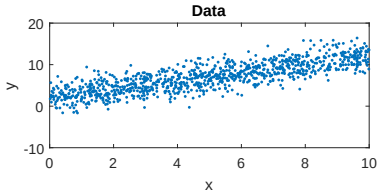
EMAT30007 Applied Statistics Lecture 7: Linear Models (Multiple Linear Regression)

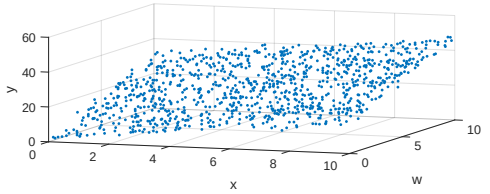
Ksenia Shalonova & Nikolai Bode

Linear models

- *Last lecture* — looked at simple linear regression ($Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$):



- *This lecture* — look at *multiple linear regression* (more than one predictor):



Structure of multiple linear regression models

- Good news: sLMs are a special case of multiple linear regression and with minor extensions everything we've looked at so far still applies.
- For p predictors and data tuples $\{(x_{1,i}, x_{2,i}, ..., x_{p,i}, Y_i), i = 1, ..., n\}$, we assume the relationship: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_p x_{p,i} + \epsilon_i$.
- As before, we assume $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma)$.
- The *matrix notation*, $Y = X\beta + \epsilon$, now becomes very convenient:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & \dots & x_{p,n} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- As before, the *random variable notation* is:
$Y_i \overset{i.i.d}{\sim} N(\beta_0 + \beta_1 x_{1,i} + ... + \beta_p x_{p,i}, \sigma)$. Or shorter, $Y \sim N(X\beta, \sigma I)$.

Fitting LMs to data (quantify pattern)

- Model fitting using *Maximum Likelihood Estimation* (MLE) proceeds in the same way as for sLMs seen in lecture 6 and is equivalent to *Ordinary Least Squares* (OLS) fitting.
- The *likelihood function* is given by:
$L(\beta|X) = \prod_{i=1}^{n} f_N(Y_i, \mu = \beta_0 + \beta_1 x_{1,i} + ...\beta_p x_{p,i}, \sigma)$, where f_N is the probability density function for the normal distribution evaluated at Y_i.
- The exact equations for parameter MLEs are (recall '^' notation for MLEs):
$\hat{\beta} = (X'X)^{-1}X'Y$,
$\hat{\sigma}^2 = \frac{1}{n - Number\ of\ \beta\ parameters}(Y - X\hat{\beta})'(Y - X\hat{\beta})$
- *Fitted values* for the response are: $\hat{Y} = X\hat{\beta}$
- Parameter estimates for one explanatory variable describe the relationship between this variable and the response variable when all other explanatory variables are held fixed.

Assumptions of LMs

The assumptions for multiple linear regression models are the same as for sLMs:

- *Linearitiy*: response variable is a linear combination of explanatory variables.
- *Normality*: errors follow a normal distribution.
- *Homoscedasticity*: variance of response variable (or errors) is constant.
- *Independence*: the errors are uncorrelated (ideally statistically independent).
- *Weak exogeneity*: the explanantory variables can be treated as fixed values, rather than random variables.

However, there is one important addition:

- *Lack of perfect multicollinearity*: if two explanatory variables are perfectly correlated, we cannot solve the equation for parameter estimates on the previous slides. Some (but not perfect!) correlation between explanatory variables may be permissible.

Hypothesis tests on LM parameters ('can we belive the pattern?')

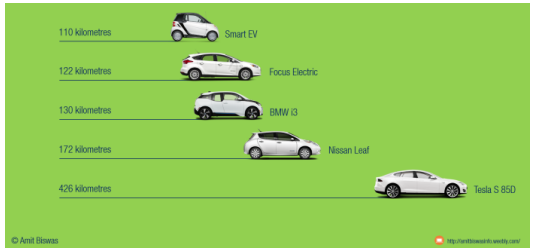
- Using a conceptually similar approach as for sLMs, we can test hypotheses about the βs and construct confidence intervals for them.
- Importantly, the test statistics depend on the entries of the matrix

$$(X'X)^{-1} = \begin{pmatrix} c_{00} & c_{01} & \dots & c_{0p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{p0} & c_{p1} & \dots & c_{pp} \end{pmatrix},$$

  as $\sigma_{\hat{\beta}_j} = \sigma\sqrt{c_{jj}}\ (j = 1, ..., p)$. Off-diagonal entries determine the covariance of estimates and are important for the variance in prediction.
- Generally, σ needs to be estimated (previous slide), so in practice we use Student's T-distributions and the test statistic: $T = \frac{\hat{\beta}_j - hypothesised\ value}{\hat{\sigma}\sqrt{c_{jj}}}$.
- MANY DETAILS OMITTED — SOFTWARE DOES THE WORK FOR US!

Model selection (finding the 'right' model)

- With the more general Linear Model formulation, we can create many different models. How to choose which model to use?
- Example: Consider the range of electric vehicles.



  Depends on battery size, driving style or ambient temperature? Or all factors?
- The 'right' model depends on the purpose: find all relevant factors? Prediction of range on a given day?

Overview of model selection approaches

There is not one correct approach to model selection. What makes a good model depends on what it is designed for. Consequently, there are many different tools and techniques.

Broadly, there are three types of model selection approaches:

- Hypothesis tests on parameters within models (include/exclude parameters).
- Measures that describe the quality or goodness of fit of models.
- Hypothesis tests comparing the fit of entire models (often related to measures from the previous point).

Fundamentally, model selection is at the heart of scientific inquiry. Statistical techniques offer one quantitative and rigorous approach.
...we'll go through a few of the most common statistical techniques.

Hypothesis tests on model parameters

- We've enountered these already (test $H_0 : \beta_j = 0$).
- Could determine our model by fitting a *full model* that includes all conceivable explanatory variables first and then removing the explanatory variables for which we cannot reject H_0.
- **Problems:**
  - *Multiple comparisons:* conducting many hypothesis tests means that some may be rejected/accepted by chance (there are ways of dealing with this, e.g. *Bonferroni correction*).
  - The model fit changes every time we remove an explanatory variable.

... while these tests are useful, we may want other approaches that allow us to test global hypotheses about models.

R-squared and adjusted R-squared

- Last lecture we looked at the *coefficient of determination*, R^2 (proportion of variance in response explained by explanatory variables).
- Could use this to distinguish between different models: the closer to 1 it is, the better the model.
- **Problem:** R^2 always increases when we include more explanatory variables.

Fact (Occam's Razor — after work by William of Occam, 1287-1347)

*Entities are not to be multiplied without necessity (Latin: Non sunt multiplicanda entia sine necessitate). **Basically:** prefer simpler models (parsimony).*

- *Adjusted R-squared*, R^2_a: unlike R^2, this takes into account ('adjusts for') the sample size n and the number of model parameters.
- **Warning:** I caution against relying on R^2_a or R^2. They can only be usefully applied in particular circumstances.

F-test on linear models

- Suppose we want to test a more general, global hypothesis about a linear model:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_a : At\ least\ one\ of\ the\ \beta_j \neq 0$$

  ... i.e. we compare our specified model to a *constant model* that only has an intercept, β_0.
- This is called the *F-test* or *Analysis of Variance* (Global). It uses the F-distribution and the test statistic can be expressed in terms of R^2 (we skip the details).
- **Warnings:**
  - The test is very specific (we'll see a more general test in a moment). It asks if the improvement in model fit is larger than expected under H_0.
  - It makes several assumptions (e.g. normally distributed errors), so check model assumptions hold.
  - It doesn't tell us whether our model is correct.

Quality measures based on the likelihood (AIC and BIC)

- Techniques so far rely on the variance in the response explained by models
- Recall the analogy for MLE of maximising $P(data|parameters)$. Could use the maximum likelihood of models , $\hat{L}$, to compare them: highest $\hat{L}$ indicates 'best' model.
- **Warning:** the likelihood faces the same problem as R^2 — it will always improve when more parameters are added to the model.
- **Solution:** penalty for parameters in measures for relative model quality.
  - *Akaike Information Criterion:* $AIC = 2k - 2ln(\hat{L})$, where k is the number of model parameters, including the intercept, but typically excluding the error variance (as this is included in all models). Model with smallest AIC is 'best'.
  - *Bayesian Information Criterion:* $BIC = ln(n)k - 2ln(\hat{L})$. Same idea, but penalty for parameters is stronger (based on different assumptions).
- **Warning:** candidate models must be fitted to the same response data.

Likelihood-ratio test for nested models

- The *Likelihood-ratio test* is a much more general version of the F-test.
- Consider a linear model, M_1 with parameters $\beta_{M_1} = \{\beta_1, \beta_2, ..., \beta_p\}$.
- The test considers a restriction M_0 of M_1, where e.g. $\beta_{M_0} = \{\beta_1, \beta_2 ..., \beta_q, 0, ..., 0\}, q < p$. We say M_0 is *nested* in M_1.
- We test the hypotheses:

$$H_0 : \beta = \beta_{M_0}$$
$$H_a : \beta = \beta_{M_1}$$

  using the test statistic $D = 2ln(\frac{\hat{L_{M_1}}}{\hat{L_{M_0}}}) = 2[ln(\hat{L_{M_1}}) - \hat{L_{M_0}}]$. Under some conditions and assuming H_0, D is asymptotically distributed as $D \sim \chi^2_{p-q}$.
- This is a flexible and very useful test (e.g. tests on individual parameters). Also works when error distribution is not normal.

University of BRISTOL
Applied Statistics 7
Spring Semester 2018

Automated or standardised model selection strategies

- Model selection is time consuming. So people have tried to come up with standardised or even automated procedures, e.g. *stepwise regression*.
- Many software packages, including Matlab, implement such procedures.
- There are many flavours, but the basic idea is:
  - Identify response Y and all potentially important explanatory variables $x_1, x_2, ...x_p$.
  - Automatically work through models defined by all possible combinations of the x_js, starting with simpler models.
  - At each step use standard hypothesis tests (or other measures) to assess if additional explanatory variables improve model.
  - Continue until some stopping criterion is reached.
- **Warning:** control of the process is relinquished to software that can be very intricate. Procedure conducts many statistical tests (multiple comparisons!). Because of these and other issues, many statisticians recommend not to use this approach.

University of BRISTOL
Applied Statistics 7
Spring Semester 2018

Summary table for model selection approaches

Warning: none of these techniques tell you anything about the correctness of a model. So checking model assumptions separately is important (cf. residuals)!

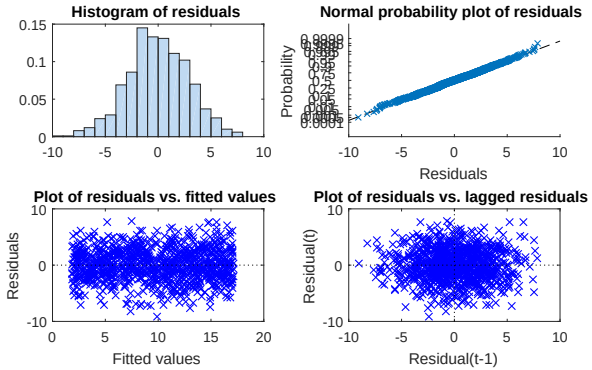
Use	Technique
Hypothesis tests for single parameters	T-test, Likelihood-ratio test
Hypothesis tests for several parameters	F-test, Likelihood-ratio test
Measures for relative quality of models	R^2, R^2_a, Likelihood, AIC, BIC

- Remember, none of these techniques are perfect. Choose your approach depending on what you are trying to achieve with your models.
- With this in mind, prediction intervals and residuals plots could also be used in model selection.

University of BRISTOL
Applied Statistics 7
Spring Semester 2018

Checking LM assumptions

Model checking works in the same way as for sLMs, using residuals.



Additionally, may want to plot each explanatory variables versus residuals.

Typical steps in LM analysis

1. Look at raw data (scatterplots of response versus different explanatory variables).
2. Decide on candidate models (e.g.which predictors are relevant? Exploration versus prediction?).
3. Model selection: find one (or a few) models to look at in more detail.
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... .

University of BRISTOL
Applied Statistics 7
Spring Semester 2018

Matlab output for LM analysis

Most Matlab output was explained in the last lecture. There is one addition:

```
Linear regression model:
    response ~ 1 + predictor1 + predictor2

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    _______    ______    __________

    (Intercept)     3.1677       1.4649    2.1624      0.037123
    predictor1      1.5614      0.19125    8.1639    8.4835e-10
    predictor2      2.8001      0.20877    13.412     8.806e-16


Number of observations: 40, Error degrees of freedom: 37
Root Mean Squared Error: 3.51
R-squared: 0.875,  Adjusted R-Squared 0.869
F-statistic vs. constant model: 130, p-value = 1.88e-17
```

Slide: F-test on linear models