Lecture 7

Go over Exam #1

ANOVA as a general linear model

Last semester, we spent several lectures looking at ANOVA, and when we started looking at regression this semester, we talked about some of the similarities between ANOVA and regression analysis in terms of the structure of the hypotheses and other factors. We want to look at this a little bit more closely now in preparation for extending our linear models to nonlinear ones, and other elements that we can add to our simple regression models.

Both ANOVA and regression are models that compare two or more variables to each other. One value is the response variable that we are trying to predict, and other variables are the variables we are the explanatory variables we are using to perform the predictions. The one significant difference between regression and ANOVA is that in regression, we treat all variables as numerical values, whereas in ANOVA, we treat all explanatory variables as categorical variables. Even when we had numerical values for the explanatory variables, we treated these as fixed levels, and not as being along a continuity.

Recall that the predictive model for ANOVA took the form, in the single factor case as $y = \mu + \alpha_i + \epsilon$. The ϵ is the random error terms, μ is the grand mean or the overall mean of all observations regardless of factor level, and finally, α_i is the adjustment to the mean for the factor level that the input belongs to; this value can be positive or negative as needed to adjust to the mean corresponding to the factor level in question. (As we will see below, this is not always how results are presented: one level may be selected as the "default" mean other than the grand mean, which will allow one α to be zero.)

As we added factors, the main effects model continued in the same vein. The two-way ANOVA predictive model was $y = \mu + \alpha_i + \beta_j + \epsilon$. This model is interpreted similarly where ϵ is the random error, μ is the grand mean, and α_i and β_j are the adjustments for an observation with ith level in the first factor, and the jth level in the second factor.

Let's consider a factor with only two values for a simple to interpret contrast. Consider the case of gender. There are traditionally only two categories allowed, so there are only two possible levels. The grand mean will fall in between the means for each of the factor levels, thus one α will be positive and one negative. If these differences are small enough, then this coefficient, similar to our slope coefficient in the regression model, is not sufficiently different from zero to be statistically significant. If they are large enough, then they will be a clear indication of a difference by level of the variable. In a case like this where there are only two categories, there is no "in-between" level, similar to a discrete variable based on counts.

We did use a numerical variable taking on values of either 0 or 1 to represent gender in our regression analysis. (We did this on the exam and a similar situation occurs in the mtcars dataset for American made or not.) We can think of this as saying one gender is "True" taking on the level 1, and one is "False" taking on the level 0. (This is not a judgment call about value since the selection of which level is 1 or 0 is arbitrary and only results in a sign flip.) In such a model, the constant in the equation represents the mean of the 0 factor level, and the slope coefficient represents the change in the mean from one factor level to the other.

Things come a bit less similar if we have three factor levels, for instance, if they are labeled 0, 1, 2 (or 1, 2, 3) in our regression model. The ANOVA model treats each factor level as independent, and so the difference between the levels can be different values. If we model the levels in regression, as described above, we are 1) explicitly ordering them (which is a problem if they are not in order or cannot be ordered), and 2) we are requiring that the distance between the first two levels be the same as the distance between the next two levels, so that even if the variables are in order, that the differences between the categories are the same is a big assumption. As we saw on the exam with education levels at the Beta company, it might not be the case that the salary jump from high school to associates degree is the same as the jump from masters to doctorate, or any other pair of consecutive levels. It's possible this assumption is approximately valid, but it is an assumption we are making by using discrete numbers to stand in for these categorical levels when there are more than two levels.

There is a way around this in regression, which we will discuss the specifics of in a couple of weeks, but it's worth thinking about the advantages and disadvantages of our various encoding strategies as we go.

Let's look at this from the opposite perspective.

Consider the model regression analysis using the mtcars data, modeling mpg using the cylinders variables as a number.

Call:
lm(formula = mpg ~ cyl, data = mtcars)

Residuals:
   Min   1Q Median   3Q   Max
-4.9814 -2.1185 0.2217 1.0717 7.5186

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.8846	2.0738	18.27	< 2e-16 ***
cyl	-2.8758	0.3224	-8.92	6.11e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.206 on 30 degrees of freedom
Multiple R-squared: 0.7262,	Adjusted R-squared: 0.7171
F-statistic: 79.56 on 1 and 30 DF, p-value: 6.113e-10

What if we then model this as an ANOVA with cylinders as a factor?

Call:
  aov(formula = mpg ~ cyl, data = mtcars)

Terms:

	cyl	Residuals
Sum of Squares	824.7846	301.2626
Deg. of Freedom	2	29

Residual standard error: 3.223099
Estimated effects may be unbalanced

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cyl	2	824.8	412.4	39.7	4.98e-09 ***
Residuals	29	301.3	10.4		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Intercept) cyl6 cyl8
26.663636 -6.920779 -11.563636

Compare the linear model when cylinders = 4 to the ANOVA intercept.
Compare the linear model when cylinders = 6 to the ANOVA cyl6. How does the step here differ by changing two cylinders?
Compare the linear model when cylinders = 8 to the ANOVA cyl8. How does the step here differ by doubling the number of cylinders (compared to 4)? Or with the step size of two compared to six cylinders?

The value of the residual standard error is similar. The regression model appears to give a better sense of the quality of the model than the ANOVA with more measurements of fit than just the F-statistic.

What happens if we have a lot of levels, say if we used weight as a factor in our model?

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
Min 1Q Median 3Q Max
-4.5432 -2.3647 -0.1252 1.4096 6.8727

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.2851	1.8776	19.858	< 2e-16 ***
wt	-5.3445	0.5591	-9.559	1.29e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446
F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10

And now ANOVA:

Call:
aov(formula = mpg ~ wt, data = mtcars)

Terms:

	wt	Residuals
Sum of Squares	1124.7955	1.2517
Deg. of Freedom	28	3

Residual standard error: 0.6459274
Estimated effects may be unbalanced

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wt	28	1124.8	40.17	96.28	0.00149 **
Residuals	3	1.3	0.42		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Intercept)	wt1.615	wt1.835	wt1.935	wt2.14	wt2.2	wt2.32	wt2.465	wt2.62
30.400	0.000	3.500	-3.100	-4.400	2.000	-7.600	-8.900	-9.400
wt2.77	wt2.78	wt2.875	wt3.15	wt3.17	wt3.19	wt3.215	wt3.435	wt3.44
-10.700	-9.000	-9.400	-7.600	-14.600	-6.000	-9.000	-15.200	-11.833
wt3.46	wt3.52	wt3.57	wt3.73	wt3.78	wt3.84	wt3.845	wt4.07	wt5.25
-12.300	-14.900	-15.750	-13.100	-15.200	-17.100	-11.200	-14.000	-20.000
wt5.345	wt5.424							
-15.700	-20.000							

Here the residual standard error is much smaller in the ANOVA model, but the coefficients are much more difficult to interpret. One reason the residual error is much smaller, though, is artificial. Because there are so many more "levels", the coefficients for each level can be targeted to the one observation that occurs in most levels. The only error that can occur, then, is if more than one observation has the same weight. Since weight is continuous, this doesn't happen often.

Another major difference is that when we considered cylinders in an engine, if we wanted to predict another value, engines don't actually come with 7 cylinders or 7.5 of them, so the levels are unique and distinct. With weight, however, if we wanted to estimate a valid weight value that did not appear in our original data, we have no way of doing it in the ANOVA model. There is no general relationship between weight and mpg that we could use to extrapolate values easily because we are treating them as distinct levels when they are not in reality.

We will see soon that in R there is a function called glm(), that stands for general linear model. It will allow us to extend our notions of a linear model beyond what we have done so far, including working with logistic models and other extensions of ordinary least squares. We can also do ANOVA with it using factor variables. Compare the glm() output for the factor(ed) variable cyl.

Call:
glm(formula = mpg ~ cyl, data = mtcars)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-5.2636	-1.8357	0.0286	1.3893	7.2364

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.6636	0.9718	27.437	< 2e-16 ***
cyl6	-6.9208	1.5583	-4.441	0.000119 ***
cyl8	-11.5636	1.2986	-8.905	8.57e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 10.38837)

Null deviance: 1126.05 on 31 degrees of freedom
Residual deviance: 301.26 on 29 degrees of freedom
AIC: 170.56

Number of Fisher Scoring iterations: 2

In future lectures, we will also look at interactions between terms in linear models. We saw some cases last semester where interactions between factors was statistically significant, so we'll want to be able to look at such interactions in the regression model as well. This is all still coming up!

In the next lecture we are going to return to ANOVA analysis to look at Factorial Designs and analysis of such designs, specifically 2^p factorial designs. Think about the relationship to regression as we proceed.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://sites.utexas.edu/sos/guided/inferential/numeric/glm/
3. https://www.theanalysisfactor.com/general-linear-model-anova-regression-same-model/
4. http://psych.colorado.edu/~carey/Courses/PSYC5741/handouts/GLM%20Theory.pdf
5. https://ajstewartlang.github.io/11_glm_anova_pt1/knitted_workshop/11_glm_anova_pt1.html
6. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm