Multiple Regression Analysis: The Problem of Inference

Jamie Monogan

University of Georgia

Intermediate Political Methodology
Objectives

By the end of this meeting, participants should be able to:

- Construct confidence intervals for individual partial regression coefficients.
- Test hypotheses about individual partial regression coefficients with $t$-ratios.
- Test hypotheses about a whole model, a set of coefficients, or a restriction on a model with $F$-ratios.
- Forecast values of the outcome given values of the predictors.
Inference on Individual Partial Regression Coefficients

- Just as with the two-variable model, we assume that the disturbances have a normal distribution.
- This allows us to conduct inferences on individual partial coefficients using the $t$-distribution just as before.
- Choose a Type I error rate $\alpha$ and then build a confidence interval or conduct a hypothesis test.
- Confidence interval for $\beta_j$: $\hat{\beta}_j \pm t_{\alpha/2}se(\hat{\beta}_j)$.
- Hypothesis test for $\beta_j$: $t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$.
- For hypothesis testing, you could calculate a $p$-value or compare to the critical value, $t_\alpha$ for a one-tailed test and $t_{\alpha/2}$ for a two-tailed test.
- In all of this, $t$ has $n - k$ degrees of freedom where $k$ is the number of parameters.
Suppose you want to test a hypothesis about more than one coefficient at once.

Estimate a *restricted* model as if the null hypothesis is true and an *unrestricted* model that does not impose that assumption.

For example, consider the unrestricted model:
\[ Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i. \]

You state the null hypothesis: \( H_0 : \beta_2 = \beta_3 = 0. \)

If the null hypothesis is true, then the restricted model is:
\[ Y_i = \beta_1 + \beta_4 X_{4i} + u_i. \]

The unrestricted model will explain more variance in \( Y_i. \) Does it have a *significantly* better fit?
General $F$ Testing

- Provided the restricted model is *nested* within the unrestricted model, we can use an $F$-ratio to test whether the null hypothesis is true.
- Record the coefficient of determination for the unrestricted model ($R^2_{UR}$) and for the restricted model ($R^2_R$) and use them to calculate the test statistic:
  \[ F = \frac{(R^2_{UR} - R^2_R)/m}{(1 - R^2_{UR})/(n - k)} \]
- In this equation $n$ is the sample size, $k$ is the number of parameters in the unrestricted model, and $m$ is the number of parameters omitted in restricted model. (In the example from the previous slide: $k = 4$ & $m = 2$.)
- This test statistic has an $F$ distribution with $m$ & $n - k$ degrees of freedom.
An Alternate Version of $F$

An equivalent way to state $F$ is:

$$F = \frac{(RSS_R - RSS_{UR})/m}{(RSS_{UR})/(n - k)}$$

If ever your restriction requires you to rescale the dependent variable, you must use this form of the test. The distribution is exactly the same.

This specification also illustrates why the test statistic has an $F$-distribution: It is a ratio of $\chi^2$ distributed statistics.
The unrestricted model is your theoretically-specified linear model.

Unrestricted model: \( Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \cdots + \beta_k X_{ki} + u_i. \)

Overall significance of multiple regression: \( H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0. \)

Restricted model: \( Y_i = \beta_1 + u_i. \)

Significance of a block of predictors: E.g., \( H_0 : \beta_2 = \beta_3 = 0. \)

Restricted model: \( Y_i = \beta_1 + \beta_4 X_{4i} + \cdots + \beta_k X_{ki} + u_i. \)

Testing a linear restriction: E.g., \( H_0 : \beta_2 + \beta_3 = 1. \)

Restricted model:
\[
Y_i = \beta_1 + X_{2i} + \beta_3(X_{3i} - X_{2i}) + \beta_4 X_{4i} + \cdots + \beta_k X_{ki} + u_i.
\]

On the last point, we would estimate the model using restricted least squares (RLS).
Post hoc t Tests

- Besides restricted least squares, we also can test linear restrictions and (within that) equality of coefficients using a post hoc t-test.
- This is very handy for inference on interaction terms. (Next week.)
- State your null hypothesis such that all parameters are on one side and a number is on the other side. For example:
  - $H_0 : \beta_2 = \beta_3$ is equivalent to $H_0 : \beta_2 - \beta_3 = 0$.
  - $H_0 : \beta_2 = 1 - \beta_3$ is equivalent to $H_0 : \beta_2 + \beta_3 = 1$.
- Let’s say in general your hypothesis is: $H_0 : a\beta_2 + b\beta_3 = c$, where $a$, $b$, & $c$ are constants (potentially $-1$ for subtraction).
- Then the standard error of your $t$ test is:
  $$se = \sqrt{a^2 \text{Var}(\hat{\beta}_2) + b^2 \text{Var}(\hat{\beta}_3) + 2ab \text{Cov}(\hat{\beta}_2, \hat{\beta}_3)}.$$  
- Your test statistic is: $t = \frac{a\hat{\beta}_2 + b\hat{\beta}_3 - c}{se}$. 
Forecasting with a multiple regression model is just as easy as for the two-variable model.

\[ \hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_{20} + \cdots + \hat{\beta}_k X_{k0}. \]

We can still compute error variances for prediction of the mean or of individual values. This is tricky in scalar notation, though.

Refer to Appendix C for notes on matrix-based computation of standard errors.
For Next Time

- Read Gujarati & Porter chap 9 (Dummy Variable Regression Models).
- Study for the midterm. Bring your questions next time.
- Estimate a model in which crimes per 100,000 residents is a function of percent metropolitan, percent poverty, and percent single parent.
- Present your results in a well-formatted table. Interpret the intercept and partial slope coefficients.
- Present the 95% percent confidence intervals for each parameter and report the results of the hypothesis test $H_0 : \beta = 0, H_1 : \beta \neq 0$ at the 95% confidence level for each parameter. Also, test whether the whole model explains a significant portion of variance in crime.
- Would a percentage point increase in metropolitan population and a percentage point increase in poverty together raise the crime rate by 10 per 100,000 residents on average? State the null and alternative hypothesis. Report the results of the hypothesis test.