Recitation 5

Seung-hun Lee

Columbia University

Multivariate Regression Models

Sampling distribution

- The estimates for the $\hat{\beta}_j$, can be obtained in a similar way in which we have obtained the OLS estimates for the single variable version.

$$\min_{\{\beta_0,\beta_1,\beta_2\}} \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}]^2$$

- After some more amount of algebra (than the single variable case), the result we get is the following

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_{1i}-\bar{X}_1)(Y_i-\bar{Y}) \sum_{i=1}^n (X_{2i}-\bar{X}_2)^2 - \sum_{i=1}^n (X_{2i}-\bar{X}_2)(Y_i-\bar{Y}) \sum_{i=1}^n (X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)}{\sum_{i=1}^n (X_{1i}-\bar{X}_1)^2 \sum_{i=1}^n (X_{2i}-\bar{X}_2)^2 - [\sum_{i=1}^n (X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)]^2} \\
\hat{\beta}_2 &= \frac{\sum_{i=1}^n (X_{2i}-\bar{X}_2)(Y_i-\bar{Y}) \sum_{i=1}^n (X_{1i}-\bar{X}_1)^2 - \sum_{i=1}^n (X_{1i}-\bar{X}_1)(Y_i-\bar{Y}) \sum_{i=1}^n (X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)}{\sum_{i=1}^n (X_{1i}-\bar{X}_1)^2 \sum_{i=1}^n (X_{2i}-\bar{X}_2)^2 - [\sum_{i=1}^n (X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)]^2}
\end{aligned}$$

- What matters at this point is how we should **interpret** these coefficients.

Multicollinearity

- We are quite likely to end up including independent variables that are highly correlated with each other. There are two
- We say two variables X_1 and X_2 are **perfectly multicollinear** if X_1 is in an exact linear relationship of some sort with X_2.
- Any multicollinearities that are not in exact linear relationship is referred to as **imperfect multicollinearity**.

Multivariate Regression Models

Multicollinearity

- **Assume that** $X_2 = cX_1$ **for some constant** c: Then we have $(X_{2i} - \bar{X}_2) = c(X_{1i} - \bar{X}_1)$. Then $\hat{\beta}_1$ changes to $\frac{0}{0}$
- **Dummy variable trap**: Say that you have the dummy variable for females and males. Let each of them be X_{1i} and X_{2i} with $X_{2i} = 1 - X_{1i}$. Then the regression can be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \iff Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2(1 - X_{1i}) + u_i$$
$$\iff Y_i = \beta_0 + \beta_2 + (\beta_1 - \beta_2)X_{1i} + u_i$$

  Therefore, by including both X_{1i} and X_{2i} in the same regression, the X_{2i} vanishes from the equation. This is why when you have dummy variables for all categories in the observation, **one of them must be left out.**

Multivariate Regression Models

Joint hypothesis tests (Why it is not straightforward)

- Suppose that you are running a two-sided test with 5 independent variables and significance level $\alpha = 5\%$ under the null hypothesis

$$H_0 : \beta_1 = ...\beta_5 = 0$$

- You reject the null hypothesis when $|t_i| \geq 1.96$ with probability 0.05.
- Now assume that each test statics are independent. Then the probability of incorrectly rejecting the null hypothesis using this approach is

$$\begin{aligned} \Pr(|t_1| > 1.96 \cup ... \cup |t_5| > 1.96) &= 1 - \Pr(|t_1| \leq 1.96 \cap .. \cap |t_1| \leq 1.96) \\ (\because \text{Independence of } t_i\text{'s}) &= 1 - \Pr(|t_1| \leq 1.96) \times .. \times \Pr(|t_5| \leq 1.96) \\ &= 1 - (0.95)^5 \\ &= 0.2262 \end{aligned}$$

- This means that the rejection rate under the null is not 5% but 22% percent - we end up rejecting the null hypothesis more than we have to.

Multivariate Regression Models

F-test

- This is a test where all parts of the joint hypothesis can be tested at once. It also has mechanism for correcting the correlation between the *t*-test statistics.
- It ultimately allows us to correctly set the significance level even for the multiple testing case.
- The usual joint hypothesis test for the regression with k variables (not including the constant term) is

  $$H_0 : \beta_1 = ... = \beta_k = 0, \ H_1 : \neg H_0$$

  where H_1 refers to the case where there is a nonzero element in any one of β_1 to β_k.
- Note that the default F-test null hypothesis for STATA is as above

Multivariate Regression Models

Other tests

- Suppose that instead of β_1 and β_2 being zero, we are just interested in whether they are equal.
- The F-test can also be used for testing this hypothesis. The setup of the hypothesis would be

$$H_0 : \beta_1 = \beta_2 \; H_1 : \beta_1 \neq \beta_2$$

- With this, you can answer various types of tests (e.g. is $\beta_1 + \beta_2 = 100$?)

Multivariate Regression Models

Interpreting the results

- Below are the results of a sample regression on multiple variables. I regress *birthweight* on *smoker, alcohol, Nprevist* (number of prenatal visits to doctor).

```
. regress birthweight smoker alcohol nprevist
```

Source	SS	df	MS
Model	76610831.2	3	25536943.7
Residual	975009173	2,996	325436.974
Total	1.0516e+09	2,999	350656.887

Number of obs	=	3,000
F(3, 2996)	=	78.47
Prob > F	=	0.0000
R-squared	=	0.0729
Adj R-squared	=	0.0719
Root MSE	=	570.47

birthweight	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smoker	-217.5801	26.6796	-8.16	0.000	-269.8923	-165.2679
alcohol	-30.49129	76.23405	-0.40	0.689	-179.9677	118.9851
nprevist	34.06991	2.854994	11.93	0.000	28.47197	39.66786
_cons	3051.249	34.01596	89.70	0.000	2984.552	3117.946

- You can see that running multivariate regression is similar in terms of the techniques involved.

Interpreting the results

- Additional complication rises from interpreting the goodness of fit. In addition to R^2, we now get the **adjusted** R^2, which is defined as

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\text{RSS}}{\text{TSS}}$$

- Since we are assuming that $k \geq 1$, adjusted R^2 is smaller than the R^2.
- As we include more variables, the $\frac{n-1}{n-k-1}$ increases, leading to further decrease in adjusted R^2.
- However, if the new variables are very relevent, $\frac{\text{RSS}}{\text{TSS}}$ decreases.
- This reduces the gap between R^2 and the adjusted R^2. If the adjusted R^2 do not decrease drastically, it is a sign that we are adding a relevant variable.

Deriving The F-statistic

- One uses t-statistics from individual hypotheses. This is calculated as

$$\frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}^2_{t_1,t_2}}\right)$$

- Another, which is useful for calculating $H_0 : \beta_1 = ... = \beta_q = 0$ hypothesis, uses R^2 from the 'unrestricted' and 'restricted' regressions.
  - Assume the following setup

$$\begin{aligned}
\text{Restricted: } & Y_i = \beta_0 + 0X_{1,i} + ... + 0X_{q,i} + \beta_{q+1}X_{q+1,i} + ... + \beta_k X_{k,i} + u_i \\
\text{Unrestricted: } & Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_q X_{q,i} + \beta_{q+1}X_{q+1,i} + ... + \beta_k X_{k,i} + u_i
\end{aligned}$$

  - Restricted regression assumes that H_0 is true and then only optimizes with respect to $\beta_{q+1}, ..., \beta_k$.
  - Unrestricted regression does not assume that H_0 is true and optimizes with respect to all slope coefficients.

Multivariate Regression Models

Deriving The F-statistic

- We use R^2 from these two regressions.

$$\frac{(R^2_{\text{Unrestricted}} - R^2_{\text{Restricted}})/q}{(1 - R^2_{\text{Unrestricted}})/(n-k-1)}$$

  - k: number of independent variables (not counting intercept)
  - q is the number of restrictions.
  - Since unrestricted models allows roles for $X_1, ..., X_q$ variables, they have higher R^2 (Restricted: They should have no role)
- Another: Using $R^2_{\text{Restricted}} = 1 - \frac{RSS_{\text{Restricted}}}{TSS}$, we can write

$$\frac{(RSS_{\text{Restricted}} - RSS_{\text{Unrestricted}})/q}{(RSS_{\text{Unrestricted}})/(n-k-1)}$$

Multivariate Regression Models

Control variables and conditional mean independence

- Assume that

$$\begin{aligned}\text{True: } & Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \\ \text{Mistake: } & Y_i = \beta_0 + \beta_1 X_i + u_i^* \\ \text{Sample: } & Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i\end{aligned}$$

- If we end up with an omitted variable bias by not including Z_i, Then, we have a problem.

$$\begin{aligned}E[u_i^*|X_i] &= E[\beta_2 Z_i + u_i|X_i] \\ &= \beta_2 Z_i + E[u_i|X_i] \neq 0\end{aligned}$$

- Assumption 2 from the classical linear regression model (refer to Recitation 3), fails and $\hat{\beta}_1$ without inclusion of Z_i is biased

Multivariate Regression Models

Control variables and conditional mean independence

- We can find W_i variable correlated with Z_i and put it in the regression
- By doing so, we achieve three things
  - The u_i term is no longer correlated with X_i ($cov(X_i, u_i) = 0$)
  - For given value of W_i, then the variable of interest X_i is no longer correlated with the omitted determinant of Y_i
  - For given W_i, X_i acts as if they are randomly assigned
- Variable W_i that achieves this is called an **effective control variable**.
- In this case, we say that the **conditional mean independence** hold,

$$E[u_i|X_i, W_i] = E[u_i|W_i]$$

- Note that W_i itself does not need to have causal relationship with Y_i

Multivariate Regression Models

Nonlinear regressions

- Not everything in world is linearly related
- For correlations like this, nonlinear regressors are necessary
- When incorporating such regressors, the interpretation of each coefficient becomes trickier.
- Quadratic relations: Think about wage and age - wages increase with age, but (usually) at a decreasing pace

$$W = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

- The marginal effect of X on W can be written as

$$\frac{\partial W}{\partial X} = \beta_1 + 2\beta_2 X$$

Multivariate Regression Models

Nonlinear regressions

- In a linear regressor only format, the marginal effect is

$$W = \beta_0 + \beta_1 X + u$$

$$\frac{\partial W}{\partial X} = \beta_1$$

- The difference is that with quadratic terms, we can express cases where *marginal* changes to W with respect to X is not a constant, but depends on some value of X
  - In the above case, if $\beta_2 > 0$, marginal increase in W increases with X (and vice versa)