UNIVERSITY OF CAMBRIDGE
Judge Business School

Updated: November 17, 2011

Lecturer: Thilo Klein
Contact: tk375@cam.ac.uk

Contest Quiz 3

Question Sheet

In this quiz we will review concepts of linear regression covered in lecture 2.

NOTE: Please round your results to *two decimal places.*

EXAMPLE: If your unrounded solution is 0.13897439, drop all decimal places except the first three. This leaves you with 0.138. If the third decimal place is 5 or above (as is the case here), round up. This gives 0.14.

Question 1: Simple linear regression

(i) Consider the linear model $Y_i = \beta_0 + \beta_1 X_i + u_i$. The variance of Y_i is given by

- (a) $\beta_0^2 + \beta_1^2 var(X_i) + var(u_i)$.
- (b) the variance of u_i.
- (c) $\beta_1^2 var(X_i) + var(u_i)$.
- (d) the variance of the residuals.

(ii) The sample average of the OLS residuals is

- (a) some positive number since OLS uses squares.
- (b) zero.
- (c) unobservable since the population regression function is unknown.
- (d) dependent on whether the explanatory variable is mostly positive or negative.

(iii) The slope estimator, $\hat{\beta}_1$, has a smaller standard error, other things equal, if

- (a) there is more variation in the explanatory variable, X.
- (b) there is a large variance of the error term, u.
- (c) the sample size is smaller.
- (d) the intercept, β_0, is small.

(iv) The regression R^2 is a measure of

- (a) whether or not X causes Y.
- (b) the goodness of fit of your regression line.
- (c) whether or not $ESS > TSS$.
- (d) the square of the determinant of R.

(v) To decide whether or not the slope coefficient is large or small,

  (a) you should analyse the economic importance of a given increase in X.
  (b) the slope coefficient must be larger than one.
  (c) the slope coefficient must be statistically significant.
  (d) you should change the scale of the X variable if the coefficient appears to be too small.

(vi) Multiplying the dependent variable by 100 and the explanatory variable by 100,000 leaves the

  (a) OLS estimate of the slope the same.
  (b) OLS estimate of the intercept the same.
  (c) regression R^2 the same.
  (d) variance of the OLS estimators the same.

(vii) In which of the following relationships does the intercept have a real-world interpretation?

  (a) the relationship between the change in the unemployment rate and the growth rate of real GDP ("Okun's Law")
  (b) the demand for coffee and its price
  (c) test scores and class-size
  (d) weight and height of individuals

(viii) Changing the units of measurement, e.g. measuring testscores in 100s, will do all of the following EXCEPT for changing the

  (a) residuals
  (b) numerical value of the slope estimate
  (c) interpretation of the effect that a change in X has on the change in Y
  (d) numerical value of the intercept

Question 2: Hypothesis tests and confidence intervals

(i) When estimating a demand function for a good where quantity demanded is a linear function of the price, you should

  (a) not include an intercept because the price of the good is never zero.
  (b) use a one-sided alternative hypothesis to check the influence of price on quantity.
  (c) use a two-sided alternative hypothesis to check the influence of price on quantity.
  (d) reject the idea that price determines demand unless the coefficient is at least 1.96.

(ii) The confidence interval for the sample regression function slope

  (a) can be used to conduct a test about a hypothesized population regression function slope.
  (b) can be used to compare the value of the slope relative to that of the intercept.
  (c) adds and subtracts 1.96 from the slope.
  (d) allows you to make statements about the economic importance of your estimate.

(iii) Under the least squares assumptions (zero conditional mean for the error term, X_i and Y_i being *i.i.d.*, and X_i and u_i having finite fourth moments), the OLS estimator for the slope and intercept

(a) has an exact normal distribution for $n > 15$.

(b) is BLUE.

(c) has a normal distribution even in small samples.

(d) is unbiased.

(iv) Consider the following regression line: $\widehat{TestScore} = 698.9 - 2.28 * STR$. You are told that the t-statistic on the slope coefficient is 4.38. What is the standard error of the slope coefficient?

(v) The construction of the t-statistic for a one- and a two-sided hypothesis

(a) depends on the critical value from the appropriate distribution.

(b) is the same.

(c) is different since the critical value must be 1.645 for the one-sided hypothesis, but 1.96 for the two-sided hypothesis (using a 5% probability for the Type I error).

(d) uses ± 1.96 for the two-sided test, but only $+1.96$ for the one-sided test.

(vi) The only difference between a one- and two-sided hypothesis test is

(a) the null hypothesis.

(b) dependent on the sample size n.

(c) the sign of the slope coefficient.

(d) how you interpret the t-statistic.

(vii) Using 143 observations, assume that you had estimated a simple regression function and that your estimate for the slope was 0.04, with a standard error of 0.01. You want to test whether or not the estimate is statistically significant. Which of the following possible decisions is the only correct one:

(a) you decide that the coefficient is small and hence most likely is zero in the population

(b) the slope is statistically significant since it is four standard errors away from zero

(c) the response of Y given a change in X must be economically important since it is statistically significant

(d) since the slope is very small, so must be the regression R^2.

Question 3: Multiple regression models

(i) In the multiple regression model, the adjusted R^2, $\bar{R}^2$

(a) cannot be negative.

(b) will never be greater than the regression R^2.

(c) equals the square of the correlation coefficient $r_{Y,\hat{Y}}$.

(d) cannot decrease when an additional explanatory variable is added.

(ii) Under imperfect multicollinearity

   (a) the OLS estimator cannot be computed.
   (b) two or more of the regressors are highly correlated.
   (c) the OLS estimator is biased even in samples of $n > 100$.
   (d) the error terms are highly, but not perfectly, correlated.

(iii) When there are omitted variables in the regression, which are determinants of the dependent variable, then

   (a) you cannot measure the effect of the omitted variable, but the estimator of your included variable(s) is (are) unaffected.
   (b) this has no effect on the estimator of your included variable because the other variable is not included.
   (c) this will always bias the OLS estimator of the included variable.
   (d) the OLS estimator is biased if the omitted variable is correlated with the included variable.

(iv) Imagine you regressed earnings of individuals on a constant, a binary variable (*Male*) which takes on the value 1 for males and is 0 otherwise, and another binary variable (*Female*) which takes on the value 1 for females and is 0 otherwise. Because females typically earn less than males, you would expect

   (a) the coefficient for *Male* to have a positive sign, and for *Female* a negative sign.
   (b) both coefficients to be the same distance from the constant, one above and the other below.
   (c) none of the OLS estimators to exist because there is perfect multicollinearity.
   (d) this to yield a difference in means statistic.

(v) When you have an omitted variable problem, the assumption that $E(u_i X_i) = 0$ is violated. This implies that

   (a) the sum of the residuals is no longer zero.
   (b) there is another estimator called weighted least squares, which is BLUE.
   (c) the sum of the residuals times any of the explanatory variables is no longer zero.
   (d) the OLS estimator is no longer consistent.

(vi) In the multiple regression model you estimate the effect on Y_i of a unit change in one of the X_i while holding all other regressors constant. This

   (a) makes little sense, because in the real world all other variables change.
   (b) corresponds to the economic principle of *mutatis mutandis*.
   (c) leaves the formula for the coefficient in the single explanatory variable case unaffected.
   (d) corresponds to taking a partial derivative in mathematics.

(vii) You have to worry about perfect multicollinearity in the multiple regression model because

(a) many economic variables are perfectly correlated.
(b) the OLS estimator is no longer BLUE.
(c) the OLS estimator cannot be computed in this situation.
(d) in real life, economic variables change together all the time.

(viii) The intercept in the multiple regression model

(a) should be excluded if one explanatory variable has negative values.
(b) determines the height of the regression line.
(c) should be excluded because the population regression function does not go through the origin.
(d) is statistically significant if it is larger than 1.96.

(ix) The following OLS assumption is most likely violated by omitted variables bias:

(a) $E(u_i|X_i) = 0$
(b) (X_i, Y_i), $i = 1, ..., n$ are *i.i.d* draws from their joint distribution
(c) there are no outliers for X_i, u_i
(d) there is heteroskedasticity

(x) In multiple regression, the R^2 increases whenever a regressor is

(a) added unless the coefficient on the added regressor is exactly zero.
(b) added.
(c) added unless there is heterosckedasticity.
(d) greater than 1.96 in absolute value.

(xi) Consider the following multiple regression models (a) to (d) below. $DFemme = 1$ if the individual is a female, and is zero otherwise; *DMale* is a binary variable which takes on the value one if the individual is male, and is zero otherwise; *DMarried* is a binary variable which is unity for married individuals and is zero otherwise, and *DSingle* is $(1 - DMarried)$. Regressing weekly earnings (*Earn*) on a set of explanatory variables, you will experience perfect multicollinearity in the following cases unless:

(a) $\widehat{Earn}_i = \hat{\beta}_0 + \hat{\beta}_1 DFemme + \hat{\beta}_2 Dmale + \hat{\beta}_3 X_{3i}$
(b) $\widehat{Earn}_i = \hat{\beta}_0 + \hat{\beta}_1 DMarried + \hat{\beta}_2 DSingle + \hat{\beta}_3 X_{3i}$
(c) $\widehat{Earn}_i = \hat{\beta}_0 + \hat{\beta}_1 DFemme + \hat{\beta}_3 X_{3i}$
(d) $\widehat{Earn}_i = \hat{\beta}_1 DFemme + \hat{\beta}_2 Dmale + \hat{\beta}_3 DMarried + \hat{\beta}_4 DSingle + \hat{\beta}_5 X_{3i}$

(xii) Consider the multiple regression model with two regressors X_1 and X_2, where both variables are determinants of the dependent variable. When omitting X_2 from the regression, then there will be omitted variable bias for $\hat{\beta}_1$

(a) if X_1 and X_2 are correlated
(b) always
(c) if X_2 is measured in percentages
(d) if X_2 is a dummy variable

(xiii) Consider the multiple regression model with two regressors X_1 and X_2, where both variables are determinants of the dependent variable. You first regress Y on X_1 only and find no relationship. However when regressing Y on X_1 and X_2, the slope coefficient $\hat{\beta}_1$ pertaining to X_1 changes by a large amount. This suggests that your first regression suffers from

(a) heteroskedasticity

(b) perfect multicollinearity

(c) omitted variable bias

(d) dummy variable trap

(xiv) Imperfect multicollinearity

(a) implies that it will be difficult to estimate precisely one or more of the partial effects using the data at hand

(b) violates one of the four Least Squares assumptions in the multiple regression model

(c) means that you cannot estimate the effect of at least one of the Xs on Y

(d) suggests that a standard spreadsheet program does not have enough power to estimate the multiple regression model

Question 4: Hypothesis tests in multiple regression

(i) The following linear hypothesis can be tested using the F-test with the exception of

(a) $\beta_2 = 1$ and $\beta_3 = \beta_4/\beta_5$.

(b) $\beta_2 = 0$.

(c) $\beta_1 + \beta_2 = 1$ and $\beta_3 = -2\beta_4$.

(d) $\beta_0 = \beta_1$ and $\beta_1 = 0$.

(ii) When testing joint hypothesis, you should

(a) use t-statistics for each hypothesis and reject the null hypothesis if all of the restrictions fail.

(b) use the F-statistic and reject all the hypothesis if the statistic exceeds the critical value.

(c) use t-statistics for each hypothesis and reject the null hypothesis once the statistic exceeds the critical value for a single hypothesis.

(d) use the F-statistics and reject at least one of the hypothesis if the statistic exceeds the critical value.

(iii) The overall regression F-statistic tests the null hypothesis that

(a) all slope coefficients are zero.

(b) all slope coefficients and the intercept are zero.

(c) the intercept in the regression and at least one, but not all, of the slope coefficients is zero.

(d) the slope coefficient of the variable of interest is zero, but that the other slope coefficients are not.

(iv) For a single restriction, the F-statistic

(a) is the square root of the t-statistic.
(b) has a critical value of 1.96.
(c) will be negative.
(d) is the square of the t-statistic.

(v) Let $R^2_{unrestricted}$ and $R^2_{restricted}$ be 0.4366 and 0.4149 respectively. The difference between the unrestricted and the restricted model is that you have imposed two restrictions. That is, there are 3 regressors (excluding the intercept) in the unrestricted model and 1 regressor in the restricted model. There are 420 observations. What is the F-statistic in this case?

(vi) If you reject a joint null hypothesis using the F-test in a multiple hypothesis setting, then

(a) a series of t-tests may or may not give you the same conclusion.
(b) the regression is always significant.
(c) all of the hypotheses are always simultaneously rejected.
(d) the F-statistic must be negative.

(vii) A 95% confidence set for two or more coefficients is a set that contains

(a) the sample values of these coefficients in 95% of randomly drawn samples.
(b) integer values only.
(c) the same values as the 95% confidence intervals constructed for the coefficients.
(d) the population values of these coefficients in 95% of randomly drawn samples.

(viii) When testing the null hypothesis that two regression slopes are zero simultaneously, then you cannot reject the null hypothesis at the 5% level, if the confidence ellipse contains the point

(a) $(-1.96, 1.96)$.
(b) $(0, 1.96)$.
(c) $(0, 0)$.
(d) $(1.96^2, 1.96^2)$.

(ix) All of the following are true, with the exception of one condition:

(a) a high R^2 or $\bar{R}^2$ does not mean that the regressors are a true cause of the dependent variable.
(b) a high R^2 or $\bar{R}^2$ does not mean that there is no omitted variable bias.
(c) a high R^2 or $\bar{R}^2$ always means that an added variable is statistically significant.
(d) a high R^2 or $\bar{R}^2$ does not necessarily mean that you have the most appropriate set of regressors.

(x) Consider a regression with two variables, in which X_{1i} is the variable of interest and X_{2i} is the control variable. Conditional mean independence requires

(a) $E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{2i})$
(b) $E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{1i})$
(c) $E(u_i|X_{1i}) = E(u_i|X_{2i})$
(d) $E(u_i) = E(u_i|X_{2i})$

(xi) What is the critical value of $F_{4,\infty}$ at the 5% significance level?