LECTURE 04:

LINEAR REGRESSION PT. 2

September 20, 2017

SDS 293: Machine Learning

Announcements

- Stats TA hours start **Monday** (sorry for the confusion)
- Looking for some refreshers on mathematical concepts?
  - The **Spinelli Center** has several coming up:
    - "Exponents & Logarithms" tonight (Sept. 20th)
    - "Trigonometry Review" on Thurs. Sept. 21st
    - …and several more!
  - Sessions run from **7-9pm in Seeyle 211**
- Evening office hours with Jordan:

Tuesdays 6-7pm

Ford 355

(will confirm on Slack each week)

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
- **3-minute activity**: Dealing with Qualitative Predictors
- Extending the Linear Model
  - Removing the additive assumption
  - Non-linear relationships
- Potential Problems

3-minute activity: the `Carseats` data set






3-minute activity: the `Carseats` data set

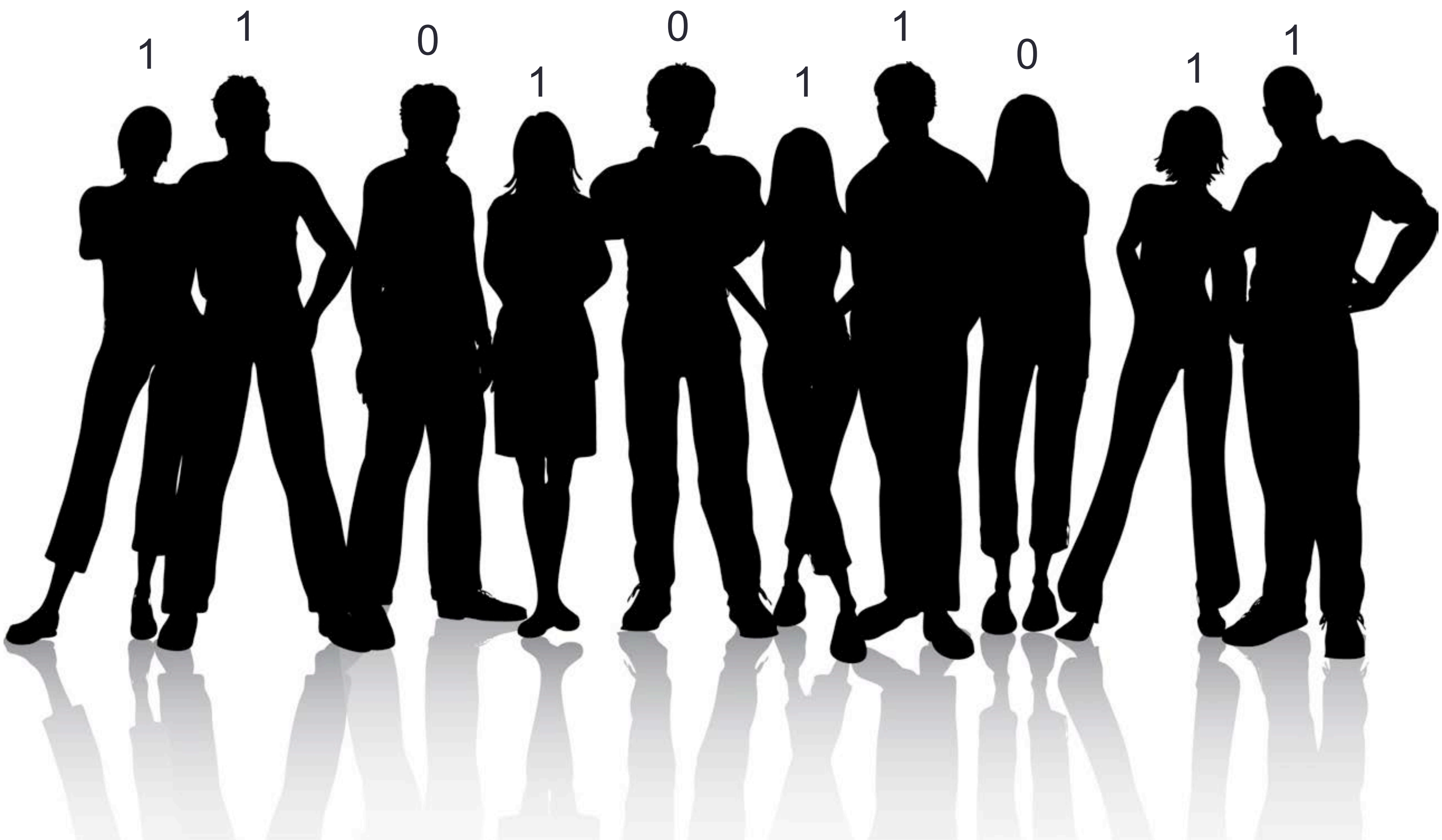
- **Description:** simulated data set on sales of car seats
- **Format:** 400 observations on the following 11 variables
  - **`Sales`**: unit sales at each location
  - **`CompPrice`**: price charged by nearest competitor at each location
  - **`Income`**: community income level
  - **`Advertising`**: 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

3-minute activity: the `Carseats` data set

1. Find a friend (or two)
2. Hypothesize 3 possible relationships between variables in this dataset (e.g. higher `Price` predicts lower `Sales`)

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

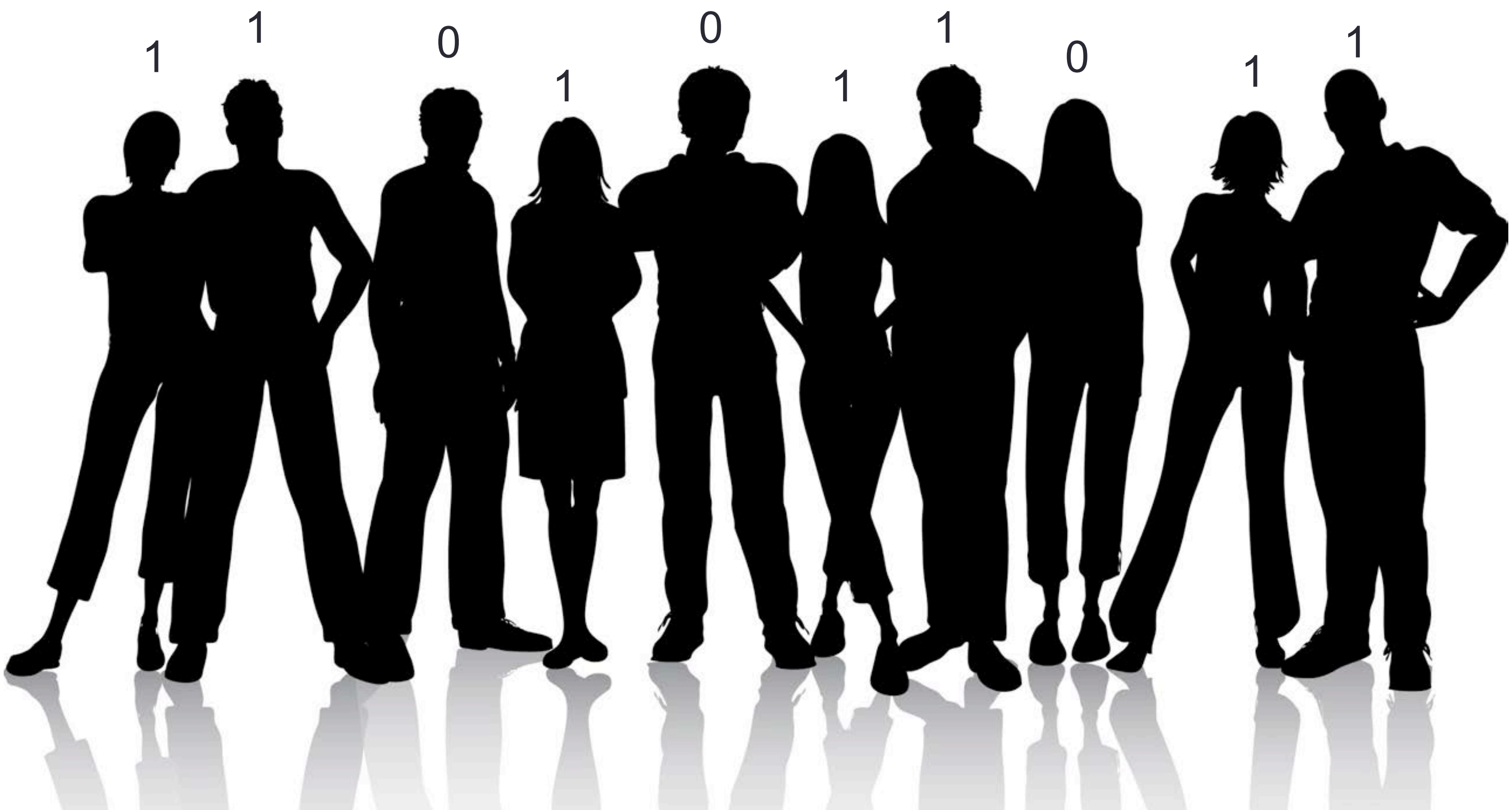
Two-level qualitative predictors



({P1:“enrolled”}, {P2:“enrolled”}, {P3:“auditing”},…)

({P1:1}, {P2:1}, {P3:0},…)

Two-level qualitative predictors



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

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:

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

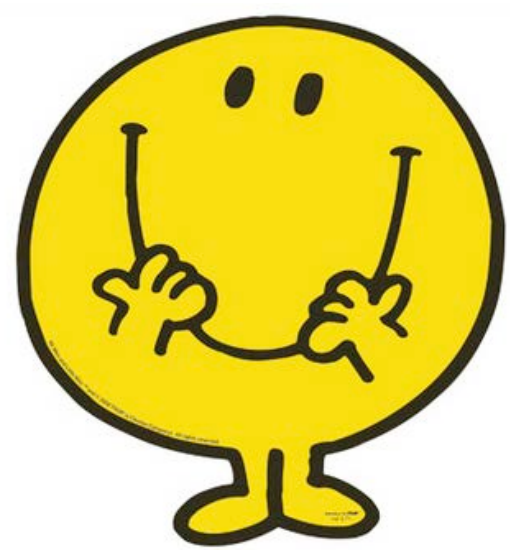
$$x_{i2} = \begin{cases} 1 \; if \; the \; i^{th} \; person \; is \; from \; Mt.Holyoke \\ 0 \; if \; the \; i^{th} \; person \; is \; not \; from \; Mt.Holyoke \end{cases}$$

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

"baseline"

Extending the linear model

The linear regression model provides nice, interpretable results and is a good starting point for many applications

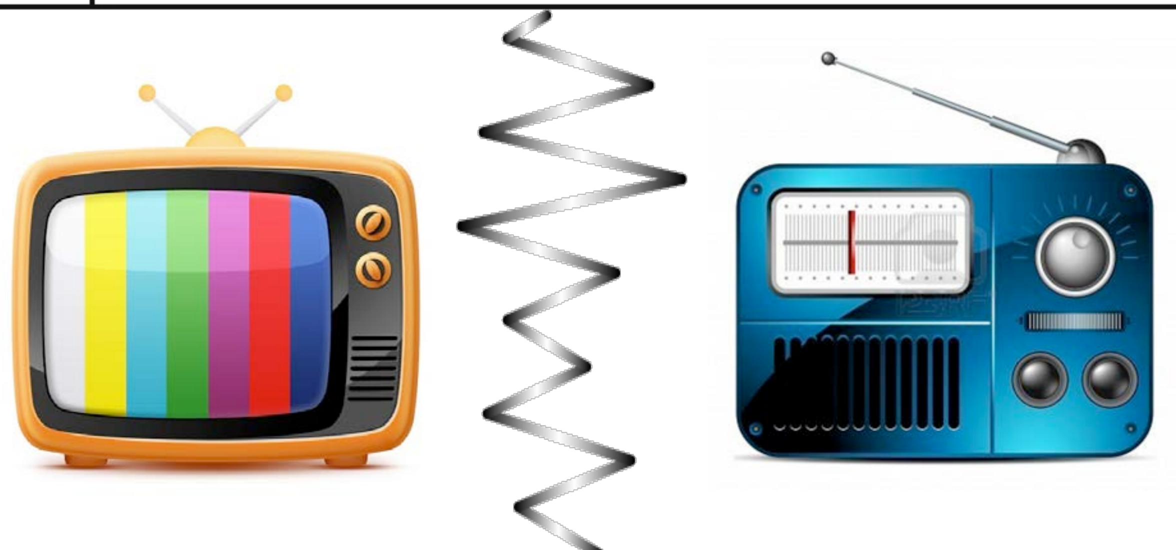


Thanks, Roger Hargreaves!

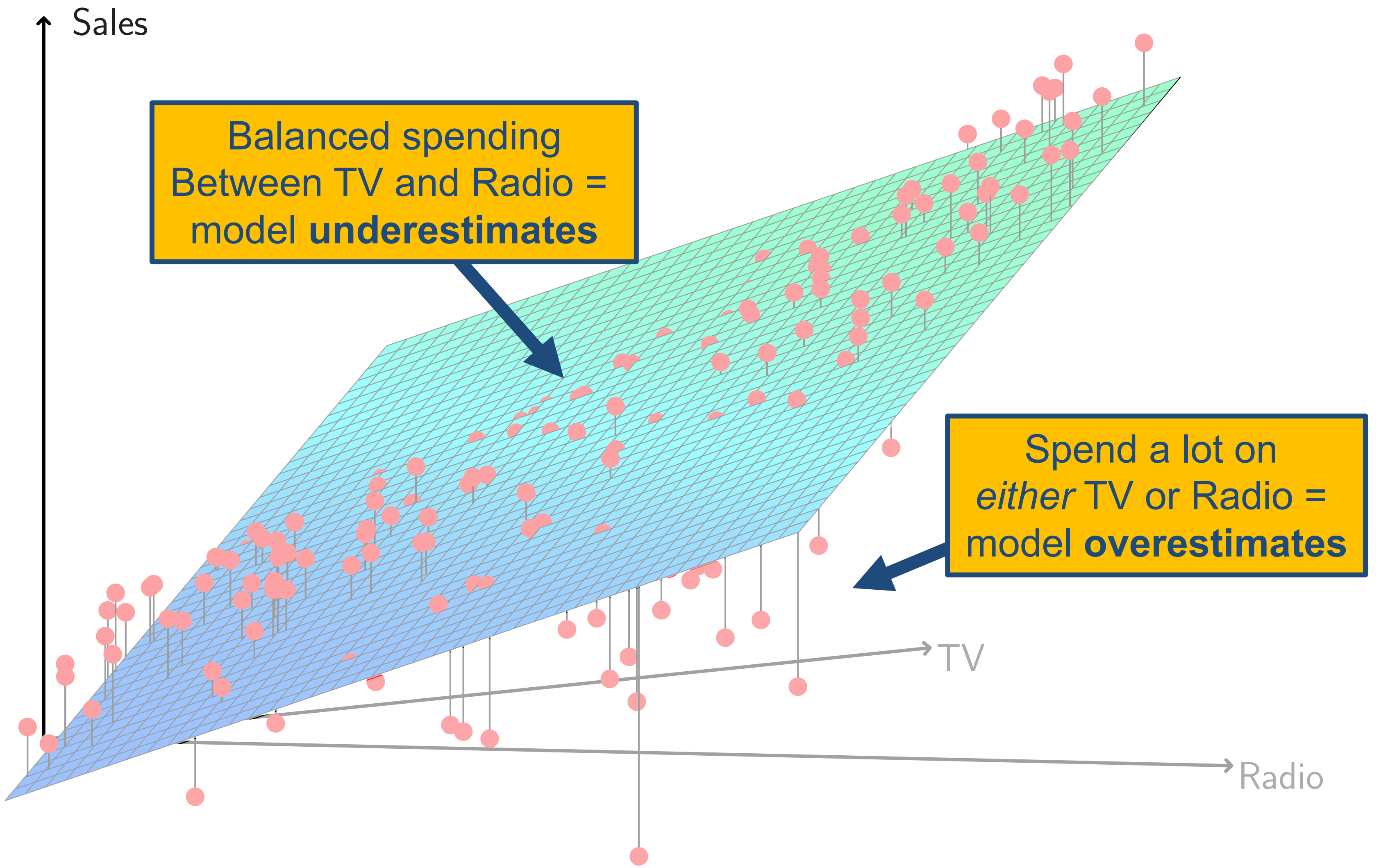
Assumption 1: independent effects

- Think back to our model of the **`Advertising`** dataset

	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~~



Reality: interaction effects



Reality: interaction effects

- Suppose that spending money on **`radio`** advertising actually increases the effectiveness of **`TV`** advertising

- This means that the slope term for **`TV`** should increase as **`radio`** increases

- In the standard linear model, we didn't account for that:

$$Y = \beta_0 + \beta_1 \times \mathbf{radio} + \beta_2 \times \mathbf{TV} + \epsilon$$

Solution: interaction terms

- **One solution**: add a new term

$$Y = \beta_0 + \beta_1 \times \mathbf{radio} + \beta_2 \times \mathbf{TV} + \boxed{\beta_3 \times (\mathbf{TV} \times \mathbf{radio})} + \epsilon$$

- **Question:** how does this fix the problem?

$$= \beta_0 + (\beta_1 + \beta_3 \times \mathbf{TV}) \times \mathbf{radio} + \beta_2 \times \mathbf{TV} + \epsilon$$

$$= \beta_0 + \tilde{\beta}_1 \times \mathbf{radio} + \beta_2 \times \mathbf{TV} + \epsilon$$

slope for **`radio`** now depends on the value of **`TV`**

Solution: Interaction terms

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	6.7502	0.248	27.23	< 0.0001
`TV`	0.0191	0.002	12.70	< 0.0001
`radio`	0.0289	0.009	3.24	0.0014
`TV`×`radio`	0.0011	0.000	20.73	< 0.0001

- p-value for **`TV`×`radio`** is very low (indicating what?)
- R^2 *without* interaction term is 89.7%; this model: 96.8%

diff. var. explained by each model → $(96.8 - 89.7)$

var. missed by first model → $(100 - 89.7)$

$$\frac{(96.8 - 89.7)}{(100 - 89.7)} = 69\%$$

of the variability that our previous model missed is explained by the interaction term

Important caveat

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	6.7502	0.248	27.23	< 0.0001
`TV`	0.0191	0.002	12.70	< 0.0001
`radio`	0.0289	0.009	3.24	0.0014
`TV`×`radio`	0.0011	0.000	20.73	< 0.0001

- In this case, p-values for all 3 predictors are significant
- May not always be true!
- ***Hierarchical principle***: if we include an interaction term, we should include the main effects too (why?)

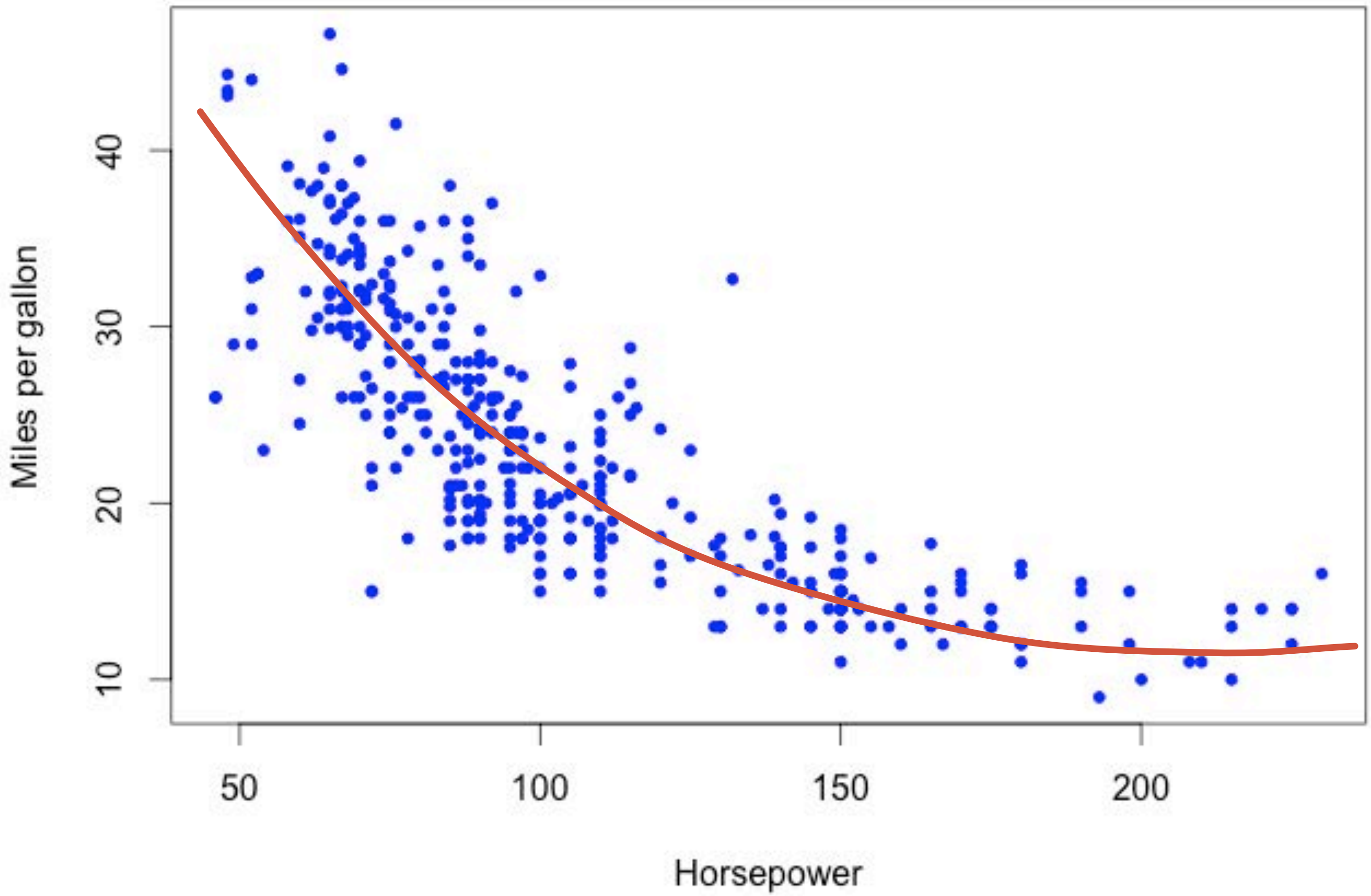
Assumption 2: linear relationships

- LR assumes that there is a straight-line relationship between the predictors and the response

- If the true relationship is far from linear:
  - the conclusions we draw from a linear model are probably flawed
  - the prediction accuracy of the model is likely going to be pretty low

Assumption 2: linear relationships

- For example, in the **`Auto`** dataset:



Solution: polynomial regression

- **Simple approach**: use polynomial transformations

$$\textbf{mpg} = \beta_0 + \beta_1 \times \textbf{horsepower} + \beta_2 \times \textbf{horsepower}^2 + \epsilon$$

- **Note**: still a linear model!

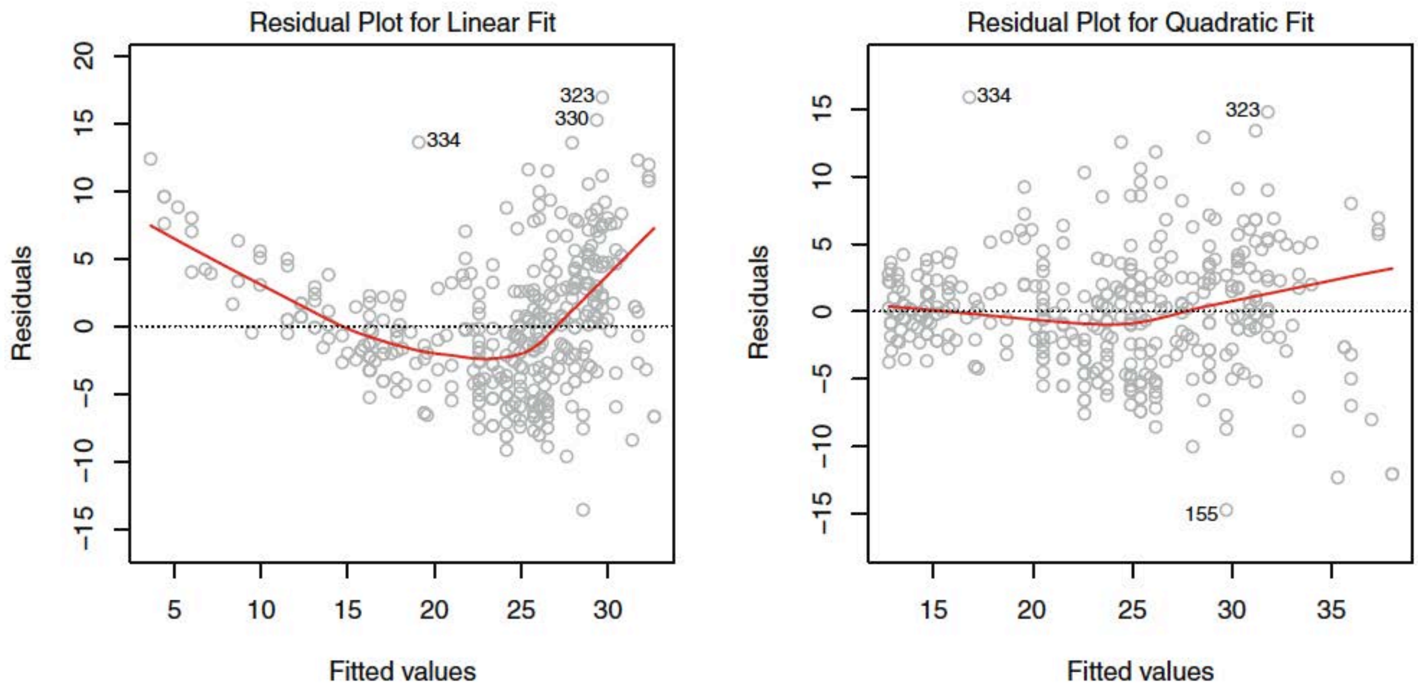
(Want more? We've got some shinier non-linear models coming up in Ch. 7)

Okay, so wait…

How to tell if you need more power

- Residuals plots can help identify problem areas in the model (by highlighting patterns in the errors)

- Ex. LR of **`mpg`** on **`horsepower`** in the **`Auto`** dataset:



Discussion: breaking LR

- What other problems might we run into when using LR?
- How could we fix them?

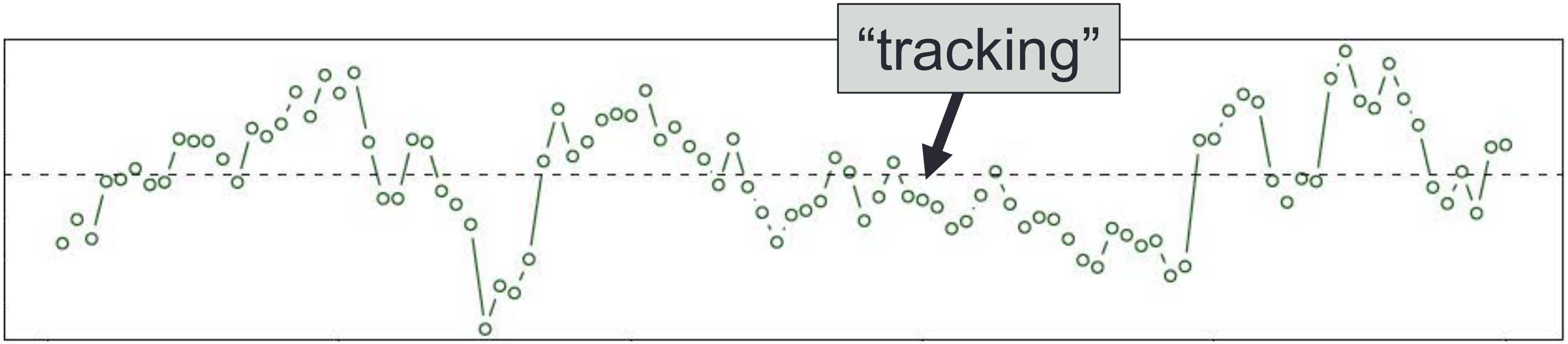
1. Correlated error terms

- LR assumes that the error terms are **uncorrelated**
- If these terms *are* correlated, the estimated standard error will tend to **underestimate** the true standard error
- What does this mean for the associated confidence intervals and p-values?

Question: when might we want to be wary of this?
(*hint*: tick, tock…)

How to tell if error terms are correlated

- In the time-sampled case, we can plot the residuals from our model as a function of time



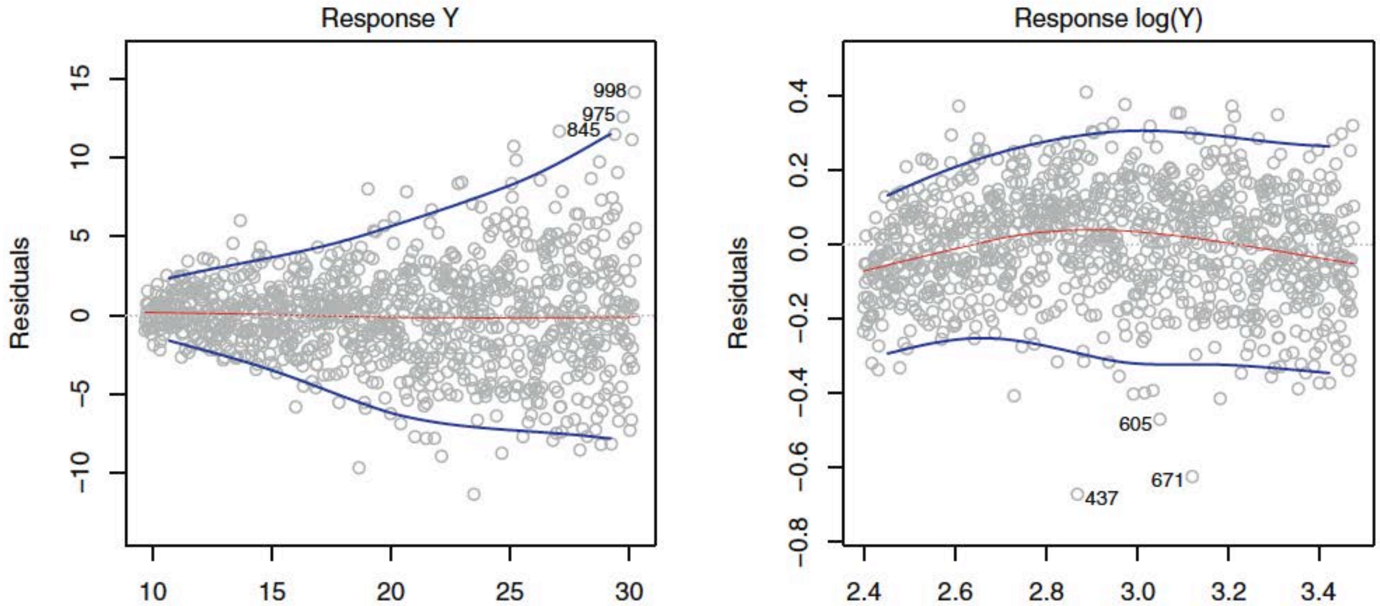
- Uncorrelated errors = no discernable pattern

2. Non-constant variance of error terms

- LR assumes that error terms have constant variance:
$$Var(e_i) = \sigma^2$$
- Often not the case (e.g. error terms might increase with the value of the response)
- Non-constant variance in errors = **heteroscedasticity**

How to identify / fix heteroscedasticity

- The residuals plot will show a funnel shape



- Options:
  - transform the response using a **concave function** (like *log* or *sqrt*)
  - **weight** the observations proportional to the inverse variance

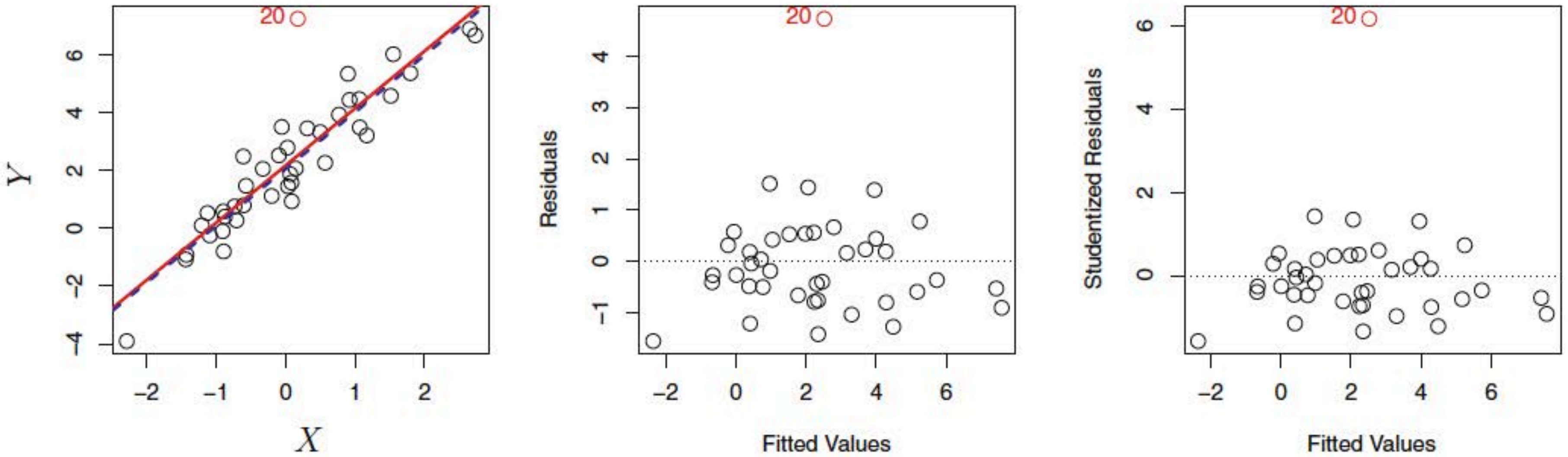
3. Outliers

- **Outlier**: an observation whose true response is **really far** from the one predicted by the model
- Sometimes indicate a problem with the model (i.e. a missing predictor), or might just be a data collection error
- Can mess with RSE and R^2, which can lead us to misinterpret the model's fit

How to identify outliers

- Residual plots can help identify outliers, but sometimes it's hard to pick a cutoff point (how far is "too far"?)

- **Quick fix**: divide each residual by dividing by its estimated standard error (*studentized residuals*), and flag anything larger than 3 in absolute value



“Studentizing”?

- Named for English statistician Wm. Gosset
- Graduated from Oxford with degrees in chemistry and math in 1988
- Published under the pseudonym “Student”

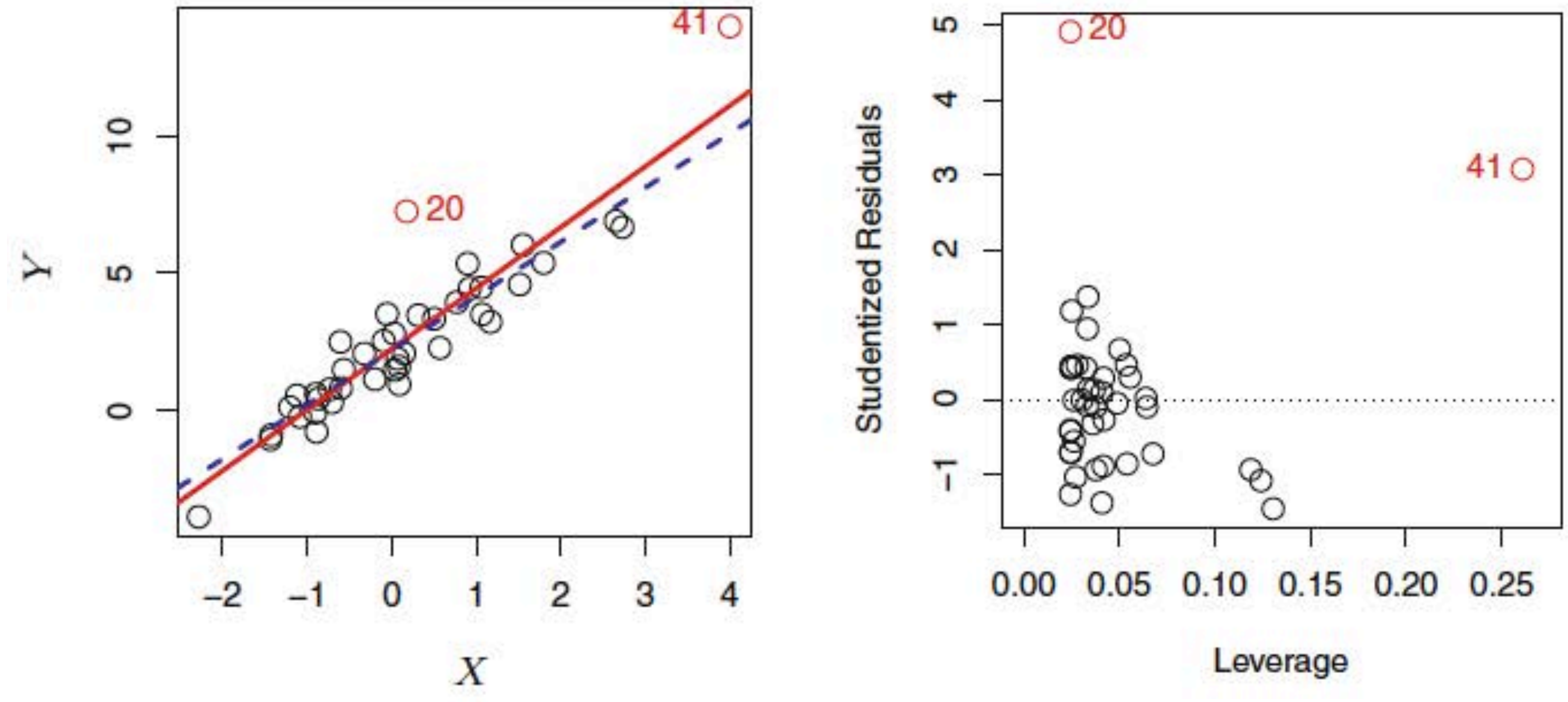
4. High leverage points

- Outliers = unusual values in the **response**
- High leverage points = unusual values in the **predictor(s)**
- The more predictors you have, the harder they can be to spot (why?)
- These points can have a major impact on the least squares line (why?), which could invalidate the entire fit

How to identify high leverage points

- Compute the leverage statistic. For SLR:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



5. Collinearity

- Problems can also arise when two or more predictor variables are closely related to one another

- Hard to isolate the individual effects of each predictor, which **increases uncertainty**

- This makes it harder to detect whether or not an effect is actually present (why?)

Detecting collinearity

- Look at the correlation matrix of the predictors
- **`Auto`** dataset: just about everything is highly correlated

	`mpg`	`cylinders`	`displacement`	`horsepower`	`weight`	`acceleration`	`year`
`mpg`	1	-0.7776175	-0.8051269	-0.7784268	-0.8322442	0.4233285	0.5805410
`cylinders`		1	0.9508233	0.8429834	0.8975273	-0.5046834	-0.3456474
`displacement`			1	0.8972570	0.9329944	-0.5438005	-0.3698552
`horsepower`				1	0.8645377	-0.6891955	-0.4163615
`weight`					1	-0.4168392	-0.3091199
`acceleration`						1	0.2903161
`year`							1
`origin`							

- **Caveat**: this won't help you find interactions between multiple variables when no single pair is highly correlated (called **multicollinearity)**

Approaches for dealing with collinearity

1. **Drop one** of the problematic variables from the regression (linearity implies they're redundant)
2. **Combine** the collinear variables together into a single predictor

Lab: Linear Regression

- To do today's lab in R: **`car`**, **`MASS`** and **`ISLR`** packages
- To do today's lab in python: **`numpy`**, **`pandas`** and **`statsmodels`** libraries
- Instructions and code can be found at:

  R version: [course website]/labs/lab2-r.html

  Python version: [course website]/labs/lab2-py.html

- Original version can be found beginning on p. 109 of ISLR

Assignment 1

- To get credit for today's lab, please post a response to the prompt posted to #lab2
- Assignment 1 posted on course website and Moodle
- Problems from ISLR 3.7 (p. 120-123)
  - Conceptual: 3.1, 3.4, and 3.6
  - Applied: 3.8, 3.10
- Due **Wednesday September 27 by 11:59pm**