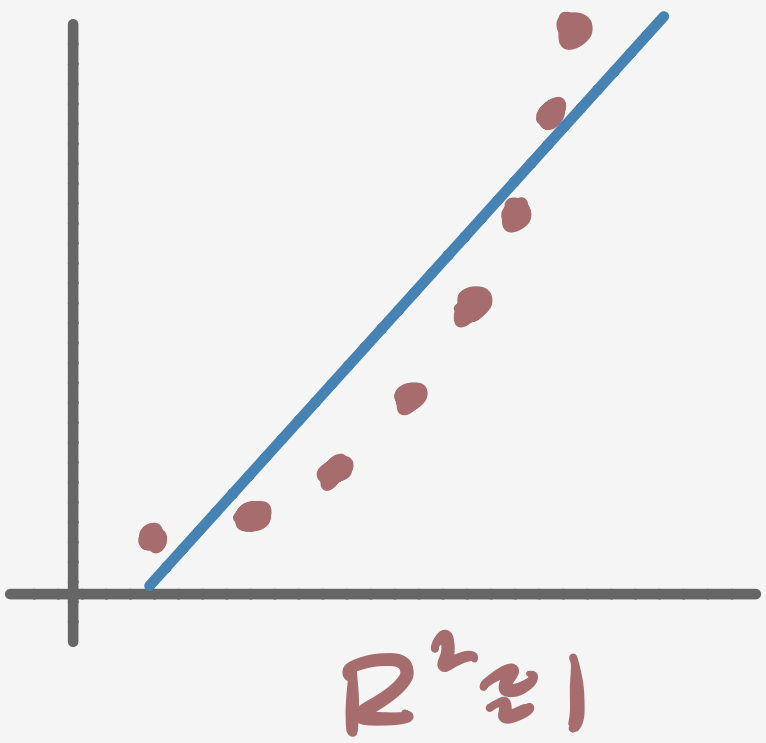
The Coefficient of Determination

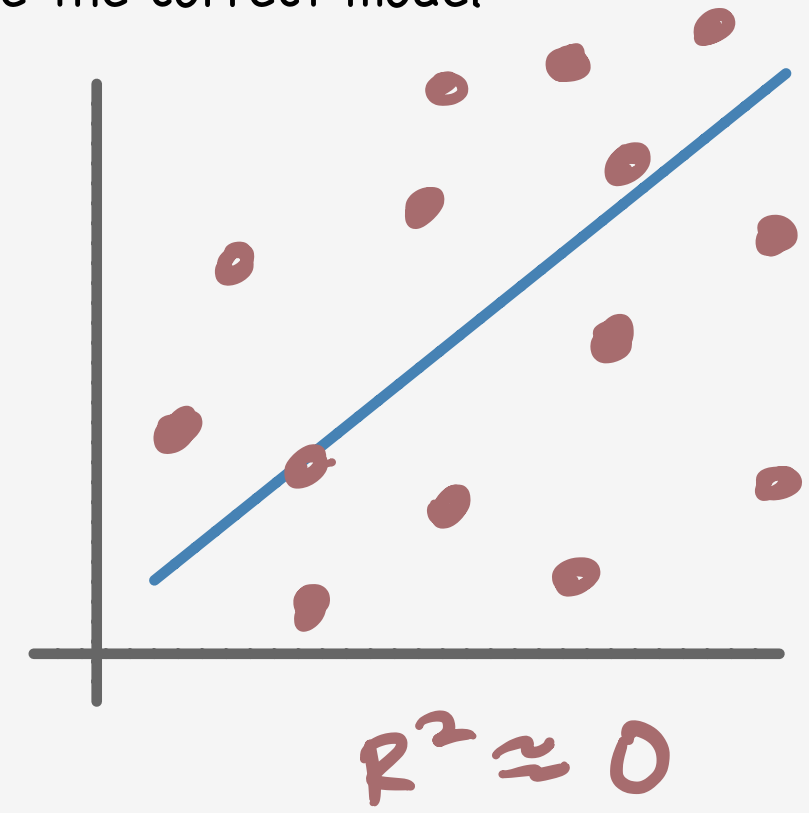


The Coefficient of Determination

Note: R^2 is the proportion of total variation in the data that is explained by the model

R^2 does **NOT** tell you that you necessarily have the correct model





Inference about Parameters

The parameters in the simple linear regression model have distributions. From these distributions we can construct confidence intervals for the parameters, conduct hypothesis tests, etc.

We'll focus on inference about the slope parameter $\hat{\beta}$, because this allows us to answer questions like: *Is there really a relationship between the feature and the response*?

The distribution for the estimate of the slope is given by

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right) \qquad SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Inference about Parameters

Confidence Intervals: $100(1-\alpha)\%$ CI FOR β

$$\hat{\beta} \pm t_{\alpha/2, n-2} \times SE(\hat{\beta})$$

Hypothesis Testing:

$H_0: \beta = c$

$H_1: \beta \neq c$

$$t = \frac{\hat{\beta} - c}{SE(\hat{\beta})}$$

(TEST STAT)

COMPARE TO $t_{\alpha/2, n-2}$ OR COMPUTE P-VAL.

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 21 In-Class Notebook

Let's:

- Do some stuff!