Week 12 Live Session - Multiple OLS Regression Inference

w203 Instructional Team

Announcements

Useful functions in R:

model coefficients: coefficients(fit)

predicted values: fitted(fit)

residuals: residuals(fit)

heteroskedasticity-robust covariance matrix for model parameters: vcovHC(fit)

heteroskedasticity-robust standard errors and hypothesis tests: coeftest(fit, vcov = vcovHC)

CIs for model parameters (at 95%): confint(fit, level=0.95) (Warning: not robust to heteroskedasticity)

For heteroskedasticity-robust confidence intervals, get the variance of each coefficient from vcovHC, take the square root to get the standard error, get the proper t critical values from qt, and construct manually.

Variance of OLS Estimators

Recall the expression for the variance of each OLS slope coefficient:

$$var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

For each component of this equation, explain (1) what it means, and (2) why it moves the standard error of β_j up or down.

1. σ^2
2. $\frac{1}{SST_j}$
3. $\frac{1}{1-R_j^2}$

Component 3 has a special name: the Variance Inflation Factor. You can find the variance inflation factor for each variable in a linear model using the vif function in the car package. Interpreting VIFs depends very much on context, but a VIF of 10 would usually be considered very high.

To get the variance of each coefficient in R, we would typically get the diagonal elements of the robust covariance matrix, diag(vcovHC(model))

To get the standard error of a coefficient, take the square root of the variance.

R Exercise

In this analysis, we will use the mtcars dataset which is a dataset that was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). The dataset is automatically available when you start R. For more information about the dataset, use the R command: help(mtcars)

Some useful libraries for multivariate ols regression:

```
library(car)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
```

Q1.1: Using the mtcars data, run a multiple linear regression to find the effect of displacement (disp), gross horsepower (hp), weight (wt), and rear axle ratio (drat) on the miles per gallon (mpg).

Q1.2: For each of the 6 CLM assumptions, assess whether the assumption holds. Where possible, demonstrate multiple ways of assessing an assumption. When an assumption appears violated, state what steps you would take in response.

a. Linear population model
b. Random Sampling
c. No perfect multicollinearity
d. Zero-conditional mean
e. Homoskedasticity
f. Normality of Errors

Q1.3: In addition to the above, assess to what extent (imperfect) multicollinearity is affecting your inference.

Q1.4 Interpret your slope coefficients, and note which ones are significantly different from zero. Whether or not you detected heteroskedasticity above, be conservative in this step and use robust standard errors.

Alternate specification

Next, alter your previous model by taking the log of mpg. Fit the model as before and examine your diagnostic plots.

Q1.5 How does the log transform affect which CLM assumptions hold.

Q1.6 Which model has a better fit.

Q1.7 (As time allows) Report the results of both models in a nicely formatted regression table.

More about Multicollinearity

A common problem with multivariate regression is collinearity. If two or more predictor variables are highly correlated, and they are both entered into a regression model, it increases the standard error of each one and you get very unstable estimates of the slope. We usually assess the collinearity by variance inflation factor (VIF).

Ways to Detect Multicollinearity

We begin by regressing a particular independent variable on all other independent variables.

1. As the squared correlation (r2) increases toward 1.0, the magnitude of potential problems associated with multicollinearity increases correspondingly.
2. Tolerance (1-R2) One minus the squared multiple correlation of a given IV from other Ivs in the equation. Tolerance values of 0.10 or less Indicate that there may be serious multicollinearity.
3. The Variance Inflation Factor [VIF=1/(1-R2)] VIF Is the reciprocal of the Tolerance. Any VIF of 10 or more provides evidence of serious multicollinearity.
4. Condition Number (k) The square root of the ratio of the largest eigenvalue to the smallest eigenvalue. k of 30 or larger indicate that there may be serious multicollinearity.