EDP308: STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020
RAZ: Rebecca A. Zárate, MA

Overview

- Predictions
- The Line of Best Fit
- Regression Line Equation
  - Home Price Example
- Multiple Regression
  - Home Price Example
  - Graduate School Example
  - Attractiveness Example
- Regression in R
  - Simple Linear Regression
  - Multiple Linear Regression

Regression and Prediction

Predictions

- Of course we can not claim causation with a correlation, but we can try to use that correlation to make predications
  - Ex. If you studied X number of hours, you most likely get Y number of questions correct

Would higher or lower values of r result in more accurate predications?

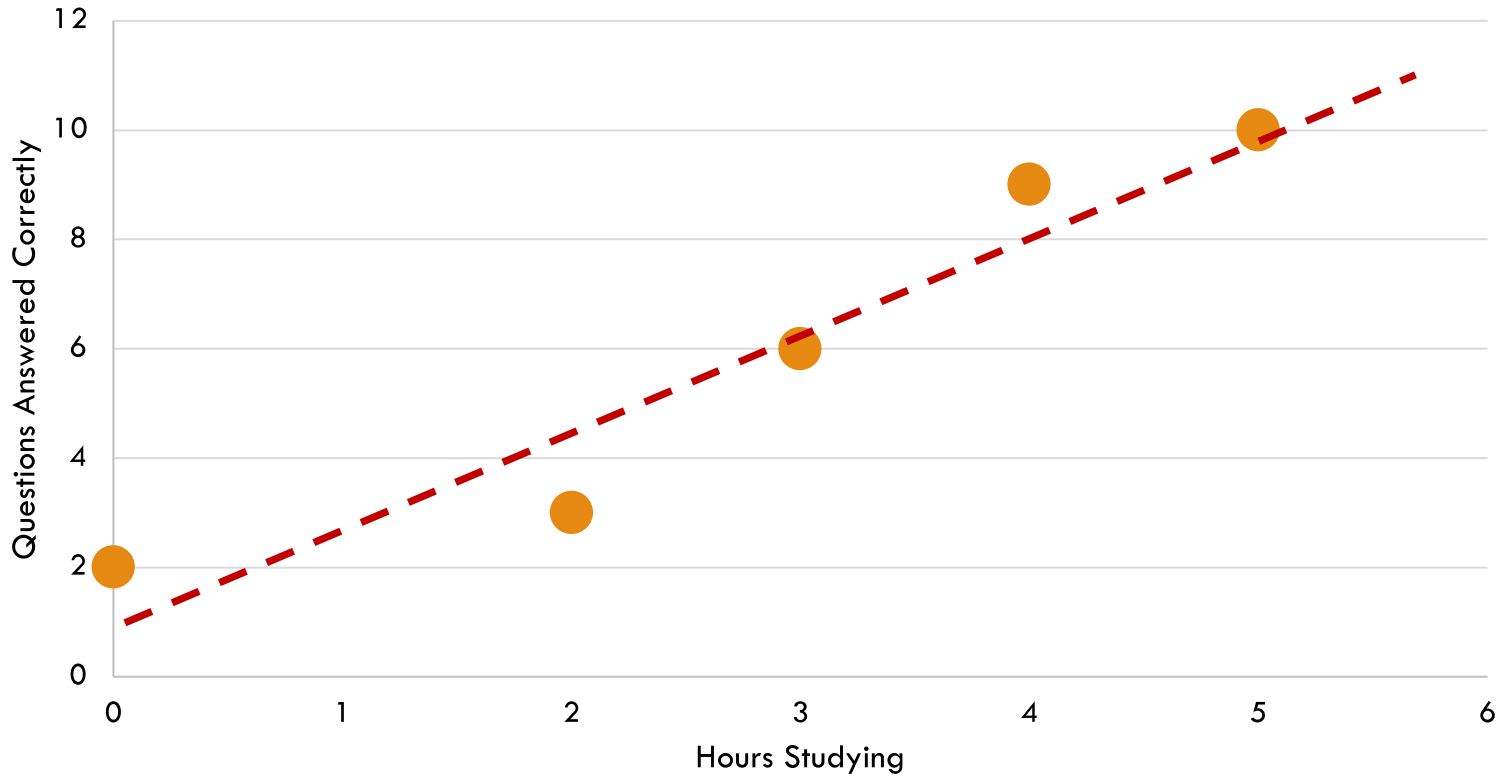
Correlation vs. Regression

- Correlations describe relationships between variables
- Regression uses information from correlations to make predictions
- So if you know how many hours someone studied, you can try to predict how many questions they answered correctly

Regression Analysis

- Regression Analysis is used to measure the linear association between quantitative variables
  - Correlation: describes the strength of a relationship between two variables
  - Regression line: used to predict values for our response variable, Y, using our explanatory variable X

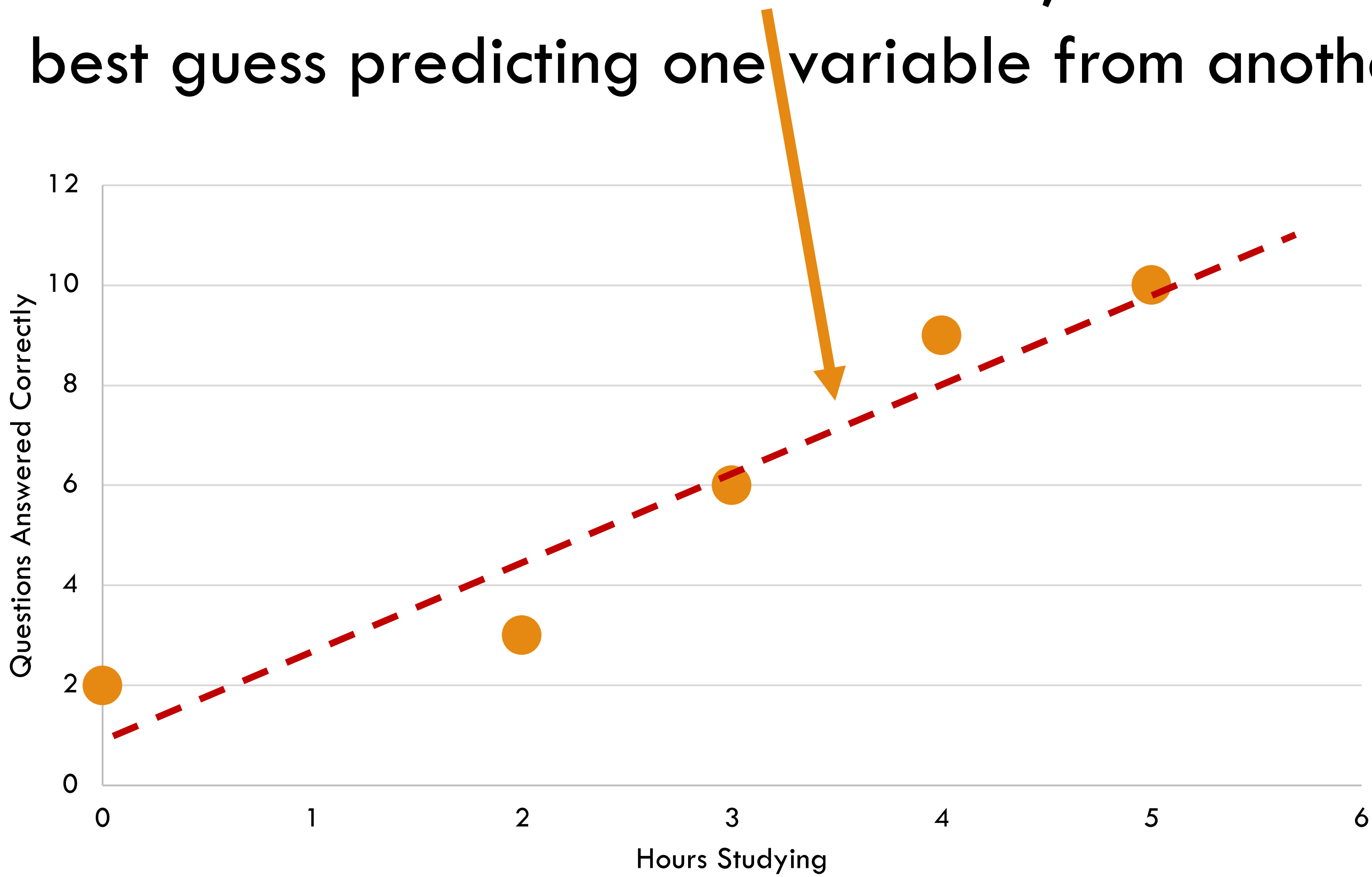
Studying and Questions Correct



If someone studied for 3 hours, how many questions would you predict they will answer correctly?

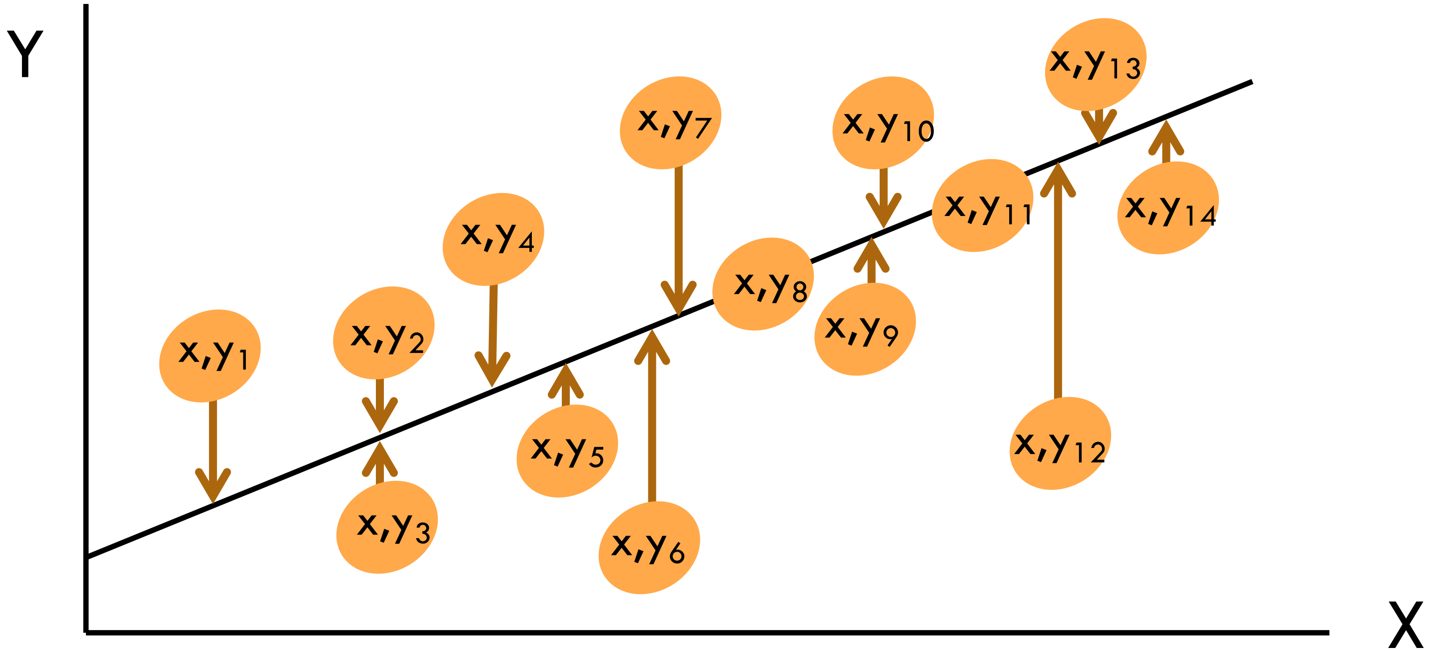
The Line of Best Fit

- This is known as “the line of best fit,” and it is out best guess predicting one variable from another.



Line of Best Fit

- The Line of Best Fit is a regression statistic that attempts to find a line that crosses through the data points or minimizes the <u>distance of points from the line</u>.

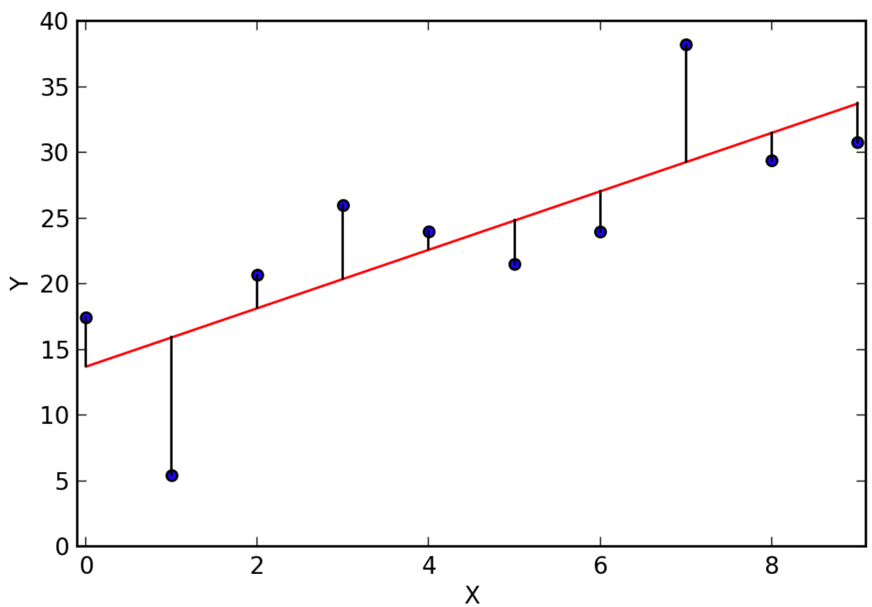


Residue

- The observed values in our data will usually be different from the values predicted from the regression line
- The difference between our observed values and the predicted values from the regression line are called "residuals".

Residual = Observed – Predicted

Do we want the residuals to be as small as possible or as big as possible?

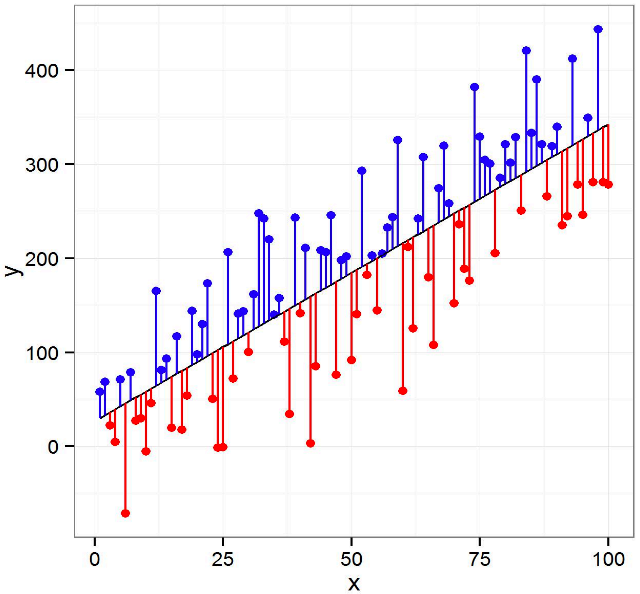


Ordinary Least Squares

- We want residuals to be as small as possible, the technique for doing this is called "Ordinary Least Squares"

Residual = Observed – Predicted

- Observed = Your Actual Score
- Predicted Score = My predication of your score based on how many hours you studied
- Residual = How off my prediction was



(We will not be doing the Ordinary Least Squares part.)

Regression Line Equation

Regression Line Equation

- After using the Ordinary Least Squares technique, it always turns out that the regression line is:

$$\hat{y} = \beta_0 + \beta_1 * x$$

- Where:

Y = b + m*x

  - x is the observed value of the explanatory variable
    - Ex. How many hours you studied.
  - β_1 (beta 1, coefficient) is how much y is predicted to change if x increases by 1 unit. (i.e. β_1 is the slope of the regression line)
    - Ex. For every additional hour you spend studying, your score goes up by x1.5
  - β_0 (beta zero) is the predicted value of y when $x = 0$. (i.e. β_0 is the y-intercept of the regression line)
    - Ex. Your score if you didn't study at all.
  - $\hat{y}$ (y-hat) is the predicted value for our response variable.
    - Ex. The score I predict you will get based on how many hours you studied

Regression Equation Example

Suppose the regression line equation to predict the selling price of a house from the size of the home in square feet is:

$$\widehat{price} = 9161 + 77 * size$$

- What is the explanatory variable, and what is the response variable?
- What is the predicted price of a home that is 0 square feet?
- If a house could increase its size by 1 square foot, how much would we expect the selling price of the home to change?
- What is the predicted price of a house that was 2000 square feet?
- If this house (2000 square feet) actually sold for \$160,000, did our regression line overestimate, or underestimate the selling price? What was the residual?

Regression Equation Example

$$\widehat{price} = 9161 + 77 * size$$

- What is the explanatory variable, and what is the response variable?
  - Size of the home (in square feet), selling price of the home (with the hat)
- What is the predicted price of a home that is 0 square feet?
  - $\beta_0 = \$9161$
- If a house could increase its size by 1 square foot, how much would we expect the selling price of the home to change?
  - $\beta_1 = \$77$
- What is the predicted price of a house that was 2000 square feet?
  - $\$163{,}161 = 9161 + 77 * 2000$
- If this house (2000 square feet) actually sold for \$160,000, did our regression line overestimate, or underestimate the selling price? What was the residual?

$$Residual = Observed - Predicted$$
$$-3{,}161 = 160{,}000 - 163{,}161$$

Finding the Regression Line Equation

Using Ordinary Least Squares, the regression line equation is:

$$\hat{y} = \beta_0 + \beta_1 * x$$

Where:

- $\beta_1 = r\left(\frac{s_y}{s_x}\right)$
- $\beta_0 = \bar{y} - (\beta_1 \bar{x})$

Note: Because of the breakdown of the equations, it is usually best to find β_1 first, and to find β_0 second.

Finding the Regression Line Equation

The table shows the number of hours spent studying (explanatory), and the number of questions correct on a quiz (response).

You are given the following information:

$$\bar{X}_{Hours} = 2.8,\ s_{Hours} = 1.92,$$
$$\bar{Y}_{Questions} = 6,\ s_{Questions} = 3.54$$
$$r = .96$$

1. Find and interpret β_1.
2. Find and interpret β_0.
3. State the regression equation.

Hours	Questions
0	2
2	3
3	6
4	9
5	10

Finding the Regression Line Equation

- Find and interpret β_1

$$\beta_1 = r\left(\frac{s_y}{s_x}\right)$$
$$= .96 * \left(\frac{3.54}{1.92}\right) \approx 1.77$$

For every additional hour you study, you are predicted to get another 1.77 questions correct.

Finding the Regression Line Equation

- Find and interpret β_0

$$\beta_0 = \bar{y} - (\beta_1 \bar{x})$$
$$= 6 - (1.77 * 2.8) \approx 1.044$$

If you did not study at all, you are predicted to get 1.044 questions correct.

Finding the Regression Line Equation

- State the regression equation:

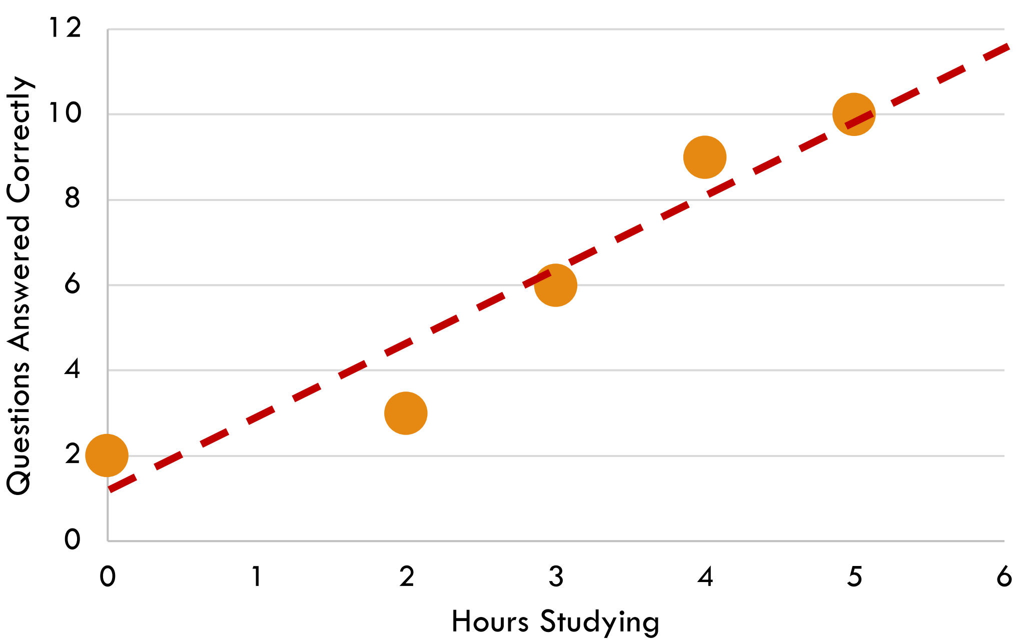
$$\widehat{Questions} = 1.044 + 1.77(Hours)$$

The number of questions you get right increases by 1.77 for every hours you spend studying. If you did not study at all, you'd get ~1 question right

Regression Equation Example

$$\widehat{Questions} = 1.044 + 1.77(Hours)$$

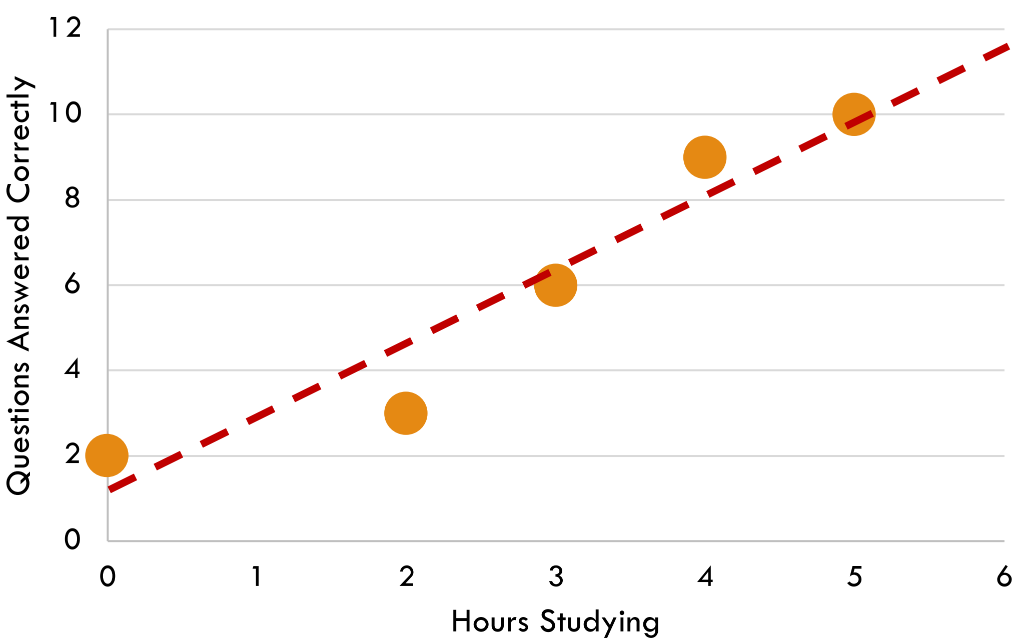
- What is your predicated y-hat ($\hat{y}$) value, i.e. your predicted score if you studied for for:
  - Four hours?
  - Six hours?
  - Zero hours?



Regression Equation Example

$$\widehat{Questions} = 1.044 + 1.77(Hours)$$

- What is your predicated y-hat ($\hat{y}$) value, i.e. your predicted score if you studied for for:
  - Four hours?
    - ~8
  - Six hours?
    - ~11.5
  - Zero hours?
    - ~1



Multiple Regression

Multiple Regression

- Simple linear regression using one predictor variable is all well and good, but rarely are outcomes of interest explained by just one variable.
- With multiple regression, we can add in more than one explanatory variable to try to predict a certain response variable outcome.

Multiple Regression

- The examples are endless…
  - What variables predict the number of asthma related emergency department visits?
  - What variables predict the risk of a person will have a heart attack in the next year?
  - What variables predict your level of happiness?

What outcome would you like to run a regression on?
What are you trying to predict?
What variables do you think are important?

Multiple Regression

- Linear regression can include multiple explanatory variables (in *multiple regression*) as in:

$$\widehat{price} = \beta_0 + \beta_1 * size + \beta_2 * bedrooms$$

  - β_0 is the predicted selling price when **all** predictors (house size and number of bedrooms) are equal to 0.
  - β_1 is the change in predicted selling price when house size increases by 1 square foot, **holding the number of bedrooms constant.**
  - β_2 is the change in predicted selling price when the number of bedrooms increases by 1, **holding house size constant.**

Graduate School

- Suppose the regression line equation predicts the chances of getting into graduate school from the following variables: GPA (in 4.0 format), Research Experience (in years), Publications (number of), and GRE (in GRE format)

$$grad\ \widehat{adm}\ score = 0 + 3 * GPA + 4.5 * Research + 8 * Pub + 1 * GRE$$

  - $\beta_1(GPA) = 3$
  - $\beta_2(Research\ Exp.) = 4.5$
  - $\beta_3(Publications) = 8$
  - $\beta_4(GRE) = 1$

*Numbers are made up.

Graduate School

- What are the explanatory variables, and what is the outcome variable?
- What is the predicted admission score if the applicant had no publications, no research experience, no GRE scores, and a GPA of 2.0?
- Compute the graduate admissions score for the following scenarios:

Scenario 1:
GPA = 3.5
Research Experience = 1.5 years
Publications = 1
GRE = 160

Scenario 2:
GPA = 3.8
Research Experience = 0.5 years
Publications = 0
GRE = 167

$$grad\ \widehat{admission}\ score = 0 + 3 * GPA + 4.5 * Research\ Exp. + 8 * Publications + 1 * GRE$$

Graduate School

- What are the explanatory variables, and what is the outcome variable?
  - Explanatory: GPA, Research Experience, Publications, GRE
  - Response: Graduate Admission Score
- What is the predicted admission score if the applicant had no publications, no research experience, no GRE scores, and a GPA of 2.0?
  - Predicted admission score = 6
- Compute the graduate admissions score for the following scenarios:

Scenario 1:
GPA = 3.5
Research Experience = 1.5 years
Publications = 1
GRE = 160

Scenario 2:
GPA = 3.8
Research Experience = 0.5 years
Publications = 0
GRE = 167

$$grad\ \widehat{admission}\ score = 0 + 3 * GPA + 4.5 * Research\ Exp. + 8 * Publications + 1 * GRE$$

Graduate School

Scenario 1:
GPA = 3.5
Research Experience = 1.5 years
Publications = 1
GRE = 160

Scenario 2:
GPA = 3.8
Research Experience = 0.5 years
Publications = 0
GRE = 167

Scenario 1:

$$185.25 = 0 + 3 * 3.5 + 4.5 * 1.5 + 8 * 1 + 1 * 160$$

Scenario 2:

$$180.650 + 3 * 3.8 + 4.5 * 0.5 + 8 * 0 + 1 * 167$$

$$grad\ \widehat{admission}\ score = 0 + 3 * GPA + 4.5 * Research\ Exp. + 8 * Publications + 1 * GRE$$

Multiple Regression Example: Attractiveness

- What do you find to be the most important features about a partner?
  - Intelligence?
  - Physical appearance?
  - Humor?
  - Athleticism?
  - Spiritual?

Create a regression equation for how attracted you are to someone based on the variables that are important to you.

$$\widehat{attractiveness} = \beta_0 + \beta_1 * Intelligence + \beta_2 * Physical \ \ldots \ \beta_n$$

Up Next…

- For our very last topic in this course, we will switch gears. The statistical tests we've looked at so far use quantitative data. Next we'll see what we can do with some categorical data…

Contingency Tables
Chi-Squared Tests

Regression in R

Studying and Correct Answers in R

```
# The data
hours <- c(0,2,3,4,5)
questions <- c(2,3,6,9,10)

# Correlations are quick with "cor()"
cor(hours, questions)

# To run a simple linear regression in R, use the "lm()" function
regression_model <- lm(questions ~ hours)

# Then look at the output from the regression using the "summary()" function
summary(regression_model)
```

Studying and Correct Answers R Output

The Intercept: How many questions you'd get right if you didn't study at all.

The Coefficient/Slope, The Explanatory Variable: How many more questions you are expected to get right for each additional hour you study.

```
Call:
lm(formula = questions ~ hours)

Residuals:
      1       2       3       4       5 
 0.9189 -1.5946 -0.3514  0.8919  0.1351 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   1.0811     1.0256   1.054   0.3692  
hours         1.7568     0.3121   5.629   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.201 on 3 degrees of freedom
Multiple R-squared:  0.9135,	Adjusted R-squared:  0.8847 
```

R-Squared: How much of the variance is accounted for by the model, here about 91%.

p-values for the predictor variable

Chance of Admission (Multiple Regression)

- Here we are looking at a (real) dataset of 500 Indian students applying to graduate school. We are going to run a regression to try to determine what increases your chances of getting into grad school.
  - Variables:
    - Explanatory Variables
      - GRE Score (out of 340)
      - TOEFL Score (out of 120)
      - University Rating (out of 5)
      - Statement of Purpose, SOP (out of 5)
      - Letters of Recommendation, LOR (out of 5)
      - GPA (out of 10)
      - Research Experience (binary variable, yes = 1, no = 0)
    - Outcome Variable
      - Chance of Admission (ranging from 0 to 1)

Which variables do you think will be most explain of getting into graduate school?

Data Source: https://www.kaggle.com/mohansacharya/graduate-admissions/discussion/82214

Chance of Admission (Multiple Regression)

```r
########################################
###### Multiple Linear Regression #######
########################################

# The data
admissions <- read.csv("admission_predict.csv")

# Look at the variables
names(admissions)

# To run a simple linear regression in R, use the "lm()" function
admissions_model <- lm(ChanceofAdmit ~ GREScore + SOP + LOR + GPA + Research, data = admissions)

# Then look at the output from the regression using the "summary()" function
summary(admissions_model)
```

Multiple regression in R is the same as simple, but you just add more predictor variables with the “+” sign

Chance of Admission (Multiple Regression) Output

Which explanatory variables were significant? Which had the biggest affect on the chances of getting into grad school?

```
Call:
lm(formula = ChanceofAdmit ~ GREScore + SOP + LOR + GPA + Research,
    data = admissions)

Residuals:
      Min        1Q    Median        3Q       Max
-0.270994 -0.023461  0.007897  0.035155  0.164322

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.337952   0.102417 -13.064  < 2e-16 ***
GREScore     0.002677   0.000449   5.963 4.73e-09 ***
SOP          0.006175   0.004250   1.453 0.146874
LOR          0.017902   0.004144   4.320 1.88e-05 ***
GPA          0.130108   0.009275  14.028  < 2e-16 ***
Research     0.023930   0.006666   3.590 0.000364 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06068 on 494 degrees of freedom
Multiple R-squared:  0.817,	Adjusted R-squared:  0.8151
```

Chance of Admission (Multiple Regression) Output

All but the statement of purpose (SOP) were significant. GPA appears to have the strongest affect on your chances of getting into graduate school.

About 81% of the variance in the chance for admission is explained by the model.

```
Call:
lm(formula = ChanceofAdmit ~ GREScore + SOP + LOR + GPA + Research,
    data = admissions)

Residuals:
      Min        1Q    Median        3Q       Max
-0.270994 -0.023461  0.007897  0.035155  0.164322

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.337952   0.102417 -13.064  < 2e-16 ***
GREScore     0.002677   0.000449   5.963 4.73e-09 ***
SOP          0.006175   0.004250   1.453 0.146874
LOR          0.017902   0.004144   4.320 1.88e-05 ***
GPA          0.130108   0.009275  14.028  < 2e-16 ***
Research     0.023930   0.006666   3.590 0.000364 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06068 on 494 degrees of freedom
Multiple R-squared:  0.817,     Adjusted R-squared:  0.8151
```