## WELCOME TO ANALYTICS FOR STRATEGY

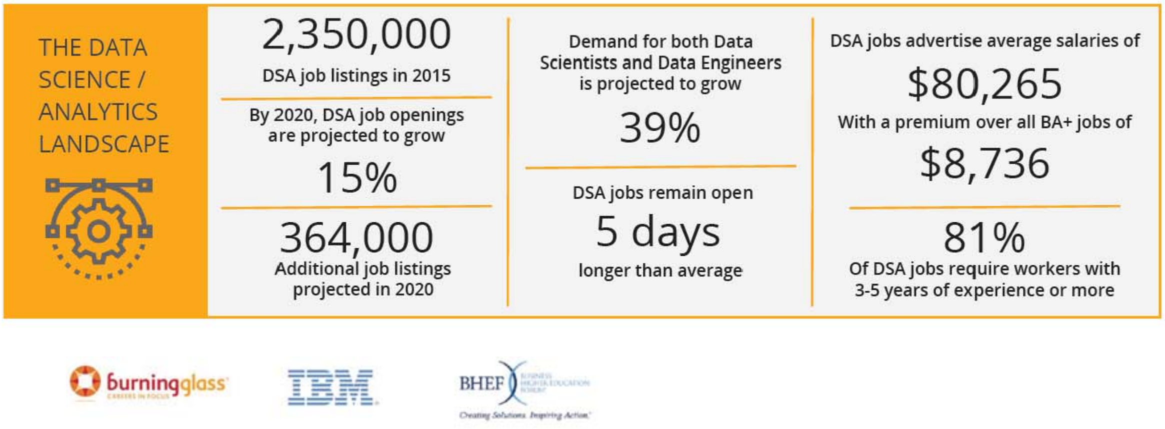
- Please turn on video and mute mic
- Slides are posted to Canvas → Modules → Week 1
- We will be using R today! **Open RStudio** on your computer
- **Assignment #0** is due today (see Canvas)
- Private-chat the VCM (Liz Laurie) if you have Zoom issues

# Course Introduction

**Analytics for Strategy**

# Analytics and data science are disrupting the job market

Influential Burning Glass/IBM report on data science jobs:



THE DATA SCIENCE / ANALYTICS LANDSCAPE

2,350,000
DSA job listings in 2015

By 2020, DSA job openings are projected to grow
15%

364,000
Additional job listings projected in 2020

Demand for both Data Scientists and Data Engineers is projected to grow
39%

DSA jobs remain open
5 days
longer than average

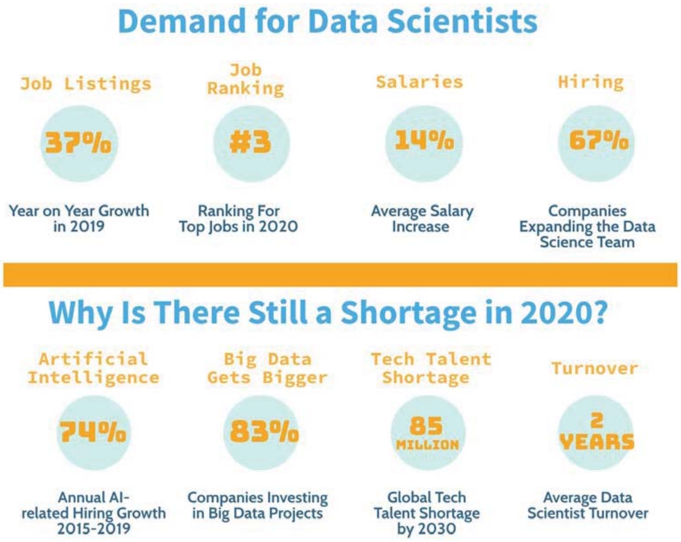
DSA jobs advertise average salaries of
$80,265
With a premium over all BA+ jobs of
$8,736

81%
Of DSA jobs require workers with 3-5 years of experience or more

Source: IBM and Burning Glass (2017). “The Quant Crunch.” https://www.ibm.com/downloads/cas/3RL3VXGA

# Analytics and data science are disrupting the job market

State of the job market in 2020



Demand for Data Scientists

Job Listings
37%
Year on Year Growth in 2019

Job Ranking
#3
Ranking For Top Jobs in 2020

Salaries
14%
Average Salary Increase

Hiring
67%
Companies Expanding the Data Science Team

Why Is There Still a Shortage in 2020?

Artificial Intelligence
74%
Annual AI-related Hiring Growth 2015-2019

Big Data Gets Bigger
83%
Companies Investing in Big Data Projects

Tech Talent Shortage
85 MILLION
Global Tech Talent Shortage by 2030

Turnover
2 YEARS
Average Data Scientist Turnover

Source: DuBois, Jen (2020). “The Data Scientist Shortage in 2020.” *quanthub*. https://quanthub.com/data-scientist-shortage-2020/

# Not just for Silicon Valley anymore

“Data analytics is the oxygen of Wall Street.”

– Tsvi Gal, Morgan Stanley

“data analytics is […] like digital: everyone's going to need to have a base level understanding of it.”

– Bhushan Sethi, PwC

“it's not that we need to have everyone out there go and become the next great data analyst, but we need to have people who understand how to consume the data”

– Steve Kern, Gates Foundation

# Analytics for Strategy

- This is an ***advanced analytics course*** that tackles a ***broad range of strategy questions***
- Fair warning: more analytics than strategy!

# What this course is about

- Using data analytics for decision-making
  - Generating actionable insights
  - Understanding which conclusions an analysis doesn’t support (must know your methods)

- Implementation of statistical analyses using R
- Critical interpretation of statistical analyses
- Just enough theory to use the methods correctly

# Course outline

| Techniques | Week | Substantive Topic |
|---|---|---|
| **Advanced linear regression** | 1 | Introduction |
| | 2 | Aggregate customer analysis |
| | 3 | Aggregate customer analysis |
| **ML classification and clustering** | 4 | Individual customer analysis |
| | 5 | Market segmentation (customers) |
| | 6 | Market segmentation (products) |
| **A/B experiments and prescriptive analytics** | 7 | Product introductions |
| | 8 | Employee management |
| | 9 | Employee management |
| | 10 | Analytics for personal decisions, review |

# Course outline

| Techniques | Week | Technical Sub-Topic |
|---|---|---|
| **Advanced linear regression** | 1 | Regression, multiple comparisons |
| | 2 | Categorical variables, F-tests |
| | 3 | Slope dummies, interactions |
| **ML classification and clustering** | 4 | Binary classification with logit |
| | 5 | K-means analysis |
| | 6 | Hierarchical clustering |
| **A/B experiments and prescriptive analytics** | 7 | Bias, randomization, A/B experiments |
| | 8 | Instrumental variables |
| | 9 | Difference-in-differences |
| | 10 | Fixed effects |

# Regression as a building block

- In **predictive analytics:** analytics for predicting an outcome
  - E.g. how much less likely is a new Facebook user to churn if they added 10+ friends on day 1?

- In **prescriptive analytics:** analytics for making decisions about how to influence an outcome
  - E.g. how much less likely is a new Facebook user to churn if *Facebook itself* devotes page space to friend suggestions instead of ads on day 1?

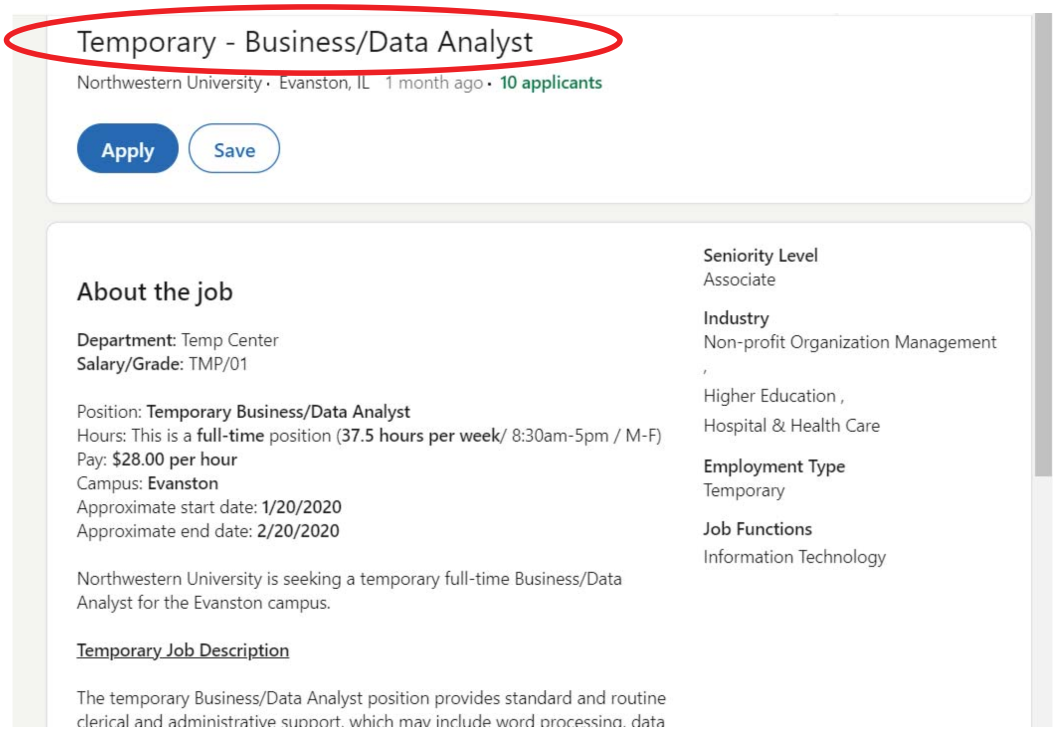
# Regression is a building block in predictive analytics (first half of quarter)

- Predictive analytics is the primary focus of *classical machine learning*
- Regression is a building block for most supervised learning methods:
  - LASSO, ridge regression, regularized regression, splines, classification (logit, multinomial), kernel smoothing, …
- First half of the quarter: basics of predictive analytics

# Regression is a building block in prescriptive analytics (second half of quarter)

- Prescriptive analytics is the focus of *causal inference* methods
  - And the emergent field of *causal machine learning*
- Regression is a building block for:
  - Controlled A/B tests, difference-in-differences, regression discontinuity, instrumental variables, fixed effects, …
- Second half of the quarter: deep dive into prescriptive analytics

# Predictive vs. prescriptive example: LinkedIn posting efficacy



Temporary - Business/Data Analyst

Northwestern University · Evanston, IL 1 month ago · 10 applicants

About the job

**Department:** Temp Center
**Salary/Grade:** TMP/01

Position: **Temporary Business/Data Analyst**
Hours: This is a **full-time** position (**37.5 hours per week**/ 8:30am-5pm / M-F)
Pay: **$28.00 per hour**
Campus: **Evanston**
Approximate start date: **1/20/2020**
Approximate end date: **2/20/2020**

Northwestern University is seeking a temporary full-time Business/Data Analyst for the Evanston campus.

Temporary Job Description

The temporary Business/Data Analyst position provides standard and routine clerical and administrative support, which may include word processing, data

**Seniority Level**
Associate

**Industry**
Non-profit Organization Management
,
Higher Education ,
Hospital & Health Care

**Employment Type**
Temporary

**Job Functions**
Information Technology

# LinkedIn posting efficacy: background

- Postings with a job title have a 10% higher view-to-apply rate

- **Prescriptive question:** should LinkedIn require more detailed postings?
  - Will deter some new listings
  - Only worth it if the increase in the view-to-apply rate is large enough to compensate for deterrence

# LinkedIn posting efficacy: predictive vs. prescriptive distinction

- **Predictive:** accounting for other features of the postings, job posting A (has a title) will have a 10% higher rate than job posting B (no title)

- **Prescriptive:** job posting B will get a 10% higher rate by adding a job title while making no other changes
  - Requires knowing that adding a title *causes* the 10% increase all by itself

# LinkedIn posting efficacy: prescriptive analytics answer

- Does adding a title *cause* the 10% increase?
  - No: more detailed postings tend to come from already-more-attractive companies (larger, better-known employers)

- After applying prescriptive analytics, find that adding a title only causes a 2.4% increase
  - The remaining 7.6 percentage points reflect other differences between listings with vs. without a title

# Rest of this week

1. Course logistics: syllabus, Canvas

2. Regression review
   - Linear approximation
   - Coefficient interpretation
   - Dummy variables

3. Statistical inference
   - Statistical significance
   - Multiple comparisons

# Syllabus

- Lots of helpful information in the Syllabus. Let’s go over it together…

# Tour of Canvas

- R resources
- Weekly modules
  - Data and sample code
  - Readings
- Assignments
- Announcements
- Discussions (replaces emails to professor)

# Class liaison

- Email me if interested!

# Rest of this week

1. Course logistics: syllabus, Canvas

2. **Regression review**
   - **Linear approximation**
   - **Coefficient interpretation**
   - **Dummy variables**

3. Statistical inference
   - Statistical significance
   - Multiple comparisons

# What is a regression?

- A mathematical expression of a relationship between predictor variables and a dependent variable of interest

- In technical terms, a regression equation mathematically describes a real-world **data generating process**

# Regressions express a relationship

- The real-world relationship is approximated additively:

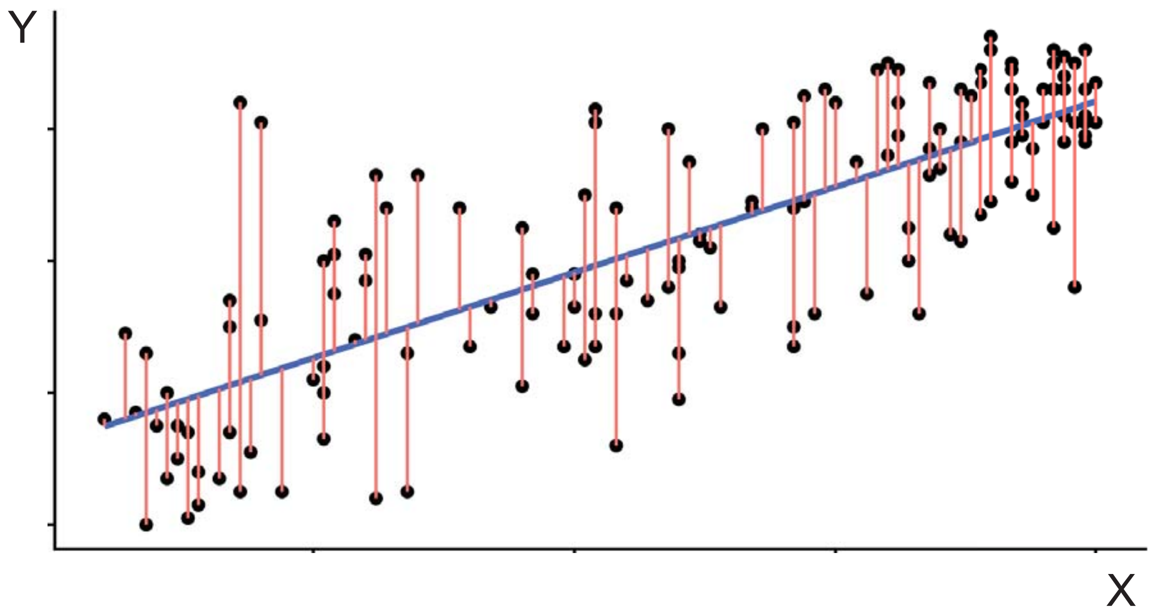
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$
- Synonyms for…
  - X's: RHS variables, regressors, independent variables, explanatory variables, predictors, input variables, attributes, features
  - Y: LHS variable, dependent variable, outcome, output variable, label, target
  - ε: errors, noise terms, stochastic terms

# Regression error term

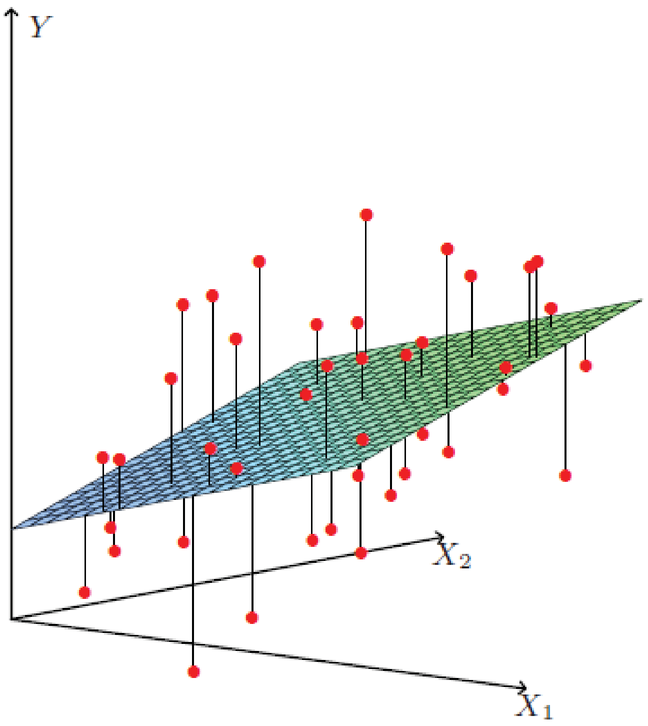
- Regressions do not perfectly determine the relationship between X and Y
- Sometimes we get it “wrong” because the real world is inherently uncertain and probabilistic
- Regression equation includes an **error term** (ε) that makes up the discrepancy between Y and Xβ

- Regression works by minimizing the sum of squared errors (actually residuals)

# Regressions with 1 X-variable



# Regressions with multiple X-variables



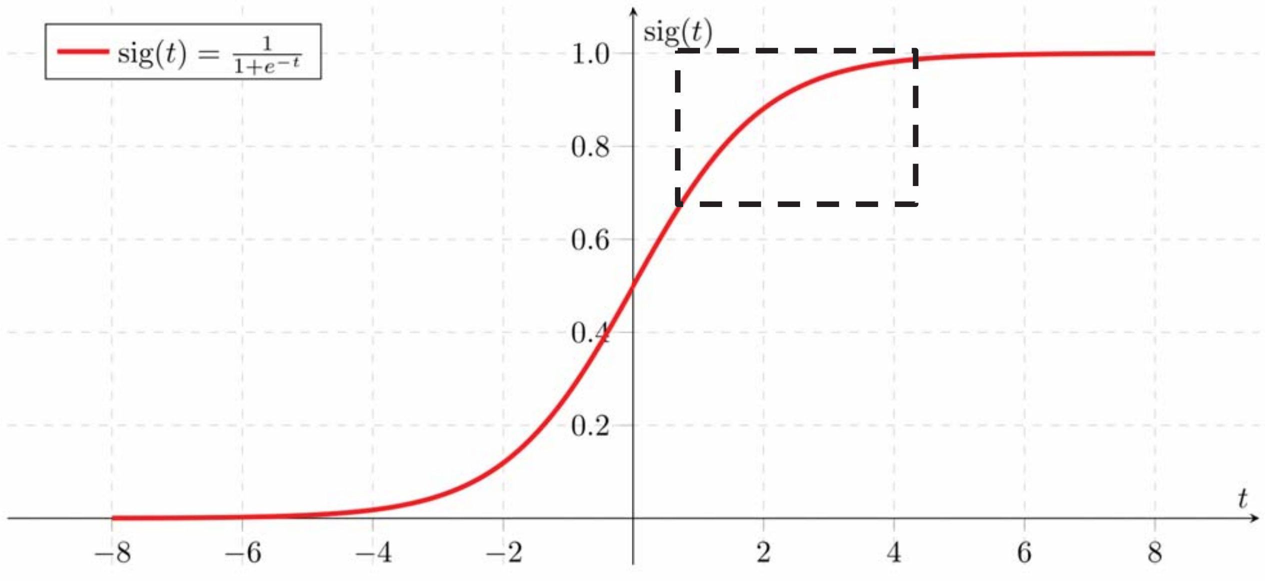
# Regression error term: true error vs. regression residuals

- True error term is the difference between Y and Xβ *when we have the right β*
- In practice, never know if we got β right
  - Can only check **regression residuals**, the difference between Y and X$\hat{\beta}$ using the estimated $\hat{\beta}$
  - Important distinction for causal inference
  - (Note: to avoid clunky PPT features, slides will use β for both estimated $\hat{\beta}$ and true β)

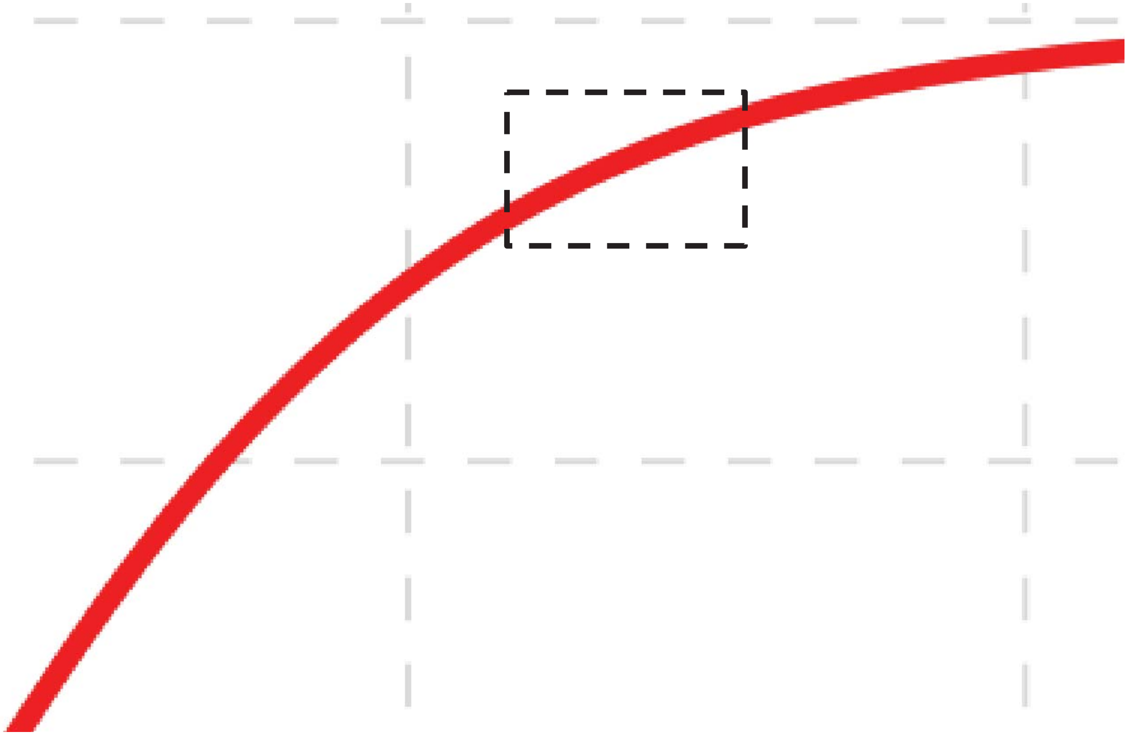
# Why a *linear* regression?

- Most real-world relationships are more complicated than a straight line
- But…
- Linear models are
  - Simple to work with
  - Building blocks for more complex nonlinear models
  - Able to accommodate curvature with additional X-variables
  - Good “first-order approximations” for most nonlinear relationships

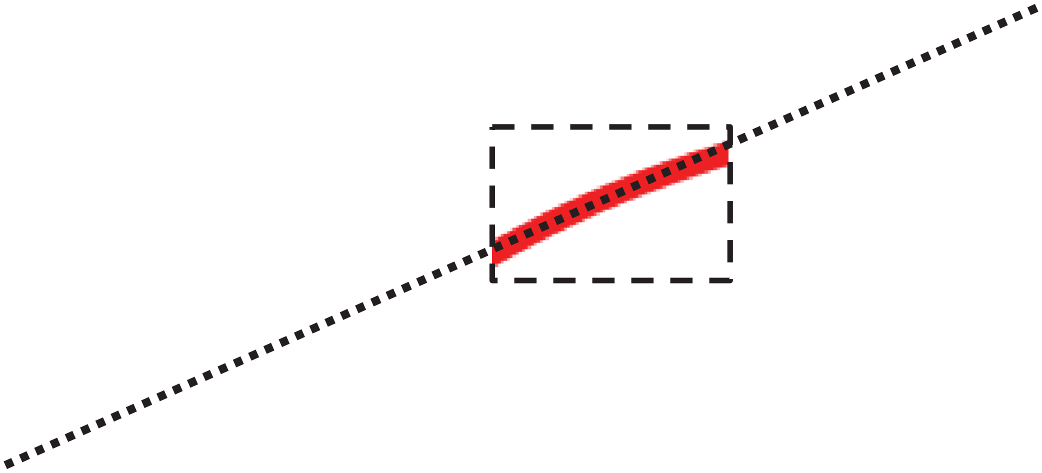
# Why a *linear* regression?



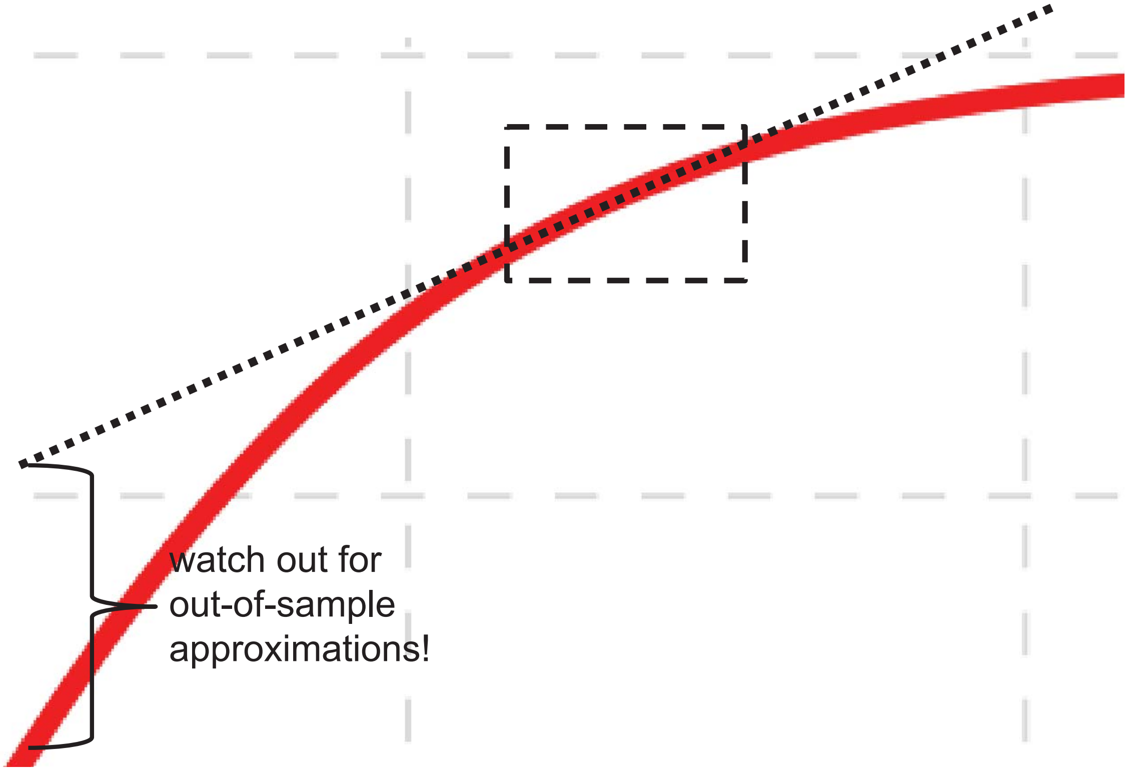
# Why a *linear* regression?



# Why a *linear* regression?



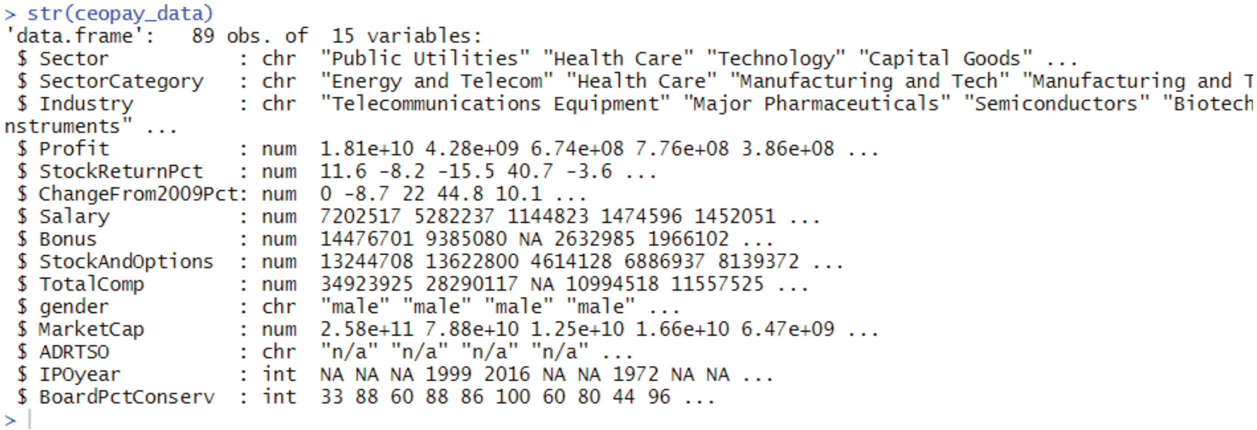
# Why a *linear* regression?



watch out for
out-of-sample
approximations!

# Let’s see a regression!

- CEO compensation data: ceo_pay_usatoday2010.csv



```
> str(ceopay_data)
'data.frame':   89 obs. of  15 variables:
 $ Sector          : chr  "Public Utilities" "Health Care" "Technology" "Capital Goods" ...
 $ SectorCategory  : chr  "Energy and Telecom" "Health Care" "Manufacturing and Tech" "Manufacturing and T
 $ Industry        : chr  "Telecommunications Equipment" "Major Pharmaceuticals" "Semiconductors" "Biotech
nstruments" ...
 $ Profit          : num  1.81e+10 4.28e+09 6.74e+08 7.76e+08 3.86e+08 ...
 $ StockReturnPct  : num  11.6 -8.2 -15.5 40.7 -3.6 ...
 $ ChangeFrom2009Pct: num  0 -8.7 22 44.8 10.1 ...
 $ Salary          : num  7202517 5282237 1144823 1474596 1452051 ...
 $ Bonus           : num  14476701 9385080 NA 2632985 1966102 ...
 $ StockAndOptions : num  13244708 13622800 4614128 6886937 8139372 ...
 $ TotalComp       : num  34923925 28290117 NA 10994518 11557525 ...
 $ gender          : chr  "male" "male" "male" "male" ...
 $ MarketCap       : num  2.58e+11 7.88e+10 1.25e+10 1.66e+10 6.47e+09 ...
 $ ADRTSO          : chr  "n/a" "n/a" "n/a" "n/a" ...
 $ IPOyear         : int  NA NA NA 1999 2016 NA NA 1972 NA NA ...
 $ BoardPctConserv : int  33 88 60 88 86 100 60 80 44 96 ...
> 
```

# Group exercise (4min)

- Go to Canvas → Modules → Week 1
- Download ceo_pay_usatoday2010.csv to computer
- Open RStudio and load in the data:

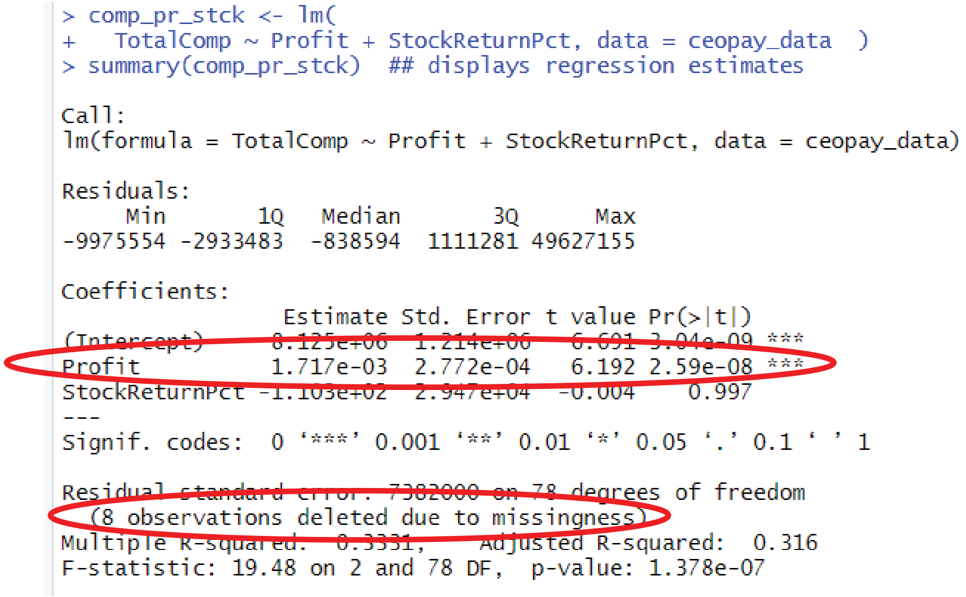
```
ceopay_data <- read.csv(
   "filepathhere/ceo_pay_usatoday2010.csv")
```

- Tips: only use forward slashes in the file path; use `read.csv()` rather than `load()`
- Group component: since this is the first time loading data in class, help each other with data setup in RStudio

- If all breakout group members finish early, take the rest of the 5 min as a break

# Let’s see a regression!

- Regress total pay on profit and stock return



```
> comp_pr_stck <- lm(
+   TotalComp ~ Profit + StockReturnPct, data = ceopay_data  )
> summary(comp_pr_stck)  ## displays regression estimates

Call:
lm(formula = TotalComp ~ Profit + StockReturnPct, data = ceopay_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-9975554 -2933483  -838594  1111281 49627155 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    8.125e+06  1.214e+06   6.691 3.04e-09 ***
Profit         1.717e-03  2.772e-04   6.192 2.59e-08 ***
StockReturnPct -1.103e+02  2.947e+04  -0.004    0.997    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7382000 on 78 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.3331,	Adjusted R-squared:  0.316 
F-statistic: 19.48 on 2 and 78 DF,  p-value: 1.378e-07
```

# Interpreting linear regression coefficients

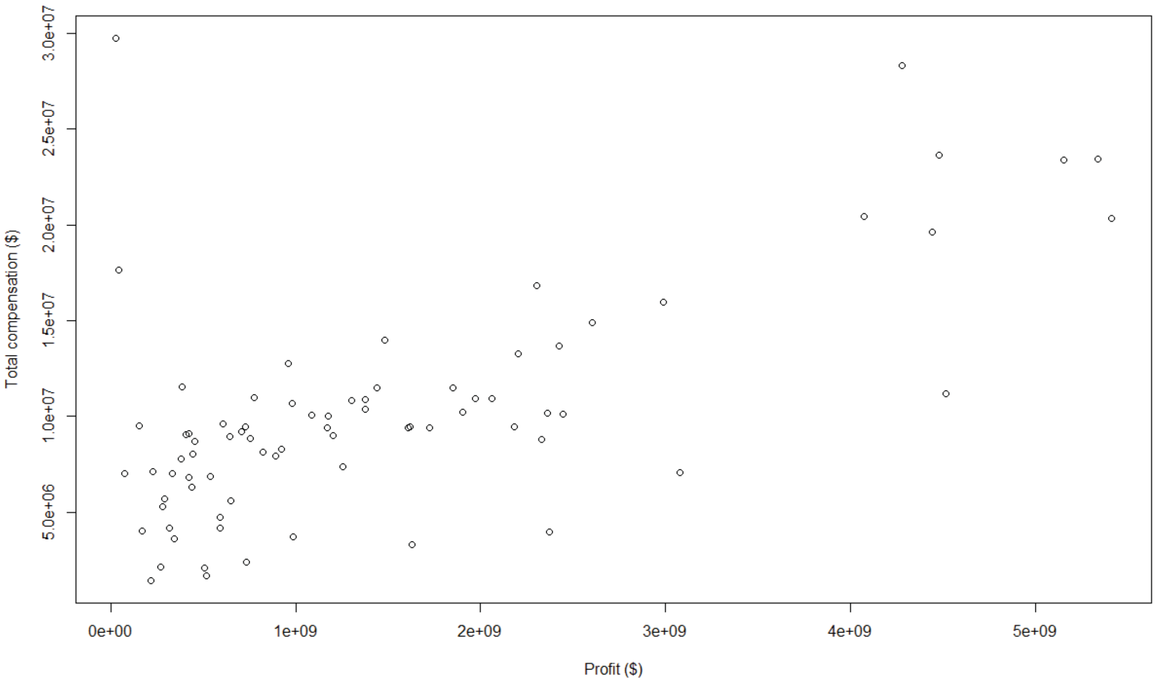
- Regression equation:

  $Pay = \beta_0 + \beta_1 Profit + \beta_2 StockReturn + \varepsilon$
- Estimated coefficients give us:

  $Pay = 8.1M + 0.0017 \times Profit \; - 110.3 \times StockReturn$

- Let’s draw this together in R…

# Interpreting linear regression coefficients



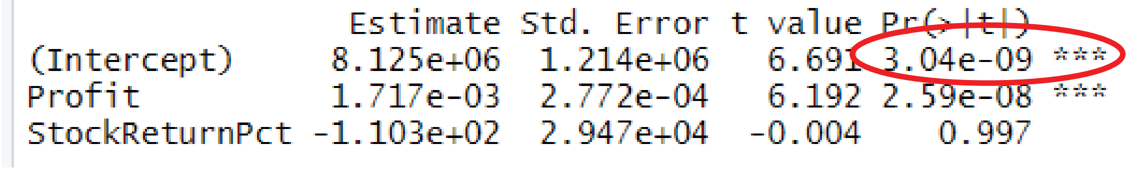
# Interpreting linear regression coefficients

- For an estimated equation
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

- **Constant term $\beta_0$** = Y-intercept *when all X's are 0*
- **Slope coefficient $\beta_1$** = change in Y for a 1-unit increase in $X_1$ while *holding $X_2$ constant*
- **Slope coefficient $\beta_2$** = change in Y for a 1-unit increase in $X_2$ while *holding $X_1$ constant*

# Interpreting the intercept coefficient

- **$\beta_0$ = 0** means regression line passes *through the origin* (intersection of the X-axis and the Y-axis)
- **$\beta_0$ > 0** means regression line crosses the Y-axis $\beta_0$ units *above* the X-axis
- **$\beta_0$ < 0** means regression line crosses the Y-axis $\beta_0$ units *below* the X-axis

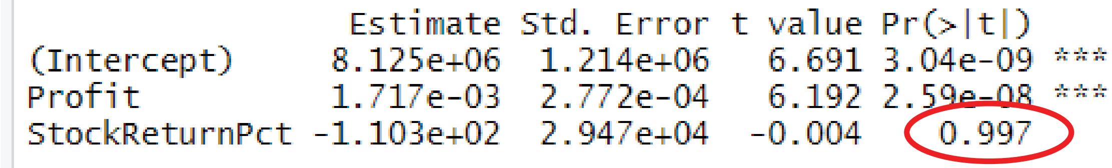


```
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     8.125e+06  1.214e+06   6.691 3.04e-09 ***
Profit          1.717e-03  2.772e-04   6.192 2.59e-08 ***
StockReturnPct -1.103e+02  2.947e+04  -0.004    0.997    
```

- Small p-value means we *can reject the null hypothesis* that Y=0 when all the X-variables are zero

# Interpreting slope coefficients

- $\beta_1$ = 0 means a flat line, Y *doesn't change* when $X_1$ increases
- $\beta_1$ > 0 means Y *increases* when $X_1$ increases
- $\beta_1$ < 0 means Y *decreases* when $X_1$ increases



```
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     8.125e+06  1.214e+06   6.691 3.04e-09 ***
Profit          1.717e-03  2.772e-04   6.192 2.59e-08 ***
StockReturnPct -1.103e+02  2.947e+04  -0.004    0.997    
```

- Large p-value means we *cannot reject the null hypothesis* that Y doesn't change when StockReturnPct changes
  - At best, could estimate a **precise zero** (tight confidence interval that includes zero but excludes large effects)

# Interpreting slope coefficients

- Estimated coefficients give us:

  Pay = 8.1M + 0.0017 × Profit - 110.3 × StockReturn

- Slope coefficient on profit tells us:
  - For every additional $1 in profit and *holding stock return equal*, we expect the CEO to be paid an additional $0.0017
- Interpretation in sensible units:
  - For every additional $10M in profit and *holding stock return equal*, we expect the CEO to be paid an additional $17K

# Interpreting slope coefficients

- Let's see this in R for the slope coefficient on Profit

# Interpreting slope coefficients

| Pro-fit | Stock Return (%) | Expected CEO pay calculation | Expected CEO pay |
|---|---|---|---|
| $1M | 5 | 8,125,406 + 0.0017×1,000,000 - 110.3×5 | $8,126,554.5 |
| $11M | 5 | 8,125,406 + 0.0017×11,000,000 - 110.3×5 | $8,143,554.5 |
| | | *difference (pay gain for $10M profit gain)* | $17K |
| | | | |
| $1M | 5 | 8,125,406 + 0.0017×1,000,000 - 110.3×5 | $8,126,554.5 |
| $1M | 6 | 8,125,406 + 0.0017×1,000,000 - 110.3×6 | $8,126,444.2 |
| | | *difference (pay gain for 1 point stock return gain)* | - $101.30 |

# Interpreting slope coefficients when the X-variable shows up again

- Suppose we instead estimated:

Pay = 5.959M + 0.0036×Profit - 0.1250×(Profit in millions)$^2$

- Slope coefficient on profit tells us:
  - For every additional $10M in profit and *holding profit squared equal*, we expect the CEO to be paid an additional $36K
- But it makes no sense that profit changes while profit squared remains constant!

# Interpreting slope coefficients when the X-variable shows up again

- Suppose we instead estimated:

Pay = 5.959M + 0.0036×Profit - 0.1250×(Profit in millions)$^2$

- We need to account for *all X-variables involving profit*
- Coefficients *on all variables involving profit* tell us:
  – Going from \$0 to \$10M in profit and *holding non-profit X-variables equal*, we expect the CEO to be paid an additional

    $( 0.0036 \times 10{,}000{,}000 - 0.1250 \times (10)^2 )$

    $- (0.0036 \times 0 - 0.1250 \times (0)^2 )$

    $= \$35{,}987.50$

# Rest of this week

1. Course logistics: syllabus, Canvas

2. Regression review
   – Linear approximation
   – Coefficient interpretation
   – **Dummy variables**

3. Statistical inference
   – Statistical significance
   – Multiple comparisons

Digression: assignment for next week

# Assignment for next week

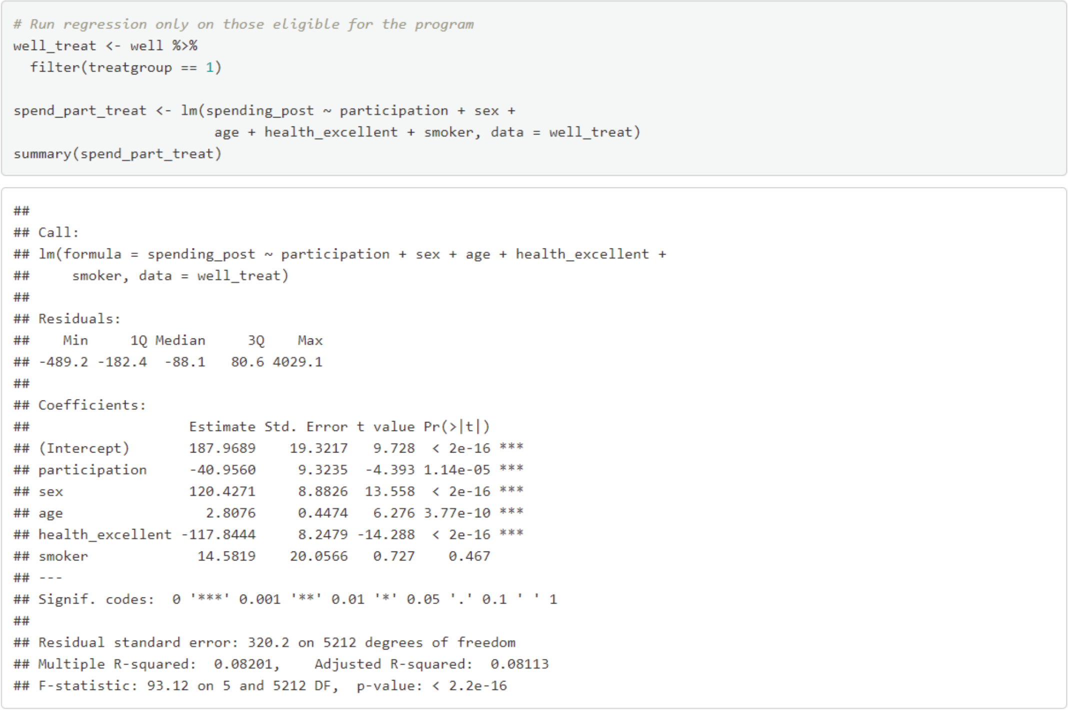
- Tutoring company looking to reach new clients (college students)
- Partnering with a university to help the university identify students at risk of a bad grade in intermediate microeconomics

- Your job: explore the factors associated with intermediate micro grade using regression tools
  - No attempts at causality (yet!)

# Reference slide: assignment formatting

1. Use a .Rmd notebook (see Assignment #0)
2. Put all code into code chunks using ```
3. Put comments, explanations, answers (anything except code and output) outside the code chunks
4. Silence unnecessary output (see setup files)
5. Execute the code and output to HTML
6. Submit the .Rmd file and the HTML output (option: print HTML to PDF)

# Reference slide: assignment formatting example



```
# Run regression only on those eligible for the program
well_treat <- well %>%
  filter(treatgroup == 1)

spend_part_treat <- lm(spending_post ~ participation + sex +
                        age + health_excellent + smoker, data = well_treat)
summary(spend_part_treat)
```

```
##
## Call:
## lm(formula = spending_post ~ participation + sex + age + health_excellent +
##     smoker, data = well_treat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -489.2 -182.4  -88.1   80.6 4029.1
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       187.9689    19.3217   9.728  < 2e-16 ***
## participation     -40.9560     9.3235  -4.393 1.14e-05 ***
## sex               120.4271     8.8826  13.558  < 2e-16 ***
## age                 2.8076     0.4474   6.276 3.77e-10 ***
## health_excellent -117.8444     8.2479 -14.288  < 2e-16 ***
## smoker             14.5819    20.0566   0.727    0.467
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 320.2 on 5212 degrees of freedom
## Multiple R-squared:  0.08201,    Adjusted R-squared:  0.08113
## F-statistic: 93.12 on 5 and 5212 DF,  p-value: < 2.2e-16
```

Participation effect estimate is ~ -30 to -45

# Key points so far

- Strategy decisions can require prediction (correlation is sufficient) or prescription (causation is needed)
- Regression is a linear approximation of a real-world relationship
- Regression can handle more complex relationships using special predictor variables (e.g. dummies)

- Later this week: making inferences from a regression