Business Analytics

Regression Diagnostics

Prof. Dr. Martin Bichler

Decision Sciences & Systems

Department of Informatics

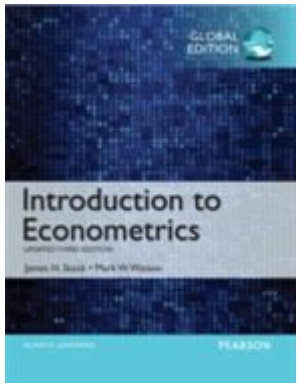
Technische Universität München

Course Content

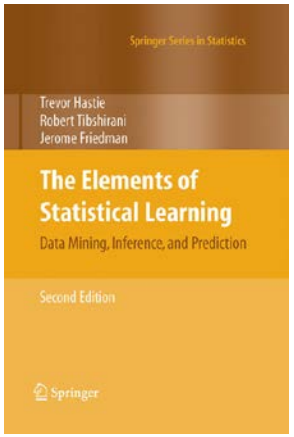
- Introduction
- Regression Analysis
- **Regression Diagnostics**
- Logistic and Poisson Regression
- Naive Bayes and Bayesian Networks
- Decision Tree Classifiers
- Data Preparation and Causal Inference
- Model Selection and Learning Theory
- Ensemble Methods and Clustering
- High-Dimensional Problems
- Association Rules and Recommenders
- Neural Networks

Recommended Literature

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



- **The Elements of Statistical Learning**
  (Trevor Hastie, Robert Tibshirani, Jerome Friedman)
  - http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf
  - Section 3: Linear Methods for Regression



- **An Introduction to Statistical Learning: With Applications in R**
  (Gareth James, Trevor Hastie, Robert Tibshirani)
  - http://www-bcf.usc.edu/~gareth/ISL/
  - Section 3: Linear Regression

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”

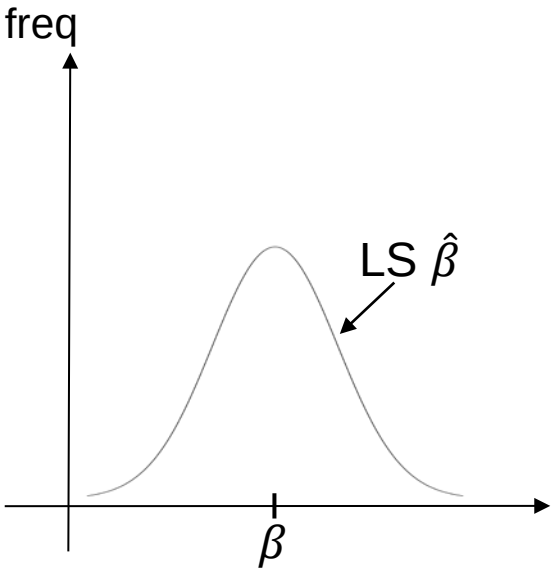
Consistent

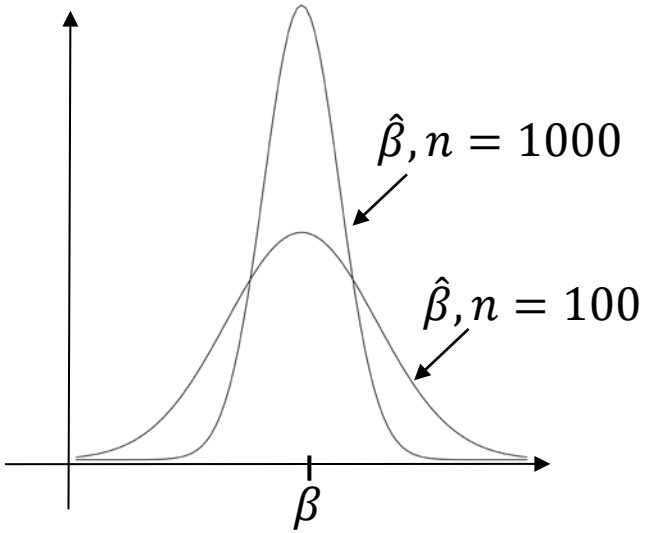
$var(\hat{\beta})$ decreases with increasing sample size n

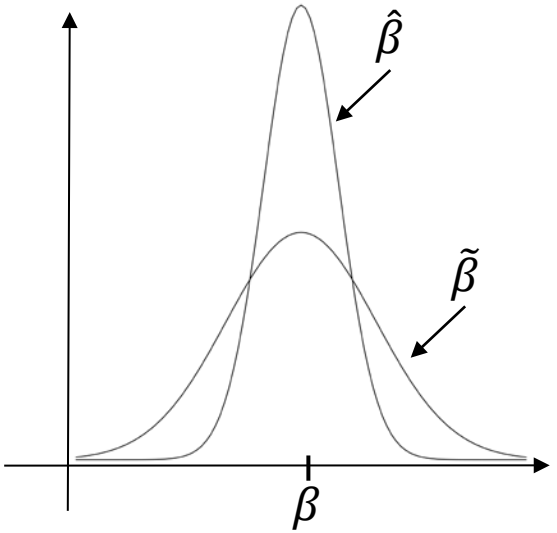
Efficient

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

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 (does not apply to predictors)

2) **No multicollinearity** of predictors (data redundancy, overfit, large std. errors, imprecise estimates)

   No linear dependency between predictors

3) **Homoscedasticity** (for efficiency & accurate hypoth. testing)

   The residuals exhibit constant variance

4) **No autocorrelation** (underesti. std. error, error not random)

   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 ($cov(\varepsilon, X) = 0$))

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 ($cov(\varepsilon, X) = 0$))

When Linearity Does Not Hold: Try to Reformulate

- Polynomial regression (still a linear model):

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

- Transform either X or Y or both variables, e.g.:

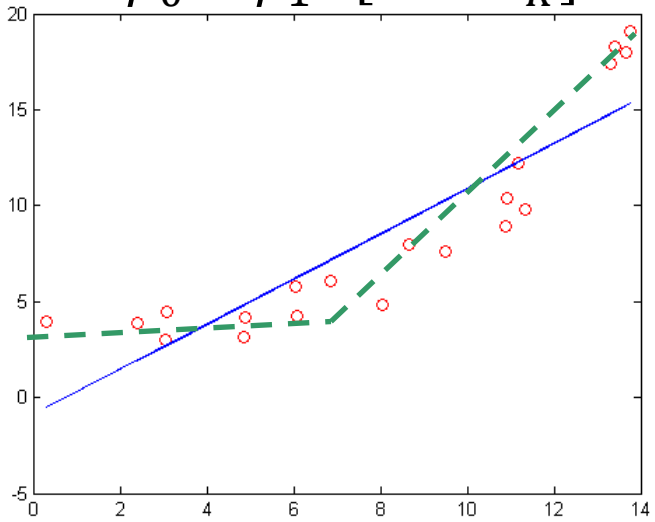
$$Log(Y) = \beta_0 + \beta_1 Log(X) + \varepsilon$$

- Piecewise linear regression:

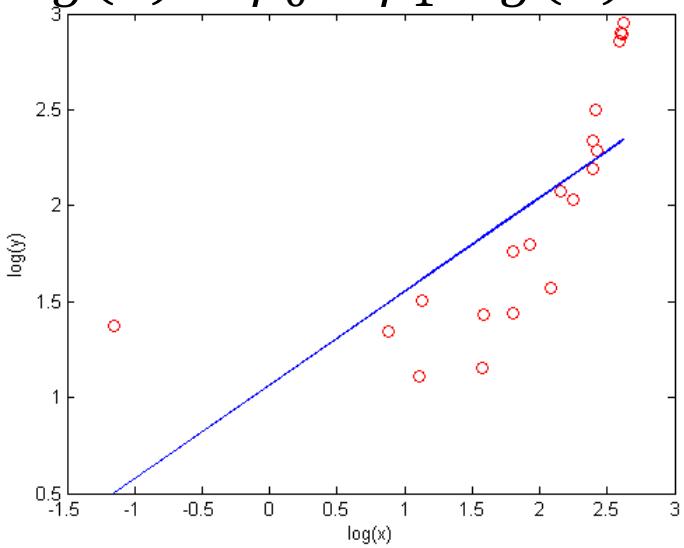
$$Y = \beta_0 + \beta_1 X[X > X_K] + \varepsilon$$

where $[X > X_K] = 0$ if $X \leq X_K$ and $[X > X_K] = 1$ if $X > X_K$

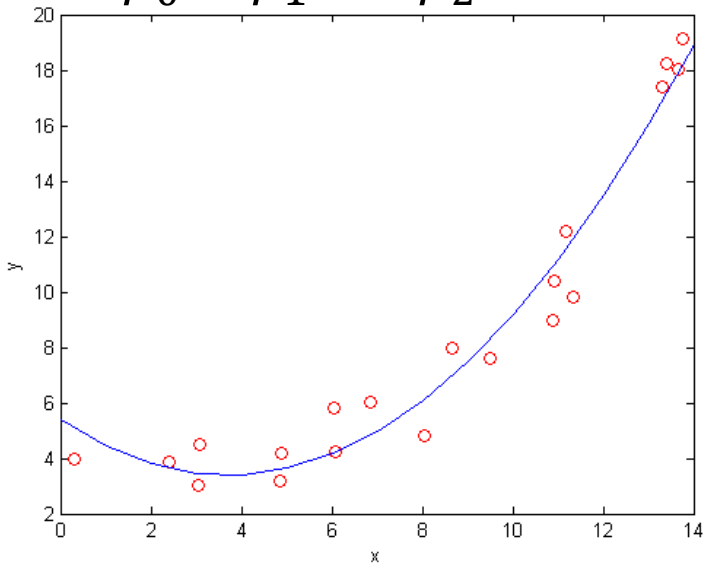
$$Y = \beta_0 + \beta_1 X[X > X_K] + \varepsilon$$



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



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



Outliers

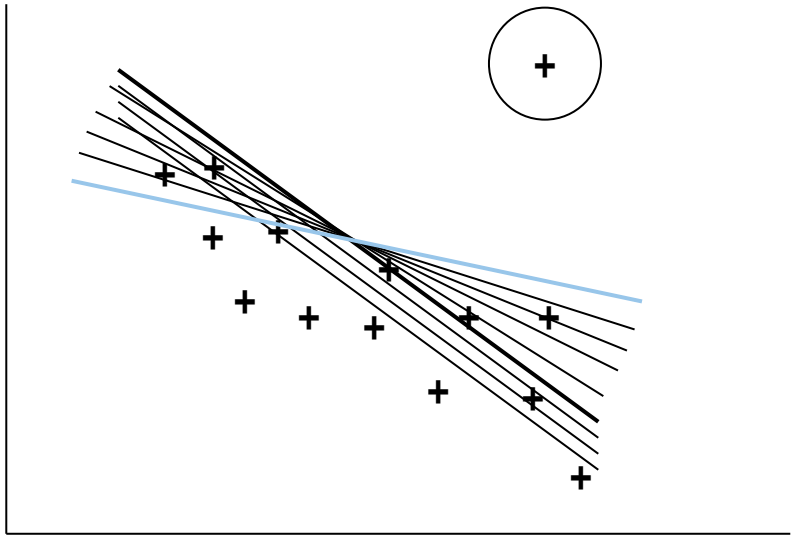
- An outlier is an observation that is unusually small or large

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 ($cov(\varepsilon, X) = 0$))

Multicollinearity

- The rank of the data matrix $\mathbf{X}$ is p, the number of columns
- $p < n$, the number of observations
- If there is no exact linear relationships among independent variables

  $rank(\mathbf{X}) = p$, $\mathbf{X}$ has full column rank
- If $rank(\mathbf{X}) < p$, $\mathbf{X}$ is singular -> impossible to calculate the inverse
  - *Remember:* $\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$

- 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*)

Check - 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

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

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
X1            1.5511     0.7448   2.083   0.0708
X2            0.5102     0.7238   0.705   0.5009
X3            0.1019     0.7547   0.135   0.8959
X4           -0.1441     0.7091  -0.203   0.8441
Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-Squared: 0.9824,     Adjusted R-squared: 0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

Large p-values

Big R-squared

```
> round(cor(cement.df),2)
      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

```
> vif(fit)
       X1        X2        X3        X4
 38.49621 254.42317  46.86839 282.51286
```

Drop X4

```
> vif(fit)
      X1       X2       X3
3.251068 1.063575 3.142125
```

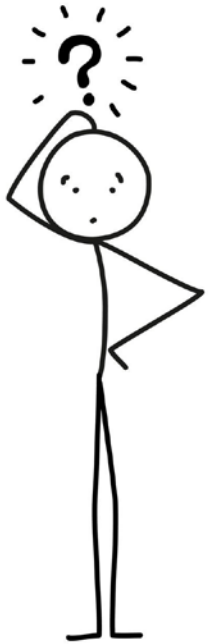
VIF's now small

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.19363    3.91330  12.315 6.17e-07 ***
X1           1.69589    0.20458   8.290 1.66e-05 ***
X2           0.65691    0.04423  14.851 1.23e-07 ***
X3           0.25002    0.18471   1.354    0.209

Residual standard error: 2.312 o      s of freedom
Multiple R-Squared: 0.9823,           squared: 0.9764
F-statistic: 166.3 on 3 and 9 D        3.367e-08
```

X1, X2 now signif

R-squared hardly decreased

Please explain the intuition behind the VIF.



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**

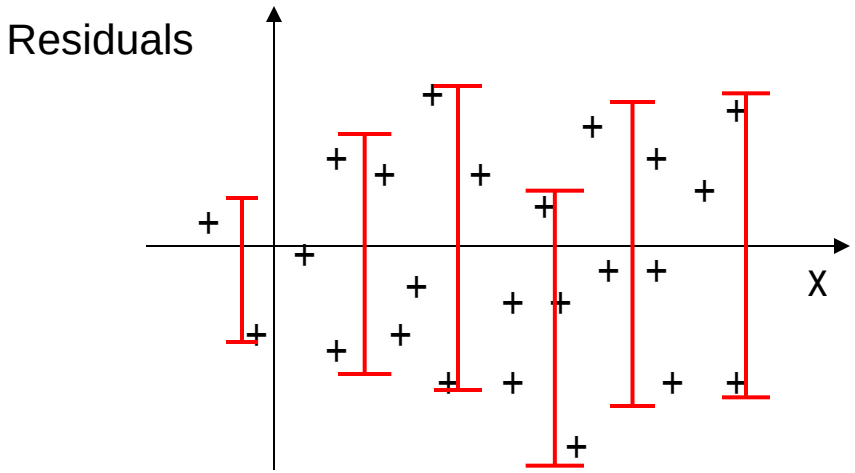
   There is no correlation between the i^{th} and j^{th} residual terms

5) **Expected value** of the residual vector, given X, is 0 ($E(\varepsilon|X) = 0$)

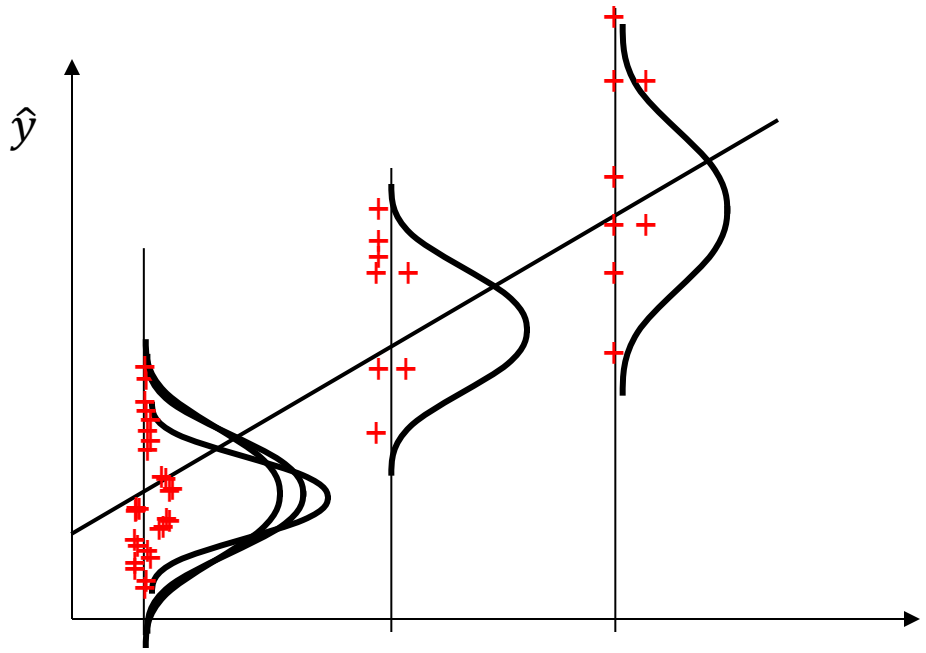
   (i.e. exogeneity ($cov(\varepsilon, X) = 0$))

Homoscedasticity

- When the requirement of a constant variance is not violated we have homoscedasticity

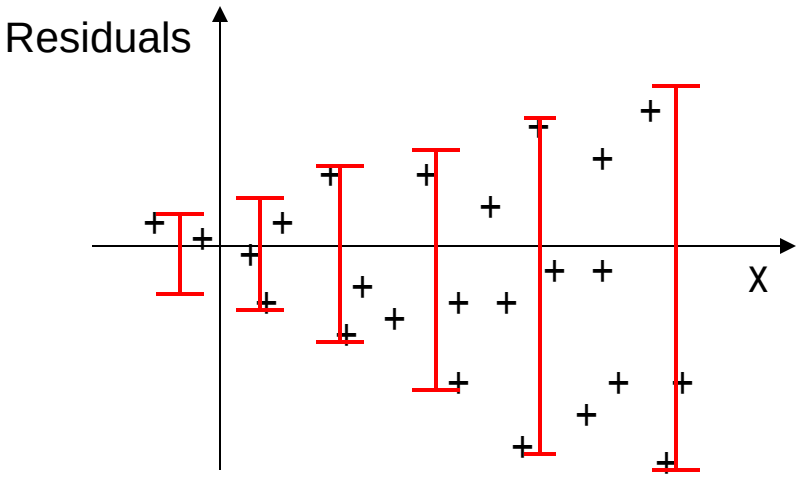


The spread of the data points does not change much.

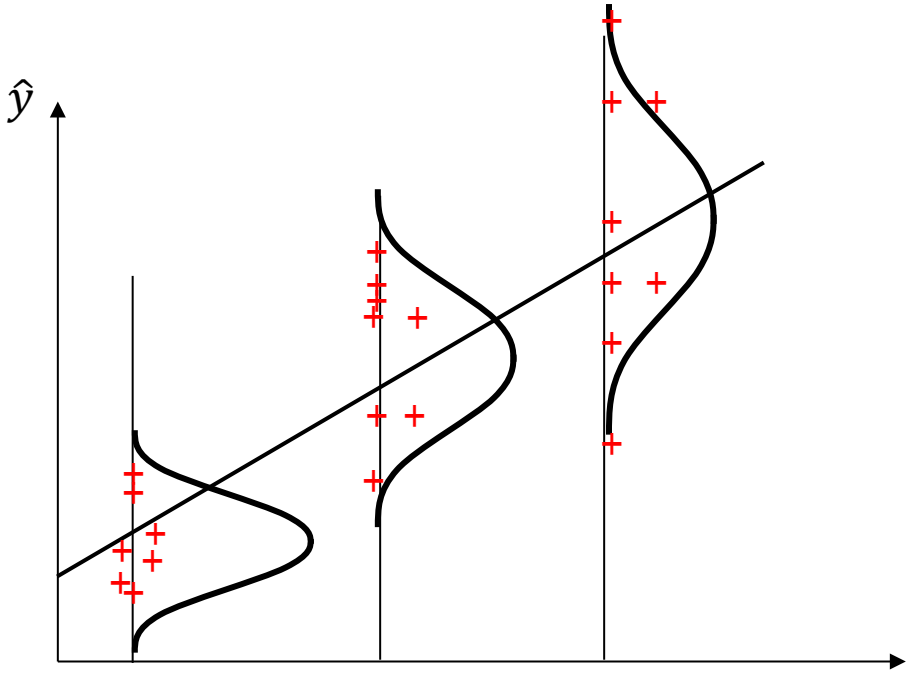


Heteroscedasticity

- When the requirement of a constant variance is violated we have heteroscedasticity ($var(\varepsilon_i|x_{1i}, \dots, x_{pi})$ not constant)
- Breusch-Pagan test or White test are used to check for heteroscedasticity



The spread increases with x



White Test

- The test is 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

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

  or only $e^2 = \beta_0 + \beta_1 \hat{y} + \beta_2 \hat{y}^2 + v$ to have less degrees of freedom

- $\chi^2 = nR^2$, where n is the sample size and R^2 is the unadjusted R-squared from the auxiliary OLS regression.
  - The statistic χ^2 has an asymptotic chi-square (χ^2) distribution with $d.f. = p$, 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 $\chi^2 > {\chi_{cr}}^2$
- If there is heteroscedasticity, the estimated $Var(\hat{\beta})$ is biased and OLS might not be efficient anymore.

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 ($cov(\varepsilon, X) = 0$))

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		

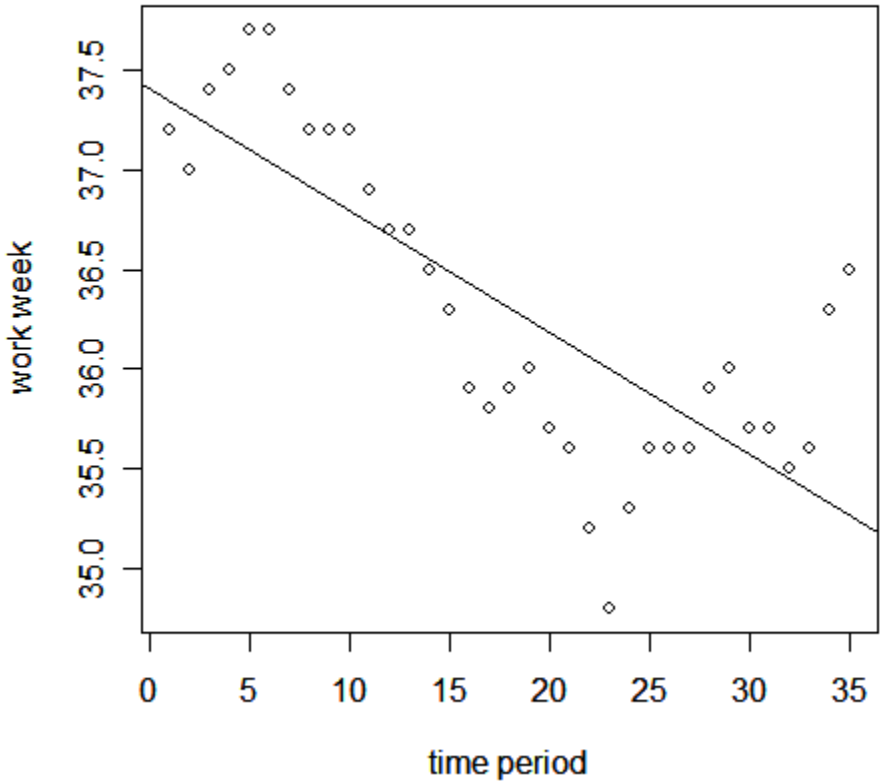
Forecasting Linear Trend using the Multiple Regression

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

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.20297 -0.28361  0.04711  0.30798  1.23048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.40874    0.17502 213.744  < 2e-16 ***
x           -0.06112    0.00848  -7.208  2.9e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5067 on 33 degrees of freedom
Multiple R-Squared: 0.6116,     Adjusted R-squared: 0.5998
F-statistic: 51.95 on 1 and 33 DF,  p-value: 2.901e-08
```

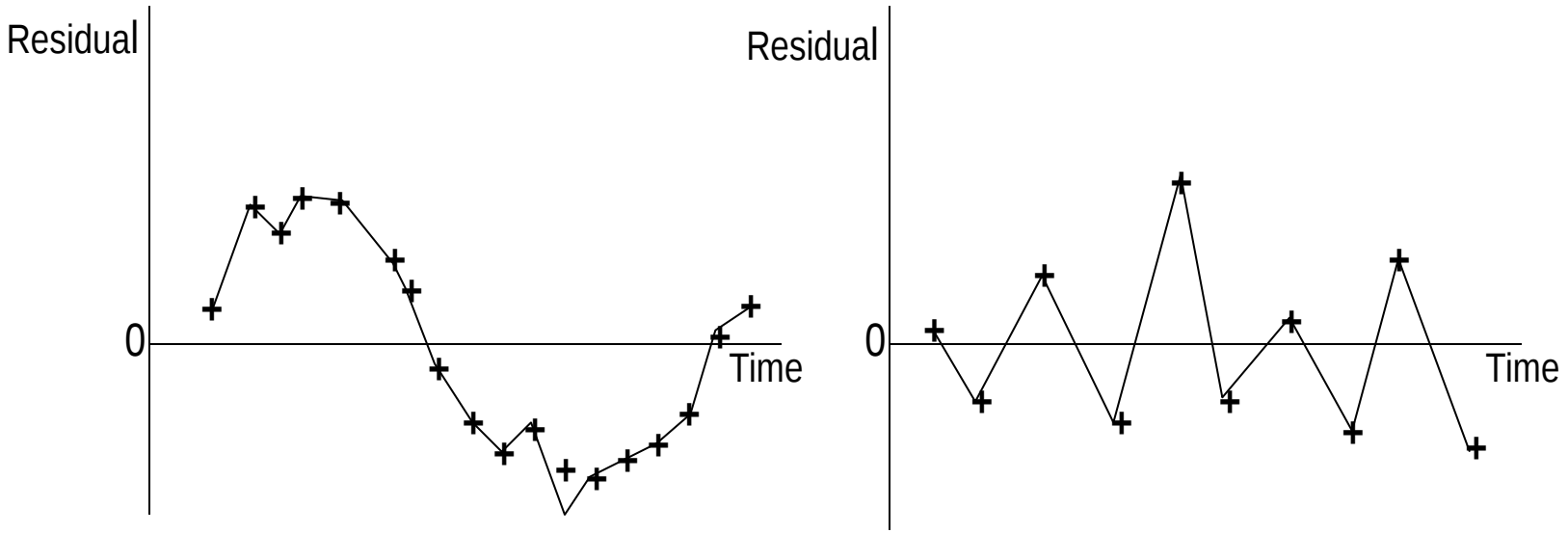
Probability that the null hypothesis ($\beta = 0$) is true

Hours Worked Data - A Linear Trend Line



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

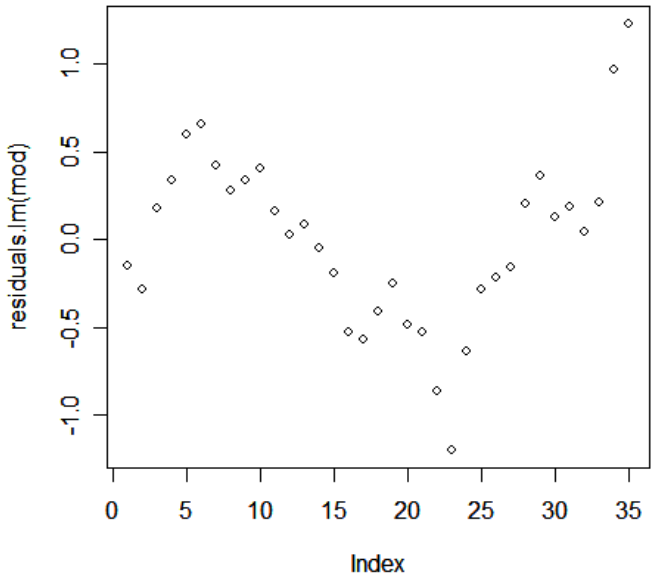
- Reasons leading to autocorrelation:
  - Omitted an important variable
  - Functional missfit
  - Measurement error in independent variable

- 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
- $DW = 4$ – perfect negative autocorrelation

Test for Autocorrelation in R




```
> plot(residuals.lm(mod))
> library(car)
> durbin.watson(mod)

lag Autocorrelation  D-W Statistic p-value
   1       0.7705505     0.2775895       0
 Alternative hypothesis: rho != 0
```

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 ($cov(\varepsilon, X) = 0$))

Endogeneity due to Omitted Variables

- Endogeneity means ($corr(\varepsilon_i, X_i) \neq 0$) => $E(\varepsilon_i|X_i) \neq 0$
- 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 x_2 are correlated and x_2 affects y, this leads to endogeneity.

Why is this a problem?

Omitted Variables

- 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?

Omitted Variables

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%

School of 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%

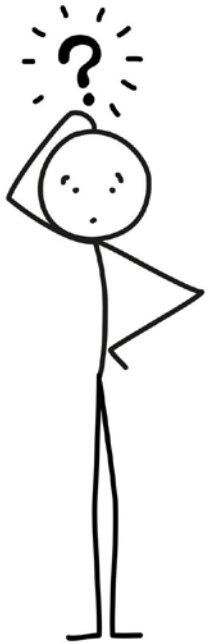
Explanations?

- 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 ...?

Check out the Wikipedia entry for Simpson‘s paradoxon!



Omitted Variable Bias

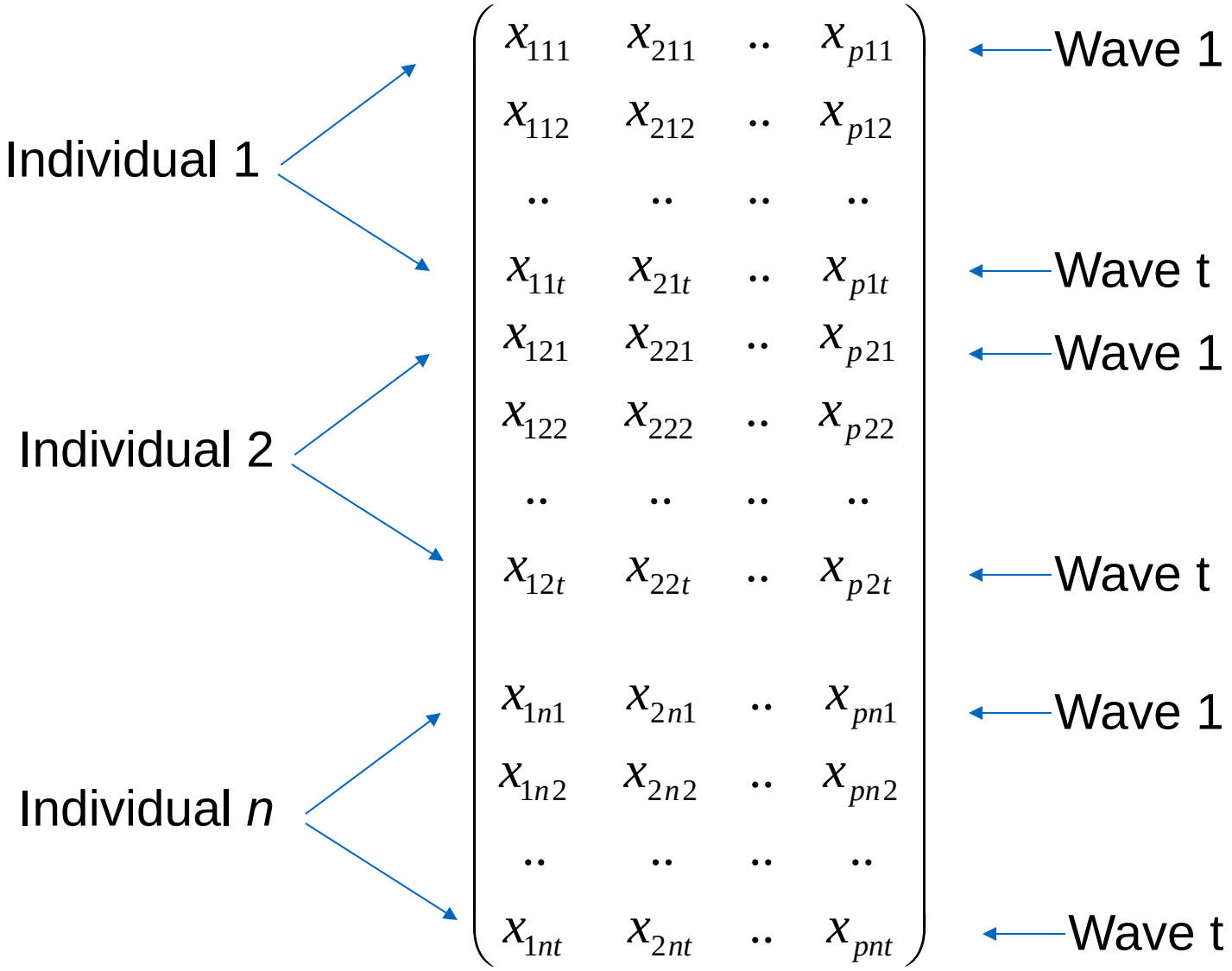
- **Endogeneity** is given when an independent variable is correlated with the error term and the covariance is not null
  - A reason for endogeneity might be that relevant variables are omitted from the model
  - For example, enthusiasm or willingness to take risks of an individual describe unobserved heterogeneity

- Various techniques have been developed to address endogeneity in panel data
  - 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 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
  - There might be omitted variables (aka. confounder) describing important characteristics of individuals
- A panel data set, or longitudinal data set, is one where there are repeated observations on the same units
  - They may make it possible to overcome a problem of an omitted variable bias caused by unobserved heterogeneity
  - 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

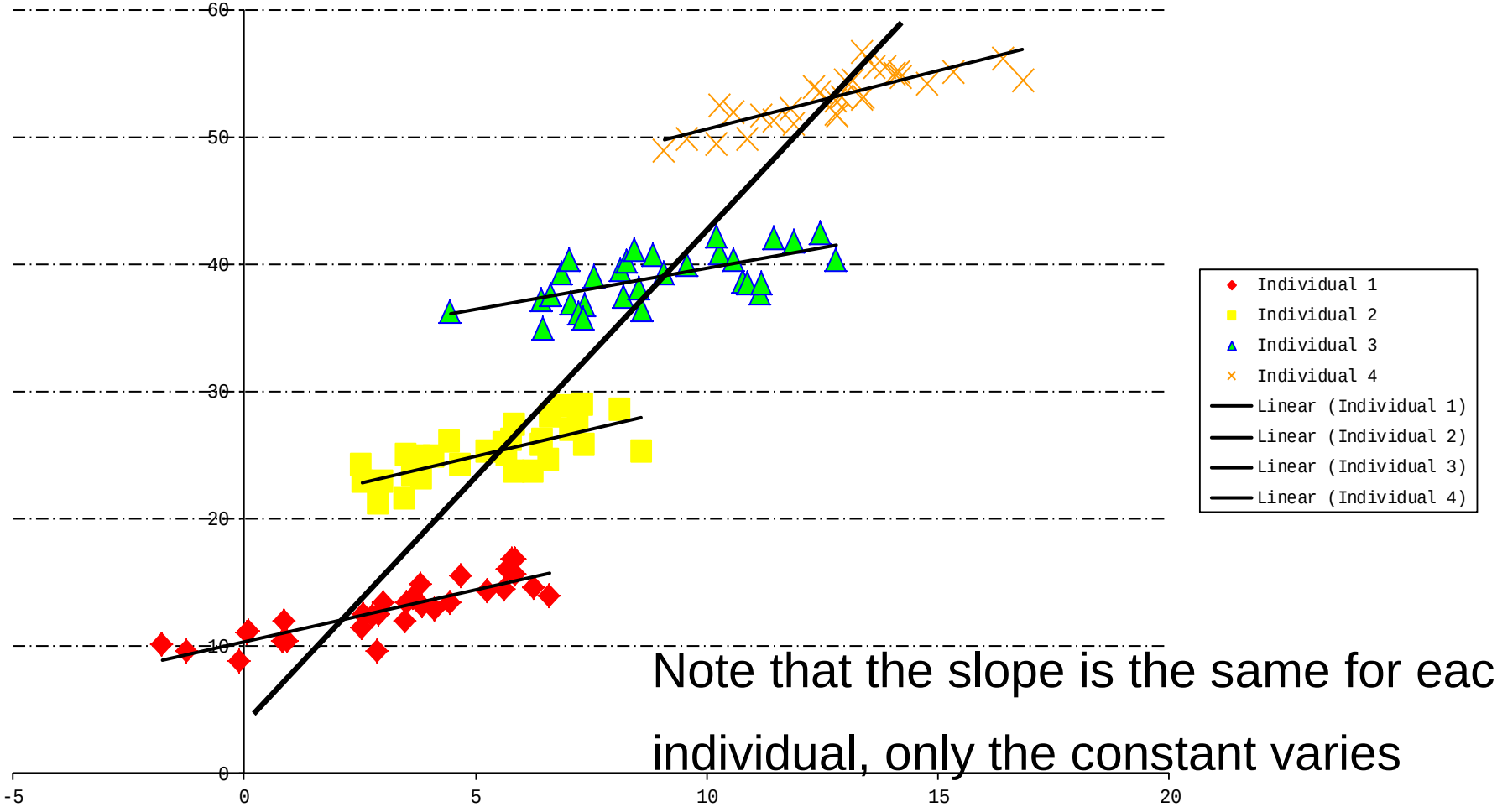


Treatment of Individual Effects

- There are two main options for the treatment of individual effects in panel data:
  - **Fixed effects** – assume λ_i are constants (there is **endogeneity**)
    - Individual-specific effects are **correlated to the other covariates**
    - see also https://en.wikipedia.org/wiki/Fixed_effects_model
  - **Random effects** – assume λ_i are drawn independently from some probability distribution
    - Individual-specific effects are **uncorrelated to the other covariates**
    - see also https://en.wikipedia.org/wiki/Random_effects_model
  - The Hausman test can help decide on one or the other

- Specific packages in R are available for fixed, random, and mixed effects models, which combine both (e.g., plm)

Fixed Effects Models

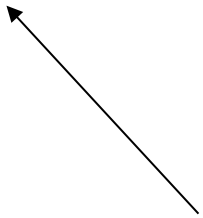


Note that the slope is the same for each individual, only the constant varies

The Fixed Effect Model

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

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



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

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

Lag one period and subtract

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

Constant and individual effects eliminated

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

Transformed equation

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

Alternative Estimators

- Within estimator (for more than two periods)
  - 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

- Least squares dummy variable estimator
  - uses a dummy variable for each individual

Random Effects Model

- The **fixed effect assumption** is that the individual specific effect is correlated with the independent variables ($cov(\lambda_i, x_{jit}) \neq 0$).

- The **random effects assumption** (in a random effects model) is that the individual specific effects are *uncorrelated* with the independent variables ($cov(\lambda_i, x_{jit}) = 0$, but λ_i might be correlated).

Original equation

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

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

λ_i is part of error term in random effects models

The Hausman Test

- Test of whether the fixed effects (FE) or random effects (RE) model is appropriate, i.e. a test for endogeneity.
- The test takes into account the covariance matrix of the FE and RE estimators as well as the estimates and follows a chi-square (χ^2) distribution.
- Null hypothesis: no correlation between $\boldsymbol{X}$ and the residuals.

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 ($cov(\varepsilon, X) = 0$))