Business Analytics & Machine Learning

Regression Diagnostics

Prof. Dr. Martin Bichler & Prof. Dr. Jalal Etesami

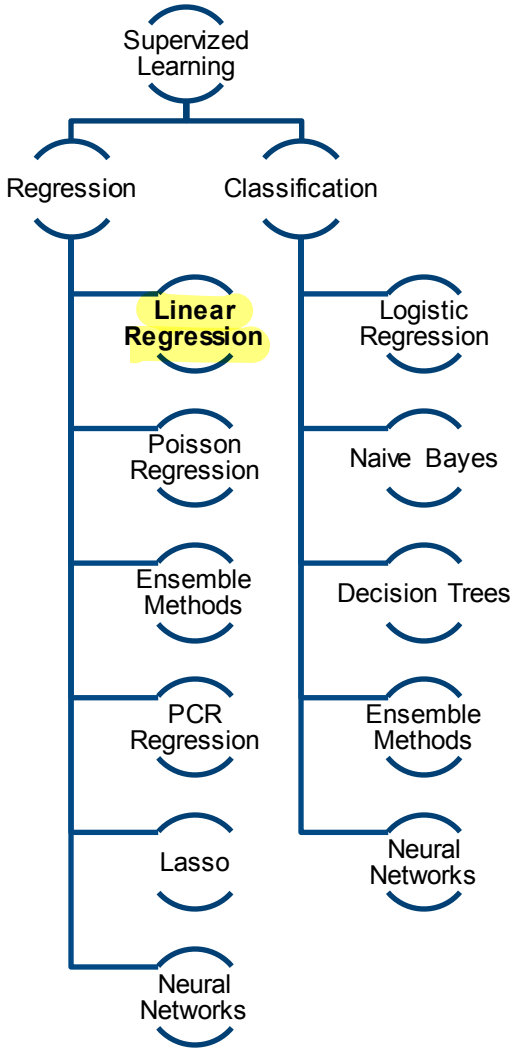
Department of Computer Science

School of Computation, Information, and Technology

Technical University of Munich

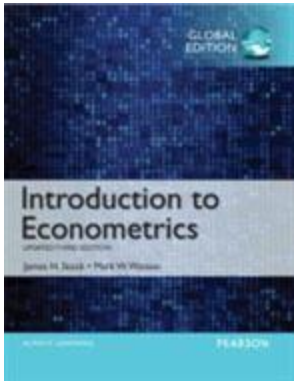
Course Content

- Introduction
- Regression Analysis
- **Regression Diagnostics**
- Logistic and Poisson Regression
- Naive Bayes and Bayesian Networks
- Decision Tree Classifiers
- Data Preparation and Causal Inference
- Model Selection and Learning Theory
- Ensemble Methods and Clustering
- Dimensionality Reduction
- Association Rules and Recommenders
- Convex Optimization
- Neural Networks
- Reinforcement Learning

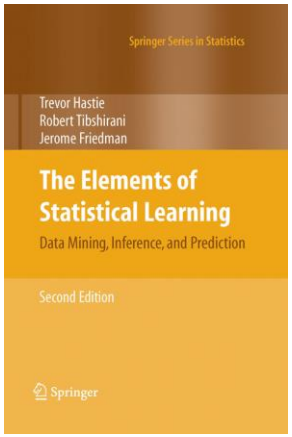


Recommended Literature

- **Introduction to Econometrics**
  - James H. Stock and Mark W. Watson
  - Chapter 6, 7, 10, 17, 18



- **The Elements of Statistical Learning**
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - http://web.stanford.edu/~hastie/Papers/ESLII.pdf
  - Section 3: Linear Methods for Regression



- **An Introduction to Statistical Learning: With Applications in R**
  - Gareth James, Trevor Hastie, Robert Tibshirani
  - Section 3: Linear Regression

Multiple Linear Regression

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

y	$\boldsymbol{X}$	β	ε
$(n \times 1)$	$(n \times (p+1))$	$((p+1) \times 1)$	$(n \times 1)$

Reminder: Least Squares Estimation

$\mathbf{X}$ is $n \times (p+1)$, y is the vector of outputs
RSS(β) = $(y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta)$

If $\mathbf{X}$ is full rank, then $\mathbf{X}^T\mathbf{X}$ is positive definite

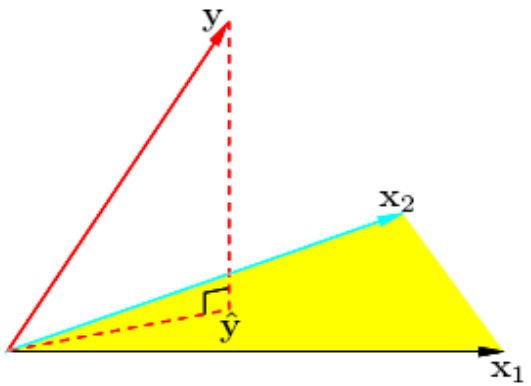
$$RSS = (y^T y - 2\beta^T \mathbf{X}^T \mathrm{y} + \beta^T \mathbf{X}^T \mathbf{X}\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T \mathrm{y} + 2\mathbf{X}^T \mathbf{X}\beta = 0$$ First-order condition

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

$$\hat{y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

"Hat" or projection matrix H



Source: Hastie et al. 2016, p. 46

Agenda for Today

The linear regression model for **a random sample** (no bias in the sampling) is computationally simple and „best“ if certain assumptions are satisfied.

Today, we will discuss the **main assumptions** and introduce selected tests that can help you check whether you can assume that the assumptions are satisfied in your data set.

There are several alternative tests that have been developed for each assumption. We introduce one of these tests, which will enable you to use the multiple linear regression in applications.

Gauss-Markov Theorem

The Gauss-Markov theorem states that in a linear regression model in which the errors
- have expectation zero and
- are uncorrelated and
- have equal variances,

the **best linear unbiased estimator** (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator.

- „unbiased“ means $E(\hat{\beta}_j) = \beta_j$
- „best“ means giving the **lowest variance** of the estimate as compared to other linear unbiased estimators
  - restriction to unbiased estimation is **not always the best**

    (will be discussed in the context of the ridge regression later)

Bias, Consistency, and Efficiency

Unbiased

$E(\hat{\beta}) = \beta$

Expected value for estimator "is true"

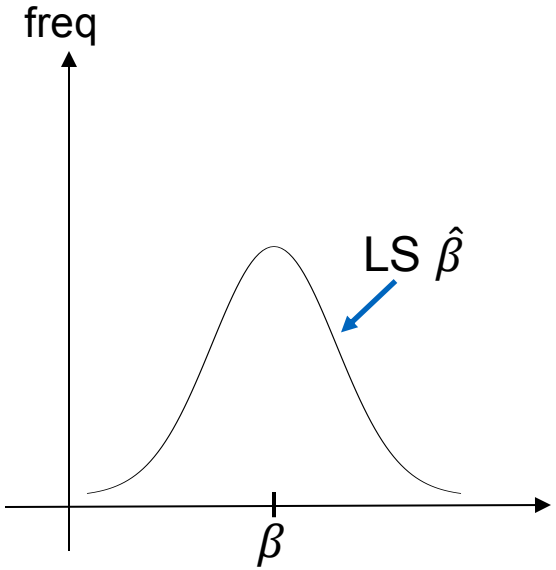
Consistent

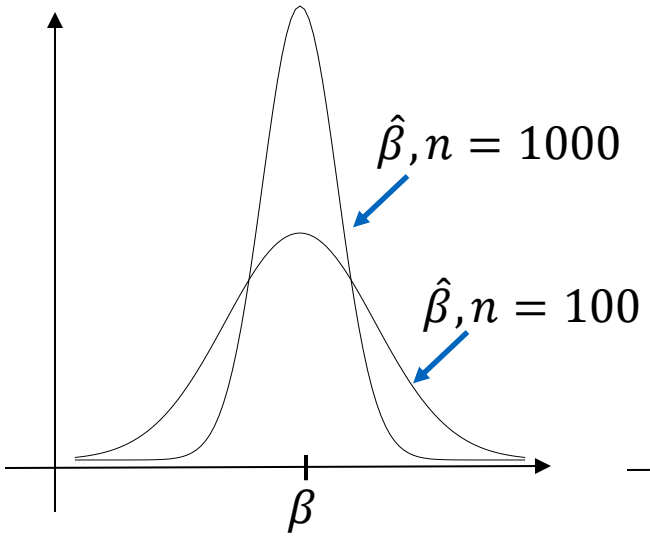
$\text{var}(\hat{\beta})$ decreases with increasing sample size n

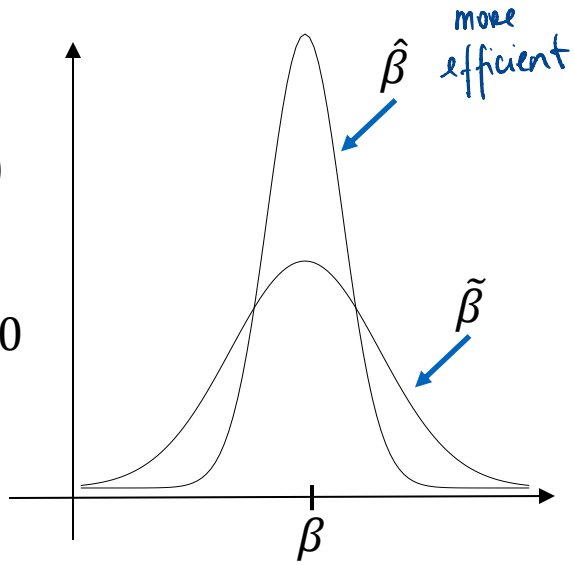
Efficient

$\text{var}(\hat{\beta}) < \text{var}(\tilde{\beta})$

estimator $\hat{\beta}$ has lower variance than any other estimator, $\tilde{\beta}$







Gauss-Markov Assumptions in Detail

The OLS estimator is the **best linear unbiased estimator (BLUE)**, iff

1) **Linearity**
   Linear relationship in parameters β

2) **No multicollinearity** of predictors
   No linear dependency between predictors

3) **Homoscedasticity**
   The residuals exhibit constant variance

4) **No autocorrelation**
   There is no correlation between the i^{th} and j^{th} residual terms

5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$) (i.e., exogeneity ($\text{cov}(\varepsilon, X) = 0$))

Gauss-Markov Assumptions in Detail

The OLS estimator is the **best linear unbiased estimator (BLUE)**, iff

1) **Linearity**
   Linear relationship in parameters β

2) **No multicollinearity** of predictors
   No linear dependency between predictors

3) **Homoscedasticity**
   The residuals exhibit constant variance

4) **No autocorrelation**
   There is no correlation between the i^{th} and j^{th} residual terms

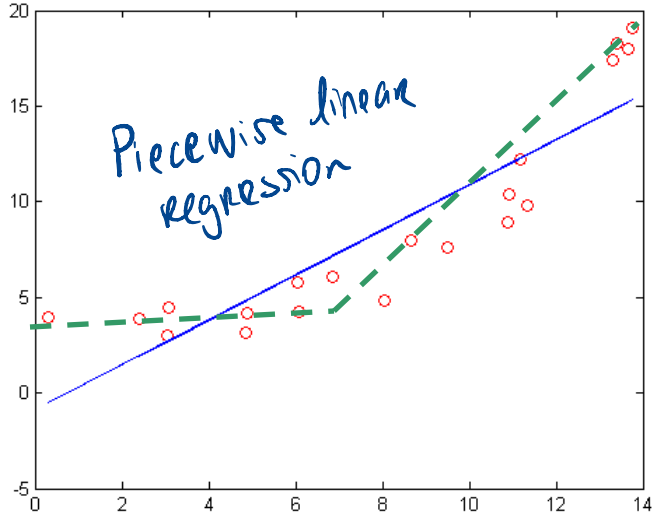
5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$) (i.e., exogeneity ($\text{cov}(\varepsilon, X) = 0$))

When Linearity Does Not Hold: Try to Reformulate

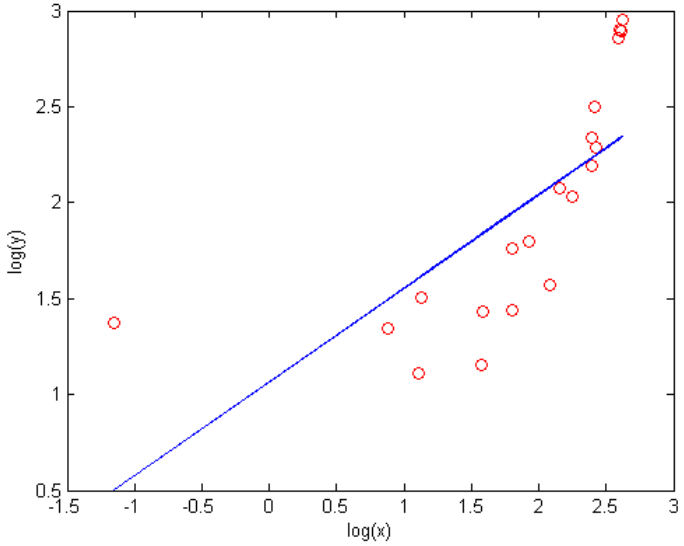
For non-linear regressions the OLS estimator from the last class might not be appropriate. Often, you can adapt your data such that you can still use the multiple linear regression. The following reformulations lead to models that are again linear in β:

- Polynomial regression (still a linear model):
  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \varepsilon$
- Transform either only X, only Y, or both variables, e.g.:
  $\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$
- Piecewise linear regression (aka. segmentation):
  $Y = \beta_0 + \beta_1 X[X > X_K] + \varepsilon$
  where $[X > X_K] = 0$ if $X \leq X_K$ and $[X > X_K] = 1$ if $X > X_K$

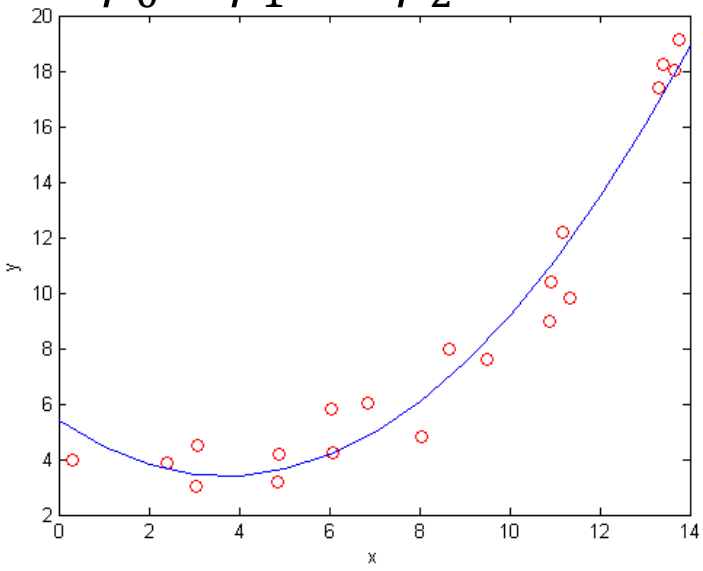
$$Y = \beta_0 + \beta_1 X[X > X_K] + \varepsilon$$



$$Log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$$



$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \varepsilon$$



Outliers

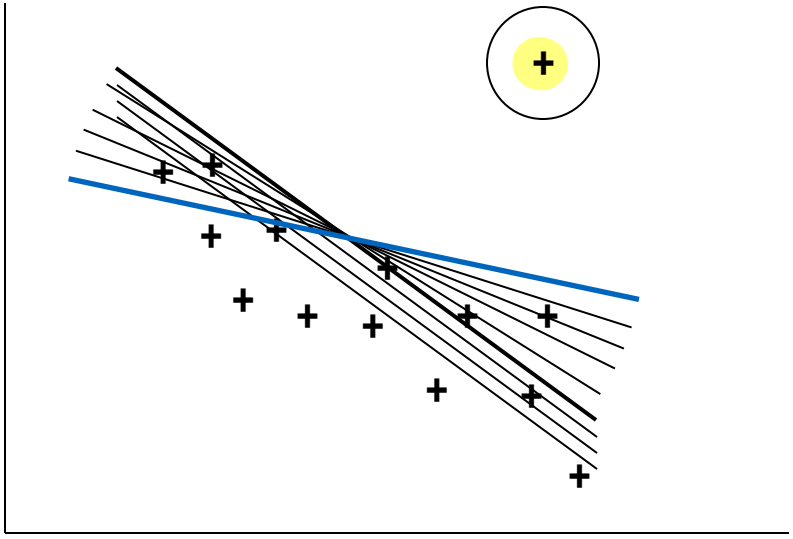
An outlier is an observation that is unusually small or large.

Several possibilities need to be investigated when an outlier is observed:

- There was an error in recording the value.
- The point does not belong in the sample.
- The observation is valid.

Identify outliers from
the scatter diagram.

There are also methods
for “robust” regression.



Gauss-Markov Assumptions in Detail

The OLS estimator is the **best linear unbiased estimator (BLUE)**, iff

1) **Linearity**
   Linear relationship in parameters β

2) **No multicollinearity** of predictors
   No linear dependency between predictors

3) **Homoscedasticity**
   The residuals exhibit constant variance

4) **No autocorrelation**
   There is no correlation between the i^{th} and j^{th} residual terms

5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$) (i.e., exogeneity ($\text{cov}(\varepsilon, X) = 0$))

Multicollinearity

no linear relathionship between x's

Independent variables must not be linearly dependent. If two independent variables were dependent, one could easily omit one.

To check for linear dependencies of two columns in a matrix, we can use the **rank**.

- The rank of the data matrix $\mathbf{X}$ is p, the number of columns
- $p < n$, the number of observations
- If there is an exact linear relationship among independent variables $\text{rank}(\mathbf{X}) = p$, $\mathbf{X}$ has full column rank
- If $\text{rank}(\mathbf{X}) < p$, $\mathbf{X}$ is singular -> impossible to calculate the inverse
  - Remember: $\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$

if one column is a linear combination of another, the rank will be lower than the number of cols

Also, **high correlation** between independent variables leads to issues wrt. the significance of predictors.

Check for Multicollinearity

A basic check of multicollinearity is to calculate the **correlation coefficient** for each pair of predictor variables.

- Large correlations (both positive and negative) indicate problems.
  - Large means greater than the correlations between predictors and response.
- It is possible that the pairwise correlations are small, and yet a linear dependence exists among three or even more variables.

Alternatively use the *variance inflation factor* (*VIF*).

Variance Inflation Factor

- $\text{VIF} = \frac{1}{1-R_k^2}$, where the R_k^2 value here is the value when the predictor in question (k) is set as the dependent variable.

- For example, if the VIF = 10, then the respective R_k^2 would be 90%.
  This would mean that 90% of the variance in the predictor in question can be explained by the other independent variables.

- Because so much of the variance is captured elsewhere, removing the predictor in question should not cause a substantive decrease in overall R^2.
  The rule of thumb is to remove variables with VIF scores greater than 10.

$y \rightarrow \{x_1, x_2, \dots x_n\}$

$x_k \leftarrow \{x_1, x_2, \dots, x_p\} / \{x_k\}$

$k \in \{1, \dots, p\}$

- try to regress x_k on the other values
  - if large (~ 1.0) → there is collinearity
  - if not, no relation ✓

  (Remember R^2 measure)

Consequence - Non-Significance

If a variable has a non-significant t-value, then either

- the variable is not related to the response, or
  (→ Small t-value, small VIF, small correlation with response)
- the variable is related to the response, but it is not required in the regression because it is strongly related to a third variable that is in the regression, so we don't need both
  (→ Small t-value, big VIF, big correlation with response).

The usual remedy is to drop one or more variables from the model.

Example

Y	X1	X2	X3	X4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Example

Big R^2

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.982
Model:	OLS	Adj. R-squared:	0.974
Method:	Least Squares	F-statistic:	111.5
Date:	Mon, 06 Nov 2023	Prob (F-statistic):	4.76e-07
Time:	11:08:39	Log-Likelihood:	-26.918
No. Observations:	13	AIC:	63.84
Df Residuals:	8	BIC:	66.66

	coef	std err	t	P>|t|	[0.025	0.975]
const	62.4054	70.071	0.891	0.399	-99.179	223.989
X1	1.5511	0.745	2.083	0.071	-0.166	3.269
X2	0.5102	0.724	0.705	0.501	-1.159	2.179
X3	0.1019	0.755	0.135	0.896	-1.638	1.842
X4	-0.1441	0.709	-0.203	0.844	-1.779	1.491

$H_0: X_n = 0 \Rightarrow$ we want to reject this

Large p-values

```
data.corr()
      Y     X1     X2     X3     X4
Y     1.00  0.73   0.82  -0.53  -0.82
X1    0.73  1.00   0.23  -0.82  -0.25
X2    0.82  0.23   1.00  -0.14  -0.97
X3   -0.53 -0.82  -0.14   1.00   0.03
X4   -0.82 -0.25  -0.97   0.03   1.00
```

Big correlation

Data

```
from statsmodels.stats.outliers_influence import variance_inflation_factor as vif

for index, variable_name in enumerate(features.columns):
    if variable_name != "const":
        print(f"VIF for variable {variable_name} is {vif(features, index)}")
```

X1	X2	X3	X4
38.49621	254.42317	46.86839	282.51286

Very large!

Very large!

Drop X4

```
for index, variable_name in enumerate(features.columns):
    if variable_name != "const":
        print(f"VIF for variable {variable_name} is {vif(features, index)}")
```

```
      X1          X2          X3
3.251068 1.063575 3.142125
```

VIF's now small

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.982
Model:	OLS	Adj. R-squared:	0.976
Method:	Least Squares	F-statistic:	166.3
Date:	Mon, 06 Nov 2023	Prob (F-statistic):	3.37e-08
Time:	11:08:39	Log-Likelihood:	-26.952
No. Observations:	13	AIC:	61.90
Df Residuals:	8	BIC:	64.16

R^2 hardly decreased ✓

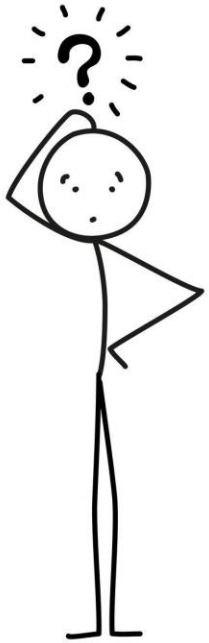
	coef	std err	t	P>|t|	[0.025	0.975]
const	48.1936	3.913	12.315	0.000	39.341	57.046
X1	1.6959	0.205	8.290	0.000	1.233	
X2	0.6569	0.044	14.851	0.000	0.557	
X3	0.2500	0.185	1.354	0.209	-0.168	0.668

smaller p values are better

X1, X2 now signif

Can you explain the intuition behind the VIF?

identify which variables are relevant for the regression, checking for multicollinearity between them.



Gauss-Markov Assumptions in Detail

The OLS estimator is the **best linear unbiased estimator (BLUE)**, iff

1) **Linearity**
   Linear relationship in parameters β

2) **No multicollinearity** of predictors
   No linear dependency between predictors

3) **Homoscedasticity**
   The residuals exhibit constant variance

4) **No autocorrelation**
   There is no correlation between the i^{th} and j^{th} residual terms

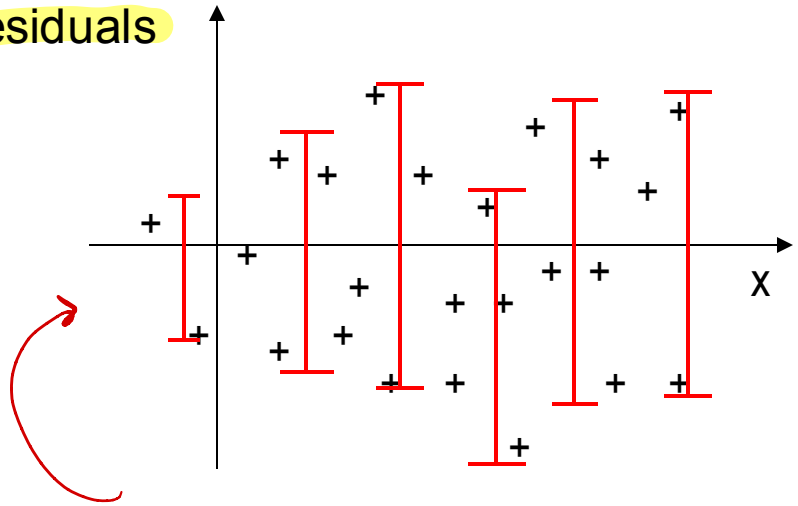
5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$) (i.e., exogeneity ($\text{cov}(\varepsilon, X) = 0$))

Homoscedasticity

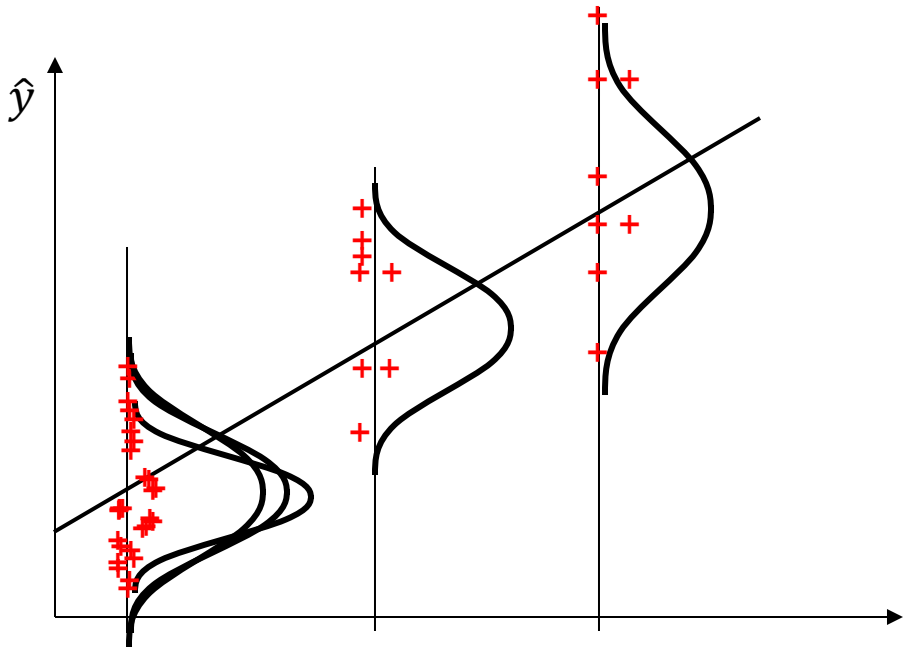
↳ residual distribution does not change much with the changes in the independent variables

- When the requirement of **a constant variance** is not violated, we have homoscedasticity.
- We also assume residuals to be **normally distributed**.
- If the data is not homoscedastic, a different estimator (e.g., weighted least squares) might be better than OLS.

Residuals

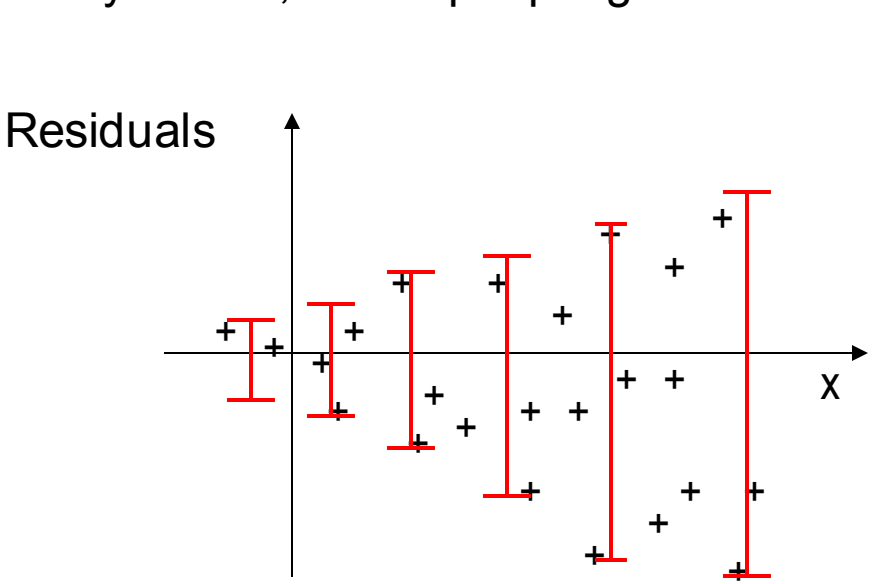


The spread of the data points does not change much.

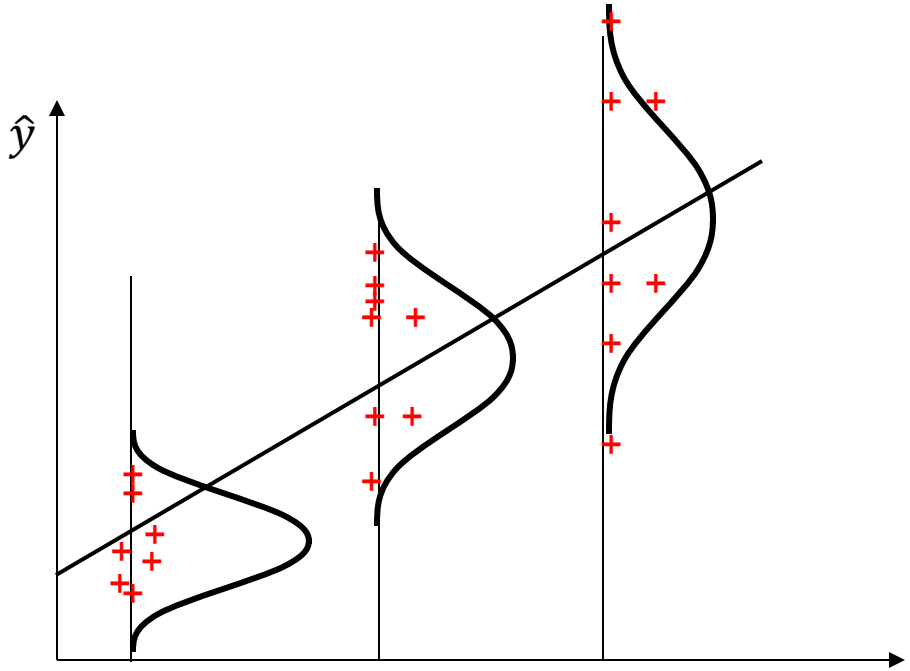


Heteroscedasticity

- When the requirement of a constant variance is violated, we have heteroscedasticity ($\mathrm{var}(\varepsilon_i|x_{1i},\dots,x_{pi})$ not constant).
- Heteroscedasticity leads to biased error terms and p-values of significance tests.
- **Example**: annual income when predicted by age. Entry level jobs are often paid very similar, but as people grow older their income distribution increases.



The spread increases with x.



Glejser Test

Apart from visual inspection, the Glejser test is a simple statistical test. It regresses the residuals on the explanatory variable that is thought to be related to the heteroscedastic variance.

1. Estimate the original regression with OLS and find the sample residuals e_i.
2. Regress the absolute value $|e_i|$ on the explanatory variable that is associated with heteroscedasticity.

$$|e_i| = \gamma_0 + \gamma_1 X_i + v_i$$

$$|e_i| = \gamma_0 + \gamma_1 \sqrt{X_i} + v_i$$

$$|e_i| = \gamma_0 + \gamma_1 \frac{1}{X_i} + v_i$$

→ if there is a relation, it's bad (heteroscedascity)
→ Variance should not change on the residuals given x_i's changes

3. Select the equation with the highest R^2 and lowest standard errors to represent heteroscedasticity.
4. Perform a t-test on γ_1. If γ_1 is statistically significant, the null hypothesis of homoscedasticity can be rejected. We test if one of the independent variables is significantly related to the variance of our residuals.

White Test

$x_1, x_2, x_1^2, x_2^2, x_1 x_2 \ \ldots$

The White test assumes more complex relationships, and also models interaction terms. You do not have to choose a particular X and do not need normally distributed residuals.

The test is also based on an auxiliary regression of e^2 on all the explanatory variables (X_j), their squares (X_j^2), and all their cross products.

$y \leftarrow \{x_1, x_2, x_3\}$

$$e^2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + v$$

$\#\{x_1, x_2, x_3, x_3 x_2, x_1 x_3, \ldots\} = 3^3 = 27$

With more than 2 independent variables, you need to analyze the product of each independent variable with each other independent variable. A large R^2 counts against homoskedasticity.

The test statistic is nR^2, where n is the sample size and R^2 is the unadjusted.

- The statistic has an asymptotic chi-square (χ^2) distribution with $\mathrm{df} = p$, where p is the no. of all explanatory variables in the auxiliary model.
- H_0: All the variances σ_i^2 are equal (i.e., homoscedastic)
- Reject H_0 if $\chi^2 > \chi_{cr}^2$

Gauss-Markov Assumptions in Detail

The OLS estimator is the **best linear unbiased estimator (BLUE)**, iff

1) **Linearity**
   Linear relationship in parameters β

2) **No multicollinearity** of predictors
   No linear dependency between predictors

3) **Homoscedasticity**
   The residuals exhibit constant variance

4) **No autocorrelation**
   There is no correlation between the i^{th} and j^{th} residual terms

5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$) (i.e., exogeneity ($\text{cov}(\varepsilon, X) = 0$))

Applications of Linear Regressions to Time Series Data

Average hours worked per week by manufacturing workers:

Period	Hours	Period	Hours	Period	Hours	Period	Hours
1	37.2	11	36.9	21	35.6	31	35.7
2	37.0	12	36.7	22	35.2	32	35.5
3	37.4	13	36.7	23	34.8	33	35.6
4	37.5	14	36.5	24	35.3	34	36.3
5	37.7	15	36.3	25	35.6	35	36.5
6	37.7	16	35.9	26	35.6		
7	37.4	17	35.8	27	35.6		
8	37.2	18	35.9	28	35.9		
9	37.3	19	36.0	29	36.0		
10	37.2	20	35.7	30	35.7		

Forecasting Linear Trend using the Multiple Regression

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$
where: Y_i = data value for period i
$\hat{Y} = 37.416 - 0.0614 X_i$

OLS Regression Results

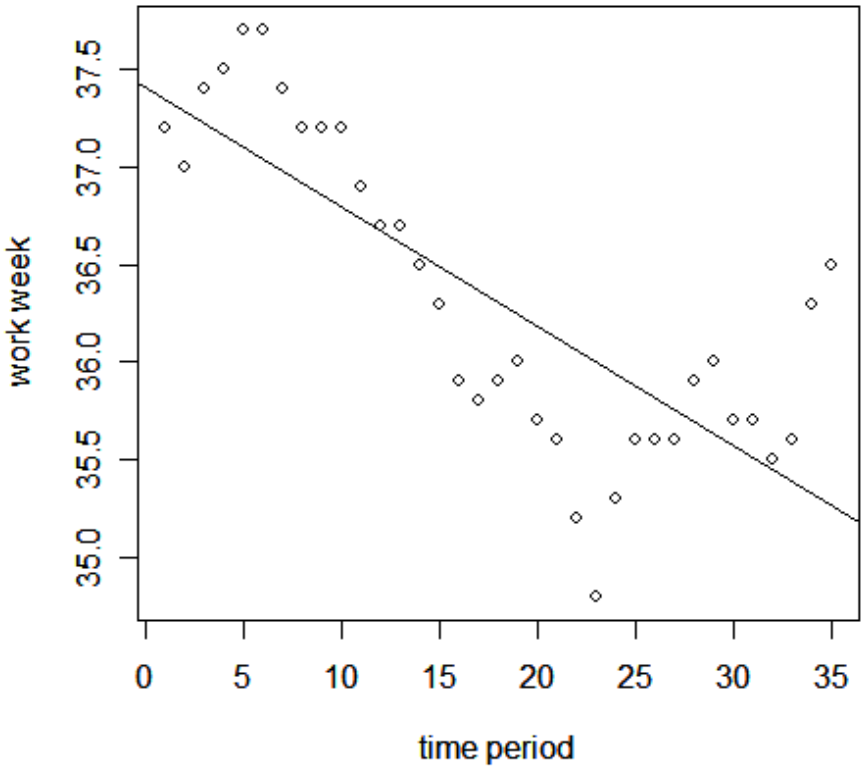
Dep. Variable:	y	R-squared:	0.612
Model:	OLS	Adj. R-squared:	0.600
Method:	Least Squares	F-statistic:	51.95
Date:	Mon, 06 Nov 2023	Prob (F-statistic):	2.90e-08
Time:	11:53:29	Log-Likelihood:	-24.835
No. Observations:	35	AIC:	53.67
Df Residuals:	33	BIC:	56.78
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	37.416	0.175	213.744	0.000	37.053	37.765
x1	-0.0614	0.008	-7.208	0.000	-0.078	-0.044

Omnibus:	0.812	Durbin-Watson:	0.278
Prob(Omnibus):	0.666	Jarque-Bera (JB):	0.155
Skew:	0.047	Prob(JB):	0.925
Kurtosis:	3.312	Cond. No.	42.3

p-value for hypothesis ($\beta = 0$)

Hours Worked Data - A Linear Trend Line

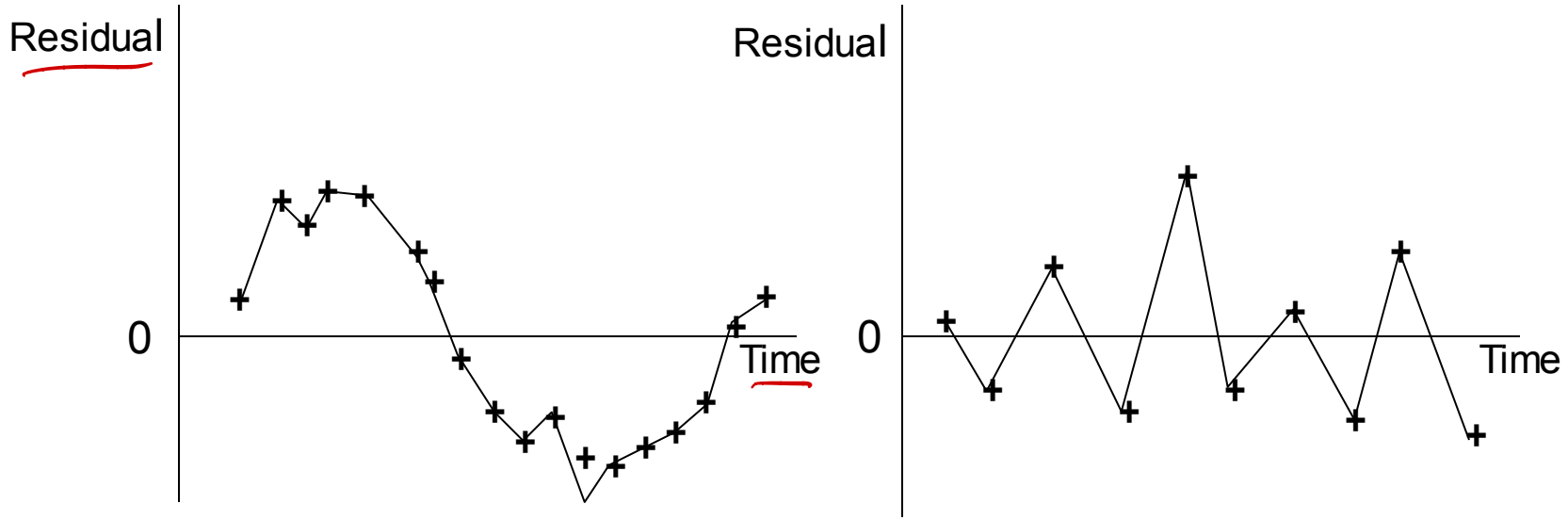


correlation between residuals at different points of time or across observations. It indicates whether there is a systematic pattern of correlation in the residuals, which violates the assumption of independence.

Autocorrelation

Examining the residuals over time, no pattern should be observed if the errors are independent.

Autocorrelation can be detected by graphing the residuals against time, or Durbin-Watson statistic.



Autocorrelation

Reasons leading to autocorrelation:

- Omitted an important variable
- Functional misfit
- Measurement error in independent variable

Use Durbin-Watson (DW) statistic to test for first order autocorrelation. DW takes values within [0, 4]. For no serial correlation, a value close to 2 (e.g., 1.5-2.5) is expected.

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- $DW = 2$ – no autocorrelation
- $DW = 0$ – perfect positive autocorrelation → $e_i - e_{i-1} = 0$
- $DW = 4$ – perfect negative autocorrelation → $e_i + e_{i-1} = 0$

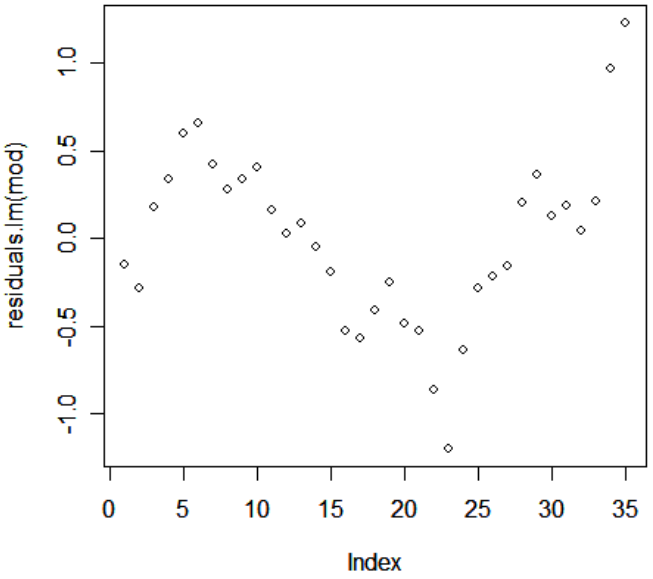
Test for Autocorrelation in R

```
from statsmodels.formula.api import ols
from statsmodels.stats.stattools import durbin_watson

model = ols('target ~ period', data=data).fit()
durbin_watson(model.resid)
```

0.2775895

- there might be a variable which is important and wasn't collected



Modeling Seasonality

A regression can estimate both the trend and additive seasonal indexes.

- Create dummy variables which indicate the season.
- Regress on time and the seasonal variables.
- Use the multiple regression model to forecast.

For any season, e.g., season 1, create a column with 1 for time periods which are season 1, and zero for other time periods (only season -1 dummy variables are required).

Dummy Variables

Trend variable

Seasonal variables

	Quarterly Input Data				
	Sales	t	Q1	Q2	Q3
Year 1	3497	1	Spring	0	0
	3484	2	0	Summer	0
	3553	3	0	0	Fall
	3837	4	Not Spring	Not Summer	Not Fall
Year 2	3726	5	1	0	0
	3589	6	0	1	0

Modelling Seasonality

The model which is fitted (assuming quarterly data) is

$$y = \beta_0 + \beta_1 t + \beta_2 Q_1 + \beta_3 Q_2 + \beta_4 Q_3$$

Only 3 quarters are explicitly modelled.

Otherwise:

$Q_1 = 1 - (Q_2 + Q_3 + Q_4)$, for all 4 quarters → Multicollinearity

Allows to test for seasonality.

Gauss-Markov Assumptions in Detail

The OLS estimator is the **best linear unbiased estimator (BLUE)**, iff

1) **Linearity**
   Linear relationship in parameters β

2) **No multicollinearity** of predictors
   No linear dependency between predictors

3) **Homoscedasticity**
   The residuals exhibit constant variance

4) **No autocorrelation**
   There is no correlation between the i^{th} and j^{th} residual terms

5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$) (i.e., exogeneity ($\text{cov}(\varepsilon, X) = 0$))

Gender Bias?

Consider the acceptance rates for the following groups of men and women who applied to college.

Counts	Accepted	Not accepted	Total
Men	198	162	**360**
Women	88	112	**200**
Total	**286**	**274**	**560**

Percents	Accepted	Not accepted
Men	55%	45%
Women	44%	56%

A higher percentage of men were accepted: Is there evidence of discrimination?

Gender Bias?

Consider the acceptance rates broken down by type of school.

Computer Science

Counts	Accepted	Not accepted	Total
Men	18	102	**120**
Women	24	96	**120**
Total	**42**	**198**	**240**

Percents	Accepted	Not accepted
Men	15%	85%
Women	**20%**	80%

Management

Counts	Accepted	Not accepted	Total
Men	180	60	**240**
Women	64	16	**80**
Total	**244**	**76**	**320**

Percents	Accepted	Not accepted
Men	75%	25%
Women	**80%**	20%

Explanations?

Within each school a higher percentage of women were accepted!

- There was no discrimination against women.
- Women rather applied to schools with lower acceptance rates.
- Men applied in schools with higher acceptance rates.

This is an example of **Simpson's paradox**:

- When the **omitted (aka. confounding) variable** (Type of School) is ignored the data seem to suggest discrimination against women.
- However, when the type of school is considered, the association is reversed and suggests discrimination against men.
- see also https://en.wikipedia.org/wiki/Simpson%27s_paradox.

But we often do not have all relevant variables in the data ...?

Endogeneity due to Omitted Variables

Endogeneity means $(\mathrm{corr}(\varepsilon_i, X_i) \neq 0) \Rightarrow E(\varepsilon_i | X_i) \neq 0$

- Simple test: analyze the correlation of the residuals and an independent variable.
- Reason for endogeneity: measurement errors, variables that affect each other, **omitted variables**(!)

Omitted (aka. confounding) variables:

- True model: $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i$
- Estimated model: $y_i = \beta_0 + \beta_1 x_1 + u_i$
- Now $u_i = \beta_2 x_2 + e_i$. If x_1 and u_i are correlated and x_2 affects y, this leads to endogeneity.

Why is this a problem?

Omitted Variable Bias

- A reason for endogeneity might be that relevant variables are omitted from the model.
- For example, enthusiasm or willingness to take risks of an individual describe unobserved heterogeneity.
- Can we control for these effects when estimating our regression model?

Various techniques have been developed to address endogeneity in **panel data**.

- In panel data, the same individual is observed multiple times!
- see https://en.wikipedia.org/wiki/Omitted-variable_bias
- see https://en.wikipedia.org/wiki/Endogeneity_(econometrics)

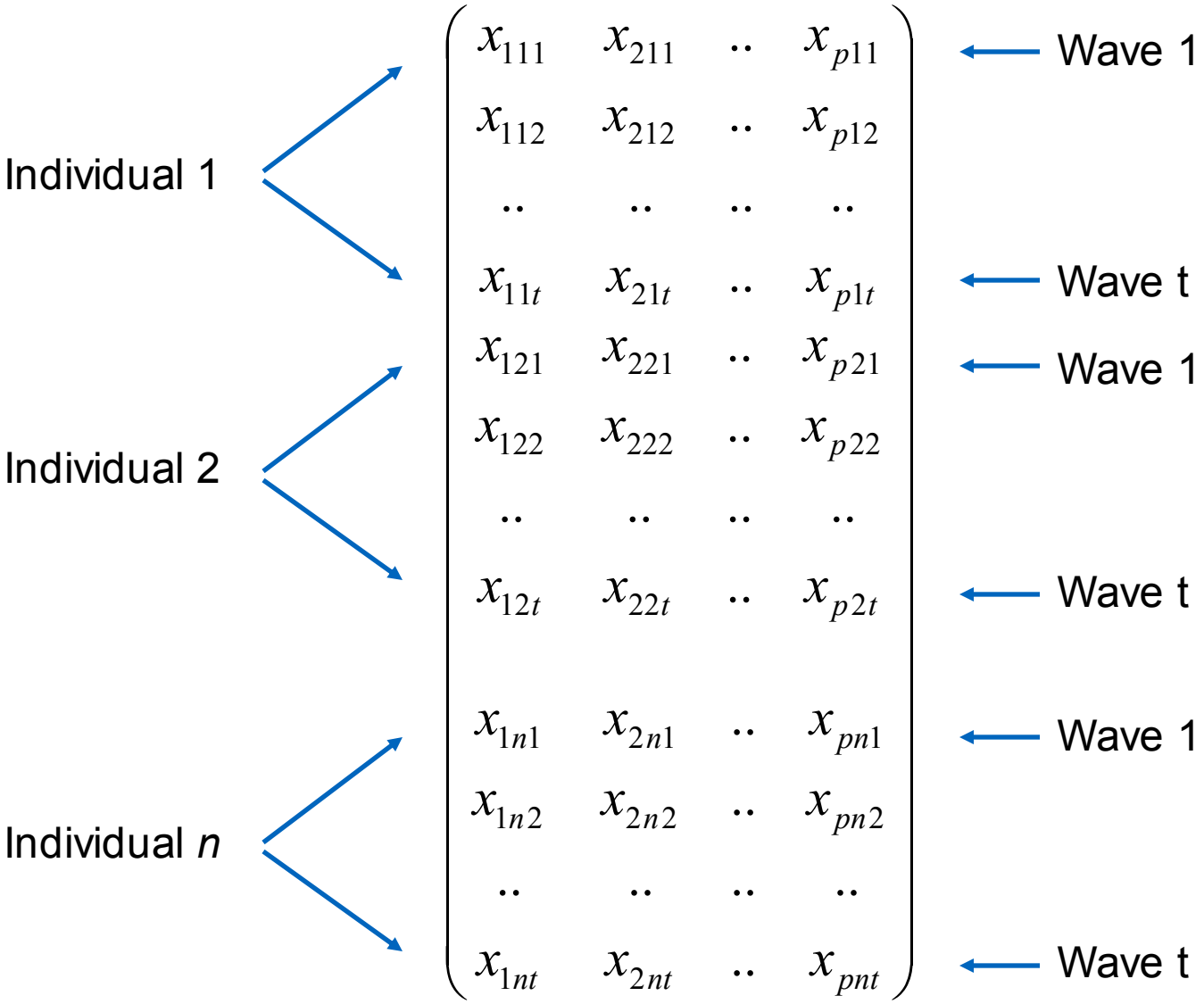
Panel Data vs. Cross-Section Data

Cross-section data collected in **observational studies** refers to data observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time.

A **panel data set**, or longitudinal data set, is one where there are repeated observations on the same units, which makes it possible to overcome an omitted variable bias.

- A balanced panel is one where every unit is surveyed in every time period.
- In an unbalanced panel some individuals have not been recorded in some time period.

The Panel Data Structure



Modeling Fixed Effects

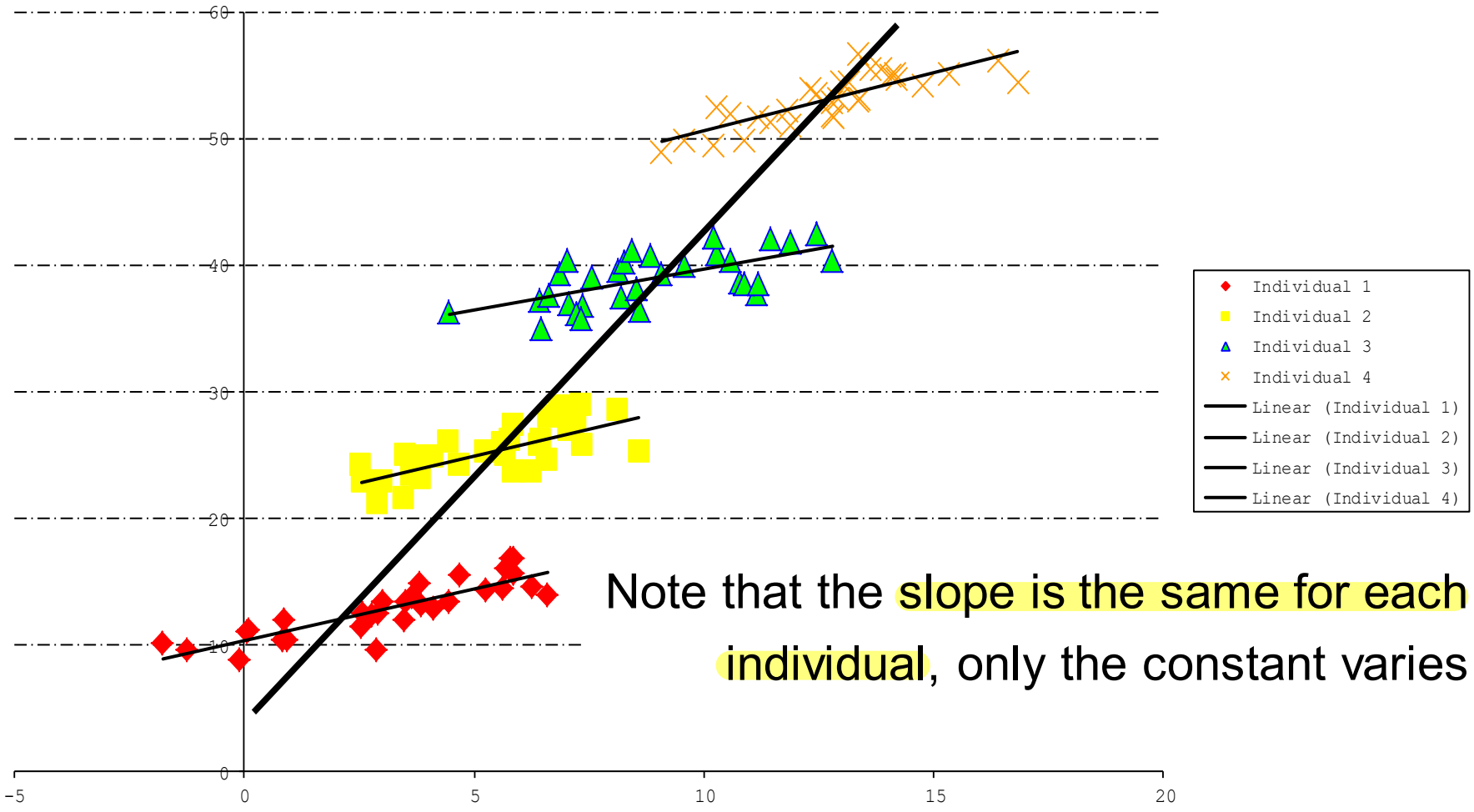
Fixed effects
- assume λ_i are constants (there is **endogeneity**)
- effects are **correlated with the other covariates***
- see also https://en.wikipedia.org/wiki/Fixed_effects_model

Random effects (not discussed here)
- assume λ_i are drawn independently from some probability distribution
- effects are **uncorrelated with the other covariates**
- see also https://en.wikipedia.org/wiki/Random_effects_model

Specific packages in R are available for fixed, random, and mixed effects models, which combine both (e.g., plm).
We only deal with fixed effects in this class. You might want to test, which effect is most likely in applications.

*We talk about a factor if it is a categorical variable and a covariate if it is continuous.

Fixed Effects Models



Note that the slope is the same for each individual, only the constant varies

The Fixed Effect Model

Treat λ_i (the individual-specific heterogeneity) as **a constant** for each individual.

$$y_{it} = (\beta_0 + \lambda_i) + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_p x_{pit} + \varepsilon_{it}$$

λ is part of constant, but varies by individual i

Various estimators for fixed effect models:
first differences, within, between, least squares dummy variable estimator.

First-Differences Estimator

Eliminating unobserved heterogeneity by taking first differences

Original equation

$$y_{it} = \beta_0 + \lambda_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_p x_{pit} + \varepsilon_{it}$$

difference between two subsequent time points

Lag one period and subtract

$$y_{it} - y_{it-1} = \beta_0 + \lambda_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_p x_{pit} + \varepsilon_{it}$$
$$-\beta_0 - \lambda_i - \beta_1 x_{1it-1} - \beta_2 x_{2it-1} - \ldots - \beta_p x_{pit-1} - \varepsilon_{it-1}$$

Constant and individual effects eliminated

$$y_{it} - y_{it-1} = \beta_1 \left(x_{1it} - x_{1it-1} \right) + \beta_2 \left(x_{2it} - x_{2it-1} \right) + \ldots$$
$$+\beta_p \left(x_{pit} - x_{pit-1} \right) + \left(\varepsilon_{it} - \varepsilon_{it-1} \right)$$

Transformed equation

$$\Delta y_{it} = \beta_1 \Delta x_{1it} + \beta_2 \Delta x_{2it} + \ldots + \beta_p \Delta x_{pit} + \Delta \varepsilon_{it}$$

then we can use OLS to compute the β's

How to Estimate a Model with Fixed Effects

Least squares **dummy variable estimator**

- uses a dummy variable for each individual (or firm, etc.), which we assume to have a fixed effect.

Within estimator

- take deviations from individual means and apply least squares

$$y_{it} - \bar{y}_i = \beta_1(x_{1it} - \bar{x}_{1i}) + \cdots + \beta_p\left(x_{pit} - \bar{x}_{pi}\right) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

  relies on variations within individuals.

Gauss-Markov Assumptions in Detail

The OLS estimator is the **best linear unbiased estimator (BLUE)**, iff

1) **Linearity**
   Linear relationship in parameters β

   transformation

2) **No multicollinearity** of predictors
   No linear dependency between predictors

   no linear relationship between x's
   remove variables

3) **Homoscedasticity**
   The residuals exhibit constant variance

   - gleicher Test, White Test
   - hetero ... vs homo
   - Corr between x's and ε's

4) **No autocorrelation**
   There is no correlation between the i^{th} and j^{th} residual terms

   - compute residuals
   - compute relations between residuals

5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$) (i.e., exogeneity ($\text{cov}(\varepsilon, X) = 0$))

   no relation between residuals and x's
   - fixed effect model
   → first-differences model