MPO1: Quantitative Research Methods

Session 5: Precision of OLS estimators, Multiple regression models, Multicollinearity, F-tests for goodness of fit

Thilo Klein

University of Cambridge
Judge Business School

Gauss-Markov Theorem

Efficiency of OLS - The Gauss-Markov Theorem

- OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions
    - of $Y_1, ..., Y_n$, in bivariate regression
- Under assumptions 1-4 (mean zero conditional distributions of disturbances, i.i.d. sampling, no outliers, homoscedasticity):
- the OLS estimators have the *smallest variance* among all linear estimators (i.e., of all estimators that are linear functions of $Y_1, ..., Y_n$)
    - Aside: proof available in standard texts (if you are interested)

CAMBRIDGE Judge Business School

Gauss-Markov Theorem

Efficiency of OLS esimators (2)

- Under all five assumptions - i.e., including normally distributed errors:
- OLS estimators have the smallest variance among all consistent estimators (i.e., linear or nonlinear functions of $Y_1, ..., Y_n$)
  - Aside: proof available in standard texts
- This is a strong result: OLS is a better choice than any other consistent estimator
- An estimator that is not consistent is a very poor choice, so OLS really is the best we can do - *if all five assumptions hold*

CAMBRIDGE Judge Business School

Gauss-Markov Theorem

Limitations of OLS

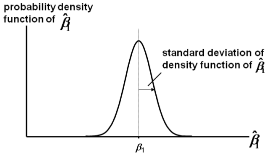
- OLS is more sensitive to outliers than some other estimators
- Recall that to estimate the population mean, if there are outliers, then the sample median is preferred to the sample mean
  - the median is less sensitive to outliers - it has smaller variance than OLS (mean) when there are outliers
- Similarly, in regression, if there are outliers, then there are other estimators that are more efficient (have smaller variances)
  - Q: What are outliers? How can we treat them?
  - Aside: Robust statistics
- All said, OLS is the most popular estimator in applied regression analysis

CAMBRIDGE Judge Business School

Precision of OLS estimators

Variance of OLS estimators under homoscedasticity

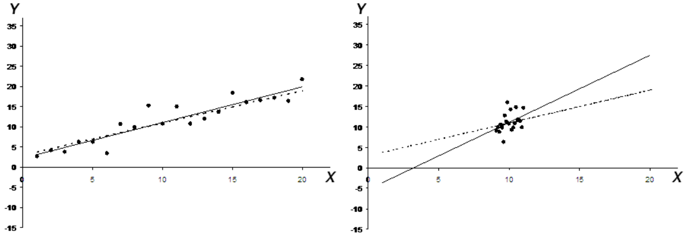
- Simple linear regression model: $Y = \beta_0 + \beta_1 X + u$
- Focus on the slope coefficient : more "interesting" and useful (why?). All arguments apply to the intercept as well
- Variances (of the sampling distributions) of regression coefficients (under homoscedasticity)
- $\sigma_{\hat{\beta}_1}{}^2 = \frac{\sigma_u{}^2}{nVar(X)}$
  - $\sigma_{\hat{\beta}_0}{}^2 = \frac{\sigma_u{}^2}{n}\{1 + \frac{\bar{X}^2}{Var(X)}\}$



Precision of OLS estimators

Variance of OLS estimators under homoscedasticity (2)

- Larger $V(X)$ (and larger n), lower is V(OLS estimators)



CAMBRIDGE Judge Business School

Precision of OLS estimators

Variance of OLS estimators under heteroscedasticity

- ${\sigma_{\hat{\beta}_1}}^2 = \frac{Var[(X_i - E(X))u_i]}{n[Var(X)]^2}$
  - Aside: derived in textbooks
- For comparison, Recall, for i.i.d. random variable Y:
  - $E(\bar{Y}) = \mu_Y \qquad Var(\bar{Y}) = \frac{{\sigma_Y}^2}{n}$

CAMBRIDGE Judge Business School

Precision of OLS estimators

So, some dispersions we are concerned with

- Simple linear regression model: $Y = \beta_0 + \beta_1 X + u$
- ${\sigma_{\hat{\beta}_1}}^2 = \frac{{\sigma_u}^2}{nVar(X)}$ (homoscedastic case)
- ${\sigma_u}^2$ estimated with ${s_u}^2 = \frac{n}{n-2}Var(e)$ (Why the $\frac{n}{n-2}$ term?)
  - $E[Var(e)] = \frac{n-2}{n}{\sigma_u}^2$ (proof available in textbooks)
  - So if we define ${s_u}^2 = \frac{n}{n-2}Var(e)$, then $E[{s_u}^2] = E[\frac{n}{n-2}Var(e)] = {\sigma_u}^2$: unbiased
- $s.e.(\hat{\beta}_1) = \sqrt{\frac{{s_u}^2}{nVar(X)}} = \sqrt{\frac{Var(e)}{(n-2)Var(X)}}$
- $s.e.(\hat{\beta}_0) = \sqrt{\frac{{s_u}^2}{n}\left(1 + \frac{\bar{X}^2}{Var(X)}\right)} = \sqrt{\frac{Var(e)}{n-2}\left(1 + \frac{\bar{X}^2}{Var(X)}\right)}$

CAMBRIDGE Judge Business School

Hypothesis testing on regression coefficients

Null and Alternate hypotheses

- Objective: test a hypothesis, e.g., $\beta_1 = \beta_1^*$, and reach a probabilistic conclusion whether this hypothesis is correct or incorrect, relative to an alternative
- General setup
- Null hypothesis and two-sided alternative:
  - $H_0 : \beta_1 = \beta_1^* \; Vs. \; H_a : \beta_1 \neq \beta_1^*$
- Null hypothesis and one-sided alternative:
  - $H_0 : \beta_1 = \beta_1^* \; Vs. \; H_a : \beta_1 > \beta_1^*$
  - $H_0 : \beta_1 = \beta_1^* \; Vs. \; H_a : \beta_1 < \beta_1^*$

CAMBRIDGE Judge Business School

Hypothesis testing on regression coefficients

Approach

- General approach: construct test-statistic, and compute p-value (or compare with critical value from t or $N(0,1)$)
- In general: test statistic $= \frac{\text{estimator - hypothesised value}}{\text{std. error of estimator}}$
- to test β_1^*: $t = \frac{\hat{\beta}_1 - \beta_1^*}{s.e.(\hat{\beta}_1)}$
    - Recall: $s.e.(\hat{\beta}_1) =$ the square root of the estimator of the variance of the sampling distribution of $\hat{\beta}_1$
- For $H_0 : \beta_1 = \beta_1^* \; Vs. \; H_a : \beta_1 \neq \beta_1^*$: Reject at 5% significance level if $|t| > 1.96$ (if large sample)
- Q: What is a confidence interval for β_1?

CAMBRIDGE Judge Business School

Multiple Regression estimators

Multiple regression with two explanatory variables: example

- A model for automobile sales : where registrations depend on “list price” and “rebate”
- Registrations $= \beta_0 + \beta_1 Price + \beta_2 Rebate + u$
- Or a model for MPO1 marks : where marks depend on “individual work” and “group work”
  Marks $= \beta_0 + \beta_1 \mathrm{IW} + \beta_2 \mathrm{GW} + u$

CAMBRIDGE Judge Business School

Multiple Regression estimators

Multiple regression with two explanatory variables: example

- Population regression function

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

- Sample regression function

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

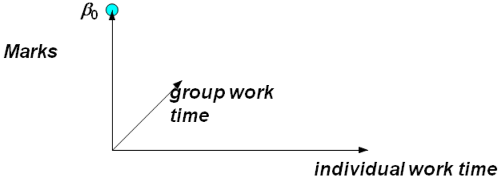
- Residual

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$$

Multiple Regression estimators

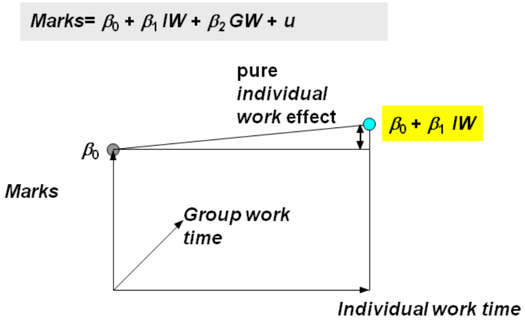
Multiple regression in pictures (1)

$$Marks = \beta_0 + \beta_1 IW + \beta_2 GW + u$$



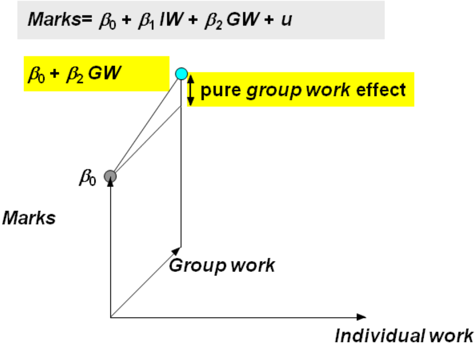
Multiple Regression estimators

Multiple regression in pictures (2)



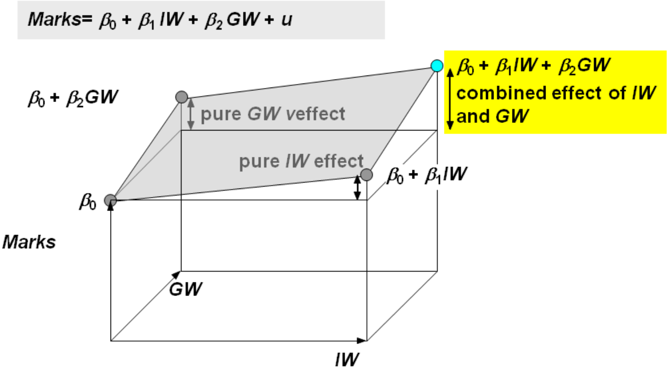
Multiple Regression estimators

Multiple regression in pictures (3)



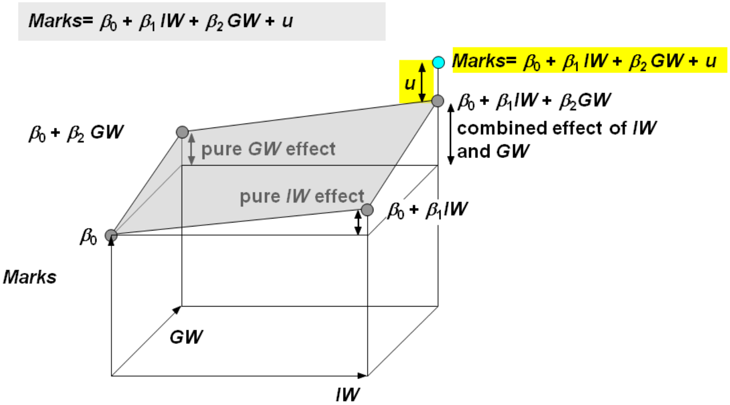
Multiple Regression estimators

Multiple regression in pictures (4)



Multiple Regression estimators

Multiple regression in pictures (5)



Multiple Regression estimators

Multiple regression with two explanatory variables: example

- Residual sum of squares in terms of unknown estimators

$$RSS = \sum e_i^2 = \sum (Y_i - \hat{Y}_i) = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

- Minimising error

$$\frac{\partial\, RSS}{\partial \hat{\beta}_0} = 0; \quad \frac{\partial\, RSS}{\partial \hat{\beta}_1} = 0; \quad \frac{\partial\, RSS}{\partial \hat{\beta}_2} = 0$$

CAMBRIDGE Judge Business School

Multiple Regression estimators

Multiple regression with two explanatory variables: example

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$
- $\hat{\beta}_1 = \frac{Cov(X_1,Y)Var(X_2) - Cov(X_2,Y)Cov(X_1,X_2)}{Var(X_1)Var(X_2) - [Cov(X_1,X_2)]^2}$
- $\hat{\beta}_2 = \frac{Cov(X_2,Y)Var(X_1) - Cov(X_1,Y)Cov(X_1,X_2)}{Var(X_1)Var(X_2) - [Cov(X_1,X_2)]^2}$
- Derivations in textbooks, if interested
- Expressions simpler for general models with matrix notation

CAMBRIDGE Judge Business School

R^2 and adjusted R^2

R^2: Coefficient of determination

- $\sum(Y_i - \bar{Y})^2$ is the total sum of squares (TSS)
- $\sum(\hat{Y}_i - \bar{Y})^2$ is the explained sum of squares (ESS)
- $\sum e_i^2$ is the residual sum of squares (RSS)
- TSS=ESS+RSS

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2}$$

- R^2 always increases with the number of regressors.
- Cannot compare 'larger' and 'smaller' models with this measure of "fit"

CAMBRIDGE Judge Business School

R^2 and adjusted R^2

Adjusted R^2

- "Adjusted R^2": makes comparison possible by "penalising" inclusion of more regressors

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

- $\bar{R}^2$ *can* fall when unrelated regressors are included
- $\bar{R}^2 < R^2$
- If n large, the two *can* be close
- Can R^2 ever be negative?
  - Yes, if the best-fit model is inappropriate and fits the data worse than a horizontal line at the mean Y value

CAMBRIDGE Judge Business School

Multiple Regression estimators: properties

Multiple regression estimators: desirable properties

- If the model is correctly specified, and the "Gauss Markov Assumptions" are not violated, OLS estimators of the multiple regression model coefficients ($\hat{\beta}_k$) are:
  - Unbiased
  - Efficient
  - Consistent

CAMBRIDGE Judge Business School

Multiple Regression estimators: properties

Precision of Multiple regression estimators

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$
- Population variance of a slope coefficient, say $\hat{\beta}_1$:
  $\sigma_{\hat{\beta}_1}{}^2 = \frac{\sigma_u{}^2}{nVar(X_1)} \times \frac{1}{1-r_{X_1,X_2}{}^2}$
  - Recall: in the simple linear model $Y_i = \beta_0 + \beta_1 X_i + u_i$
    $\sigma_{\hat{\beta}_1}{}^2 = \frac{\sigma_u{}^2}{nVar(X)}$
    - $E[Var(e)] = \frac{n-2}{n}\sigma_u{}^2$
    - $s_u{}^2 = \frac{n}{n-2}Var(e)$
    - $\sigma_{\hat{\beta}_1}{}^2 = \frac{\frac{n}{n-2}Var(e)}{nVar(X)}$
- Sample estimate of the variance of a slope coefficient, $\hat{\beta}_1$ for $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$:
- $s.e.(\hat{\beta}_1)^2 = \frac{s_u{}^2}{nVar(X_1)} \times \frac{1}{1-r_{X_1,X_2}{}^2} = \frac{Var(e)}{(n-3)Var(X_1)} \times \frac{1}{1-r_{X_1,X_2}{}^2}$

CAMBRIDGE Judge Business School

Multicollinearity

What is Multicollinearity

- Situation when two or more predictor variables are highly (linearly) correlated
  - i.e., $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + v_i = 0$; Variance of v_i is small
  - Some of the β_ks may be zero in the above
- Multicollinearity does not reduce predictive power or reliability of the *model as a whole*
- But reduces precision of estimators relating to individual predictors (why?)

CAMBRIDGE Judge Business School

Multicollinearity

Diagnosing Multicollinearity

- $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + u_i$
- Population variance of the OLS estimator for a typical regression coefficient, e.g., $\hat{\beta}_k$:
- $\sigma_{\hat{\beta}_k}{}^2 = \frac{\sigma_u{}^2}{nVar(X_k)} \times \frac{1}{1-R_k{}^2}$
- $R_k{}^2$ is the R^2 for the regression of X_k against all other explanatory variables in the model:
    - $X_k = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{k-1} X_{k-1} + \gamma_{k+1} X_{k+1} + \cdots + \gamma_K X_K + \nu_i$
- If there is no linear relation between X_k and the other explanatory variables in the model, $R_k{}^2 \approx 0$
- Diagnostic for multicollinearity is *related to* $R_k{}^2$

CAMBRIDGE Judge Business School

Multicollinearity

Variance inflation factor

- Variance Inflation Factor$_k = \frac{1}{1-{R_k}^2}$
- Assesses the degree to which variance (s.e. of the coefficient) is inflated because regressor k is not orthogonal to the other regressors
- However, the sampling distribution of VIF is not known
- Rule of thumb: Consider multicollinearity a significant problem if average $VIF > 1$ or individual $VIF > 10$ (for any regressor)

CAMBRIDGE Judge Business School

Multicollinearity

Alleviating Multicollinearity

- $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$
- ${\sigma_{\hat{\beta}_1}}^2 = \frac{{\sigma_u}^2}{nVar(X_1)} \times \frac{1}{1-{r_{X_1,X_2}}^2}$
  - Reduce σ_u^2 by including further relevant variables in the model
  - Increase the number of observations, n
  - Increase $Var(X_.)$
  - Reduce r_{X_1,X_2}
  - Combine the correlated variables
  - Drop some of the correlated variables

CAMBRIDGE Judge Business School

χ^2 and F Distributions

Chi-squared Distribution ${\chi_K}^2$

- If $Y_i \sim N(0,1)$, then
- $\sum_{i=1}^{K} {Y_i}^2 \sim {\chi_K}^2$ distribution, with K degrees of freedom

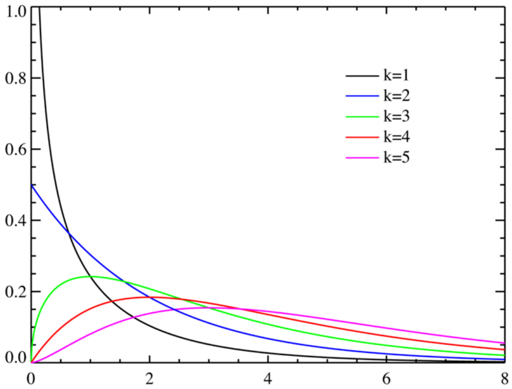
$$pdf: \; f(y,K) = \begin{cases} \frac{1}{2^{K/2}\Gamma(K/2)} y^{(K/2)-1} e^{-y/2} & for\ y > 0 \\ 0 & for\ y \leq 0 \end{cases}$$

  - $\Gamma(\cdot)$ is the Gamma function
- $E(\sum_{i=1}^{K} {Y_i}^2) = K$

CAMBRIDGE Judge Business School

χ^2 and F Distributions

Chi-squared Distribution $\chi_K{}^2$



χ^2 and F Distributions

F Distribution

- If $U_1 \sim {\chi_{df_1}}^2$, $U_2 \sim {\chi_{df_2}}^2$ and U_1, U_2 are independent, then

$$X = \frac{U_1/df_1}{U_2/df_2} \sim F_{df_1,df_2}$$

- pdf of an F distributed random variable, X with df_1 and df_2 degrees of freedom is:

$$f(x) = \frac{\sqrt{\frac{(df_1\, x)^{df_1}\ df_2^{df_2}}{(df_1\, x + df_2)^{df_1+df_2}}}}{x\,\mathrm{B}\left(\frac{df_1}{2}, \frac{df_2}{2}\right)}$$

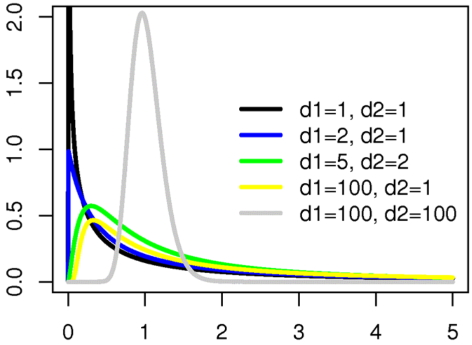
    - $B(\cdot,\cdot)$ is the Beta function

- $E(X) = \frac{df_2}{df_2 - 2}$ for $df_2 > 0$

CAMBRIDGE Judge Business School

χ^2 and F Distributions

F-distribution



F Tests of fit

F-test of R^2

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u_i$$

$$H_0 : \beta_1 = \cdots = \beta_K = 0 \quad H_a : \text{at least one } \beta \neq 0$$

$$\begin{aligned}
\frac{ESS/(K-1)}{RSS/(n-K)} &= \frac{\frac{ESS}{TSS}/(K-1)}{\frac{RSS}{TSS}/(n-K)} \\
&= \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K)
\end{aligned}$$

Application

CAMBRIDGE Judge Business School

F Tests of fit

Another application: incremental contribution of a set of variables

- $Y = \beta_1 + \beta_2 X_2 + u : \quad RSS_1$
- $Y = \beta_1 + \beta_2 X_2 + +\beta_3 X_3 + \beta_4 X_4 + u : \quad RSS_2$
- $H_0 : \beta_3 = \beta_4 = 0; \quad H_a : \beta_3 \neq 0$ or $\beta_4 \neq 0$ or both β_3 and $\beta_4 \neq 0$

$$\frac{\text{Increase in ESS}}{\text{cost in d.f.}} \Big/ \frac{\text{remaining RSS}}{\text{d.f. remaining}} \sim F(\text{cost, d.f. remaining})$$

$$\frac{(RSS_1 - RSS_2)/(df_1 - df_2)}{RSS_2/df_2} \sim F(df_1, df_2)$$

- Note: $F_{1,n}$ is the squared Student t_n distribution

CAMBRIDGE Judge Business School