Lecture 5: Linear Models (Simple Linear Regression)

EMAT30007 Applied Statistics

Nikolai Bode

Department of Engineering Mathematics

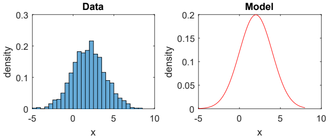
Lecture 5, Topic 1 – The simple Linear (regression) Model (sLM)

Content covered

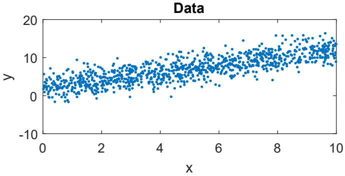
- Structure of sLMs
- Represenations of sLMs
- Fitting sLMs to data

Introduction to this part of the module

- . . . in the first half, we looked at statistical models with a *fixed mean*, e.g.:



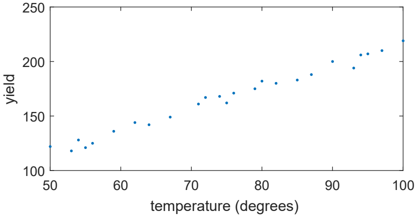
- In the second half, we start to look at models with a *varying mean*, e.g.:



- We still consider the same basic elements of statistical analysis:
  1. Quantify patterns in data.
  2. Test if patterns can be believed (i.e. they don't arise by chance).

Motivating example

- Example: suppose a chemical reaction produces higher yields of a product, the higher the ambient temperature :

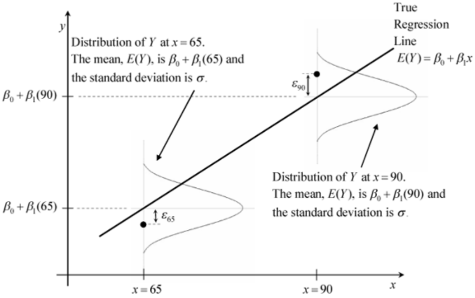


- *Simple linear regression* is what we can use to investigate if a relationship between two variables exists when we don't know about the underlying process.
- We model a linear relationship between two variables (straight line) and (1) quantify this pattern before (2) testing if we 'can believe it'.

Structure of simple linear regression models

- For n observed data pairs $\{(x_i, Y_i), i = 1, ..., n\}$, *simple linear regression* assumes we have the relationship: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ Linear combination
- Y is called the *response/ dependent variable* and x the *explanatory/ independent/ predictor variable*. β_0 and β_1 are *model parameters*.
- ϵ_i are error terms (spread around the regression line) and crucially, we assume $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma)$ in simple linear regression.

Errors follow a normal distribution



Alternative representations for simple linear regression models

- *Algebraic notation*: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ (as on the previous slide).
- *Matrix notation*: $Y = X\beta + \epsilon$, where

Vandermonde matrix

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- Because of the error term, ϵ, the response Y is a random variable. So, $Y_i = E(Y_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and thus $E(Y_i) = \beta_0 + \beta_1 x_i$. This leads to the following notation:

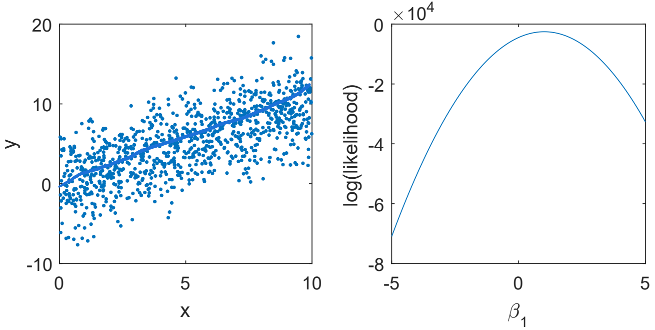
  deterministic

- *Random variable notation*: $Y_i \overset{i.i.d}{\sim} N(\beta_0 + \beta_1 x_i, \sigma)$.
- Since the mean of the response is a linear function of the explanatory variables, they're often called: '*simple linear models*' (sLMs).

Fitting simple LMs to data (quantify pattern)

- Given data, we want to estimate the parameters of the model.
- In this course, we always use *Maximum Likelihood Estimation* (MLE). That means, we find the parameter values that maximise the likelihood of our model. Analogy: maximise $P(data|parameters)$.
- The *likelihood* for an sLM is $L(\beta|X) = \prod_{i=1}^{n} f_N(y_i, \mu = \beta_0 + x_i\beta_1, \sigma)$, where $f_N(Y_i, \mu = \beta_0 + x_i\beta_1, \sigma)$ is the probability density function for the normal distribution with mean $\mu = \beta_0 + x_i\beta_1$ and standard deviation σ evaluated at Y_i.
- For LMs, it has been shown that MLE is equivalent to *Ordinary Least Squares* (OLS).
- For LMs exact equations for these parameter estimates exist (estimates denoted with '^'): $\hat{\beta} = (X'X)^{-1}X'Y$, $\hat{\sigma}^2 = \frac{1}{n-2}(Y - X\hat{\beta})'(Y - X\hat{\beta})$
- Fitted values for the response are: $\hat{Y} = X\hat{\beta}$

Likelihood function revisited



In $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, assuming we know $\beta_0 = 0$ and σ, we can see that the likelihood is maximal for $\beta_1 = 1$. Seems plausible given the data on the left.

Lecture 5, Topic 2 – Assumptions of sLMs

Content covered

- Assumptions of sLMs
- Residuals and residual plots

Assumptions of simple LMs

All statistical models make assumptions. Linear models are no exception. They assume, most importantly:

- *Linearitiy*: response variable is a linear combination of the explanatory variables (not as restrictive as it seems, see lecture 8).
- *Normality*: errors follow a normal distribution.
- *Constant error variance (homoscedasticity)*: variance of response variable (or errors) does not depend on the value of the explanatory variables. If this assumption is invalid it's called *heteroscedasticity*.
- *Independence*: the errors are uncorrelated (ideally statistically independent). This means that the response variable observations are *conditionally independent*. Need this for the product in the likelihood function.
- *Weak exogeneity*: the explanantory variables can be treated as fixed values, rather than random variables.

It is important to check that the most important assumptions hold (approximately) when fitting LMs to data (see below).

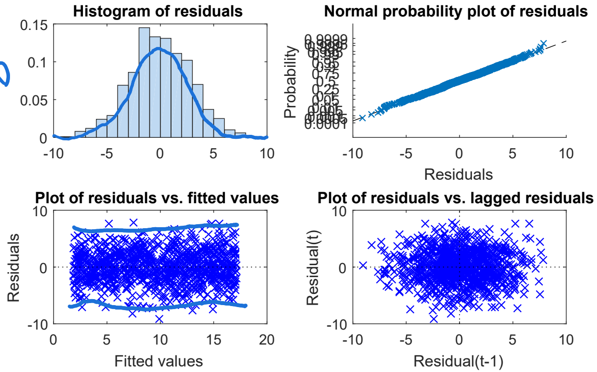
Checking sLM assumptions: residuals

To check if the model assumptions hold, we look at the errors, or *residuals*: $\hat{\epsilon} = Y - X\hat{\beta}$. Hypothesis testing on residuals is possible, but we will focus on residual plots for model checking. Perfect residual plots look like this:

Normal distribution with mean 0

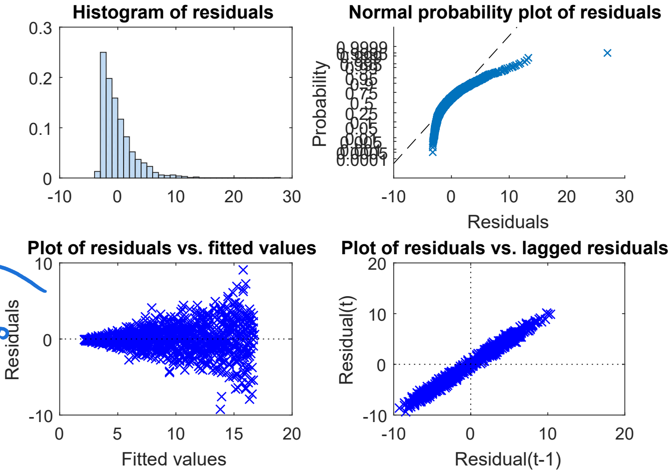
Quantile plot.

Fixed range at y values



Residual plots gone wrong

Note: these plots are not all from the same data.



Variance depends on fitted values

Lecture 5, Topic 3 – Inference using sLMs

Content covered

- Hypothesis tests and sLMs
- Confidence intervals and sLMs
- Goodness of fit for sLMs
- Estimation and prediction for sLMs

Hypothesis tests on sLM parameters ('can we believe the pattern?')

- How to test if the $x_i s$ contribute information for the prediction of Y in $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$?
- One way of doing this is to test the hypothesis that Y does not change as the explanatory x changes. In other words, we test the hypotheses:
$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$
- Fortunately, 'maths boffins' have found that $\hat{\beta}_1$ follows a normal distribution with mean β_1 and standard error $\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}} \approx \frac{s}{\sqrt{SS_{xx}}}$, where $SS_{xx} = \sum (x_i - mean(x))^2$ and $s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$
- Since σ is usually unknown, we need to use a Student's T-test on $T = \frac{\hat{\beta}_1 - hypothesised\ value}{s/\sqrt{SS_{xx}}}$ with degrees of freedom based on the number of data points and model parameters ($df = (n-2)$ for 2 parameters).
- IN PRACTICE, SOFTWARE DOES THIS FOR US!

Confidence intervals for sLM parameters

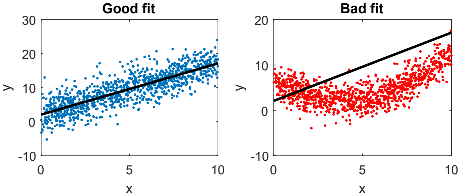
- As an alternative for inference on prameter estimates, we can also compute *Confidence Intervals*.
- The $(1-\alpha)100\%$ Confidence Interval for the gradient β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1} \text{ where } s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

  where $t_{\alpha/2}$ is based on $(n-2)$ degrees of freedom. SS_{xx} and s are defined on the previous slide. $t_{\alpha/2}$ is obtained from the Student's T-distribution in the usual way.
- Some statistical software provides these values by default, but often (e.g. in Matlab), only the standard errors for parameter estimates (see previous slide) are provided.

Goodness of fit for sLMs

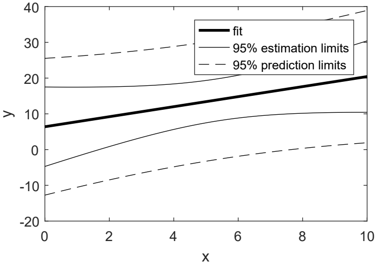
- It's good practice to always check the fit visually. Example:



- Looking at residual plots is also useful (e.g. fitted values versus residuals, or explanatory variables versus residuals).
- *Coefficient of Determination*, R^2 (`R-squared` in Matlab): $R^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$, where $SS_{yy} = \sum (Y_i - mean(Y))^2$ and $SSE = \sum (Y_i - \hat{Y}_i)^2$.
- R^2 can be interpreted as the proportion of the variance in the response variable that is explained by (or attributed to) the explanatory variable.
- There are other measures, similar to R^2 and there are a few issues making it problematic for assessing goodness of fit. We'll revisit this later in the course.

Estimation and prediction for sLMs

- *Estimation:* estimate mean value of Y over many data points.
- *Prediction:* predict Y for a particular value of x. This leads to higher error bounds (add error in mean to variation around mean).
- Example (we'll skip the calculation details):



Prediction limits wider than estimation limits

- **Warning:** careful with predictions far away from mean of explanatory variable or outside of region covered by data.

Typical steps in sLM analysis

1. Look at raw data (scatterplots).
2. Decide on model (e.g.intercept yes/no?).
3. Fit model using MLE.
4. Check model assumptions hold (residual plots).
5. Perform hypothesis tests on model parameters.
6. Interpret findings. Depending on use of model, look at goodness of fit, estimation, prediction... .

Matlab output for sLM analysis

Matlab does most of the work for us with one short command, e.g.:

```
>> fitlm(data, 'response~predictor')

ans =


Linear regression model:
    response ~ 1 + predictor

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     2.2859      0.18512    12.348     1.0648e-32
    predictor       1.4782     0.033151     44.59    9.2918e-240


Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 2.99
R-squared: 0.666,  Adjusted R-Squared 0.665
F-statistic vs. constant model: 1.99e+03, p-value = 9.29e-240
```

Slide: “Fitting simple LMs to data”

Slide: “Hypothesis tests on parameters”

tStat = Estimate/SE

pValue from Student’s T-distribution

Error s.d. estimate

Slide: “Goodness of fit for LMs”