Stat 139: Statistical Sleuthing through Linear Models

Section 9: Intermediate Regression

TF: Reagan Rose (rrose@g.harvard.edu)

November 11, 2015

Review: Diagnostics for Assumptions

Recall that linear regression makes a number of assumptions about the distribution of the response variables as well as the relationship between Y and X. As noted earlier, you should always conduct the following checks before doing a regression analysis:

1. **Scatterplot:** Plotting Y against X allows for checking of linearity, constant variance, and (non-) independence
2. **Histogram:** A histogram of errors $e_i = Y_i - \hat{Y}_i$ allows for checking of Normality (of residuals), at least indirectly

In general, there is little that an analyst can do against violations of independence. We will now discuss a time-honored way of combating nonlinearity and heteroskedasticity: **transformations of variables**.

When the scatterplot reveals nonlinearities or nonconstant variance, the most common tactic is to transform either the Y or X variables to restore the assumptions. The problem is, of course, that removing one issue may lead to others; for example, a nonlinear plot with constant variance can be transformed into a linear plot with nonconstant variance, due to changes in the Y variable. We will explore how to go about finding suitable transformations in the practice problems, but one should always try:

$$\log(Y), 1/Y, Y^n (n > 1), e^Y.$$

Moreover, one can simultaneously transform both Y and X variables, or simply X or Y alone.

Multiple Regression

Motivation

Last week we talked about simple linear regression, where we had some outcome variable Y and some predictor variable X, and we were interested in modeling the *linear relationship* between X and Y. This concept is extremely important in statistics, and is a building block for a lot of other statistical concepts, but in the real world people **don't really ever use simple linear regression**. Why? Because, as analysts/statisticians/data scientists and/or whatever else we want to call ourselves, we're generally interested in understanding entire **systems** rather than just two variables. An easy way to do this is by **multiple linear regression**, which is just a simple extension of simple regression.

With multiple regression, our goal is to understand **how a number of predictor variables are (jointly) related to an outcome variable.**

Model

Assume we have a dataset with $i = 1, \ldots, n$ observations, each with one value of a response variable, Y_i, and p values of predictor variables $X_{i1}, \ldots, X_{ip}$. This is the classical setup of most datasets. Our multiple regrssion equation is of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i,$$

where, as always, $\epsilon_i \sim^{iid} N(0, \sigma^2)$ represents some noise that exists in our dataset.

Interpretation

Recall from the lecture notes that each regression coefficient β_j represents the "effect" of the predictor variable X_j on the outcome, **controlling for all other variables**. We could also call this the **main effect** of variable X_j. The main thing to note here is that **each coefficient represents the individual effect of a single variable**. If you want to incorporate interaction effects in your regression model, you need to add new β coefficients for those interactions.

Hypothesis Testing

In general, whenever multiple regression is done, R (or basically any other software) will automatically run an **F-test** and will report the results back to you. The purpose of doing the F-test is to answer the question: *Should I even be running a regression? Is there even a relationship to model here?*. If the F-test comes back non-significant, it should tell you that **there's not a whole lot going on in the data**, so any regression you run might have an unusual risk of picking up spurious effects, or results that are merely noise.
The F test tests the hypotheses

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_a : \text{at least one } \beta_j \text{ is not zero,}$$

using the test statistic

$$F = \frac{SSM/df_M}{SSE/df_E}.$$

This should look familiar to you. In particular, it should look like an ANOVA. And in fact, we'll see next week that this F-test is mathematically identical to the ANOVA under certain conditions.

Sidebar on R-Squared and Parsimony

This section is a bit of a digression, and is intended for a curious reader hoping to gain more intuition about multiple regression. Recall that the R^2 in a regression analysis is a measure of goodness-of-fit for the model. Specifically, R^2 represents the **amount of variability in the data that can be explained by the regression**, with higher R^2 values indicating that we have a better understanding of the relationships between the variables we're analyzing. So clearly R^2 is important.

But you may notice in R that when you run multiple regression it actually gives you two values for R^2, the **multiple $\boldsymbol{R^2}$** and the **adjusted $\boldsymbol{R^2}$**. And they're different numbers. This is because, mathematically, as we add more and more predictors to a multiple regression model, we are **guaranteed** to get higher and higher values of R^2 (ask me for the proof if you're interested). So if we only look at the **multiple$\boldsymbol{R^2}$**, we're guaranteed to see higher values as we add predictors to the model, *even if the predictors are not significant.* This is misleading, and might cause us to think we should include some predictors in the model when really we should not. So some smart person came up with **adjusted$\boldsymbol{R^2}$**, which modifies the regular multiple R^2 so it isn't guanteed to improve with more predictors. This is all related to the concept of **parsimony**, which says we should use th simplest model possible to explain the data, and is a critical principle in staitsitics.

Exercises

Question 1. Visualizing transformations and residuals
Download the VARIETIES.CSV file from Canvas. It contains one predictor variable X and 5 different outcomes variables $Y_1, \dots, Y_5$ that we are interested in. For each of the Y_i variables, do the following:

1. Create a scatterplot of x and y_i. Comment on whether the assumptions of linear regression hold.
2. Find the transformation that best restores linearity, and run a regression on the transformed data.
3. Retransform the fitted line and overlay it on the scatterplot from (a).
4. Calculate the residuals from the model in part (b) and generate some residual plots.Use these to comment on the assumptions of the model.

Question 2. Multiple regression with binary predictor Download the HEIGHT.CSV file from Canvas. This dataset contains shows the heights of adults (in inches), along with the heights of each subjects mother and father, and the subject's gender (male or female).

1. Run a complete multiple regression analysis, making sure to check assumptions and model fit. Interpet the results of the regession.
2. Use your model to predict the height of a female whose mother is 5'0 and father is 6'0. Provide 95% confidence and prediction intervals for your estimates and interpret these intervals.

Question 3. R-Squared Toy Example
Come up with a "toy" example that shows why Adjusted R^2 is better for evaluating model fit than Multiple R^2.