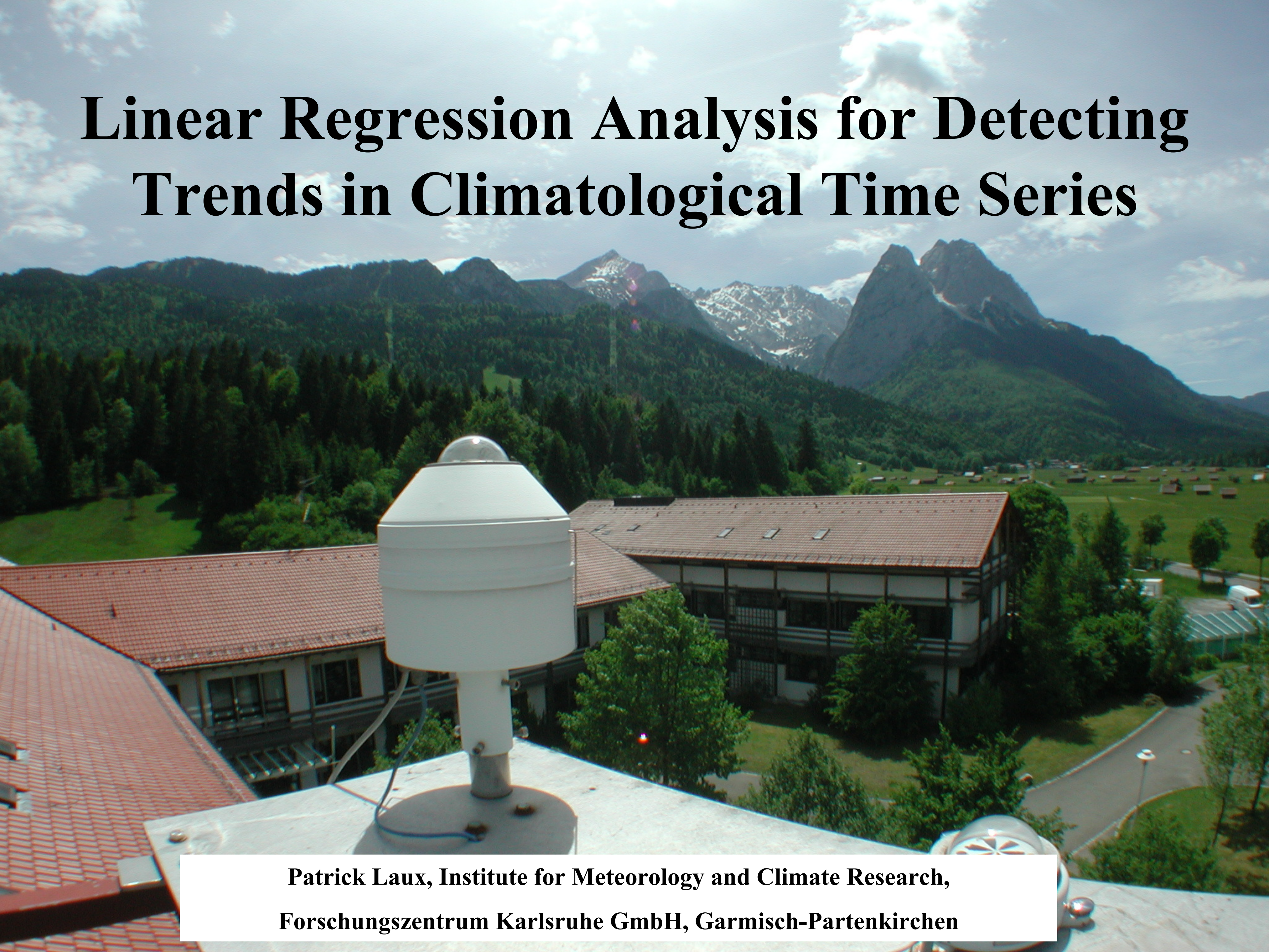


Linear Regression Analysis for Detecting Trends in Climatological Time Series

Patrick Laux, Institute for Meteorology and Climate Research,

Forschungszentrum Karlsruhe GmbH, Garmisch-Partenkirchen

Contents

1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB

Aim: To introduce the statistical concept of **1) linear regression for trend analysis** and show how it can be used to model the response of a variable to changes in an explanatory variable, **2) theory of statistical significance tests**.

Practical exercises: Exercises will follow using rainfall observation data from the Volta Basin (West Africa).

Prerequisistes: Minimal statistical knowledge, but some basic mathematics and computer skills.

1. Introduction

Linear Regression Analysis (LRA):

- Modeling functional relationship of two or more variables
- Correlation coefficient just quantifies the magnitude and the direction of the relationship, **not** the functional relationship!

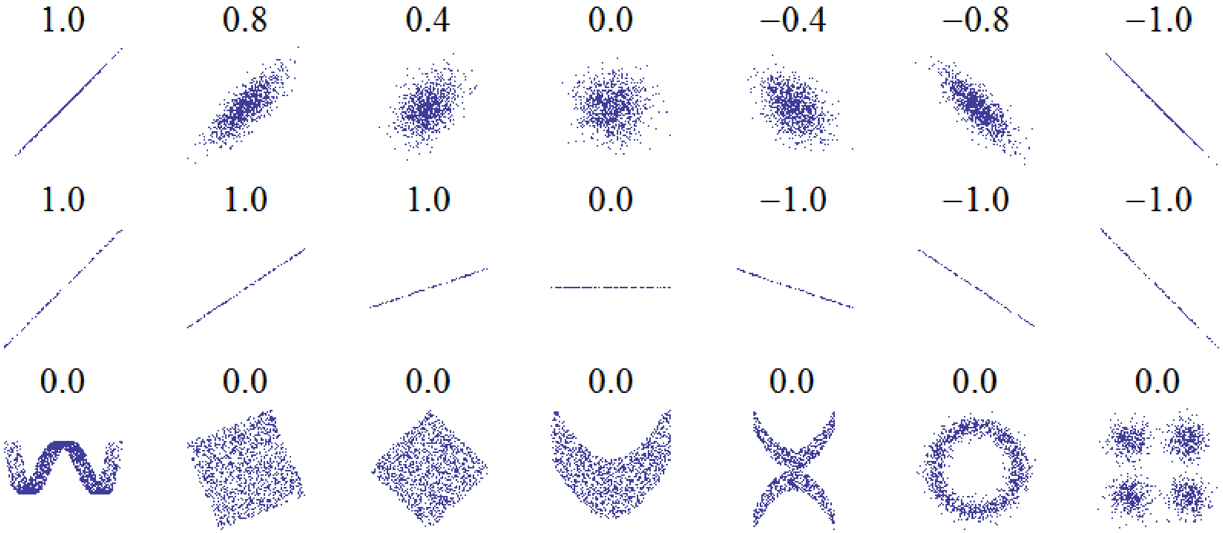
Example

Hypothesis: Body height (x_i) is an important factor in determining the body weight (y_i).

0.9419 → strong positive relationship!

i	Weight	Height
1	64	1.62
2	74	1.83
3	86	1.91
4	69	1.76
5	84	1.88
6	69	1.69
7	74	1.78

Correlation Coefficient

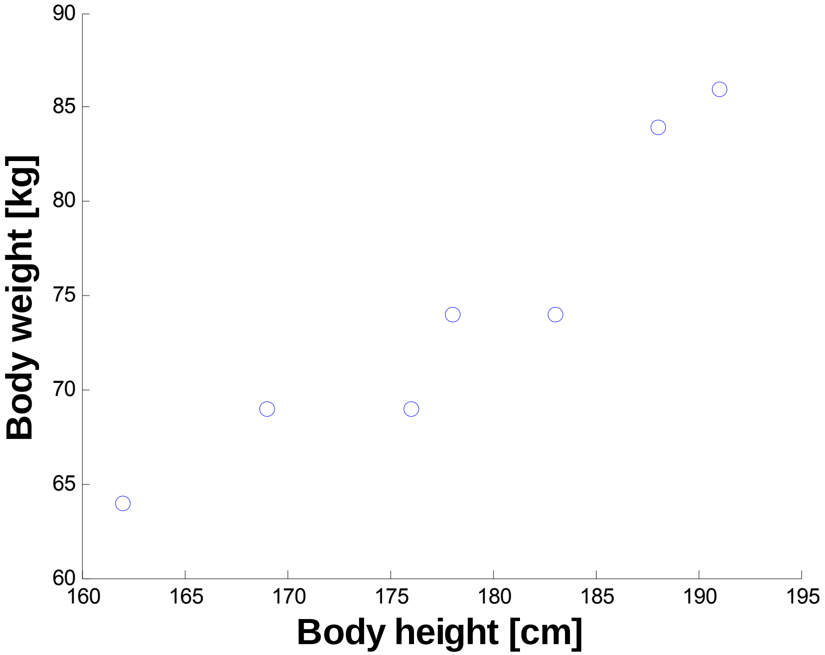


Correlation Coefficient (CC) reflects the **noisiness** and **direction** of a linear relationship (top row), but not the **slope** of that relationship (middle), nor many aspects of **nonlinear** relationships (bottom).

1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

Regression = Functional relationship

Body weight = -57.8 + 0.741 Body height



1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

1. Introduction

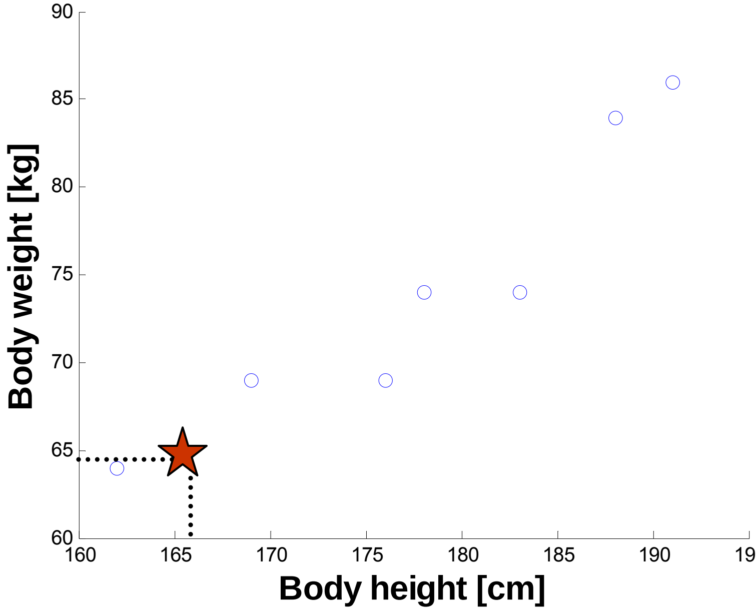
Linear Regression Analysis (LRA):

- Modeling functional relationship of two or more variables
- Correlation coefficient just quantifies the magnitude and the direction of the relationship!
- Predicting values, which are not measured

Prediction

What is the body weight of a person with a height of 165cm?

Body weight = - 57.8 + 0.741 * 165 = 64.5



1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

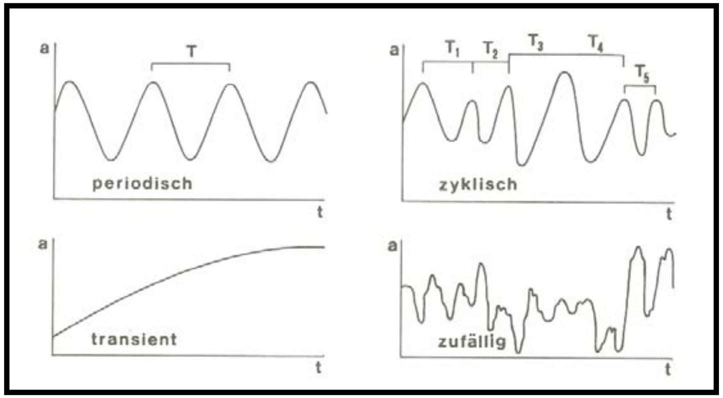
1. Introduction

Linear Regression Analysis (LRA):

- Modeling functional relationship of two or more variables
- Correlation coefficient just quantifies the magnitude and the direction of the relationship!
- Predicting values, which are not measured
- **Assessing linear trends in time series**

Time Series

- A **time series** is a sequence of data points, measured typically at successive times, spaced at (often uniform: e.g. δt = 1d) time intervals.
- Superposition of four components:
  1. Seasonal (periodical)
  2. Cyclical
  3. Transient (Trend)
  4. Stochastic



Transient component can often be described as a linear function (linear regression)!

Modeling Steps - General

1. Model identification

- Descriptive statistics (mean, variance, etc)
- Plotting the data (scatter plot)

Modeling Steps - General

2. Model estimation

- Fitting the model to the sample data
- Estimating the confidence intervals

Modeling Steps - General

3. Model validation

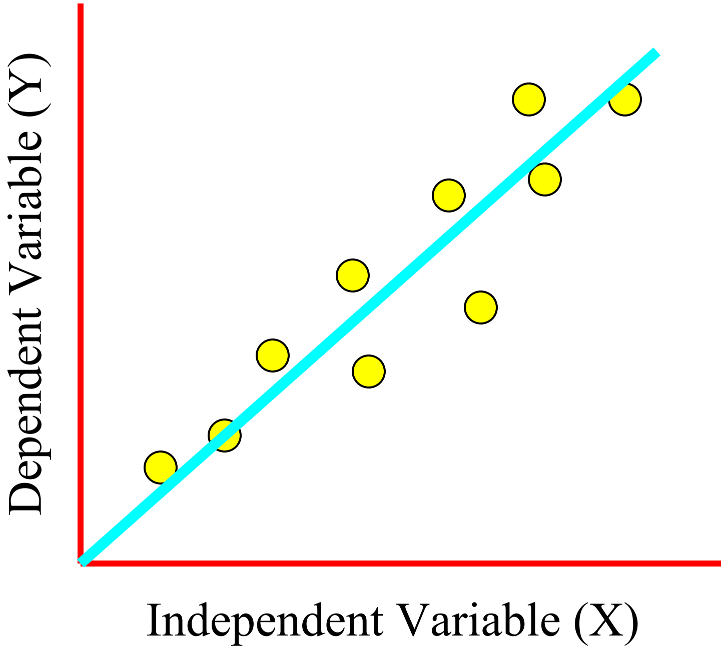
- The model fit is critically assessed by carefully analysing the residuals (errors) of the fit
- Further diagnostics

Modeling Steps - General

4. Application (Prediction)

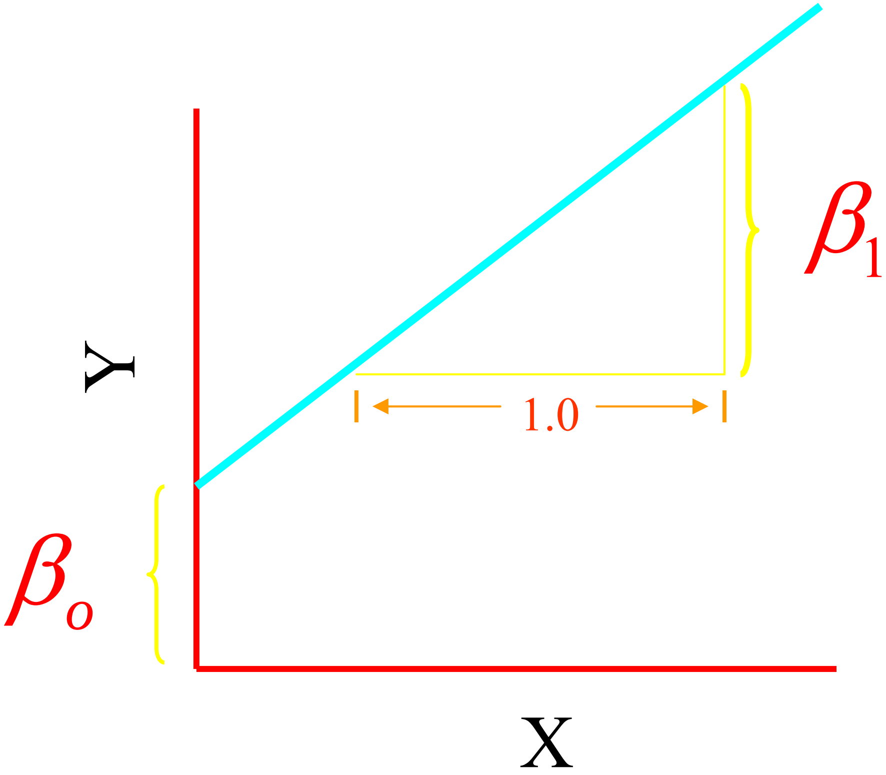
- The model is used to make predictions in new situations
- Ultimate test for any model (real skill)

2. Brief mathematical description of LRA



Linear regression describes the <u>linear</u> relationship between a predictor variable, plotted on the *x*-axis, and a response variable, plotted on the *y*-axis

1. Introduction
2. **Brief mathematical description of LRA**
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!



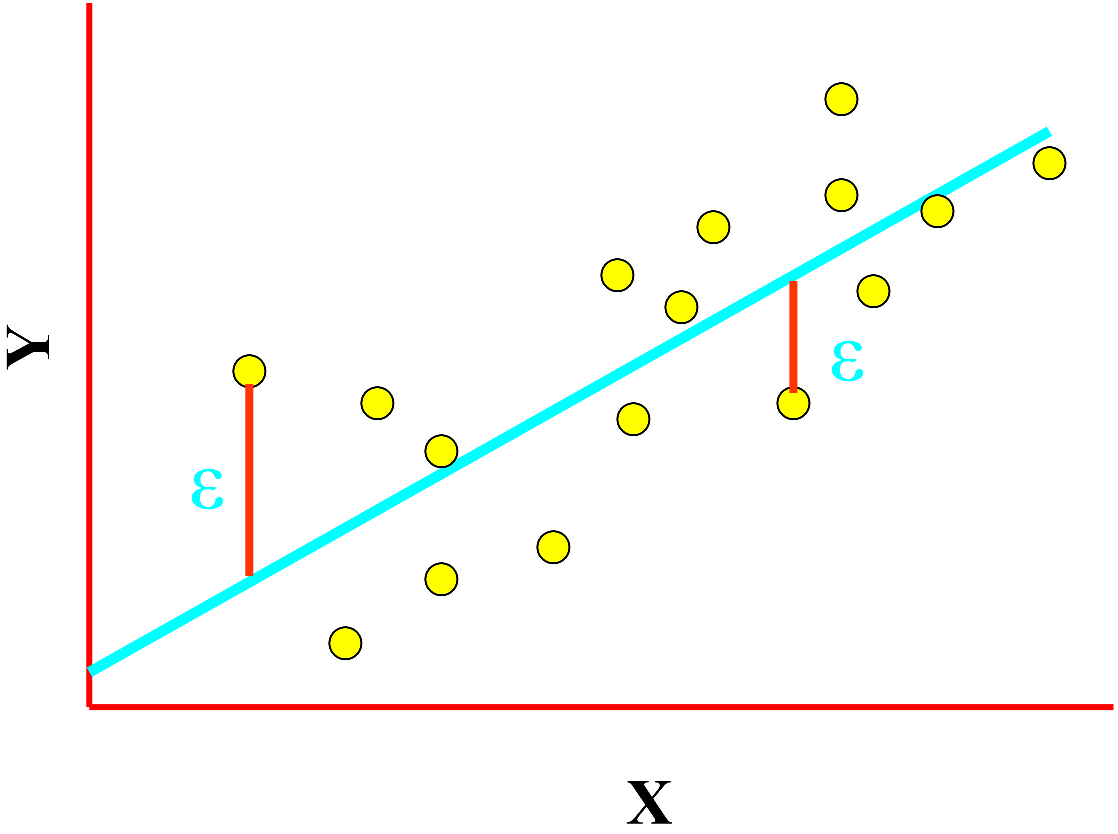
$$Y = \beta_o + \beta_1 X$$

β_0 – Intercept

β_1 - Slope

1. Introduction
2. **Brief mathematical description of LRA**
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

Point estimation



$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

1. Introduction
2. **Brief mathematical description of LRA**
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

How to built a linear model that fits
to the measured data?

The Ordinary Least Square Method (OLS)

Ordinary Least Squares (OLS) Regression

Model line: $\hat{Y} = \beta_0 + \beta_1 X$

Residual (ε) = $Y - \hat{Y}$

Sum of squares of residuals = $\sum (Y - \hat{Y})^2$

We must find values of β_o and β_1 that minimise

$$\min \sum (Y - \hat{Y})^2$$

1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

Least squares estimators for β1 and β0

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

1. Introduction
2. **Brief mathematical description of LRA**
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

Regression Statistics

$$\overline{X} = \frac{\sum X}{n}$$

$$\overline{Y} = \frac{\sum Y}{n}$$

n = number of observations

Descriptive Statistics

$$\mathrm{Var}(X) = \frac{\boxed{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}{n-1} \longrightarrow S_{xx}$$

$$\mathrm{Var}(Y) = \frac{\boxed{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}{n-1} \longrightarrow S_{yy}\,(SST)$$

$$\mathrm{Covar}(X, Y) = \frac{\boxed{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}}{n-1} \longrightarrow S_{xy}$$

Regression Statistics

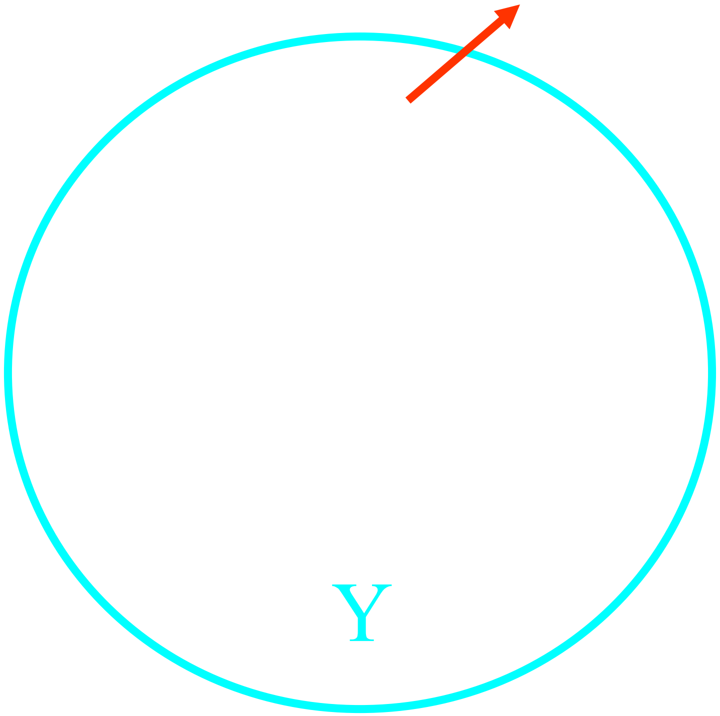
$$\mathrm{SST} = \sum (\mathrm{Y} - \overline{\mathrm{Y}})^2$$

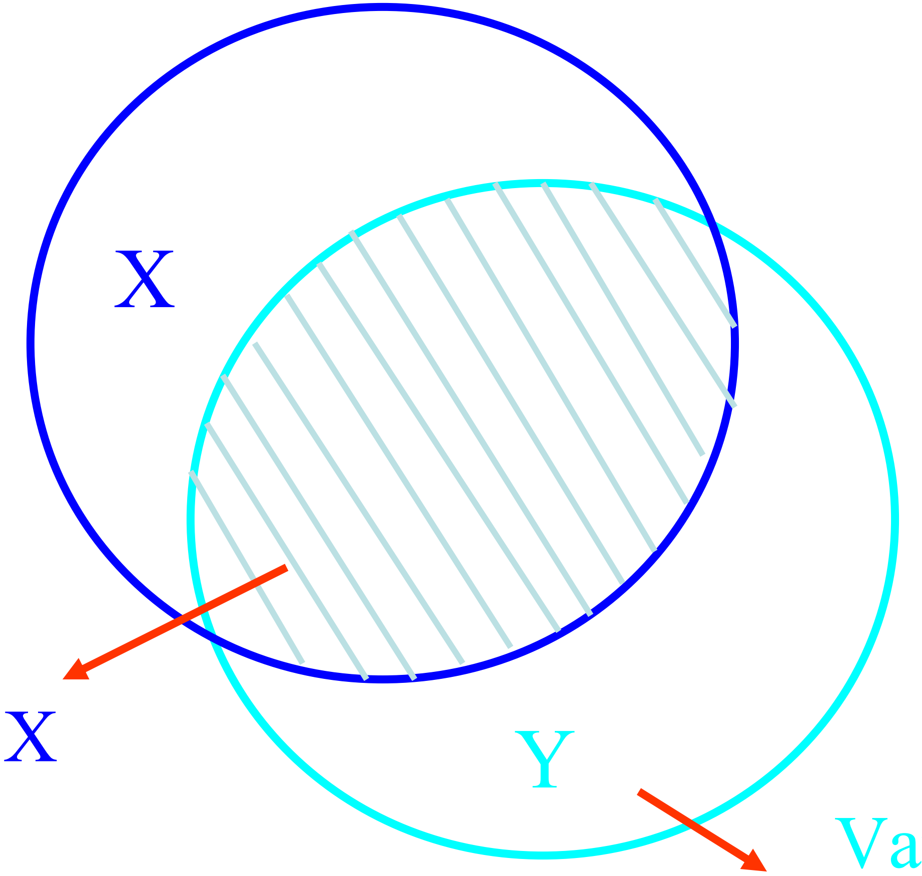
$$\mathrm{SSR} = \sum (\hat{\mathrm{Y}} - \overline{\mathrm{Y}})^2$$

$$\mathrm{SSE} = \sum (\mathrm{Y} - \hat{\mathrm{Y}})^2$$

1. Introduction
2. Brief mathematical description of LRA
3. **Regression statistics**
4. Inference statistics
5. Practical exercises using MATLAB!

Variance to be explained by predictors (SST)





Variance explained by X (SSR)

Variance NOT explained by X (SSE)

1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. Inference statistics
5. Practical exercises using MATLAB!

Regression Statistics

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$

1. Introduction
2. Brief mathematical description of LRA
3. **Regression statistics**
4. Inference statistics
5. Practical exercises using MATLAB!

Regression Statistics

$$R^2 = \frac{SSR}{SST}$$

Coefficient of Determination

to judge the adequacy of the regression model

1. Introduction
2. Brief mathematical description of LRA
3. **Regression statistics**
4. Inference statistics
5. Practical exercises using MATLAB!

Regression Statistics

$$R = \sqrt{R^2}$$

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Correlation Coefficient

Regression Statistics

$$S_e^2 = \frac{\sum (Y - \hat{Y})^2}{n - 2}$$

$$S_e = \sqrt{S_e^2}$$

Standard Error *for the regression model*

Confidence Interval on Regression Coefficients

$$b_1 - t_{\alpha/2,(n-2)}\sqrt{\frac{S_e^2}{S_{xx}}} \le \beta_1 \le b_1 + t_{\alpha/2,(n-2)}\sqrt{\frac{S_e^2}{S_{xx}}}$$

SE

Confidence Interval for the slope β_1

Confidence Interval on Regression Coefficients

$$b_0 - t_{\alpha/2,(n-2)} \underbrace{\sqrt{S_e^2\left(\frac{1}{n}+\frac{\overline{X}^2}{S_{xx}}\right)}}_{\text{SE}} \le \beta_0 \le b_0 + t_{\alpha/2,(n-2)} \underbrace{\sqrt{S_e^2\left(\frac{1}{n}+\frac{\overline{X}^2}{S_{xx}}\right)}}_{\text{SE}}$$

Confidence Interval for the intercept β_0

Example - Confidence Interval

1. Identify a sample statistic, e.g. the regression slope b1 calculated from sample data.

2. Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels, but any percentage can be used.

3. Calculate the margin of error for b1, use a t score for the critical value, with degrees of freedom (DF) equal to $n - 2$:

Margin of Error

- Compute alpha (α): $\alpha = 1 - (\text{confidence level} / 100) = 1 - 99/100 = 0.01$
- Find the critical probability (p^*): $p^* = 1 - \alpha/2 = 1 - 0.01/2 = 0.995$
- Find the degrees of freedom (df): df = $n - 2 = 7 - 2 = 5$.
- The critical value is the t score having 5 degrees of freedom and a cumulative probability equal to 0.995. From the t Distribution (tabulated values), we find that the critical value is 4.032.
- ME = critical value * SE = 4.032 * 0.12 = 0.477

Steps

4. Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* $\underline{+}$ ME. And the uncertainty is denoted by the confidence level.

b_1= 0.741

ME = 0.477

We are 99% confident that the true slope of the regression line is in the range $0.2645 \leq \beta_1 \leq 1.2184$.

Inferential Statistics

Comprises the use of statistics to make inferences from the sample data to unknown aspects of the population (general condition)!

Statistical Hypothesis Testing:

- make a *null hypothesis* (H_0) and *an alternative hypothesis* (H_1) *and* set a *significance level*. This is the (low) probability α at which we will reject H_0
- calculate the statistic and its degrees of freedom
- look up its predicted value in statistical distribution tables
- if the observed statistic is larger than the tabulated value we reject H_0

Hypotheses Tests for Regression Coefficients

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

$$t_{emp(n-2)} = \frac{b_1 - \beta_1}{S_e(b_1)} = \frac{b_1 - \beta_1}{\sqrt{\frac{S_e^2}{S_{xx}}}}$$

1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. **Inference statistics**
5. Practical exercises using MATLAB!

Hypothesis Tests on Regression Coefficients

$$H_0 : \beta_0 = 0$$

$$H_A : \beta_0 \neq 0$$

$$t_{emp(n-2)} = \frac{b_0 - \beta_0}{S_e(b_0)} = \frac{b_0 - \beta_0}{\sqrt{S_e^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right)}}$$

1. Introduction
2. Brief mathematical description of LRA
3. Regression statistics
4. **Inference statistics**
5. Practical exercises using MATLAB!

Hypotheses Test on the CC

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

$$t_{emp} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

We would reject the null hypothesis if $\left|t_{emp}\right| > t_{\alpha/2,n-2}$

ANOVA

$$H_0 : \beta_1 = 0$$

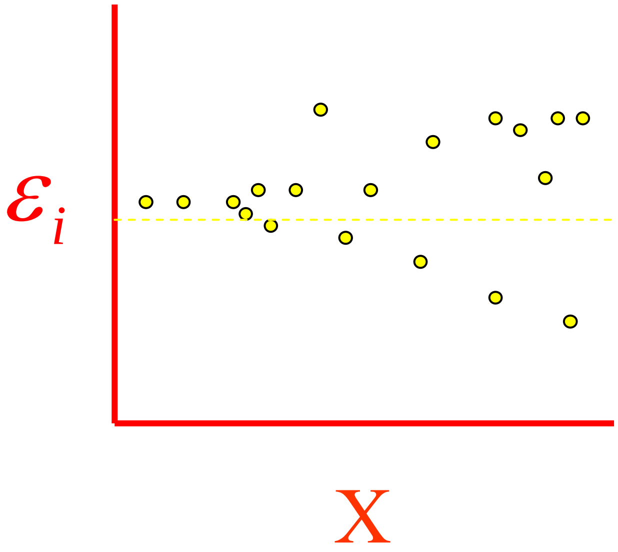
$$H_A : \beta_1 \neq 0$$

	df	Sum of Squares	Mean Squares	F_{emp}	P-value
Regression	1	SSR	SSR / df	MSR / MSE	$P(F_{emp})$
Residual	n-2	SSE	SSE / df		
Total	n-1	SST			
If $P(F_{emp})<\alpha$ then we know that we get significantly better prediction of Y from the regression model than by just predicting mean of Y.					

Assumptions of LRA

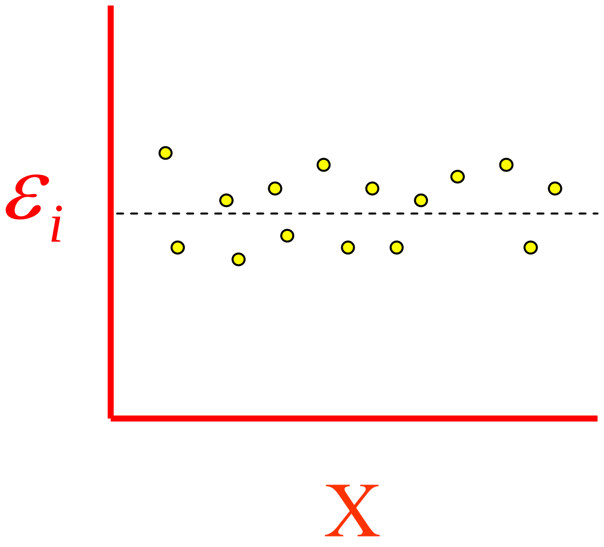
1. **Homoscedasticity** – the variance of the error terms is constant for each x_i, To check this, look at the plot(s) of the residuals versus the X value.

Here: not fulfilled (increasing in variance)!



Assumptions of LRA

2. **Linearity** – the relationship between X and Y is linear. To check this, again look at the plot of the residuals versus the X value. You don’t want to see a clustering of positive residuals or a clustering of negative residuals.



Assumptions of LRA

3. **Normality of the residuals** – residuals follow a normal distribution. > Normal probability plot of standardized residuals, histogram of residuals.

4. **Independence of error terms** – successive residuals are not correlated. If they are correlated, it is known as autocorrelation. > Durbin-Watson statistics.

Assumptions

If any of these assumptions is violated (i.e., if there is **nonlinearity**, **serial correlation**, **heteroscedasticity**, and/or **non-normality**), then the predictions, confidence intervals, and relationship yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading.

Violations of linearity are extremely serious - if you fit a linear model to data which are nonlinearly related, your predictions are likely to be seriously in error, especially when you extrapolate beyond the range of the sample data.

Violations of independence are also very serious in *time series regression* models: serial correlation in the residuals means that there is room for improvement in the model, and extreme serial correlation is often a symptom of a badly misspecified model. Serial correlation is also sometimes a byproduct of a violation of the linearity assumption as in the case of a simple (i.e., straight) trend line fitted to data which are growing exponentially over time.

Violations of homoscedasticity make it difficult to estimate the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Heteroscedasticity may also have the effect of giving too much weight to small subset of the data when estimating coefficients.

Violations of normality compromise the estimation of coefficients and the calculation of confidence intervals. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of squared error, a few extreme observations can exert a disproportionate influence on parameter estimates. Calculation of confidence intervals and various signficance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

Thank You!