Elements of Econometrics.
Lecture 28.
Revision 2

FCS, 2022-2023

Dummy variables

Binary variable for qualitative characteristics

Types of dummies

- Intercept dummy
- Slope dummy

Change of a reference category

Dummy variables trap (perfect multicolleniarity)

- If you have intercept: *L-1* dummies (*L* – number of categories)
- If you have no intercept: *L* dummies
- Include dummies for different characteristics (for example, gender and race) for each group if you have intercept
- Pay attention to interpretation of coefficients in both cases

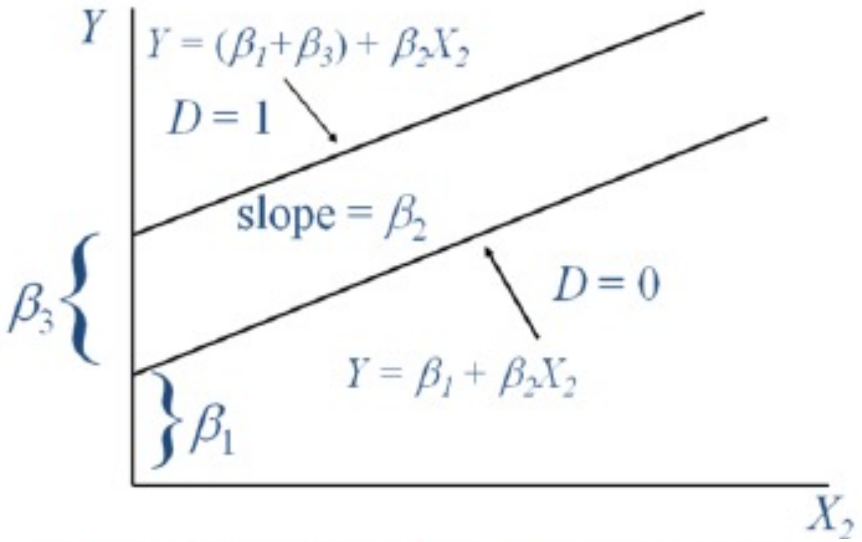
Dummy variables

Intercept dummy

Assume only shifts in constant term is affected

No change in slope (marginal effects are not affected)

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 \cdot D + u$$



Y
$Y = (\beta_1 + \beta_3) + \beta_2 X_2$
$D = 1$
slope $= \beta_2$
$D = 0$
$Y = \beta_1 + \beta_2 X_2$
β_3
β_1
X_2

Interpretation: Reference category is $D = 0$

If $D = 0$, then $Y = \beta_1 + \beta_2 X_2 + u$;

If $D = 1$, then $Y = (\beta_1 + \beta_3) + \beta_2 X_2 + u$

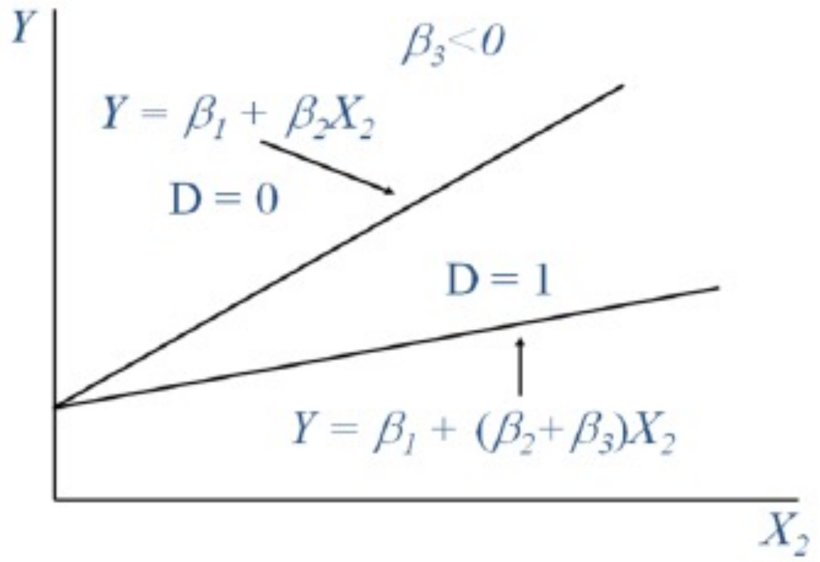
Dummy variables

Slope dummy

Assume no effects on constant term

Changes in slope (marginal effects are affected)

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 (D \cdot X_2) + u$$



Interpretation: Reference category is $D = 0$

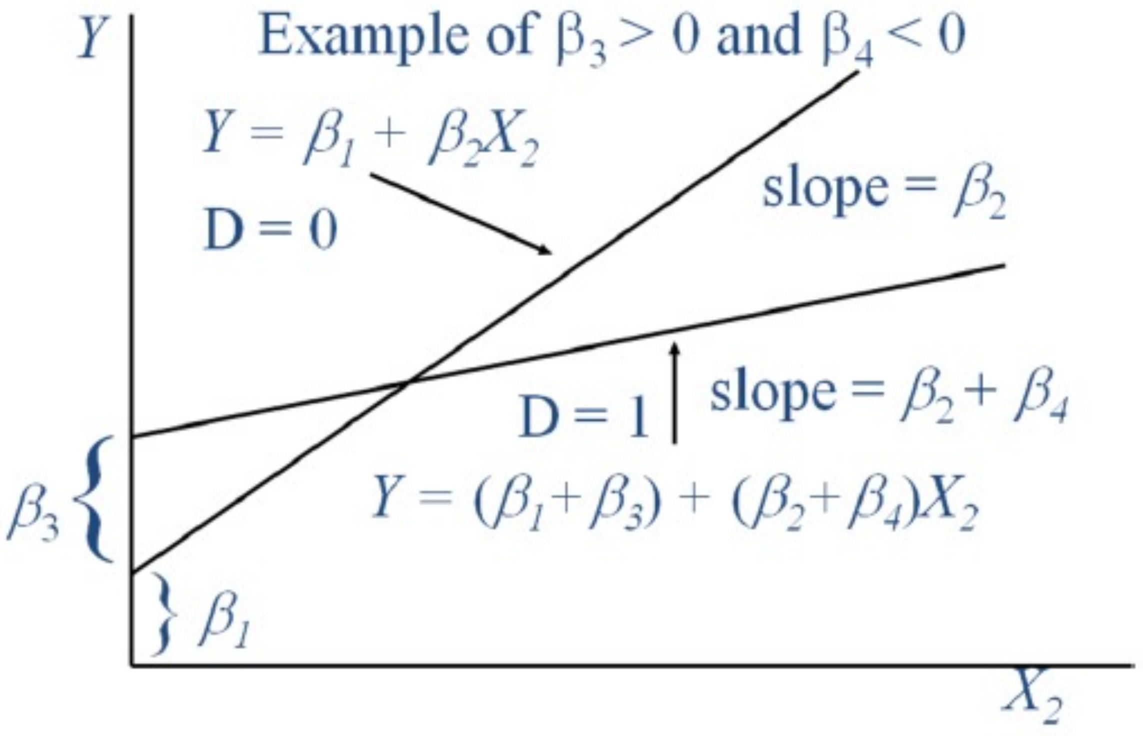
If $D = 0$, then $Y = \beta_1 + \beta_2 X_2 + u$;

If $D = 1$, then $Y = \beta_1 + (\beta_2 + \beta_3) X_2 + u$

Dummy variables

- Combine slope and intercept dummy

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 \cdot D + \beta_4 (D \cdot X_2) + u$$


Y
Example of $\beta_3 > 0$ and $\beta_4 < 0$
$Y = \beta_1 + \beta_2 X_2$
D = 0
slope = β_2
D = 1
slope = $\beta_2 + \beta_4$
$Y = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_2$
β_3
β_1
X_2

If $D = 0$, then $Y = \beta_1 + \beta_2 X_2 + u$;

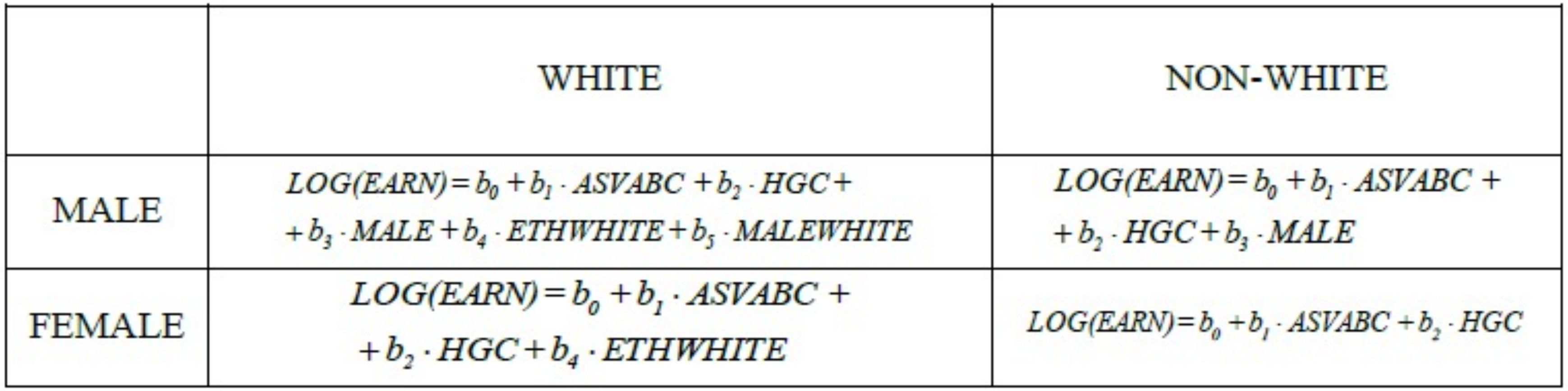
If $D = 1$, then $Y = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_2 + u$.

Dummy variables of interaction

- Example: assume 2 qualitative characteristics affects simultaneously dependent variable

$$Male = \begin{cases} 1 \text{ for male} \\ 0 \text{ for female} \end{cases} \qquad ETHWHITE = \begin{cases} 1 \text{ for White ethnicity} \\ 0 \text{ for otherwise} \end{cases}$$

$$LOG(EARN) = \beta_0 + \beta_1 \cdot ASVABC + \beta_2 \cdot HGC + \beta_3 \cdot MALE + \beta_4 \cdot ETHWHITE + \beta_5 \cdot MALEWHITE + u$$



	WHITE	NON-WHITE
MALE	$LOG(EARN) = b_0 + b_1 \cdot ASVABC + b_2 \cdot HGC +$ $+ b_3 \cdot MALE + b_4 \cdot ETHWHITE + b_5 \cdot MALEWHITE$	$LOG(EARN) = b_0 + b_1 \cdot ASVABC +$ $+ b_2 \cdot HGC + b_3 \cdot MALE$
FEMALE	$LOG(EARN) = b_0 + b_1 \cdot ASVABC +$ $+ b_2 \cdot HGC + b_4 \cdot ETHWHITE$	$LOG(EARN) = b_0 + b_1 \cdot ASVABC + b_2 \cdot HGC$

Try to find out whether there are ethnic variations in effects of the gender of a respondent on earning?

Remember: OLS estimators are invariant in case of linear transformations

Dummy variables

Testing

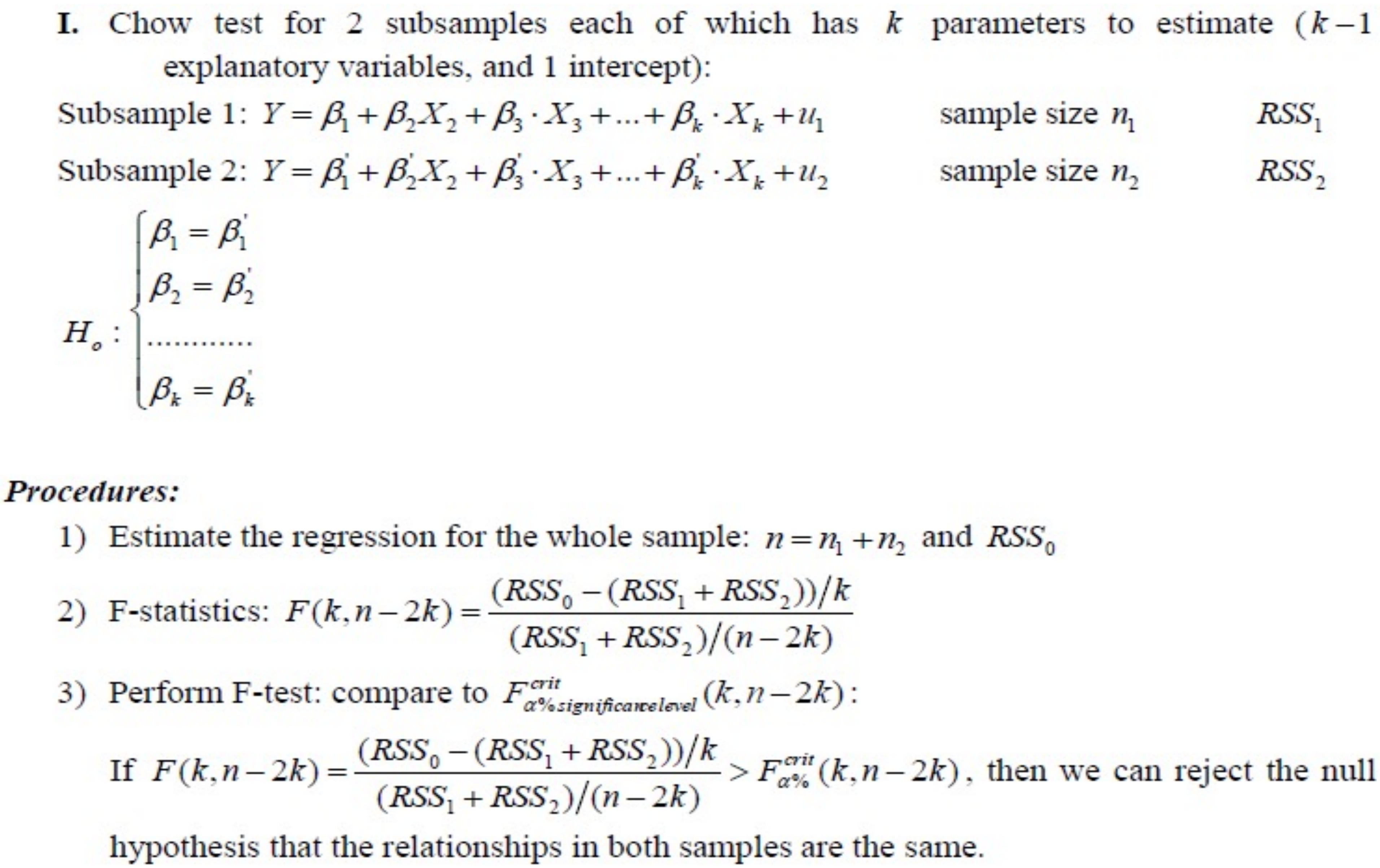
1. t-test
   1. Individual influence on slope/intercept
   2. Ho: $\beta_k = 0$
2. F-test
   1. Joint explanatory power of all dummies
   2. Ho: $\beta_1 = \beta_2 = ... = \beta_k$
   3. Ha: at least one non zero

$$F = \frac{(RSS_{no\ dumies} - RSS_{dummies}) / the\ number\ of\ dummies}{RSS_{dummies} /(the\ number\ of\ observations - the\ total\ number\ of\ parameters\ estimated)}$$

$$\overset{H_0}{\sim} F(number\ of\ dummies, (number\ of\ observations - the\ total\ number\ of\ parameters\ estimated))$$

3. Chow test
   1. Use dummied or separate regression?
   2. Ho: coeff. are the same for all subsamples
   3. Ha: at least one differs

Chow test



I. Chow test for 2 subsamples each of which has k parameters to estimate ($k-1$ explanatory variables, and 1 intercept):

Subsample 1: $Y=\beta_1+\beta_2X_2+\beta_3\cdot X_3+...+\beta_k\cdot X_k+u_1$ sample size n_1 RSS_1

Subsample 2: $Y=\beta_1^{'}+\beta_2^{'}X_2+\beta_3^{'}\cdot X_3+...+\beta_k^{'}\cdot X_k+u_2$ sample size n_2 RSS_2

$$H_o:\begin{cases}\beta_1=\beta_1^{'}\\ \beta_2=\beta_2^{'}\\ \text{...........}\\ \beta_k=\beta_k^{'}\end{cases}$$

Procedures:

1) Estimate the regression for the whole sample: $n=n_1+n_2$ and RSS_0
2) F-statistics: $F(k,n-2k)=\dfrac{(RSS_0-(RSS_1+RSS_2))/k}{(RSS_1+RSS_2)/(n-2k)}$
3) Perform F-test: compare to $F^{crit}_{\alpha\% significancelevel}(k,n-2k)$:

   If $F(k,n-2k)=\dfrac{(RSS_0-(RSS_1+RSS_2))/k}{(RSS_1+RSS_2)/(n-2k)}>F^{crit}_{\alpha\%}(k,n-2k)$, then we can reject the null hypothesis that the relationships in both samples are the same.

Sum up

- Dummy are often used
- We can use dummy as regressand (Y) but we will discuss it later
- You can use either dummy restrictions test or Chow subsample test to check the dummy effect
- Don’t forget about dummy trap
- If you have intercept, coefficients before dummies show absolute changes for the chosen categories, if you have not got intercept, then coefficients show absolute levels
- Change of a reference category does not change conceptually your model

VARIABLE MISSPECIFICATION

Consequences of variable misspecification

		True model	
		$\boldsymbol{Y = \beta_1 + \beta_2 X_2 + u}$	$\boldsymbol{Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u}$
Fitted model	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$	**Correct specification, no problems**	***Omission of a relevant variable*** **Coefficients are biased (in general). Standard errors are invalid.**
	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$	***Including of irrelevant v.*** **Coefficients are unbiased, but inefficient. Standard errors are valid**	**Correct specification, no problems**

There are two types of Variable Misspecification: Omission of a relevant variable and Including an irrelevant one.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ – true model $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$ – fitted model (X$_3$ omitted)

$$\hat{\beta}_2 = \frac{\sum\left(X_{2i} - \bar{X}_2\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_{2i} - \bar{X}_2\right)^2} =$$

$$= \beta_2 + \beta_3 \frac{\sum\left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right)}{\sum\left(X_{2i} - \bar{X}_2\right)^2} + \frac{\sum\left(X_{2i} - \bar{X}_2\right)\left(u_i - \bar{u}\right)}{\sum\left(X_{2i} - \bar{X}_2\right)^2}$$

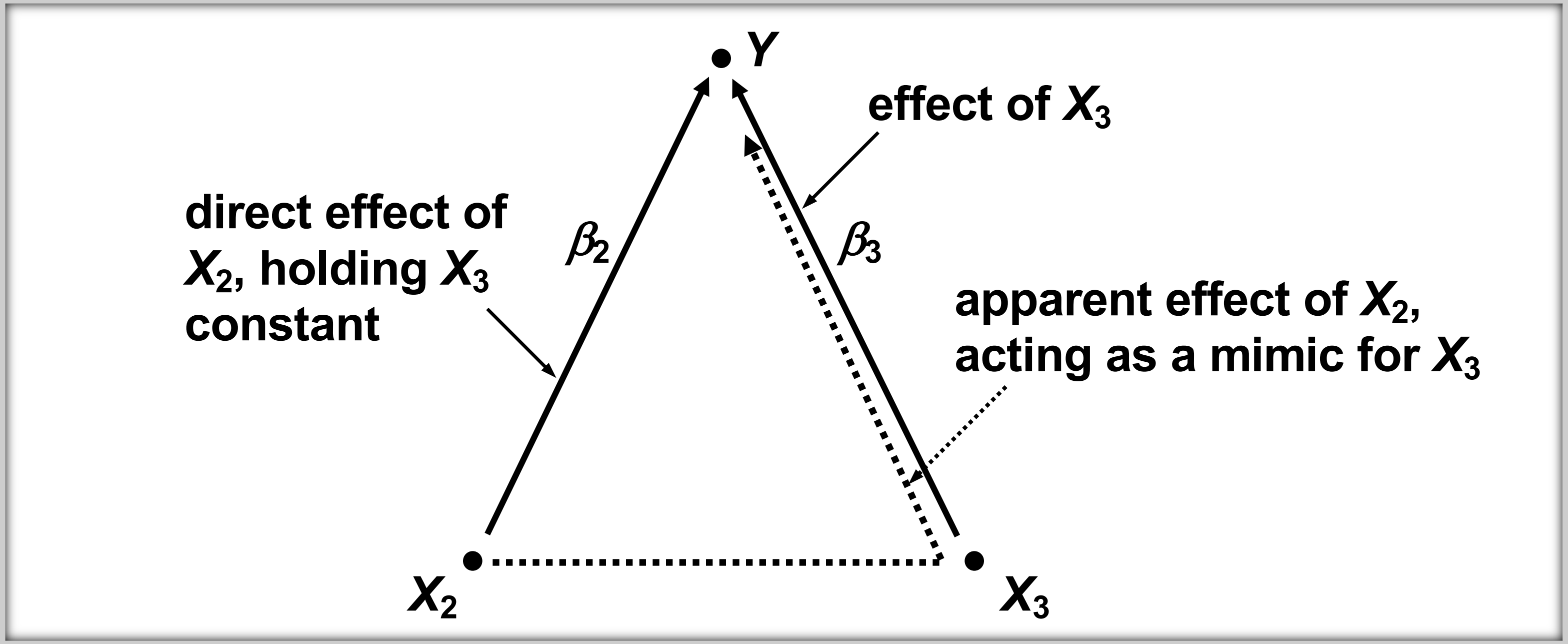
$$E\left(\hat{\beta}_2\right) = \beta_2 + \beta_3 \frac{\sum\left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right)}{\sum\left(X_{2i} - \bar{X}_2\right)^2} + 0$$

The expected value of $\hat{\beta}_2$ is equal to the true value β_2 plus bias term. As a consequence of the misspecification, the standard errors, *t* tests and *F* test are invalid.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$$\boldsymbol{Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u} \qquad \hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$$

$$E(\hat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$



Y
effect of *X*₃
direct effect of *X*₂, holding *X*₃ constant
β_2
β_3
apparent effect of *X*₂, acting as a mimic for *X*₃
*X*₂
*X*₃

The reason is that, in addition to its direct effect β_2, X_2 has an apparent indirect effect as a consequence of acting as a proxy for the missing X_3.

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

$$\boldsymbol{Y = \beta_1 + \beta_2 X_2 + u}$$

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

$$Y = \beta_1 + \beta_2 X_2 + 0X_3 + u$$

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

The estimator of β_2 in the multiple regression model is less efficient than the alternative one in the simple regression model. The standard errors remain valid, because the model is formally correctly specified, but they will tend to be larger than those obtained in a simple regression, reflecting the loss of efficiency.

PROXY VARIABLES

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

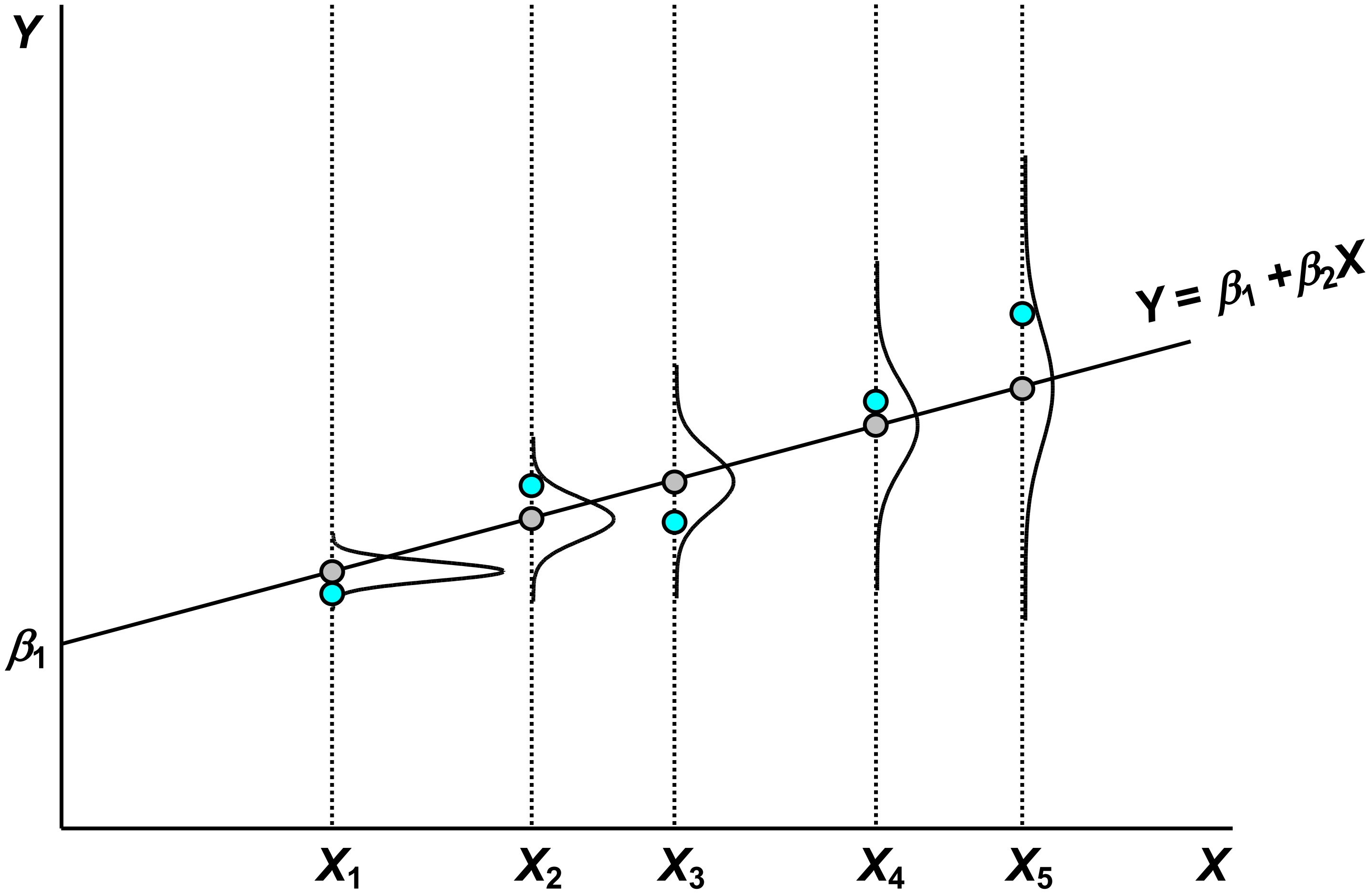
$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u \end{aligned}$$

Comparison of regression with *Z* instead of X_2

1. The estimates for β_3, ..., β_κ are the same
2. S.e. and *t* for *the* estimates *of* β_3, ..., β_κ are the same
3. R^2 is the same
4. Impossible to obtain an estimate of β_2, unless μ is known
5. *t* statistic for *Z* is the same as that for X_2
6. Impossible to obtain an estimate of β_1

Suppose that a variable *Y* depends on a set of explanatory variables X_2, ..., X_k, and there are no data on X_2. Regression of *Y* on X_3, ..., X_k would yield biased estimates and invalid standard errors and tests. Suppose that Z is linearly related with X_2 and there is data for Z.

A.4 (G-M 2) violatoin: Heteroscedasticity, $\sigma_{u_i}^2 \neq \sigma_u^2$



Consequences of heteroscedasticity:

1. Standard errors of the regression coefficients are estimated wrongly and the *t* tests (and *F* test) are invalid.

2. OLS estimators are inefficient (though still unbiased).

A.4 violation. Heteroscedasticity

Consequences (why it is a problem?)

1. **Origins**
    I. Data nature – true heteroscedasticity
    II. Model misspecification – apparent heteroscedasticity

2. **Consequences**
    1. OLS estimation – inefficient
        1. Not BLUE
        2. Unbiased, not efficient
    2. S.E. is based on incorrect formula
    3. *t* and *F* tests are invalid

Note that heteroscedasticity is not connected with endogeneity (*X* is non-stochastic!)

- Wrong s.e. logic:
    - S.e are estimates of s.d.
    - S.e. assume distribution of *u* is homoscedastic
    - Assumption is wrong
    - S.e. biased (downwards or upwards, depending on data specifics)

A.4 violation. Heteroscedasticity

Detection

1. Goldfeld-Quandt test

Assumption:

Ho: homoscedasticity $\sigma_{u_i}^2 = \sigma_u^2$

H1: heteroscedasticity $\sigma_i = \gamma \cdot x_i$

Important: very restrictive assumption about variance formula

1. Arrange all observations using the proposed factor *X*
2. Divide the sample into three parts (3/8;2/8;3/8 gives the highest power of the test)
3. *F=SSR(highest)/SSR(lowest)*

HETEROSCEDASTICITY DETECTION:

GENERAL CASE

General: White test for detection of any form of association between σ_i^2 and the regressors. Since σ_i^2 is unobservable, $\hat{u}_i^{\,2}$ is used as a proxy.

The White test consists of two steps:

1. Regressing the squared residuals on the explanatory variables in the model, their squares, and their cross-products

2. Test statistic nR^2 is calculated. Under the H_0 of homoscedasticity, it is distributed as a chi-squared statistic with degrees of freedom equal to the number of regressors (in large samples).

F-test for the auxiliary regression can be also done.

So, after estimating the model $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$

$$\begin{gathered}
\textit{the regression} \\
\hat{u}_i^{\,2} = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i}^{\;2} + \beta_5 X_{3i}^{\;2} + \beta_6 X_{2i} X_{3i} + v_i \\
\textit{is estimated, and } \chi^2 \textit{ test or } F-\textit{test done.}
\end{gathered}$$

A.4 violation. Heteroscedasticity

Remedial Measures

1. Weighted Least Squares
2. White’s heteroscedasticity consistent s.e.
3. Log model can decrease a degree of heteroscedasticity

A.4 violation. Heteroscedasticity

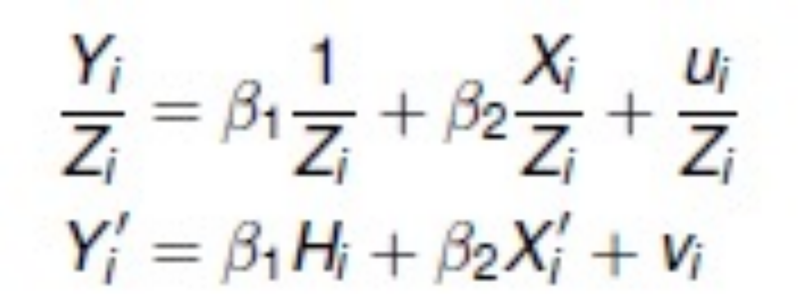
Remedy (how to cope with it?)

1. Weighted Least Squares (special of Generalized Least Squares)

<u>Assumption:</u>

$$Y = \beta_1 + \beta_2 X_i + u - \text{true model}$$

$$\sigma_{u_i}^2 = \lambda^2 Zi^2$$



$$\frac{Y_i}{Z_i} = \beta_1 \frac{1}{Z_i} + \beta_2 \frac{X_i}{Z_i} + \frac{u_i}{Z_i}$$

$$Y_i' = \beta_1 H_i + \beta_2 X_i' + v_i$$

Transformed model:

$$\text{var}(v_i) = \text{var}(\frac{u_i}{Z_i}) = \frac{Z_i^{\,2} \lambda^2}{Z_i^{\,2}} = \lambda^2$$

- Homoscedastic
- OLS on transformed model is BLUE
- Special case $Z_i = X_i$

A.4 violation. Heteroscedasticity

3. **White’s heteroscedasticity consistent s.e.**

In 1980 White showed how to get heteroscedasticity consistent s.e. based on residuals

$$\text{Standard errors: } s_{\hat{\beta}_2} = \sqrt{\frac{\sum_{i=1}^{n} x_i{}^2 \hat{u}_i^2}{\left(\sum_{j=1}^{n} x_j{}^2\right)^2}} = \sqrt{\sum_{i=1}^{n} a_i{}^2 \hat{u}_i^2}$$

Advantages:

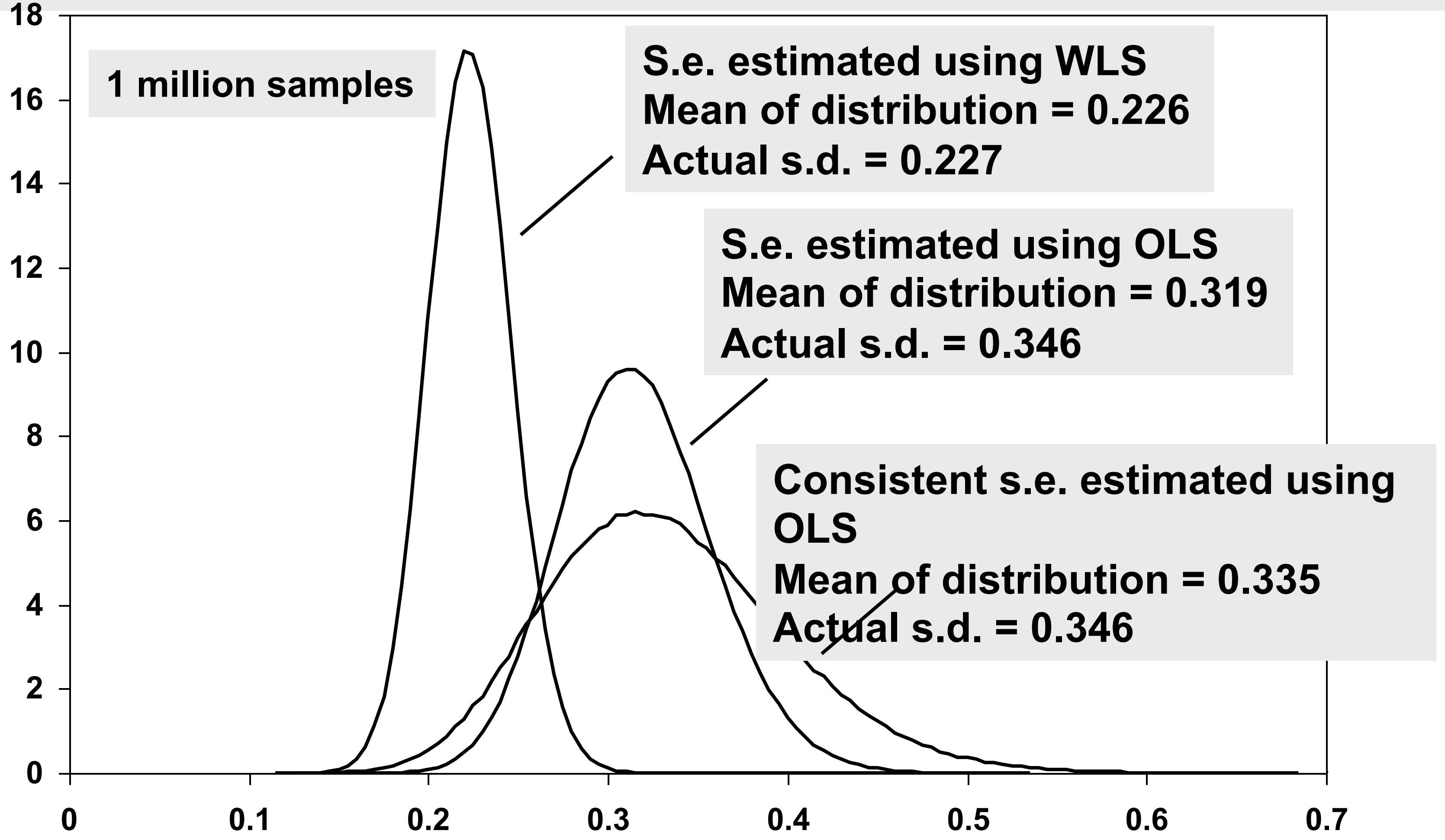
- Do not need to specify a type of heteroscedasticity
- Easy: build in STATA and Eviews

Disadvantages:

- Designed for large samples (how for small?)

HETEROSCEDASTICITY: MONTE CARLO ILLUSTRATION

$$Y_i = 10 + 2.0X_i + u_i \quad X_i = \{5,6, ..., 54\} \quad u_i = X_i\varepsilon_i \quad \varepsilon_i \sim N(0,1)$$



1 million samples

S.e. estimated using WLS
Mean of distribution = 0.226
Actual s.d. = 0.227

S.e. estimated using OLS
Mean of distribution = 0.319
Actual s.d. = 0.346

Consistent s.e. estimated using OLS
Mean of distribution = 0.335
Actual s.d. = 0.346

Consistent standard errors are valid only in large samples. For small samples, their properties are unknown. They can mislead even more than the OLS standard errors.