

# Exam 1 Study Guide 209

Vinny

February 2026

I will test over the the different regressions we went over. You will need to make a prediction, calculate a residual, test the assumptions by looking at residual graphs, interpret the parameters, etc...

HINT: Time was dominated by the linear model, the log-log model and models with indicator variables (for categorical predictors) so focus on those.

If I bold something in the slides it means it's important. Expect a mix of short answer, multiple choice & true/false, and questions on regressions. There will be graphs. I would advise against memorizing the formulas (see section 10 for my comments on the math you need to know) and focusing on the concepts.

I will NOT test on coding R. You will be required to see summary output of a model and will need to be able to correctly write the estimated regression equation using that output. Eg see notes on indicators for an example of this.

There are useful practice problems (and solutions to some) in the textbook (link is in the syllabus) which I advise you leverage. I would also advise checking which questions have solutions and focusing on working on them first.

## 1 Introduction

- When looking at a data set you can identify...
  1. what are the observations
  2. what are the variables
- Populations vs Samples
  - An observation is a single thing from a sample, a sample is a collection of things from the population, the population is all of the things
- Parameters vs Statistics
- Types of variables

- Numeric
  - \* Continuous
  - \* Discreet
- Categorical
  - \* Nominal
  - \* Ordinal

## 2 Data Visualization

- Why do we graph data?
- How do we choose which type of graph to use?
  - Identify appropriate graphs for different types of data
- Association vs Independence
- Describe the distribution for...
  - One categorical variable
    - \* Using a bar chart mention
      - Proportion in each category
      - Something qualitative (Eg largest category is roughly x2 as large as the next biggest category)
  - One numeric variable
    - \* Using a histogram be comfortable identifying...
      - the number of modes
      - symmetry vs skew
      - outliers
    - \* Using a boxplot be comfortable identifying...
      - Symmetry vs skew
      - Outliers
      - Understanding a boxplot can't show modes
  - Two categorical variables
    - \* Be comfortable reading the different types of bar charts
    - \* Advantages and limitations of the different options
  - Two numeric variables
    - \* Scatterplots
      1. Direction
      2. Strength
      3. Outliers
      4. Form

- Numeric and Categorical
  - \* Boxplots
    - Identify the 5 number summary using a boxplot
    - Building a boxplot from the 5 number summary
    - Discussing relative sizes of IQR between categories/boxplots

### 3 Accessibility

- What is a very common pair of colors that colorblind people struggle with?
- What are the 4 main challenges colorblindness presents
- Strategies to mitigate those problems
  - certain color pallettes do better
  - Redundant Coding
    - \* Define
    - \* Give examples of redundant coding
  - What advantage does looking at a graph you are building in grey scale give?
- Fonts
  - Don't use small fonts (ie Don't use `small fonts` )
  - Dyslexic friendly fonts
  - What are serfs in fonts? *HINT*
- The competing interests of accessibility and aesthetics
- Alt Text
  - Why does it exist?
  - Be comfortable describing a graph like you are writing the alt text

### 4 ggplot2

- Understand coding best practices (eg save early, save often)

### 5 Numeric Summaries

- Drawbacks compared to data visualizations
- Moment Statistics
  - Mean
  - St. Dev. (formula not needed)

- Assumptions (ie symmetric, no outliers in the distribution)
- Order Statistics
  - percentiles
  - 5 number summary
  - IQR ( $Q_3 - Q_1$ )
  - What “robust” to outliers means
- Relative advantages and disadvantages between moment stats and order stats
- In what situations is the mean larger/smaller than the median (heavily skewed right/left respectively, or outliers in those directions)
- When to choose one over the other
- What do I mean when I talk about “processed” data?
- Conditional statistics
  - It’s just calculating a statistic(s) for a particular group
  - eg we find the distribution of income for only women

## 6 Correlation

- What does Pearson’s correlation coefficient,  $r$ , measure?
  - What’s the range of  $r$ ?
  - What’s the units of  $r$ ?
  - How, if at all, does  $r$  change if we convert our measurement units (eg mpg to kpl)
  - Does a large  $r$  imply a linear fit?
  - Type of data  $r$  can be used with?
- What is Spearman’s correlation?
  - What type of data can Spearman’s cor. be applied to?
  - When do we choose Spearman’s over Pearson’s?
  - What does a “monotonic” relationship mean
- Rank graphs by how strong their (Pearson) correlations are
- Understand that correlation can be either positive negative or zero, and what does having a negative correlation mean? 0 correlation?
- Ecological Fallacy (group aggregated statistics don’t apply on the individual level)
- Correlation vs Causation
- Lurking variable
  - Be prepared to give an example

## 7 Simple Linear Regression

- What are the two main goals of simple linear regression?
- Be comfortable talking about what an explanatory variable and response variable are
- Difference between the regression equation and an estimated regression equation
- Difference between  $\hat{y}$  and  $y$ ; similar between  $\hat{\beta}_1$  and  $\beta_1$  and again  $\hat{\beta}_0$  and  $\beta_0$
- Use a best-fit-line to make a prediction
- Use said prediction and the actual value to calculate the residual
- Interpret your coefficients
- 4 Assumptions. Be comfortable checking the ones you can with a residual vs predicted plot
  - Normality (roughly equal spread of points above and below the 0 line, no patterns)
  - X, Y's relationship is correctly identified (as a linear model)
  - Independent and Identically Distributed
  - Be prepared to give examples of when one assumption goes wrong
- Extrapolation and why that does not bring joy
- $R^2$ 
  - Interpretation
  - Relationship to Pearson's correlation coefficient  $r$
  - Appreciation for the fact it boils down the entire scatterplot/linear regression to a single number

## 8 Regression with Categorical Predictors

- What's a good (best?) value for guessing a prediction for a group? (hint: it's the mean)
- Indicator variables
  - What are they?
  - Why do we use them?
  - How to make them
- Make a prediction using a model with a categorical predictor, and calculate a residual
- Interpretations of it's parameters (HINT: I'll only be using the format with  $\beta$ 's, the same as R)
  - What's a baseline category? (The default category associated with  $\beta_0$ , and the category the other categories are compared to when estimating  $\beta_1, \beta_2$ , etc..)

## 9 Multiple Linear Regression (MLR)

- Explain why we use MLR over simple linear regression
  - SLR only allows for one explanatory variable which is too limiting
  - MLR allows us to better understand the system as a whole since we can account for differing effects
- How do we interpret the coefficient of a quantitative (numeric) explanatory variable?
- How do we interpret the coefficient of an indicator (eg a nominal/categorical explanatory variable)?
- Assumptions
  - Our model is appropriate (linear model)
  - Residuals are normally distributed with mean 0
  - IID
    - \* There is where homoskedasticity assumption is (need same spread to be identically distributed)
  - NOTE: Both IID and normality assumption reference the residuals!
  - NOTE: If an assumption violation occurs we usually violate multiple assumptions, in my experience
  - Justify/explain why we color residuals by explanatory variables

## 10 Transformations

- Why do we transform our data? What is the goal?
- Name a popular transformation
- Log() transformation and relationship to heteroskedasticity
- Identify whether two variables' relationship might benefit from a log transformation.
  - By looking at a scatterplot similar to the in-class notes
  - By looking at a lot of large outliers in both the x and y directions
  - By looking at a residual by predicted plot and seeing the spread of the data “balloon” out on the right hand side
- Back-transforming the log transformation
- Make a prediction using a log-log model

- Need to put the prediction on the linear scale!!
- Drawback of using a transformed model

## 11 Math You'll Need to be Comfortable With

- Taking a log of a number and back transforming it (eg  $e^{\log(x)}$ )
- Using a linear model equation to make a prediction
- Given an actual observation and a prediction find a residual
- Calculating a five number summary
  - Min, Max are easy
  - Median is the middle number
  - $Q_1$  and  $Q_3$  are the medians of the lower and upper halves of the data, respectively
- Calculating IQR given  $Q_1$  and  $Q_3$
- Going from  $r$  to  $R^2$  and back again
  - Be careful on if we need a positive or negative square root
  - Decision is based on the direction of the graph