LECTURE 4:
MULTIPLE REGRESSION

Jan Zouhar

Introductory Econometrics

What is Multiple Regression?

- multiple regression = regression with multiple explanatory variables
- **notation & terminology** that we'll use in all the formulas:
  - k number of explanatory variables
  - n number of observations
  - y the explained (dependent) variable
  - $x_1, x_2, \ldots, x_k$ explanatory variables
  - i subscript that indicates the *observation number*
  - j subscript that indicates the *explanatory variable*
  - x_{ij} the ith observation of x_j
  - if we **regress** y **on** $x_1, x_2, \ldots, x_k$, it means we work with the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

  - *note*: in literature, matrix/vector notation is heavily used in multiple regression. I'll try to avoid it in this course. However, sometimes we'll use the matrix symbol $\mathbf{X}$ to refer to all the available data on explanatory variables.

Why Do We Need Multiple Regression?

Reason 1

- we've seen the motivation for multiple regression in our discussion of the *omitted variable bias* in the previous lecture
- if we need to estimate the effect of x on y, believing we have a problem with the $x \leftarrow z \rightarrow y$ relationship, we can solve the problem with regressing y on x *and* z rather than on x only
- primary drawback of the simple regression:
  - *statistics*: breach of the $\mathrm{E}[u \mid x] = 0$ assumption (SLR.4)
  - *economics*: difficult to draw *ceteris paribus* conclusions about how x affects y (causal interpretation)
- multiple regression allows us to explicitly control for many other factors which simultaneously affect the dependent variable
- once we control for a factor, the *ceteris paribus* condition with respect to this factor is automatically fulfilled (see later)

Why Do We Need Multiple Regression? (cont'd)

- **example**: wages vs. education
  - imagine we want to measure the (causal) effect of an additional year of education on a person's wage
  - if we want to the model $wage = \beta_0 + \beta_1 educ + u$ and interpret β_1 as the *ceteris paribus* effect of *educ* on *wage*, we have to assume that *educ* and *u* are uncorrelated (SLR.4)
  - consider a different model now: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$, where *exper* is a person's working experience (in years)
    - since the equation contains experience explicitly, we will be able to measure the effect of education on wage, *holding experience fixed*
    - this is still far from "complete" *ceteris paribus*, but we're definitely getting closer
    - we'll still have to rule out the correlation between *educ* and *u*
    - by including *exper* in our equation, we effectively pulled it out of *u*
  - note that in the new model, our primary interest is still the value of β_1, *exper* is only a *control variable* here

Why Do We Need Multiple Regression? (cont'd)

Reason 2

- multiple regression analysis is also useful for generalizing functional relationships between variables
- **example**: consumption vs. income
  - suppose family consumption (*cons*) is a quadratic function of family income (*inc*):

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$

  - in reality, consumption is dependent on only one observed factor, income; formally, we can treat this as a linear regression model with two variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

    where $x_1 = inc$ and $x_2 = inc^2$.
  - note that the interpretation of the coefficients has to be adjusted
  - obviously, β_1 is not the effect of a unit change in *inc* on *cons* now; it makes no sense to measure the effect of *inc* while holding inc^2 fixed

Simple Regression vs. Multiple Regression

- as you'll see, most of the properties of the simple regression model directly extend to the multiple regression case (perhaps even those you would not expect to)

  → basically, we'll keep talking about the same principles
- we derived many of the formulas for the simple regression model; however, with multiple variables, formulas can get pretty messy
- therefore, I'll just give you the results in most cases; most of the derivations can be found in Wooldridge or other textbooks
- as far as the interpretation of the model is concerned, there's a new important fact:
  - the coefficient β_j captures the effect of jth explanatory variable, *holding all the remaining explanatory variables fixed*
  - this is true even if we don't have even a single pair of observations where "all the remaining explanatory variables" are identical

  → this brings us closer to the *ceteris paribus* setting used in laboratory experiments in natural sciences (see lecture 1)

Estimation

- recall that in simple regression, we derived the formulas for estimates in three different ways, taking on the *descriptive*, *causal* and *forecasting* approach
- we could do the same thing with multiple explanatory variables
  - *descriptive approach*: conditional expectation:

    $$\mathrm{E}[y \mid x_1,\dots,x_n] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

  - *causal approach*: structural model:

    $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

    - + assumptions about u: $\mathrm{E}[u \mid x_1,\dots,x_n] = 0$

  } estimation using the idea of *sample analogue*

  - *forecasting approach*: best fit of the approximate relationship

    $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

  } best fit using OLS

Estimation (cont'd)

- as in simple regression, the resulting estimates are identical
- similarly as before, we can define:
  - **population regression model**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

  - **sample regression model**:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$$

  - **fitted values of** y:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$$

  - **residuals**:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$$

Estimation (cont'd)

- **sample analogue**:
  - from the population model (+ assumptions about u), we know

$$\begin{aligned} \mathsf{E}u &= 0 \\ \mathsf{E}[x_1 u] &= 0 \\ &\vdots \\ \mathsf{E}[x_k u] &= 0 \end{aligned}$$

  - the sample analogue of this is

$$\left.\begin{aligned} \tfrac{1}{n}\textstyle\sum \hat{u}_i &= 0 \\ \tfrac{1}{n}\textstyle\sum x_{i1}\hat{u}_i &= 0 \\ &\vdots \\ \tfrac{1}{n}\textstyle\sum x_{ik}\hat{u}_i &= 0 \end{aligned}\right\} \text{normal equations}$$

  - substituting $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$ gives you a system of $k + 1$ linear equations with $k + 1$ unknowns $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

Estimation (cont'd)

- **OLS:**
  - with OLS, we minimize the residual sum of squares, SSR
  - as with a single variable, we have to set all partial derivatives to zero:

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = -2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\frac{\partial SSR}{\partial \hat{\beta}_1} = -2\sum x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\vdots$$

$$\frac{\partial SSR}{\partial \hat{\beta}_k} = -2\sum x_{ik}\underbrace{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})}_{\hat{u}_i} = 0$$

  - we're back at the same system of equations
  - we won't solve this explicitly here (this can only be done in matrix notation)
  - instead, we'll let *Gretl* do the job

Three Facts About The Fitted Values and Residuals

1. The sample average of the residuals is zero (see the first of *normal equations*). Consequently, sample average of the fitted values equals the sample average of y.

$$\bar{\hat{y}} = \frac{1}{n}\sum \hat{y}_i = \frac{1}{n}\sum (y_i - \hat{u}_i) = \underbrace{\frac{1}{n}\sum y_i}_{\bar{y}} - \underbrace{\frac{1}{n}\sum \hat{u}_i}_{0} = \bar{y}$$

2. The sample covariance between each independent variable and the residuals is zero (see *normal equations* again). Consequently, the sample covariance between the fitted values and the residuals is zero.

$$\frac{1}{n-1}\sum \hat{y}_i\hat{u}_i = \frac{1}{n-1}\sum (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})\hat{u}_i =$$

$$= \frac{1}{n-1}\left(\hat{\beta}_0 \underbrace{\sum \hat{u}_i}_{0} + \hat{\beta}_1 \underbrace{\sum x_{i1}\hat{u}_i}_{0} + \ldots + \hat{\beta}_k \underbrace{\sum x_{ik}\hat{u}_i}_{0} \right) = 0$$

3. The point $(\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$ always lies on the regression "line", i.e. $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_k\bar{x}_k$ (this follows immediately from 1).

Goodness of Fit

- recall that before we defined:
  - **total sum of squares** (SST): $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - **explained sum of squares** (SSE): $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
  - **residual sum of squares** (SSR): $SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2$
- we define them exactly the same way for the multiple regression model
- it is straightforward to show that once again

$$SST = SSE + SSR$$

- therefore, we can still use

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

- however, there is an interesting property here:
  - what happens to the R^2 when we add a regressor to the model?

Goodness of Fit (cont'd)

The effect of an additional regressor on R^2

- intuitively, R^2 should go up; we'll show this is mathematically true
- imagine we estimate two regression models:

$$y = \beta_0 + \beta_1 x_1 + u \qquad (1)$$

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + u \qquad (2)$$

- as we know, OLS tries to minimize $SSRs$ in both models (we'll denote them SSR_1 and SSR_2)
- let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimates of parameters in model (1)
- if in model (2) I take

$$\hat{\gamma}_0 = \hat{\beta}_0, \quad \hat{\gamma}_1 = \hat{\beta}_1, \quad \hat{\gamma}_2 = 0,$$

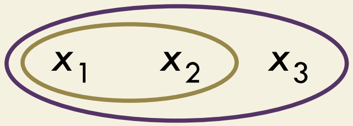
  I'll have $SSR_1 = SSR_2$, and the R^2's will be equal
- unless x_2 is completely useless in explaining y, OLS will come up with something better than me, resulting in a higher R^2

Goodness of Fit (cont'd)

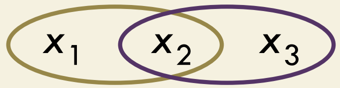
- economists sometimes like to choose models based on R^2 (we'll talk about this in more detail later on)
- this means that you'll always end up adding more variables when comparing **nested models**

Nested vs. non-nested models:

- nested models:
  - model **1**: regress y on x_1, x_2
  - model **2**: regress y on x_1, x_2, x_3



- non-nested models:
  - model ***A***: regress y on x_1, x_2
  - model ***B***: regress y on x_1, x_3



- because of this, econometricians came up with **adjusted R-squared**, which enables a comparison of nested models (even though this should be done with care)

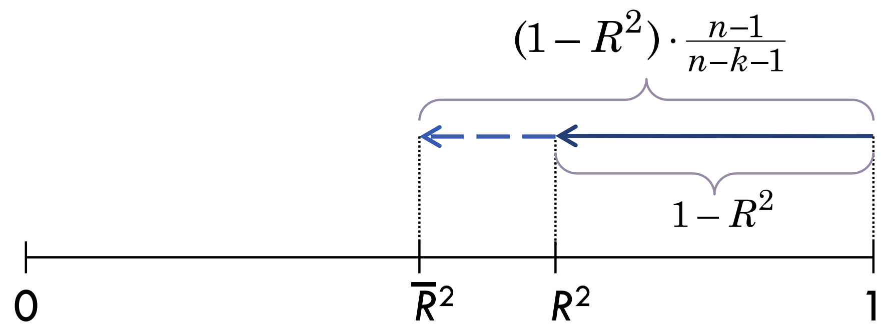
Goodness of Fit (cont'd)

Adjusted R-squared ($\bar{R}^2$)

- the basic idea is to take R^2 and penalize for additional regressors
- as I'll be telling you, additional regressors can cause trouble especially when we have few observations → the correction has to account for this
- the resulting formula is

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

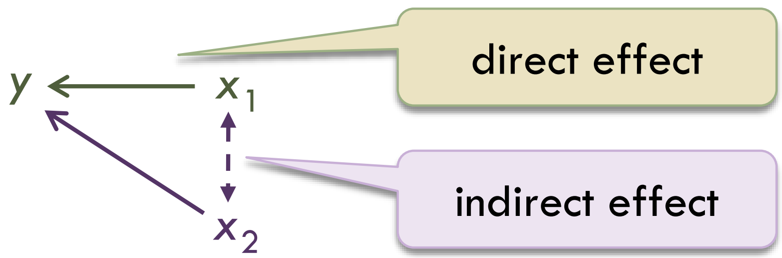
this is greater than 1 and grows with k



- as you can see, $\bar{R}^2 < R^2$ unless $R^2 = 1$ (this is *very* unlikely) or $k = 0$ (no regressors) or $n \leq k + 1$ (too few obs., but OLS doesn't work then)

A "Partialling Out" Interpretation

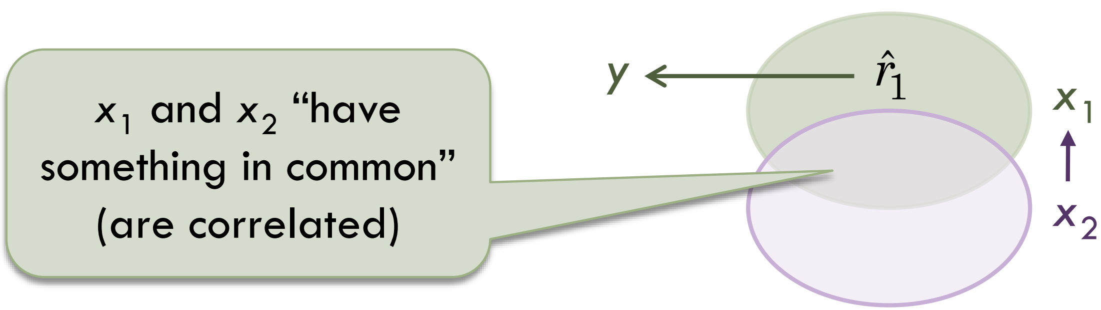
- consider a regression of y on x_1, x_2
  - when we talked about the omitted variable bias, we saw that the bias in the estimate of β_1 (when regressing y on x_1 only) was due to the indirect effect $x_1 \rightarrow x_2 \rightarrow y$



  - we would somehow like to show that this problem has been fixed here, that the path $x_1 \rightarrow x_2 \rightarrow y$ has been "blocked" in $\hat{\beta}_1$ by adding x_2 explicitly in the model, and that $\hat{\beta}_1$ contains only the direct effect $x_1 \rightarrow y$
  - in other words, we want to show that $\hat{\beta}_1$ captures the partial effect of x_1 on y, after the effect of the other variables has been accounted for

A "Partialling Out" Interpretation (cont'd)

- if I need to know the "net" effect of x_1 with x_2 being "partialled out", I can proceed as follows:
  1. first, I run the regression of x_1 on x_2, and save the residuals (I'll denote these as $\hat{r}_1$)
     - the residuals represent whatever is left in x_1 after we subtract all that x_1 has in common with x_2
  2. next, I run the regression of y on $\hat{r}_1$
- fortunately, the $\hat{\beta}_1$ from $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{r}_1$ is numerically identical to $\hat{\beta}_1$ from $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ (see Wooldridge, page 77), so that the latter is already "partialled out"

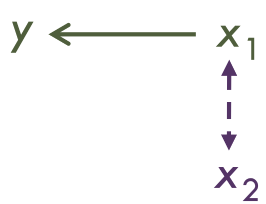


Simple Vs. Multiple Regression Estimates

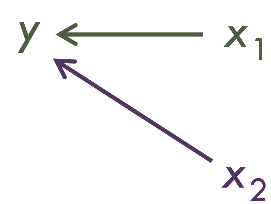
- imagine we estimate two regression models:

$$y = \beta_0 + \beta_1 x_1 \qquad\quad + u$$
$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + u$$

- is it possible that $\hat{\beta}_1$ and $\hat{\gamma}_1$ are identical?
  - theoretically, yes, but one of the two situations would have to arise:
    1. the partial effect of x_2 on y is zero in the sample, i.e., $\hat{\gamma}_2 = 0$
    2. x_1 and x_2 are uncorrelated in the sample
  - in both cases, there's no indirect effect (no $x_1 \rightarrow x_2 \rightarrow y$ path)



no partial effect of x_2 on y



x_1 and x_2 uncorrelated

The Expected Value of OLS Estimators

- now we'll talk about some statistical properties of the OLS estimators (i.e., expected values, variances, sampling distributions)
- remember that statistical properties have nothing to do with a particular sample, but rather with the property of estimators when random sampling is done
- again, we'll need a set of assumptions about the model

Assumption **MLR.1** (linear in parameters) :

The population model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are are the unknown parameters (constants) of interest, and u is an unobservable random error or random disturbance term.

The Expected Value of OLS Estimators (cont'd)

Assumption **MLR.2** (random sampling):

We have a random sample of size n, $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$, $i = 1, \dots, n$ following the population model defined in MLR.1.

Assumption **MLR.3** (no perfect collinearity):

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear relationships* among the independent variables.

- note that SLR.3 was telling us that there is sample variation in x
- now, not only do we need variation in all explanatory variables, but we need them to vary *separately*

The Expected Value of OLS Estimators (cont'd)

- estimating this equation clearly is a problem:
$$y = \beta_0 + \beta_1 x + \beta_2 x + u$$
- this is the same as estimating
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
where $x_1 = x_2$
- it doesn't help if x_1 and x_2 are scaled differently: if $x_2 = cx_1$, we have
$$y = \beta_0 + \beta_1 x_1 + (c\beta_2)x_1 + u,$$
which is no better than before
- it doesn't even help if $x_2 = cx_1 + d$, we're back at the same problem:
$$y = (\beta_0 + d\beta_2) + \beta_1 x_1 + (c\beta_2)x_1 + u,$$
- mostly, if you encounter a relationship like this in your data, you've done something wrong, such as estimating
$$\log(cons) = \beta_0 + \beta_1 \log(inc) + \beta_2 \log(inc^2) + u$$
- *Quizz*: what's the problem with this equation?

The Expected Value of OLS Estimators (cont'd)

Assumption **MLR.4** (zero conditional mean of u):

The error u has an expected value of zero, given any value of the independent variables. In other words, $E[u \mid x_1, \ldots, x_k] = 0$.

- note that this implies $E[u \mid x_j] = 0$ for any j, i.e., all the explanatory variables are uncorrelated with u
  - remember all the implications of this on causality issues
- again, we can't test this assumption with statistical means
- possible violations
  - correlation of x_j and u (can be sometimes argued from outside)
  - misspecification of the model form:
    - omitting an important variable
    - using the wrong functional form (using the level-level form instead of log-log or log-level etc.)

The Expected Value of OLS Estimators (cont'd)

Theorem: Unbiasedness of OLS

Under the assumptions MLR.1 through MLR.4, the OLS estimators are *unbiased*. In other words,

$$\mathrm{E}[\hat{\beta}_j] = \beta_j, \quad j = 0, 1, \ldots, k$$

for any values of the population parameter β_j.

- for a proof, would need an explicit formula for $\hat{\beta}_j$, which we haven't derived here; then, the proof is similar as in the simple regression case
- *note*:
  - we cannot use the unbiasedness property to say things like: "my estimate of β_1, namely 3.5, is unbiased"
  - unbiasedness = a property of the *estimator*, not the *estimate*!
  - it tells us that if we collected multiple random samples, OLS doesn't systematically overestimate or underestimate the real values

Including Irrelevant Variables in a Regression Model

- suppose we regress y on x_1 and x_2, even though x_2 has no partial effect on y in our population, i.e., $\beta_2 = 0$ in the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- obviously, it's not a clever thing to include x_2 in the regression model, but sometimes we just don't know x_2 is irrelevant
- the question is, did we cause any harm to the estimate of β_1?
  - in terms of *unbiasedness*, the answer is *no*
    - we know that all OLS estimates are unbiased for *any values of* β_2, including zero
    - note that we know that *not including* an important variable may cause a bias (*omitted variable bias*)
  - however, including irrelevant variables (or *overspecifying the model*) reduces the *accuracy* of the estimated coefficients
    - more precisely, including x_2 in the equation above typically increases the sampling variance of $\hat{\beta}_1$

Including Irrelevant Variables in a Regression Model

- suppose we regress y on x_1 and x_2, even though x_2 has no partial effect on y in our population, i.e., $\beta_2 = 0$ in the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- obviously, it's not a clever thing to include x_2 in the regression model, but sometimes we just don't know x_2 is irrelevant
- the question is, did we cause any harm to the estimate of β_1?
  - in terms of *unbiasedness*, the answer is *no*
    - we know that all OLS estimates are unbiased for *any values of* β_2, including zero
    - note that we know that *not including* an important variable may cause a bias (*omitted variable bias*)
  - however, including irrelevant variables (or *overspecifying the model*) reduces the *accuracy* of the estimated coefficients
    - more precisely, including x_2 in the equation above typically increases the sampling variance of $\hat{\beta}_1$

The Variance of the OLS Estimators

- in order to describe the variance nicely, we need another assumption (note that MLR.1 through MLR.5 are collectively known as the **Gauss-Markov assumptions** for cross-sectional regression)

Assumption **MLR.5** (homoskedascticity):

$\mathrm{Var}[u \mid x_1,\ldots,x_k] = \sigma^2.$

Theorem: Sampling variances of the OLS estimators

Under assumptions MLR.1 through MLR.5,

$$\mathrm{var}[\hat{\beta}_j \mid x_1,\ldots,x_k] = \frac{\sigma^2}{SST_j(1-R_j^2)}, \qquad j = 1,\ldots,k$$

where $SST_j = \Sigma(x_{ij} - \overline{x}_j)^2$ is the total sample variation in x_j and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

The Variance of the OLS Estimators (cont'd)

- *note*: for $j = 1$, the denominator in the formula for conditional variance contains SST and R^2 from the regression of x_1 on $x_2,\dots,x_k$ (rather than from the "original" regression of y on the x's; y plays no role here)

The more the total variation in x_j, the more accurate the OLS estimates of β_j will be.

What can help: adding more observations increases SST_j.

Remember that σ^2 is the variance of the error term *u*. If the variance of *u* diminishes, the accuracy of OLS estimators grows.

What can help: adding explanatory variables (taking some factors out of *u*)

$$\frac{\sigma^2}{SST_j(1-R_j^2)}$$

If x_j is uncorrelated with other independent variables, this R-squared is zero. With increasing correlation between the *x*'s, the accuracy of OLS estimators diminishes. The possible linear relationship between the *x*'s is called **multicollinearity**.

Multicollinearity

- it is obvious from the formula that for given σ^2 and SST_j, the smallest variance of the OLS estimator is obtained when R_j^2 is zero
  - this happens if, and only if, x_j has zero sample correlation between and *every other* explanatory variable
- remember that R_j^2 tells us the fraction of x_j's sample variance that can be explained with a linear combination of the remaining x's
- let's make clear that we are not talking about *exact* linear relationship between the x's (i.e., *perfect multicollinearity*), which is ruled out by MLR.3
  - exact linear relationship means that e.g. x_j can be expressed as a linear combination of the remaining x's
  - then, $R_j^2 = 1$, and the denominator in the variance formula is zero
  - in practice, violating MLR.3 is either due to an *extremely* bad luck in collecting the data, or (more likely) due to a mistake in putting up the model (see Exercise 4.2 in the tutorials)

Multicollinearity (cont'd)

- an R_j^2 close to 1 does not violate MLR.3, but reduces the accuracy of $\hat{\beta}_j$
- if the linear relationship between the x_j and the remaining x's gets stronger, R_j^2 approaches 1 and the resulting $\mathrm{var}[\hat{\beta}_j]$ grows above all limits
- sometimes the expression $\dfrac{1}{1-R_j^2}$ is called the **variance inflation factor** (VIF_j)
  - this is the terminology that *Gretl* uses; the variance formula becomes:

$$\mathrm{var}[\hat{\beta}_j \mid x_1,\ldots,x_k] = \frac{\sigma^2}{SST_j} \cdot VIF_j$$

- we can see that: $R_j^2 \to 1 \quad \Rightarrow \quad VIF_j \to \infty \quad \Rightarrow \quad \mathrm{var}[\hat{\beta}_1] \to \infty$
- *questions*:
  - in my sample, $R_j^2 = 0.83$. Is it too much?
  - what values of R_j^2 indicate a problem with multicollinearity?
- there's no clear answer (for instance, σ^2 and SST_j matter as well)

Multicollinearity: Is There a Way Out?

- the simple answer is: *not quite*
- remember that we can try to reduce the $\hat{\beta}_j$ variance by:
  - *adding more variables* (reduces σ^2): this can only make the collinearity problem even worse
  - *adding more observations* (increases SST_j): this typically doesn't change R_j^2 (but, on the other hand, it cannot do any harm either)
- **example**: imagine we are interested in estimating the effect of various school expenditure categories (teacher salaries, instructional materials, athletics,…) on student performance
  - it is likely that expenditures on the individual categories are highly correlated (wealthier schools spend more on everything)
  - therefore, it will be difficult to separate the effect of a single category
  - perhaps we are asking a question that may be too subtle for the available data to answer with any precision
  - on the other hand, assessing the effect of total expenditures might be relatively simple (i.e., changing the scope of the analysis might help)

Multicollinearity: Is There a Way Out? (cont'd)

- on the other hand, sometimes we don't really care about multicollinearity among the **control variables**
  - **example**: in our wages vs. education exercise, we (more-or-less) developed the equation to estimate the returns to schooling
  - we decided to regress *wages* on:
    - *education* → this is the key variable
    - *work experience*
    - *age*
    - *industry*
    - …

    (*work experience*, *age*, *industry*, …: control variables, needed for the model to be "correct")
  - we're only really interested in $\beta_{education}$; therefore, we don't mind if the coefficients on the control variables are not quite precise
  - the only thing that really matters is $R^2_{education}$, multicollinearity among the control variables doesn't spoil this

Estimating Standard Errors of the OLS Estimators

- as in the simple regression case, we want to estimate $\text{var}[\hat{\beta}_j]$ or its square root, $\text{sd}(\hat{\beta}_j)$, but the formula we know contains the variance of the random error, σ^2

→ first, we need to estimate σ^2:

Theorem: Unbiased estimation of σ^2

Under the Gauss-Markov assumptions MLR.1 through MLR.5,

$$\hat{\sigma}^2 = \frac{SSR}{n-k-1}$$

is an unbiased estimator of σ^2.

- the logic behind this estimate is the same as with simple regression
- the term $n - k - 1$ is the **degrees of freedom** (df) of the regression:

$$df = \text{no. of observations} - \text{no. of estimated parameters}$$

Estimating Standard Errors of the OLS Estimators (cont'd)

- the square root of $\hat{\sigma}^2$ is called the **standard error of the regression**
- it's important to distinguish between the *standard deviation* and the *standard error* of $\hat{\beta}_1$
  - the standard deviation of $\hat{\beta}_1$ is the square root of the conditional variance of $\hat{\beta}_1$ (for brevity, we typically omit the conditioning in the formulas)

$$\text{sd}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{SST_j(1 - R_j^2)}}$$

  - the standard error of $\hat{\beta}_1$ is the thing we can calculate in practice

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)}}$$

  - this is the standard error reported by *Gretl* and other stat. packages
  - *note*: se relies on $\hat{\sigma}^2$; therefore, se is a valid estimator of sd only if the homoskedasticity assumption is fulfilled

Gauss-Markov Theorem

- Gauss-Markov theorem justifies the use of OLS for multiple regression; it states that OLS is, in certain sense, the best among possible competing estimators
  - we already know one justification of OLS: under MLR.1 through MLR.2, OLS is *unbiased*
  - however, many different unbiased estimators can be developed
  - so what is it that makes OLS so good? intuitively, we would like the use the most *accurate* estimator; i.e., the estimator with the smallest variance
  - the Gauss-Markov theorem shows that, within a certain class of unbiased estimators, OLS is the one that exhibits the smallest variance among all competing estimators

Theorem: Gauss Markov theorem.

Under the assumptions MLR.1 through MLR.5, OLS estimator is the **best linear unbiased estimator** (BLUE) of the regression coefficients.

Gauss-Markov Theorem (cont'd)

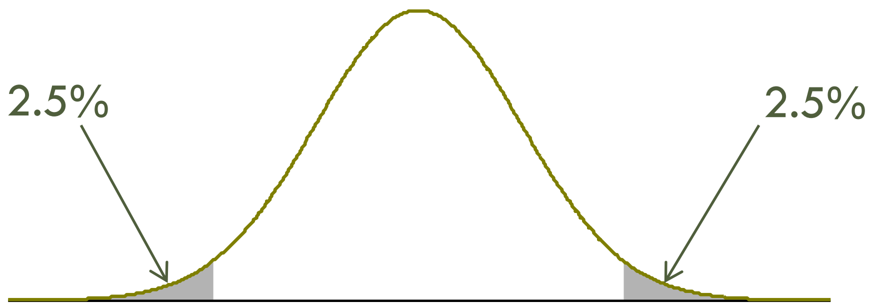
- on the meaning of BLUE:
  - **B**est:
    - "best" actually means "the one with the lowest variance" (or, more generally, the one with the lowest **mean squared error**)
  - **L**inear:
    - an estimator of multiple regression coefficients is **linear**, if the estimate of each of the regression coefficients can be calculated as a *linear combination of the values of the dependent variable* (y), i.e. there exist values w_{ij}, such that the estimate of β_j equals

$$\sum_{i=1}^{n} w_{ij} y_i$$

    - OLS can be shown to be a linear estimator
  - **U**nbiased
    - i.e., the expected value of the estimate of β_j is the real value β_j
  - **E**stimator

Sampling Distribution of the OLS Estimator

- the discussion of the OLS sampling distribution in multiple regression will be almost identical to the simple regression case
- recall we need to know the sampling distribution of the estimates in order to carry out hypothesis testing (next lecture)



- we'll start with the model satisfying the MLR.1 through MLR.5 assumptions only; this will only allow for *asymptotical* (or *large-sample*) results
- for small samples, asymptotic analysis is useless; we'll have to add the normality assumption as with simple regression

Sampling Distribution of the OLS Estimator (cont'd)

- once again, we'll work with the standardized estimators: $\dfrac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)}$

Theorem: Asymptotic normality of the OLS estimators

Under the assumptions MLR.1 through MLR.5, as the sample size increases, the distributions of standardized estimators converge towards the standard normal distribution *Normal*(0,1).

- for small samples, we need an additional assumption:

Assumption **MLR.6** (normality):

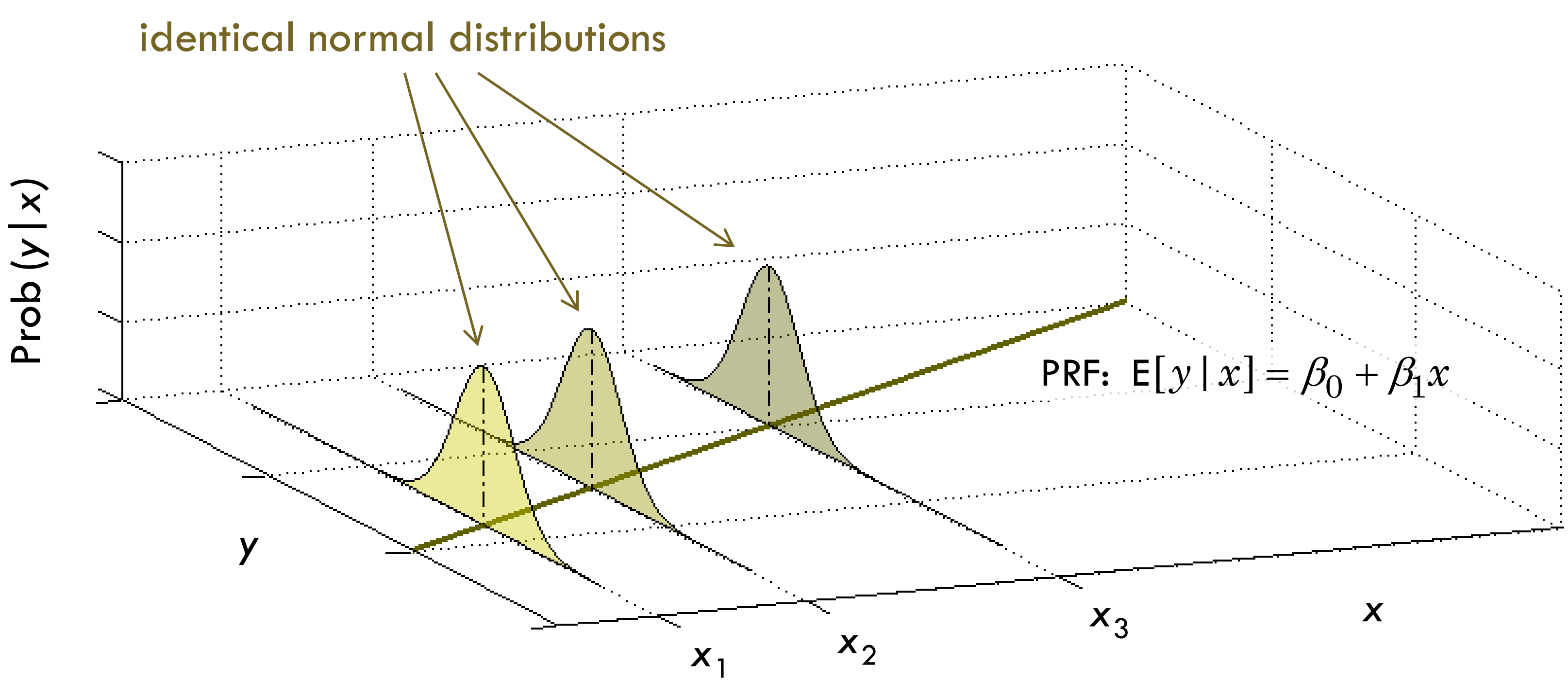
The population error u is *independent* of the explanatory variables and is normally distributed with zero mean and variance σ^2:

$$u \sim Normal(0, \sigma^2).$$

Sampling Distribution of the OLS Estimator (cont'd)

- MLR.6 implies both MLR.4 and MLR.5 (why?)
- a succinct way to put the population assumptions (all but MLR.2) is:

$$y \mid x \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$$



Sampling Distribution of the OLS Estimator (cont'd)

- with MLR.6, we can derive *exact* (as opposed to *asymptotical*) sampling distribution of OLS:

Theorem: Sampling distributions under normality.

Under the assumptions MLR.1 through MLR.6, conditional on the sample values of the explanatory variables,

$$\hat{\beta}_j \sim \text{Normal}(\beta_j, \text{var}\,\hat{\beta}_j)$$

which implies that $\left(\hat{\beta}_j - \beta_j\right) / \text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1)$.

Moreover, it holds $\left(\hat{\beta}_j - \beta_j\right) / \text{se}(\hat{\beta}_j) \sim t_{n-k-1}$ (Student's *t* distribution).

- *note*: the last formula is especially important, as the standardized estimates can easily be computed, given a *hypothesized value* of β_j

Lecture 4:
Multiple Regression

Jan Zouhar

Introductory Econometrics