Study Guide for Midterm 2

ECON 441

1 Modeling steps (big picture)

You start out with a question and want to measure the effect of a variable (RHS variable) on another variable (LHS variable). Here are the steps of the every empirical analysis:

- *Economic modeling*
  For now it only consists of thinking about the factors that influence the LHS variable. In more advanced analyses, you would incorporate more assumptions as well (e.g. sign, boundedness).
- *Econometric modeling*
  Choosing an econometric model you know (simple and multivariate linear model, including quadratic terms or not, logarithmic variables) that describes how you think your data set came to be (the **data generating process**)
  You also need to think about what is the parameter of interest (for us it is usually the slope parameters) for your question
- *Estimation*
  Choose an estimator that you know (you know only one: OLS) that can estimate the population parameter of interest with good properties (consistent, maybe unbiased, asymptotically normal, has lower variance)
- Implementation and interpretation: We run the estimator in Stata using the sample data and interpret our results (we pretend that the estimates are the true population parameters)
- *Inference*
  You test your empirical hypotheses (e.g. significance), and write your paper

2 Multivariate linear regression: Standard model

2.1 Basic definitions/vocabulary in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

Random variables:

Y	$\mathbf{X}$	U
Dependent variable	Independent variable	Unobservable
Left-hand side variable (LHS)	Right-hand side variable (RHS)	Error term
Explained variable	Explanatory variable	
Outcome variable	Covariates	
Predicted	Predictor	
Regressand	Regressor	
Response variable	Variable of interest + control variables	

In our models, we observe the realizations of the Y and the $\mathbf{X}$ variables.[1]

2.1.1 The population level

The **population** is the largest set of subjects that are important for us to answer our empirical question. The whole population is inherently unobservable most of the time.

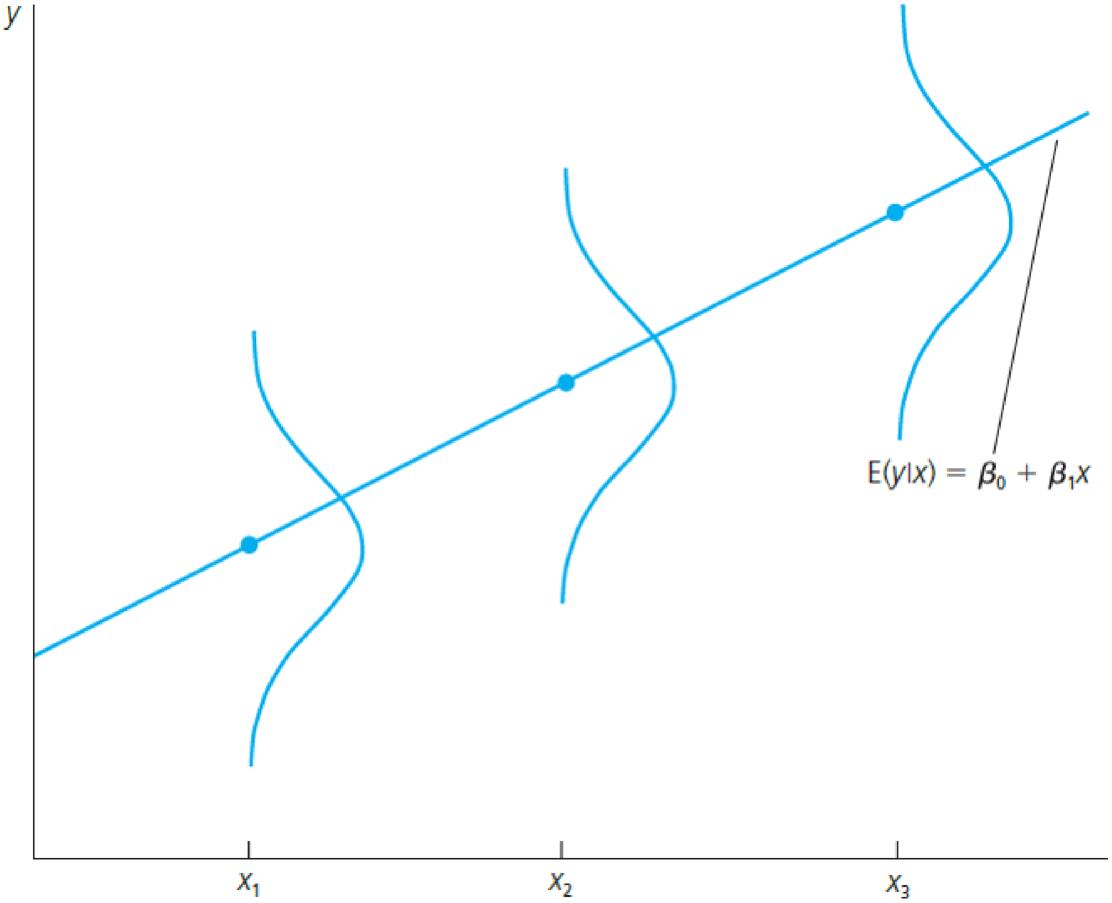
Population regression function (=conditional expectation):

$$E[Y|\mathbf{X}] = \beta_0 + \beta_2 X_2 + \beta_3 X_3$$

where $\mathbf{X}$ is a vector of (X_1, X_2, X_3). β_0 is the **intercept**, β_1 is the **slope parameter** corresponding to the variable X_1. Together with the variance of the unobservables (usually denoted by σ_U^2), these are **population parameters**, meaning they 'live' in the population, we will never observe/calculate them exactly, we may only estimate them.

Under the linear model assumptions (see below) $Y - E[Y|\mathbf{X}] = U$, the picture that you have in mind is

[1] Everything here is understood to be indexed by i, an occurrence in the population or an observation in the sample. I omit the subscripts in order to make the formulas more legible.



2.1.2 The causal effect in the linear model

The unitary effect of X on Y is the change in Y induced by increasing the X by 1 unit, while keeping every other (observable or unobservable) factors constant (=the same). This is also called a **causal effect** or **ceteris paribus effect**. You calculate this by writing up the model equation for the two X values, and take the difference:

$$Y_1 = \beta_0 + \beta_1(X_1 + 1) + \beta_2 X_2 + U \qquad (1)$$

$$Y_0 = \beta_0 + \beta_1(X_1) + \beta_2 X_2 + U \qquad (2)$$

$$\Delta Y = Y_1 - Y_0 = \beta_1 \qquad (3)$$

We often denote change in Y by ΔY and change in X by ΔX. The equation above says that the unitary effect of X on Y in the linear model is the slope parameter β_1.

The interpretation of the intercept parameter: If every RHS variable takes the value 0, the expected value of the LHS variable is β_0. Note that this interpretation does not make sense sometimes (0 year-old employee), so then you have to say 'no sensible interpretation exists'.

Usually our parameter of interest is the slope parameter, the intercept is just a bonus.

2.1.3 The sample level (estimation)

The **sample** is an observable subset of the population. Besides the assumptions that make the population regression function work (see below), you need to specify how you sample from the population to have a complete econometric model of the data generating process (origin story of your data set).

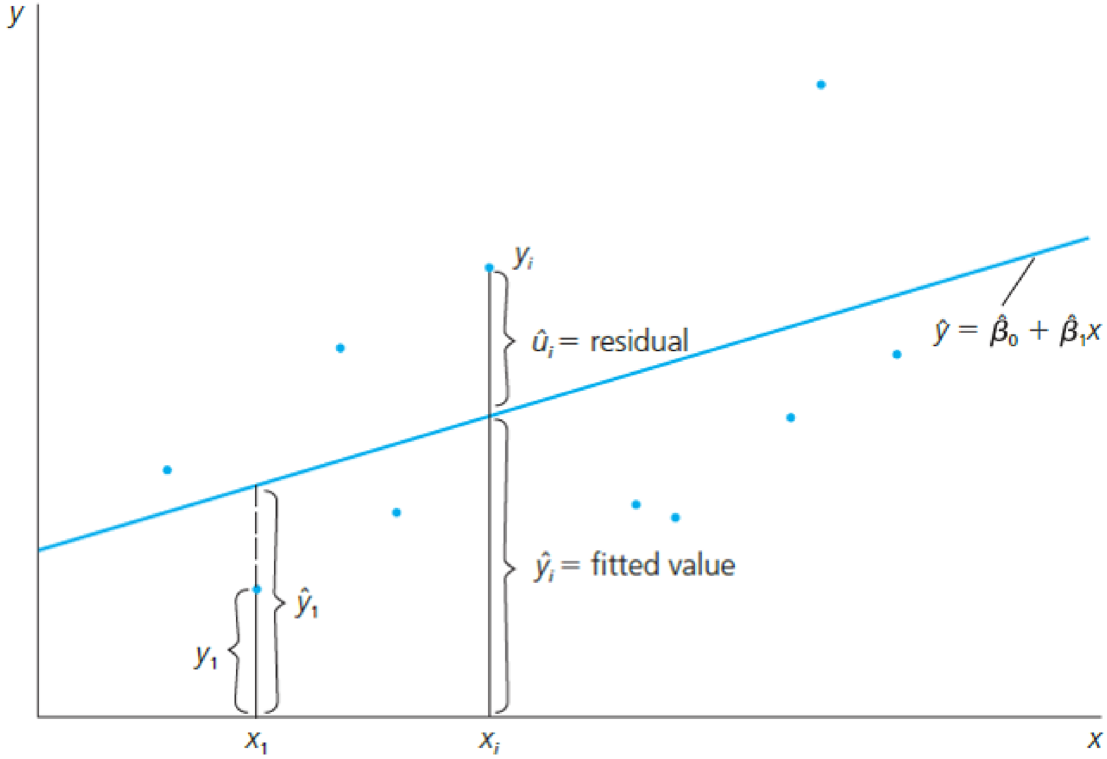
In this class, we are fitting a line on the scatter plot of the observables in our *sample* $(y, \mathbf{x})$, and we are getting the **estimates** of the population parameters, denoted by $\hat{\beta}_0, \hat{\beta}_1, \ldots$. Also, we define the **fitted value** as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3,$$

and we call the discrepancy between the fitted value (can be calculated) and the LHS variable is the **residual** $(\hat{u})$

$$\hat{u} = y - \hat{y}$$

The sample regression line is not exactly the population regression line that we wanted, but we are hoping it is close enough. The picture in mind should be


y
y_i
$\hat{u}_i$ = residual
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y}_i$ = fitted value
$\hat{y}_1$
y_1
x_1
x_i
x

The particular way we are fitting the line, which produces the estimates of the parameters of interest is called the Ordinary Least Squares **(OLS) estimator**. There could be many other ways to fit a line. This is the simplest one and produces very well-behaved estimates under our assumptions.

Expected outcomes:

1. Understand the difference between population and sample
2. Use the vocabulary correctly in interpretation questions

2.2 Model assumptions and failures

The multivariate regression model consists of 4 assumptions for this class. The first 2 establish a meaningful population regression function, the second 2 is about the sample you have, and make it possible to estimate the meaningful population parameters.

1. Linearity
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + U$$
(so we have k variables on the RHS)
2. Exogeneity
$$E[U|X_1, X_2, X_3, ..., X_k] = 0$$
3. Random sampling: the sample is randomly selected from the population
4. No perfect multicollinearity (=Sufficient variation): the RHS variables cannot perfectly predict each other in any combination when put in a linear model.

The following table summarizes the role of these assumptions, and what happens when they fail. Cells that are marked with asterisks are not pat of the midterm material or sometimes the whole course.

	Linearity	Exogeneity	Random sampling	No perfect multicollinearity
Symptom of failure	Scatter plot of (y_i, x_i) is 'curving'	No symptoms	No symptoms	Stata doesn't run 'regress'
Intuition when fails	Relationship is not additive, or the effect of X on Y is not the same for every value of X, or The LHS variable is bounded, only takes certain values	There is a third factor (in U) influencing Y and an X at the same time	If you observe someone from the population depends on their outcome*	Included a variable on the RHS that doesn't provide new information compared to the RHS variables already in the regression
Consequences of failure	We are estimating something that does not have a clear meaning	OLS estimates are biased and inconsistent - you do not estimate the causal effect, non-causal interpretation only.	OLS estimates are biased and inconsistent*	You cannot calculate the OLS estimates (Stata does not run 'regress')
Fixes	Have log-ed variables on the RHS and LHS Include quadratic terms on the RHS, later: limited dependent variable models	Include the omitted third factor (if observed) (later in the course: Instrumental variables)	Instrumental variables, model missingness*	Take out the variable that is predicted (Stata automatically does it for you)

More vocabulary and interpretation decisions:

- *exogeneity* is the opposite of *endogeneity* (when you find a *confounding* third factor)
- Multicollinearity: when **multi**ple RHS variables **co**-vary **lin**early, if they have a perfect linear relationship, we have the failure of the 'No perfect multicollinearity' assumption
- No exogeneity = endogeneity $\implies$ no ceteris paribus/causal interpretation, we only measure correlation instead of causation

In the standard linear model (level-level model without quadratic terms):

Exogeneity	A 1 unit increase in X_j causes a β_j unit change in Y
Endogeneity	Holding all other *OBSERVABLE* factors constant, a 1 unit increase of X_j corresponds to a β_j unit change in Y on average in the sample.

Note1: always 'plug in' into 'X' and 'Y' and into their 'unit'-s the words that are particular for your application
Note2: if there are no other observables than the variable of interest, don't say 'Holding other observable factors constant' at the beginning.
Note3: feel free to rearrange the sentences, but I want to see 'unit' and 'causes' vs 'on average' for the causal and the non-causal interpretations.

Also note:

Exogeneity	A ΔX unit increase in X_j causes a $\beta_j \cdot \Delta X$ unit change in Y
Endogeneity	Holding all other *OBSERVABLE* factors constant, a ΔX unit increase of X_j corresponds to a $\beta_j \cdot \Delta X$ unit change in Y on average in the sample.

So in the standard linear model (without quadratic terms) you do not need to know what was the value of X, before the change of ΔX occurred to calculate the effect. This is because of the homogenous effect assumption we implicitly impose with the linear model (see table).

Expected outcomes:

1. Know (understand) the table to be able to decide if you have causal interpretation or not
2. to be able to decide if the assumptions of the model are plausible or not
3. to be able to say what happens with your results if one or more assumptions fail
4. Be able to write down and recognize the standard linear model, and interpret the coefficients

2.3 Goodness-of-fit (R^2)

Given an estimated model, the (quadratic) variation in the LHS variable (=SST)

$$SST = \sum_{i}^{N} (y_i - \bar{y})^2$$

can be decomposed into explained part of the variation (SSE)

$$SSE = \sum_{i}^{N} (\hat{y}_i - \bar{y})^2$$

and the residual part of the variation (not explained by our model)

$$SSR = \sum_{i}^{N} (y_i - \hat{y}_i)^2$$

(here $\bar{y}$ is the sample mean of the y_i-s.)

The ratio of the explained (SSE) and total variation (SST) is the most famous goodness-of-fit measure,

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

Say, $R^2 = 0.32$. Interpretation: Our model explains 32% of the variation of the LHS variable.

If you have a prediction exercise, the higher the better, but not necessarily if you would like to establish causal relationships. Always have to report it with a regression. You can increase it to 100% if you include a lot of variables, but of course your standard errors will also blow up due to multicollinearity.

Expected outcomes:

1. Be able to calculate the R^2 using the SSR and SSE from the Stata output or if given
2. Be able to write down the 1-sentence interpretation

3 Including some non-linearities

Remember:

- Multiplicative relationship $\implies$ take logs

- Heterogenous effect[2] (effect of X on Y gradually increasing/decreasing with the values of X) $\implies$ Include quadratic terms

It's rare we do the two things together. It is ok to take log/include squared terms for only a select number of RHS variables.

3.1 Curvature in an otherwise additive model (quadratic terms)

Leading example 1: house prices vs house age.
First, when the age of the house is low, if we increase it, the price will decrease. After 100 years, if you further increase the age of the house, the price increases. *The effect of X is going from negative to positive/growing as you increase X.*

Leading example 2: wage equation with age. First, when you are 25, increasing your wage tend to have a positive effect on your wage, however, holding other factors constant (e.g. doing the same kind of job) increasing your age by 1 year tend to have a negative effect on your wage. *The effect of X is going from positive to negative/decreasing, as you increase X.*

Case	Scatter plot shape	Math name	Sign of coefficient on X^2
Example 1	U-shaped	convex	positive
Example 2	Inverted-U shape	concave	negative

When I ask you to interpret the coefficient on the quadratic term, I ask you to tell me if you have a convex or concave relationship, and what it means in the context of the application (see the example explanations above).

Moreover, you will need to comment on the unitary effect. For the purposes of this subsection, I denote the estimated coefficient on the quadratic term by $\hat{\beta}_2$, and the coefficient on the level term by $\hat{\beta}_1$

The unitary effect of X_1 (to increase the level of the first RHS variable from x_1 value to $x_1 + 1$ value) is

$$e(x_1) = \hat{\beta}_1 + \hat{\beta}_2(2x_1 + 1).$$

For interpretation then, if you argued for exogeneity plausibly, you would say

" A 1 unit increase in X_1 from the x_1 level causes a $e(x_1)$ unit change in Y."

Note that we could also construct the corresponding non-causal interpretation of the same kind (you modify the sentence corresponding to the endogenous case in the very same way).

[2] Heterogenous (different/changing) is the opposite of homogenous (the same/constant).

You may want to know[3] that most people will talk about the partial effect, defined as

$$\frac{\partial Y}{\partial X_1} = \hat{\beta}_1 + 2\hat{\beta}_2 x_1.$$

Expected outcomes:

- Argue for concave/convex cases from economic intuition, and be able to connect the intuition with the sign of the coefficient on the quadratic term
- Be able to calculate the unitary effect and write down the interpretation
- Recognize the interpretation of the unitary effect from a question

3.2 Multiplicative relationships: logarithmic models

When someone suggests/you think that

- changing X by 1 unit will not increase Y by some units, but rather by some constant percentage (log-level).
- you need to change X by some percentage (not by 1 unit) to get some constant increase in Y (level-log)
- the X-elasticity of Y is constant, or if you change X by some percentage, it will induce a constant percentage change in Y (log-log)

Remember the "Table of Interpretations":

Model	LHS	Var. of interest	Causal interpetation
level-level	Y	X	A ΔX unit change in X causes a $\beta \cdot \Delta X$ unit change in Y.
log-level	$\log(Y)$	X	A ΔX unit change in X causes a $100 \cdot \beta \cdot \Delta X$ percent change in Y.
level - log	Y	$\log(X)$	A ΔX percent change in X causes a $\frac{\beta}{100} \cdot \Delta X$ unit change in Y.
log-log	$\log(Y)$	$\log(X)$	A ΔX percent change in X causes a $\beta \cdot \Delta X$ percent change in Y.

Note again, you can easily modify the sentence corresponding to the endogeneity case from the interpretation table of the second section (standard model) to get a non-causal interpretation of these estimated coefficients.

[3] I am not going to ask you the interpretation of the partial effect

Expected outcomes:

1. Be able to recognize and write up log-level, level-log and log-log models in problems
2. Interpret coefficients in these models

4 Model selection

How to decide how many RHS variables are enough?
If you put too few: high omitted variable bias (severe endogeneity).
If you put too many: high variance (severe, eventually perfect multicollinearity).

We defined the adjusted R^2, which is an adjusted version of our measure of fit: we penalize the high fit if it is achieved by a high number of RHS variables. So R^2 always increases if you add a new RHS variable, but if the new addition does not give enough new information about the LHS variable, the adjusted R^2 will decrease.

How to use it?
If model 1 has higher adjusted R^2 than model 2, we think model 1 is better.

5 Properties of OLS and the homoskedasticity assumption

During the sampling process we select observations from the population to be part of our particular sample. Since we do not calculate the coefficient estimates using the whole population, only a particular sample, OLS will yield a **sample estimate** $\hat{\beta}$ that is different from the **population parameter** β (the number we would get if we would use OLS on the population data).

Moreover, just as we created our particular sample, we could have selected another group of n observations from the population. The OLS **estimator** (a procedure) would have yielded yet another number as the estimate for that sample. There are many possible n-samples (samples consisting of n observations), and each could potentially come with its unique $\hat{\beta}$ estimate value.

The **distribution** of the OLS estimator is the histogram of these $\hat{\beta}$ values (based on the relative frequencies) when we collect the estimates from all possible n-samples. Under random sampling, this is the same as thinking about the probability distribution of the $\hat{\beta}$ that will be calculated from tomorrow's (not yet known, uncertain) random sample. The properties of the estimator will describe this probability distribution, which is needed to do inference on our estimates.

5.1 Unbiasedness

What does it mean?
The expected value (mean) of the distribution of $\hat{\beta}$ is the true population parameter β:

$$E[\hat{\beta}] = \beta. \tag{4}$$

We say that in this case the bias of the estimator (Bs) is zero:

$$Bs[\hat{\beta}] = E[\hat{\beta}] - \beta = 0. \tag{5}$$

If this is true, we have that our estimator's distribution is centered around the true value (even in low sample sizes). *What are the assumptions needed to guarantee it?* We need the first 3 assumptions: linearity, exogeneity, no perfect multicollinearity

5.2 Consistency

What does it mean?
As we increase the sample size, the probability that we draw a random sample that gives an estimate very close to the true value is more-and-more likely (approaching 1). Consistency is a minimum requirement for an estimator; note that it is a large sample/**asymptotic property**.

What are the assumptions needed to guarantee it?
We need the 4 assumptions of the linear model: linearity, exogeneity, no perfect multicollinearity and random sampling.

Misc (optional)
Definition: An estimator $\hat{\beta}$ consistently estimates the population parameter β, if $P[|\hat{\beta} - \beta| > \epsilon] \to 0$ as the sample size, $n \to \infty$.

If for the variance of the estimator ($V[\hat{\beta}]$) and the bias ($Bs[\hat{\beta}] = E[\hat{\beta}] - \beta$) we have that

$$Bs[\hat{\beta}] \to 0 \tag{6}$$

$$V[\hat{\beta}] \to 0 \tag{7}$$

as the sample size increases to infinity, **then** $\hat{\beta}$ is consistently estimating β.

5.3 The variance of OLS (informal)

What does it mean?
It measures the precision of our estimator; it is the variance of the distribution of the possible $\hat{\beta}$-s ($\sim$ how 'fat' the distribution is.) This is an important statistic, as if you take a square-root (to get the standard deviation) and multiply by $\sqrt{n}$ (the n is the sample size), then you get the standard error of the estimator.

What terms influence the variance?

Component	**Relationship with $V[\hat{\beta}_k]$ and $se(\hat{\beta}_k)$**
(Conditional) Variance of error term (σ_U^2)	Positive
Sample size (n)	Negative
Variance of X_k	Negative
Multicollinearity for X_k	Positive

Positive: if the component increases (gets more 'severe') then the standard error for the coefficient on X_k increases
Negative: if the component increases, the standard error for the kth coefficient decreases

Multicollinearity measure for X_k: the R^2 of the regression when you regress X_k on the rest of the RHS variables. If this R^2 is 1, we have perfect multicollinearity – the Se() formula gives infinite variance.

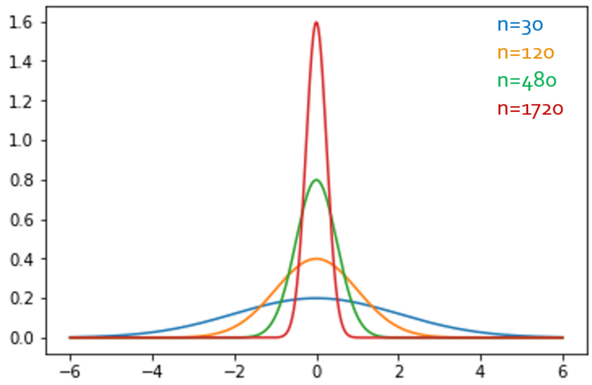
What assumptions do we need?
We used the homoskedasticity assumption to derive the variance of the $\hat{\beta}$ in the single regression case. This 5th assumption is only needed for this calculation; it does not affect consistency or unbiasedness. Modern regression analysis calculates standard errors without this 5th assumption. Note however that every calculation uses Assumption 1-4 of the linear model and assumes large sample.

5.4 Asymptotic normality

What does it mean?
The distribution ($\sim$ histogram) of the $\hat{\beta}$ looks like the bell-curve if we have large sample. (Really as the sample size increases to infinity.)

What are the assumptions needed to guarantee it?
We need the 4 assumptions of the linear model: linearity, exogeneity, no perfect multicollinearity, random sampling. This is an asymptotic (=as n grows to infinity) property, so we need a large sample to make use of it (for example do inference).

5.5 The picture in mind



Here the true value of the parameter is zero. The picture depicts the pdfs (∼ histograms) of the parameter estimates as the sample size grows.

You should be able to see

1. This estimator is unbiased, as for every sample size the pdfs are centered around the true value (zero)
2. The variance is diminishing as n grows (the pdf-s are getting leaner until you can only see a needle)
3. Looks like the estimator is (asymptotically) normal, because the pdf-s resemble the Gaussian bell-curve (they are exactly Gaussians)

Please remember: consistency is ensured by points 1. and 2. (But you can see that consistency holds in the picture as well.)

5.6 Homoskedasticity vs. Heteroskedasticity

The homoskedasticity assumption can be thought of as the 5th assumption of the linear model. It is just like the other 4 assumptions, so it could be added to our big table in section 2.2. What homoskedasticity fails, we say we have heteroskedasticity.

What does it mean? It means that the conditional variance of the unobservables (U) is the same, regardless at what level we fix the RHS variables (X-s). Mathematically:

$$V[U|X_1, X_2, X_3, ..., X_k] = \sigma_U^2 \tag{8}$$

Why is it useful?
We can derive the Gauss-Markov theorem, and so OLS is optimal in some sense.
It is easy to calculate the standard errors of OLS, and some tests need this assumption in some statistical package implementations (e.g. F-test in Excel...no comment.)

Symptom of failure
On the scatter plot of the LHS and the RHS variable the cloud has increasing/decreasing/uneven width as we increase the level of the RHS variable.

Intuition when fails
Almost always - especially if you have money on the RHS.

Consequences of failure
OLS standard errors, hence testing results become invalid.

Fixes
Use the heteroskedasticity-robust or 'White' standard errors.

I mentioned in 2 minutes that there are tests developed for detecting heteroskedasticity: the White and Breusch-Pagan tests (optinal, but important, especially if you want to do time series later).

5.7 Gauss-Markov theorem

Under Assumptions 1-5 (add homoskedasticity) of the linear model, OLS is the Best Linear Unbiased Estimator (OLS is BLUE).
Best: it has the lowest standard errors – not just unbiased, it is also the most precise!
Linear: easy to handle and compute

6 Inference

We talked about confidence intervals and their meanings. You should be able to interpret confidence intervals.
We talked about the t and F-tests. You should be able to conduct these tests with Stata or by hand after I provide you with the necessary outputs. Make sure you can also interpret the results.

Remember: for our testing procedures to be valid we need that the OLS properties above hold: we need large samples and either homoskedasticity or heteroskedasticity-robust (White) standard errors.

For this material, see the separate testing handout!

7 Qualitative variables on the RHS

Qualitative variable: The value of the variable is not a number. You need to encode it into numbers.

Dummy variable: a variable that can take only 2 values, 0 and 1. When you have a qualitative variable with only 2 possible values, it is naturally encoded by a dummy.

First thing to do (or think about) when you want to include a qualitative variable is to count how many possible values it can take. I denote this number by K.

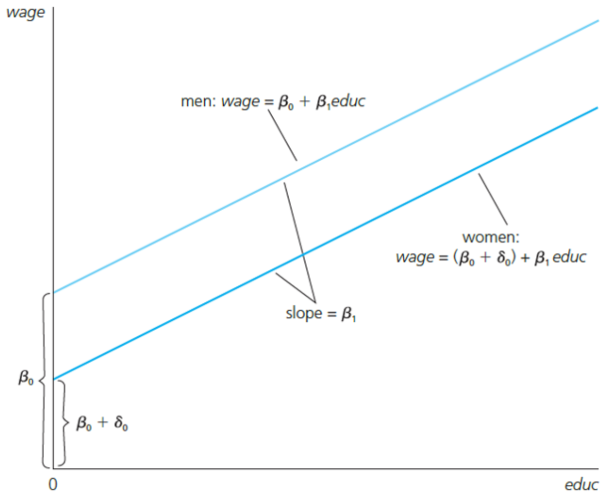
7.1 Modeling

If you have only 2 possible values in your data set for the variable: just encode the values into 0 and 1, and put the variable name into the RHS

If you have K>2 possible values for the qualitative variable, you need to create K-1 dummies corresponding to K-1 possible values. The Kth possible value is going to be the **reference category**, which will be omitted from the regression. When I ask you to write down the new model, your regression model now will include this $K-1$ dummies on the RHS.

7.2 Interpretation

The meaning of including these qualitative variables is that we allow for a different intercept value for the different categories. As shown in the picture, by including a gender dummy, we have two parallel regression lines: one for men and one for women. Note that the slopes (the effect of the other RHS variables) are the same for both groups. (See figure below.)


wage
men: $wage = \beta_0 + \beta_1 educ$
women:
$wage = (\beta_0 + \delta_0) + \beta_1 educ$
slope = β_1
β_0
$\beta_0 + \delta_0$
0
educ

7.2.1 Interpretation of a single dummy variable

The coefficient gives you the difference in the (expected value of the) LHS variable between two groups: when the dummy takes the value 1 and when it takes the value 0.

Let us call the category for which the dummy is zero Group0, and the category for which it is 1 Group1. Also, let us call the coefficient on the dummy variable δ. You need to be able to interpret in the following 4 'environments':

Given that δ is positive/negative...

	Level-level	**Log-level**
Endogeneity	Y is $\lvert\delta\rvert$ unit higher/lower for Group1 than for Group0 on average in our sample.	Y is $\lvert\delta\rvert \cdot 100$ percent higher/lower for Group1 compared to Group0 on average in our sample.
Exogeneity	Due to the differences in the X (qualitative variable), Y is expected to be $\lvert\delta\rvert$ unit higher/lower for Group1 than in Group0.	Due to the differences in the X (qualitative variable), Y is expected to be $\lvert\delta\rvert \cdot 100$ percent higher/lower for Group1 than in Group0.

Note that you cannot log a dummy variable, and should not log any qualitative variable.

7.2.2 Interpretation for $K > 2$

You have a coefficient δ_l that multiplies the dummy corresponding to the value (category or group) l. The δ_l gives the difference in the (expected value of the) LHS variable when we compare the group corresponding to the coefficient to the *reference category* (the omitted category).

You need to be able to interpret in the following 4 'environments'

Given that δ is positive/negative...

	Level-level	**Log-level**
Endogeneity	Y is $\lvert\delta_l\rvert$ unit higher/lower for Group$l$ than in the reference group on average in our sample.	Y is $\lvert\delta_l\rvert \cdot 100$ percent higher/lower for Group1 compared to the reference group on average in our sample.
Exogeneity	Due to the differences in the X (qualitative variable), Y is expected to be $\lvert\delta_l\rvert$ unit higher/lower for Group1 than for the reference group.	Due to the differences in the X (qualitative variable), Y is expected to be $\lvert\delta_l\rvert \cdot 100$ percent higher/lower for Group1 than in the reference group.

You should try to make your sentences as 'English' as possible. For example, you should say what is the reference group in the interpretation, instead of putting 'reference group'.

In addition, the exercise may ask you to compare the expected Y of Group$l$ to that of Group$k$, where neither of these categories is the reference group. Then your interpretation sentence will be the same as above, except you get that the effect (the actual number) is $\delta_l - \delta_k$, not just δ_l.

7.3 Estimation and inference

We are not stepping out of the linear model's framework, so the estimation procedure remains the same: OLS.

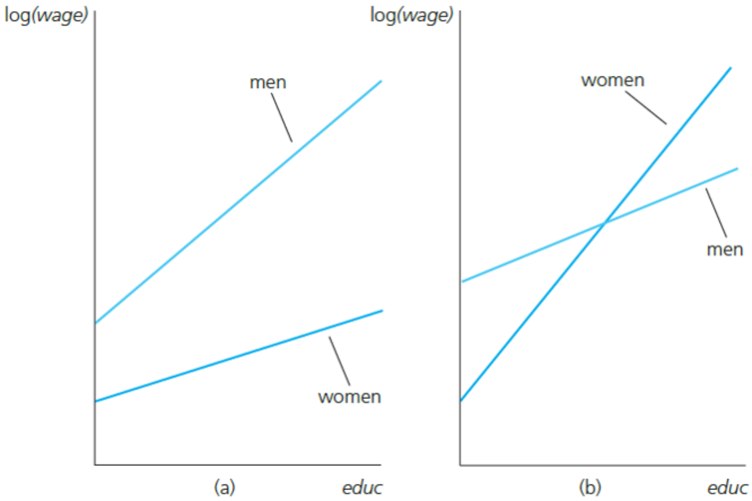
When you have $K = 2$ (a single dummy), you can use the usual t-test for testing significance.

When you have $K > 2$, we need to use the F-test for joint significance testing.

7.4 Interaction terms

Interaction terms are generated when we multiply two RHS variables (and put the product in the regression).

It is particularly interesting when you put the interaction term of a dummy variable and a quantitative variable on the RHS. The interpretation of this is that you allow the effect of the quantitative variable to differ between Group0 and Group1 as well (not just the intercept). This corresponds to the following picture in our running example:



You have separate regression lines for men and women, but now they do not have to be parallel either. Return to education can vary between the two groups.

While I do not require you to be able to write nice interpretation sentences for the coefficients on interaction terms, you need to know the basic interpretation above and how to run a regression with them

Also, make sure to understand that we need to test the significance of these terms if the exercise is asking about testing if the effect of a quantitative variable differs across categories of a qualitative variable.

8 Stata

You will need to be able to read Stata output, including R^2, coefficients, SSR, SSE, SST and adjusted R^2.

You will need to remember the following commands and their usage:

Importing data sets (command: use)

```
use ”path/filename.dta”
```

Generating a new variable (command: generate)

```
generate newvar_name= log(oldvar) generate newvar_name= oldvar2̂
```

Regressing (command: regress) using heteroskedasticity-robust (White) standard errors

```
regress LHS_variable RHS_var1 RHS_var2 RHS_var3, robust
```

Getting basic descriptives of variables (command: summarize)

```
summarize varname1 varname2
```

Conducting the F or t-tests (command: test)

Including qualitative variable on the RHS with possibly string entries (Stata will generate $K-1$ dummy variables and include them on the RHS; prefix: xi)

```
xi: regress lhs_var var1 var2 i.qualitative_var var3 var4, robust
```

(the usages are only examples here, please consult the log-files to see how to use each command for some tasks we had in the lab sessions)