MA478 - Lesson1

Clark

Contact Information

COL Nick Clark

nicholas.clark@westpoint.edu

Thayer Hall Room 229

Admin Stuff

-Course Text

-R/Rstudio

-Course Overview

- This course introduces modeling beyond that gained in MA376. You will learn statistical models for analyzing quantitative and qualitative data.
- Methods will generally be taught in the generalized linear model framework and may include binomial and multinomial regression, count regression, robust regression, and panel regression.
- You will also be exposed to techniques for handling problems that arise when analyzing real data, such as missing data, outliers, and influential observations.
- You will focus on understanding data, implementing advanced regression modeling techniques, and developing intuition from analyzed data.
- You should understand when certain statistical methods are appropriate, how to use them using the R statistical programming language, and how to visually represent and explain results.
- The course also addresses issues of exploratory data analysis, data preparation, model development, model selection, and model validation.

-Course Objectives

- Understand generalized linear models and other advanced regression methods.
- Be able to independently learn and appropriately apply advanced statistical techniques.
- Be able to critically read and interpret data science literature that applies advanced statistical techniques and understand ethical issues in statistical modeling.
- Be able to successfully use R statistical programming software to analyze data.
- When faced with a real-world problem, be able to select and ethically execute appropriate statistical modeling techniques to gain insight into the problem to help solve it.

-Graded Events

-My Expectations of you

- Complete readings before class
- Complete assignments by their assigned due dates
  - Late work is NOT accepted unless the student has made alternate arrangements ahead of time
- Bring your course textbooks and laptop to each class
  - Stay off of your laptop unless using it for class
- Participate in class

-What you can expect from me

- Publish lesson objectives prior to each lesson
- Provide suggested problems for each lesson
- Lesson sheets to guide discussion
- Create interactive classroom environment
- Timeline in grading assignments
- Available for AI
- High energy!

Introductions

Recall in previous classes we discussed some different regression models. For instance, we had a linear regression model that we formulated like:

Typically we would use a linear regression model when we wanted to explain a continuous random variable and we met a variety of conditions that are all contained in the ϵ_i term.

In MA376, we talked about another regression model, the *logistic* regression model that we formulated like:

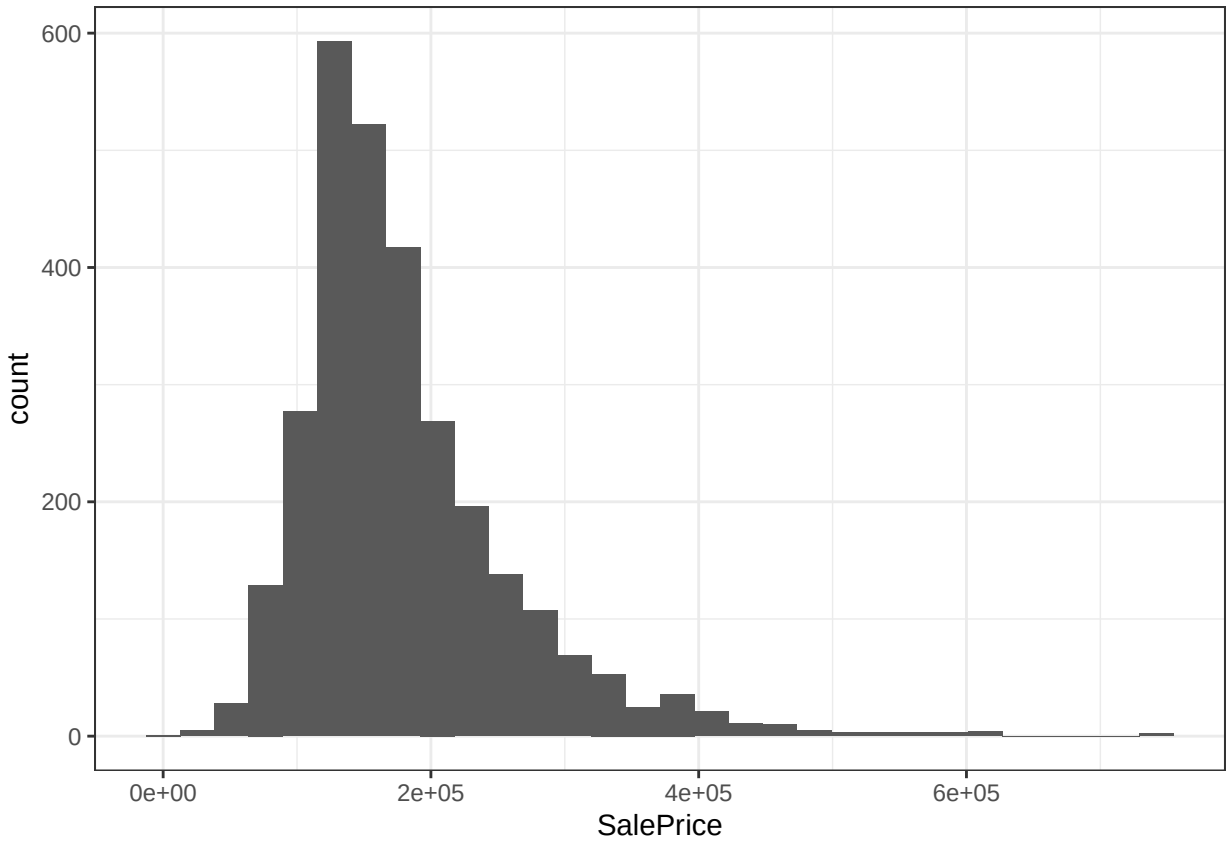
In both these cases we typically started with our response variable and worked backwards to figure out what type of model to build. However, as situations that we are modeling become more and more complex, thinking about building a model all in one-swoop is not ideal. The advent of *generalized linear models (GLMs)* allow us to build a model by thinking of three sepearte components.

The first is the **random component**. The random component specifies the conditional distribution of the response variable. Sometimes this is easy, if our data are binary than clearly the distribution of our data are:

However, we could have other situations that we might be interested in modeling. For instance, perhaps we want to model the number of people that die of COVID in a given county. In this case, maybe we want the distribuiton of our data to be:

Or, perhaps we want to model the cost of housing in Ames, Iowa.

```
#install.packages(AmesHousing)
library(AmesHousing)
library(tidyverse)
ames_raw %>%
  ggplot(aes(x=SalePrice)) +
  geom_histogram() +
  theme_bw()
```



In this case, perhaps our data come from a *gamma distribution* or a *inverse-Gaussian distribution.*

The second part of a GLM is the systematic component, or the linear predictor. This should look familiar to us, but here we are building out a linear combination of our explanatory variables that relates not directly to our a combination of our parameters.

Here, note, that we aren't really thinking of the form of our response variable, rather we are thinking about what covariates would we expect to impact the expected value of our response variable.

For instance, with our Ames Housing data, perhaps we want to use x_1 as square footage, and x_2 as central air conditioning. Then, regardless of our choice of distribution of y we can write:

Where η is some function of the expected value of y.

This takes us to the third part of a GLM, the **link function**. The link function connects the random component, or more precisely the expectation of the random component, with the linear predictor. The only requirement of a link function is that it is monotonic, and differentiable.

Let's put this all together with a familiar example. Let's say that we want to model whether a Cadet passes or fails the IOCT and we want to use height and sex as predictors.

First, we come up with the random component:

Next we build out our linear predictor:

Finally, we come up with a way to link our linear predictor with the expected value of our random component:

Even more familiar. Let's say we want to model ACFT scores using height and sex as predictors. If we assume $y_i \sim N(\mu_i, \sigma)$ let's put this model into the GLM framework.

Next lesson:

- Understand how least squares model fitting applies to estimating the linear model
- Understand error variance, sum-of-squares decomposition, R-squared, and residuals in a linear model, their purpose, and applications
- Apply R statistical programming to estimate, evaluate, and interpret linear models

Chapter 2.1-2.4 of Agresti. Notation is dense, but this is review of MA376.

Pages 1-14 of Faraway.

Install the faraway library in R. Look at the mtcars dataset and explore the relationship between mpg and engine type.