Lectures 4&5 – Program

1. Residuals and diagnostics

2. Variable selection

Assumptions for linear regression

$y_i = \eta_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$

1. Linearity:

   $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

2. Constant variance (homoscedasticity):

   $\text{Var}(\varepsilon_i) = \sigma^2 \qquad \text{all } i$

3. Uncorrelated errors:

   $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j$

4. Normally distributed errors:

   $\varepsilon_i \sim N(0, \sigma^2)$

- Serious violations of 1) can have "catastrophic" consequences.

- Even if 2) or 3) are violated, estimators are unbiased.
  Confidence intervals and p-values will be wrong, however.

- Violations of 4) need not be serious.
  Confidence intervals and p-values are still valid for large samples.
  Outliers may be a problem, however.

Residuals

Population model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Fitted model:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$$

Residuals $\widehat{e}_i = y_i - \widehat{y}_i$

Standardised residuals

$$\widehat{e}_i' = \widehat{e}_i / k_i$$

These are similar to the unstandardised residuals, but have equal variances.

Diagnostics - Plot of residuals

Plots of residuals may be used to check:

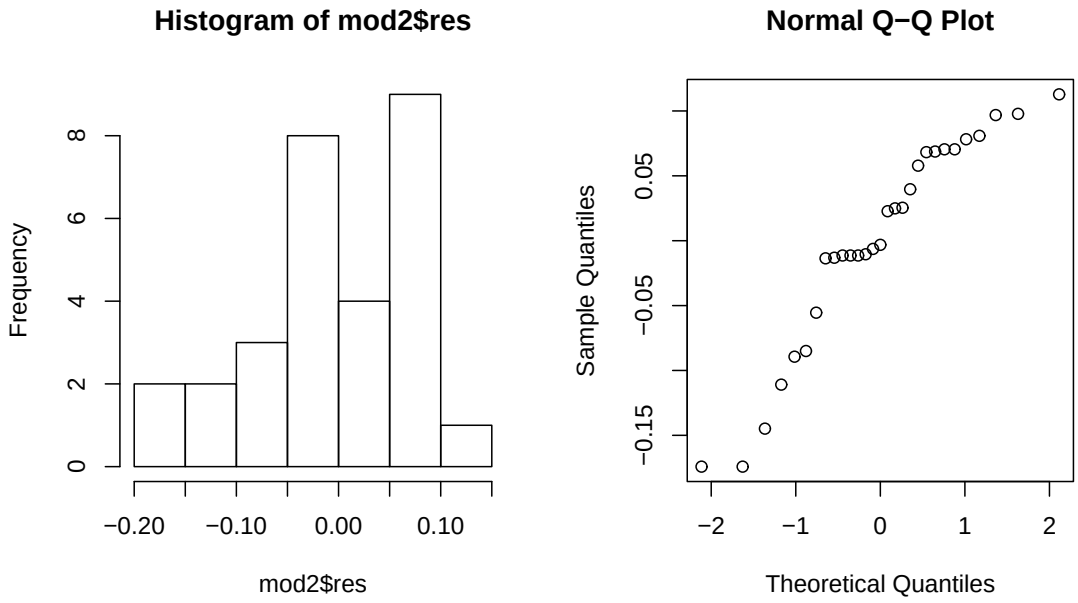
- Normal errors (including outliers)
- Constant variance
- Linearity
- Uncorrelated errors

Normal errors

- Histogram of $\widehat{e}_i$'s. (Symmetric?)
- QQ-plot of $\widehat{e}_i$'s. (Straight line?)
- Box-plot of $\widehat{e}_i$'s. (Outliers?)
- Descriptive statistics of $\widehat{e}_i$'s.

The plots and statistics are useful for detecting deviation normality, including *outliers*.

Example, nicotine content



Histogram not quite symmetric

Some deviation from straight line.

R commands:

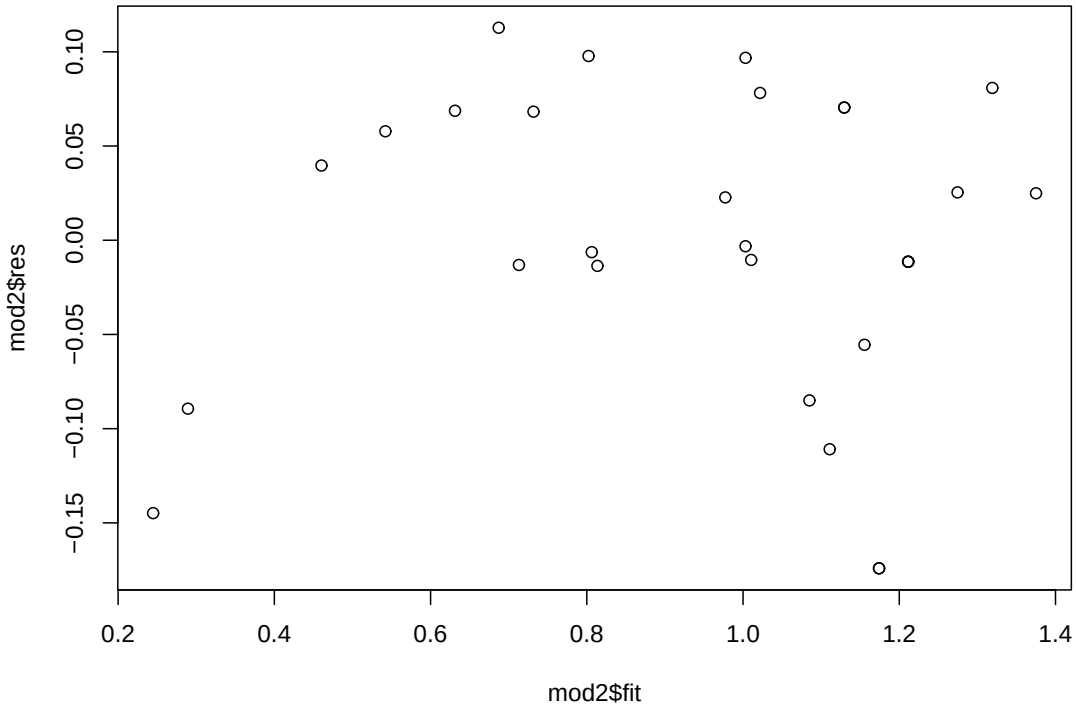
```
mod2<-lm(nicot~co+tar, data=sigarett)
hist(mod2$res)
qqnorm(mod2$res)
```

Constant variance

- Plot of $\widehat{e}_i$ versus $\widehat{y}_i$
- Plot of $|\widehat{e}_i|$ (or $\sqrt{|\widehat{e}_i|}$) versus $\widehat{y}_i$

Larger dispersion of $\widehat{e}_i$ for some $\widehat{y}_i$ indicates heteroscedasticity.

Example, nicotine content



Some indication of heteroscedasticity (or perhaps curvature)

R commands:

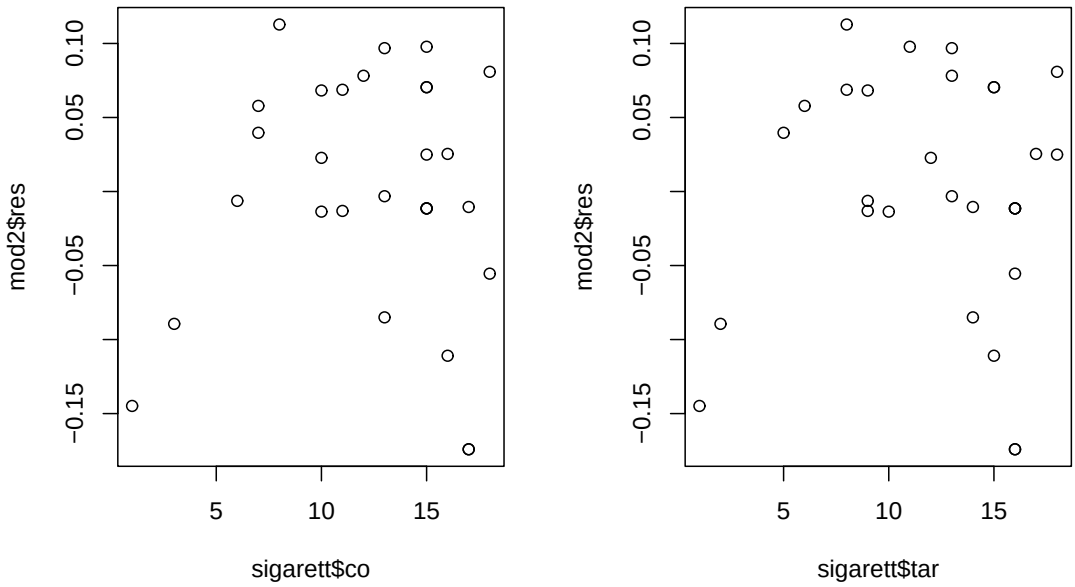
```
mod2<-lm(nicot~co+tar, data=sigarett)
plot(mod2$fit,mod2$res)
```

Linearity

- Plot of $\widehat{e}_i$ versus each covariate x_{ij}

A systematic pattern of the residuals
(e.g. a curvature) indicate deviation from linearity

Example, nicotine content



Some indication of curvature

R commands:

```
mod2<-lm(nicot~co+tar, data=sigarett)
plot(sigarett$co ,mod2$res)
plot(sigarett$tar ,mod2$res)
```

Correlated errors (time series)

Example: y_i = temperature day no i

Possible model: $y_i = \beta_0 + \beta_1 x_i + \gamma y_{i-1} + \varepsilon_i$

Temperature today depend on temperature yesterday

Possible plots:

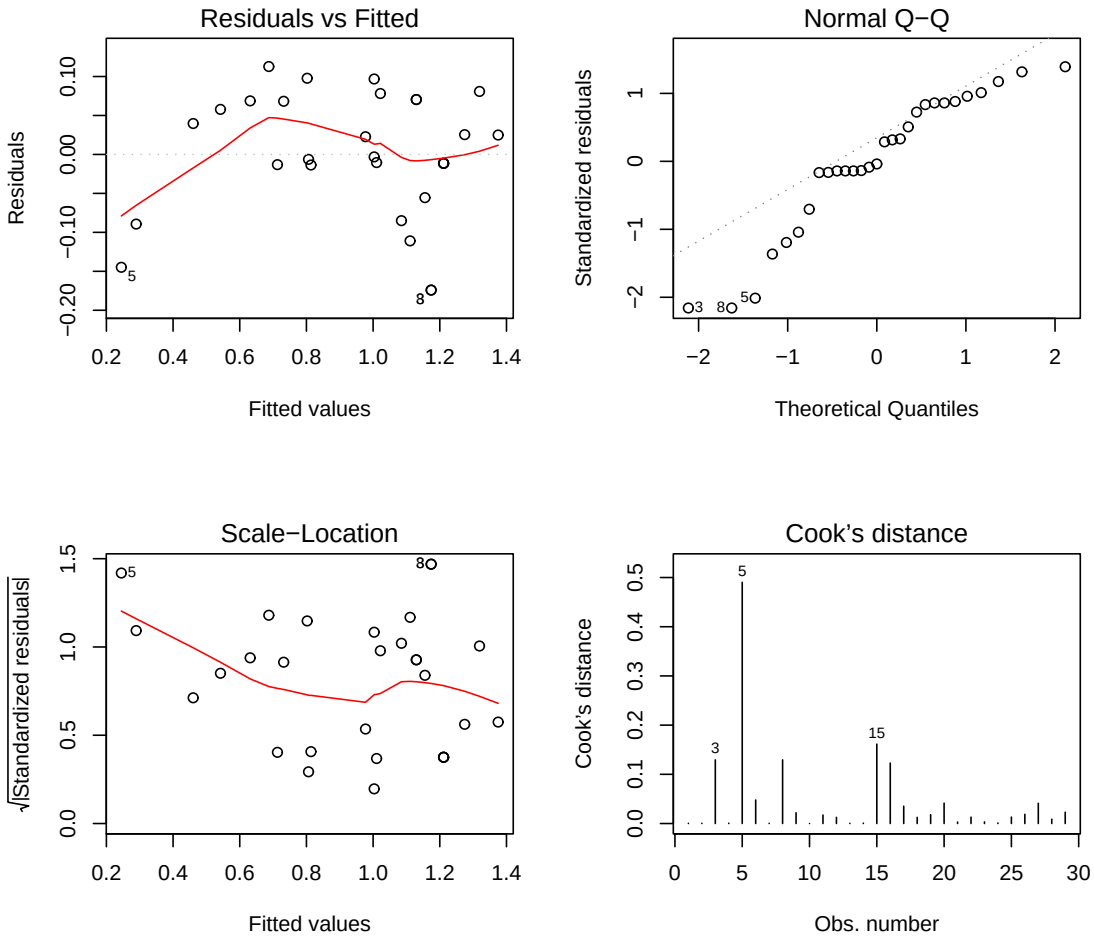
- Plot $\widehat{e}_i$ versus observation number i

- Plot $\widehat{e}_i$ versus previous residual $\widehat{e}_{i-1}$

Diagnostic plots in R

R has some "ready made" resudual plots:

```
mod2<-lm(nicot~co+tar, data=sigarett)
plot(mod2, 1:4)
```



Cook's distance is a measure of the influence each observation

The importance of the model assumptions

- Without linearity of the covariates we have a wrong specification of the systematic part of the model:
  - The effect of a covariate may be wrongly estimated
  - A covariates may be important, but we do not know
  - Serious nonlinearity jeopardizes the analysis

- If the variances are not equal and/or the errors are correlated:
  - The estimates of the β_j's will be unbiased
  - The error variance is wrongly estimated
  - Confidence intervals and p-values are flawed

- If the errors are not normal – but the other model assumptions are true:
  - Estimates of standard errors are valid
  - Test statistics are not exactly t- and F-distributed, but for large n they are approximately so
  - The distributional assumptions are not critical

- A few outliers may have large influence on the estimates. How these are treated may be critical for the conclusions on the relations between covariates and response

Model breakdown and possible improvements

Non-linearity:

- Transform x_i, e.g. log(x_i)
- Transform y_i, e.g. log(y_i)
- Include second order term(s) and/or interaction(s)

Heteroscedasticity:

- Transform y_i, typically log-transform
- More advanced: Use weighted least squares (with weights from the residuals in an unweighted regression)

Model breakdown and possible improvements, cont.

Dependent responses

- Include covariate indicating observation number i:

  $y_i = \beta_0 + \beta_i x_i + \beta_2\, i + \varepsilon_i$

- Include last observation y_{i-1} as covariate:

  $y_i = \beta_0 + \beta_i x_i + \beta_2 y_{i-1} + \varepsilon_i$

  (maybe also y_{i-2}, y_{i-3}, etc.)

- Use time series models

- Other types of dependent data (families, litters, classes in school, etc.):

  Other types of corrections needed.

Model breakdown and possible improvements, cont.

Non-normality

- Transform y_i, e.g. to $\log(y_i)$
- For large n the problem can be ignored
- Use bootstrap

Outliers

- Check the coding of the observations
- Run the regression without outliers. How different are the estimates?

  If the difference is large, you have a problem. Do not ignore it!

Pros and cons in model fitting

- When we know where the model assumptions are problematic, improvements may be possible.

- If several assumptions are violated, it may be difficult to improve all.

- After many improvements we may end up with a well specified, but complex model.

- If the improvements are small, it might be preferable to go for the simpler one.

- Principle of parsimony.

- Avoid over parameterizations.

Selection of variables

Two objectives

- simple model
- good empirical fit

These objectives may be conflicting and
a trade-off is necessary.

We will take a look at criteria and algorithms
that take both considerations into account.

Model with p covariates

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

2^p possibilities to combine the covariates

- $p = 10$: $2^{10} = 1024$ different sub-models
- $p = 20$: $2^{20} \approx 10^6$ different sub-models

For each numeric covariate one may also include e.g. a quadratic term.
Further one may take interactions into account by including products of covariates.

Except for small values of p it is not feasible to investigate all possible models.

Forward selection

1. Fit all p models with only one covariate.
2. Choose the covariate that "contributes most".
3. Run $p-1$ regressions with this covariate and another one.
4. Choose the model that "fits" best.
5. Continue until "no improvement".

There is a variant called **stepwise regression**.

Since covariates that have been included on an earlier stage need not continue to be important later on, step 4 can be supplemented with deletion of covariates that no longer contribute.

Backward selection

1. Fit the model with all p covariates.
2. Compare the model with all covariates with the p different models where one covariate has been deleted.
3. Leave out the "least important" covariate.
4. Compare the model now obtained with the $p-1$ different models where one more covariate has been deleted.
5. Leave out the "least important" covariate.
6. Continue in this way until a model is obtained that only contains "important" covariates.

Criteria for inclusion / exclusion

The squared multiple correlation coefficient

$$R_p^2 = 1 - \frac{SS_{unexp}}{SS_{total}}$$

measures the proportion of the variation explained by the model.

We could try to choose the model with largest R_p^2.

But then we would end up with a model including all covariates.

The criterion must somehow penalize inclusion of covariates.

Possibilities:

- Adjusted R^2
- Cross validated R^2
- Akaike information criteria (AIC)
- Significance

Significance

- Forward:
  Include most significant covariate
  (lowest p-value)

- Backward:
  Exclude least significant covariate
  (largest p-value)

The focus of such a method is *not* on prediction, and that can be a drawback.

Using level 5% often leads to "tighter" models than other criteria.

Adjusted R^2

$$R^2_{adj} = 1 - \frac{SS_{unexp}/(n-p-1)}{SS_{total}/(n-1)}$$

penalizes including more covariates.

Can be used for model selection.

Estimated residual variance:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} SS_{unexp}$$

Using adjusted R^2 is the same as choosing the model having smallest $\widehat{\sigma}^2$.

Cross validation

A drawback with R_p^2 and R_{adj}^2 is that the observations are used both to:

- estimate $\widehat{\beta}_j$'s
- evaluate the predictions of the y_i's:
  $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$

Idea:

- Estimate the regression model without using the observation y_i
- Predict y_i using the obtained estimates. Denote this prediction $\widehat{y}_i^{-i}$.

Cross validated R^2

$$R^2_{cross} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i^{-i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Since R^2_{cross} has a maximum over the different models considered, it can be used for model selection.

There are several ways to perform the cross validation:

- Delete only observation i when computing $\widehat{y}_i^{-i}$
- Split the data in k parts and use the parts *not* containing i when computing $\widehat{y}_i^{-i}$

There is a formula for calculating R^2_{cross} when exactly one observation is deleted. Thus, it is not necessary to do all n auxiliary regressions where one observation is deleted.

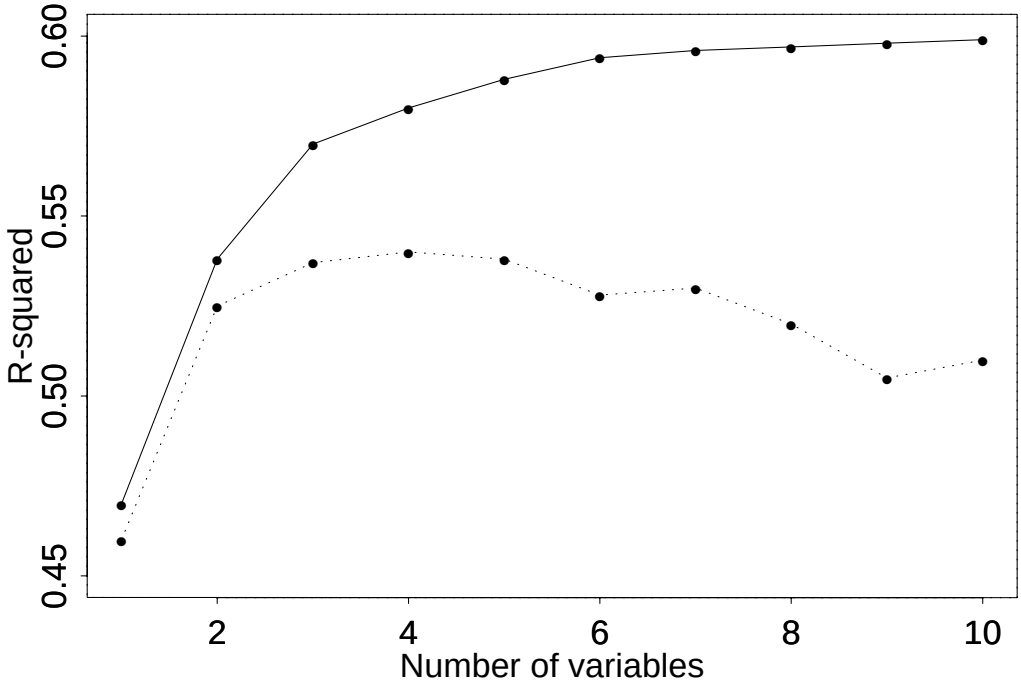
Akaike's information criterion

$$\text{AIC} = n \log\left(\frac{SS_{unexp}}{n}\right) + 2(p+1)$$

Select the model with the smallest AIC.

Example of cross validation

From Bølviken & Skovlund (page 54):



Ordinary R_p^2 (solid line) increases with p, while R_{cross}^p (dotted line) attains a maximum at $p = 4$.

Automatic or manual selection?

Automatic stepwise algorithms are often implemented in statistical software packages.

Can they be trusted?

- Depends on the criterion used
- Cross validation and R^2 may include too many covariates
- Some covariates have intrinsic meaning and should be included for substantive reasons
- Easy to lose "contact" with data

On the other hand

- easy to use
- may get new ideas