Econometrics V Lecture 7

Review of some points

1. Best Simple Linear Regression: Among the variables $X_1,\ldots,X_K$, the variable which best predicts Y based on a simple linear regression is the variable for which the simple linear regression has the highest R^2 (equivalent to the lowest sum of squared errors)

2. Multiple Linear Regression Model: The multiple linear regression model for the mean of Y given $X_1,\ldots,X_K$ is

$$E(Y \mid X_1,\ldots,X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K \qquad (1.1)$$

where β_j =partial slope on variable X_j =change in mean of Y for each one unit increase in X_j when the other variables $X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_K$ are held fixed.

The disturbance e_i for the multiple linear regression model is the difference between the actual Y_i and the mean of Y_i given $X_{i1},\ldots,X_{iK}$ for observation i:

$e_i = Y_i - E(Y_i \mid X_{i1},\ldots,X_{iK}) = Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK})$. In addition to (1.1), the multiple linear regression model makes the following assumptions about the disturbances e_i:

Econometrics V Lecture 7

(i) Linearity assumption: $E(e_i \mid X_{i1},\ldots,X_{iK}) = 0$. This implies that the linear model for the mean of Y given $X_1,\ldots,X_K$ is the correct model for the mean.

(ii) Constant variance assumption: $Var(e_i \mid X_{i1},\ldots,X_{iK}) = \sigma_e^2$. The disturbances e_i are assumed to all have the same variance σ_e^2.

(iii) Normality assumption: The disturbances e_i are assumed to have a normal distribution.

(iv) Independence assumption: The disturbances e_i are assumed to be independent.

Econometrics V Lecture 7

3. Partial slopes vs. Marginal slopes: The coefficient β_j on the variable X_j is a partial slope. It indicates the change in the mean of Y that is associated with a one unit increase in X_j when the other variables $X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_K$ are held fixed. The partial slope differs from the marginal slope that is obtained when we perform a simple regression of Y on X_j. The marginal slope measures the change in the mean of Y that is associated with a one unit increase in X_j, not holding the other variables $X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_K$ fixed.

Econometrics V Lecture 7

4. Least Squares Estimates of the Multiple Linear Regression Model: Based on a sample $(X_{11},\ldots,X_{iK},Y_1),\ldots,(X_{n1},\ldots,X_{nK},Y_n)$, we estimate the slopes and intercept by the least squares principle --

we minimize the sum of squared prediction errors in the data, $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - \cdots - b_KX_{iK})^2$. The least squares estimates of the intercept and the slopes are the $b_0, b_1, \ldots, b_K$ that minimize the sum of squared prediction errors.

Econometrics V Lecture 7

6. Using the Residuals to Check the Assumptions of the Multiple Linear Regression Model: For multiple regression, there are several residual plots. There is (1) the residual by predicted plot of the predicted values $\hat{E}(Y_i \mid X_{i1},\ldots,X_{iK})$ versus the residuals and (2) residual plots of each variable X_j versus the residuals. To check the linearity assumption, we check if $E(\hat{e}_i)$ is approximately zero for each part of the range of the predicted values and the variables $X_1,\ldots,X_K$ in the residual plots. To check the constant variance assumption, we check if the spread of the residuals remains constant as the predicted values and the variables $X_1,\ldots,X_K$ vary in the residual plots. To check the normality assumption, we check if the histogram of the residuals is approximately bell shaped. For now, we will not consider the independence assumption.

Econometrics V Lecture 7

7. Root Mean Square Error: The root mean square error (RMSE) is approximately the average absolute error that is made when using $E(Y_i \mid X_{i1},\ldots,X_{iK})$ to predict Y_i. The

8. Confidence Interval for the Slopes: The confidence interval for the slope β_j on the variable X_j is a range of plausible values for the true slope β_j based on the sample $(X_{11},\ldots,X_{iK},Y_1),\ldots,(X_{n1},\ldots,X_{nK},Y_n)$. The 95% confidence interval for the slope is $b_j \pm t_{.025,n-(K+1)}s_{b_j}$, where s_{b_j} is the standard error of the slope obtained. The 95% confidence interval for the slope is approximately $b_j \pm 2s_{b_j}$.

Econometrics V Lecture 7

9. Hypothesis Testing for the Slope: To test hypotheses for the slope β_j on the variable X_j, we use the t-statistic $t = \dfrac{b_j - \beta_j^*}{s_{b_j}}$ where β_j^* is detailed below.

(i) Two-sided test: $H_0 : \beta_j = \beta_j^*$ vs. $H_a : \beta_j \neq \beta_j^*$. We reject H_0 if $t > t_{.025,n-(K+1)}$ or $t < -t_{.025,n-(K+1)}$.

(ii) One-sided test I: $H_0 : \beta_1 \geq \beta_1^*$ vs. $H_a : \beta_1 < \beta_1^*$. We reject H_0 if $t < -t_{.05,n-(K+1)}$

(iii) One-sided test II: $H_0 : \beta_1 \leq \beta_1^*$ vs. $H_a : \beta_1 > \beta_1^*$. We reject H_0 if $t > t_{.05,n-(K+1)}$

When $\beta_1^* = 0$, we can calculate the p-values for these two tests as follows:

Econometrics V Lecture 7

(i) Two-sided test: the p-value is Prob>|t|
(ii) One-sided test I: If *t* is negative (i.e., the sign of the t-statistic is in favor of the alternative hypothesis), the p-value is (Prob>|t)/2. If *t* is positive (i.e., the sign of the t-statistic is in favor of the null hypothesis), the p-value is 1-(Prob>|t)/2.
(iii) One-sided test II: If *t* is positive (i.e., the sign of the t-statistic is in favor of the alternative hypothesis), the p-value is (Prob>|t)/2. If *t* is negative (i.e., the sign of the t-statistic is in favor of the null hypothesis), the p-value is 1-(Prob>|t)/2.

Note that the two-sided t-test is equivalent to the partial F test.

Econometrics V Lecture 7

10. R Squared and Assessing the Quality of Prediction: The R squared statistic measures how much of the variability in the response the regression model explains. R squared ranges from 0 to 1, with higher R squared values meaning that the regression model is explaining more of the variability in the response.

$$R^2 = \frac{\text{Total Sum of Squares-Residual Sum of Squares}}{\text{Total Sum of Squares}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Econometrics V Lecture 7

R squared is a measure of a fit of the regression to the sample data. It is not generally considered an adequate measure of the regression's ability to predict the responses for new observations. A strategy for assessing the ability of the regression to predict the responses for new observations is data splitting. We split the data into two groups – a training sample and a holdout sample. We fit the regression model to the training sample and then assess the quality of predictions of the regression model to the holdout sample:

Let n_2 be the number of points in the holdout sample.

Let $\hat{Y}_1,\ldots,\hat{Y}_{n_2}$ be the predictions of Y for the points in the holdout sample based on the model fit on the training sample and the explanatory variables for the observations in the holdout sample.

$$\text{Root Mean Squared Deviation (RMSD)} = \sqrt{\frac{\sum_{i=1}^{n_2}(Y_i-\hat{Y}_i)^2}{n_2}}$$

Econometrics V Lecture 7

11. Partial F tests for comparing two regression models: Consider the regression model $E(Y \mid X_1,\ldots,X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_L X_L + \beta_{L+1} X_{L+1} + \cdots + \beta_K X_K$

Suppose we want to test whether the variables $X_{L+1},\ldots,X_K$ are useful for predicting Y once the variables $X_1,\ldots,X_L$ have been taken into account, i.e., to test

$H_0 : \beta_{L+1} = \cdots = \beta_K = 0$ vs.

H_a : at least one of $\beta_{L+1},\ldots,\beta_K$ does not equal 0

We use the partial F test. We calculate the sum of squared errors for the full model:

$E(Y \mid X_1,\ldots,X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_L X_L + \beta_{L+1} X_{L+1} + \cdots + \beta_K X_K$

and the reduced model:

$E(Y \mid X_1,\ldots,X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_L X_L$ and

calculate the test statistic:

$$F = \frac{(SSE_R - SSE_F)/(K-L)}{SSE_F/(n-K-1)}$$

Econometrics V Lecture 7

where SSE_R =sum of squared errors for reduced model and SSE_F = sum of squared errors for full model. Our decision rule for the test is

Reject H_0 if $F \geq F(.05; K-L, n-K-1)$

Accept H_0 if $F < F(.05; K-L, n-K-1)$

To test the usefulness of the model (i.e., are any of the variables in the model useful for predicting Y, we set $K = L$ in the partial F test.

Econometrics V Lecture 7

12. Prediction Intervals: The best prediction for the Y_p of a new observation p with $X_1 = X_{p1}, \ldots, X_K = X_{pK}$ is the estimated mean of Y given $X_1 = X_{p1}, \ldots, X_K = X_{pK}$.

The 95% prediction interval for the Y_p of a new observation p with $X_1 = X_{p1}, \ldots, X_K = X_{pK}$ is an interval that will contain the value of Y_p most of the time. The formula for the prediction interval is :

$\hat{Y}_p \pm t_{.025,n-K-1} * s_p$, where

$\hat{Y}_p = \hat{E}(Y \mid X_1 = X_{p1}, \ldots, X_K = X_{pK}) = b_0 + b_1 X_{p1} + \cdots + b_K X_{pK}$ and

s_p is the standard error of predictions.

When n is large (say n>30), the 95% prediction interval is approximately equal to $\hat{Y}_p \pm 2 * RMSE$.

Econometrics V Lecture 7

13. Multicollinearity: A multiple regression suffers from multicollinearity when some of the explanatory variables are highly correlated with each other.

Consequence of multicollinearity: When there is multicollinearity, it often hard to determine the individual regression coefficients. The standard errors of the regression coefficients are large and the regression coefficients can sometimes have surprising signs.

Detecting multicollinearity: One sign that there is multicollinearity is that the F test for the usefulness of the model yields a large F statistic but the t statistics for the individual regression coefficients are all small. The primary approach we use to detecting multicollinearity is to look at the variance inflation factors (VIFs). The VIF on a variable X_j is the amount by which the variance of the coefficient on X_j in the multiple regression is multiplied compared to what the variance of the coefficient on X_j would be if all variables were uncorrelated.

Econometrics V Lecture 7

Multicollinearity and prediction: If interest is in predicting Y , as long as the pattern of relationships between the variables in the sample continues for those observations where forecasts are desired, multicollinearity is not particularly problematic. But if interest is in predicting Y for observations where the pattern of relationships between the variables is different than that in the sample, multicollinearity makes predictions unreliable because the predictions are extrapolations.

Econometrics V Lecture 7

14. Multiple Regression and Causal Inference: Suppose we want to estimate the causal effect on Y of increasing a variable X_j and keeping all other variables in the world fixed

A lurking variable is a variable that is associated with both Y and X_j. The slope on X_j in a multiple regression measures the causal effect if we include *all* lurking variables in the regression in addition to X_j. However, it is often difficult to include all lurking variables in the multiple regression. Omitted variables bias is the bias in estimating the causal effect of a variable that comes from omitting a lurking variable from the multiple regression.

Econometrics V Lecture 7

More on Multicollinearity?

- Multicollinearity is the condition where the independent variables are linearly related to each other.
- Variables need not be causally related, only correlated.
- X's are usually related to some extent, it is a matter of degree.

Econometrics V Lecture 7

Perfect Multicollinearity

- Recall to estimate b, the matrix $(X'X)^{-1}$ had to exist
- This meant that the matrix X had to be of full rank
- That is, none of the X's could be a perfect linear function of any combination of the other X's
- If so, then b is undefined

Econometrics V Lecture 7
: Definition

- If multicollinearity is not perfect, then $(X'X)^{-1}$ exists and analysis can proceed.
- But multicollinearity can cause problems even if correlations are not perfect.
- As the level of multicollinearity increases, the amount of *independent* information about the X's decreases
- Problem is insufficient information in the sample

Econometrics V Lecture 7

Multicollinearity: Implications

- Our only assumption in deriving b and showing it is BLUE was that $(X'X)^{-1}$ exits.
- Thus if multicollinearity is not perfect, then OLS is unbiased and is still BLUE
- But while b will have the least variance among unbiased linear estimators, its variance will *increase*

Econometrics V Lecture 7

Multicollinearity: Implications

- Recall our multivariate equation for b in scalar notation:

$$b_k = \frac{\sum (X_{kt} - \hat{X}_{kt}) Y_t}{\sum (X_{kt} - \hat{X}_{kt})^2}$$

- As multicollinearity increases, the correlation between X and X_{hat} increases

Econometrics V Lecture 7

Multicollinearity: Implications

- Reducing X's variation from X_{hat} is the same as reducing X's variation from X_{mean} in the bivariate model
- Recall the equation for the variance of b in the bivariate model:
- $\sigma_{hat\text{-}b}^2 = \sigma_{hat\text{-}u}^2 / \sum (X\text{-}X_{hat})^2$

Econometrics V Lecture 7

Multicollinearity: Implications

- Thus as the correlation between X and X_{hat} increases, the denominator for the variance of b decreases – increasing the variance of b
- Notice if X_1 is uncorrelated with $X_2 \ldots X_n$, then the formula is the same as in the bivariate case.

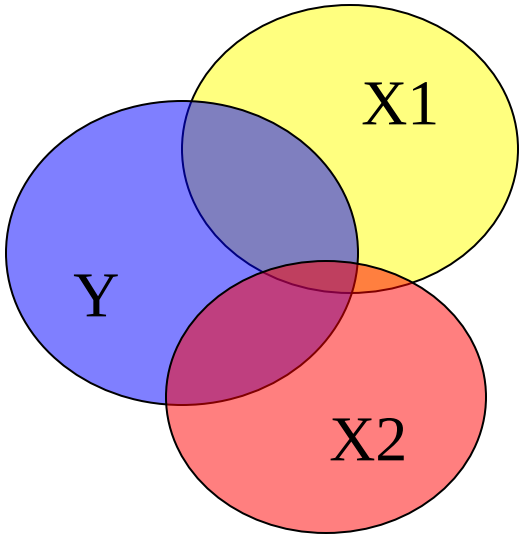
Econometrics V Lecture 7

Multicollinearity: Implications

- In practice, X's that are causing the same Y are often correlated to some extent
- If the correlation is high, it becomes difficult to distinguish the impact of X_1 on Y from the impact of $X_2 \ldots X_n$
- OLS estimates tend to be sensitive to small changes in the data.

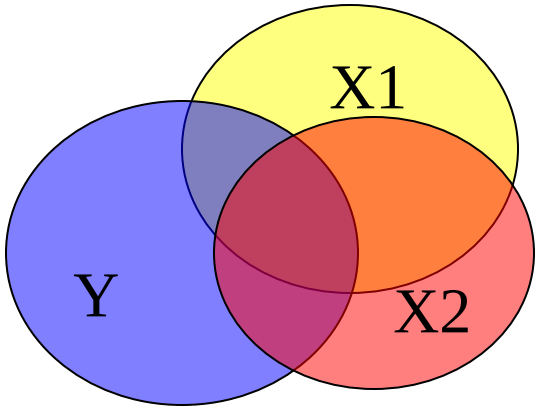
Econometrics V Lecture 7
Illustrating Multicollinearity

- When correlation among X's is low, OLS has lots of information to estimate b
- This gives us confidence in our estimates of b



Econometrics V Lecture 7
Illustrating Multicollinearity

- When correlation among X's is high, OLS has very little information to estimate b
- This makes us relatively uncertain about our estimate of b



Econometrics V Lecture 7

Multicollinearity: Causes

- Poorly constructed sampling design causes correlation among X's
- Poorly constructed measures over-aggregate information and make cases correlate
- Statistical model specification: adding polynomial terms or trend indicators.

Econometrics V Lecture 7

Multicollinearity: Causes

- Too many variables in the model – X's measure the same conceptual variable.
- X's are causally related to one another

Econometrics V Lecture 7
Multicollinearity: Warning Signs

- F-test is significant but coefficients are not.
- Coefficients are substantively large but statistically insignificant
- Standard errors of b's change when other variables included or removed, but estimated value of b does not

Econometrics V Lecture 7

Multicollinearity: Warning Signs

- Multicollinearity could be a problem any time that a coefficient is not statistically significant
- Should always check analyses for multicollinearity levels
- If coefficients are significant, then multicollinearity is not a problem.
- Its only effect is to increase the σ_b^2

Econometrics V Lecture 7

Multicollinearity: The Diagnostic

- Diagnostic of multicollinearity is the auxiliary r-squared
- Regress each X on all other X's in the model
- R-squared will show you linear correlation between each X and all other X's in the model

Econometrics V Lecture 7

Multicollinearity: The Diagnostic

- There is no definitive threshold when the auxiliary r-squared is too high
  - Depends on whether b is significant
- Tolerance for multicollinearity depends on N
  - Larger N means more information
- If aux-R^2 is above .9 and b is insignificant, you should be concerned
  - If N is small then .7 or .8 may be too high

Multicollinearity: Remedies

- Increase sample size to get more information
- Change sampling mechanism to allow greater variation in X's
- Change unit of analysis to allow more cases and more variation in X's
- correlations

Econometrics V Lecture 7

Multicollinearity: Remedies

- Disaggregate measures to capture independent variation
- Create a composite scale or index if variables measure the same concept
- Construct measures to avoid correlations