Basic econometrics

Tutorial 3

Dipl.Kfm. Johannes Metzler

Introduction

- Some of you were asking about material to revise/prepare econometrics fundamentals. First of all, be aware that I will not be too technical, only as technical as necessary to understand the methodologies used in the papers.
- We will mostly deal with cross-section and panel data papers, no time series econometrics.
- basic (German) textbook "**Ökonometrie**. Eine **Einführung.** by **Auer**, Ludwig von. 3. überarb. Aufl., 2005, ISBN: 3-540-24978-8"
- "**Guide to Econometrics** by Peter **Kennedy**, 5th Edition (only the 5th ed. has a section on panel data!), ISBN: 026261183X„: focus on intuition
- Apart from that, for quicker reference, econometrics lecture notes are useful, e.g. the panel data part from **Alan Duncan** (Nottingham) - http://www.nottingham.ac.uk/%7Elezad/courses/cspd6.pdf
- I personally find Prof. Winter's lecture notes very useful, containing all the necessary information - for anyone who has not attended his courses yet. User and Password:

Introduction

"Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories and evaluating and implementing government and business policy" (Wooldridge)

$$Y = f(x_1, x_2, \ldots, x_k, \varepsilon)$$

Y = dependent variable

$x_1, \ldots, x_k$ = (some) determinants of Y, explanatory variables

ε = random error term

- More generally, we can say that regression analysis is concerned with studying the distribution of Y given some X's
- The error term capture the effects of all the determinants of y that are not in x1…xk. That also means that the relation is not unique (we do not have a unique value of Y given certain values of X's) but we have a certain distribution of values for Y → the relationship is described in probabilistic terms

Introduction

Different data that we deal with

- Cross-sectional data: observe individual units (people, households, countries) at one certain point in time
  - E.g. The price of a car depending on its characteristics
- Time-series data: observe one unit over time
  - E.g. GDP growth of a country depending on its inflation, no. of employed people, technological advances, exports, etc.
- Panel data: observe individual units over time
  - E.g. household surveys: track people's consumption patterns in multiple time periods

Simple regression model

Let's start considering only one explanatory variable

Imagine a linear relation between Y and X:

$$Y = \alpha + \beta X + \varepsilon$$

If the assumption that $E(\varepsilon \mid X)=0$ then:

$$E(Y|X) = \alpha + \beta X$$

That is, a unit increase in X changes the expected value of Y by the amount β

For any given value of X the distribution of Y is centred about E(Y|X).

- What does it mean that $E(\varepsilon \mid X) = 0$? The average value of the error term doesn't depend on x…
- if y = exam score, x = class attendance, the things that can be in the error term do not depend on x. What can be included in the error term? Let's say ability. So if we run this regression we are implicitly assuming that the average level of ability doesn't depend on the # of classes attended.

Regression analysis

What are we searching for?

- We want to find a transformation of the X's , $f(X)$, that gives us the best approximation of Y

Which is the best approximation?

The one that minimizes the expected error of prediction

$$\text{Min } \mathrm{E}[l(\mathrm{Y} - f(\mathrm{X}))]$$

Which $l(.)$ and which $f(.)$?

Quadratic loss function: $l(.) = (\mathrm{Y} - f(\mathrm{X}))^2$

→ OLS: Ordinary Least Squares

Linear transformation of the X's: $\mathrm{Y} = \alpha + \beta\, \mathrm{X}$

Simple regression model: results

The problem:

$$\min_{\alpha,\beta} \mathrm{E}[\mathrm{Y} - (\alpha + \beta\, \mathrm{X})]$$

Solving the minimization problem we obtain the following condition for the estimated parameters:

$$\hat{\beta} = \frac{E[X(Y - E(Y))]}{E[X(X - E(X))]} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\alpha} = E(Y) - \hat{\beta} * E(X)$$

With the estimated coefficients we can obtain the fitted values for Y when X = X_i

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} * X_i$$

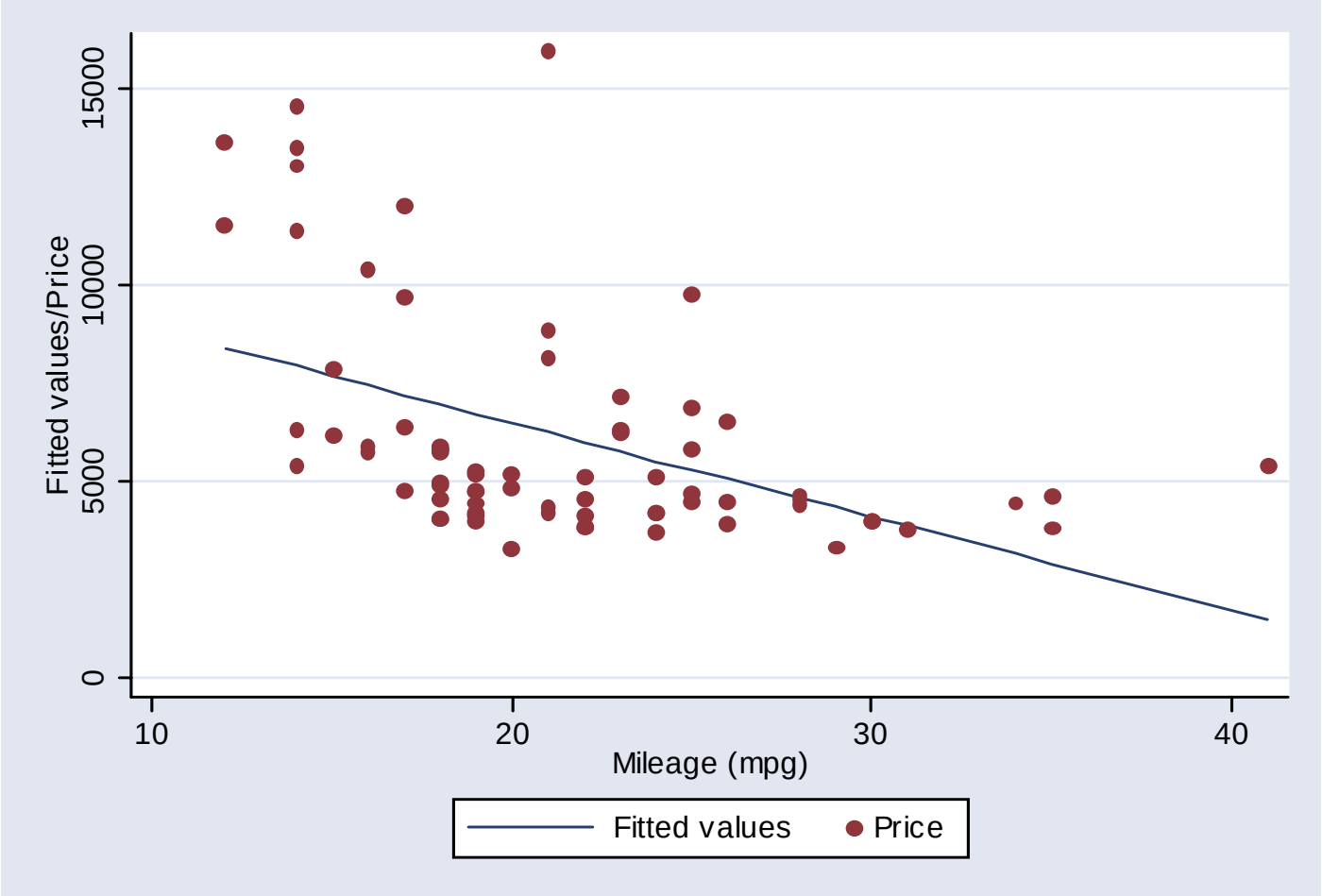
The fitted value for Y ($\hat{Y}_i$) is the value we predict for Y when X = X_i

- remember that $\hat{Y}_i$ is the SAMPLE regression function and it is the estimated version of the POPULATION regression function that we suppose to be existing but unknown → different samples will generate different coefficients

Simple regression model

A simple cross-sectional regression:

- Explain the price of a car with its mileage (miles per gallon)

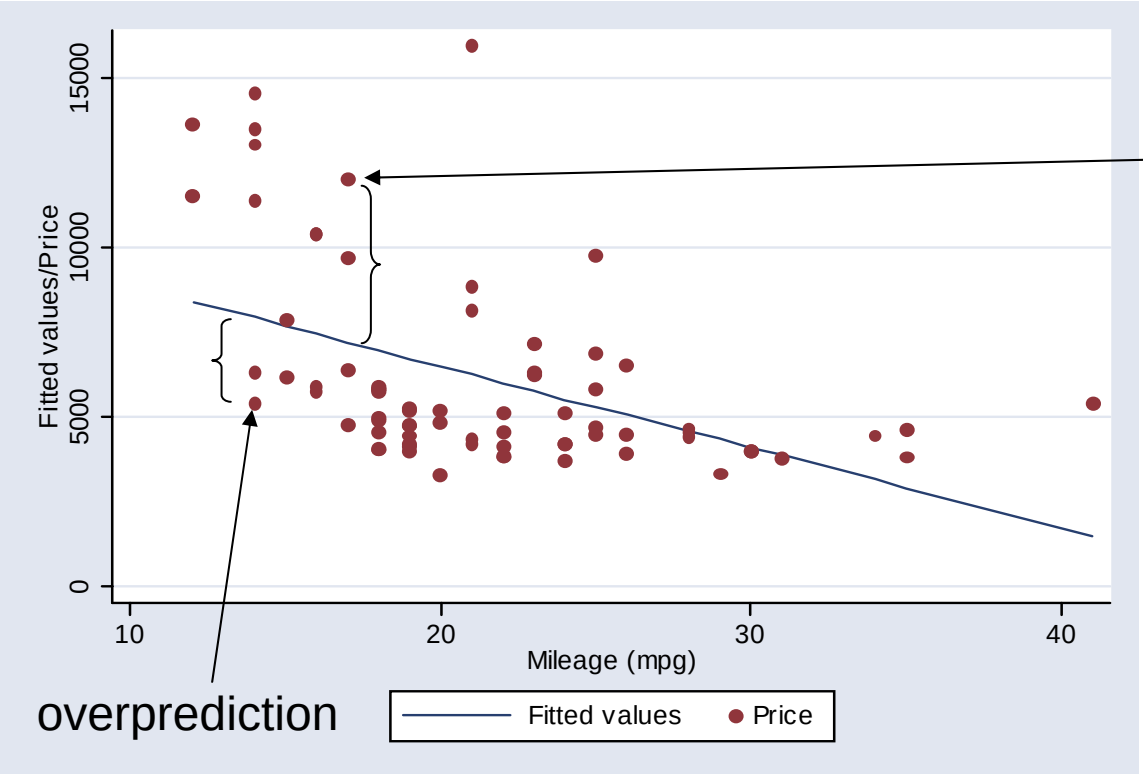


Simple regression model

By construction, each fitted value is on the regression line

The OLS residual ($\hat{U}_i$) associated with each observation is the difference between the actual dependent value Y_i and its fitted value $\hat{Y}_i$.

If $\hat{U}_i$ is positive (negative), the line underpredicts (overpredicts) Y_i



Basic diagnostics

We want to measure of how well the explanatory variable X explains the dependent variable Y (goodness-of-fit)

SST (total sum of squares) = measure of total sample variation in the Y_i

SSE (explained sum of squares) = sample variation in the $\hat{Y}_i$

SSR (residual sum of squares) = sample variation in the $\hat{U}_i$

SST = SSE + SSR

1 = SSE/SST + SSR/SST

R^2 = SSE/SST = 1 – SSR/SST

- R^2 it's interpreted as the <u>fraction of the sample variation in Y that is explained by X</u>
- R^2 is bounded between 0 and 1: a value close to 0 indicate a poor fit of the OLS line to the data. When all the real data are lying on the OLS line, then R^2 = 1 (perfect fit)

Basic diagnostics

The estimated OLS line depends on the sample of observation we got.

- It could be, then, that the "real" β is zero, but because of the sample, we estimate a coefficient different from zero.
- The statistic $\frac{\hat{\beta} - \beta}{s_{\beta}} \sim t_{n-1}$ is distributed as a t distribution with $n-1$ degrees of freedom
- We can then test the (null) hypothesis that $\beta = 0$

1. Look at the value of the t statistic
2. Look at the conf. interval
3. Look at the p value (prob. of falsely rejecting the H0)

```
Number of obs =      74
R-squared     =  0.2196
Adj R-squared =  0.2087
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -238.8943   53.07669    -4.50   0.000    -344.7008   -133.0879
       _cons |   11253.06   1170.813     9.61   0.000     8919.088    13587.03
------------------------------------------------------------------------------
```

Simple regression model: including nonlinearities

Linear relationships btw variables are not enough

We require linearity in the coefficients, not necessarily in the covariates

- Polynomial model
- Logarithmic model
- Interaction terms

Simple regression model: including nonlinearities

Polynomial model:

The regressors are power of the same explanatory variable

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon$$

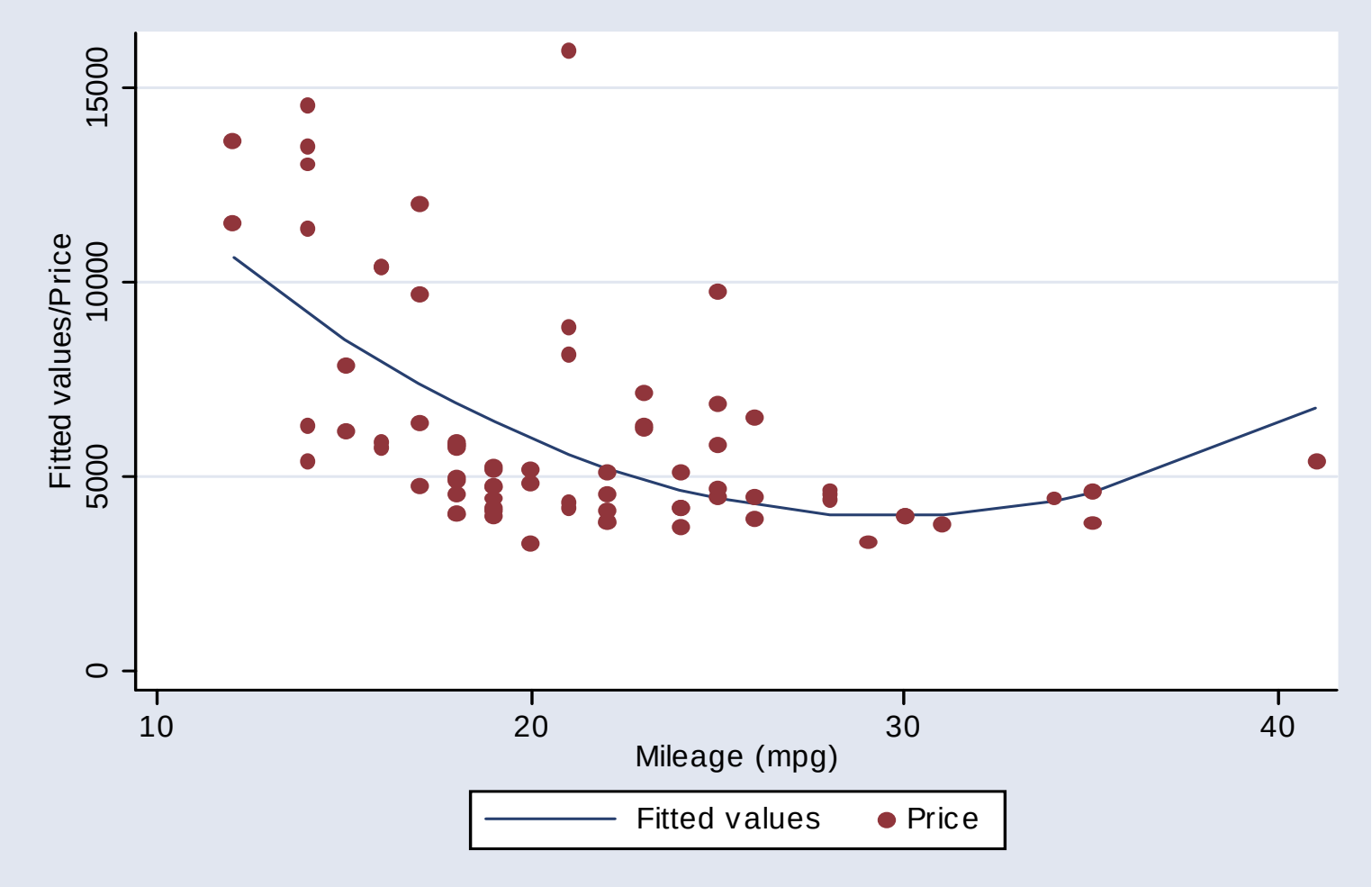
Increasing the power included in the regression gives more flexibility

Adding too many regressors can reduce the precision of the estimated coefficients

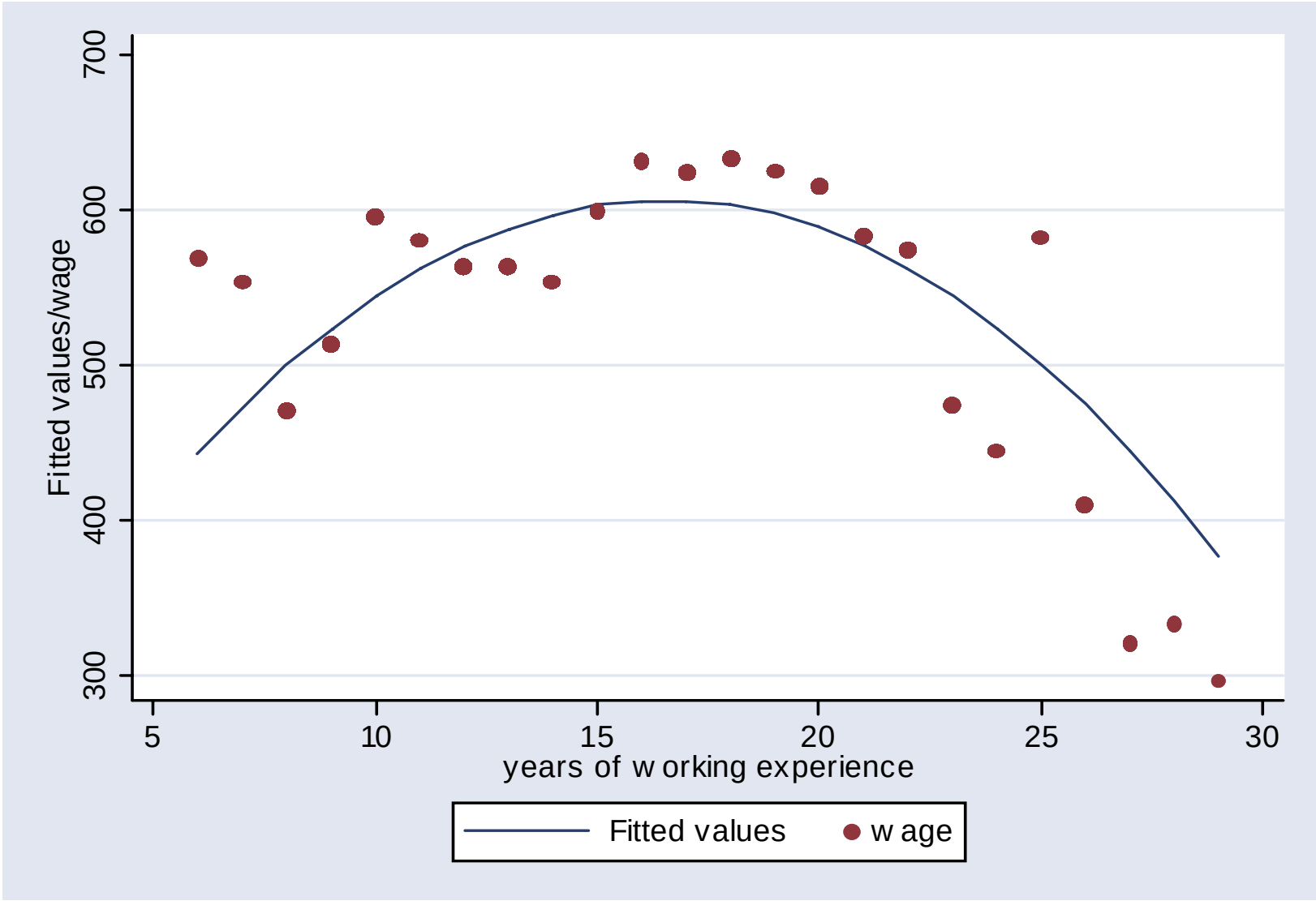
The coefficients on β_k gives information on the concavity or convexity of the line

Simple regression model

$$\text{Price} = \alpha + \beta_1 mpg + \beta_2 mpg^2 + \varepsilon$$



Simple regression model



Simple regression model:
including nonlinearities

Logarithmic model:

Two cases are particularly interesting:

log – level $\log(y) = \alpha + \beta x + \varepsilon$

log – log $\log(y) = \alpha + \beta \log(x) + \varepsilon$

Interpretation of the coefficients:

log – level → a unit change in X is associated with (100*β) % change in Y

log – log → a 1% change in X is associated with a β% change in Y; β is the elasticity of Y w.r.t. X

Simple regression model:
including nonlinearities

	Dependent variable	Explanatory variable	Interpretation of the coefficient
level - level	y	x	$\Delta y = \beta \Delta x$
level – log	y	log(x)	$\Delta y = 0.01\beta \, \%\Delta x$
log – level	log(y)	x	$\%\Delta y = 100 \, \beta \Delta x$
log – log	log(y)	log(x)	$\%\Delta y = \beta \, \%\Delta x$

Simple regression model: including nonlinearities

Interaction model and dummies:

- Example we want to estimate the effects of schooling on earnings

$$\log(wage) = \alpha + \beta \cdot educ + \varepsilon$$

- We can imagine, though, that women and men have different entry wages. How do we deal with that?

$$\log(wage) = \alpha + \beta_1 \cdot educ + \beta_2 D + \varepsilon \qquad D = \begin{cases} 0 \text{ if man} \\ 1 \text{ if woman} \end{cases}$$

- α is the entry wage of males with no year of schooling
- $\alpha + \beta_2$ = entry wage for females with no schooling
- β_1 = % change in wages for both males and females for each year of schooling

Simple regression model: including nonlinearities

Interaction model and dummies:

$$\log(wage) = \alpha + \beta_1 \cdot educ + \beta_2 D + \varepsilon$$

- In this specification, the effect of an additional year of schooling is the same for women and men

- We can imagine, however, that the entry wage is the same, but that one year of schooling has a different effect on wages for women and men

$$\log(wage) = \alpha + \beta_1 \cdot educ + \beta_2 D \cdot educ + \varepsilon$$

- Finally, we can imagine that both the intercept and the slope are different for women and men

$$\log(wage) = \alpha + \beta_0 D + \beta_1 \cdot educ + \beta_2 D \cdot educ + \varepsilon$$

  - b1 is the effect of an additional year of schooling for males
  - b1 + b2 is the effect of and additional year of schooling for females
  - b2 measures the difference in the effect of an additional year of schooling on wages for females vs. males

Multiple regression analysis

The previous discussion can be extended to the case with more than one explanatory variable

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Of course we will have $k+1$ parameters to estimate. The OLS regression line is then given by:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

The betas have the **partial effect** interpretation

- partial effect (or ceteris paribus) means that the coefficient on x_1 measure the change in y due to a one-unit increase in x_1, holding all the other independent variables fixed

Multiple regression analysis : basic diagnostics

R^2 is computed in the same way (R^2 = SSE/SST = 1 – SSR/SST), but:

- cannot be used to compare models with different dependent variables
- never decreases when an additional regressor is added
- Adjusted $R^2 = 1 - \frac{SSE / N - k - 1}{SST / N - 1}$

N = no. of units/observations
k = no. of explanatory variables
N – k – 1 = degrees of freedom

We actually may be interested in testing if *all* the coefficient are *jointly* equal to zero → *F* test

the statistic $\frac{R^2}{1 - R^2} \frac{N - k}{k - 1}$ is distributed as a *F* distribution

Multiple regression analysis

```
Number of obs =      69
F(  5,    63) =   10.29
Prob > F      =  0.0000
R-squared     =  0.4497
Adj R-squared =  0.4060

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   -111.268    81.4019    -1.37   0.177    -273.9368    51.40068
      weight |   5.463167   1.246622     4.38   0.000     2.971991    7.954342
      length |  -119.9012   38.96983    -3.08   0.003    -197.7762   -42.02619
  gear_ratio |   908.5008   984.0873     0.92   0.359    -1058.041    2875.042
       rep78 |   868.6331   308.3117     2.82   0.006     252.5213    1484.745
       _cons |   8843.388   6680.521     1.32   0.190    -4506.568    22193.34
------------------------------------------------------------------------------
```

Unbiasedness & Consistency

- What do we want from an estimator?
- The **unbiasedness** property of the estimators means that, if we have many samples for the random variable and we calculate the estimated value corresponding to each sample, the average of these estimated values approaches the unknown parameter
  - we want the expected value of the estimator to be equal to the population characteristic.
- An estimator is said to be **consistent** if it converges in probability to the unknown parameter
  - Intuitively: if the estimated coefficient differs only by an arbitrarily small amount from the true value of the parameter in the population
  - Or: a consistent estimator is one that is bound to give an accurate estimate of the population characteristic if the sample is large enough, regardless of the actual observations in the sample.
- Note that consistency is not the same as unbiasedness. Consistency says that the bias and variance tend to zero, not that either ever attains zero.

Assumptions of OLS unbiasedness

- Population model is linear in parameters:

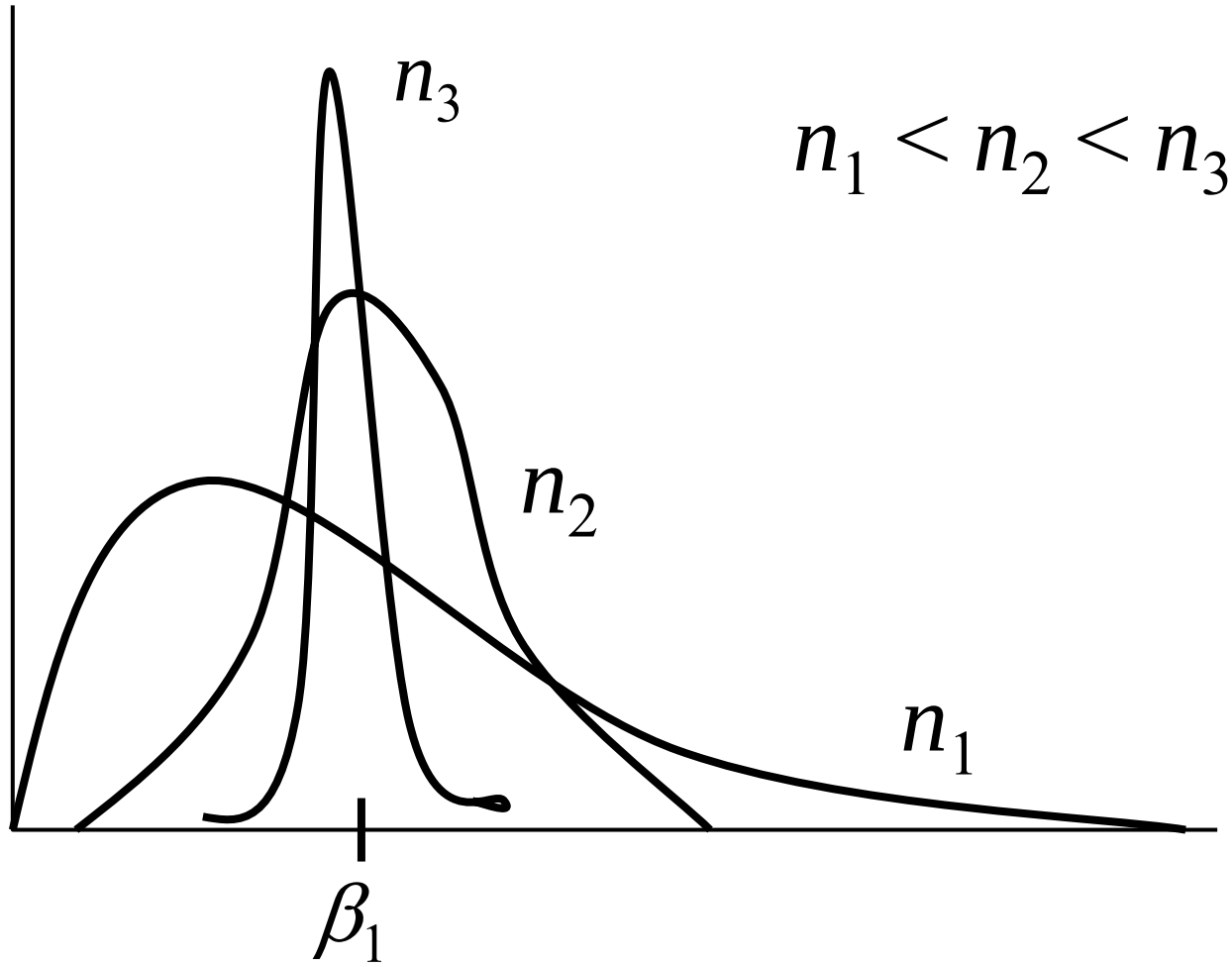
  $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + u$
- We can use a random sample of size n, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i=1, 2, \ldots, n\}$, from the population model, so that the sample model is

  $y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \ldots + b_kx_{ik} + u_i$
- **$E(u/x_1, x_2, \ldots x_k) = 0$, implying that all of the explanatory variables are exogenous (zero conditional mean)**
- None of the x's is constant, and there are no exact linear relationships among them

OLS Consistency

- Under the stated assumptions OLS is BLUE (the best linear unbiased estimator), but in other cases it won't always be possible to find unbiased estimators
- Thus, in most cases, the desired property is consistency, meaning as $n \to \infty$, the distribution of the estimator collapses to the parameter value

Sampling Distributions as $n \uparrow$



$$n_1 < n_2 < n_3$$

A Weaker Assumption

For unbiasedness, we assumed a zero conditional mean – $\mathbf{E(u/x_1, x_2,\ldots,x_k) = 0}$

For consistency, we can have the weaker assumption of zero mean and zero correlation

- $E(u) = 0$ (always the case if our model has a constant)
- **$Cov(x_j,u) = 0$, for $j = 1, 2, \ldots, k$** *(exogeneity)*

Without this assumption, OLS will be biased and inconsistent!

The second assumption is very critical and often a point of concern (you will see in the papers…)

Violation of exogeneity

What if **Cov(x_j,u) = 0, for j = 1, 2, ..., k** is violated?

- E.g. Cov($\mathbf{x}_j$, **u**) = 0 for j = 1,2,...K-1
- But Cov ($\mathbf{x}_k$, **u**) $\neq$ 0 → Then $\mathbf{x}_k$ is potentially endogenous.
- Least squares estimation will result in biased and inconsistent estimates for <u>all</u> the β_j. (Note: it is the case that even if only one of the explanatory variables is endogenous all the coefficient estimates will be biased and inconsistent.)
- E.g. unobserved variable
  - Wage = f(age, educ, ability)
  - Possible correlation between education and ability (why?)
  - Ability is unobserved, disappears in the error term

  → Cov (educ, u(ability)) $\neq$ 0

Violation of exogeneity

What if **Cov(x_j,u) = 0, for j = 1, 2, ..., k** is violated?

Possible solutions:

1. Try to find a suitable **proxy** for the unobserved variable
   - E.g. IQ in the case of ability
2. Use **panel data**
   - Assume the unobserved variable does not change over time and use a fixed effects model
3. Leave the unobserved variable in the error term but use a different estimation method that recognises the presence of the omitted variable (**Instrumental variables** method)
4. Use experiments
   - True versus natural experiments

2-4 are also suited to establish **causality** (as opposed to only showing statistical **correlation**)

Panel Data

- Data following the same cross-section units over time
- Panel data can be used to address omitted variable bias
  - Assume the unobserved variable does not change over time and use a fixed effects model
- Suppose the population model is
  - $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$
- Here the error has a time-constant component, $\upsilon_{it} = a_i + u_{it}$
  - E.g. assume a_i is individual ability which does not change over time
  - If a_i is correlated with the x's, OLS will be biased, since we a_i is part of the error term
- With panel data, the unobserved fixed effect can be differenced out

Panel Data

First differencing

- Subtract one period from the other, to obtain

  $\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + \ldots + \beta_k \Delta x_{ik} + \Delta u_i$

- The fixed effect has disappeared. This model has no correlation between the *x*'s and the error term, so no bias → estimate the differenced model

Fixed effects estimation

- Consider the average over time of

  $y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$

- The average of a_i is a_i, so when subtracting the mean, a_i will be differenced out just as when doing first differences

- This method is also identical to including a separate intercept for every individual

Experiments

- We want to estimate the effect of a "treatment", e.g. a job training
- To estimate the treatment effect, we could just compare the treated units before and after treatment
- However, we might pick up the effects of other factors that changed around the time of treatment
- Therefore, we use a control group to "difference out" these confounding factors and isolate the treatment effect
- Diff-in-diff estimation in this context is only appropriate if treatment is allocated totally randomly in the population. This would be a **true** experiment, which is hard to do.
- However, in the social sciences this method is usually applied to data from **natural** experiments, raising questions about whether treatment is truly random.
  - Natural experiments use arbitrary variation in a variable to imitate a true experiment
  - E.g. does an increase in minimum wage reduce employment? The „experiment": In April 1992 the minimum wage in New Jersey was raised from $4,25 to $5,05 per hour. Comparison to Pennsylvania where the minimum wage stayed the same.

Instrumental Variables

Consider the following regression model:

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

- Variation in the endogenous regressor X_i has two parts
  - the part that is uncorrelated with the error ("good" variation)
  - the part that is correlated with the error ("bad" variation)
- The basic idea behind instrumental variables regression is to isolate the "good" variation and disregard the "bad" variation
- Identify a valid instrument: A variable Z_i is a valid instrument for the endogenous regressor X_i if it satisfies two conditions:
  1. Relevance: $\text{corr}(Z_i, X_i) \neq 0$
  2. Exogeneity: $\text{corr}(Z_i, e_i) = 0$
- E.g. use parents' education as an instrument for own education

Instrumental Variables

The most common IV method is two-stage least squares (2SLS)

- Stage 1: Decompose X_i into the component that can be predicted by Z_i and the problematic component

$$X_i = \alpha_0 + \alpha_1 Z_i + \mu_i$$

- Stage 2: Use the predicted value of X_i from the first-stage regression to estimate its effect on Y_i

$$y_i = \gamma_0 + \gamma_1 \text{X-hat}_i + \nu_i$$

Complications

Many possible complications:

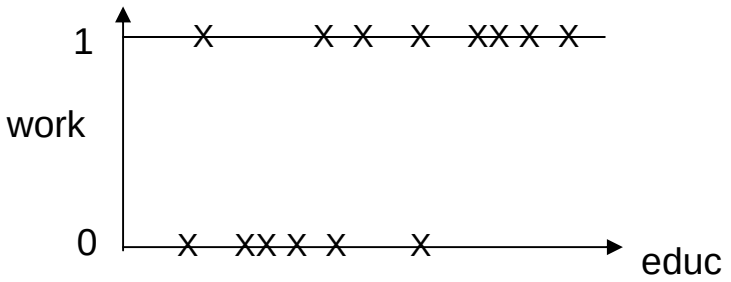
- Inconsistent OLS estimations
  - Missing data / omitted variables
  - Measurement error in explanatory or dependent variables
  - Wrong functional form of the regression
  - Non-random samples
- Consistent but inefficient estimations (not the smallest variance)
  - Heteroskedasticity (non-constant variance) of the error terms
  - Autocorrelation of the error terms
- Other complications
  - Lagged variables
  - Outliers
  - …

Discrete dependent variables

What if the dependent variable (y) is a dummy variable?

- Example: we want to analyze the determinants of the participation to the labour force (1 – work, 0 – no work)

→ the interpretation of β_j would be the change in the probability of finding work when x_j changes (linear probability model)



- But: OLS may yield values outside [0,1]
- OLS is only a starting point
- → **probit** or **logit** use cumulative distribution functions which can be interpreted in terms of probabilities
- The coefficients of the logit/probit model DO NOT have the same interpretation as in the linear regression model
- The sign tells us the direction of the effect, but the coefficient does not tell u generally to what extent the explanatory variable alters the dependent variable
- Careful when interpreting coefficients:
  - Marginal effect for the average person (person with average education)
  - Average effect over all people