Week 2

Cogs 109: Data Analysis and Modeling

Fall 2018

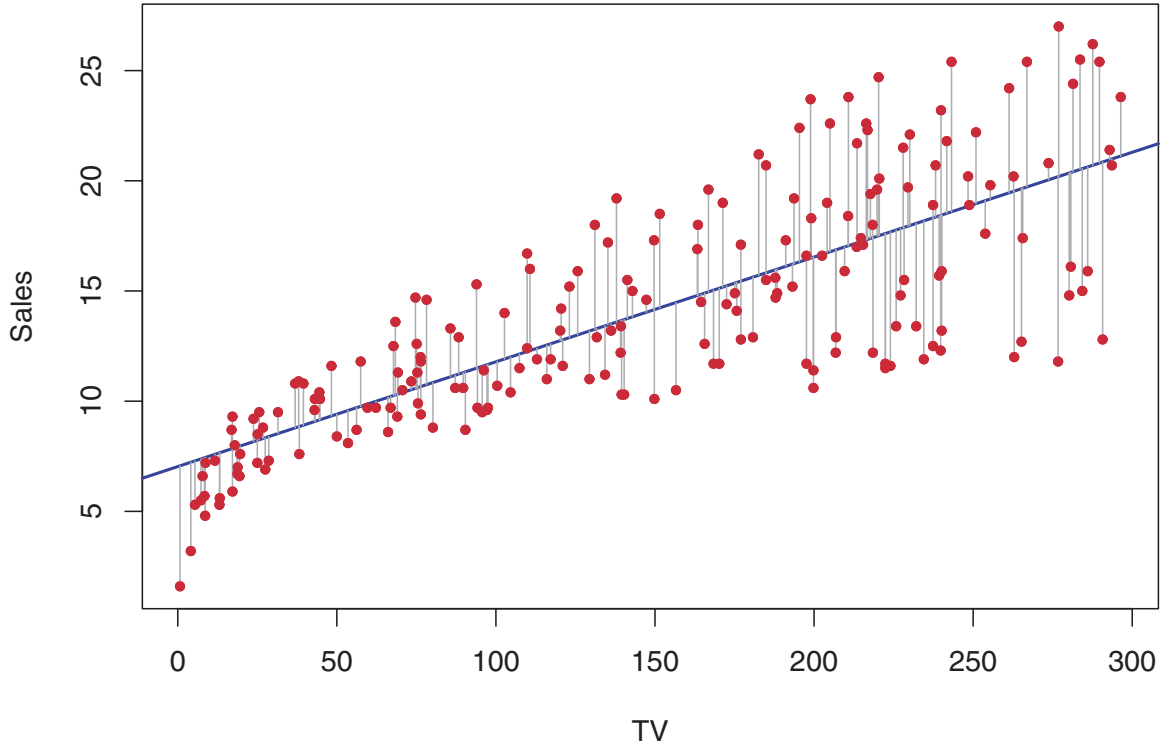
Prof. Eran Mukamel

Simple Linear Regression (1 predictor)

$$Y \approx \beta_0 + \beta_1 X.$$

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

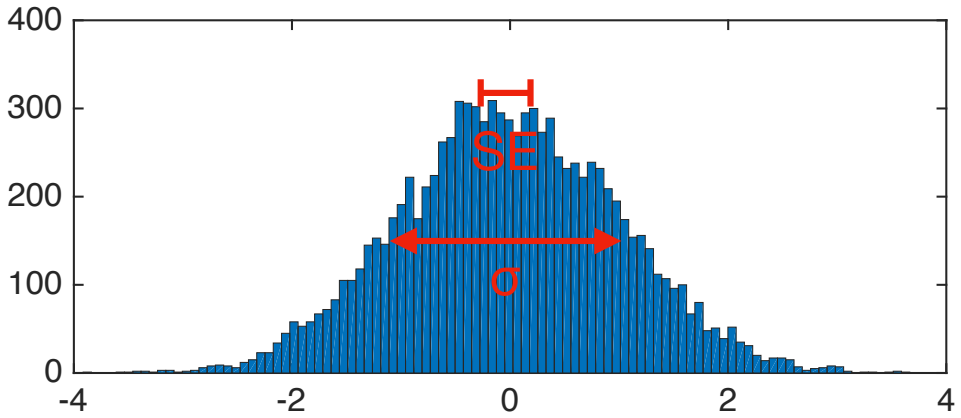


To find good estimates of beta0 and beta1, we need to minimize the *residual sum of squares (RSS)*. This is called ***Least Squares Regression***

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2, \quad = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

How good is our estimate?

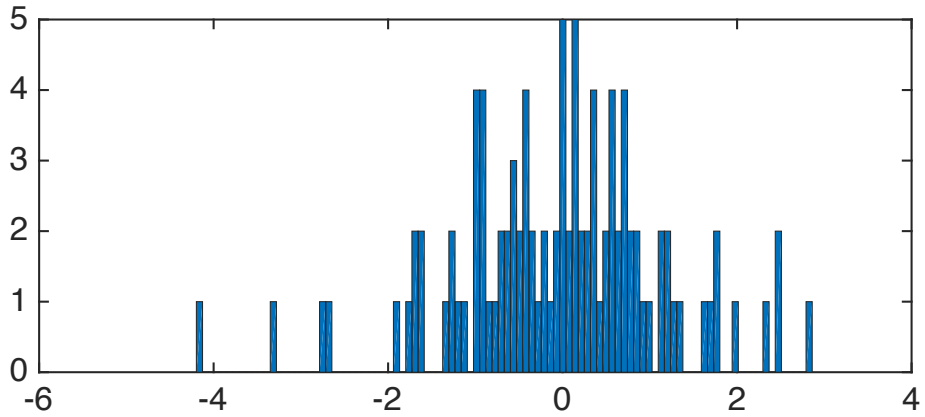
- Recall the difference between standard deviation (SD) and standard error (SE)
    - SD (σ) is the width of a distribution of observations
    - SE is the uncertainty of our estimate of a parameter based on a a set of *n* observations



$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

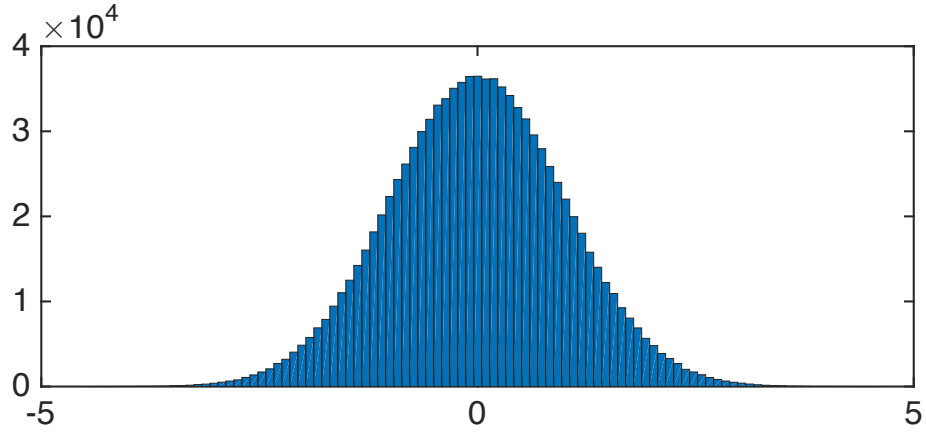
Example: Same SD, different SE

$\sigma = 1$



n=100

SE = 0.1



n = 1,000,000

SE = 10^{-3}

Standard error for regression parameters

- Depends on: σ, n, and also the distribution of x values
- When x values are more spread out, we have greater leverage for estimating parameters, so the SE is lower.

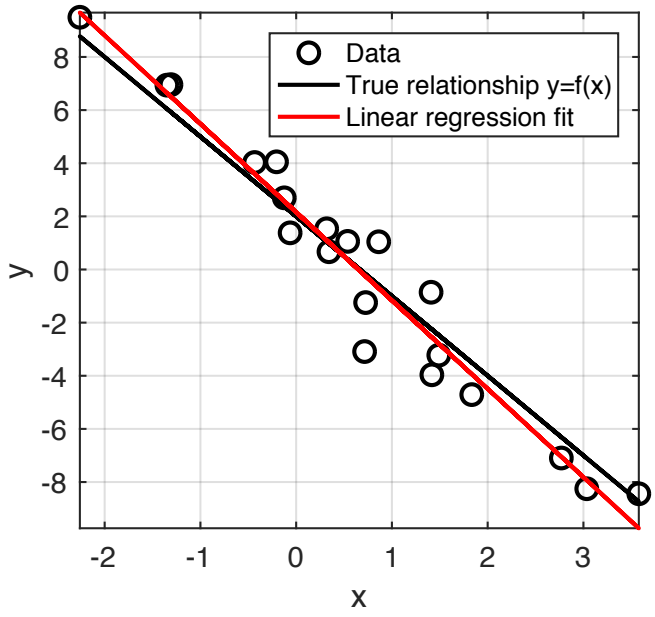
$$\sigma^2 = \mathrm{Var}(\epsilon)$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

Fitting a regression in MATLAB

- Use “fitlm” (fit linear model)

n = 20, σ = 1



$$y = f(x) = 2 - 3x + \varepsilon$$

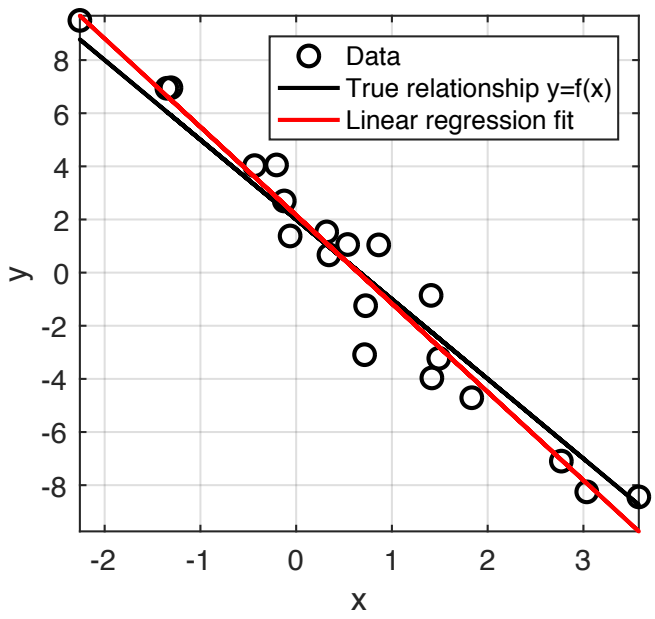
```
>> b = fitlm(x,y);
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    ________    _________

    (Intercept)      2.1584     0.28063      7.6915    4.2709e-07
    x1              -3.3248     0.17672     -18.814    2.7589e-13
```

$$\hat{y} = \hat{f}(x) = 2.1584 - 3.3248x$$

n = 20, σ = 1

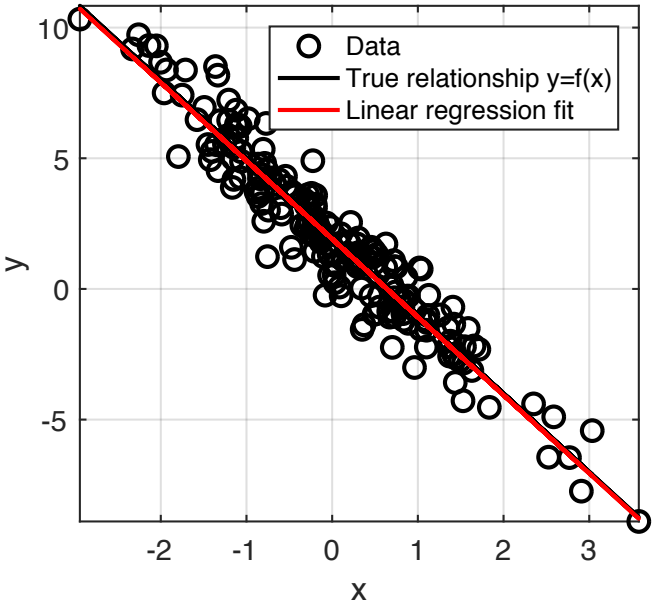


Linear regression model:
y ~ 1 + x1

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	2.1584	0.28063	7.6915	4.2709e-07
x1	-3.3248	0.17672	-18.814	2.7589e-13

n = 200, σ = 1



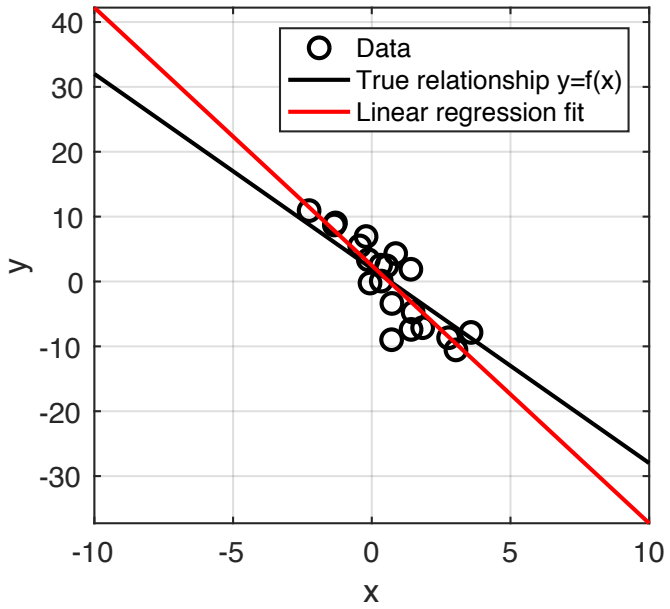
Linear regression model:
y ~ 1 + x1

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	1.912	0.06932	27.582	9.6104e-70
x1	-2.9906	0.06384	-46.846	4.3111e-109

Effect of "leverage"

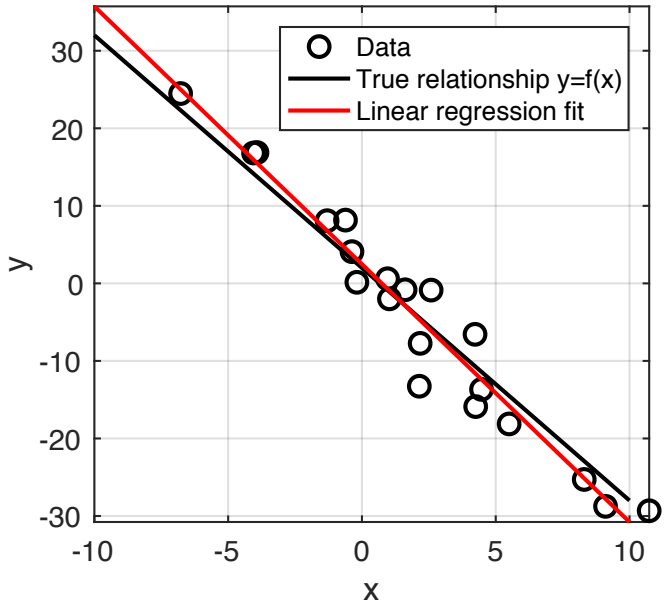
n = 20, σ = 3



If the x values are more spread-out, they provide more precise information about the slope

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	2.4753	0.84188	2.9402	0.0087498
x1	-3.9745	0.53015	-7.4969	6.1029e-07



Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	2.4753	0.84188	2.9402	0.0087498
x1	-3.3248	0.17672	-18.814	2.7589e-13

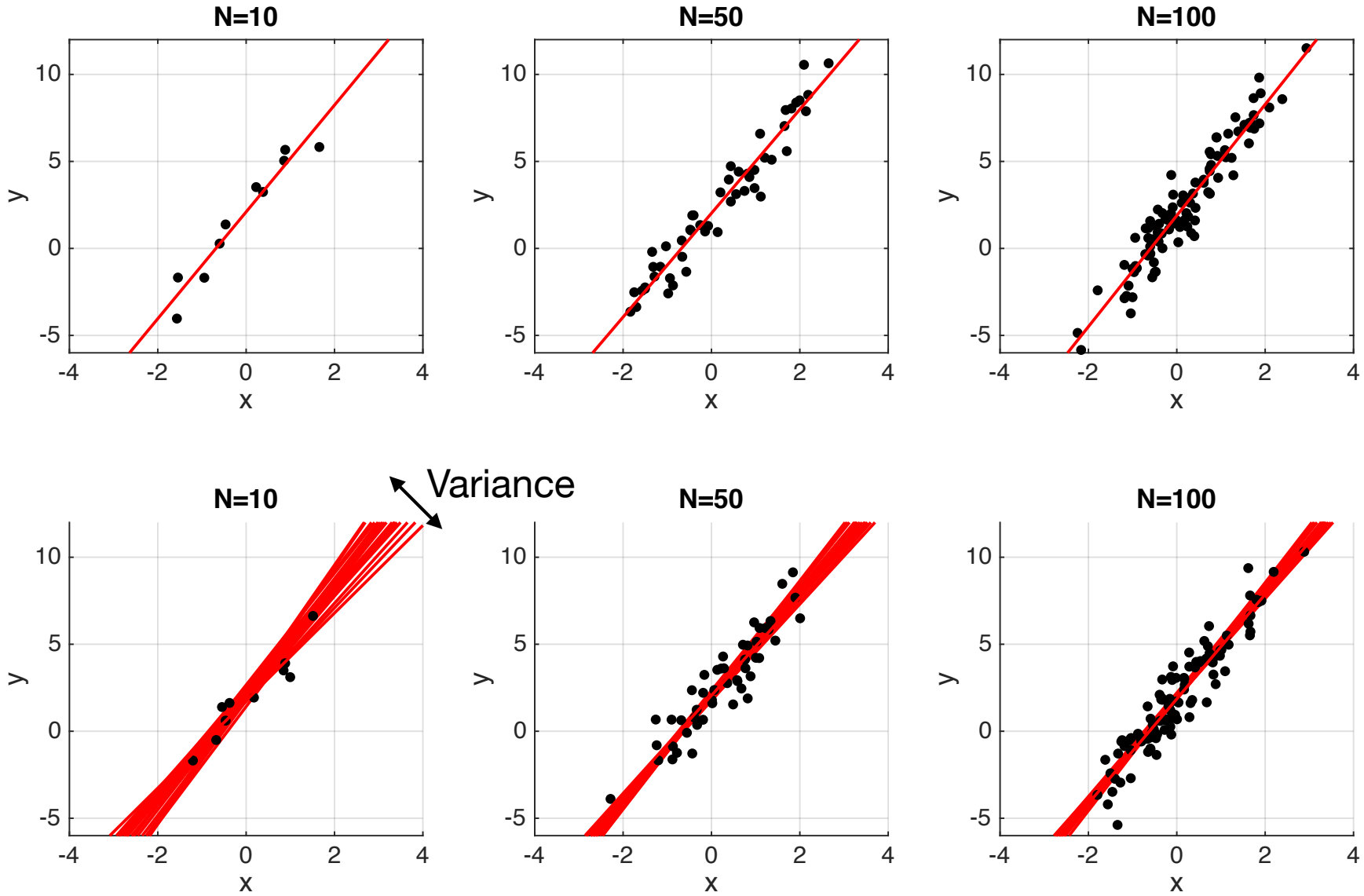
Standard error for regression parameters

- Depends on: σ, n, and also the distribution of x values
- When x values are more spread out, we have greater leverage for estimating parameters, so the SE is lower.

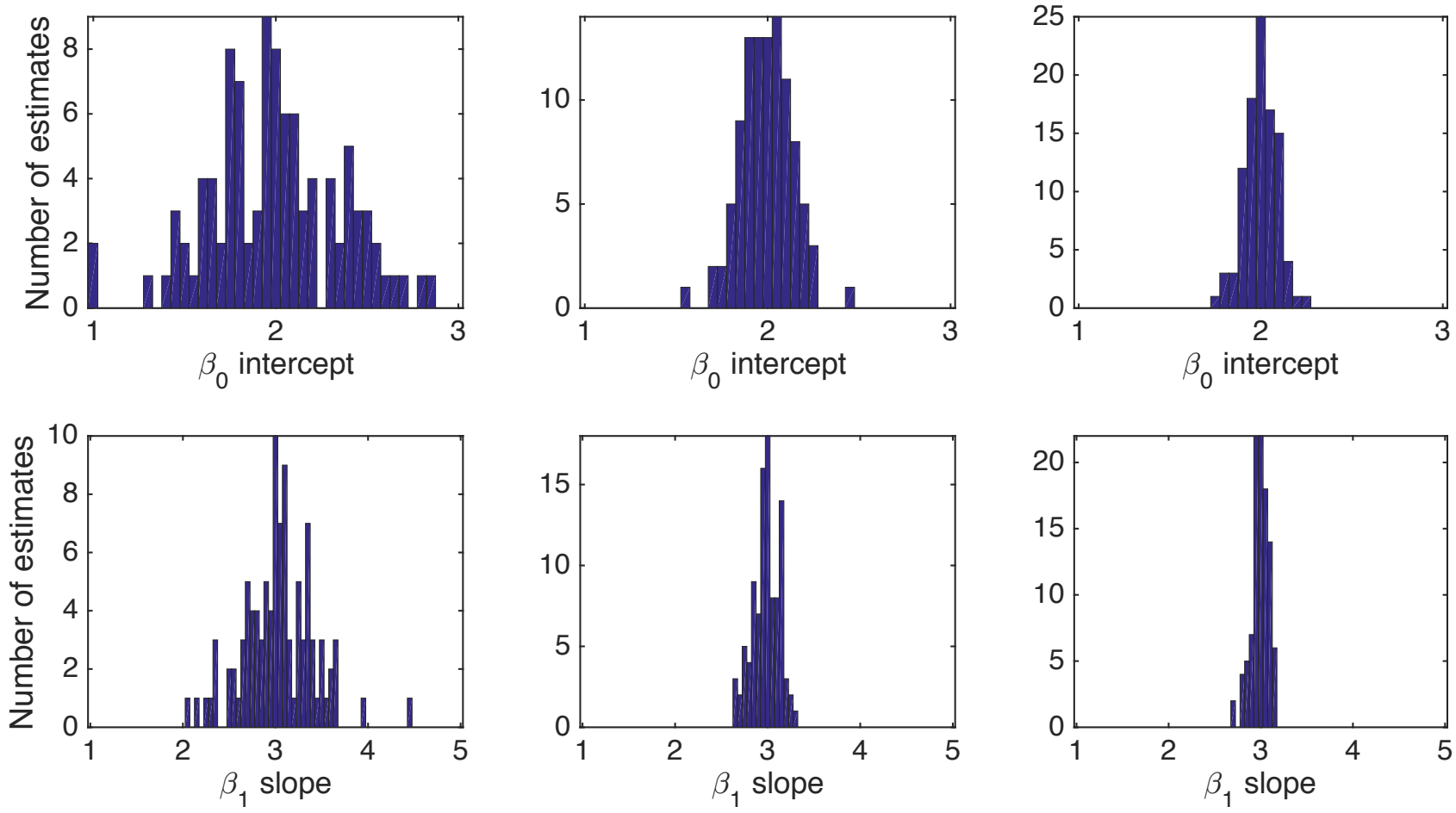
$$\sigma^2 = \mathrm{Var}(\epsilon)$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Total sum of squares for x values (predictors): Measures how "spread out" the x values are

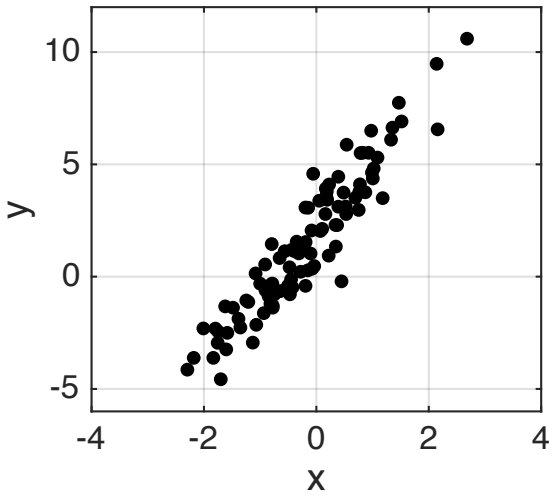

N=10
N=50
N=100
Variance
N=10
N=50
N=100
y
x
10
5
0
-5
-4
-2
0
2
4

Estimates of the parameters get better (lower variance) with higher N

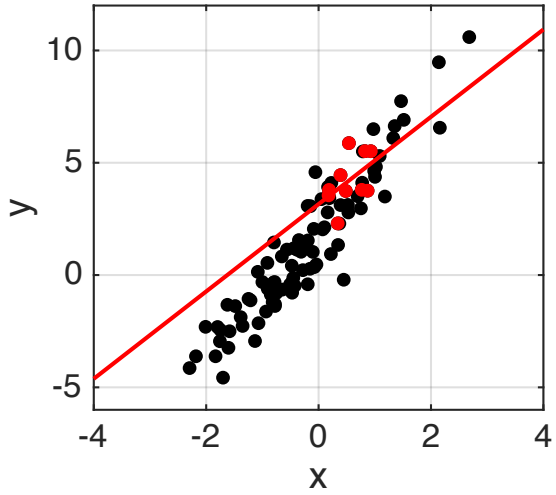


Example: Leverage

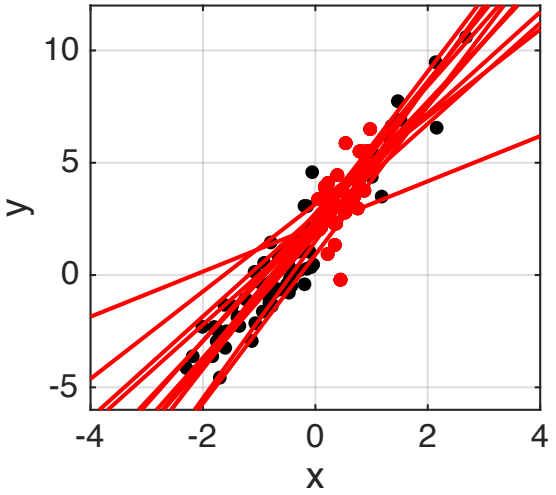
n=100



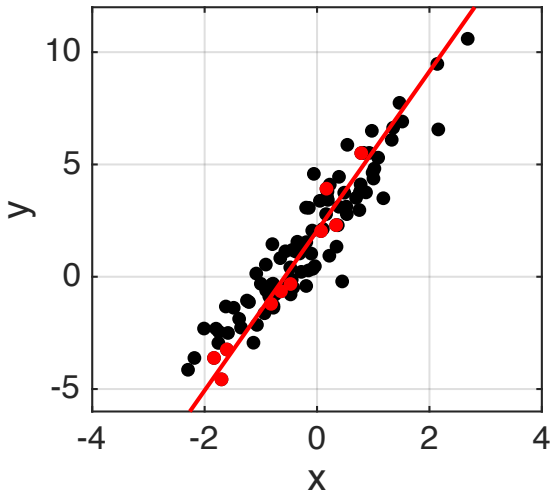
Fit using n=10 points
from a narrow range of x values [0,1]



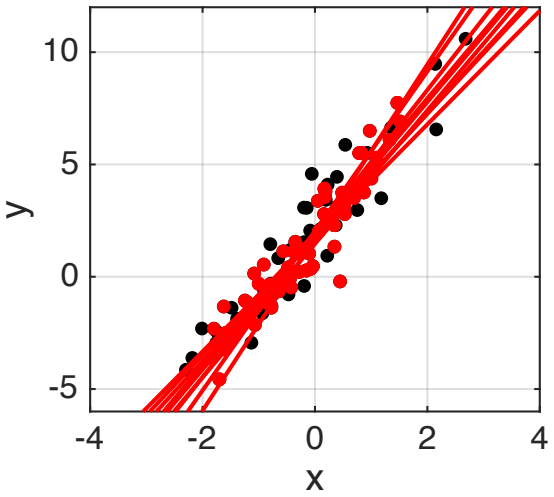
Many fits



Fit using n=10 points
from a broad range of x values [-2,2]



Many fits



Confidence intervals

- The regression coefficients are our best guess about the true values of the parameters
- The confidence interval defines a range of values that could be consistent with the data
- 95% confidence interval is given by ± 2 SE:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

- Note: This is based on some assumptions, but it is a good approximation as long as n is not too small

Using regressions for inference: Testing hypotheses

- Recall that inference means answering questions — or, testing hypotheses — about the relationships in the data
- Most basic question: Is there are a relationship between X and Y?
- To test this, use a null hypothesis H0 and an alternative hypothesis H1.

H_0 : There is no relationship between X and Y

H_a : There is some relationship between X and Y.

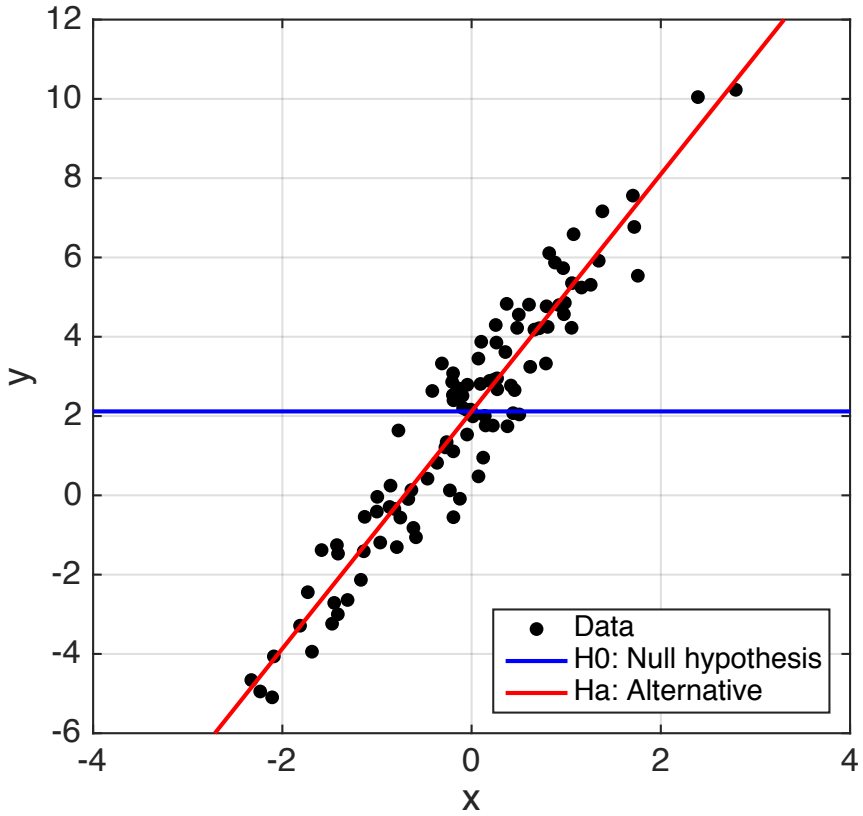
Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

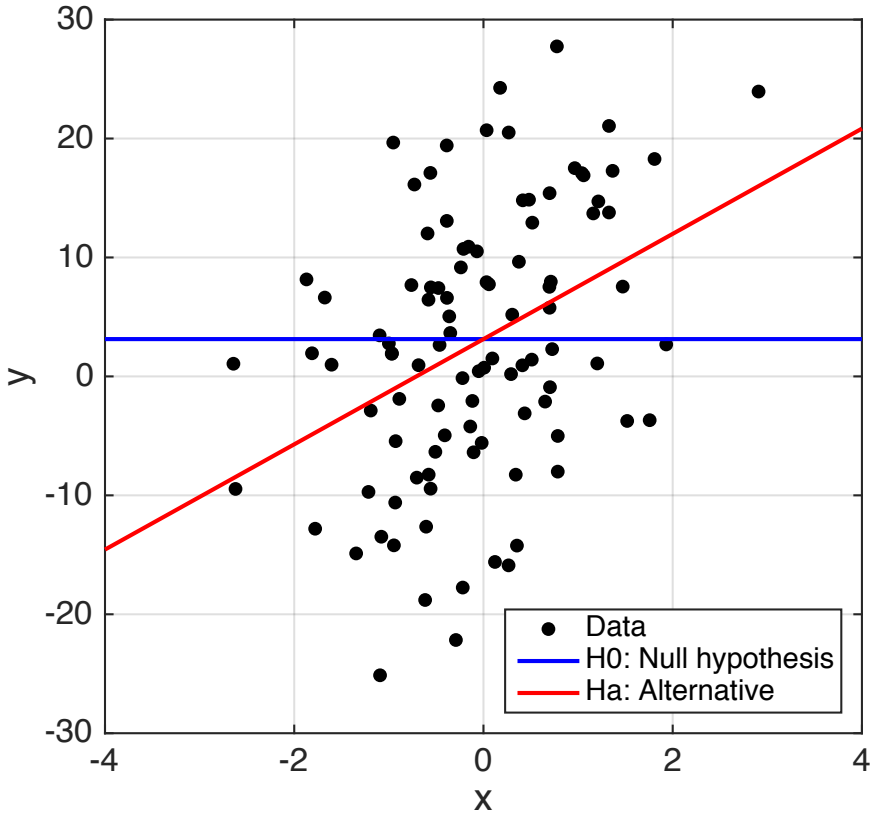
versus

$$H_a : \beta_1 \neq 0,$$

Low noise:
Ha fits much better than H0



High noise:
Ha does not improve much upon H0

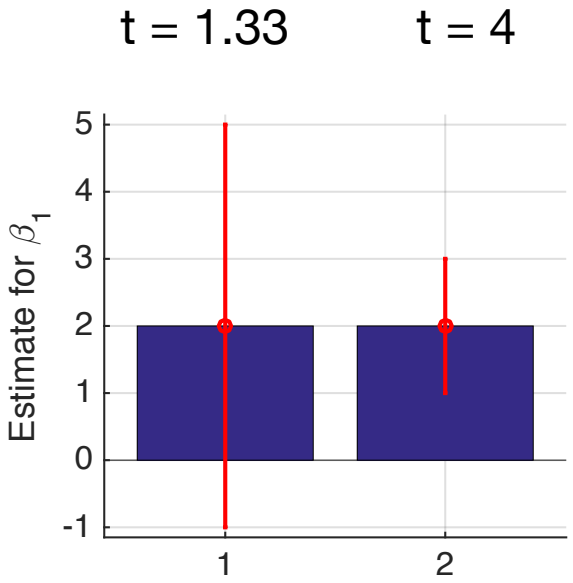


Testing hypotheses using t-statistics and p-values

- t-statistic: “The slope of the best fit regression line is t standard errors away from 0”

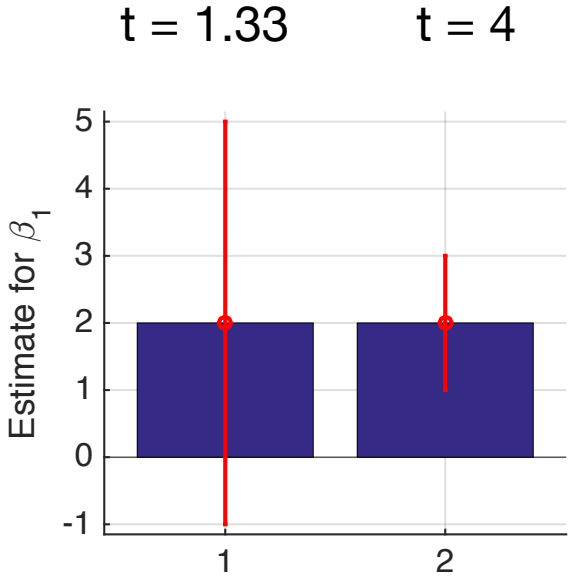
$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- If |t| < 2, the 95% confidence interval for the slope includes beta1=0
  - We can’t reject the null hypothesis!
- If |t| > 2, the 95% confidence interval for the slope does not include beta1=0
  - We can reject H0



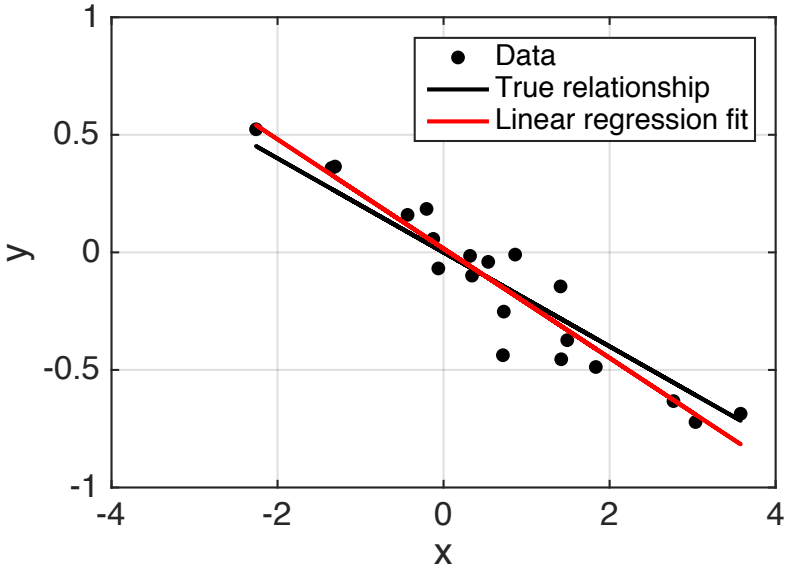
Testing hypotheses using t-statistics and p-values

- p-value: The **p**robability that we would have observed the data if the null hypothesis (H0) were true
  - $|t| > 1.96$ corresponds to $p < 0.05$
  - $|t| > 2.58$ corresponds to $p < 0.01$
- Note: This is NOT the "probability that the alternative hypothesis is true"
  - There are many reasons the null hypothesis may be false



Summary tables show parameter estimates, SE, t-statistic and p-value

$n = 20, \sigma = 0.1$



	Estimate	SE	tStat	pValue
(Intercept)	0.011361	0.025885	0.4389	0.66596
x1	0.16244	0.026655	6.0942	9.2986e-06

- Alison runs a study in which 10 first graders are given different amounts of apple juice in the morning. She measures their ability to concentrate (measured in minutes) later that day, and runs a regression to test whether the amount of apple juice is related to concentration.
- Consider the summary table below:
  - What is the null hypothesis?
  - What is the probability that these data would be observed if the null hypothesis were true?
  - What is the 95% confidence interval for the slope of the relationship between juice and concentration?
  - What can Alison conclude?
- Now suppose that Alison repeats this experiment in 100 different first grade classrooms. In each classroom she tests 10 children and fits a regression. She reports that in 7 of the classrooms there is a significant effect of juice on concentration ($p<0.05$). Do you agree that these data support providing juice to the kids in those 7 classes to boost their performance?

	Estimate	SE	tStat	pValue
(Intercept)	3.0000	0.025885	0.4389	0.66596
Juice (mL)	0.56244	0.026655	2.0537	0.040000

- Alison runs a study in which 10 first graders are given different amounts of apple juice in the morning. She measures their ability to concentrate (measured in minutes) later that day, and runs a regression to test whether the amount of apple juice is related to concentration.

- Consider the summary table below:

  - What is the null hypothesis? **H0: There is no effect of juice; the slope, beta1, is zero**

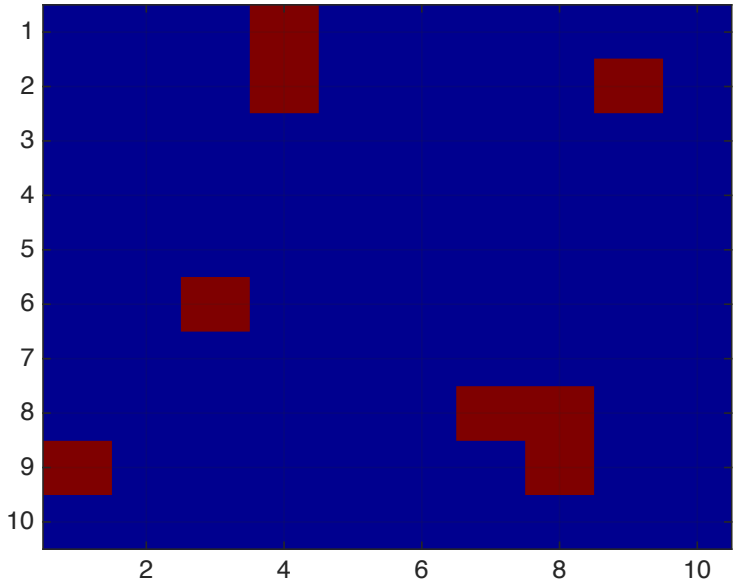
  - What is the probability that these data would be observed if the null hypothesis were true? **The p-value is 0.04, so there is a 4% probability that these data would be observed**

  - What is the 95% confidence interval for the slope of the relationship between juice and concentration? **The 95% CI is equal to beta1 ± 2SE = 0.56 ± 2*0.027 = [0.51 - 0.62]**

  - What can Alison conclude? **There is a significant effect of juice on concentration, with each additional unit of juice increasing concentration by ~0.56±0.05 minutes**

	Estimate	SE	tStat	pValue
(Intercept)	3.0000	0.025885	0.4389	0.66596
Juice (mL)	0.56244	0.026655	2.0537	0.040000

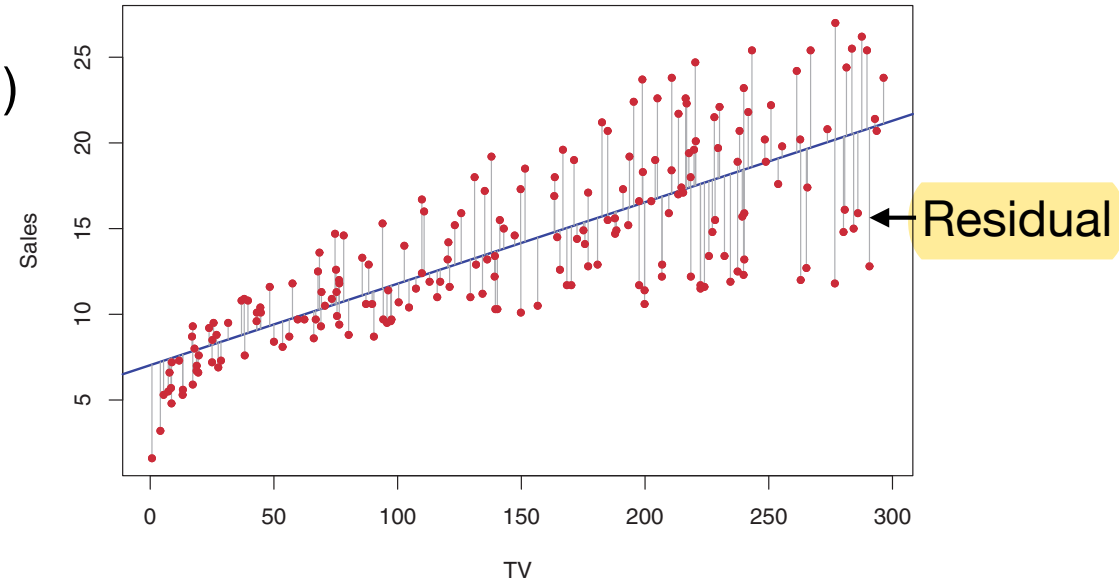
- Now suppose that Alison repeats this experiment in 100 different first grade classrooms. In each classroom she tests 10 children and fits a regression. She reports that in 7 of the classrooms there is a significant effect of juice on concentration ($p<0.05$). Do you agree that these data support providing juice to the kids in those 7 classes to boost their performance?
  - **NO!**
  - **If we define a "significant effect of juice" using a p-value threshold of 0.05, there is a 5% chance of false positive detection for EACH class**
  - **If we repeat this procedure in 100 classrooms, we can expect around ~100*0.05 = 5 false positives**
  - **This is called the problem of multiple comparisons, and we will come back to it.**



Model accuracy

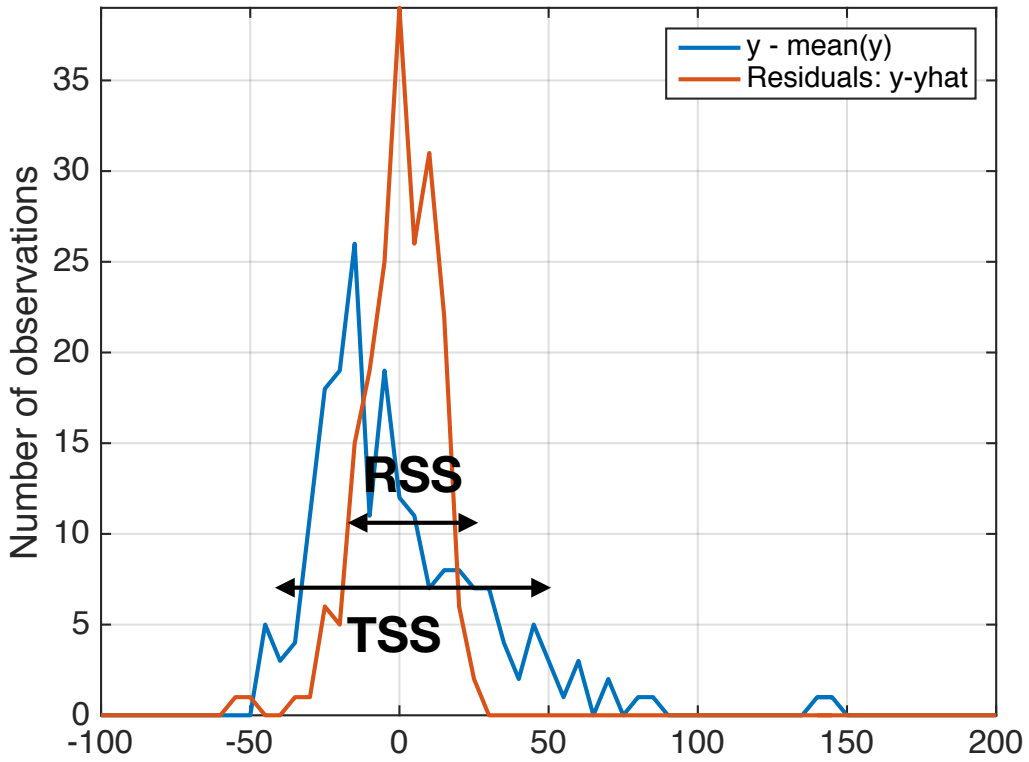
- R^2 is the proportion of the total variance (TSS: total sum of squares) that is explained by the model.
    - The part that is unexplained is the RSS: Residual sum of squares
- R^2 ranges from 0 to 1 (or 100%)

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$



- For a simple linear regression (with one predictor variable), R^2 = r^2 where r is the Pearson correlation coefficient

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$


y - mean(y)
Residuals: y-yhat
Number of observations
35
30
25
20
15
10
5
0
-100
-50
0
50
100
150
200
RSS
TSS

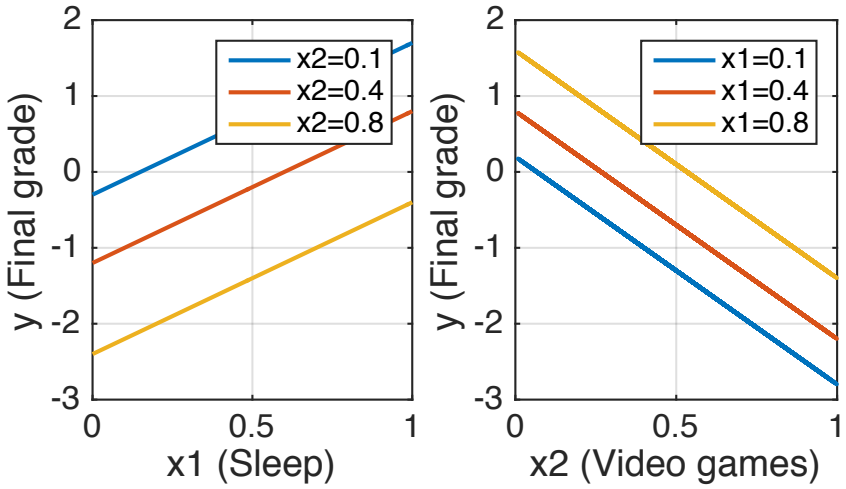
Multiple regression

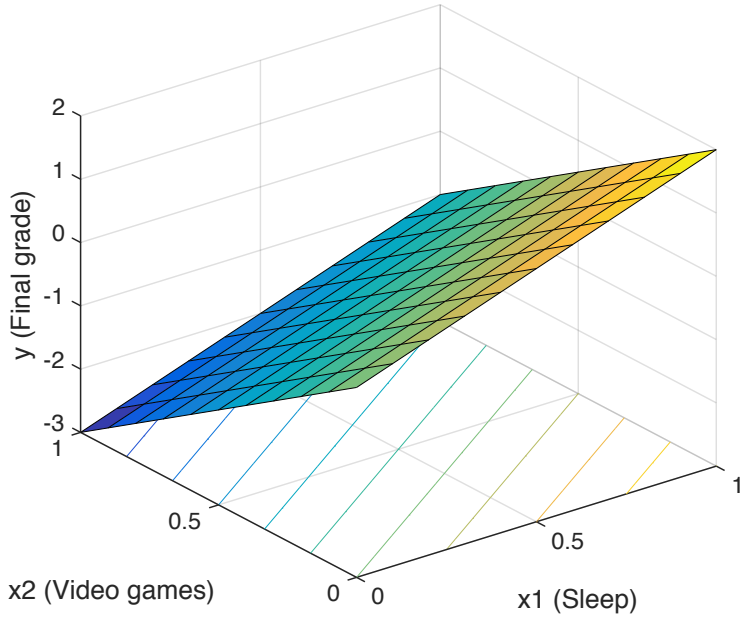
- Your grade in a class is influenced by multiple factors: Study time, Prior knowledge, How much sleep you're getting…
- You could fit a separate regression model for each factor
- However, we would like to understand how the outcome (grade) depends on *all* the factors. We'd like one unified model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Grade **Sleep time** **Study time** **Prior knowledge**

- A multiple linear regression corresponds to a flat plane in a p-dimensional space
- If you choose a fixed value for X2, the effect of X1 is linear
- If you choose a fixed X1, X2 has a linear effect





Simple vs. Multiple regressions

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	< 0.0001

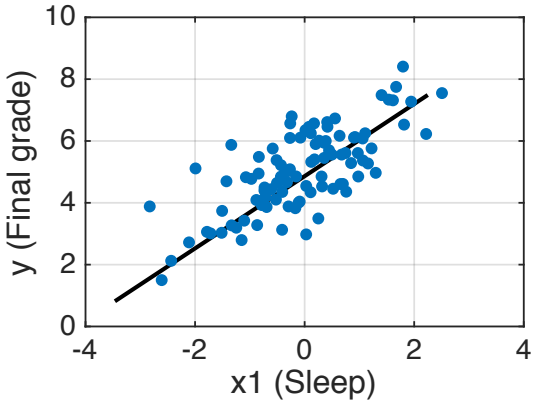
When each factor is considered separately, newspaper and radio advertising both seem to have a positive effect on sales

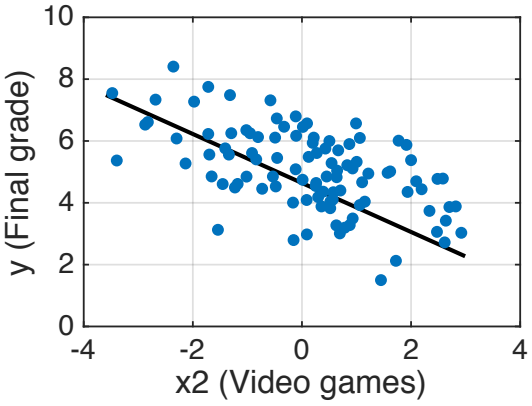
Multiple regression with TV, radio and newspaper as factors

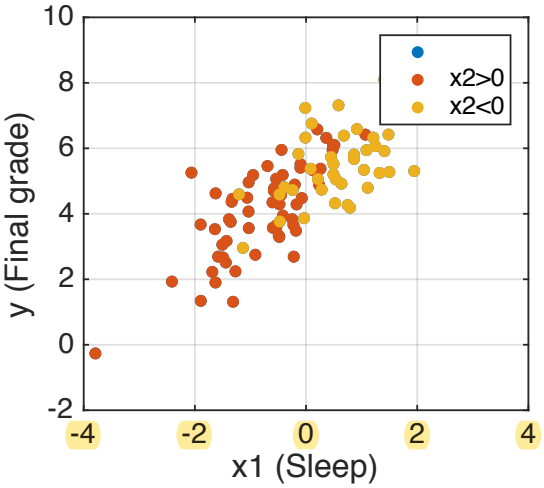
	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

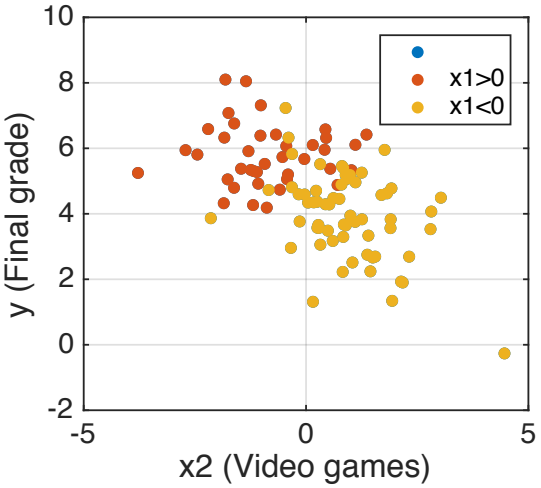
After controlling for the effects of radio and TV, newspaper advertising has no significant effect

Why not fit separate models?









- Both sleep and video games appear significantly correlated with final grade
- However, sleep and video games are (negatively) correlated with each other.
- If we hold sleep fixed, there is no additional correlation with video games
- If we hold video games fixed, there is still an additional correlation with sleep

What questions can we ask using a multiple regression?

1. Is there a significant effect of *any of the predictors*?

2. Which of the predictors has a significant effect?

    - More precisely: Which predictors have a significant effect *if we hold all the other variables fixed*?

3. How can we use the model to predict y, and how accurate will the prediction be?

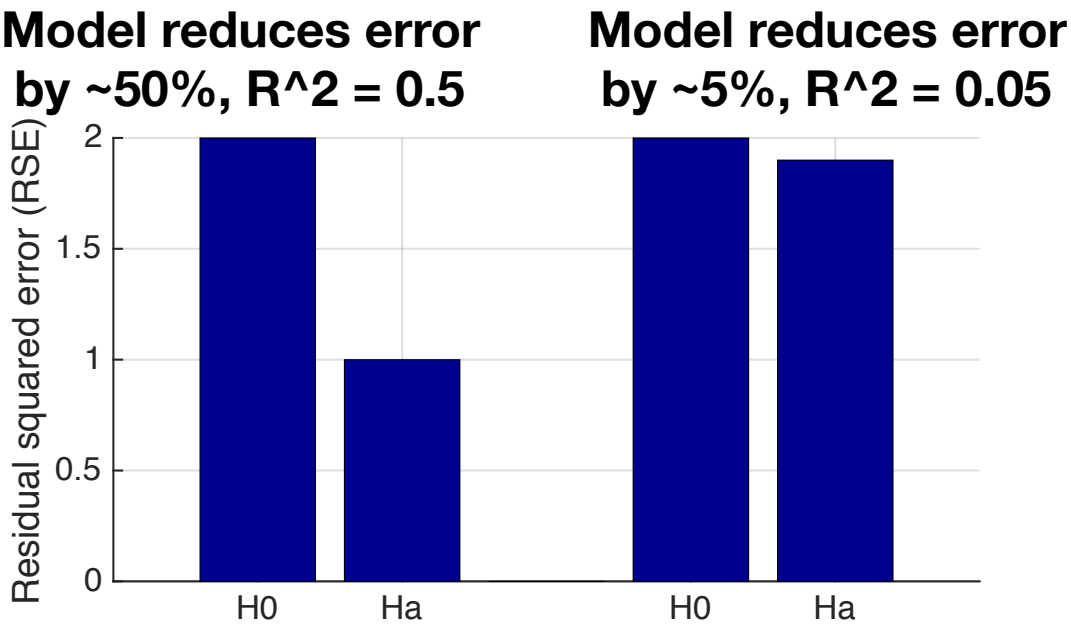
Question #1: Is there a significant effect of *any* predictor?

- We always start by testing the simplest (most uninteresting!) null hypothesis

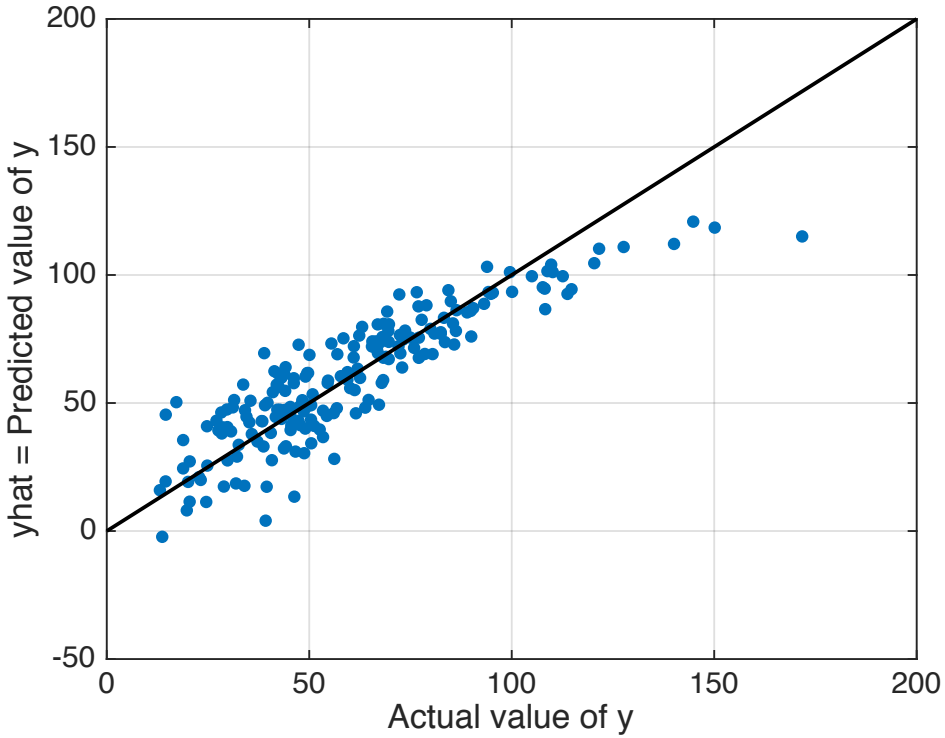
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

- If we can't rule this out, there's not much point going further.

- To test this, we look at R^2 = the proportion of variance explained by the model



R^2 is the squared correlation between y and yhat

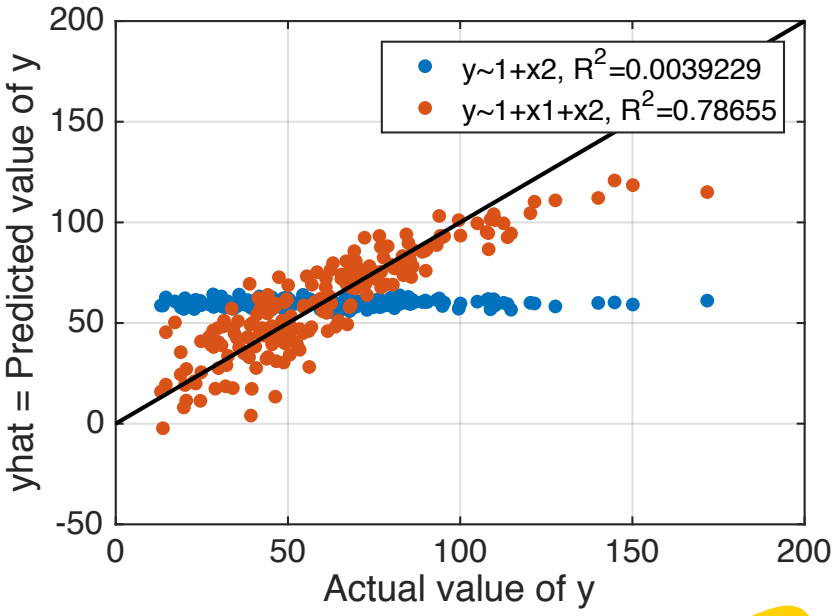


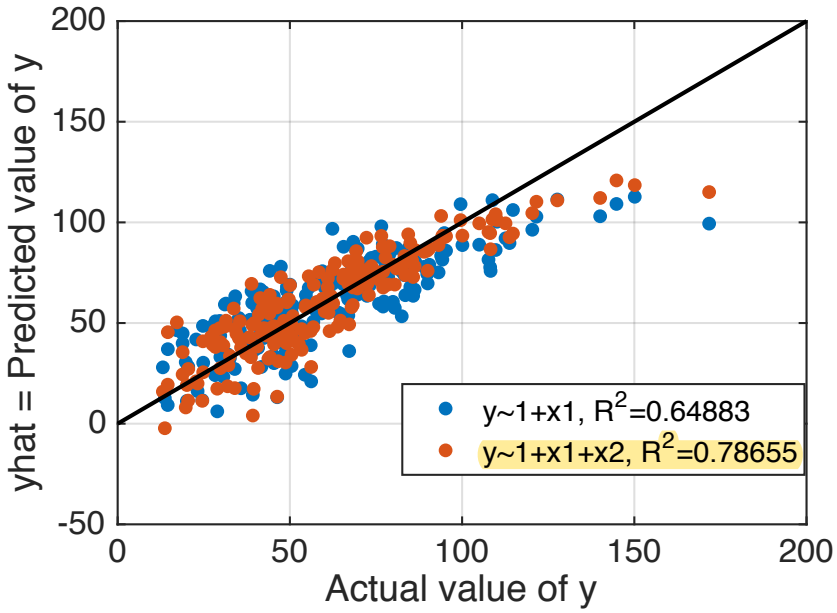
R^2 = corr(y,yhat)^2

R^2 = 1 - var(y-yhat)/var(y)

```
Number of observations: 200, Error degrees of freedom: 197
Root Mean Squared Error: 13
R-squared: 0.787,  Adjusted R-Squared 0.784
F-statistic vs. constant model: 363, p-value = 8.61e-67
```

Question #2: Which predictors are significant?





Linear regression model:
y ~ 1 + x1 + x2

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	60.029	0.92651	64.79	8.6976e-135
x1	-29.859	1.111	-26.876	8.114e-68
x2	10.808	0.95863	11.274	4.5253e-23

Regression with categorical predictors

- The linear regression framework can be applied using either quantitative or categorical predictors, or a combination of both
- To do this, we have to represent categorical predictors by numbers called dummy variables
- Why dummy? Because the numbers themselves don't mean anything
    - Example 1: Chocolate = 1, Vanilla = 0
    - Example 2: Chocolate = 0, Vanilla = 1
    - Example 3: Chocolate = -1, Vanilla = 1
- In all three cases, the math may look different but the results are equivalent

Categorical variable with 2 values (levels)

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

TABLE 3.7. *Least squares coefficient estimates associated with the regression of* `balance` *onto* `gender` *in the* `Credit` *data set. The linear model is given in (3.27). That is, gender is encoded as a dummy variable, as in (3.26).*

Alternative parameterization of the same data

First parameterization

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Second parameterization

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

- The values of beta0, beta1 will be different
- However, the predictions (yhat) and statistics (p-value, t-value) will be the same

Categorical predictors with >2 values (levels)

- Need a separate dummy variable for each new level.
- For a factor with p levels, we would have p-1 dummy variables (plus the intercept)
- Example: Race = African-American/Caucasian/Asian

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases} \qquad x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

Baseline

Extra effect of being Asian

Extra effect of being Caucasian

Regression results will have separate coefficients, p-values for each dummy variable

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260