Diagonstic steps

We test for $\neq$ using $\neq$:
- Planarity: Residuals vs. fitted values, Added variable plots, GAM plots, Box-Cox plot
- Constant Variance: Funnel plots, weighted least squares
- Normality of Errors: QQ plots, Weisberg-Bingham test (and many more)
- Outliers: ...
- Independence: ...

R-hint of the day

Applying a function to a data frame
```r
> data(cherry.df)
> apply(cherry.df,2,mean)
  diameter height volume
13.24839 76.00000 30.17097
> round(sapply(1:31,function(k)
  {cherry.df[k,1]/12}
},2)
[1] 0.69 0.72 0.73 0.88 0.92 0.92 0.92 0.92 0.92
[11] 0.94 0.95 0.95 0.98 1.00 1.07 1.07 1.11 1.14 1.15
[21] 1.17 1.18 1.21 1.33 1.36 1.44 1.46 1.49 1.50 1.50
[31] 1.72
```

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.
- Introduce diagnostic tools to check for independence of errors.

Outliers and high-leverage points

- An outlier is a point that has a larger or smaller $y$ value than the model would suggest:
  - Can be due to a genuine large error $\epsilon$;
  - 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

Example: The education data (without urban)

Measuring leverage

Fitted values $\hat{y}$ are related to response $y$ by the equation

$$\hat{y} = Hy, \quad \text{where } H = X (X^T X)^{-1} X^T,$$

$$\hat{y}_i = h_{i1} y_1 + \cdots + h_{i2} y_2 + \cdots + h_{in} y_n.$$  

- $h_{ij}$ depend explanatory variables $X$, $h_i$ is called hat matrix diagonals (HMD's), measures the influence $y_i$ has on $\hat{y}_i$.  
- Also distance between average $x_i$, and average $\bar{x}$, measures how extreme observations are.

Interpreting the HMD's

Example: The education data (without urban)

> hatvalues(educ.lm)

- 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.
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?
- The internally studentised residuals are approximately standard normal distributed if the model is OK and there are no outliers.
- The externally studentised residuals has a \( t \)-distribution.
- Thus, studentised residuals should be between \(-2\) and \(2\) with approximately 95\% probability.

Studentised residuals: Calculating it in R

```r
#Load the MASS library
> library(MASS)
# internally studentised (standardised in R)
> stdres(educ.lm)[50]
3.089699
# externally studentised (studentized in R)
> studres(educ.lm)[50]
3.424107
```

What does studentised mean?

- 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.

Recognising outliers
Interpreting LR plots

High-leverage outlier

Leverage-residual plots

> plot(educ.lm,which=5)

False Line

True Line

Fitted Line

Residuals vs Fitted

Residuals vs Leverage

Outlier

Potential High-leverage Outlier

No big studentised residuals, no big HMD’s

One big studentised residuals, no big HMD’s

No big studentised residuals, one big HMD’s

Leverage

HMD
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
\[ \text{dfbeta}_j = \frac{\hat{\beta}_j - \bar{\beta}_j}{\text{se}(\bar{\beta}_j)} \]

Problematic: when |dfbeta| > 1. This is a criterion coded into R.

Coefficient measures: DFFITS

DFFITS: Standardised difference in fitted values:
\[ \text{dffits}_j = \frac{\hat{y}_j - \bar{y}_j}{\text{se}(\bar{y}_j)} \]

Problematic: when |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 \(qf(.5,k+1,n-k-1)\) (median of \(F\)-distribution), roughly 1 in most cases.

Influence measures of lm(formula = educ ~ 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>
</tbody>
</table>

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.
Diagnostic steps

We test for $\leftrightarrow$ using $\leftrightarrow$:

- **Planarity**: Residuals vs. fitted values, Added variable plots, GAM plots, Box-Cox plot
- **Constant Variance**: Funnel plots, weighted least squares
- **Normality of Errors**: QQ plots, Weisberg-Bingham test (and many more)
- **Outliers**: Leverage-Residual plots, influence measures

Independence

- One of the regression assumptions is that the errors are independent.
- Data that is collected sequentially over time often has errors that are not independent.
- If the independence assumption does not hold, then the standard errors will be wrong and the tests and confidence intervals will be unreliable.
- We need to be able to detect lack of independence.

Types of dependence

- If large positive errors have a tendency to follow large positive errors, and large negative errors a tendency to follow large negative errors, we say the data has positive autocorrelation.
- If large positive errors have a tendency to follow large negative errors, and large negative errors a tendency to follow large positive errors, we say the data has negative autocorrelation.

Diagnostics: Positive Autocorrelation

If the errors are positively autocorrelated:

- Plotting the residuals against time will show long runs of positive and negative residuals.
- Plotting residuals against the previous residual (i.e. $e_i$ vs. $e_{i-1}$) will show a positive trend.
- A correlogram of the residuals will show positive spikes, gradually decaying.

Diagnostics: Negative Autocorrelation

If the errors are negatively autocorrelated:

- Plotting the residuals against time will show alternating positive and negative residuals.
- Plotting residuals against the previous residual (i.e. $e_i$ vs. $e_{i-1}$) will show a negative trend.
- A correlogram of the residuals will show alternating positive and negative spikes, gradually decaying.

Residuals against time

```r
res <- residuals(lm.obj)
plot(1:length(res), res, xlab="time", ylab="residuals", type="b")
abline(h=0, lty=2)
```
Residuals against time: Plot

Residuals against their predecessor

\[
\text{res} \leftarrow \text{residuals(lm.obj)}
\]
\[
\text{n} \leftarrow \text{length(res)}
\]
\[
\text{plot.res} \leftarrow \text{res[-1]} \quad \text{# element 1 has no predecessor}
\]
\[
\text{prev.res} \leftarrow \text{res[-n]} \quad \text{# last residual has no successor}
\]
\[
\text{plot(prev.res,plot.res,xlab="previous residual", ylab="residual")}
\]

Residuals against their predecessor: Plot

Correlogram

\[
\text{acf(residual(lm.obj))}
\]

- The autocorrelation function (acf, aka correlogram) investigates the correlation between a residual and another residual k time units apart.
- This is also called lag k autocorrelation.

Correlogram: Plot

Durbin-Watson test

- We can also do a formal hypothesis test, the Durbin-Watson test, for independence.
- The test assumes the errors follow a model of the form
  \[
  e_i = \varrho e_{i-1} + \epsilon_i
  \]
  where the \( \epsilon_i \)'s are independent, normal and have constant variance. \( \varrho \) is the lag 1 correlation.
- This is the autoregressive model of order 1.
Durbin-Watson test

- When $\rho = 0$ then the errors are independent.
- **DW** test statistic is
  \[
  DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \approx 2(1 - \hat{\rho})
  \]
- Value of **DW** is between 0 and 4;
- Values of **DW** around 2 are consistent with independence;
- Values close to 4 indicate negative autocorrelation;
- Values close to 0 indicate positive autocorrelation.

Durbin-Watson test in R

```r
> library(car)
> dwt(model, simulate=T, max.lag=1)
```

Does advertising increase sales? Residual plots

- Residuals against time
- Comparing Lag 1 residuals
Does advertising increase sales? Residual plots

![Correlogram of residuals](image)

**Durbin-Watson test**

```r
> library(R330)
> library(car)
> data(ad.df)
> ad.lm <- lm(sales~spend+prev.spend,data=ad.df)
> dwt(ad.lm,max.lag=2)

<table>
<thead>
<tr>
<th>lag</th>
<th>Autocorrelation</th>
<th>D-W Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.44249700</td>
<td>1.103870</td>
<td>0.006</td>
</tr>
<tr>
<td>2</td>
<td>0.07550052</td>
<td>1.789442</td>
<td>0.612</td>
</tr>
</tbody>
</table>

Alternative hypothesis: rho[lag] != 0
```

**Remedy**

- If we detect serial correlation, we need to fit special time series models to the data.
- For full details see STATS 326/726.

**Fitting the new model**

```r
> time <- 1:35
> ad.lm2 <- lm(sales~spend+prev.spend+time,data=ad.df)
> dwt(ad.lm2,max.lag=2)

<table>
<thead>
<tr>
<th>lag</th>
<th>Autocorrelation</th>
<th>D-W Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1234779</td>
<td>1.619324</td>
<td>0.152</td>
</tr>
<tr>
<td>2</td>
<td>-0.3925537</td>
<td>2.608298</td>
<td>0.070</td>
</tr>
</tbody>
</table>

Alternative hypothesis: rho[lag] != 0
```

- Lag 1 problem resolved, time is significant variable in model.

**Diagostic steps**

We test for <= using <=:

- **Planarity**: Residuals vs. fitted values, Added variable plots, GAM plots, Box-Cox plot
- **Constant Variance**: Funnel plots, weighted least squares
- **Normality of Errors**: QQ plots, Weisberg-Bingham test (and many more)
- **Outliers**: Leverage-Residual plots, influence measures
- **Independence**: Correlogram, Durbin-Watson test