Intro to Social Science Data Analysis

Lecture 10: Comparing Proportions & Simple Linear Regression

Christopher Gandrud

November 5, 2012

Assignment 3

Due: Friday 16 November

Have a data set with three variables of the following type:

- 1 numeric variable,
- 1 dummy variable,
- 1 multinomial variable.

There should be more than 50 observations per variable & variable category.

Assignment 3

Due: Friday 16 November

Have a data set with three variables of the following type:

- 1 numeric variable,
- 1 dummy variable,
- 1 multinomial variable.

There should be more than 50 observations per variable & variable category.

Assignment 3

Due: Friday 16 November

Have a data set with three variables of the following type:

- 1 numeric variable,
- 1 dummy variable,
- 1 multinomial variable.

There should be more than 50 observations per variable & variable category.

Find the answers to these questions.

Numeric Continuous Variable

- What do you predict the population mean of this variable is?
- Create two groups of this variable based on the dummy variable. Are the population means of these two groups likely to be different? (extra points if you can show this graphically)

Categorical Variables

- Do the two groups of the dummy variable have values on the multinomial variable that are independent of one another?

Find the answers to these questions.

Numeric Continuous Variable

- What do you predict the population mean of this variable is?
- Create two groups of this variable based on the dummy variable. Are the population means of these two groups likely to be different? (extra points if you can show this graphically)

Categorical Variables

- Do the two groups of the dummy variable have values on the multinomial variable that are independent of one another?

Find the answers to these questions.

Numeric Continuous Variable

- What do you predict the population mean of this variable is?
- Create two groups of this variable based on the dummy variable. Are the population means of these two groups likely to be different? (extra points if you can show this graphically)

Categorical Variables

- Do the two groups of the dummy variable have values on the multinomial variable that are independent of one another?

Quick Quiz (1)

Create a hypothesis test to examine whether infant morality rates are on average different in OECD countries compared to non-OECD countries.

Quick Quiz (2)

What conditions do we need to meet in order to use the Central Limit Theorem to assume that our sampling distribution is normally distributed?

If we have a small sample size (< 50) what alternative sampling distribution could we use?

Quick Quiz (3)

What is a p-value?

What two steps do you take to calculate a p-value?

How does a p-value compare to a confidence interval?

Last class we learned how to draw inferences from sample means.

This is useful for continuous numeric variables, but what if we have **categorical variables**?

Today, Part I

In the first part of today's lecture we will learn how to make inferences from **sample proportions**.

Remember the Mode:

For categorical variables the best measure of central tendency is the **mode**.

A way of measuring the mode in a **meaningfully comparable way** is with **proportions**.

In general, for categorical data we are interested in **inferring population proportions**. These proportions are our **population parameter** of interest.

Quick Quiz

Imagine we have a random sample of 275 juries in a US country. Overall the juries have the following racial composition:

White	Black	Hispanic	Other	Total
205	26	25	19	275

Find the **sample proportions** for each racial group.

Example from Diaz et al. Ch. 5.

Sampling Proportions

Table: Racial Composition of Sample Juries

	White	Black	Hispanic	Other	Total
Sample Count	205	26	25	19	275
Sample Proportion	0.75	0.09	0.09	0.07	1

Question

Is the racial composition of the juries similar to the racial composition of the county's population?

Sampling Proportions vs. Population Proportions

Table: Racial Composition of Sample Juries & County's Registered Voters

	White	Black	Hispanic	Other	Total
Sample Count	205	26	25	19	275
Sample Proportion	0.75	0.09	0.09	0.07	1
Registered Voters	0.72	0.07	0.12	0.09	1

They are different, but are they **statistically different** or is the difference simply due to **sampling error**?

Null Hypothesis:

H_0 : The jurors are randomly sampled from the county's population. There is no racial bias in jury selection.

Alternative Hypothesis:

H_A : The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

How do we test these hypotheses?

1st the Test Statistic

Last week we used the following equation for a **test statistic**:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of the point estimate}}$$

Let's use a similar strategy to find the test statistic for the proportions.

What is the null value?

Null Value for Proportions

Our null value is the **expected frequencies in the sample** if the **null hypothesis is true**.

Table: Expected Racial Composition if Null Hypothesis is True

	White	Black	Hispanic	Other	Total
Sample Count	205	26	25	19	275
Registered Voters	0.72	0.07	0.12	0.09	1
Expected Frequency	198	19.25	33	24.75	275

The Test Statistic (1)

Now we can calculate the test statistic for *white* jurors.

$$Z_{white} = \frac{205 - 198}{\sqrt{198}} = 0.5$$

The Test Statistic (2)

We can also calculate the test statistic for the other racial groups.

$$Z_{black} = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$

$$Z_{hispanic} = \frac{25 - 33}{\sqrt{33}} = -1.39$$

$$Z_{other} = \frac{19 - 24.75}{\sqrt{24.25}} = -1.16$$

χ^2 (1)

Our hypotheses were about how whether **all** of the sample proportions were different from the population proportions.

How can we combine these four test statistics together?

χ^2 test statistic:

$$\chi^2 = Z_1^2 + Z_2^2 \ldots Z_n^2$$

For our example this would be:

$$\chi^2 = 0.5^2 + 1.54^2 + -1.39^2 + -1.16^2 = 5.89$$

Note: χ^2 is pronounced “ki squared”.

χ^2 Distribution

We can't assume that the χ_2 statistic follows a normal or t distribution if the null hypothesis is true.

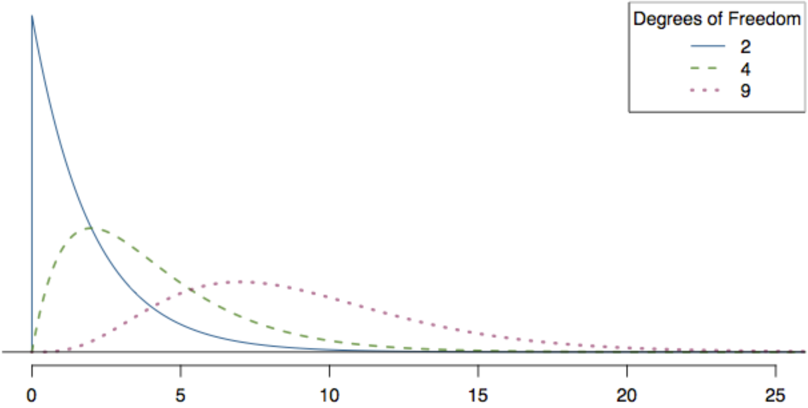
Instead, we use a χ^2 **distribution**.

It's only parameter is the **degrees of freedom** (df).

If k is the number of categories then

$$df = k - 1$$

The χ^2 distribution with various degrees of freedom



Diaz et al. (2011, 216)

Our Example

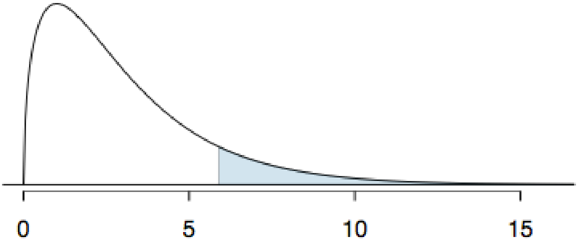
In our example we have:

$$\chi^2 = 5.89$$

$$df = 4 - 1 = 3$$

What is the probability of finding data at least as favourable to the alternative hypothesis as this, if the null hypothesis was true?

The χ^2 distribution with 3 degrees of freedom



Diaz et al. (2011, 219)

Finding the p-value in R

To find the p-value in R for χ^2 of 5.89 when there are 3 degrees of freedom:

```
# Find p-value
1 - pchisq(q = 5.89, df = 3)

## [1] 0.1171
```

At the 95% significance level we fail to reject the null hypothesis that the jurors are randomly chosen from the country population.

Conditions for the χ^2 Test:

- Each case that contributes a count must be **independent** of the other cases.
- Each cell count must be **10 or greater**.

Conditions for the χ^2 Test:

- Each case that contributes a count must be **independent** of the other cases.
- Each cell count must be **10 or greater**.

We can use a similar test to examine if groups in samples are different, i.e. if they are **independent**

Do disadvantaged children who attended preschool have better life outcomes than children who did not go to preschool?

Abecedarian Study Data (Campbell et al., 2002)

Table: Two-way Contingency Table of Selected Data Age 21 Follow-Up Data from the Abecefarian Study

		Preschool	No Preschool	Total
University Enrollment	Enrolled	37	7	44
	Not Enrolled	16	44	60
Total		53	51	104

Null Hypothesis:

H_0 : There is no difference in university enrollement at age 21 between disadvantaged children who attended preschool and those who didn't. (dependent)

Alternative Hypothesis:

H_A : There is a difference in university enrollement at age 21 between disadvantaged children who attended preschool and those who didn't. (independent)

Test Statistic (1)

First find the expected counts if the null hypothesis was true.

We observe that 44 of the 104 people attend unversity. This is a proportion of:

$$\frac{44}{104} = 0.423$$

So we would expect that the number of people who attended preschool and are now in university would be:

$$0.423 * 53 = 22.42$$

General Frequency Count Equation

The general formula for finding **expected count** for a row i in column j is:

$$\text{Expected Count}_{\text{i,j}} = \frac{(\text{row } i \text{ total}) * (\text{column } j \text{ total})}{n}$$

Table: Expected vs. Observed Counts

		Pres.		No Pres.		Total
University Enrollment	Enrolled	37	(22.42)	7	(21.57)	44
	Not Enrolled	16	(30.58)	44	(29.43)	60
Total		53		51		104

Note: Expected countes in parentheses.

Test Statistic (2)

Now we find the test statistic in a similar way to what we did with the one-way table.

If f_o is the observed frequency and f_e is the expected frequency, then:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

The equation for the degrees of freedom (df) is a little different:

$$df = (\text{number of rows} - 1) * (\text{number of columns} - 1)$$

Finding χ^2 & df

$$\chi^2 = 9.48 + 9.842 + 6.951 + 7.213 = 33.486$$

$$df = (2-1) * (2-1) = 1$$

Finding the p-value in R

To find the p-value in R for χ^2 of 33.486 when there is 1 degree of freedom:

```
# Find p-value
1 - pchisq(q = 33.486, df = 1)

## [1] 7.178e-09
```

At the 95% significance level we reject the null hypothesis that there is no difference in university enrollment at age 21 between the people who attended preschool and those who didn't.

To do this in R (the easy way):

```
# Create Contingency Table
Preschool <- c(37, 16)
NoPreschool <- c(7, 44)
Data <- data.frame(Preschool, NoPreschool)

# Find chi2 and p-value
chisq.test(Data)

##
##  Pearson's Chi-squared test with Yates'
##  continuity correction
##
## data:  Data
## X-squared = 31.24, df = 1, p-value =
## 2.284e-08
```

So far we have used tools of statistical inference to determine

- determine likely population parameters from a sample, especially the mean & proportions,
- determine if groups are independent.

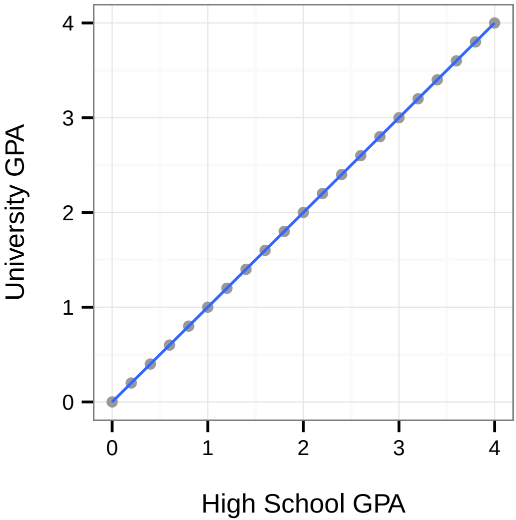
What if we want to use the value of one variable to predict the value of another variable?
Or at least describe the relationship between variables in more detail than "they are independent or not"?

Simple Linear Regression

Question

How closely does a high school GPA predict someone's university GPA?

If there was a perfect linear relationship we would expect to see data like this:



Simplest Equation

We could describe this relationship with the following equation:

If University GPA is denoted y and High School GPA is denoted x

$$y = x$$

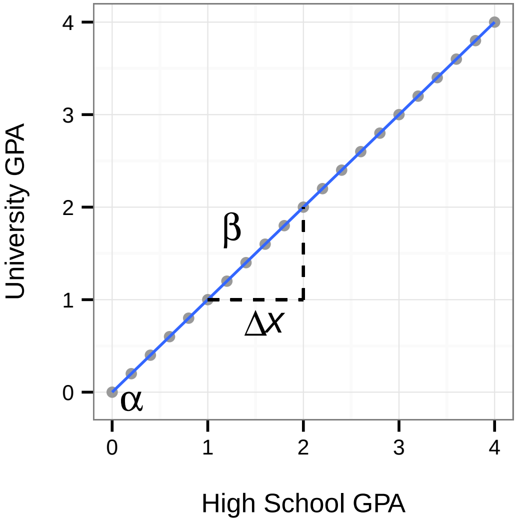
More General Equation

We could use a slightly more general equation:

If α is the line's **y-intercept** and β (**coefficient**) is the slope of the line then:

$$y = \alpha + \beta x$$

This is known as the **simple linear regression equation**.


University GPA
High School GPA
0
1
2
3
4
0
1
2
3
4
β
Δx
α

Perfectly Predicts

In our example if High School GPA perfectly predicts University GPA than we would have the full equation:

$$y = \alpha + \beta x = 0 + 1 * x = x$$

Interpretting β:

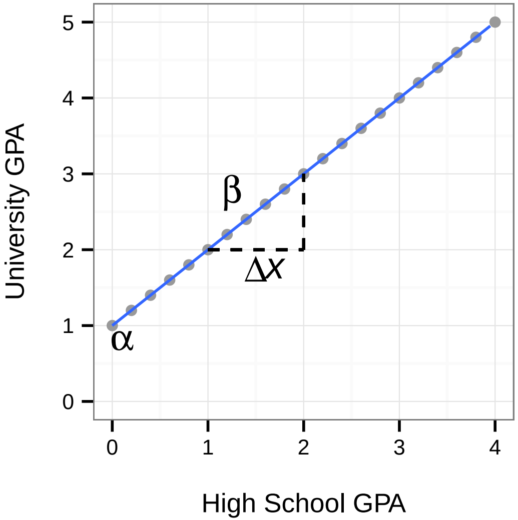
For every one unit increase in x $(+\Delta x)$ we expect β unit increase in y.

In our example, for every 1 point increase in High School GPA we expect a 1 unit increase in University GPA.

Question

What would the simple linear regression equation be if everybody's University GPA was exactly 1 point higher than their High School GPA?

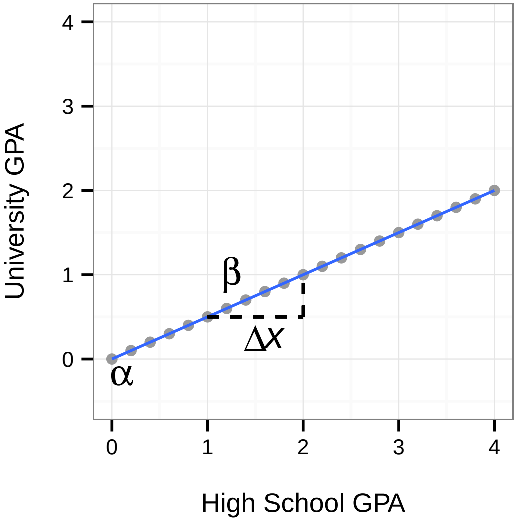
$$y = 1 + \beta x$$


University GPA
High School GPA
α
β
Δx
0
1
2
3
4
5

Question

What would the simple linear regression equation be if University GPA was half High School GPA?

$$y = 0 + 0.5(x)$$

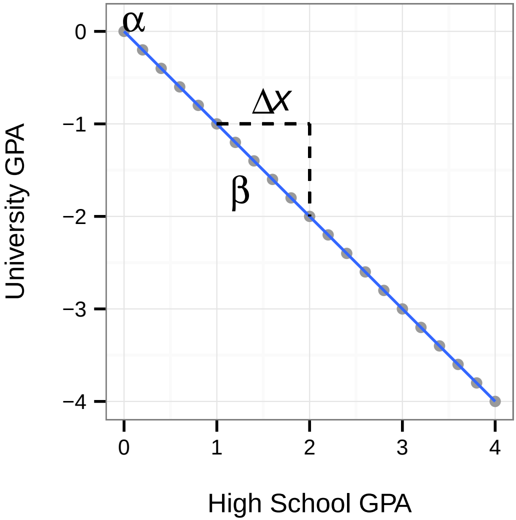


4
3
2
1
0
University GPA
β
Δx
α
0
1
2
3
4
High School GPA

Question

What would the simple linear regression equation be if University GPA was one times less than High School GPA?

$$y = 0 - 1(x)$$


α
Δx
β
0
−1
−2
−3
−4
0
1
2
3
4
University GPA
High School GPA

Negative Relationship:

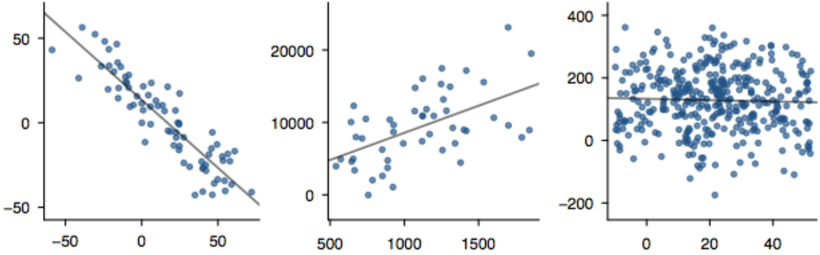
A relationship between variables x and y is negative if the regression coefficient is negative.

Positive Relationship:

A relationship between variables x and y is positive if the regression coefficient is negative.

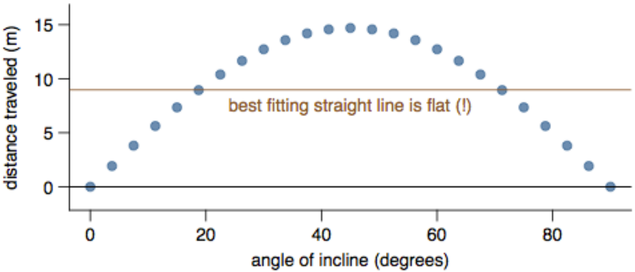
Of course, real world relationships are **rarely** perfectly linear.

More Common



Source: Diaz et. al. (2011, 216)

Be Careful of Non-Linear Relationships

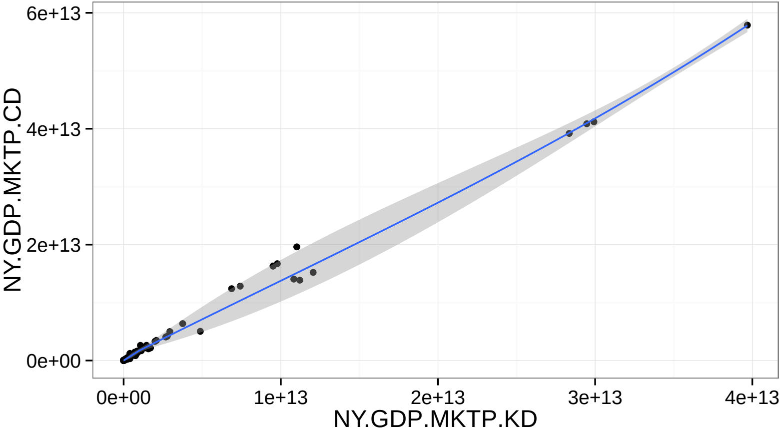


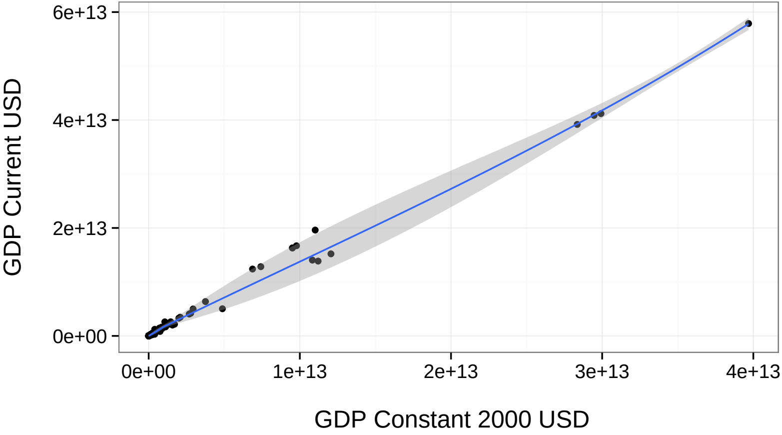
Source: Diaz et. al. (2011, 216)

Infact. . .

Infact, if you find a perfectly linear or almost perfectly linear relationship in social science research, you probably have a **problem**.

You are probably have two variables that are measuring the same thing or almost the **same thing**.


6e+13
4e+13
2e+13
0e+00
NY.GDP.MKTP.CD
0e+00
1e+13
2e+13
3e+13
4e+13
NY.GDP.MKTP.KD


GDP Current USD
6e+13
4e+13
2e+13
0e+00
0e+00
1e+13
2e+13
3e+13
4e+13
GDP Constant 2000 USD

References I

Campbell, Frances A, Craig T Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson. 2002. "Early Childhood Education: Young Adult Outcomes from the Abecedarian Project. Applied Developmental Science 6(1): 4257.

Crawley, Michael J. 2005. Statistics: An Introduction Using R. Chichester: John Wiley Sons. Ltd.

Diaz, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2011. OpenIntro Statistics. 1st ed. http://www.openintro.org/stat/downloads.php.