Mathematics 231

Lecture 34

Liam O’Brien

Announcements

- Today – Multiple Regression
  - Assumption checking
  - Model Interpretation

Example: Final Model

- Remember the birthweight “final” model we obtained?
- It contained head circumference, infant length, toxemia, and a toxemia-length interaction term.
- We examined the residual vs. predicted value plot (an added component plot) and saw no obvious assumption violations.
- The R^2 value was 0.78, indicating that about 78% of the variability in birthweights is explained by our regression model.

Final Model

```
. regress birthwt length headcirc toxemia  lentox

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  4,    95) =   85.05
       Model |  5641405.82     4  1410351.46           Prob > F      =  0.0000
    Residual |  1575336.93    95   16582.494           R-squared     =  0.7817
-------------+------------------------------           Adj R-squared =  0.7725
       Total |  7216742.75    99  72896.3914           Root MSE      =  128.77

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   35.89579   5.271313     6.81   0.000     25.43091    46.36067
    headcirc |   46.05232   7.392782     6.23   0.000      31.3758    60.72885
     toxemia |  -854.8818   379.2735    -2.25   0.026    -1607.835   -101.9288
      lentox |   21.05083   10.10944     2.08   0.040      .981055    41.12061
       _cons |  -1427.483   160.9943    -8.87   0.000    -1747.097   -1107.869
------------------------------------------------------------------------------
```

Additional Checks – Collinearity

- We know that too much collinearity is BAD.
- We also know that anytime we include an interaction variable, we are guaranteed to have it to some extent.
- We can check the degree of collinearity among the predictors by looking at the **variance inflation factors (VIF)**.
- To look at this in Stata, type "vif" after running your regression.

Variance Inflation Factors

- The variance inflation factors take each predictor and run a regression of that predictor on the other predictors in the model.
- The inverse of the VIF tells you how much variability of the predictor under consideration is not explained by the other predictors (large numbers are good).
- There is no standard cutoff for how small (1/vif) has to be in order to remove that predictor from the model.
- The VIF doesn’t work and play well with interactions.

Example

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      lentox |    145.55    0.006871
     toxemia |    143.91    0.006949
      length |      2.12    0.472598
    headcirc |      2.09    0.478004
-------------+----------------------
    Mean VIF |     73.42
```

- Toxemia and the interaction between toxemia and length have a lot of collinearity with the other predictors. We will keep it in though, since the interaction is always necessary if it’s a good predictor of the response.

Example: After Dropping Interaction

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
    headcirc |      2.03    0.492011
      length |      2.03    0.492011
-------------+----------------------
    Mean VIF |      2.03
```

Checking Homoscedasticity

- There are *many* tests around for checking the homoscedasticity assumption.
- The Cook-Weisberg (aka Breusch-Pagan) test will perform a test of heteroscedasticity (the null hypothesis is that the variances are all equal).
- To perform this test in Stata, type "hettest" after the regress command – this will test the residual vs predicted values.
- To test the residuals against each predictor, type "hettest, rhs".
  - This is problematic if there are interactions, or a lot of predictors.

Example

```
. hettest

Breusch-Pagan / Cook-Weisberg test for
  heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of birthwt

         chi2(1)      =     0.12
         Prob > chi2  =   0.7296
```

- The equal variance assumption is not violated (p > 0.10).

Checking Homoscedasticity

- There is another built-in test in Stata for checking homoscedasticity called Szroeter's test.
- The null hypothesis is still that homoscedasticity holds, but this test must be done for each predictor separately, thus making it difficult to perform in the presence of an interaction.
- Note that anytime you test homoscedascticity for more than one predictor, you are inflating the alpha level.

Multiple Comparisons

- As we went through our forward stepwise procedure when building our model, we were performing multiple tests (for the coefficients).
- This inflates the alpha level in much the same way as we encountered when looking at ANOVA.
- There is no standard way to adjust for this, but you should keep it in mind when many variables are present.

Multiple Comparisons

- Testing sets of variables at once rather than doing it one-by-one cuts down on the number of tests performed (nested F-test).
- You may also want to reduce the point at which you consider the p-value small enough to include the covariate in the model.
- The homoscedasticity tests do have multiple comparisons adjustment procedures built in.

Comparing Several Predictors at Once

- To compare more than one coefficient in Stata, go to **Statistics > General Post-estimation > Tests > Test parameters after model fitting**
- Enter the names of the variables you want to simultaneously test in the box
- You can test to see if they're all zero, or if they're all equal to each other.

Comparing Several Predictors at Once

```
. testparm lentox toxemia length

 ( 1)  lentox = 0
 ( 2)  toxemia = 0
 ( 3)  length = 0

       F(  3,    95) =   20.83
            Prob > F =    0.0000
```

Note that the null is that they are all zero, so that hypothesis is rejected here

Interpretation

- Okay, all things considered, it seems that the model we built is a good one.
- It explains a lot of the variability in the outcome.
- There are no predictors with coefficients close to zero.
- We saw no obvious violations of linearity or homoscedasticity.
- While collinearity is present, the standard errors of the regression coefficients are not too large, and the interaction term is significant.

Interpretation

- Now we need to be able to interpret the model we obtained.
- E(birthwt) = -1427 + 35.9(length) + 46.1 (headcirc) – 854.9 (toxemia) + 21.1 (toxemia*length)
- How do we make sense of how birthweight changes with the predictors?

Interpreting Interactions

- Remember that when an interaction is present, one predictor modifies the effect of the other predictor on the outcome.
- In this case, toxemia modifies the effect of infant length on birthweight (i.e., length does have an effect on birthweight, but that effect is different depending on whether toxemia is present or not).

Interpreting Interactions

- When interactions are present and at least one of the interacting variables in binary, it is often useful to consider the cases when it is 0 or 1, separately.
- For those with toxemia:

  E(birthwt) = -2282 + 57.0(length) + 46.1 (headcirc)
- For those without toxemia:

  E(birthwt) = -1427 + 35.9(length) + 46.1 (headcirc)

Interpretation

- For every 1cm increase in head circumference, birthweight increases by 46.1g, on average, holding infant length constant.
- Among those with toxemia, for every 1 cm increase in length, there is a 57g increase in birthweight, on average, holding head circumference constant.
- Among those without toxemia, for every 1 cm increase in length, there is a 35.9g increase in birthweight, on average, holding head circumference constant.

Interpretation

- Does this mean that babies exposed to toxemia have higher birthweights since an increase in length seems to increase birthweight more for them?
- No… look at the value of the intercept term.
- Those with toxemia start out with a severe disadvantage, although they catch up to those without once infant length gets to a certain point.

What is the ‘Standard’ Method for Obtaining a Multiple Regression Model?

- There is none!
- There are semi-standard methods:
    - Forward Selection
    - Backward elimination
    - Forward Stepwise
    - Backward Stepwise
- These are usually built into statistical software packages, but are rarely used in practice.

Checking Assumptions

- Regardless of how you go about generating your model, you must check the basic assumptions on linear regression:
  1. Linear relationship between your set of predictors and the response (can be done one predictor at a time).
  2. Homoscedasticity (look at all the residual versus predictor plots, or at least the residual versus fitted value plot).
  3. Normality of residuals (can look at a normal quantile plot of the residuals).
     - Type “predict resid, resid” to generate them.

Take Into Account "Lurking" Variables

- We have seen the importance of interaction terms, where the value of one predictor alters the effect of another predictor on the outcome.
- However, anytime you suspect that a variable may have some effect on the outcome and/or other predictors' relationships with the outcome, you should include that variable in the model.
- If inclusion of this variable in the model alters the values of the coefficients considerably, you want to be extra careful to ensure that it remains in the model.

Beware the “Kitchen Sink” Model

- Adding predictors to your model will never decrease the amount of variation in the response explained, but will often introduce collinearity, making the estimation of the coefficients unstable.
- For every new predictor, you lose a degree of freedom to estimate its coefficient that could be used to help explain away the error in the model.

Don’t Sacrifice Fit for Interpretability

- Sometimes you may find three- or four-way interactions to be important predictors; or you may find some bizarre transformation to work well.
- These may contribute to problems with interpretability of the model – a model that cannot be described to the applied researcher is not useful in practical terms.