Linear Regression (part 1)

Ryan Miller

Statistical Models

- In discussing ANOVA, we introduced the concept of **statistical models**, which are simplified representations of the world that involve a probability distribution
- Recall the one-way ANOVA model:

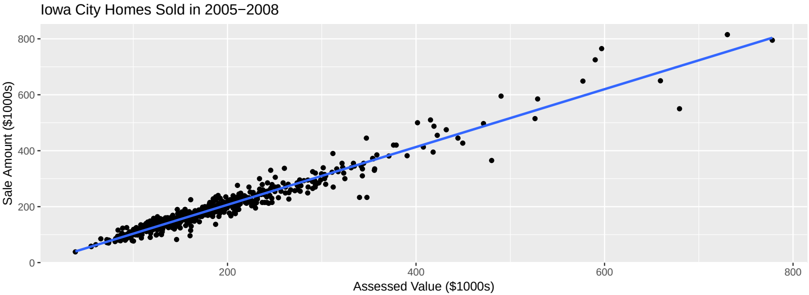
$$y_i = \mu_i + \epsilon_i$$

- Where y_i is the outcome measurement for the i^{th} case
- μ_i is the population mean of the group that the i^{th} case belongs to
- ϵ_i was an unexplained deviation for the i^{th} case
  - These deviations follow a normal distribution with a mean of zero, thereby making this a statistical model

Simple Linear Regression

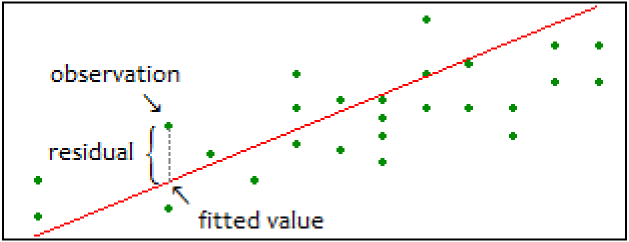
- **Simple linear regression** is another example of a statistical model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$



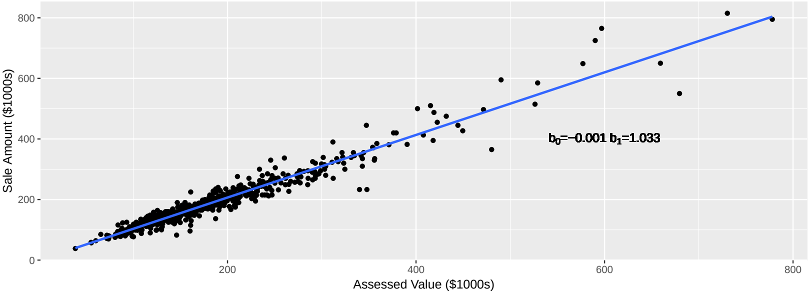
Simple Linear Regression

- To utilize the simple linear regression model, the **coefficients** (the intercept β_0 and the slope β_1) must be estimated from the data
  - This is done via **least squares estimation**, a method which *minimizes* the squared residuals:



Simple Linear Regression

- We use the notation b_0 and b_1 to denote our estimates of the model *parameters* $\{\beta_0, \beta_1\}$
  - These estimates (b_0 and b_1) describe how the x and y variables are related *in our data*
  - In the example below, what does b_0 tell you? What does b_1 tell you?



Uncertainty and Statistical Inference

- Like any estimate, the regression estimates, b_0, b_1, won't *exactly* match the population parameters, β_0, β_1
- We won't go too far into the details, but most standard software will provide confidence interval estimates for the population parameters using the t-distribution
  - We can also use the regression estimates, b_0, b_1, to perform hypothesis testing using the t-distribution
  - By default, software (Minitab included) will automatically test the null hypotheses $\beta_0 = 0$ and $\beta_1 = 0$ whenever you fit a regression model

Statistical Inference - Example

1. Load the "Professor Salaries" dataset into Minitab and fit a simple linear regression model that predicts Salary (response variable) based upon Years of Service (continuous predictor) using the "Stat -> Regression -> Fit Regression" menus
2. Think about practical meaning of the null hypotheses $\beta_0 = 0$ and $\beta_1 = 0$
3. Using the coefficients table, interpret the p-values provided in the rows labeled "Constant" and "yrs.service"

Statistical Inference - Example (solution)

- The row labeled "constant" tests whether $H_0 : \beta_0 = 0$, which corresponds to professors with zero years of experience
  - There is overwhelming evidence from these data that the salaries of newly hired professors are not zero
- The row labeled "yrs.service" tests whether $H_0 : \beta_1 = 0$, which corresponds to *no linear association* between years of service and salary
  - There is overwhelming evidence that the salary is linearly associated with years of service. More specifically, each 1 year of additional service predicts a 780 increase in annual salary.

Statistical Inference - Example

1. Continuing with the Professor Salaries dataset, click on the "Results" button in the "Fit Regression" menu and select "Expanded Tables" to add 95% confidence intervals to the coefficient table
2. Provide an interpretation of the 95% confidence interval for β_1 from the model you fit earlier

Statistical Inference - Example (solution)

- The 95% CI is (562, 997), indicating we can be 95% confident that each one-year increase in experience corresponds with an increase in annual salary between 562 and 997 in the population represented by these data (and according to this model)

Multiple Regression

- Simple linear regression is actually a special case of a broader method known as **multiple regression**
  - The one-way ANOVA model, which we'll revisit in a few moments, also happens to be a generalization of multiple regression
- Multiple regression models take the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Dummy Variables

- To connect multiple regression and one-way ANOVA, we need to introduce **dummy variables**, which are a way of representing a categorical variable using one or more binary variables (taking on values of 0 or 1)
- For a predictor with two categories, we assign one category to be the **reference category**, with a numeric value of 0
  - Data-points in the non-reference category receive a value of 1

Y	group
8.5	B
11.6	A
9.0	B
9.1	B
8.0	A
9.7	A

Y	dummy
8.5	1
11.6	0
9.0	1
9.1	1
8.0	0
9.7	0

Dummy Variables

- For a categorical predictor with k categories, $k - 1$ different dummy variables are necessary

Y	group
8.5	C
11.6	B
9.0	C
9.1	B
8.0	A
9.7	A

Y	dummy1	dummy2
8.5	0	1
11.6	1	0
9.0	0	1
9.1	1	0
8.0	0	0
9.7	0	0

Dummy Variables - Example

1. Load the "Tailgating" dataset into Minitab and display descriptive statistics showing the mean following distances, D, by Drug group
2. Fit a regression model that predicts following distance, D, based upon Drug (categorical predictor) using the "Stat -> Regression -> Fit Regression" menus
3. Which group did Minitab choose as the reference category in this model? Do you notice the sample mean of this group anywhere in the model?
4. How do you interpret the coefficient estimates of this model?

Dummy Variables - Solution

The *estimated* model is:

$$\hat{Y} = b_0 + b_1 X_{MDMA} + b_2 X_{NODRUG} + b_3 X_{THC}$$

- "Alcohol" was used as the reference category
- $b_0 = 36.83$ is the sample mean of the alcohol group, this isn't a coincidence
- $b_1 = -9.2$ is the difference between the alcohol and MDMA group means
- $b_2 = 10.5$ is the difference between the alcohol and no drug group means
- $b_3 = 5.8$ is the difference between the alcohol and the THC group means

Connection with One-way ANOVA

- Notice the "Analysis of Variance" table in the output of the previous model
- The top row labeled "Regression" represents this entire model
  - The sub-row "Drug" represents the impact of the variable Drug within this model
- In general, ANOVA can be used to assess the importance of any explanatory variable within a multiple regression model
  - In the next lecture we'll explore some of these models

Loose-ends - R-Squared

- Chapter 2 of the textbook introduced the coefficient of variation or R^2
  - R^2 summarizes how much variation in the outcome variable is explained by the explanatory variable
- We can express R^2 using sums of squares:

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

- In calculating R^2, SST refers to the null model that predicts each observation as the mean $\bar{y}$.

Loose-ends - Model Assumptions

- The regression models we've discussed so far are *statistical models* because they specify normally distributed errors
- For statistical inference to be valid, we must assess if the model's errors truly are normally distributed
- Two ways to check this assumption in Minitab are:
    - Looking at a histogram of the residuals
    - Looking at a **normal probability plot**, sometimes called a **QQ-plot**

Model Assumptions - Example

- For the tailgating dataset model that predicts D using Drug, select "Histogram of the Residuals" and "Normal Probability Plot" after hitting the "Graphs" button under the "Fit Regression Model" menu
  - Do the residuals of this model appear normally distributed?
  - No, the histogram is highly skewed, and the quantiles of residuals do not match their expected quantiles under the normal distribution

Conclusion

These notes cover Ch 9 of the textbook. Right now, you should...

1. Know the relationship between one-way ANOVA and linear regression
2. Understand how to perform on statistical inference on the parameters of linear regression model

I encourage you to read Ch 9.1 and 9.2 of the book and their examples.