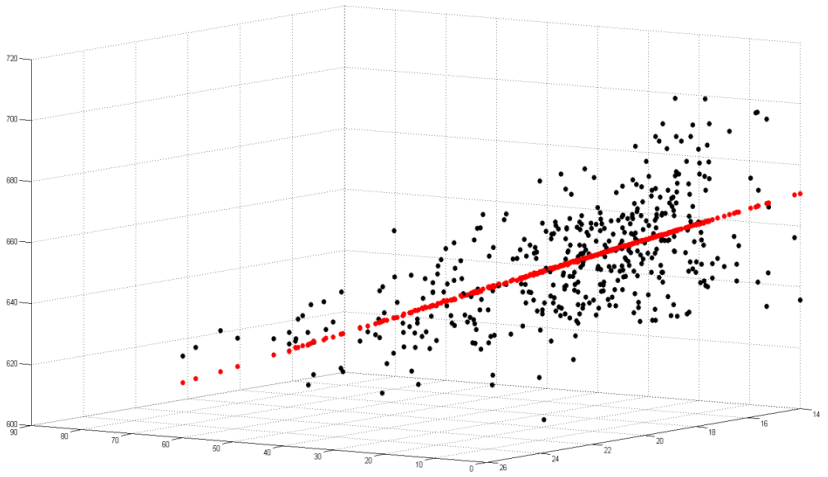
EC 3303: Econometrics I

Linear Regression with Multiple Regressors (Part 2)



Kelvin Seah

AY 2022/2023, Semester 2

Outline

1. Measures of fit
2. Least squares assumptions
3. Sampling distribution of the OLS estimator
4. Multicollinearity
5. Hypothesis tests & confidence intervals for a single coefficient

Measures of Fit in Multiple Regression

- 3 statistics which measure how well the OLS regression line describes or fits the data:
  - *regression* R^2
  - *adjusted* R^2 ($\bar{R}^2$)
  - *standard error of the regression* (*SER*)

R^2

- R^2 is the fraction of the variance of Y that is explained by the OLS regression (of Y on $X_1, X_2, \ldots, X_k$)

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2;\ TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2;\ SSR = \sum_{i=1}^{n} \hat{u}_i{}^2$$

- ***R^2 always increases*** when you add another regressor, unless the estimated coefficient on the added regressor is ***exactly zero***.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- ***R^2 always increases*** when you add another regressor. Why?
- Think about starting with one regressor and then adding another
  - With one regressor, OLS finds the values of b_0 and b_1 that minimize

$$\sum_{i=1}^{n} [Y_i - b_0 - b_1 X_{1i}]^2$$

  - Now add another regressor
  - With two regressors, OLS finds the values of b_0, b_1, and b_2 that minimize

$$\sum_{i=1}^{n} [Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}]^2$$

  - If OLS happens to choose a value of $b_2 = 0$, then the SSR will be the same whether or not the second variable is included in the regression.

$$\sum_{i=1}^{n}[Y_i - b_0 - b_1X_{1i} - b_2X_{2i}]^2$$

- But if OLS chooses a value of b_2 such that $b_2 \neq 0$, then it must be that this value of b_2 reduced the SSR relative to the regression that excluded this regressor.

- Since SSR will decrease as long as the value of b_2 chosen by OLS is not zero, $R^2 = 1 - \frac{SSR}{TSS}$ will increase if a regressor is added.

- R^2 generally ***increases*** (and ***never decreases***) when a ***new regressor is added***.

Adjusted R^2

- Since R^2 generally increases when a new regressor is added, an increase in R^2 does not necessarily mean that the added variable actually improves the regression's fit.
- R^2 gives an "inflated" measure of how well the regression line fits the data.
- To correct for this, the adjusted R^2 ($\bar{R}^2$) is used.
- $\bar{R}^2$ "penalizes" you for including another regressor – $\bar{R}^2$ does not necessarily increase when you add another regressor.

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

where k is number of regressors and n is sample size.

$$R^2 = 1 - \frac{SSR}{TSS}$$

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

1) When there is one or more regressors, $\frac{n-1}{n-k-1} > 1$, so $\bar{R}^2 < R^2$ (if *n* is large and *k* is moderate, the two will be close).

2) Adding a regressor has 2 opposing effects on $\bar{R}^2$

   a) SSR falls; which tends to increase $\bar{R}^2$

   b) $\frac{\mathrm{n-1}}{\mathrm{n-k-1}}$ increases since k increases; which tends to decrease $\bar{R}^2$

   c) Whether $\bar{R}^2$ actually increases or decreases depends on which of these 2 effects is stronger

3) $\bar{R}^2$ can be negative.

Standard Error of the Regression (*SER*)

- *SER* and the *RMSE* are measures of the spread of the observations around the regression line (measured in units of *Y*).
- *SER* is the sample standard deviation of the OLS residuals:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(\hat{u}_i - \bar{\hat{u}}\right)^2} = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i{}^2}$$

second equality holds because the sample average of the OLS residuals is 0 (i.e. $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$).

Root Mean Squared Error

- *RMSE* in the multiple regression case is exactly as defined in the single-regressor case:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i{}^2}$$

Test Score E.g.

$$\widehat{TestScore} = 698.9 - 2.28 \times STR\,, \quad R^2 = 0.051, \quad SER = 18.6$$

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL,\ R^2 = 0.426, \bar{R}^2 = 0.424, SER = 14.5$$

- Including $PctEL$ in the regression increased R^2 from 0.051 to 0.426.
  - with only the single regressor STR, only a small fraction (5.1%) of the variation in $TestScore$ is explained.
  - with $PctEL$ included, 42.6% of the variation in $TestScore$ is explained.
  - including $PctEL$ substantially improves the fit of the regression.
  - Do R^2 & $\bar{R}^2$ differ much here. Why?

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \quad R^2 = 0.051, \quad SER = 18.6$$

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL, \; R^2 = 0.426, \bar{R}^2 = 0.424, SER = 14.5$$

- SER falls from 18.6 to 14.5 when $PctEL$ is added as a regessor.
- fall in SER means that the sample standard deviation of the OLS residuals falls.
- so the spread of the observations around the regression line becomes smaller.
- predictions of test scores using the OLS regression line with $PctEL$ as an added regressor will be more accurate.

Least Squares Assumptions in Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \qquad i = 1, \ldots, n$$

- What are the conditions under which OLS will give appropriate estimates of the population coefficients $\beta_0, \beta_1, \ldots, \beta_k$?
- 4 conditions – the "Least Squares Assumptions".
- first 3 are the same as those in the single regressor case (extended to allow for multiple regressors).
- fourth assumption – no perfect multicollinearity – is new.

Least Squares Assumption #1

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

LSA #1:

$$E(u_i | X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k) = 0$$

If there are omitted factors:

(1) which are not explicitly included in the model (therefore subsumed in u_i)

(2) and which are correlated with an included regressor $X_{1i}, \ldots, X_{ki}$.

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.

Least Squares Assumption #2 & #3

- LSA #2: $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ are i.i.d
  - arises automatically if the entity (e.g. individual, district) is selected by simple random sampling.

- LSA#3: Large outliers are rare
  - As in the case of a single regressor, OLS can be sensitive to large outliers, so check your data (scatterplots!) to make sure there are no extreme values (typos or coding errors).

Least Squares Assumption #4

- LSA #4: ***no*** perfect multicollinearity
- ***Perfect multicollinearity*** is when one of the regressors is an exact linear function of the other regressors.

Example:

```
regress testscr str str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
         str |   (dropped)
       _cons |   698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

- If the regressors exhibit perfect multicollinearity, it is ***not possible to compute the OLS estimator***…

- In previous e.g., the regressors exhibit perfect multicollinearity because one of the regressors (first occurrence of STR) is an exact linear function of another (second occurrence of STR).

- Why can't we compute the OLS estimates when there is perfect multicollinearity?

- …because you are asking the regression an illogical question!

- β_1 is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant (does this make sense?).

- return to perfect (& imperfect) multicollinearity shortly…

Sampling Distribution of the OLS Estimators in Multiple Regression

- The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are calculated from a sample of data; a different sample will give a different value of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. So the OLS estimators are random variables, each with a sampling distribution.

Under the ***four*** LSAs,

- The exact (finite-sample) distribution of $\hat{\beta}_1$ has mean β_1, $Var(\hat{\beta}_1) \propto \frac{1}{n}$; so too for $\hat{\beta}_2, \ldots, \hat{\beta}_k$.
- Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is very complicated; but for large n…
  - $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (LLN)
  - $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT)
  - So too for $\hat{\beta}_2, \ldots, \hat{\beta}_k$

Conceptually, there is nothing new here.

Multicollinearity: Perfect & Imperfect

Perfect Multicollinearity

Run a regression of $TestScore_i$ on STR_i & $PctEL_i$. Adding any of the following regressors will result in perfect multicollinearity:

E.g. #2: $FracEL_i$: Fraction of English learners in the i^{th} district. Why?

- For any district i, $PctEL_i = 100 \times FracEL_i$.
- one of the regressors ($FracEL_i$) is an exact linear function of another ($PctEL_i$).
- OLS will fail because you are asking "what is the effect of a unit change in the percentage of English Learners in the district, holding constant the fraction of English Learners in the district?"
- Since the percentage of English learners & the fraction of English learners in any district must move together in a ***perfectly linear way***, this question makes no sense.

- E.g. #3: $STRGT12$: binary variable indicating whether district class size is greater than or equal to 12. Why?

  - Define $STRGT12_i$ (=1 if $STR_i \geq 12$; =0 if $STR_i < 12$).

  - no districts in our dataset with $STR_i < 12$; means that $STRGT12_i = 1$ for all i ***in our dataset.***

  - linear regression model with an intercept can be thought of as including a ***regressor*** $\boldsymbol{X_{0i}}$ that ***equals 1*** for all i.

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,$$

where X_{0i}*=1 for all i*

  - thus, $STRGT12_i = 1 = X_{0i}$ for all i ($STRGT12_i$ is an exact linear function of the constant regressor X_{0i})

No districts with STR< 12 in the dataset

```
. summarize str
```

Variable	Obs	Mean	Std. Dev.	Min	Max
str	420	19.64043	1.891812	14	25.8

Eg #4: What if we regress $TestScore$ on **a constant** (i.e. X_{0i}), and on **two binary variables** $SmallSTR$ & $LargeSTR$, where

$$SmallSTR_i = \begin{cases} 1 \; if \; the \; STR \; in \; the \; ith \; district < 20 \\ 0 \; if \; the \; STR \; in \; the \; ith \; district \geq 20 \end{cases}$$

$$LargeSTR_i = \begin{cases} 0 \; if \; the \; STR \; in \; the \; ith \; district < 20 \\ 1 \; if \; the \; STR \; in \; the \; ith \; district \geq 20 \end{cases}$$

- In this case,

$$SmallSTR_i = 1 - LargeSTR_i \;\; or$$
$$SmallSTR_i = X_{0i} - LargeSTR_i$$

…so there is perfect multicollinearity

This example is a special case of…

The Dummy Variable Trap

- Suppose you have a set of multiple binary (dummy) variables, which are ***mutually exclusive*** and ***exhaustive*** – that is, there are multiple categories and every observation must fall into ***one and only one*** category.

- If you include all these dummy variables *and* a constant in the regression, you will have perfect multicollinearity – this is called ***the dummy variable trap***.

- Suppose you partition school districts into 3 mutually exclusive & exhaustive categories: rural, suburban, urban.

- Each district must fall into one category.

Define variables:

$Rural_i$ (=1 for a rural district; =0 otherwise)

$Suburban_i$ (=1 for a suburban district; =0 otherwise)

$Urban_i$ (=1 for an urban district; =0 otherwise)

- Because each district falls into one (and only one) category, for any district i,

$$Rural_i + Suburban_i + Urban_i = 1 = X_{0i}$$

- Since, $Rural_i = X_{0i} - Suburban_i - Urban_i$, each regressor, can be written as an exact linear function of the other regressors.

Solutions to the dummy variable trap:

- Omit one of the groups, *or*
- Omit the intercept (*not advised*)

Omit one of the groups

- To estimate the regression, exclude ***one*** of the binary indicators.
- For e.g., exclude $Rural_i$ from the regression.
  - has *implications for the interpretation of the coefficients on the binary variables*
  - *If we exclude* $Rural_i$, then the coefficient on $Suburban_i$ will be the difference in the mean test scores of suburban & rural districts, holding constant the other variables in the regression. Similarly, the coefficient on $Urban_i$ will be the difference in the mean test scores of urban and rural districts, holding constant the other variables in the regression.

- In general, if you have G binary variables, the usual way to avoid the dummy variable trap is to ***exclude one of the binary indicators*** from the regression, so only $G - 1$ variables are included as regressors.

- Coefficients on the included binary variables represent the ***incremental effect of being in that category***, ***relative to being in the omitted category (base case)***, holding the other regressors constant.

Perfect Multicollinearity

- Perfect multicollinearity usually reflects a mistake in specifying the regressors, or an oddity in the data.

- If you have perfect multicollinearity, your statistical software will let you know – either by crashing, giving an error message, or "dropping" one of the variables arbitrarily.

- Solution to perfect multicollinearity is to modify your list of regressors so that you no longer have it.

Imperfect Multicollinearity

Imperfect & perfect multicollinearity are different.

- ***Imperfect multicollinearity*** occurs when two or more regressors are highly (but not perfectly) correlated.
    - Why this name?

- Imperfect multicollinearity does not pose problems for the theory of the OLS estimators

  - OLS estimators ***will still be unbiased & consistent*** as long as the four LSAs hold.

- But imperfect multicollinearity implies that ***one or more of the regression coefficients will be imprecisely estimated.***

Why?

  - 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 $Var(\hat{\beta}_1)$ will be large.

Intuition: recall single regressor case:

$$Var(\hat{\beta}_1) = \frac{1}{n} \times \frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}$$

- if there is little variation in an independent variable (X_1), what happens to $Var(\hat{\beta}_1)$?

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

Let’s turn to hypothesis tests & confidence intervals in multiple regression…

Hypothesis Tests & Confidence Intervals for a Single Coefficient in Multiple Regression

- Consider a multiple regression model with k regressors.
- focus on the sampling distribution of $\hat{\beta}_1$
- If n is sufficiently large, then by the Central Limit Theorem,

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_1)}} \xrightarrow{d} N(0,1)$$

- In practice, we don't know $\sqrt{Var(\hat{\beta}_1)}$ so we estimate it using $SE(\hat{\beta}_1)$.

- Hypotheses on β_1 can be tested using the usual *t*-statistic, and confidence intervals are constructed as before:

95% CI for β_1

$$\left[\hat{\beta}_1 - 1.96SE\left(\hat{\beta}_1\right), \hat{\beta}_1 + 1.96SE\left(\hat{\beta}_1\right)\right]$$

- these apply too for hypothesis tests and CIs on $\beta_2, \ldots, \beta_k$.

- approach to hypothesis tests and to confidence intervals for a single coefficient in multiple regression ***remain the same*** as for the single regressor case.

- Note: the OLS estimators of the slope coefficients are generally not independent and so neither are their *t-statistics*.

Example: California Data

1) After controlling for the percentage of English learners in the district, is there any evidence that class size has an effect on test score?

2) What is a 95% confidence interval for the effect on test scores of a 1 unit change in *STR*?

```
reg testscr str pctel, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
```

OLS regression line with standard errors:

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} \times PctEL$$

1) After controlling for the percentage of English learners in the district, is there any evidence that class size has an effect on test score?

1) Formulate hypothesis: $H_0: \beta_1 = 0;\ \ H_1: \beta_1 \neq 0$

2) Compute the t-statistic: $t^{act} = \frac{\widehat{\beta^{act}}_1 - \beta_{1,0}}{SE(\widehat{\beta}_1)} = \frac{-1.10-0}{0.43} = -2.54$

3) Calculate the p-value: $2\Phi(-|t^{act}|) = 2\Phi(-2.54) = 0.011$

- Since $p - value = 0.011 < 0.05$, reject H_0 at the 5% level.
- Alternatively, since $|t^{act}| = 2.54 > 1.96$ (5% two-sided critical value), reject H_0 at the 5% level.

2) What is a 95% confidence interval for the effect on test scores of a 1 unit change in the STR?. Note: some rounding error is possible.

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} \times PctEL$$

a. $\{-1.73, -0.26\}$

b. $\{-1.73, -0.95\}$

c. $\{-0.95, -0.26\}$

d. $\{-1.95, -0.26\}$

Adding Expenditures Per Pupil to the Regression

- Reducing *STR* (i.e. hiring more teachers) costs money.
- superintendent asks "If I reduce *STR* while holding expenditures per pupil constant, how will this affect test scores?"
- in other words, she wants to know whether reducing class sizes will be useful if the reduction is achieved through budget cuts in other areas.

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

A school can lower class size by spending less on other areas (e.g. on hiring quality teachers)

Hwa Chong today

Over 80% of the school's close to 400 teachers hold advanced degrees (Honours or higher). About 35 staff members have PhDs, or are pursuing doctoral studies.

But Hwa Chong could choose to hire more teachers (reducing class size) by sacrificing teacher quality.

- Using OLS to estimate the relationship between $TestScore$ and STR, $Expn$, $PctEL$ with the 420 observations:

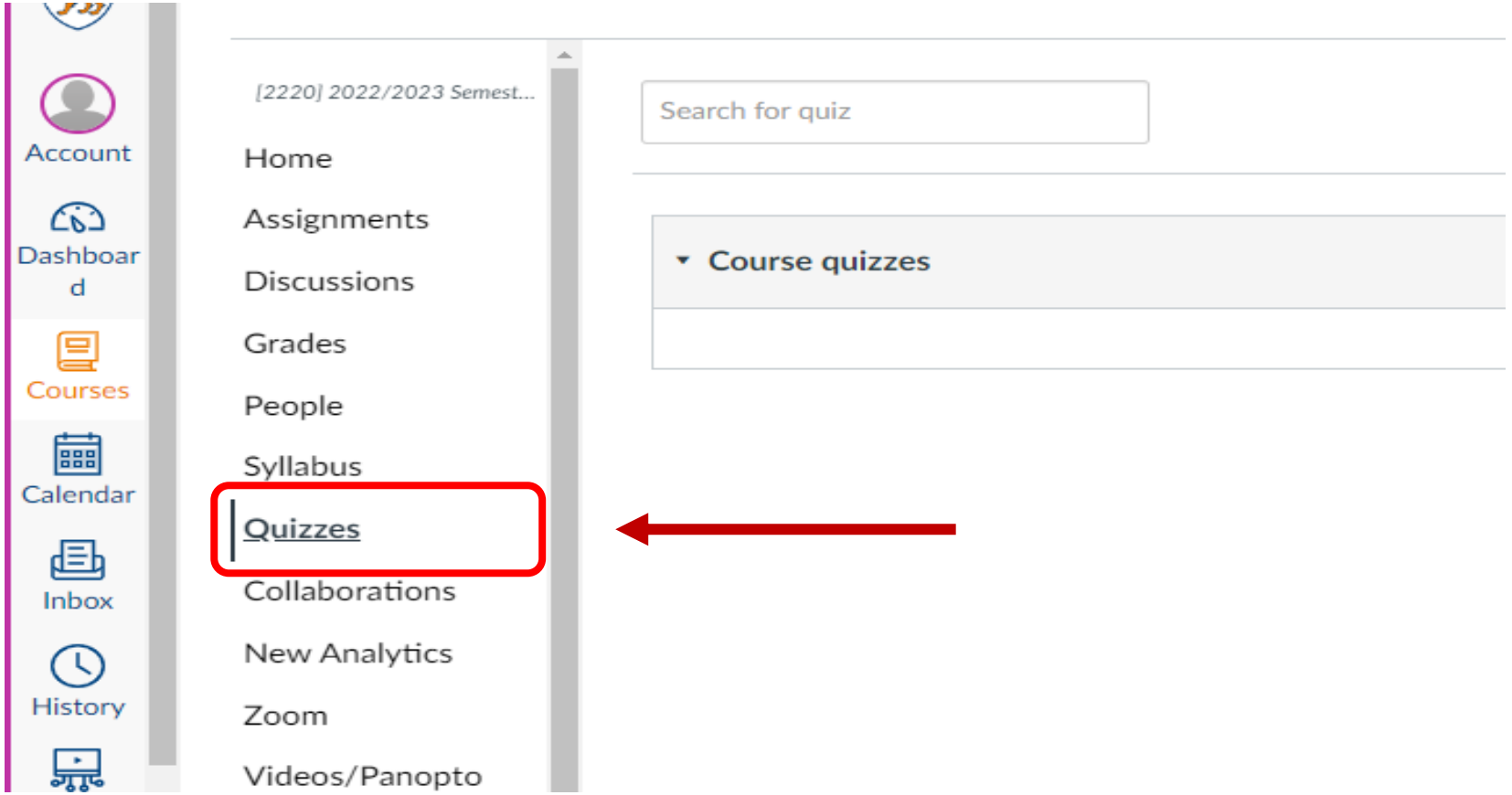
$$\widehat{TestScore} = \underset{(15.7)}{649.6} - \underset{(0.48)}{0.29} \times STR - \underset{(1.59)}{3.97} \times Expn - \underset{(0.032)}{0.656} \times PctEL$$

- Holding expenditures per pupil (and the percentage of English learners) constant, changing *STR* is estimated to have a very small effect on test scores.
  - hypothesis that the population coefficient on *STR* is equal to 0 cannot be rejected at even the 10% level.
- there is no evidence that reducing class size improves test scores if overall expenditures per pupil are held constant (i.e. if reducing class size is only made possible by budget cuts in other areas).

Homework 1

- Homework 1 will be posted on 28 February (Tuesday), 12pm and will be due on 3 March (Friday), 7pm.
- It will be done through **Canvas Quiz**.
- Go to Canvas, on the left panel click on “Quizzes” (see next slide) and you will be able to access the Homework.
- Covers Lectures 1-4.
- Students need not complete the homework in 1 sitting. They can save the homework and submit it only later (but by the due date).

Canvas Quiz


Account
Dashboard
Courses
Calendar
Inbox
History
[2220] 2022/2023 Semest...
Home
Assignments
Discussions
Grades
People
Syllabus
Quizzes
Collaborations
New Analytics
Zoom
Videos/Panopto
Search for quiz
Course quizzes

Midterm Test

- March 7 (Tuesday), 4.15pm-5.15pm (week 8)
- 1-hour
- Venue: MPSH 2A
- Seating plan in MPSH 2A will be provided nearer the date
- Please arrive at least 5 mins early
- Covers lecture 1 to lecture 4 (inclusive)
- 20 MCQ questions
- Closed book
- Bring a 2B pencil with you
- Bring a calculator with you
- Normal table will be provided

Additional Consultation

- Every Friday, 4-6pm (regular consultation)
- plus 2 March, Thursday, 4-6pm (additional consultation)
- https://nus-sg.zoom.us/j/83502944248?pwd=K2c1WENobVBXV0YrcEliUlh4Z2s5dz09
- Meeting ID: 835 0294 4248
- Passcode: 714459