Aims of today’s lecture

- To describe some techniques for selecting the explanatory variables for a regression.
- To describe the consequences of making an incorrect choice.
- To apply these techniques to an example.
Variable selection

- Often there are several (perhaps a large number) of potential explanatory variables available to build a regression model. Which ones should we use?

- We could, of course, use them all. However, sometimes this turns out to be not such a good idea.
Over-fitting

If we put too many variables in the model, including some unrelated to the response, we are over-fitting. Consequences are:

▶ Fitted model is not good for prediction of new data - prediction error is underestimated.

▶ Model is too elaborate, models “noise” that will not be the same for new data.

▶ Variances of regression coefficients inflated.
Under-fitting

If we put too few variables in the model, leaving out variables that could help explain the response, we are under-fitting. Consequences are:

- Fitted model is not good for prediction of new data - prediction is biased.
- Regression coefficients are biased
- Estimate of error variance is too large.
Example

- Suppose we have some data which follow a quadratic model

\[ Y = 1 + x/2 + x^2/4 + \varepsilon, \]

where the \( x \)'s are uniformly distributed on \([0, 1]\) and \( \varepsilon \) is standard normal distributed.

- The next slides shows the data, with the true data as a dotted line.
Data and true relationship

Plot of $y$ vs. $x$
Suppose we fit a straight line. This is under-fitting, since we are not fitting the squared term. The fitted line (in green) is shown on the next slide.

Alternatively, we could fit a 6-degree polynomial. This is over-fitting, since there are unnecessary terms in $x^3$, $x^4$, $x^5$ and $x^6$. The fitted polynomial is shown in blue on the next slide. Fit using `lm(y~poly(x,6))`. 
Data and true relationship

Plot of $y$ vs. $x$

- True quadratic
- Linear
- Sixth degree
Points to note

- Straight line is biased: Cannot capture the curvature in the true regression.

- 6-degree line: too variable, attracted to the errors which would be different for a new set of data.

- Moral: For good models we need to choose variables wisely to avoid over-fitting and under-fitting. This is called variable selection.
Uses of regression

The two main uses are

1. To explain the role(s) of the explanatory variables in influencing the response.

2. To construct a prediction equation for predicting the response.

Consequences of over/under fitting are different in each case!
Consider the example of heart disease (D), and two risk factors, alcohol (A) and smoking (S).

Studies have found an association between A and D (a significant regression coefficient if we regress D on A).

There is also an association between A and S.
Possible explanations for the significant alcohol coefficient:

1. Alcohol consumption causes heart disease.

2. Alcohol consumption does not cause heart disease but is associated with smoking that does.
To decide among these, we can fix $S$ and see if $A$ is related to $D$ for fixed $S$. This is measured by the coefficient of $A$ in the model including $S$. Leaving $S$ out gives a biased estimate of the appropriate coefficient.

- Variables like $S$ are called **confounders**, omitting them leads to misleading conclusions.

- Thus, under-fitting is potentially more serious than over-fitting when interpreting coefficients - see example in the next lecture.
The situation is simpler when we are predicting. We choose the model that will give us the smallest prediction error. This is often not the full model.

We will discuss methods for estimating the prediction error later in the lecture, and in the next lecture.
Variable selection

- If we have $k$ variables, and assuming a constant term in each model, there are $2^k - 1$ possible subsets of variables, not counting the null model with no variables. How do we select a subset for our model?

- Two main approaches:
  - All possible regressions (APR, this lecture)
  - Stepwise methods (SWR, next lecture)
All possible regressions

- For each subset of variables, define a criterion of model goodness which tries to balance over-fitting (model too complex) with under-fitting (model does not fit very well).

- Calculate the criterion for each of the $2^k - 1$ models (subsets)

- Pick the best one according to the criterion.

- One difficulty: there are several possible criteria, and they do not always agree.
Possible criteria: $R^2$

- Since $R^2$ increases as we add more variables, picking the model with the biggest $R^2$ will always select the model with all the variables. This will often result in over-fitting!

- However, $R^2$ is OK for choosing between models with the same number of variables.

- We need to modify $R^2$ to penalise overly complicated models. One way is to use the adjusted $R^2$ ($p =$ number of coefficients in model)

$$
\bar{R}_p^2 = 1 - \frac{n - 1}{n - p} (1 - R_p^2)
$$
Suppose we have two models: model A with $p - 1$ variables and model B with an additional $q$ variables (we say A is a submodel of B).

Then the adjusted $R^2$ is defined so that

$$\bar{R}_p^2 < \bar{R}_{p+q}^2$$

if and only if $F > 1$,

where $F$ is the $F$-statistic for testing that model A is adequate.
Possible criteria: Residual mean square (RMS)

- Recall the estimate of the error variance \( \sigma^2 \): estimated by 
  \[ s^2 = \frac{RSS}{n - p}, \]
  sometimes called the residual mean square (RMS)

- Choose model with the minimum RMS.

- We can show that this is equivalent to choosing the model with the biggest adjusted \( R^2 \).
Possible criteria: AIC and BIC

- These are criteria that balance goodness of fit, as measured by RSS, against model complexity, as measured by the number of regression coefficients.

- AIC (Akaike Information Criterion) is, up to a constant depending on $n$

$$AIC = n \log(RMS_p) + 2p$$

- Alternative version is

$$AIC = \frac{RSS_p}{RMS_{full}} + 2p$$
Possible criteria: AIC and BIC

- BIC (Bayesian Information Criterion) is

\[
BIC = \frac{RSS_p}{RMS_{\text{full}}} + p \log(n)
\]

- Small values indicate good models.

- AIC tends to favour more complex models than BIC.
Criteria based on prediction error

▶ Our final set of criteria use an estimate of prediction error to evaluate models.

▶ They measure how well a model predicts new data.
Estimating prediction error: Cross-validation

- If we have plenty of data, we split the data into 2 parts
  - The training set, used to fit the model and construct the predictor;
  - The test set, used to estimate the prediction error.

- Test set error (prediction error) estimated by

\[
TSE = \frac{1}{n} \sum_{\text{test set}} (y_i - \hat{y}_i)^2,
\]

where \( \hat{y}_i \) is the predicted value obtained from the training set.

- Choose model with smallest prediction error.

- NB: Using training set to estimate prediction error underestimates the error (old data).
If we do not have plenty of data, we randomly split the data into 10 parts. One part acts as a test set, the rest as the training set. We compute the prediction error from the test set as before.

Repeat another 9 times, using a different 10th as the test set each time. Average the estimates to get a good estimate of prediction error.

Repeat for different “random splits”.

This is 10-fold cross-validation. Can do 5-fold, or n-fold, but 10-fold seems to be best.
Estimating prediction error: Mallow’s $C_p$

- Suppose we have a model with $p$ regression coefficients. Mallow’s $C_p$ provides an estimate of how well the model predicts new data, and is given by

$$C_p = \frac{RSS_p}{RMS_{\text{full}}} + 2p - n.$$  

- The subscript $\text{full}$ refers to the full model with $k$ variables. Small values of $C_p$ with $C_p$ about $p$ are good.

- **Warning:** $C_{k+1} = k + 1$ always, so do not take this as evidence that the full model is good unless all the other $C_p$’s are larger.

- Note the similarity to AIC.
Example: The fatty acid data

Example: The fatty acid data

The **R330**-function `allpossregs` does the business: e.g., for the fatty acid data.

```r
> library(R330)
> data(fatty.df)
> fatty.lm <- lm(ffa~age+skinfold+weight,data=fatty.df)
> allpossregs(ffa~age+skinfold+weight,data=fatty.df)
```

<table>
<thead>
<tr>
<th>rssp</th>
<th>sigma2</th>
<th>adjRsq</th>
<th>Cp</th>
<th>AIC</th>
<th>BIC</th>
<th>CV</th>
<th>age</th>
<th>skinfold</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.910</td>
<td>0.051</td>
<td>0.380</td>
<td>2.406</td>
<td>22.406</td>
<td>24.397</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.794</td>
<td>0.047</td>
<td>0.427</td>
<td>2.062</td>
<td>22.062</td>
<td>25.049</td>
<td>0.107</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.791</td>
<td>0.049</td>
<td>0.394</td>
<td>4.000</td>
<td>24.000</td>
<td>27.983</td>
<td>0.116</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```
Example: The fatty acid data
Example: The evaporation data

Hydrological Cycle

Units: Thousand cubic km for storage, and thousand cubic km/yr for exchanges
Example: The evaporation data

- This data set was discussed in Tutorial 2.
- The response variable was evap, the amount of moisture evaporating from the soil in a 24 hour period.
- There are 10 explanatory variables, six measurements of temperatures, three measurements of humidity, and the average wind speed.
Example: The evaporation data
Example: The evaporation data

- There are strong relationships between the variables, so we probably do not need them all. We can perform an all possible regressions analysis using the following code.

```r
library(R330)

data(evap.df)

evap.lm <- lm(evap~.,data=evap.df)

allpossregs(evap~.,data=evap.df)
```
Example: The evaporation data

Call:
`lm(formula = evap ~ ., data = evap.df)`

---

Coefficients:

|         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------|----------|------------|---------|----------|
| (Intercept) | -54.074877 | 130.720826 | -0.414 | 0.68164 |
| avst     | 2.231782  | 1.003882   | 2.223   | 0.03276 * |
| minst    | 0.204854  | 1.104523   | 0.185   | 0.85393  |
| maxst    | -0.742580 | 0.349609   | -2.124  | 0.04081 * |
| avat     | 0.501055  | 0.568964   | 0.881   | 0.38452  |
| minat    | 0.304126  | 0.788877   | 0.386   | 0.70219  |
| maxat    | 0.092187  | 0.218054   | 0.423   | 0.67505  |
| avh      | 1.109858  | 1.133126   | 0.979   | 0.33407  |
| minh     | 0.751405  | 0.487749   | 1.541   | 0.13242  |
| maxh     | -0.556292 | 0.161602   | -3.442  | 0.00151 ** |
| wind     | 0.008918  | 0.009167   | 0.973   | 0.33733  |

---

Residual standard error: 6.508 on 35 degrees of freedom
Multiple R-squared:    0.8463, Adjusted R-squared:    0.8023
F-statistic: 19.27 on 10 and 35 DF,  p-value: 2.073e-11
Example: The evaporation data

<table>
<thead>
<tr>
<th>rssp</th>
<th>sigma2</th>
<th>adjRsq</th>
<th>Cp</th>
<th>AIC</th>
<th>BIC</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>3071.255</td>
<td>69.801</td>
<td>0.674</td>
<td>30.519</td>
<td>76.519</td>
<td>80.177</td>
<td>301.047</td>
</tr>
<tr>
<td>2101.113</td>
<td>48.863</td>
<td>0.772</td>
<td>9.612</td>
<td>55.612</td>
<td>61.098</td>
<td>220.819</td>
</tr>
<tr>
<td>1879.949</td>
<td>44.761</td>
<td>0.791</td>
<td>6.390</td>
<td>52.390</td>
<td>59.705</td>
<td>206.121</td>
</tr>
<tr>
<td>1696.789</td>
<td>41.385</td>
<td>0.807</td>
<td>4.065</td>
<td>50.065</td>
<td>59.208</td>
<td>207.489</td>
</tr>
<tr>
<td>1599.138</td>
<td>39.978</td>
<td>0.813</td>
<td>3.759</td>
<td>49.759</td>
<td>60.731</td>
<td>209.986</td>
</tr>
<tr>
<td>1552.033</td>
<td>39.796</td>
<td>0.814</td>
<td>4.647</td>
<td>50.647</td>
<td>63.448</td>
<td>215.114</td>
</tr>
<tr>
<td>1521.227</td>
<td>40.032</td>
<td>0.813</td>
<td>5.920</td>
<td>51.920</td>
<td>66.549</td>
<td>231.108</td>
</tr>
<tr>
<td>1490.602</td>
<td>40.287</td>
<td>0.812</td>
<td>7.197</td>
<td>53.197</td>
<td>69.654</td>
<td>243.737</td>
</tr>
<tr>
<td>1483.733</td>
<td>41.215</td>
<td>0.808</td>
<td>9.034</td>
<td>55.034</td>
<td>73.321</td>
<td>257.513</td>
</tr>
<tr>
<td>1482.277</td>
<td>42.351</td>
<td>0.812</td>
<td>11.000</td>
<td>57.000</td>
<td>77.115</td>
<td>275.554</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>avst</th>
<th>minst</th>
<th>maxst</th>
<th>avat</th>
<th>minat</th>
<th>maxat</th>
<th>avh</th>
<th>minh</th>
<th>maxh</th>
<th>wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Example: The evaporation data
Example: The evaporation data

Call:
`lm(formula = evap ~ maxat + maxh + wind, data = evap.df)`

---

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|)  |
|------------|----------|------------|---------|-----------|
| (Intercept)| 123.9018 | 24.6244    | 5.032   | 9.60e-06  *** |
| maxat      | 0.222768 | 0.059113   | 3.769   | 0.000506  *** |
| maxh       | -0.342915| 0.042776   | -8.016  | 5.31e-10  *** |
| wind       | 0.015998 | 0.007197   | 2.223   | 0.031664  *  |

---

Residual standard error: 6.69 on 42 degrees of freedom
Multiple R-squared: 0.805, Adjusted R-squared: 0.7911
F-statistic: 57.8 on 3 and 42 DF, p-value: 5.834e-15
Thank you

YOU'VE BEEN AT THIS FOR HOURS, WHAT ON EARTH ARE YOU DRAWING?

A BLANK!

2010-2012 SOMETHINGOROTHERWISE.COM