

IDS 702: MODULE 1.9

SPECIAL PREDICTORS, F-TESTS, AND MULTICOLLINEARITY

DR. OLANREWAJU MICHAEL AKANDE

SPECIAL PREDICTORS

SPECIAL PREDICTORS: HIGHER ORDER TERMS

- We have already seen that the relationships between a response variable and some of the predictors can be potentially nonlinear.
- Sometimes our outcome of interest can appear to have quadratic or even higher order polynomial trends with some predictors.
- Whenever this is the case, we should look to include squared terms or higher order powers for predictors to capture trends.
- In the baseline salary example, we included squared terms for both age and experience.
- General practice: include all lower order terms when including higher order ones (even if the lower order terms are not significant). This aids interpretation.
- As we have seen before, the best way to present results when including quadratic/polynomial trends is to plot the predicted average of Y for different values of X .

SPECIAL PREDICTORS: INDICATOR/DUMMY VARIABLES

- From the Harris Trust and Savings Bank example, we have also seen how to include binary variables in a MLR model with the variable sex .
- In the example, we could actually have used the variable f_{sex} (where 1=female and 0=male) instead of sex to give us the same exact results.
- That means that we also could have made a variable equal to 1 for all males and 0 for all females, instead.
- The value of that coefficient would be 767 instead of -767 like we had. All other statistics stay the same (SE, t-stat, p-value). Other coefficients also remain the same.
- Turns out that we cannot include indicator variables for the two values of the same binary variable when we also include the intercept.

SPECIAL PREDICTORS: INDICATOR/DUMMY VARIABLES

- It is not possible to estimate all three of these parameters in the same model uniquely.
- The exact same problem arises for any set of predictors such that one is an exact linear combination of the others.
- Example: Consider a regression model with dummy variables for both males and females, plus an intercept.

$$y_i = \beta_0 + \beta_1 M_i + \beta_2 F_i + \epsilon_i = \beta_0 * 1 + \beta_1 M_i + \beta_2 F_i + \epsilon_i$$

- Note that $M_i + F_i = 1$ for all cases. Thus,

$$y_i = \beta_0 * (M_i + F_i) + \beta_1 M_i + \beta_2 F_i + \epsilon_i = (\beta_0 + \beta_1) M_i + (\beta_0 + \beta_2) F_i + \epsilon_i.$$

We can estimate $(\beta_0 + \beta_1)$ and $(\beta_0 + \beta_2)$ but not all three uniquely.

- Side note: there is no need to mean center dummy variables, since they have a natural interpretation at zero.

SPECIAL PREDICTORS: INDICATOR/DUMMY VARIABLES

- What if a categorical variable has $k > 2$ levels?
- Make k dummy variables, one for each level.
- Use only $k - 1$ of the levels in the regression model, since we cannot uniquely estimate all k at once if we also include an intercept (see previous slide).
- Excluded level is called the baseline.
- R will actually do this for you automatically; that is, make the $k - 1$ dummy variables and set the first level as the baseline.
- Values of coefficients of dummy variables are interpreted as changes in average Y over the baseline.
- We will go through an example soon.

SPECIAL PREDICTORS: INTERACTION TERMS

- Sometimes the relationship of some predictor with Y depends on values of other predictors. This is called an **interaction effect**.
- Sometimes, the question we wish to answer would require including interactions in the model, even though they might not be significant.
- An example of interaction effect for the Harris Bank dataset would be if the effect of age on baseline income was different for male versus female.
- That is, what if older males are paid more starting salaries than younger males but the reverse is actually the case for females?
- How do we account for such interaction effects? Make an interaction predictor: **multiply one predictor times the other predictor**. Ideally, one of them should be a factor variable.
- General practice is to include all main effects (each variable without interaction) when including interactions.

TESTING IF GROUPS OF COEFFICIENTS ARE EQUAL TO ZERO

- With so many variables (polynomial terms, dummy variables and interactions) in a linear model, we may want to test if multiple coefficients are equal to zero or not.
- We can do so using an F test (a nested F test in this case).
- First, we fit a MLR model with all p predictors. That is,

$$M_1 : y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i; \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

- We can compute the sum of squares of the errors (SSE_1) or residual sum of squares (RSS_1) for the FULL model, that is,

$$RSS_1 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

TESTING IF GROUPS OF COEFFICIENTS ARE EQUAL TO ZERO

- Now suppose we want to test that a particular subset of q of the coefficients are zero.

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \dots = \beta_p = 0.$$

- We fit a reduced model that uses all the variables except the last q , that is,

$$M_0 : y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{i(p-q)} + \epsilon_i; \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

- Let's call the residual sum of squares for that model RSS_0 .

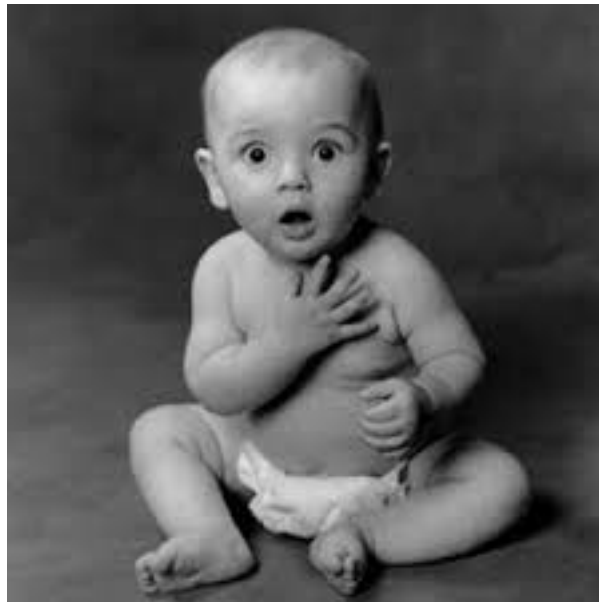
Which of the two RSS values would be larger? Why?

- Then the appropriate F-statistic is

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n - p - 1)}.$$

TESTING IF GROUPS OF COEFFICIENTS ARE EQUAL TO ZERO

- To calculate the p-value, look for the area under the F curve with q degrees of freedom in the numerator, and $(n - p - 1)$ degrees of freedom in the denominator.
- Guess what? As is the case with pretty much everything else we do in this class, this is so easy to do in R!



MULTICOLLINEARITY

THE PROBLEM OF MULTICOLLINEARITY

- Just like we had with the dummy variables, you cannot include two variables with a perfect linear association as predictors in regression.
- Example: suppose the true population line is

$$\text{Avg. } y = 3 + 4x.$$

- Suppose we try to include x and $z = x/10$ as predictors in our own model,
- Example: suppose the true population line is

$$\text{Avg. } y = \beta_0 + \beta_1 x + \beta_2 z,$$

and estimate all coefficients. Since $z = x/10$, we have

$$\text{Avg. } y = \beta_0 + \beta_1 x + \beta_2 \frac{x}{10} = \beta_0 + \left(\beta_1 + \frac{\beta_2}{10} \right) x$$

- We could set β_1 and β_2 to ANY two numbers such that $\beta_1 + \beta_2/10 = 4$. The data cannot pick from the possible combinations.

THE PROBLEM OF MULTICOLLINEARITY

- In real data, when we get “close” to perfect colinearities we see standard errors inflate, sometimes massively.
- When might we get close:
 - Very high correlations ($|\rho| > 0.9$) among two (or more) predictors in modest sample sizes.
 - When one or more variables are nearly a linear combination of the others.
 - Including quadratic terms as predictors without first mean centering the values before squaring.
 - Including interactions involving continuous variables.

THE PROBLEM OF MULTICOLLINEARITY

- How to diagnose:
 - Look at a correlation matrix of all the predictors (including dummy variables). Look for values near -1 or 1.
 - If you are suspicious that some predictor is a near linear combination of others, run a regression of that predictor on all other predictors (not including Y) to see if R squared is near 1.
 - If the R squared is near 1, you should think about centering your variables or maybe even excluding that variable from your regression in some cases.
 - Take a look at the **variance inflation factor**.
 - Variance inflation factor measures how much the multicollinearity between a variable and other variables in the model inflates the variance of the regression coefficient for that variable.

VARIANCE INFLATION FACTOR

- $$\text{VIF}_j = \frac{1}{1 - R_{X_j|X_{-j}}^2}$$

where $R_{X_j|X_{-j}}^2$ is the R-squared from the regression of predictor X_j on all other predictors $(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p)$.

- Since R-squared always lies between 0 and 1,
 - the denominator $1 - R_{X_j|X_{-j}}^2 \leq 1$
 - which implies that $\text{VIF} \geq 1$
- Generally, VIF of
 - 1 = not correlated. **Why?**
 - between 1 and 5 = moderately correlated.
 - greater than 5 = highly correlated.
- Typically, we start to get worried when $\text{VIF} > 10$.

WE SEE MULTICOLLINEARITY... SO WHAT?

- Multicollinearity is really only a problem if standard errors for the involved coefficients are too large to be useful for interpretation, and you actually care about interpreting those coefficients.
- In the Harris Bank example,
 - The main coefficient of interest is the one for `sex`.
 - The remaining variables are really just "control variables". That is, those variables may be correlated with both `bsal` and `sex`, and so we want to account for their effects in our model.
 - Recall that the correlation between `age` and `exper` was actually 0.8.
 - Even with this correlation, it is still okay to keep both in the model since we want to simply account for them but do not care about interpreting either.
- Another scenario is prediction: including highly correlated predictors can increase prediction uncertainty.

WHAT TO DO ABOUT MULTICOLLINEARITY?

- What if you do want to interpret the coefficients involved in the multicollinearity, and the SEs are inflated substantially because of it?
- Easiest remedy: remove one of the "offending" predictors.
- Keep the one that is easiest to explain or that has the largest T-statistic.
- Better remedy:
 - Mean center (or scale) your variables. It helps but may not always solve the problem.
 - Use a Bayesian regression model with an informative prior distribution on the parameters (take STA 602).
 - Get more data! Multicollinearity tends to be unimportant in large sample sizes.

WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!