Information on the Final Exam of ECON3334
Fall, 2019, HKUST

1. Duration: **3 hours.**

2. The format of the final will be similar to that of the mid-term with more questions.

3. What to bring to the exam: (1) pens or pencils (2) **hand-written note** (ONE piece of A4-size paper on which you can write on **both sides**) (3) a basic (non-programmable) calculator (4) your student ID.

4. Required readings after the mid-term:
   Chapters 5 (excluding section 5.6, "Using the t-Statistic in Regression When the Sample size is Small"), Chapter 6, and Chapter 7 in Stock and Watson and lecture notes. Note that Topic 8 (Nonlinear regression function) and Topic 9 (A guide for empirical studies) in the syllabus will not be required.

   Required reading for the final:
   The required reading above AND the required reading specified in the Reviews for the midterm exam. The final exam is cumulative. Emphasis will be on the materials that are covered in class.

   The mathematical appendices are not required, though reading them might be helpful. In the exam, you should be able to solve all the problems without reading the mathematical appendices.

   Stata knowledge is not required for the final exam.

5. Please write your answers in the exam clearly. If the Peter can read what you write, and can follow your work easily, he will likely be a much nicer grader.

6. The rest of this document is a quick summary of key points for the materials after the mid-term. Again this is just a quick summary and is NOT a substitute for the required reading.

HAPPY STUDYING and GOOD LUCK WITH ALL YOUR FINALS!

A quick summary of key points for the materials after the midterm

There are five topics we have covered after the midterm exam: inference of simple regression, estimation of multiple regression, and inference of multiple regression.

Simple Regression: Inference

1. Hypothesis testing

- The objective is to test a hypothesis, like $\beta_1 = 0$, using data to reach a tentative conclusion whether the (null) hypothesis is correct or incorrect.
  - Null hypothesis and two-sided alternative: H_0: $\beta_1 = \beta_{1,0}$ vs. H_1: $\beta_1 \neq \beta_{1,0}$, where $\beta_{1,0}$ is the hypothesized value under the null.
  - Null hypothesis and one-sided alternative: H_0: $\beta_1 = \beta_{1,0}$ vs. H_1: $\beta_1 < \beta_{1,0}$
- General approach: construct t-statistic, and do one of three steps: (1) compare to N(0,1) critical value or, (2) compute p-value or, (3) construct the confidence interval. After calculating the t-statistics, the rest details of three steps are exactly the same as what we did in Chapter 3.
- In general: $\boxed{t = \frac{\text{estimator - hypothesized value}}{\text{standard error of the estimator}}}$, where the *SE* of the estimator is the square root of an estimator of the variance of the estimator.

  For testing β_1, $\boxed{t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}}$, where SE($\hat{\beta}_1$) = the square root of an estimator of the variance of the sampling distribution of $\hat{\beta}_1$.
- Formula for SE($\hat{\beta}_1$) is a bit nasty. Basically it is an estimator of $\sqrt{\text{var}(\hat{\beta}_1)}$. Stata automatically reports it.

2. Confidence interval

- Recall that a 95% confidence is, equivalently:
  - The set of points that cannot be rejected at the 5% significance level;
  - A set-valued function of the data (an interval that is a function of the data) that contains the true parameter value 95% of the time in repeated samples.
- Because the t-statistic for β_1 is N(0,1) in large samples, construction of a 95% confidence for β_1 is just like the case of the sample mean. 95% confidence interval for β_1= $\{\hat{\beta}_1 \pm 1.96 * SE(\hat{\beta}_1)\}$

3. Binary X

- Sometimes a regressor is binary:
  - X = 1 if small class size, = 0 if not
  - X = 1 if female, = 0 if male
  - X = 1 if treated (experimental drug), = 0 if not

  Binary regressors are sometimes called "dummy" variables.
- So far, β_1 has been called a "slope," but that doesn't make sense if X is binary. How do we interpret regression with a binary regressor? Interpreting regressions with a binary regressor:

  $Y_i = \beta_0 + \beta_1 X_i + u_i$, where X is binary (X_i = 0 or 1):
  - When $X_i = 0$, $Y_i = \beta_0 + u_{i,}$, the mean of Y_i is $\beta_{0,}$ that is, $E(Y_i|X_i=0) = \beta_0$
  - When $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$, the mean of Y_i is $\beta_0 + \beta_1$, that is, $E(Y_i|X_i=1) = \beta_0 + \beta_1$
  - $\beta_1 = E(Y_i|X_i=1) – E(Y_i|X_i=0)$ = population difference in group means

- t-statistics, confidence intervals are constructed as usual. This is another way (an easy way) to do difference-in-means analysis.

4. Heteroskedasticity and homoskedasticity

- If var(u|X=x) is constant – that is, if the variance of the conditional distribution of u given X does not depend on X – then u is said to be homoskedastic. Otherwise, u is heteroskedastic.
- There are two types of standard errors:
  1) Homoskedasticity-only standard errors – these are valid only if the errors are homoskedastic.
  2) The usual standard errors –it is conventional to call these heteroskedasticity – robust standard errors (or White standard error), because they are valid whether or not the errors are heteroskedastic.
- The main advantage of the homoskedasticity-only standard errors is that the formula is simpler. But the disadvantage is that the formula is only correct in general if the errors are homoskedastic.
- If the errors are either homoskedastic or heteroskedastic and you use heteroskedastic-robust standard errors, you are OK. If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, your standard errors will be wrong (the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$ is inconsistent if there is heteroskedasticity). So, in practice you should always use heteroskedasticity-robust standard errors.

5. Gauss-Markov Theorem

- In addition to the three OLS assumptions (i) E(ui|Xi)=0; (ii) (Xi, Yi) are iid and (iii) outliers are unlikely, if we further assume (iv) var (ui|Xi)= constant (homoskedasticity), then OLS estimaors has the least variance (most efficient, best) among all other possible linear estimators. I.e., OLS is BLUE under (i), (ii), (iii) and (iv).

Multiple Regression – Estimation

1. Omitted variable bias (OVB)

- The error u arises because of factors that influence Y but are not included in the regression function; so, there are always omitted variables. Sometimes, the omission of those variables can lead to bias in the OLS estimator.
- The bias in the OLS estimator that occurs as a result of an omitted factor is called omitted variable bias. For omitted variable bias to occur, the omitted factor "X_2" must be:
  1) A determinant of Y (i.e. X_2 is part of u); and
  2) Correlated with the regressor X (*i.e.* corr(X_2,X) $\neq$ 0)

  Both conditions must hold for the omission of X_2 to result in omitted variable bias.
- Sometime, we can judge the direction of OVB. Consider a simple case where the true model is

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

  But we instead run

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \tilde{u}_i, \text{ where } \tilde{u}_i = \beta_2 X_{2i} + u_i$$

  In this case, we can show that

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\beta_2 \frac{cov(X_{1i}, X_{2i})}{var(X_{1i})}}_{OVB}$$

In this context, the sign or direction of OVB is completely determined by β_2 and covariance between X_1 and X_2.

2. Multiple regression model

- Consider a special case of two regressors:

  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \; i = 1,\ldots,n$

  - Y is the dependent variable
  - X_1, X_2 are the two independent variables (regressors)
  - (Y_i, X_{1i}, X_{2i}) denote the i^{th} observation on Y, X_1, and X_2.
  - β_0 = unknown population intercept
  - β_1 = effect on Y of a change in X_1, holding X_2 constant
  - β_2 = effect on Y of a change in X_2, holding X_1 constant
  - u_i = the regression error (omitted factors)

3. OLS estimators in multiple regression

- With two regressors, the OLS estimator solves:

$$\min_{b_0, b_1, b_2} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus. This yields the OLS estimators of β_0, β_1 and β_2.

4. Measure of fit in multiple regression

- Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual
- We have three measures of fit:
  1) SER = sample std. deviation of $\hat{u}_i$ (with d.f. correction)
  2) R^2 = fraction of variance of Y explained by X.
  3) $\bar{R}^2$ = “adjusted R^2” = R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$
- The R^2 always increases when you add another regressor – a bit of a problem for a measure of “fit”. The (the “adjusted R^2”) corrects this problem by “penalizing” you for including another regressor – the $\bar{R}^2$ does not necessarily increase when you add another regressor.

  Adjusted R^2: $\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1} \right) \frac{SSR}{TSS}$.

  Note that $\bar{R}^2 < R^2$, however if n is large the two will be very close.

5. OLS assumptions in multiple regression

- Four Assumptions:
  1) the conditional mean of u given the included X's is zero: $E(u|X_1 = x_1,\dots, X_k = x_k) = 0$
     - This has the same interpretation as in regression with a single regressor.
     - If an omitted variable (1) belongs in the equation (so is in u) and (2) is correlated with an included X, then this condition fails
     - Failure of this condition leads to omitted variable bias
     - The solution – if possible – is to include the omitted variable in the regression.
  2) $(X_{1i},\dots,X_{ki}, Y_i)$, $i = 1,\dots,n$, are i.i.d.
     - This is satisfied automatically if the data are collected by simple random sampling.
  3) large outliers are rare (finite fourth moments)
     - This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).
  4) There is no perfect multicollinearity
     - Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

6. Distribution of OLS estimators in multiple regressions

- Conceptually, there is nothing new. We can show that each $\hat{\beta}_1, \hat{\beta}_2,\dots, \hat{\beta}_k$ is normally distributed:

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$$

where $E\left(\hat{\beta}_j\right) = \beta_j$, and $var\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j}$

7. Multicollinearity

- Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.
  - Example 1: you include STR twice. You have to drop one of them.
  - Example 2: dummy trap. You include a constant, a female dummy and male dummy as regressors. You have to drop one of them.
  - Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- Imperfect multicollinearity occurs when two or more regressors are very highly correlated.
  - Why this term? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are collinear – but unless the correlation is exactly ±1, that collinearity is imperfect.
  - Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.
  - Intuition: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant – so the data are pretty much uninformative about what happens when X_1 changes but X_2 doesn't, so the variance of the OLS estimator of the coefficient on X_1 will be large.

- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.

Multiple regression: inferences

1. Hypothesis testing and confidence interval for a single coefficient

■ In the multiple regression, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$, $i = 1,…,n,$ sometimes we are interested in a single coefficient, say β_1. We can conduct hypothesis testing for H$_0$*:* $\beta_1 = \beta_1^*$ or construct a confidence interval for β_1.

■ Similar to the simple regression, one can show that $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\mathrm{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ by using CLT. Thus hypotheses on β_1 can be tested using the usual *t*-statistic: $t = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)}$, where $SE(\hat{\beta}_1)$ is the standard error, the estimator for $\sqrt{\mathrm{var}(\hat{\beta}_1)}$. The confidence interval is constructed as: $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$. So too for $\beta_2,…, \beta_k$.

■ $\hat{\beta}_1$ and $\hat{\beta}_2$ are generally not independently distributed – so neither are their *t*-statistics.

2. Tests of joint hypothesis

■ A joint hypothesis specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients. For example, H$_0$: $\beta_1 = 0$ and $\beta_2 = 0$. vs. H$_1$: either $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or both. In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.

■ A "common sense" idea is to reject if either of the individual *t*-statistics exceeds 1.96 in absolute value. But this "one at a time" test isn't valid: the resulting test rejects too often under the null hypothesis (more than 5%)! There are two solutions to this problem. First, use a different critical value in this procedure – not 1.96 (this is the "Bonferroni method) (this method is rarely used in practice however and not required for our course). Second, use a different test statistic that test both β_1 and β_2 at once: the *F*-statistic (this is common practice).

■ The *F*-statistic tests all parts of a joint hypothesis at once. Formula for the special case of the joint hypothesis $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*$ in a regression with two regressors:

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

where $\hat{\rho}_{t_1,t_2}$ estimates the correlation between t$_1$ and t$_2$. We reject when *F*-statistic is large. It is easy to see that the *F*-statistic is large when t_1 and/or t_2 is large. The *F*-statistic corrects (in just the right way) for the correlation between t_1 and t_2. The formula for more than two β's is nasty unless you use matrix algebra.

- We reject when *F*-statistic is large. The question is how large is large enough. We can show that under the null, the *F*-statistic follows a $F_{q,\infty}$ distribution, or equivalently, $\frac{\chi^2}{q}$ distribution, where χ^2 is the chi-squared distribution with *q* degrees of freedom. The critical values are obtained from the $F_{q,\infty}$ distribution. We can also compute the *p*-value using the *F*-statistic: *p*-value = tail probability of the $\frac{\chi^2}{q}$ distribution beyond the *F*-statistic actually computed.
- When the errors are homoskedastic, there is a simple formula for computing the "homoskedasticity-only" *F*-statistic:
  - Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
  - Compare the fits of the regressions – the R^2's – if the "unrestricted" model fits sufficiently better, reject the null

  For example, are the coefficients on *STR* and *Expn* zero? Then,
  - Unrestricted regression (under H_1): $TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$
  - Restricted regression (that is, under H_0): $TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i$
  - The number of restrictions under H_0 is $q = 2$.
  - The fit will be better (R^2 will be higher) in the unrestricted regression.

  By how much must the R^2 increase for the coefficients on *Expn* and *PctEL* to be judged statistically significant? We look at *F*-statistic. Under the assumption of homoskedasticity, we have a simple formula for the homoskedasticity-only *F*-statistic:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

where:

$R^2_{restricted}$ = the R^2 for the restricted regression

$R^2_{unrestricted}$ = the R^2 for the unrestricted regression

q = the number of restrictions under the null

$k_{unrestricted}$ = the number of regressors in the unrestricted regression.

The bigger the difference between the restricted and unrestricted R^2's – the greater the improvement in fit by adding the variables in question – the larger is the homoskedasticity-only *F*.

If the errors are homoskedastic, then the homoskedasticity-only *F*-statistic has a large-sample distribution that is $\frac{\chi^2}{q}$. But if the errors are heteroskedastic, the large-sample distribution is a mess and is not $\frac{\chi^2}{q}$.

3. Testing single restrictions involving multiple coefficient

- Consider the null and alternative hypothesis: H_0: $\beta_1 = \beta_2$ vs. H_1: $\beta_1 \neq \beta_2$. This null imposes a single restriction ($q = 1$) on multiple coefficients – it is not a joint hypothesis with multiple restrictions (compare with $\beta_1 = 0$ and $\beta_2 = 0$).
- Here are two methods for testing single restrictions on multiple coefficients:
  - Rearrange ("transform") the regression: Rearrange the regressors so that the restriction becomes a restriction on a single coefficient in an equivalent regression; or,
  - Perform the test directly: Some software, including Stata, lets you test restrictions using multiple coefficients directly
- Method 1: Rearrange ("transform") the regression. For example, consider the question:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

We add and subtract $\beta_2 X_{1i}$:

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

$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$ where $\gamma_1 = \beta_1 - \beta_2$, $W_i = X_{1i} + X_{2i}$

Thus, the following two systems are equivalent.

(a) Original system:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0: \beta_1 = \beta_2 \text{ vs. } H_1: \beta_1 \neq \beta_2$$

(b) Rearranged ("transformed") system:

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i, \text{ where } \gamma_1 = \beta_1 - \beta_2 \text{ and } W_i = X_{1i} + X_{2i}$$

$$H_0: \gamma_1 = 0 \text{ vs. } H_1: \gamma_1 \neq 0.$$

The testing problem is now a simple one: test whether $\gamma_1 = 0$ in specification (b).

- Method 2: consider the example above; in Stata, you can just write down: test (X_1=0) (X_2=0)

4. Confidence sets for multiple coefficients

- In the multiple regression: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$, $i = 1,\ldots,n,$ sometimes we are interested in a joint confidence set for β_1 and β_2? A 95% joint confidence set is:
  - A set-valued function of the data that contains the true parameter(s) in 95% of hypothetical repeated samples.
  - The set of parameter values that cannot be rejected at the 5% significance level.
- You can find a 95% confidence set as the set of (β_1, β_2) that cannot be rejected at the 5% level using an F-test. Let $F(\beta_1^*, \beta_2^*)$ be the (heteroskedasticity-robust) F-statistic testing the hypothesis that $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*$: 95% confidence set = $\{\beta_1^*, \beta_2^*: F(\beta_1^*, \beta_2^*) < 3.00\}$
  - 3.00 is the 5% critical value of the distribution $F_{2,\infty}$
  - This set has coverage rate 95% because the test on which it is based (the test it "inverts") has size of 5%. 5% of the time, the test incorrectly rejects the null when the null is true, so 95% of the

time it does not; therefore the confidence set constructed as the nonrejected values contains the true value 95% of the time (in 95% of all samples).

- The confidence set based on the *F*-statistic is an ellipse.

5. Model specification

■ In the tests-score/class size example, we want to get an unbiased estimate of the effect on test scores of changing class size, holding constant student and school characteristics. To do this we need to think about what variables to include and what regressions to run – and we should do this before we actually sit down at the computer. This entails thinking beforehand about your model specification.

■ A general approach to variable selection and "model specification" :

- Specify a "base" or "benchmark" model.
- Specify a range of plausible alternative models, which include additional candidate variables.
- Does a candidate variable change the coefficient of interest (β_1)?
- Is a candidate variable statistically significant?
- Use judgment, not a mechanical recipe…
- Don't just try to maximize R^2!

■ It is easy to fall into the trap of maximizing the R^2 and adjusted R^2

- But this loses sight of our real objective, an unbiased estimator of the class size effect.
- A high R^2 (or adjusted R^2) means that the regressors explain the variation in Y
- A high R^2 (or adjusted R^2) does not mean that you have eliminated omitted variable bias.
- A high R^2 (or adjusted R^2) does not mean that you have an unbiased estimator of a causal effect (β_1).
- A high R^2 (or adjusted R^2) does not mean that the included variables are statistically significant – this must be determined using hypotheses tests.