

LECTURE 03:

LINEAR REGRESSION PT. 1

September 18, 2017

SDS 293: Machine Learning

Announcements

Need help with



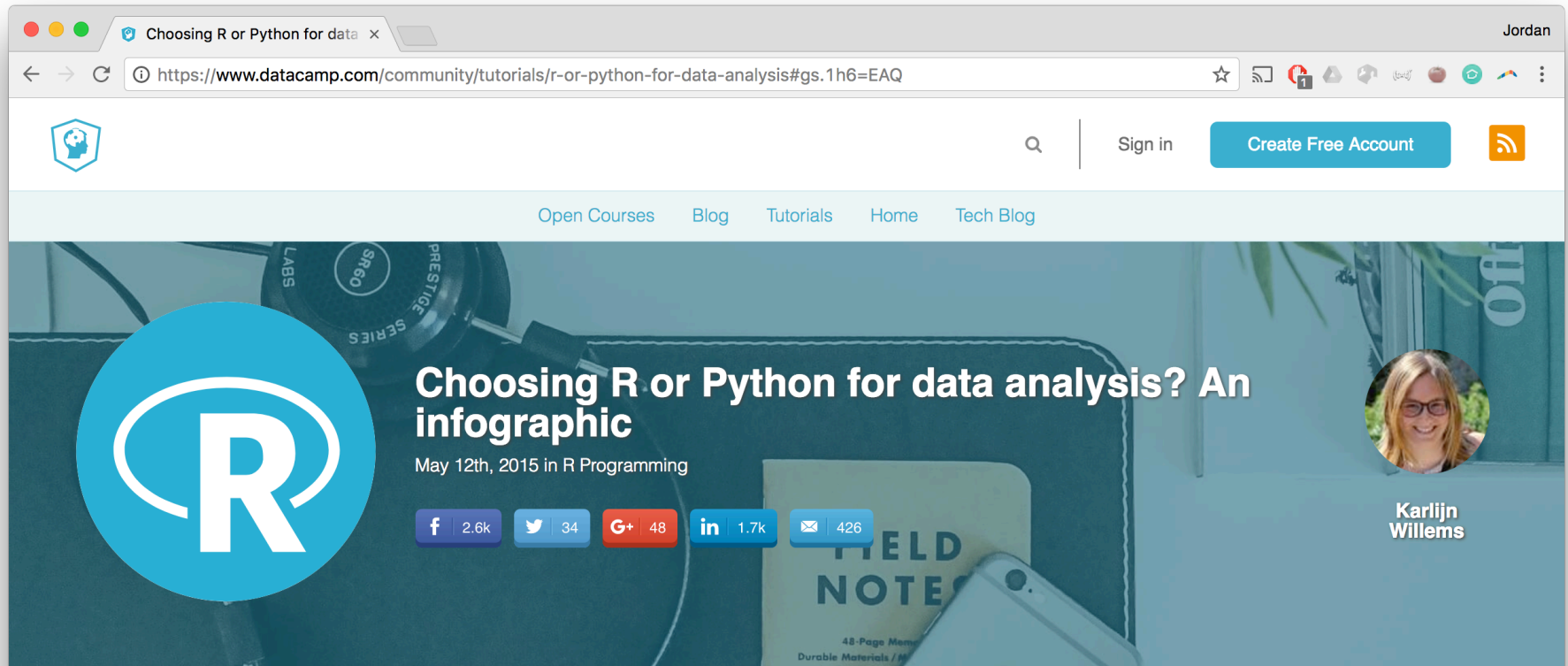
Visit the Stats TAs!

Sunday – Thursday evenings

7 – 9 pm in Burton 301

(SDS293 alum available every night 😊)

Question: R or python?



The screenshot shows a web browser window with the URL <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis#gs.1h6=EAQ>. The page features a navigation bar with links for 'Open Courses', 'Blog', 'Tutorials', 'Home', and 'Tech Blog'. The main content area has a large blue circle with the R logo on the left. The article title is 'Choosing R or Python for data analysis? An infographic', dated 'May 12th, 2015 in R Programming'. Social sharing buttons are visible for Facebook (2.6k), Twitter (34), Google+ (48), LinkedIn (1.7k), and Email (426). The author's name, 'Karlijn Willems', is displayed next to her profile picture.

I think you'll agree with me if I say: It's HARD to know whether to use Python or R for data analysis. And this is especially true if you're a newbie data analyst looking for the right language to start with. It turns out that there are many good resources that can help you to figure out the strengths and weaknesses of both languages. They often go into great detail, and provide a tailored answer to questions such as "What should I use for Machine Learning?", or "I need a fast solution, should I go for Python or R?". In today's post, I present to you our new Infographic **"Data Science Wars: R vs Python"**, that highlights in great detail the differences

Question: R or python?

Purpose

R focuses on better, user friendly data analysis, statistics and graphical models.

Python emphasizes productivity and code readability.

Used By?

R has been used primarily in academics and research. However, R is rapidly expanding into the enterprise market.

Python is used by programmers that want to delve into data analysis or apply statistical techniques, and by developers that turn to data science.

"The closer you are to statistics, research and data science, the more you might prefer R."

"The closer you are to working in an engineering environment, the more you might prefer Python."

“Supportive, statistically strong, but slow”



“Optimized, operational, but sometimes opaque”

And the winner is... it depends.

1 What problems do you want to solve?

2 What are the net costs for learning a language?*

* it will cost time to learn a new system that is better aligned for the problem you want to solve, but staying with the system you know may not be made for that kind of problem.

3 What are the commonly used tool(s) in your field?

4 What are the other available tools in your field and how do these relate to the commonly used tool(s)?

Outline

- Motivation
- Running Example: **Advertising**
- Simple Linear Regression
 - Estimating coefficients
 - How good is this estimate?
 - How good is the model?
- Multiple Linear Regression
 - Estimating coefficients
 - Important questions
- Dealing with Qualitative Predictors
- Extending the Linear Model
 - Removing the additive assumption
 - Non-linear relationships
- Potential Problems

Motivation

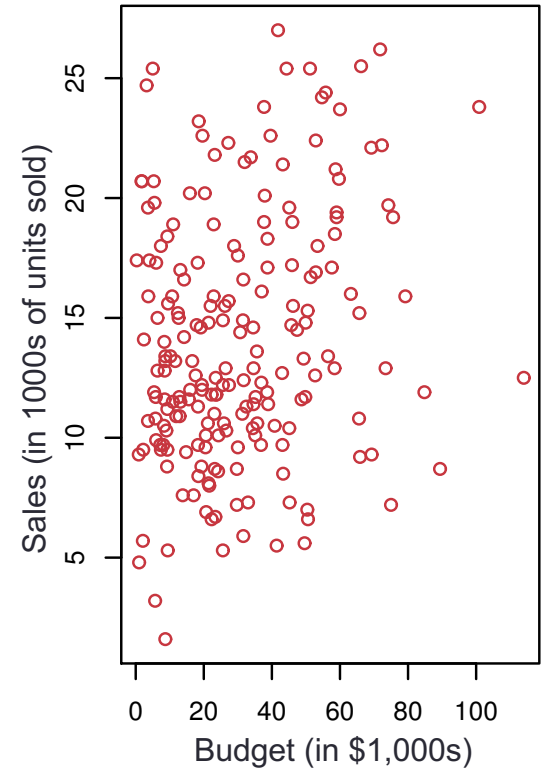
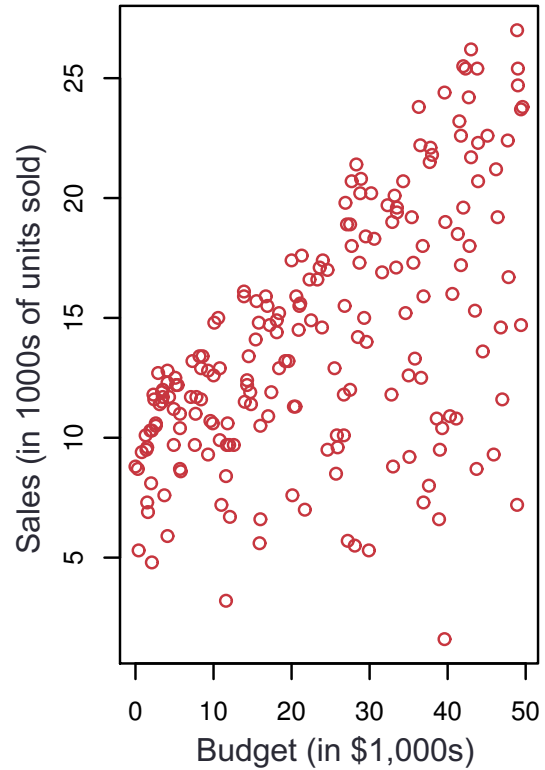
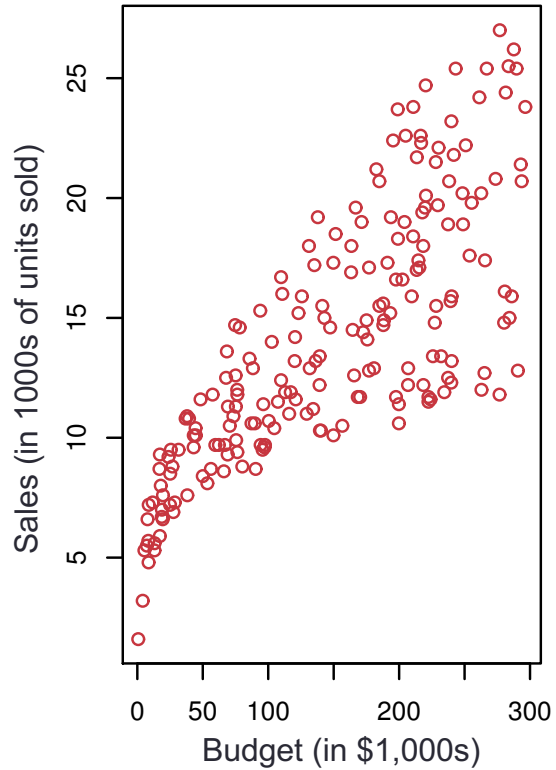
Why start a ML course with linear regression?



Running example: advertising



Last year's advertising budget



Your task



Questions you might ask

1. Is there a **relationship** between budget and sales?
2. How **strong** is the relationship?
3. Which **media** contribute to sales?
4. How accurately can we **estimate the effect**?
5. How accurately can we **predict future sales**?
6. Is the relationship **linear**?
7. Is there **synergy** among the advertising media?



Linear Regression

Simple linear regression

- Straightforward approach for predicting a **quantitative response** on the basis of a **single predictor**
- **Assumption:** there is a (roughly) linear relationship between X (the predictor) and Y (the response)

the response $\rightarrow Y \approx \underbrace{\beta_0 + \beta_1 X}_{\text{a linear function of the predictor}}$

is approximately modeled as

“intercept” \downarrow β_0 “slope” \downarrow $\beta_1 X$

$$\mathbf{sales} \approx \beta_0 + \beta_1 \times \mathbf{TV}$$

Simple linear regression

- **Reality:** β_0 and β_1 are unknown

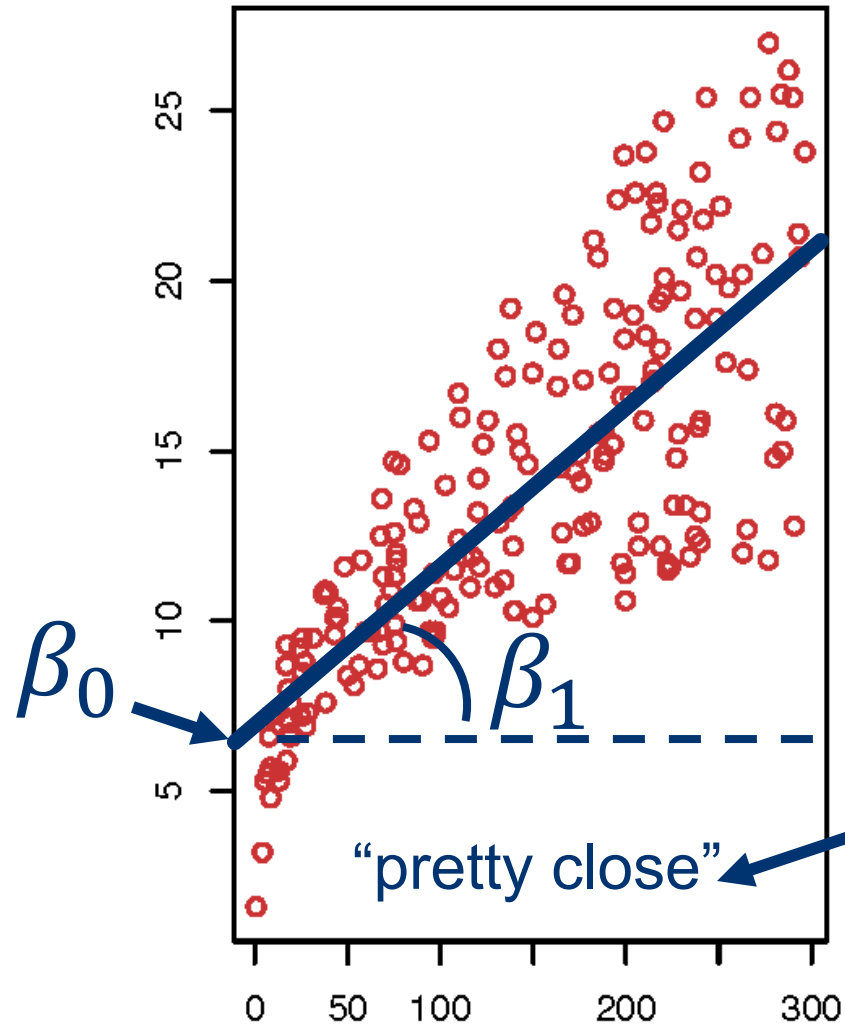
- What we **do** know:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

- **Goal:** find *estimated* coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Simple linear regression



minimizes
RSS
(other ways in Ch. 6)

Def. *residuals* and *RSS*

- Back to our hypothetical model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- Def. *residual*: $\epsilon_i = y_i - \hat{y}_i$

(difference between *observed* and *predicted* responses)

- Def. *residual sum of squares (RSS)*:

$$RSS = \epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_n^2$$

$$RSS = (y_1 - \hat{\beta}_0 + \hat{\beta}_1 x_1)^2 + \dots + (y_n - \hat{\beta}_0 + \hat{\beta}_1 x_n)^2$$

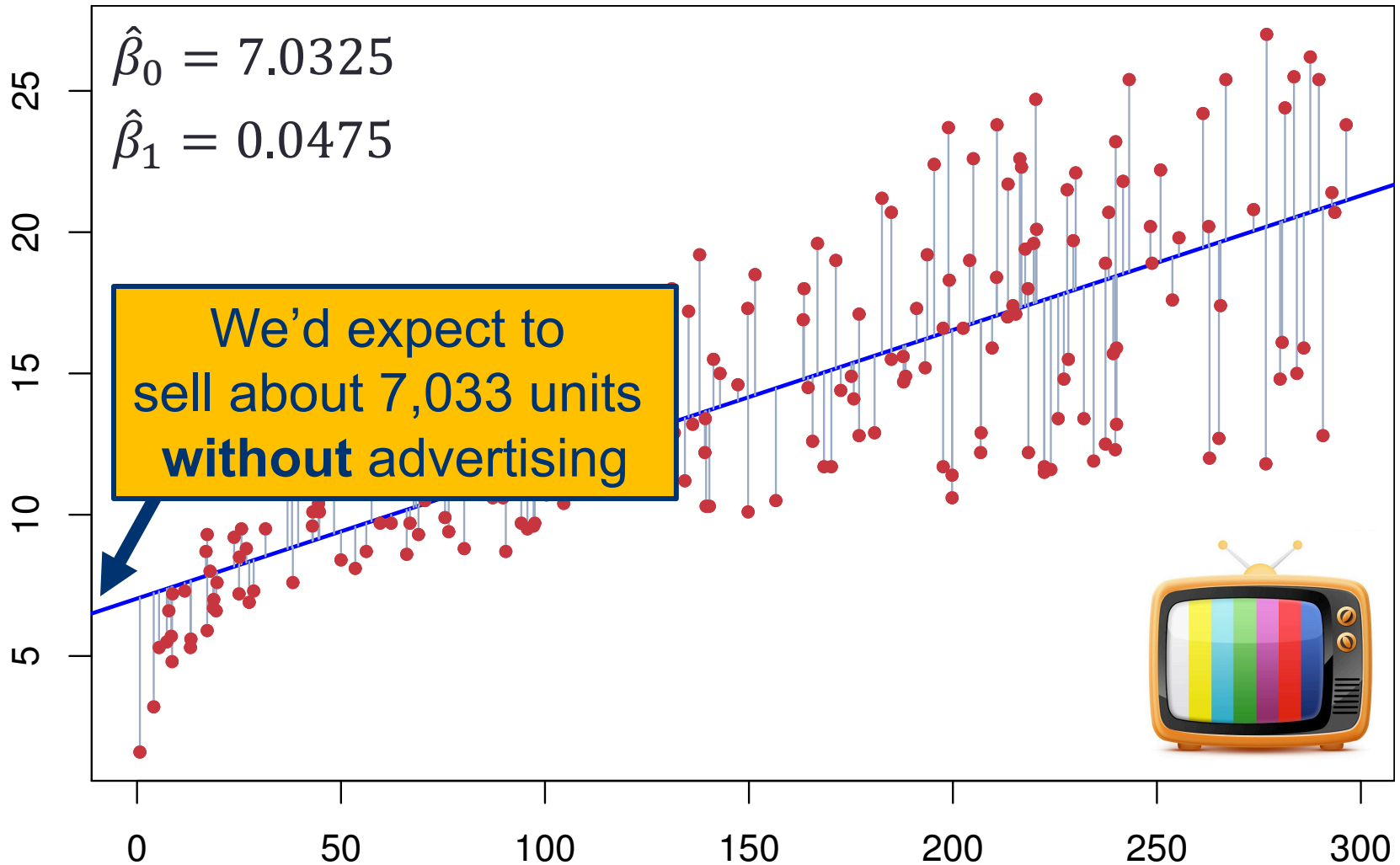
Minimizing RSS: least squares

- **Goal:** $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize RSS
- Dusting off our calculus (or looking it up), minimizers are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where \bar{x} and \bar{y} are the mean values of the sample

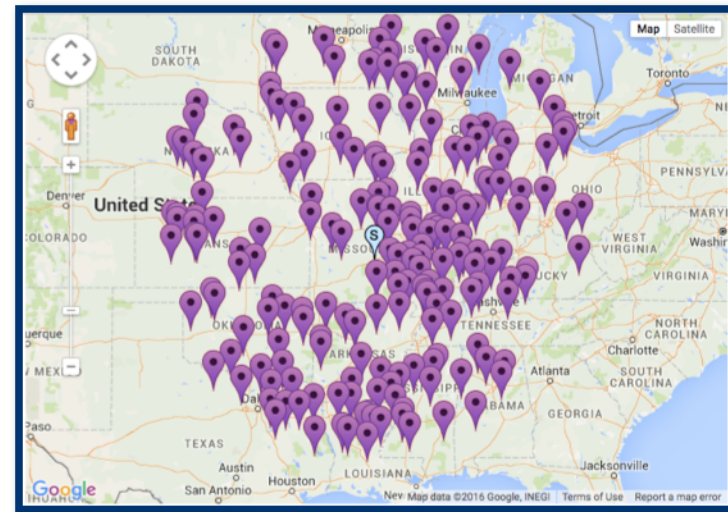
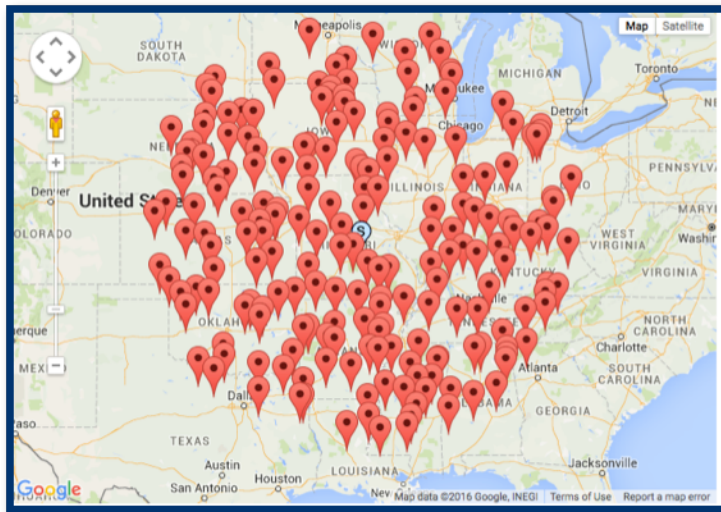
Advertising example



➔ Every additional \$1,000 \approx 47.5 additional units sold

How good is this estimate?

- **Assumption:** $Y \approx \beta_0 + \beta_1 X$
- We **estimated** $\hat{\beta}_0$ and $\hat{\beta}_1$ from the available data
- Consider this:



Standard error

- **Idea:** borrow the concept of standard error (SE):

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

- σ is the **standard deviation** of the population
- n is the **number of samples**
- **Note:** the error gets smaller as the sample size increases

Standard error of $\hat{\beta}_1, \hat{\beta}_0$

- **Idea:** use the **standard deviation of ϵ** for σ (why?)

- Start with the slope:

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

What happens as x spreads out?

- And now the intercept:

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

What happens when the mean of x is 0?

Just one problem...

- **Idea:** use the **standard deviation of ϵ** for σ



usually don't have this information

What **do** we know about ϵ ?



Residual standard error

- **Idea:** estimate standard deviation of ϵ using RSS to get *residual standard error*:

$$RSE = \sqrt{\frac{RSS}{(n-2)}}$$

- Now we can finally estimate SE, which can be used to compute *confidence intervals*
- In linear regression, the 95% confidence intervals are:

$$\hat{\beta}_0 \pm 2 \times SE(\hat{\beta}_0) \text{ and } \hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$$

Using SE for hypothesis testing

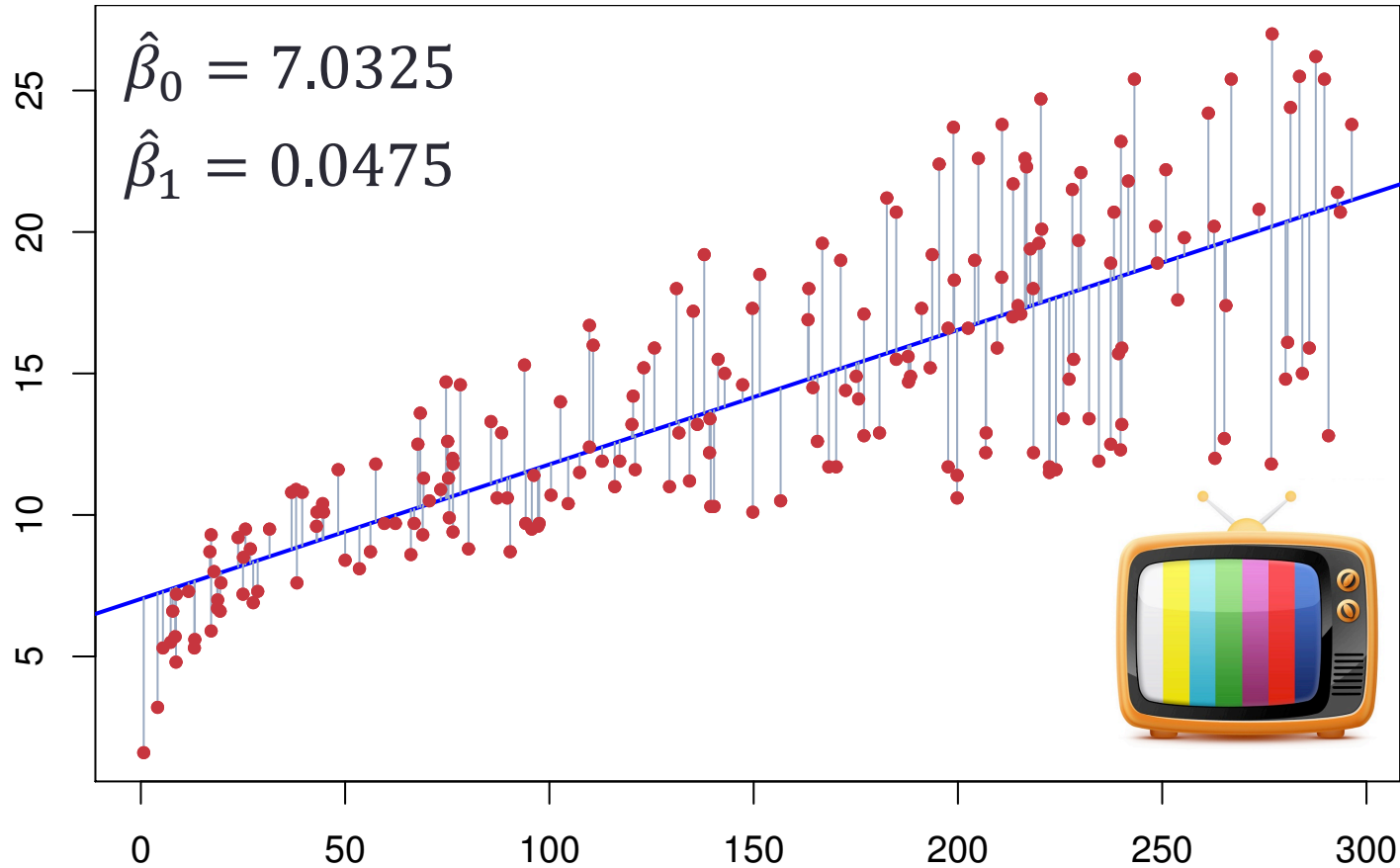
- **Goal:** determine if sales are related to advertising budget
- If there is **NO relationship**, what is the true value of β_1 ?

no relationship = no slope

$$\beta_1 = 0$$

- **To test:** compute the probability that we observed our (estimated) β_1 by chance, assuming a true value of 0
- If this probability is **small**, we say a relationship exists

Advertising example



	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

How good is this model?

- RSE is (roughly) the amount the response will deviate from the *true* regression line
- RSE is an **absolute** measure, given in the same units as the response variable
- **Question:** how do you know what a “good” RSE is?




How good is this model?

- **Alternate approach:** measure the *proportion* of variance explained by the model
- R^2 is one such measure:

$$R^2 = 1 - \frac{RSS}{\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{total variance in the response}}}$$

variance
not explained
after regression



TV and sales

Quantity	Value
Residual standard error	3.26
R^2	0.612

- What does the RSE tell us?
- What does R^2 tell us?

Discussion

Question: how could we handle multiple predictors?



Option 1: SLR for each predictor

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	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

	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

What **problems** do you see with this approach?

Option 2: extend the linear model

- Give each variable its own slope, e.g.

$$\begin{aligned}\mathbf{sales} \approx \beta_0 + \beta_1 \times \mathbf{TV} + \\ \beta_2 \times \mathbf{radio} + \\ \beta_3 \times \mathbf{newspaper} + \epsilon\end{aligned}$$

- Each slope captures the average effect on Y of an increase in one predictor, *holding all others constant*
- Estimate coefficients using least squares (same as SLR!)

Advertising example

	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

- What does this tell us?
- Do you notice anything unexpected?



What happened to newspaper ads?

- Let's look at the correlation between all the dimensions

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

In SLR, **newspaper** spending was “getting credit”
for **radio** spending's work!

Questions we ask in MLR

- Is **at least one** of the predictors useful in predicting the response?
- Do all the predictors help to explain the response, or is only a **subset** of the predictors useful?
- How well does the model **fit** the data?
- Given some predictor values, what response should we predict, and how **accurate** is our prediction?

Q1: is at least one predictor useful?

- SLR: test to see if the slope was 0 (no effect)
- MLR: test whether ALL of the slopes are 0 (no effect)
- To do this, we compute the F-statistic:

$$F = \frac{(TSS - RSS)}{p} \times \frac{(n - p - 1)}{RSS}$$

where p is the # of predictors and n is the sample size

- Value close to 1 \rightarrow no effect
- **Question:** why look at the F-statistic and not just at the p-values for each predictor in turn? (*hint: lots of predictors?*)

Q2: do we need them **all**?

- Now we know that at least one predictor has an effect: which one(s) is it?
- Determining which predictors are associated with the response is referred to as *variable selection*
- Some classic approaches:
 - Exhaustive search
 - Forward selection
 - Backward selection
 - Mixed selection
- More detail in Ch. 6

Q3: How well does the model fit the data?

- Just like in SLR, we can use RSE and R^2 to measure how well our model fits the data
- Using the MLR model we created using all 3 predictors:

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

- **Question:** what would happen to the R^2 value if we remove **newspaper** from the model?

Q4: How confident are we?

- Now that we have a model, making a prediction is a piece of cake (just plug and chug!)
- Need to consider 3 kinds of uncertainty:
 1. How far off are the coefficients? → confidence intervals
 2. How far from linear is the true relationship? → ignore this for now
 3. How much will any *specific* prediction vary from the true value, even if we had perfect coefficients? → prediction intervals

Quick activity: the Carseats data set



Quick activity: the `Carseats` data set

- **Description:** simulated data set on sales of car seats
- **Format:** 400 observations on the following 11 variables
 - `Sales`: unit sales at each location
 - `CompPrice`: price charged by nearest competitor at each location
 - `Income`: community income level
 - `Advertising`: local advertising budget for company at each location
 - `Population`: population size in region (in thousands)
 - `Price`: price charged for car seat at each site
 - `ShelveLoc`: quality of shelving location at site (Good | Bad | Medium)
 - `Age`: average age of the local population
 - `Education`: education level at each location
 - `Urban`: whether the store is in an urban or rural location
 - `USA`: whether the store is in the US or not

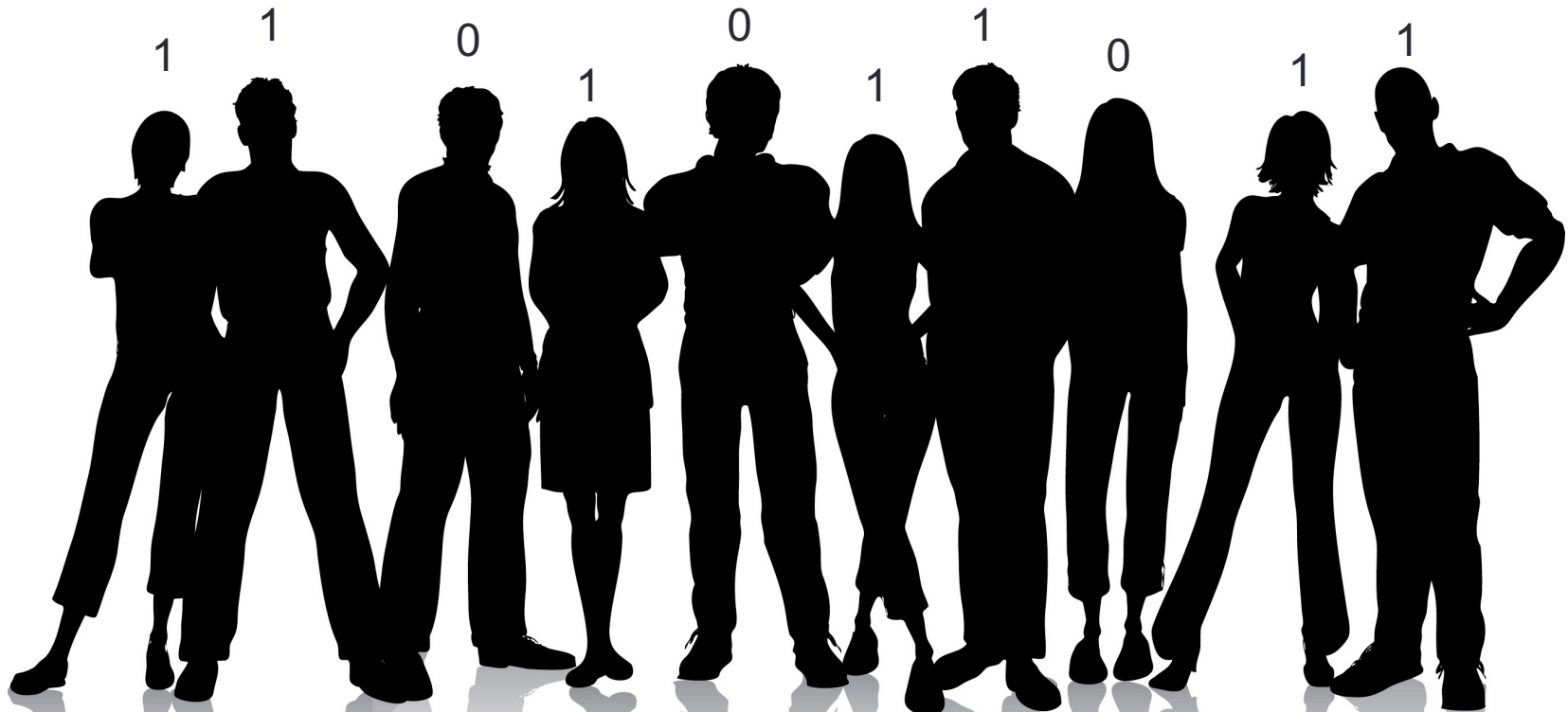
Quick activity: the Carseats data set

1. Find some friends (groups of 3-4 are ideal)
2. Hypothesize 5 possible relationships between variables in this dataset (e.g. **Price** predicts **Sales**)

Question: could you test that hypothesis with the techniques you know right now?

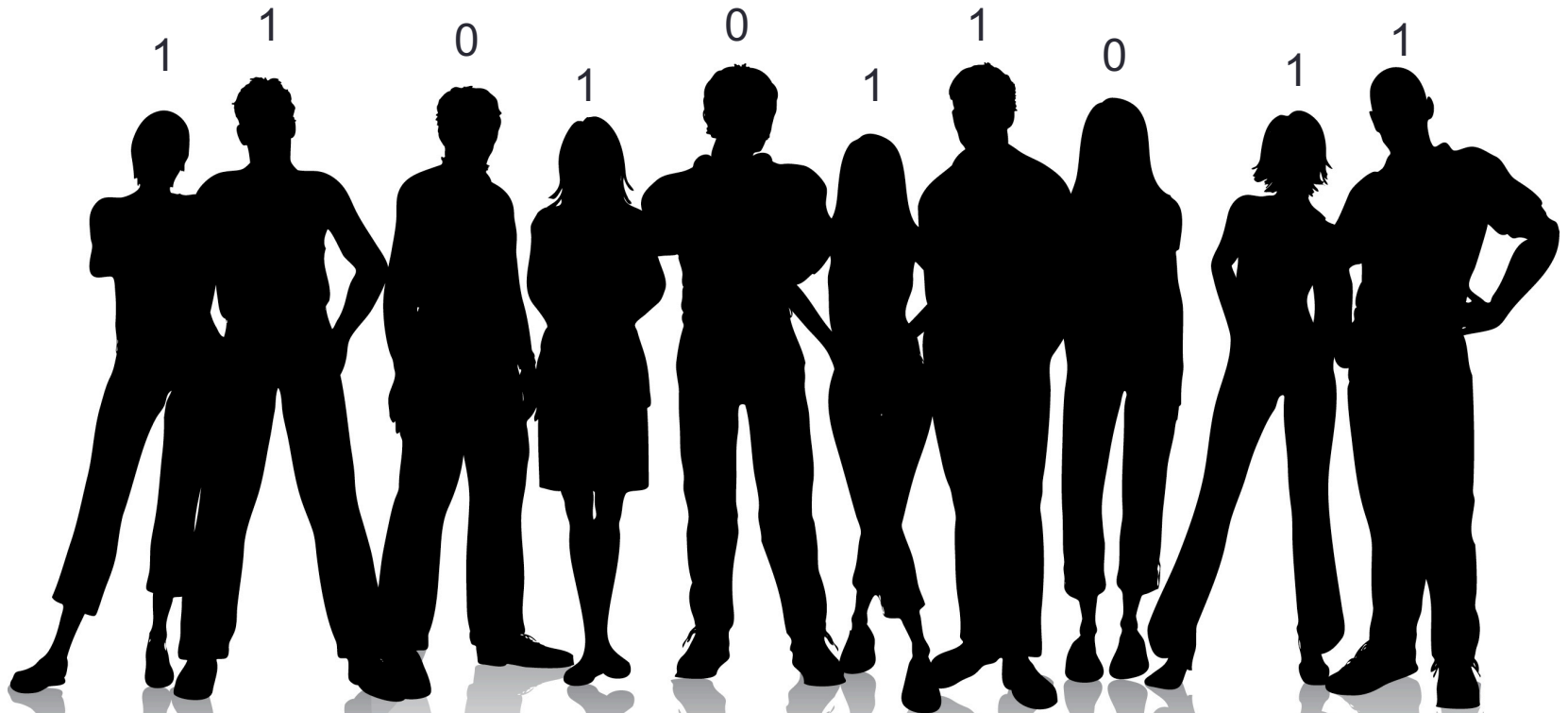


Two-level qualitative predictors



({P1:"enrolled"}, {P2:"enrolled"}, {P3:"auditing"},...)
({P1:1}, {P2:1}, {P3:0},...)

Two-level qualitative predictors



$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if enrolled} \\ \beta_0 + \epsilon_i & \text{if auditing} \end{cases}$$

A note on dummy variables

- The decision to code enrolled students as 1 and auditing students as 0 is arbitrary
- It has no effect on model fit, or on the predicted values
- It **does** alter interpretation of the coefficients
 - If we swapped them, what would happen?
 - If we used $(-1, 1)$, what would happen?

Multi-level predictors

- Need dummy variables for *all but one level*
- For example:

$$x_{i1} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ person is from Amherst} \\ 0 & \text{if the } i^{\text{th}} \text{ person is not from Amherst} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ person is from Mt. Holyoke} \\ 0 & \text{if the } i^{\text{th}} \text{ person is not from Mt. Holyoke} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i^{\text{th}} \text{ person is from Amherst} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i^{\text{th}} \text{ person is from Mt. Holyoke} \\ \beta_0 + \epsilon_i & \text{if } i^{\text{th}} \text{ person is from Smith} \end{cases}$$



Introduction to python

- Today's walkthrough was run using Jupyter:



- This allows me to build “notebooks” to combine step-by-step code and instructions/descriptions
- Want to learn more? Check out the “**Jupyter Notebook Tutorial: The Definitive Guide**” on DataCamp!

Coming up

- Next class: finish linear regression
- Assignment 1 comes out Wednesday
 - Due **Wednesday Sept. 27 by 11:59pm**