Elements of Econometrics. Lecture 4. Multiple Linear Regression Model.

FCS, 2022-2023

MULTIPLE REGRESSION MODEL A ASSUMPTIONS

A.1 The model is linear in parameters and correctly specified.

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

A.2 There does not exist an exact linear relationship among the regressors in the sample.

A.3 The disturbance term has zero expectation (Gauss-Markov 1)

A.4 The disturbance term is homoscedastic (G-M 2)

A.5 The values of the disturbance term have independent distributions (G-M 3)

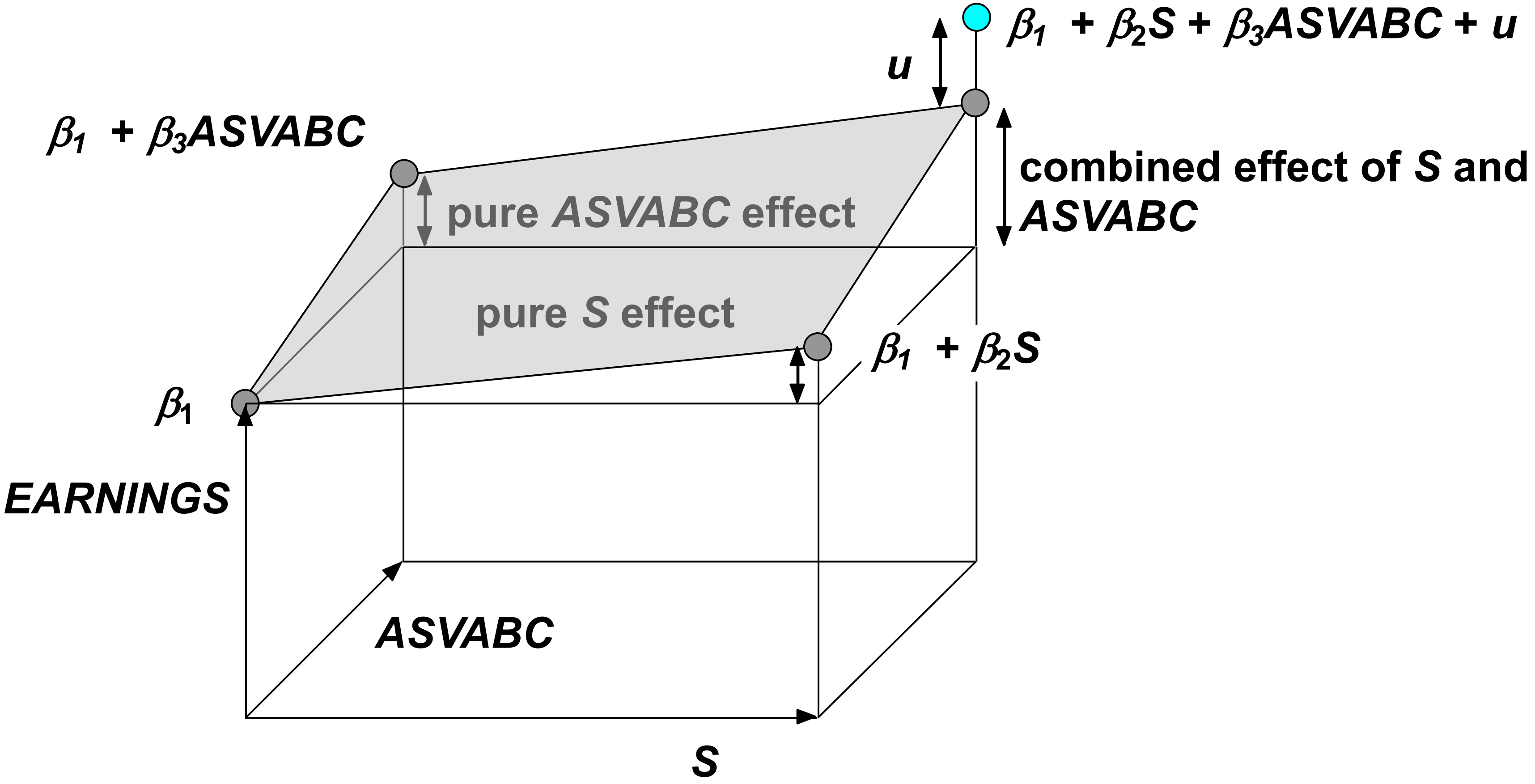
A.6 The disturbance term has a normal distribution

For the purposes of statistical inference, the assumption of normality can be replaced by a large sample size

Special case to be considered first: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 ASVABC + u$$



$\beta_1 + \beta_2 S + \beta_3 ASVABC + u$
u
$\beta_1 + \beta_3 ASVABC$
combined effect of S and $ASVABC$
pure $ASVABC$ effect
pure S effect
$\beta_1 + \beta_2 S$
β_1
$EARNINGS$
$ASVABC$
S

We assume that the effects of *S* (years of schooling) and *ASVABC* (indicator of abilities) on *EARNINGS* are linear and additive.

The impact of a difference in *S* on *EARNINGS* is supposed to be not affected by the value of *ASVABC*, and vice versa.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

True model	Fitted model
$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}$$

We define *RSS*, the sum of the squares of the residuals, and choose *b*$_1$, *b*$_2$, and *b*$_3$ so as to minimize it.

$$SSR = \sum \hat{u}_i^2 = \sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2 \rightarrow \min$$

The first order conditions are:

$$\frac{\partial SSR}{\partial b_1} = 0 \qquad \frac{\partial SSR}{\partial b_2} = 0 \qquad \frac{\partial SSR}{\partial b_3} = 0$$

$$\frac{\partial SSR}{\partial b_1} = \sum -(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \Rightarrow$$

$$\Rightarrow b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 \bar{X}_3$$

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$SSR = \sum Y_i^2 + nb_1^2 + b_2^2 \sum X_{2i}^2 + b_3^2 \sum X_{3i}^2 - 2b_1 \sum Y_i$$

$$-2b_2 \sum X_{2i}Y_i - 2b_3 \sum X_{3i}Y_i + 2b_1b_2 \sum X_{2i}$$

$$+2b_1b_3 \sum X_{3i} + 2b_2b_3 \sum X_{2i}X_{3i}$$

$$\frac{\partial SSR}{\partial b_2} = 2b_2 \sum X_{2i}^2 - 2 \sum X_{2i}Y_i + 2b_1 \sum X_{2i} + 2b_3 \sum X_{2i}X_{3i} = 0$$

$$\frac{\partial SSR}{\partial b_2} = 2b_2 \sum X_{2i}^2 - 2 \sum X_{2i}Y_i + 2(\bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3) \sum X_{2i} + 2b_3 \sum X_{2i}X_{3i} = 0$$

$$\frac{\partial SSR}{\partial b_3} = 2b_3 \sum X_{3i}^2 - 2 \sum X_{3i}Y_i + 2(\bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3) \sum X_{3i} + 2b_2 \sum X_{2i}X_{3i} = 0$$

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES:

OLS ESTIMATORS

$$b_1 = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3$$

$$b_2\widehat{\mathrm{Var}}(X_2) + b_3\widehat{\mathrm{Cov}}(X_2, X_3) = \widehat{\mathrm{Cov}}(X_2, Y)$$

$$b_2\widehat{\mathrm{Cov}}(X_2, X_3) + b_3\widehat{\mathrm{Var}}(X_3) = \widehat{\mathrm{Cov}}(X_3, Y)$$

$$b_2 = \frac{\widehat{\mathrm{Cov}}\ (X_2, Y)\widehat{\mathrm{Var}}\ (X_3) - \widehat{\mathrm{Cov}}(X_3, Y)\widehat{\mathrm{Cov}}(X_2, X_3)}{\widehat{\mathrm{Var}}(X_2)\widehat{\mathrm{Var}}(X_3) - \left[\widehat{\mathrm{Cov}}(X_2, X_3)\right]^2} = \Delta_1/\Delta$$

$$b_3 = \frac{\widehat{\mathrm{Cov}}(X_3, Y)\widehat{\mathrm{Var}}(X_2) - \widehat{\mathrm{Cov}}(X_2, Y)\widehat{\mathrm{Cov}}(X_2, X_3)}{\widehat{\mathrm{Var}}(X_2)\widehat{\mathrm{Var}}(X_3) - \left[\widehat{\mathrm{Cov}}(X_2, X_3)\right]^2} = \Delta_2/\Delta$$

We thus obtain three equations in three unknowns.

Solving for *b*$_1$, *b*$_2$, and *b*$_3$, we obtain the expressions shown above.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES:

PROPERTIES OF OLS ESTIMATORS

$$\hat{\beta}_2 = \beta_2 + \frac{\widehat{\mathrm{Cov}}(X_2, u)\widehat{\mathrm{Var}}(X_3) - \widehat{\mathrm{Cov}}(X_3, u)\widehat{\mathrm{Cov}}(X_2, X_3)}{\widehat{\mathrm{Var}}(X_2)\widehat{\mathrm{Var}}(X_3) - \left[\widehat{\mathrm{Cov}}(X_2, X_3)\right]^2}$$

$$E(\hat{\beta}_2) = \beta_2 + \frac{\widehat{\mathrm{Var}}(X_3)\mathrm{E}(\widehat{\mathrm{Cov}}(X_2, u)) - \widehat{\mathrm{Cov}}(X_2, X_3)\mathrm{E}(\widehat{\mathrm{Cov}}(X_3, u))}{\Delta} =$$

$$= \beta_2 + \mathrm{E}(\sum a_{i2}^* u_i) = \hat{\beta}_2 + \sum a_{i2}^* \mathrm{E}(u_i) = \beta_2 \quad \mathit{Similarly}, \mathrm{E}(\hat{\beta}_3) = \beta_3$$

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}(\bar{Y} - \hat{\beta}_2\bar{X}_2 - \hat{\beta}_3\bar{X}_3) = \mathrm{E}(\beta_1 + \beta_2\bar{X}_2 + \beta_3\bar{X}_3 + \bar{u} - b_2\bar{X}_2 - b_3\bar{X}_3) =$$
$$= \beta_1 + \beta_2\bar{X}_2 + \beta_3\bar{X}_3 - \bar{X}_2\mathrm{E}(\hat{\beta}_2) - \bar{X}_3\mathrm{E}(\hat{\beta}_3) = \beta_1$$

The OLS estimators are unbiased if the assumptions are valid. They are also efficient.

EXAMPLE OF MLR ESTIMATION, EARNINGS FUNCTION

```
Dependent Variable: EARNINGS
Method: Least Squares
Included observations: 570
   Variable          Coefficient        Std. Error         t-Statistic        Prob.

 C                   -13.38022          2.434830           -5.495340          0.0000
S                    1.307620           0.184170           7.100061           0.0000
ASVABC               0.183290           0.048992           3.741218           0.0002

    R-squared        0.185923           Mean dependent var         13.68988
  Adj. R-squared     0.183051           S.D. dependent var         9.702960
S.E. of regression   8.770041           Akaike info criterion      7.185809
Sum squared resid    43610.02           Schwarz criterion          7.208681
Log likelihood       -2044.956          F-statistic                64.74713
Durbin-Watson stat   1.784141           Prob(F-statistic)          0.000000
```

$$\widehat{EARNINGS} = -13.38 + 1.31S + 0.18ASVABC$$

It indicates that hourly earnings increase (on average, others equal) by $1.31 for every extra year of schooling and by $0.18 for every extra point of abilities

PRECISION OF THE MULTIPLE REGRESSION COEFFICIENTS (k=3)

True model

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Fitted model

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma_u^2}{\sum (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma_u^2}{\sum {x_{2i}}^2} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

$$E\left(\frac{1}{n}\sum \hat{u}_i^2\right) = \frac{n-k}{n}\sigma_u^2$$

$$s_u^2 = \frac{1}{n-k}\sum \hat{u}_i^2$$

$$\text{s.e.}(\hat{\beta}_2) = \sqrt{\frac{s_u^2}{\sum (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r^2_{X_2,X_3}}}$$

$$\text{s.e.}(\hat{\beta}_3) = \sqrt{\frac{s_u^2}{\sum (X_{3i} - \bar{X}_3)^2} \times \frac{1}{1 - r^2_{X_2,X_3}}}$$

$$\text{s.e.}(\hat{\beta}_2) = s_u \times \frac{1}{\sqrt{n}} \times \frac{1}{\sqrt{\Sigma {x_{2i}}^2 / n}} \times \frac{1}{\sqrt{1 - r^2_{X_2,X_3}}}$$

EXAMPLE OF MLR ESTIMATION, EARNINGS FUNCTION

$$\widehat{EARNINGS} = -13.38 + 1.31S + 0.18ASVABC$$

Dependent Variable: EARNINGS

Included observations: 570

Variable	Coefficient	Std. Error
S	1.308	0.1842
S.E. of regression		8.770041

$$\text{s.e.}(\hat{\beta}_2) = s_u \times \frac{1}{\sqrt{n}} \times \frac{1}{\sqrt{\Sigma x_{2i}{}^2/n}} \times \frac{1}{\sqrt{1 - r^2_{X_2,X_3}}}$$

$$= 8.77 \times \frac{1}{\sqrt{570}} \times \frac{1}{\sqrt{2.36^2 * 569/570}} \times \frac{1}{\sqrt{1 - 0.284}} =$$

$$= 8.77 * 0.042 * 0.424 * 1.18 = 0.1842$$

Auxiliary regression of HGC on ASVABC:

R-squared 0.284 S.D. dependent var 2.36

MULTIPLE LINEAR REGRESSION WITH (k-1) REGRESORS, VECTOR-MATRIX FORM

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \ldots + \hat{\beta}_k X_{ki}$$

$$\hat{Y} = \begin{pmatrix} \hat{Y}_1 \\ \ldots \\ \hat{Y}_n \end{pmatrix} = Xb = \begin{pmatrix} 1 & X_{21} & \ldots & X_{k1} \\ 1 & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{2n} & \ldots & X_{kn} \end{pmatrix} \begin{pmatrix} b_1 \\ \ldots \\ b_k \end{pmatrix}$$

$$e = Y - \hat{Y} = Y - Xb \qquad SSR = e'e \Rightarrow \min$$

MULTIPLE LINEAR REGRESSION WITH (k-1) REGRESORS, VECTOR-MATRIX FORM

$$SSR = e'e = (Y - Xb)'(Y - Xb) =$$
$$= Y'Y - Y'Xb - b'X'Y + b'X'Xb =$$
$$= \ Y'Y - 2b'X'Y + b'X'Xb \Rightarrow \min$$

$$\frac{\partial SSR}{\partial b} = -2X'Y + 2X'Xb = 0$$
$$X'Xb = X'Y \Rightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

$$\mathrm{Var}(\hat{\beta}) = {\sigma_u}^2(X'X)^{-1}$$

$$\mathrm{Var}(\hat{\beta}_i) = \frac{{\sigma_u}^2}{\Sigma {x_j}^2}\frac{1}{1 - {R_i}^2}$$

PRECISION OF THE MULTIPLE REGRESSION COEFFICIENTS

True model

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Fitted model

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

$$\sigma_{b_j}^2 = \frac{\sigma_u^2}{\sum\left(X_{ji} - \bar{X}_j\right)^2} \times \frac{1}{1 - R_j^2} = \frac{\sigma_u^2}{\sum {x_{ji}}^2} \times \frac{1}{1 - R_j^2}$$

$$\mathrm{E}\left(\frac{1}{n}\sum \hat{u}_i^2\right) = \frac{n-k}{n}\sigma_u^2$$

$$s_u^2 = \frac{1}{n-k}\sum \hat{u}_i^2$$

$$\text{s.e.}\left(\hat{\beta}_j\right) = \sqrt{\frac{s_u^2}{\sum\left(X_{ji} - \bar{X}_j\right)^2} \times \frac{1}{1 - R_j^2}}$$

Where R_j^2 *is* determination coefficient of the regression of X_j on all X_m $(m \neq j)$

t TESTS OF HYPOTHESES RELATING TO REGRESSION COEFFICIENTS

True model

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Fitted model

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

Null hypothesis $H_0\colon \beta_i = \beta_i^0$

Alternative (two-sided) hypothesis $H_1\colon \beta_i \neq \beta_i^0$

Test statistic

$$t_2 = \frac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)}; \quad t_3 = \frac{\hat{\beta}_3 - \beta_3^0}{\text{s.e.}(\hat{\beta}_3)}; \; \ldots \; ; \quad t_k = \frac{\hat{\beta}_k - \beta_k^0}{\text{s.e.}(\hat{\beta}_k)}$$

Reject *H*$_0$ if $|t| > t_{\text{crit}}$

d.f. = n-k

CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS

Model $Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$

Null hypothesis: $H_0: \beta_2 = \beta_2^0$

Alternative hypothesis: $H_1: \beta_2 \neq \beta_2^0$ ***d.f. = n-k***

Reject *H*₀ if $\dfrac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)} > t_{\text{crit}}$ **or** $\dfrac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)} < -t_{\text{crit}}$

Reject *H*₀ if $\hat{\beta}_2 - \beta_2^0 > \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$ **or** $\hat{\beta}_2 - \beta_2^0 < -\text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$

Reject *H*₀ if $\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} > \beta_2^0$ **or** $\hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} < \beta_2^0$

Do not reject *H*₀ if $\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} \leq \beta_2 \leq \hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$

$(\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}; \hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}})$ ***- Confidence interval; same for i≠2***

Multiple Linear Regression Model:

F TEST OF GOODNESS OF FIT FOR THE WHOLE EQUATION

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

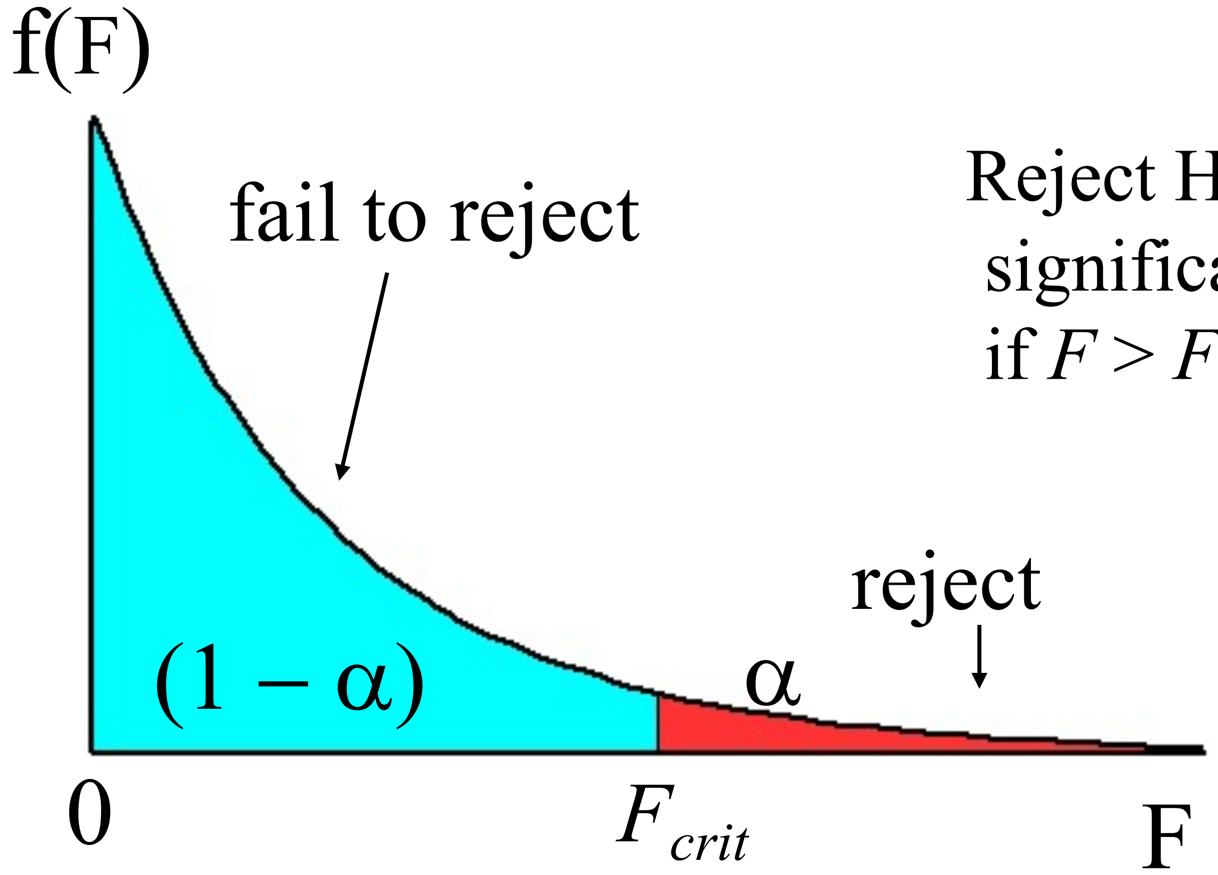
$$H_0 : \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \text{ at least one } \beta \neq 0$$

$$F(k-1, n-k) = \frac{(SSR_r - SSR_{ur})/(k-1)}{SSR_{ur}/(n-k)} = \frac{(SST - SSR)/(k-1)}{SSR/(n-k)} =$$

$$= \frac{SSE/(k-1)}{SSR/(n-k)} = \frac{\frac{SSE}{SST}/(k-1)}{\frac{SSR}{SST}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

$$F\text{(cost in d.f., d.f. unrestricted)} = \frac{\text{reduction in } SSR \;/\; \text{cost in d.f.}}{SSR \text{ unrestricted} \;/\; \text{degrees of freedom unrestricted}}$$

The *F* test and *F* statistic (continued)



Reject H_0 at α significance level if $F > F_{crit}$

F TEST OF GOODNESS OF FIT

Demonstration that *F* = *t* ^{2} IN THE SLR MODEL

$$F(k-1,n-k)=\frac{SSE/(k-1)}{SSR/(n-k)}=\frac{\frac{SSE}{SST}/(k-1)}{\frac{SSR}{SST}/(n-k)}=\frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

$$F=\frac{SSE}{SSR/(n-2)}=\frac{\sum\left(\hat{Y}_i-\bar{Y}\right)^2}{\sum\hat{u}_i^2/(n-2)}$$

$$=\frac{\sum\left([\hat{\beta}_1+\hat{\beta}_2X_i]-[\hat{\beta}_1+\hat{\beta}_2\bar{X}]\right)^2}{s_u^2}=\frac{1}{s_u^2}\sum\hat{\beta}_2^2(X_i-\bar{X})^2$$

$$=\frac{\hat{\beta}_2^2}{s_u^2}\sum(X_i-\bar{X})^2=\frac{\hat{\beta}_2^2}{s_u^2/\sum(X_i-\bar{X})^2}=\frac{\hat{\beta}_2^2}{\left(s.e.(\hat{\beta}_2)\right)^2}=t^2$$

The *F* test does not have its own role in the SLR model; it will do in the multiple regression.

R^2 and Adjusted R^2

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

Determination coefficient *R^2* always grows if an explanatory variable has been added, either significant or not. The adjusted coefficient was introduced which may increase or decrease:

$$R^2{}_{adj} = \bar{R}^2 = 1 - \frac{SSR/(n-k)}{SST/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k} =$$

$$R^2 - (1 - R^2)\frac{k-1}{n-k}$$

The $R^2{}_{adj}$ coefficient increases if and only if the absolute value of *t*-statistic of the added variable coefficient is greater than 1 (prove as an exercise before the next class). The $R^2{}_{adj}$ coefficient is not widely used for econometric analysis though available in the regression printouts.

R^2 and $\bar{R}^2$: What They Tell You—and What They Don't

The R^2 and $\overline{R}^2$ tell you whether the regressors are good at predicting, or "explaining," the values of the dependent variable in the sample of data on hand. If the R^2(or $\overline{R}^2$) is nearly one, then the regressors produce good predictions of the dependent variable in that sample, in the sense that the variance of the OLS residual is small compared to the variance of the dependent variable. If the R^2 (or $\overline{R}^2$) is nearly zero, the opposite is true.

The $\boldsymbol{R^2}$ and $\boldsymbol{\bar{R}^2}$ do NOT tell you whether:
1. an included variable is statistically significant;
2. the regressors are a true cause of the movements in the dependent variable;
3. there is omitted variable bias; or
4. you have chosen the most appropriate set of regressors.