Introduction to Machine Learning – Linear Regression

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Reminder!

- First homework is out today!
- Homework and quizzes are *week long* assignments; expect to spend 5-7 hours on then (this is standard for a college class)
- Chapter 2.3 in the textbook has refreshers for R and Python

Plan for Today

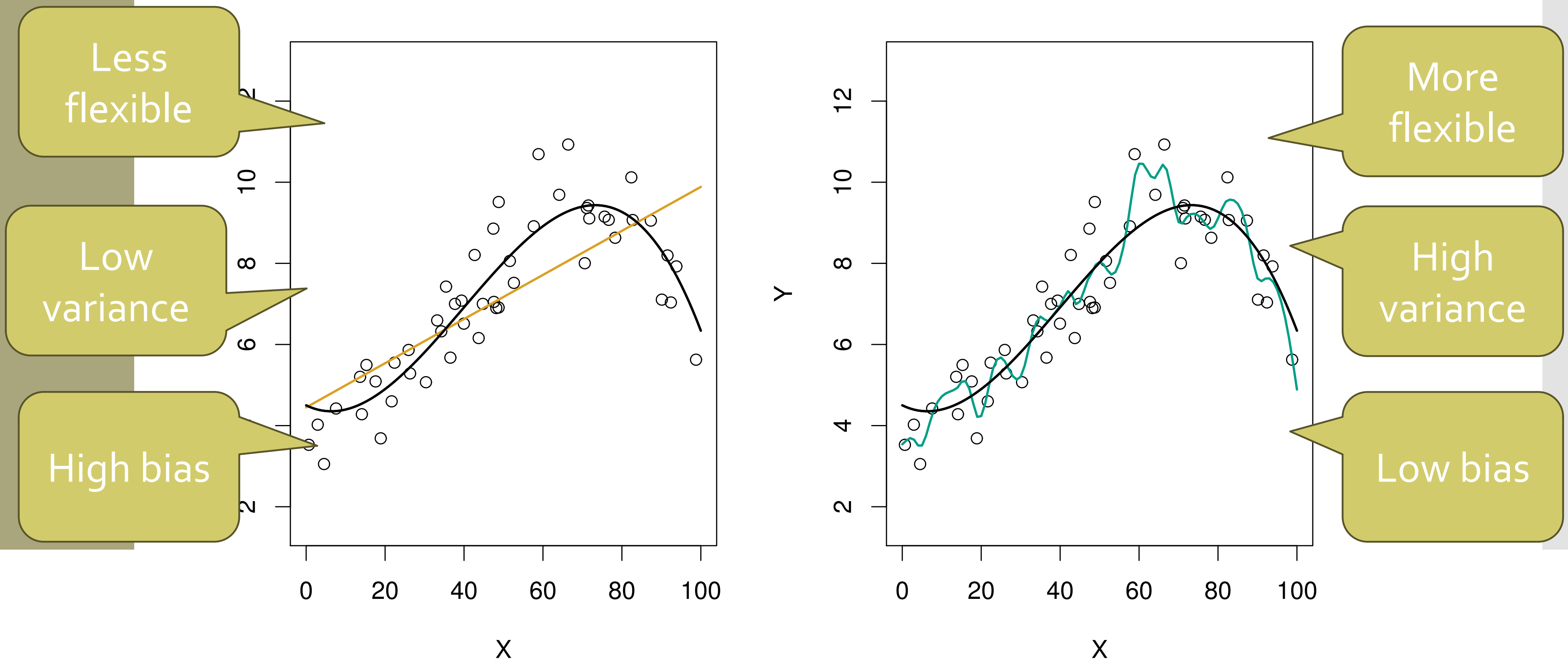
- Introduction to Linear Regression:
  - Simple linear regression
  - Multiple linear regression

Variance: the amount the model would change if we had different training data

Bias: the error introduced by approximating a complex phenomenon using a simple model

In general, more flexible methods have **higher variance** and **lower bias**

Warm Up



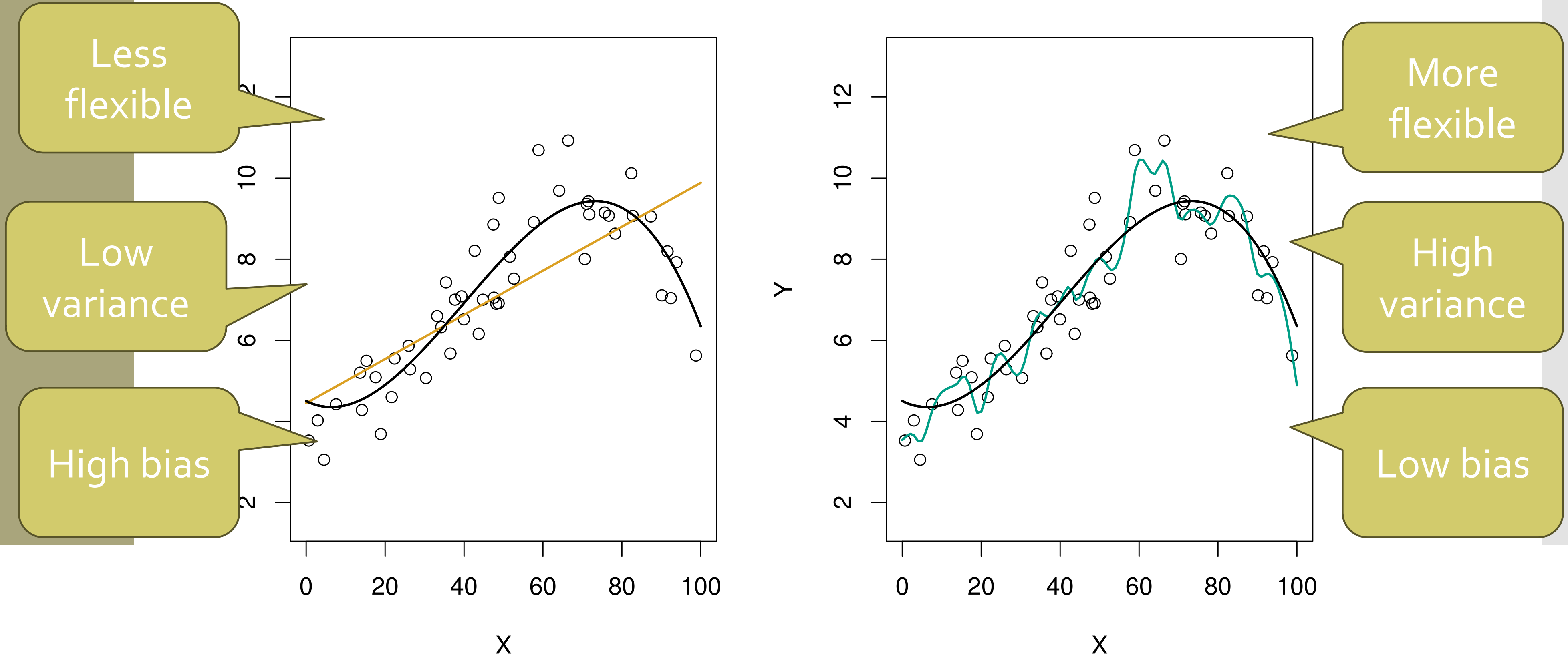
Practice: Work with a small group. Sketch your own example of a more and less flexible regression.

Warm Up

Variance: the amount the model would change if we had different training data

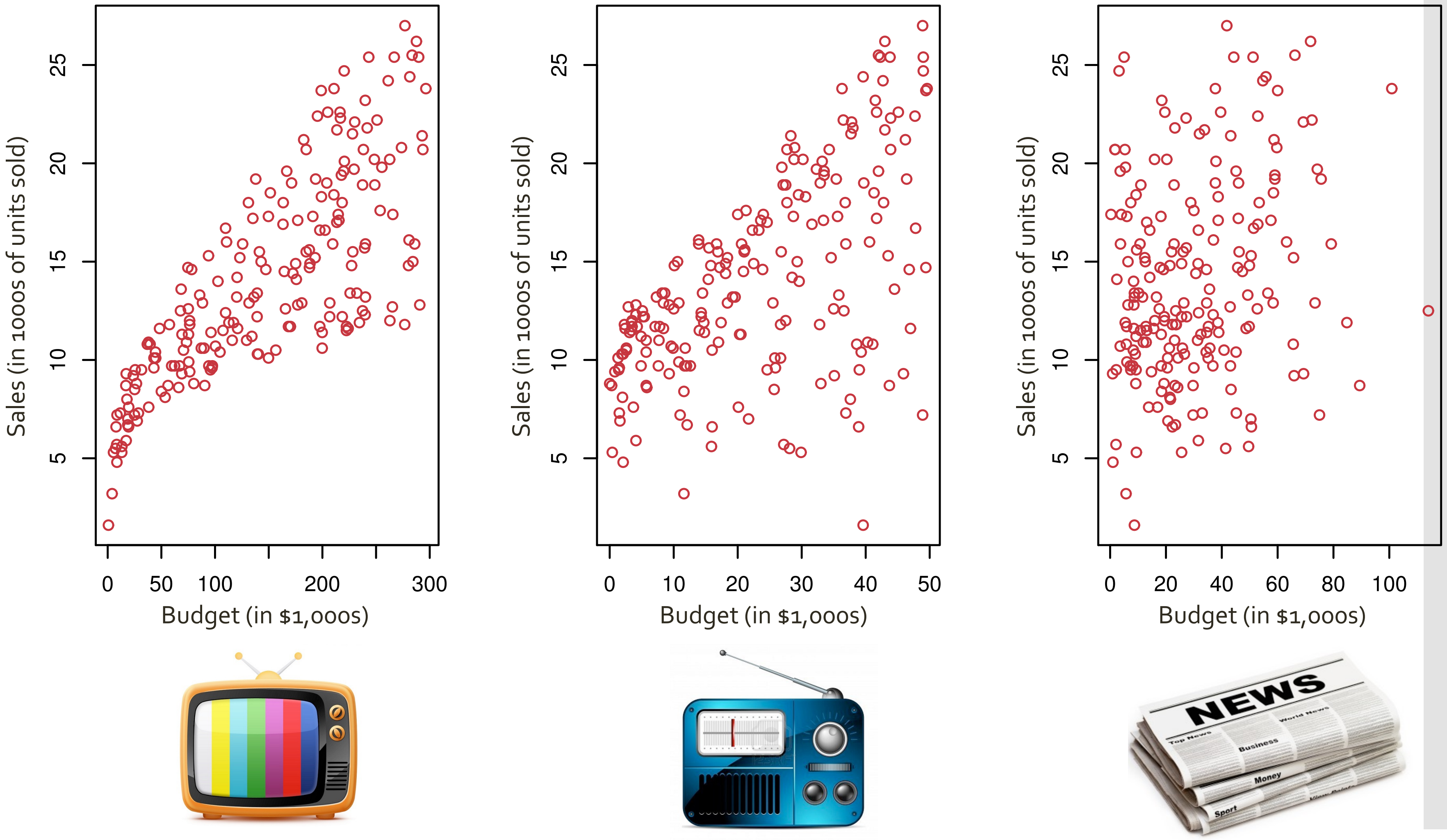
Bias: the error introduced by approximating a complex phenomenon using a simple model

In general, more flexible methods have **higher variance** and **lower bias**

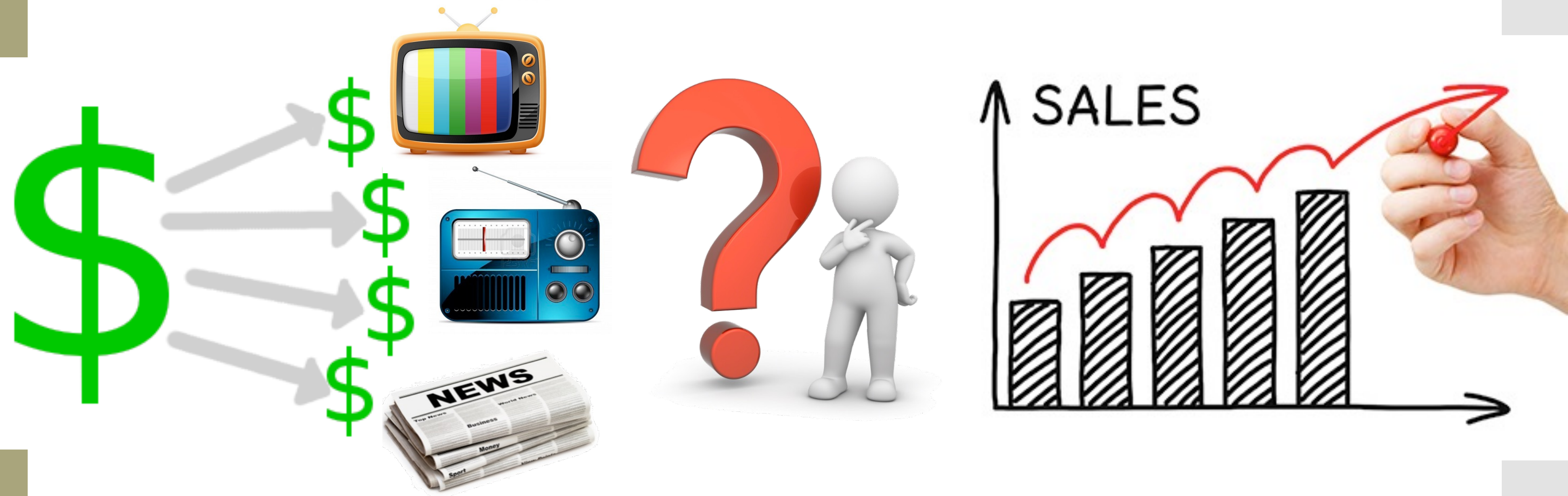


Running example: advertising

Last year's advertising budget



Your task



Questions you might ask

1. Is there a **relationship** between budget and sales?

Questions you might ask

1. Is there a **relationship** between budget and sales?
2. How **strong** is the relationship?

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?

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**?

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**?

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**?

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?

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)

$$\underset{\text{the response}}{Y} \underset{\text{is approximately modeled as}}{\approx} \underbrace{\overset{\text{“intercept”}}{\beta_0} + \overset{\text{“slope”}}{\beta_1} X}_{\text{a linear function of the predictor}}$$

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

Simple linear regression

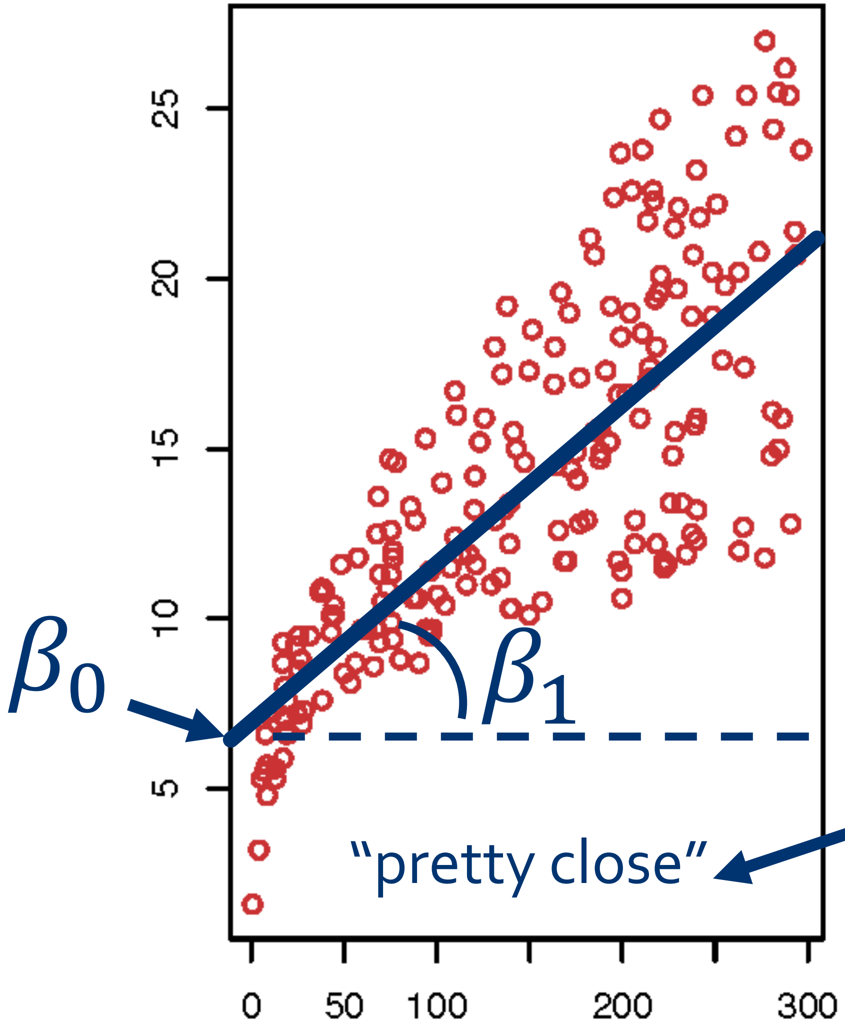
- **Reality**: β_0 and β_1 are unknown
- What we **do** know:

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

- **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)*:

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

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

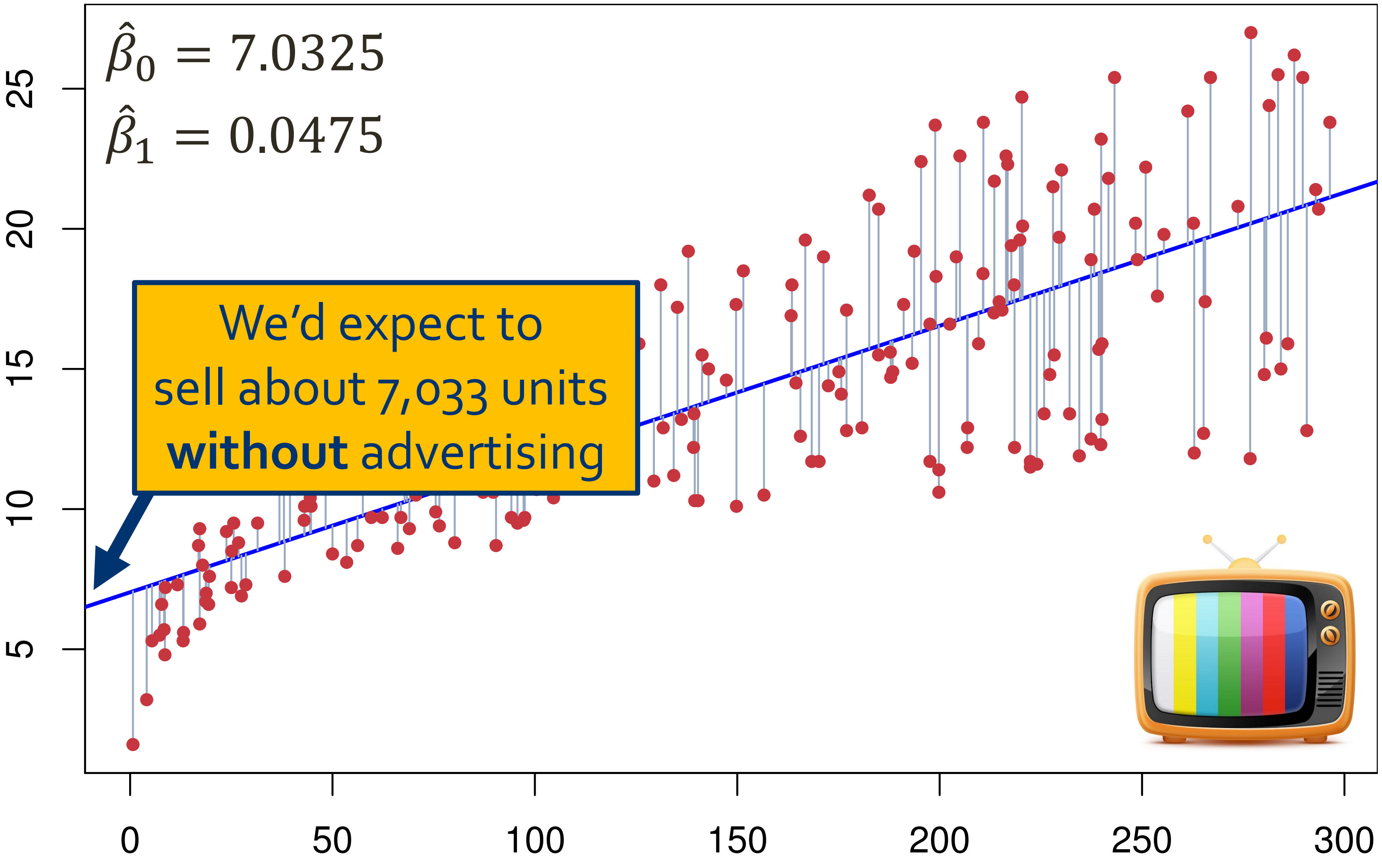
Minimizing RSS: least squares

- **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

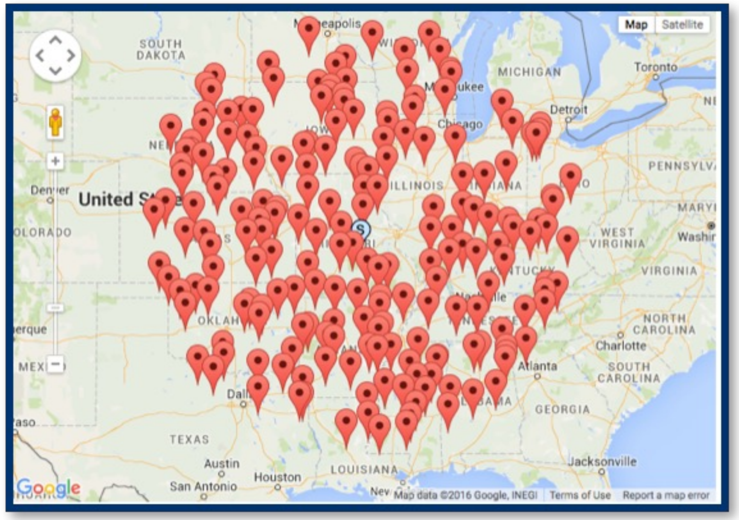
Advertising example

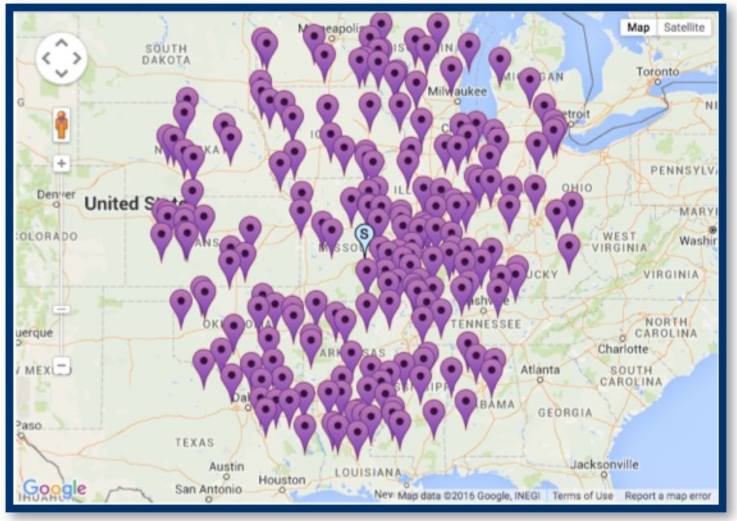


→ Every additional $1,000 ≈ 47.5 additional units sold

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 $\underbrace{\textbf{standard deviation of } \epsilon}_{\text{usually don't have this information}}$ for σ

What **do** we know about $\boldsymbol{\epsilon}$?

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:

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

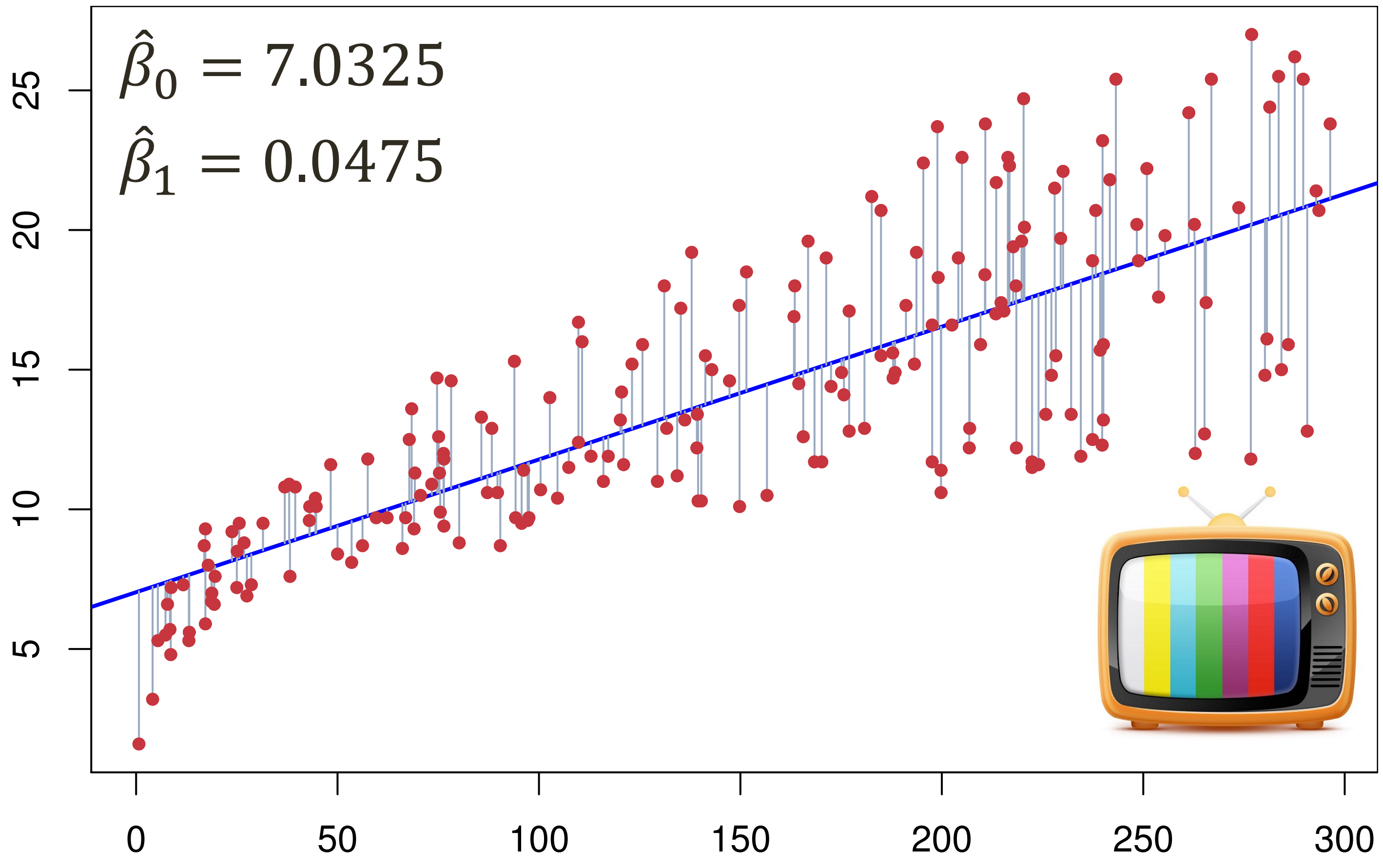
Using SE for hypothesis testing

- **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	$\boxed{< 0.0001}$

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:

variance not explained after regression

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

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.

$$\texttt{\textbf{sales}} \approx \beta_0 + \beta_1 \times \texttt{\textbf{TV}} + \beta_2 \times \texttt{\textbf{radio}} + \beta_3 \times \texttt{\textbf{newspaper}} + \epsilon$$

- 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 → 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