Week 1: Class introduction and review of linear regression

2015

Basics

- Who am I?
- Who are you?
- 146 students registered:
  - 40% distance, 20% not distance
  - 109 (75%) graduate, 34 undergraduate
- Of those that completed the assessment test (202):
  - 77 took e100
  - 90 took intro to stats more than 3 years ago
  - Average score: 43% (stdev 31%)
- Who are your teaching staff?
  - On campus: Hide, Neha
  - Online: Kela, Winnie
- Vote for sections!!!!
- Winners so far:
  - Hide: Wed 7:40-8:30pm, Tues 6:40-7:30
  - Neha: Mon 6:40-7:30
  - Kela: Mon 7:40-8:30
  - Winnie: Tues 4:10-5:00 (winner is Tues 7:40-8:30, but overlaps heavily with Kela)
- One section (probably mondays) will be recorded

General philosophy for the course

- Conceptual level – not computing by hand!
  - Learn how to read results, choose correct tests, think critically about findings
  - Not dwelling on the math….
- Flipping the classroom?
  - Read assignment before, finish quiz by 9am day of class.
  - Posted Monday mornings
    - This Monday link will appear on homepage
    - Once class list settled, will be through canvas site

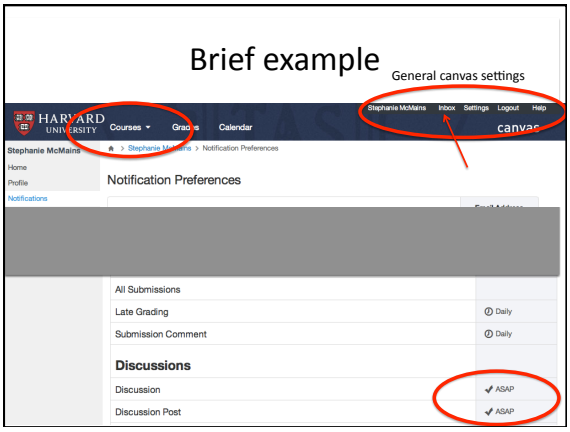
Important stuff!

- READ the SYLLABUS!!!
  - Lots of good stuff there ☺
- Use the COURSE discussion BOARD
  - Your place to post questions and get responses from teaching staff and other students.
  - Highly suggest to edit settings so get email whenever someone posts – lots of clarifications about homeworks
  - **Start a new thread when you have a new question or thought!**
- http://www.extension.harvard.edu/resources/career-academic-resource-center
  - Have lots of workshops, career advice…..

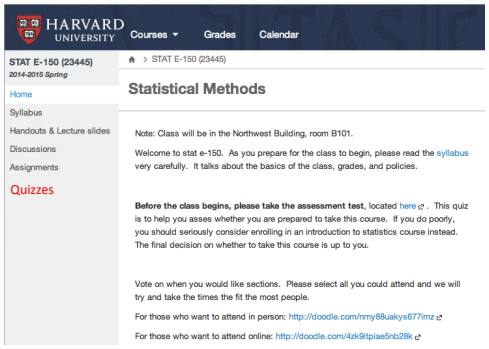
Canvas!!

- I'm new to it also.
- Having issues?
  - On course homepage there is Canvas 101 for students
    - Great videos on how to use site
    - Example: Want to learn how to set notifications when people post to the discussion board?
      - https://canvas.harvard.edu/courses/541/pages/setting-your-notification-preferences?module_item_id=2515
  - Great time to post to discussion, someone else probably confused.

Brief example



Course website



Grades!

- Undergrad:
  - average of your homework, midterm, and final grade
- Graduate:
  - Also have a final project that is worth 20% of your grade, the remaining 80 is average of your homework, midterm, and final
- Extra Credit: weekly quizzes
  - no excuses or makeups on these
- Homeworks: allowed two late submissions
  - can't be more than a week late
  - Graded with a check minus/check/check plus system
  - Only a subset of questions graded
    - a check plus does not mean you got everything perfectly
    - Look at homework solutions for exactly how to say things, to make sure you know what the best answer is.

Getting help

- Discussion board your first place!
- If you think you need more extensive help, you can make an appointment with me or a TA
- Office hours?

Online option for the first time!

- Lectures recorded and posted the next day
  - Camera shy?
- I will also post my handouts the next day
  - Since it is not live, you will have to post questions to the discussion board
- Comments on anything I can do to make recordings or online portion better are welcome
- Two sections through blackboard collaborate
  - Will also record one classroom section for posting.
- Proctored exam
  - will need to supply your own proctor that is approved by Extension school.

Now for statistics!!

Statistics: Focus on models

- Why make models of the world around us?
  - Make predictions:
    - How much money are you likely to make with your major?
  - Understand relationships:
    - How is overall happiness related to annual income?
  - Assessing differences
    - Are men or women generally more happy in life?
- Statistics allows us to say how confident we are with our predictions, or by how much something might differ
- Models are a simplification of the world!
  - For instance, many things affect the price of a used car, but there are probably a few important ones we could use to build a simple model.

Basic terminology

Can we use the number of miles that a used car has been driven to predict the price that is being asked for the car? Would it be better to base our price predictions on the age of the car in years? Or both?

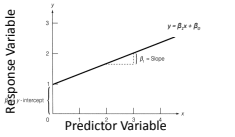
- What are the observational/experimental units?
  - Cars
- The variables?
  - Miles, age, price sold for
- Two types of variables:
  - Quantitative:
    - Price, age, miles
  - Categorical:
    - Color, manual/automatic
- Roles variables play:
  - Response (dependent): the outcome of interest
    - Price
  - Explanatory (independent): the variables whose relationship to the response is being studied, often called Predictors in regression
    - Age, miles

Basic terminology

- What if we wanted to investigate this question?
  - We collect data and fit models in order to understand populations and parameters
    - Population:
      - all used cars sold in the northeast
    - Parameters: the average age of a car that sells for more than $6,000
- We could never collect information on all cars sold…
- The collected data (say the 100 last cars sold on cars.com within 100 miles) make up our:
  - sample.
  - A characteristic of a sample, such as the average price for cars 10 years old, is called a statistic!
- Sample statistics are used to estimate population parameters

Simple linear regression

- Used to summarize the relationship between two quantitative variables:
  - 1 predictor
  - 1 response



- How is salary related to years of experience?, How does the number of accidents on your record affect the cost of car insurance? How is blood pressure related to coffee consumption?

Example

- Medical researchers have noted that adolescent females are more likely to deliver low-birthweight babies than are adult females. Because LBW babies tend to have higher mortality rates, studies have been conducted to examine the relationship between birthweight and the mother's age.
- Here is some example data consistent with the literature

Observation	1	2	3	4	5	6	7	8	9	10
Maternal Age (in years)	15	17	18	15	16	19	17	16	18	19
Birthweight (in grams)	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

Terminology review

- What are the observational units?
  - babies
- Which is the response variable?
  - Birth weight
- What kind of variable is it?
  - quantitative
- Which is the explanatory variable?
  - Maternal age
- What kind of variable is it?
  - quantitative

Observation	1	2	3	4	5	6	7	8	9	10
Maternal Age (in years)	15	17	18	15	16	19	17	16	18	19
Birthweight (in grams)	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

Can we use simple linear regression with this data?

- Do we meet the data requirements?
  - Yes! Two quantitative variables
- We will be able to use our model to answer:
  - What is the relationship between the variables?
  - What does the slope of the linear model tell us?
  - When is it appropriate to use the model to make predictions?
- How do we begin?
  - 4 step process

The Four-Step Process for statistical modeling:

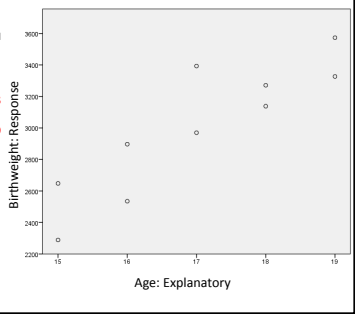
1. **Choose** a form for the model
   Identify the variables and their types:
   Examine graphs to help identify the appropriate model
2. **Fit** the model to the data
   Use the sample data to estimate the values of the model parameters
3. **Assess** how well the model fits the data
   Verify assumptions
   Examine the residuals
   Investigate significance, refine model
4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

19

Choose: Look at your data!

- In simple linear regression, you want to plot your two variables and make sure they are linearly related.
- What does the plot tell you about the strength and direction of the linear relationship?
- **The scatter diagram shows that there is a fairly strong *positive* linear relationship between the two variables**
- In context?
- **higher birthweights are associated with older mothers**



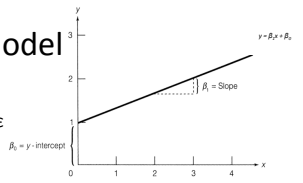
The Four-Step Process for statistical modeling:

1. **Choose** a form for the model
   Identify the variables and their types:
   Examine graphs to help identify the appropriate model
2. **Fit** the model to the data
   Use the sample data to estimate the values of the model parameters
3. **Assess** how well the model fits the data
   Verify assumptions
   Examine the residuals
   Investigate significance, refine model
4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

21

Simple linear model



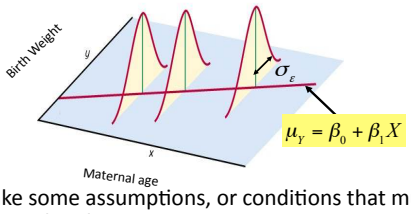
Takes the form:

$$Y=\beta_0 + \beta_1 X + \varepsilon$$

- Y= the response Variable
- X= the explanatory, predictor, or sometimes independent variable
- ε= the random error, what we don't explain from our model
- β_0= where the regression line crosses the y-axis
- β_1= the slope of the regression line
  - = change in y/change in x, or rise/run
  - = average change in y for every unit increase in x
- Conceptually, Y= the mean value of Y for a given value of X, and ε represents the error, or deviation from the mean.

Real world isn't a straight line

- It has variability, or error

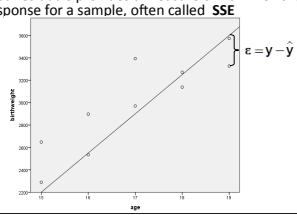


- Make some assumptions, or conditions that must be satisfied for the model to make sense. – will come back to, for now lets go on to Fit!

Fitting a simple linear model

- We will use the method of least squares to find the line the best fits the data.
  - Which provides the best estimates for β_0 and β_1.
- We can use the vertical distance between the observed value of y and the predicted value of y (y with the hat) for each value of x. This difference is called the **residual**.
  - Residual=observed - predicted
- You want the residuals to be small, and random, some positive, some negative.
- The sum of the squared residuals provides a measure of how well the line predicts the actual response for a sample, often called **SSE**

- The least squares line is where the SSE is minimized
- And as we have seen, take the form $Y=\beta_0 + \beta_1 X + \varepsilon$



SPSS output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.884[a]	.781	.754	205.30844

a. Predictors: (Constant), age

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1201970.450	1	1201970.450	28.515	.001[b]
	Residual	337212.450	8	42151.556		
	Total	1539182.900	9			

a. Dependent Variable: birthweight
b. Predictors: (Constant), age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1163.450	783.138		-1.486	.176
	age	245.150	45.908	.884	5.340	.001

a. Dependent Variable: birthweight

Birthweight^ = -1163.45 + 245.15age

The Four-Step Process for statistical modeling:

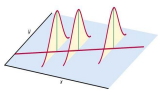
1. **Choose** a form for the model
   Identify the variables and their types:
   Examine graphs to help identify the appropriate model
2. **Fit** the model to the data
   Use the sample data to estimate the values of the model parameters
3. **Assess** how well the model fits the data
   Verify assumptions
   Examine the residuals
   Investigate significance, refine model
4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

26

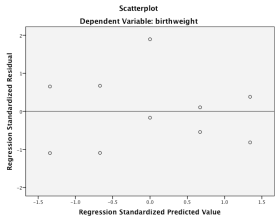
Assess the model: verify assumptions

- **Linearity** - the scatterplot shows a general linear pattern
- Conditions that deal with the distribution of ERRORS
  - **Zero Mean** - the distribution of the errors is centered at zero – always true with least squares regression!
  - **Constant Variance** - the variability of the errors is the same for all values of the predictor variable
  - **Normality** – In order to use standard distributions for confidence intervals and hypothesis tests, we often need to assume the random errors follow a normal distribution.
  - **Independence** - the errors are independent of each other
- **Random** – the data are obtained using a random process, like random sampling from a population of interest.



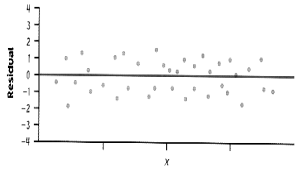
Check assumptions about errors by looking at the RESIDUALS

- A residual plot, in SPSS made by plotting standardized residuals versus standardized predicted values
- What is good?



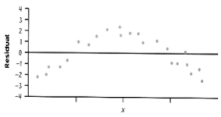
Lets familiarize ourselves with residual plots

- A uniform scatter of the points around zero
  - You don't want a systematic pattern

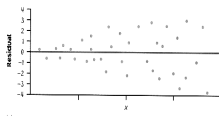


When something goes wrong…. The plot can thicken.

- A curved pattern can indicate the relationship is not linear – but you knew that, since you ALWAYS plot your data first ☺
  - What assumption violate?
  - Linearity



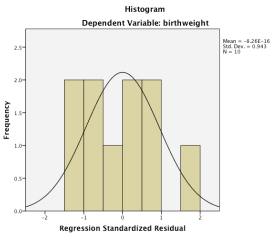
- Sometimes there is more spread for some value than others, what does this mean?
  - Predictions will be less accurate for certain values of the predictor
- What assumption does it violate?
  - Equal/constant variance



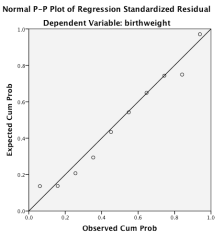
The plot thickens!!

What's left? Normality

- Can check with histograms and 'normal' probability plots



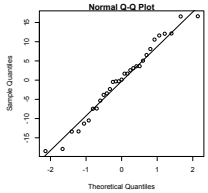
Might be more familiar to you, but can be very misleading, depends on the bins a lot.



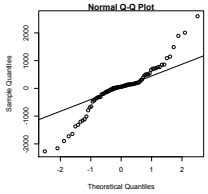
Not familiar, but once you get use to them, they are easier to interpret

Examples with more data points

- What looks good?



Good



Bad

Long tails

The Four-Step Process for statistical modeling:

1. **Choose** a form for the model
   Identify the variables and their types:
   Examine graphs to help identify the appropriate model
2. **Fit** the model to the data
   Use the sample data to estimate the values of the model parameters
3. **Assess** how well the model fits the data
   Verify assumptions
   Examine the residuals
   Investigate significance, refine model
4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

33

Use model

- The least squares line we get is:
  - Birthweight^ = -1163.45 + 245.15age
- What does the value 245.15 represent, in context?
- A baby's birthweight is expected **to increase on average** by 245.15**g** for each additional **year** in the age of the mother.
- Important aspects of answer:
  - Use variable names, not x,y
  - Use units of original problem, weight measured in grams, mothers age in years.
  - State direction of relationship (positive – increases), based on sign of slope
  - Increase is on average – since we have the nasty little ε to think about.
- What does the value -1163.45 represent, in context?
- **If the mother's age is 0 years, the child's birthweight is expected to be -1163.45 g.**
  - Doesn't make sense on any level!!! Usually doesn't. Careful about interpreting.

Interpreting model

- The least squares line we get is:
  - Birthweight^ = -1163.45 + 245.15age
- What would you expect the baby of a mother who is 16 to weigh?

  weight = – 1163.45 + 245.15 age
  = – 1163.45 + 245.15(16)
  = – 1163.45 + 3922.4
  = 2758.95
- What was the birthweight of a baby for a mother who was 16 years old?

  2897 g
- What is the residual?

  2897 – 2758.95 = 138.05g

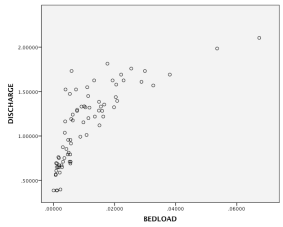
Observation	1	2	3	4	5	6	7	8	9	10
Maternal Age (in years)	15	17	18	15	16	19	17	16	18	19
Birthweight (in grams)	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

Simple linear regression: a summary of the basics!

- Data requirements: 2 quantitative variables
- Assumptions of model: linearity, constant variance, mean of zero, normality, independence of errors, and randomness of sample
  - Check these with residual plot, histogram of errors, and a normal probability plot (NPP)
- How to write regression equation
- How to interpret the betas, calculate residuals

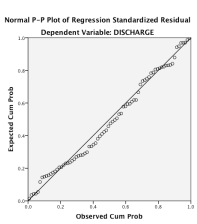
What happens when the data isn't so nice? Transformations

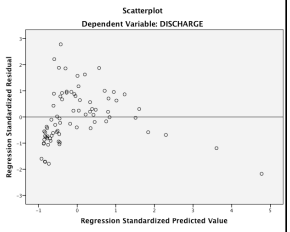
- What can we do when our data looks like this?!



Happily run our linear model…

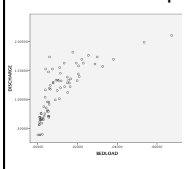
- Something doesn't look right….



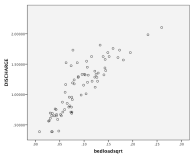


Transformation: an art

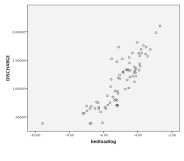
- Square root and log are the two most common, when things don't look linear, or when 'the plot thickens'



Original



Sqrt of predictor

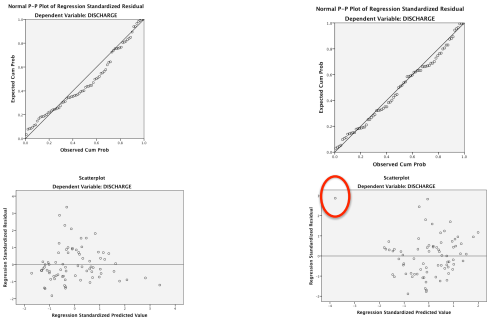


Natural log of predictor

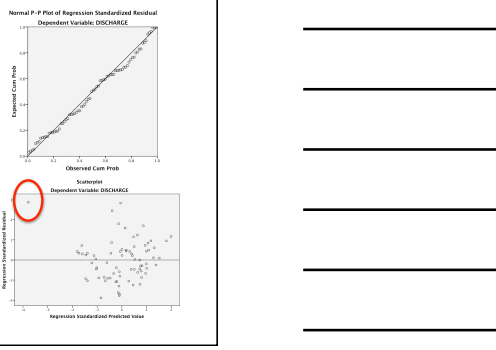
Looks better?

- Remember, nothing is perfect

Square root

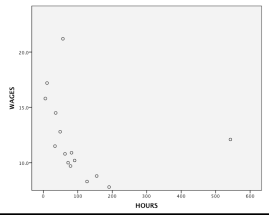


Natural log



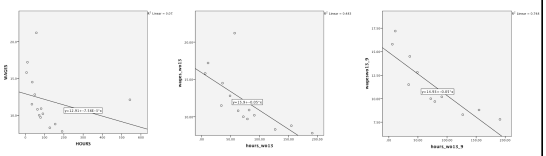
Outliers and influential points

- Outlier: doesn't fit vertically with other plots
- Influential point: Doesn't fit vertically and horizontally – can 'pull' regression line



Influential points

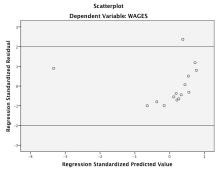
- Give equal weight to each point, sometimes one point can make a big difference
- When point has more extreme values of predictor, they can have more influence
- Will learn about more official measure later (leverage)



Wages^=12.91-.00758hours Wages^=15.9-.05hours Wages^=14.93-.05hours

Outlier: large residuals

- Standardized residual larger than 2: concerning
  - Larger than 3: very concerning
- What should you do?
  - Investigate!! Was data entered correctly?
    - In this case, they found that the person had been fired but never deleted from the system
- Can't just get rid of points we don't like……



Summary

- Know we know the basics of simple linear regression
  - the assumptions and how to check them
  - How to write the fitted regression equation
  - How to interpret the betas
- What to do if the relationship isn't linear, how to tell relationship may not be linear from residuals
- How to keep an eye out for outliers and influential points

For next week

- What you need to do by next week.
- Take assessment test if you haven't
- Vote on section times by Thursday evening
- GET SPSS!!!!
- Get the book and READ IT!
- Take quiz on reading by 9:00AM on wed. It will be posted Monday morning.
- HW1 is due by midnight Next wed. Bring questions to sections next week.