Statistical Models: Purpose

- Describe relationship between explanatory variables and response.
- Use model to make predictions for a set of observations for explanatory variables.
- This course is particularly interested in Multiple Linear Regression Models.
Multiple Linear Regression Models: Assumptions

- The relationship between the response and the explanatory variables is linear.

- The errors are independent, normal and have constant variance.
Multiple Linear Regression Models: Diagnostics

Non-Linearity: Residuals vs. Fitted-values, Residuals vs. Explanatory variables, Added Variable Plots, GAM plots

Constant Variance: Funnel plots and Weighted RSS

Outliers: Diagnostics 3 (TODAY)

Independence: Diagnostics 4

Normality of Errors: Diagnostics 5
Aims of today’s lecture

- Introduce terminology for outliers and high-leverage points.
- Introduce a broad band of diagnostic tools to identify and treat such extraordinary data points.
Outliers and high-leverage points

- An **outlier** is a point that has a larger or smaller \( y \) value that the model would suggest.
  - Can be due to a genuine large error \( \varepsilon \);
  - Can be caused by typographical errors in recording the data.

- A **high-leverage point** is a point with extreme values of the explanatory variables.
Outliers

- The effect of an outlier depends on whether it is also a high-leverage point.

- A high-leverage outlier
  - Can attract the fitted plane, distorting the fit, sometimes extremely;
  - In extreme cases may not have a big residual;
  - In extreme cases can increase $R^2$.

- A low-leverage outlier
  - Does not distort the fit to the same extent;
  - Usually has a big residual;
  - Inflates standard errors, decreases $R^2$. 
Outliers

No high-leverage points
No outliers

Low-leverage outlier
Big residual

High-leverage point
Not an outlier

High-leverage outlier
Example: The education data (without urban)

High leverage point

![Graph showing educational data with a high leverage point. The x-axis represents percap, the y-axis represents educ, and the z-axis represents under18. A red arrow indicates a high leverage point.]
An outlier too?

Number of residents per 1000 under 18

per capita income

Residual somewhat extreme
Measuring leverage

It can be shown (see e.g. STATS 310) that the fitted value of case \( i \) is related to the response data \( y_1, \ldots, y_n \) by the equation

\[
\hat{y} = H y, \quad \text{where} \quad H = X \left( X^T X \right)^{-1} X^T,
\]

\[
\hat{y}_i = h_{i1} y_1 + \cdots + h_{ii} y_i + \cdots + h_{in} y_n.
\]

- The \( h_{ij} \) depend on the explanatory variables. The quantities \( h_{ii} \) are called hat matrix diagonals (HMD’s) and measure the influence \( y_i \) has on the \( i^{\text{th}} \) fitted value.

- They can also be interpreted as the distance between the \( X \)-data for the \( i^{\text{th}} \) case and the average \( x \)-data for all the cases. Thus, they directly measure how extreme the \( x \)-values of each point are.
Interpreting the HMD’s

- Each HMD lies between 0 and 1.
- The average HMD is \((k + 1)/n\).
- An HMD larger than \(3(k + 1)/n\) is considered extreme.
Example: The education data (without urban)

\[
> \text{hatvalues(educ.lm)}
\]

Hat matrix values for educ.lm

\[
\begin{align*}
\frac{k + 1}{n} &= \frac{3}{50} \\
\frac{3(k + 1)}{n} &= \frac{9}{50}
\end{align*}
\]
Studentised residuals

- How can we recognise a big residual? How big is big?

- The actual size depends on the units in which the $y$-variable is measured, so we need to standardize them.

- Can divide by their standard deviations.

- Variance of a typical residual $e_i$ is

$$ \text{Var}(e_i) = (1 - h_{ii}) \sigma^2, $$

where $h_{ii}$ is the $i^{th}$ diagonal entry of the hat matrix $H$. 
Studentised residuals

- **Internally studentised (or standardised in R):**

\[
\hat{e}_i = \frac{e_i}{\sqrt{(1 - h_{ii})s^2}}
\]

where \( s^2 \) is the usual estimate of the residual variance \( \sigma^2 \).

- **Externally studentised (or studentised in R):**

\[
\hat{e}_i = \frac{e_i}{\sqrt{(1 - h_{ii})s_i^2}}
\]

where \( s_i^2 \) is the estimate of \( \sigma^2 \) after deleting the \( i^{th} \) data point.
Studentised residuals

- How big is big?

- Both types of studentised residual are approximately distributed as standard normals when the model is OK and there are no outliers. (in fact the externally studentised one has a $t$- distribution).

- Thus, studentised residuals should be between $-2$ and $2$ with approximately 95% probability.
#Load the MASS library
> library(MASS)

# internally studentised (standardised in R)
> stdres(educ.lm)[50]
  50
3.089699

# externally studentised (studentized in R)
> studres(educ.lm)[50]
  50
3.424107
What does studentised mean?
Recognising outliers

- If a point is a low influence outlier, the residual will usually be large, so large residual and a low HMD indicates an outlier.

- If a point is a high leverage outlier, then a large error usually will cause a large residual.

- However, in extreme cases, a high leverage outlier may not have a very big residual, depending on how much the point attracts the fitted plane. Thus, if a point has a large HMD, and the residual is not particularly big, we cannot always tell if a point is an outlier or not.
High-leverage outlier
Leverage-residual plots

```r
> plot(educ.lm, which=5)
```

![Leverage-residual plots](image)
Interpreting LR plots

- Low-leverage Outlier
- High-leverage Outlier
- Potential High-leverage Outlier
- OK
No big studentised residuals, no big HMD’s
One big studentised residuals, no big HMD’s

Plot of $y$ vs. $x$, Example 2

Residuals vs Leverage

- Fitted Line
- True Line

3(k+1)/n=0.2

Cook’s distance

0.5
1

Leverage

Standardized residuals

0 2 4 6 8
0 10 20 30 40
No big studentised residuals, one big HMD’s

Plot of y vs. x, Example 3

Residuals vs Leverage
3(k+1)/n=0.2
One big studentised residuals, one big HMD’s

Plot of $y$ vs. $x$, Example 4

Residuals vs Leverage

$3(k+1)/n = 0.2$
Three big studentised residuals, one big HMD’s

Plot of $y$ vs. $x$, Example 5

Residuals vs Leverage

3(k+1)/n=0.2
HMD Summary

- Hat matrix diagonals
  - Measure the effect of a point on its fitted value;
  - Measure how outlying the $x$-values are (how “high-leverage” a point is);
  - Are always between 0 and 1 with bigger values indicating high-leverage;
  - Points with HMD’s more than $3(k + 1)/n$ are considered “high-leverage”
Influential points

- How can we tell if a high-leverage point/outlier is affecting the regression?

- By deleting the point and refitting the regression: a large change in coefficients means the point is affecting the regression.

- Such points are called influential points.

- We do not want our analysis to be driven by one or two points!
“Leave one out” measures

- We can calculate a variety of measures by leaving out each data point in turn, and looking at the change in key regression quantities such as:
  - Coefficients
  - Fitted Values
  - Standard errors

- We discuss each in turn.
### Example: Education data

<table>
<thead>
<tr>
<th></th>
<th>With point 50</th>
<th>Without point 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const</td>
<td>-557.451</td>
<td>-298.714</td>
</tr>
<tr>
<td>percap</td>
<td>0.072</td>
<td>0.059</td>
</tr>
<tr>
<td>under18</td>
<td>1.555</td>
<td>0.933</td>
</tr>
</tbody>
</table>
Coefficient measures: DFBETAS

**DFBETAS:** Standardised difference in coefficients

\[ \text{dfbetas} = \frac{\hat{\beta}_j - \hat{\beta}_j[-i]}{\text{se}(\hat{\beta}_j)} \]

**Problematic:** when \(|\text{dfbetas}| > 1\). This is a criterion coded into R.
Coefficient measures: DFFITS

**DFFITS**: Standardised difference in fitted values:

\[
\text{dffits} = \frac{\hat{y}_j - \hat{y}_j[-i]}{\text{se}(\hat{y}_j)}
\]

Problematic: when

\[
|\text{dffits}| > 3 \sqrt{\frac{k + 1}{n - k - 1}}.
\]
Coefficient measures: Covariance Ratio and Cook’s $D$

**Cov Ratio:** Measures change in the standard errors of the estimated coefficients.

**Problem:** when Cov Ratio greater than $1 + 3(k + 1)/n$ or smaller than $1 - 3(k + 1)/n$.

**Cook’s $D$:** Measures overall change in the coefficients.

**Problem:** when greater than $q_{f}(0.5, k+1, n-k-1)$ (median of $F$-distribution), roughly 1 in most cases.
Influence measures of
\[
\text{lm(formula = educ} \sim \text{percap + under18, data = educ.df)}:
\]

<table>
<thead>
<tr>
<th></th>
<th>dfb.1_</th>
<th>dfb.prcp</th>
<th>dfb.un18</th>
<th>dffit</th>
<th>cov.r</th>
<th>cook.d</th>
<th>hat</th>
<th>inf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0120</td>
<td>-0.01794</td>
<td>-0.00588</td>
<td>0.0233</td>
<td>1.121</td>
<td>1.84e-04</td>
<td>0.0494</td>
<td>***</td>
</tr>
<tr>
<td>10</td>
<td>0.0638</td>
<td>-0.16792</td>
<td>-0.02222</td>
<td>-0.3631</td>
<td>0.803</td>
<td>4.05e-02</td>
<td>0.0257 *</td>
<td>***</td>
</tr>
<tr>
<td>44</td>
<td>0.0229</td>
<td>0.00298</td>
<td>-0.02948</td>
<td>-0.0340</td>
<td>1.283</td>
<td>3.94e-04</td>
<td>0.1690 *</td>
<td>***</td>
</tr>
<tr>
<td>50</td>
<td>-2.3688</td>
<td>1.50181</td>
<td>2.23393</td>
<td>2.4733</td>
<td>0.821</td>
<td>1.66e+00</td>
<td>0.3429 *</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>1.0000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.7579</td>
<td>0.82</td>
<td>8.00e-01</td>
<td>0.18</td>
<td>1.18</td>
</tr>
</tbody>
</table>
# Plotting influence

There will be seven plots
par(mfrow=c(2,4))
# Plot the measures using R330
influenceplots(educ.lm)
Remedies for outliers

- Correct typographical errors in the data.
- Delete a small number of points and refit (do not want fitted regression to be determined by one or two influential points)
- Report existence of outliers separately: they are often of scientific interest
- Do not delete too many points (1 or 2 max)
Summary: Doing it in R

- Leverage-Residual plot
  ```r
  plot(educ.lm, which=5)
  ```
- Full diagnostic display
  ```r
  par(mfrow=c(2,2))
  plot(educ.lm)
  ```
- Influence measures
  ```r
  influence.measures(educ.lm)
  ```
- Plots of influence measures
  ```r
  par(mfrow=c(2,4))
  influenceplot(educ.lm)
  ```
Statistical outliers

http://xkcd.com/539/