Chapter 1

Introduction

04/18/2021

Review of Multiple Linear Regression

Learning Objectives: After finishing this chapter, you should be able to:

- Identify cases where linear least squares regression (LLSR) assumptions are violated.
- Generate exploratory data analysis (EDA) plots and summary statistics.
- Use residual diagnostics to examine LLSR assumptions.
- Interpret parameters and associated tests and intervals from multiple regression models.
- Understand the basic ideas behind bootstrapped confidence intervals.

Introduction to Beyond Multiple Linear Regression - Examples

- Ecologists count species, criminologists count arrests
- Cancer specialists count cases
- Political scientists seek to explain who is a Democrat
- Pre-med students are curious about who gets into medical school
- Sociologists study which people get tattoos.

Which are counts? What type of responses are the others?

We can model these non-Gaussian (non-normal) responses in a more natural way by fitting **generalized linear models (GLMs)** as opposed to using **linear least squares regression (LLSR)** models.

Linear Least Squares Regression (LLSR)

When models are fit to data using linear least squares regression (LLSR), inferences are possible using traditional statistical theory under certain conditions/assumptions:

- A linear relationship between the response (Y) and an explanatory variable (X)
- Observations are independent of one another
- Responses are approximately normal for each level of the X, and
- The variation in the responses is the same for each level of X.

Generalized Linear Models (GLMs)

If we intend to make inferences using GLMs, necessary assumptions are different.

- We will not be constrained by the normality assumption.
- When conditions are met, GLMs can accommodate non-normal responses such as the counts and binary data in our preceding examples
- While the observations must still be independent of one another, the variance in Y at each level of X need not be equal
- The assumption of linearity between Y and X need to be plausible.

GLMs don't Always Fit

GLMs cannot be used for models in the following circumstances:

- Medical researchers collect data on patients in clinical trials weekly for 6 months
  - Why not?
- Rat dams are injected with teratogenic substances and their offspring are monitored for defects
  - Why not?
- Musicians' performance anxiety is recorded for several performances.
  - Why not?

Why not?

Each of these examples involves correlated data:

- The same patient's outcomes are more likely to be similar from week-to-week than outcomes from different patients
- Litter mates are more likely to suffer defects at similar rates in contrast to unrelated rat pups
- A musician's anxiety is more similar from performance to performance than it is with other musicians.

Each of these examples violate the independence assumption of simpler linear models for LLSR or GLM inference.

Extending our Knowledge

- The **Generalized Linear Models** in the book's title extends least squares methods you may have seen in linear regression to handle responses that are non-normal.
- The **Multilevel Models** in the book's title will allow us to create models for situations where the observations are not independent of one another.
- These approaches will permit us to get much more out of data and may be more faithful to the actual data structure than models based on ordinary least squares. These models will allow you to expand *beyond multiple linear regression*.

Reconizing Violations of Assumptions

- We first need to be able to recognize the violation of model assumptions for inference with LLSR in the context of different studies.
- Linearity is sufficient for fitting an LLSR model, but in order to make inferences and predictions we need:
  - Observations must also be independent
  - Responses should be approximately normal at each level of the predictors
  - Standard deviation of the responses at each level of the predictors should be approximately equal
- We will be starting with a review of evaluating model assumptions

Assumptions for Linear Least Squares Regression

- We want to visualize our normality assumption
- Lets do this together in R

Plot

- What conditions can we use this plot to understand?

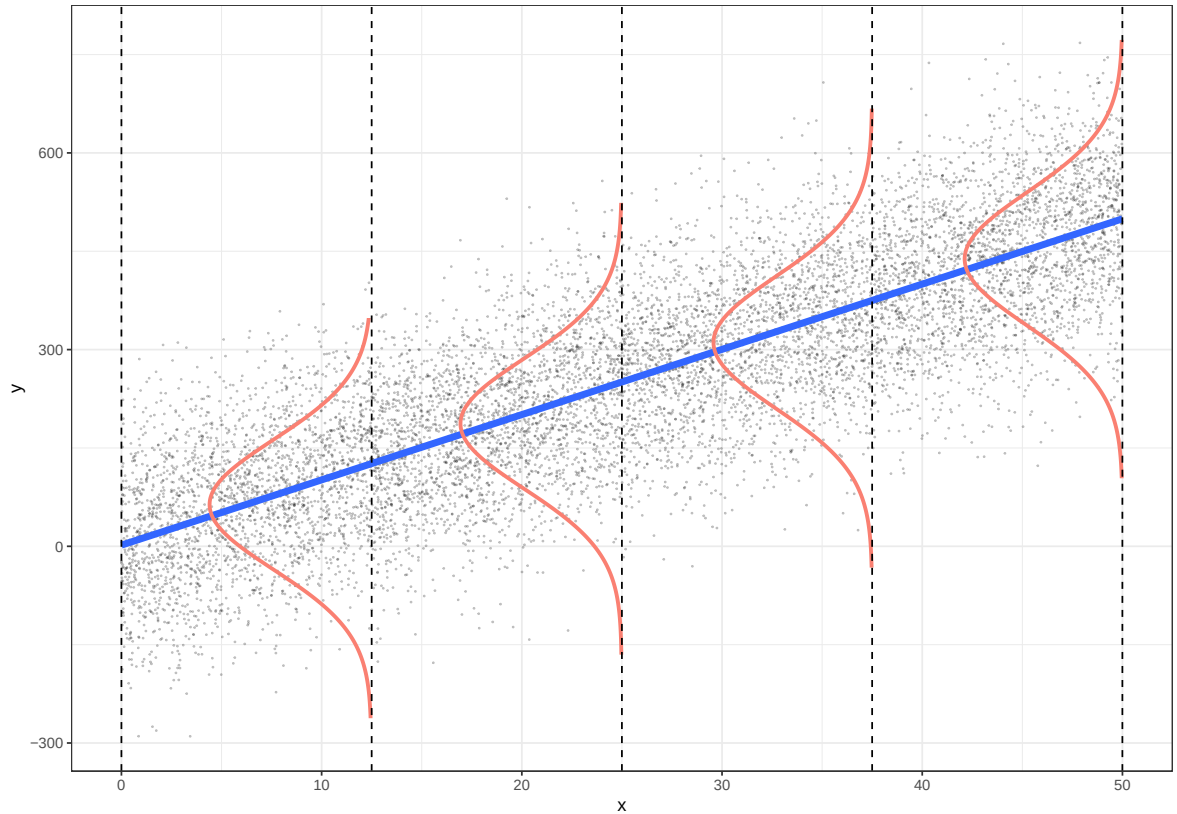

Figure 1: Assumptions for linear least squares regression (LLSR).

More on Assumptions

- The acronym LINE can be used to recall the assumptions required for making inferences and predictions with models based on simple linear regression
  - **L:** There is a linear relationship between the mean response (Y) and the explanatory variable (X),
  - **I:** The errors are independent—there's no connection between how far any two points lie from the regression line,
  - **N:** The responses are normally distributed at each level of X, and
  - **E:** The variance or, equivalently, the standard deviation of the responses is equal for all levels of X.
- These assumptions are depicted in our previous graph.

Actionable Interpretation so We can Check Assumptions

- **L:** The mean value for Y at each level of X falls on the regression line.
- **I:** We'll need to check the design of the study to determine if the errors (vertical distances from the line) are independent of one another.
- **N:** At each level of X, the values for Y are normally distributed.
- **E:** The spread in the Y's for each level of X is the same.

Checking Assumptions - Example 1

Reaction times and car radios. A researcher suspects that loud music can affect how quickly drivers react. She randomly selects drivers to drive the same stretch of road with varying levels of music volume. Stopping distances for each driver are measured along with the decibel level of the music on their car radio.

What are the response and explanatory variables?

Checking Assumptions - Example 1

- *Response variable:* Reaction time
- *Explanatory variable:* Decibel level of music

Checking Assumptions - Example 1 - Together

- Interpret the LINE conditions in our terms of our example
- Next determine if these conditions are plausibly satisfied.

Checking Assumptions - Example 1

The assumptions for inference in LLSR would apply if:

- **L:** The mean reaction time is linearly related to decibel level of the music.
- **I:** Stopping distances are independent. The random selection of drivers should assure independence.
- **N:** The stopping distances for a given decibel level of music vary and are normally distributed.
- **E:** The variation in stopping distances should be approximately the same for each decibel level of music.

Checking Assumptions - Example 1

- Potential problems with the linearity and equal standard deviation assumptions.
- If there is a threshold for the volume of music where the effect on reaction times remains the same, mean reaction times would not be a linear function of music.
- Another problem may occur if a few subjects at each decibel level took a really long time to react. In this case, reaction times would be right skewed and the normality assumption would be violated.
- We want to think of circumstances where the LLSR assumptions may be suspect.
- Later in this chapter we will describe plots which can help diagnose issues with LLSR assumptions.

Checking Assumptions - Examples 2 and 3

Crop yield and rainfall. The yield of wheat per acre for the month of July is thought to be related to the rainfall. A researcher randomly selects acres of wheat and records the rainfall and bushels of wheat per acre.

Heights of sons and fathers. Sir Francis Galton suspected that a son's height could be predicted using the father's height. He collected observations on heights of fathers and their firstborn sons [@Stigler2002].

Checking Assumptions - Examples 2 and 3

With another person, understand the examples and check assumptions:

- Determine the variables
- Interpret the LINE conditions in our terms of our example
- Next determine if these conditions are plausibly satisfied.
- Should we use LLSR?
- Make sure to specify components of the study that are suspect for violating assumptions and those that are in place to assure assumptions are met.

Checking Assumptions - Example 2

- *Response variable:* Yield of wheat measured in bushels per acre for July
- *Explanatory variable:* Rainfall measured in inches for July
- **L:** The mean yield per acre is linearly related to rainfall.
- **I:** Field yields are independent; knowing one (X, Y) pair does not provide information about another.
- **N:** The yields for a given amount of rainfall are normally distributed.
- **E:** The standard deviation of yields is approximately the same for each rainfall level.

Checking Assumptions - Example 2

- Possible problems with the linearity assumption if mean yields increase initially as the amount of rainfall increases after which excess rainfall begins to ruin crop yield.
- The random selection of fields should assure independence if fields are not close to one another.

Checking Assumptions - Example 3

- *Response variable:* Height of the firstborn son
- *Explanatory variable:* Height of the father
- **L:** The mean height of firstborn sons is linearly related to heights of fathers.
- **I:** The height of one firstborn son is independent of the heights of other firstborn sons in the study. This would be the case if firstborn sons were randomly selected.
- **N:** The heights of firstborn sons for a given father's height are normally distributed.
- **E:** The standard deviation of firstborn sons' heights at a given father's height is the same.

Checking Assumptions - Example 3

- Heights and other similar measurements are often normally distributed. There would be a problem with the independence assumption if multiple sons from the same family were selected.
- Or, there would be a problem with equal variance if sons of tall fathers had much more variety in their heights than sons of shorter fathers.

Cases With Assumption Violations - Example 4

Grades and studying. Is the time spent studying predictive of success on an exam? The time spent studying for an exam, in hours, and success, measured as Pass or Fail, are recorded for randomly selected students.

What are our variables?

Cases With Assumption Violations - Example 4

Grades and studying. Is the time spent studying predictive of success on an exam? The time spent studying for an exam, in hours, and success, measured as Pass or Fail, are recorded for randomly selected students.

- *Response variable:* Exam outcome (Pass or Fail)
- *Explanatory variable:* Time spent studying (in hours)
- Response is a binary outcome which violates the assumption of a normally distributed response at each level of X.
- What type of regression have we used for this type of response? WE will see it again in Chapter 6.

Cases With Assumption Violations - Example 5

Income and family size. Do wealthy families tend to have fewer children compared to lower income families? Annual income and family size are recorded for a random sample of families.

What are our variables?

Cases With Assumption Violations - Example 5

Income and family size. Do wealthy families tend to have fewer children compared to lower income families? Annual income and family size are recorded for a random sample of families.

- *Response variable:* Family size, number of children
- *Explanatory variable:* Annual income, in dollars
- Family size is a count taking on integer values from 0 to (technically) no upper bound.
- What assumptions could this violate?

Cases With Assumption Violations - Example 5

- The normality assumption may be problematic again because the distribution of family size is likely to be skewed, with more families having one or two children and only a few with a much larger number of children.
- Both of these concerns, along with the discrete nature of the response, lead us to question the validity of the normality assumption.
- In fact, we might consider Poisson models discussed in Chapter 4.
- Study design should also specify that families are done adding children to their family.

Cases With Assumption Violations - Example 6

Exercise, weight, and sex. Investigators collected the weight, sex, and amount of exercise for a random sample of college students.

What are our variables?

Cases With Assumption Violations - Example 6

Exercise, weight, and sex. Investigators collected the weight, sex, and amount of exercise for a random sample of college students.

- *Response variable:* Weight
- *Explanatory variables:* Sex and hours spent exercising in a typical week
- Now we are moving toward multiple linear regression, how do we check the linearity assumption?
- What assumptions are very likely to not be satisfied?

Cases With Assumption Violations - Example 6

- The standard deviation in weight for students who do not exercise for each sex is likely to be considerably greater than the standard deviation in weight for students who follow an exercise regime.
  - We can assess this potential problem by plotting weight by amount of exercise for males and females separately.
- Potential problems with the independence assumption because there is no indication that the subjects were randomly selected.

Cases With Assumption Violations - Example 7

Surgery outcome and patient age. Medical researchers investigated the outcome of a particular surgery for patients with comparable stages of disease but different ages. The ten hospitals in the study had at least two surgeons performing the surgery of interest. Patients were randomly selected for each surgeon at each hospital. The surgery outcome was recorded on a scale of 1-10.

What are our variables?

Cases With Assumption Violations - Example 7

Medical researchers investigated the outcome of a particular surgery for patients with comparable stages of disease but different ages. The ten hospitals in the study had at least two surgeons performing the surgery of interest. Patients were randomly selected for each surgeon at each hospital. The surgery outcome was recorded on a scale of 1-10.

- *Response variable:* Surgery outcome, scale 1-10
- *Explanatory variable:* Patient age, in years
- Potential Violations?

Cases With Assumption Violations - Example 7

- Outcomes for patients operated on by the same surgeon are more likely to be similar and have similar results.
  - Patient outcomes could be dependent on surgeon. Why?
- Outcomes at one hospital may be more similar. Why?
- The structure of this data suggests that the independence assumption will be violated.
- A Multilevel models, which we begin discussing in Chapter 8, can explicitly take this structure into account for a proper analysis of this study's results.

Looking forward. . .

- We reviewed possible violations of assumptions for inference in LLSR
- There may be violations of the other assumptions that we have not pointed out
- You have learned some ways to handle these violations such as applying variance stabilizing transformations or logging responses
- Other models in this text that may be more appropriate for the violations we have presented.

Case Study: Kentucky Derby

The Kentucky Derby is a 1.25-mile horse race held annually at the Churchill Downs race track in Louisville, Kentucky. Our data set `derbyplus.csv` contains the `year` of the race, the winning horse (`winner`), the `condition` of the track, the average `speed` (in feet per second) of the winner, and the number of `starters` (field size, or horses who raced) for the years 1896-2017. The track `condition` has been grouped into three categories: fast, good (which includes the official designations "good" and "dusty"), and slow (which includes the designations "slow", "heavy", "muddy", and "sloppy"). We would like to use least squares linear regression techniques to model the speed of the winning horse as a function of track condition, field size, and trends over time.

Initial Exploratory Analyses - Data Organizaiton

We can view the first few lines of data using head(dataframename) in R. We also need to creat new newvariables:

- `fast` is an **indicator variable**, taking the value 1 for races run on fast tracks, and 0 for races run under other conditions,
- `good` is another indicator variable, taking the value 1 for races run under good conditions, and 0 for races run under other conditions,
- `yearnew` is a **centered variable**, where we measure the number of years since 1896, and
- `fastfactor` replaces `fast` $= 0$ with the description "not fast", and `fast` $= 1$ with the description "fast". Changing a numeric categorical variable to descriptive phrases can make plot legends more meaningful.

Lets do this in R!

Univariate Summaries

- With any statistical analysis, our first task is to explore the data, examining distributions of individual responses and predictors using graphical and numerical summaries, and beginning to discover relationships between variables.
- What is the name for this part of statistics?
- This should *always* be done *before* any model fitting! We must understand our data thoroughly before doing anything else.
- Examine the response variable and each potential covariate individually. - Continuous variables can be summarized using histograms and statistics indicating center and spread;
  - Categorical variables can be summarized with tables and possibly bar charts.
  - Lets do this in R!

Reading our Graphs

- What are we looking for? What are we investigating with each of our plots?

Reading our Graphs

- We see that the primary response, winning speed, follows a distribution with a slight left skew, with a large number of horses winning with speeds between 53-55 feet per second.
- Plot (b) shows that the number of starters is mainly distributed between 5 and 20, with the largest number of races having between 15 and 20 starters.
- The primary categorical explanatory variable is track condition, where 88 (72%) of the 122 races were run under fast conditions, 10 (8%) under good conditions, and 24 (20%) under slow conditions.

Bivariate Summaries

- Next we need to examine numerical and graphical summaries of relationships between model covariates and responses. Why?
- Lets create these in R!

Bivariate Summaries

- We see densely packed illustrations of bivariate relationships.
- The relationship between two continuous variables is depicted with scatterplots below the diagonal and correlation coefficients above the diagonal.
- How are higher winning speeds associated with time?
- What about winning speed and number of starters?
- Year and number of starters?
- What must we remember to consider, when thinking about using the data for linear regression, about the relationship between predictors?
- How could we improve this plot?

Bivariate Summaries

- Relationships between categorical variables like track condition and continuous variables can be illustrated with side-by-side boxplots
- How do speed and condition compare? (Note the plots goes fast,good,slow from left to right)
- How do year and conditions compare?
- What is another way we can compare and understand variables?
- Lets do this in R!

Bivariate Summaries

- We can calculate summary statistics generated by subgroup.
- The mean speed under fast conditions is 53.6 feet per second
- The mean speed under good conditions is 52.7 ft/s
- The mean speed is 51.7 ft/s under slow conditions.
- Variability in winning speeds, however, is greatest under slow conditions (SD = 1.36 ft/s) and least under fast conditions (0.94 ft/s).

Bivariate Summaries

- Back to our collection of plots
- The diagonal illustrates the distribution of individual variables, using density curves for continuous variables and a bar chart for categorical variables.
- Trends observed in the last two diagonal entries match trends observed in our earlier plots.
- The year density plot is not meaningful.

Bivariate Plot Upgrades!

- By using shape or color or other attributes, we can incorporate the effect of a third or even fourth variable into the scatterplots
- Lets do this in R!

Bivariate Plot Upgrades!

- Now we can see that speeds are generally faster under fast conditions, but the rate of increasing speed over time is greater under good or slow conditions.
- Note that graphical analysis is exploratory, and any notable trends at this stage should be checked through formal modeling.

Asking Questions based on our Analylsis

- are winning speeds increasing in a linear fashion?
- does the rate of increase in winning speed depend on track condition or number of starters?
- are any of these associations statistically significant?

Asking Questions based on our Analylsis

- after accounting for other explanatory variables, is greater field size (number of starters) associated with faster winning speeds (because more horses in the field means a greater chance one horse will run a very fast time) or slower winning speeds (because horses are more likely to bump into each other or crowd each others' attempts to run at full gait)?
- how well can we predict the winning speed in the Kentucky Derby?