Assessing Model Fit

- Our model has assumptions:
  - mean 0 errors,
  - functional form of response,
  - lack of need for other regressors,
  - constant variance,
  - normally distributed errors,
  - independent errors.

- These should be checked as much as possible.

- Major tool is study of residuals.
Residual Analysis

**Definition:** The residual vector whose entries are called “fitted residuals” or “errors” is

\[ \hat{\epsilon} = Y - X\hat{\beta}. \]

- Examine residual plots to assess quality of model.
- Plot residuals \( \hat{\epsilon}_i \) against each \( x_i \), i.e. against \( S_i \) and \( F_i \).
- Plot residuals against other covariates, particularly those deleted from model.
- Plot residuals against \( \hat{\mu}_i = \) fitted value.
- Plot residuals squared against all above.
- Make Q-Q plot of residuals.
Look For

- Curvature — suggesting need of $x^2$ or non-linear model.
- Heteroscedasticity.
- Omitted variables.
- Non-normality.
Example

Here is a page of plots:

- Residual vs Sand
  - Sand Content (%)
  - Residual
  - Residual vs Fibre
  - Fibre Content (%)
  - Residual vs Fitted
  - Fitted Value
  - Q-Q Plot
  - Quantiles of Standard Normal

Richard Lockhart

STAT 350: Distribution Theory
Q-Q Plots

- Used to check normal assumption for the errors.
- Plot order statistics of residuals against quantiles of $N(0,1)$: a Q-Q plot:

$$\hat{\epsilon}(1) < \hat{\epsilon}(2) < \cdots < \hat{\epsilon}(n)$$

are the $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ arranged in increasing order — called “order statistics”. Also

$$s_1 < \cdots < s_n$$

are “Normal scores”. They are defined by the equation

$$P(N(0, 1) \leq s_i) = \frac{i}{n + 1}$$

- Plot of $s_i$ versus $\hat{\epsilon}_i$ should be near straight line for normal errors.
Conclusions from plots

- Q-Q plot is reasonably straight. So normality is OK and $t$ and $F$ tests should work well.
- The plot of residual versus fitted values is more or less OK.
- **Warning**: don’t look too hard for patterns; you will find them where they aren’t.
- The plot of residual versus Sand is ok.
- The plot of residual versus Fibre has mostly positive residuals for the middle values of Fibre suggesting a quadratic pattern.
Consequences

So, we compare

\[ Y = \beta_0 + \beta_1 S + \beta_3 F + \epsilon \]

and

\[ Y = \beta_0 + \beta_1 S + \beta_3 F + \beta_4 F^2 + \epsilon \]

Use \( t \) test on \( \beta_4 \) to test \( H_0 : \beta_4 = 0 \) in second model.

We find

\[ \hat{\beta}_4 = -0.00373 \]

\[ \hat{\sigma}_{\beta_4} = 0.001995 \]

\[ t = \frac{-0.00373}{0.001995} = -1.87 \]

based on 14 degrees of freedom.
More discussion

- So we get the marginally not significant $P$ value 0.08.
- Conclusion: evidence of need for the $F^2$ term is weak.
- We might want more data if the “optimal” Fibre content is needed.
- Notice as always: statistics does not eliminate uncertainty but rather quantifies it.
More formal model assessment tools

1. Fit larger model: test for non-zero coefficients.
2. We did this to compare linear to full quadratic model.
3. Look for outlying residuals.
4. Look for influential observations.
Standardized / studentized residuals

- Standardized residual is $\hat{\epsilon}_i / \hat{\sigma}$.
- Recall that $\hat{\epsilon} \sim MVN(0, \sigma^2(I - H))$.
- It follows that $\hat{\epsilon}_i \sim N(0, \sigma^2(1 - h_{ii}))$ where $h_{ii}$ is the $ii$th diagonal entry in $H$.
- **Jargon**: We call $h_{ii}$ the leverage of case $i$.
- We see that $\frac{\hat{\epsilon}_i}{\sigma \sqrt{1 - h_{ii}}} \sim N(0, 1)$.
Internally Studentized Residuals

- Replace $\sigma$ with the obvious estimate and find that

$$\frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \sim N(0, 1)$$

provided that $n$ is large.

- Called an **internally studentized** or **standardized** residual.

- SUGGESTION: look for studentized residuals larger than about 2.

- The original standardized residuals are also often used for this.

- The $h_{ii}$ add up to the trace of the hat matrix $= p$.

- Average $h$ is $p/n$ which should be small so usually $\sqrt{1 - h_{ii}}$ near 1.
Comments

- **Warning**: the $N(0, 1)$ approximation **requires** normal errors.
- Criticism of internally standardized residuals: if model is bad particularly at point $i$ then including point $i$ pulls the fit towards $Y_i$, inflates $\hat{\sigma}$ and makes the badness hard to see.
- Coming soon: eliminate $Y_i$ from estimate of $\sigma$ to compute slightly different residual.
Outlier Plot
Deleted Residuals

- Suggestion: for each point $i$ delete point $i$, refit the model, predict $Y_i$.
- Call the prediction $\hat{Y}_{i(i)}$ where the $(i)$ in the subscript shows which point was deleted.
- Then get **case deleted residuals**

$$Y_i - \hat{Y}_{i(i)}$$
Standardized Residuals

For insurance data residuals after various model fits:

data insure;
  infile 'insure.dat' firstobs=2;
  input year cost;
  code = year - 1975.5 ;
proc glm data=insure;
  model cost = code ;
  output out=insfit h=leverage p=fitted
      r=resid student=isr press=press rstudent=esr;
run ;
proc print data=insfit ;
run;
proc glm data=insure;
  model cost = code code*code code*code*code ;
  output out=insfit3 h=leverage p=fitted r=resid
     student=isr press=press rstudent=esr;
run ;
proc print data=insfit3 ;
run;
proc glm data=insure;
   model cost = code code*code code*code*code
              code*code*code*code code*code*code*code*code;
   output out=insfit5 h=leverage p=fitted r=resid
           student=isr press=press rstudent=esr;
run ;
proc print data=insfit5 ;
run;
<table>
<thead>
<tr>
<th>OBS</th>
<th>YEAR</th>
<th>COST</th>
<th>CODE LEVERAGE</th>
<th>FITTED</th>
<th>RESID</th>
<th>ISR</th>
<th>PRESS</th>
<th>ESR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1971</td>
<td>45.13</td>
<td>-4.5</td>
<td>42.5196</td>
<td>2.6104</td>
<td>0.36998</td>
<td>3.9881</td>
<td>0.34909</td>
</tr>
<tr>
<td>2</td>
<td>1972</td>
<td>51.71</td>
<td>-3.5</td>
<td>48.8713</td>
<td>2.8387</td>
<td>0.37550</td>
<td>3.7773</td>
<td>0.35438</td>
</tr>
<tr>
<td>3</td>
<td>1973</td>
<td>60.17</td>
<td>-2.5</td>
<td>55.2229</td>
<td>4.9471</td>
<td>0.62485</td>
<td>6.0020</td>
<td>0.59930</td>
</tr>
<tr>
<td>4</td>
<td>1974</td>
<td>64.83</td>
<td>-1.5</td>
<td>61.5745</td>
<td>3.2555</td>
<td>0.39960</td>
<td>3.7302</td>
<td>0.37758</td>
</tr>
<tr>
<td>5</td>
<td>1975</td>
<td>65.24</td>
<td>-0.5</td>
<td>67.9262</td>
<td>-2.6862</td>
<td>-0.32524</td>
<td>-2.9947</td>
<td>-0.30626</td>
</tr>
<tr>
<td>6</td>
<td>1976</td>
<td>65.17</td>
<td>0.5</td>
<td>74.2778</td>
<td>-9.1078</td>
<td>-1.10275</td>
<td>-10.1540</td>
<td>-1.12017</td>
</tr>
<tr>
<td>7</td>
<td>1977</td>
<td>67.65</td>
<td>1.5</td>
<td>80.6295</td>
<td>-12.9795</td>
<td>-1.59320</td>
<td>-14.8723</td>
<td>-1.80365</td>
</tr>
<tr>
<td>8</td>
<td>1978</td>
<td>79.80</td>
<td>2.5</td>
<td>86.9811</td>
<td>-7.1811</td>
<td>-0.90702</td>
<td>-8.7124</td>
<td>-0.89574</td>
</tr>
<tr>
<td>9</td>
<td>1979</td>
<td>96.13</td>
<td>3.5</td>
<td>93.3327</td>
<td>2.7973</td>
<td>0.37001</td>
<td>3.7222</td>
<td>0.34912</td>
</tr>
<tr>
<td>10</td>
<td>1980</td>
<td>115.19</td>
<td>4.5</td>
<td>99.6844</td>
<td>15.5056</td>
<td>2.19772</td>
<td>23.6892</td>
<td>3.26579</td>
</tr>
</tbody>
</table>
Linear Fit Discussion

- Pattern of residuals, together with big improvement in moving to a cubic model (as measured by the drop in ESS), convinces us that linear fit is bad.
- Leverages not too large
- Internally studentized residuals are mostly acceptable though the 2.2 for 1980 is a bit big.
- Externally standard residual for 1980 is really much too big.
### Cubic Fit

<table>
<thead>
<tr>
<th>OBS</th>
<th>YEAR</th>
<th>COST</th>
<th>CODE</th>
<th>LEVERAGE</th>
<th>FITTED</th>
<th>RESID</th>
<th>ISR</th>
<th>PRESS</th>
<th>ESR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1971</td>
<td>45.13</td>
<td>-4.5</td>
<td>0.82378</td>
<td>43.972</td>
<td>1.15814</td>
<td>1.21745</td>
<td>6.57198</td>
<td>1.28077</td>
</tr>
<tr>
<td>2</td>
<td>1972</td>
<td>51.71</td>
<td>-3.5</td>
<td>0.30163</td>
<td>54.404</td>
<td>-2.69386</td>
<td>-1.42251</td>
<td>-3.85737</td>
<td>-1.59512</td>
</tr>
<tr>
<td>3</td>
<td>1973</td>
<td>60.17</td>
<td>-2.5</td>
<td>0.32611</td>
<td>60.029</td>
<td>0.14061</td>
<td>0.07559</td>
<td>0.20865</td>
<td>0.06903</td>
</tr>
<tr>
<td>4</td>
<td>1974</td>
<td>64.83</td>
<td>-1.5</td>
<td>0.30746</td>
<td>62.651</td>
<td>2.17852</td>
<td>1.15521</td>
<td>3.14570</td>
<td>1.19591</td>
</tr>
<tr>
<td>5</td>
<td>1975</td>
<td>65.24</td>
<td>-0.5</td>
<td>0.24103</td>
<td>64.073</td>
<td>1.16683</td>
<td>0.59104</td>
<td>1.53738</td>
<td>0.55597</td>
</tr>
<tr>
<td>6</td>
<td>1976</td>
<td>65.17</td>
<td>0.5</td>
<td>0.24103</td>
<td>66.098</td>
<td>-0.92752</td>
<td>-0.46981</td>
<td>-1.22205</td>
<td>-0.43699</td>
</tr>
<tr>
<td>7</td>
<td>1977</td>
<td>67.65</td>
<td>1.5</td>
<td>0.30746</td>
<td>70.528</td>
<td>-2.87752</td>
<td>-1.52587</td>
<td>-4.15503</td>
<td>-1.78061</td>
</tr>
<tr>
<td>8</td>
<td>1978</td>
<td>79.80</td>
<td>2.5</td>
<td>0.32611</td>
<td>79.166</td>
<td>0.63372</td>
<td>0.34066</td>
<td>0.94039</td>
<td>0.31403</td>
</tr>
<tr>
<td>9</td>
<td>1979</td>
<td>96.13</td>
<td>3.5</td>
<td>0.30163</td>
<td>93.817</td>
<td>2.31320</td>
<td>1.22150</td>
<td>3.31229</td>
<td>1.28644</td>
</tr>
<tr>
<td>10</td>
<td>1980</td>
<td>115.19</td>
<td>4.5</td>
<td>0.82378</td>
<td>116.282</td>
<td>-1.09214</td>
<td>-1.14807</td>
<td>-6.19746</td>
<td>-1.18642</td>
</tr>
</tbody>
</table>

Now the fit is generally ok with all the standardized residuals being fine. Notice the large leverages for the end points, 1971 and 1980.
## Quintic Fit

<table>
<thead>
<tr>
<th>OBS</th>
<th>YEAR</th>
<th>COST</th>
<th>CODE</th>
<th>LEVERAGE</th>
<th>FITTED</th>
<th>RESID</th>
<th>ISR</th>
<th>PRESS</th>
<th>ESR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1971</td>
<td>45.13</td>
<td>-4.5</td>
<td>0.98322</td>
<td>45.127</td>
<td>0.00312</td>
<td>0.03977</td>
<td>0.18583</td>
<td>0.03445</td>
</tr>
<tr>
<td>2</td>
<td>1972</td>
<td>51.71</td>
<td>-3.5</td>
<td>0.72214</td>
<td>51.699</td>
<td>0.01090</td>
<td>0.03417</td>
<td>0.03924</td>
<td>0.02960</td>
</tr>
<tr>
<td>3</td>
<td>1973</td>
<td>60.17</td>
<td>-2.5</td>
<td>0.42844</td>
<td>60.232</td>
<td>-0.06161</td>
<td>-0.13462</td>
<td>-0.10780</td>
<td>-0.11685</td>
</tr>
<tr>
<td>4</td>
<td>1974</td>
<td>64.83</td>
<td>-1.5</td>
<td>0.46573</td>
<td>64.784</td>
<td>0.04641</td>
<td>0.10487</td>
<td>0.08686</td>
<td>0.09095</td>
</tr>
<tr>
<td>5</td>
<td>1975</td>
<td>65.24</td>
<td>-0.5</td>
<td>0.40047</td>
<td>65.228</td>
<td>0.01181</td>
<td>0.02520</td>
<td>0.01970</td>
<td>0.02183</td>
</tr>
<tr>
<td>6</td>
<td>1976</td>
<td>65.17</td>
<td>0.5</td>
<td>0.40047</td>
<td>64.925</td>
<td>0.24502</td>
<td>0.52270</td>
<td>0.40868</td>
<td>0.46897</td>
</tr>
<tr>
<td>7</td>
<td>1977</td>
<td>67.65</td>
<td>1.5</td>
<td>0.46573</td>
<td>68.392</td>
<td>-0.74249</td>
<td>-1.67794</td>
<td>-1.38974</td>
<td>-2.67034</td>
</tr>
<tr>
<td>8</td>
<td>1978</td>
<td>79.80</td>
<td>2.5</td>
<td>0.42844</td>
<td>78.981</td>
<td>0.81942</td>
<td>1.79036</td>
<td>1.43365</td>
<td>3.47878</td>
</tr>
<tr>
<td>9</td>
<td>1979</td>
<td>96.13</td>
<td>3.5</td>
<td>0.72214</td>
<td>96.543</td>
<td>-0.41296</td>
<td>-1.29407</td>
<td>-1.48622</td>
<td>-1.46985</td>
</tr>
<tr>
<td>10</td>
<td>1980</td>
<td>115.19</td>
<td>4.5</td>
<td>0.98322</td>
<td>115.110</td>
<td>0.08038</td>
<td>1.02486</td>
<td>4.78917</td>
<td>1.03356</td>
</tr>
</tbody>
</table>
Conclusions

- Leverages at the end are very high.
- Although fit is good, residuals at 1977 and 1978 are definitely too big.
- Overall cubic fit is preferred but does not provide reliable forecasts nor a meaningful physical description of the data.
- A good model would somehow involve economic theory and covariates, though there is really very little data to fit such models.
PRESS residuals

- Suggestion:
  \[ Y_i - \hat{Y}_{i(i)} \]

  where \( \hat{Y}_{i(i)} \) is the fitted value using all the data except case \( i \).
- This residual is called a “PRESS prediction error for case \( i \)”.
- The acronym PRESS stands for Prediction Sum of Squares.
- But: \( Y_i - \hat{Y}_{i(i)} \) must be compared to other residuals or to \( \sigma \).
- So we suggest Externally Studentized Residuals which are also called Case Deleted Residuals:
  \[
  \hat{\epsilon}_{i(i)} \frac{Y_i - \hat{Y}_{i(i)}}{\text{est'd SE not using case } i} = \frac{Y_i - \hat{Y}_{i(i)}}{\text{Case } i \text{ deleted SE of numerator}}
  \]
Computing Externally Standardized Residuals

- Apparent problem: If $n = 100$ do I have to run SAS 100 times? NO.

- **FACT 1:**

  $$Y_i - \hat{Y}_{i(i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}$$

- Recall jargon: $h_{ii}$ is the **leverage** of point $i$.

- If $h_{ii}$ is large then

  $$\left| \frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right| \gg |\hat{\varepsilon}_i|$$

  and point $i$ influences the fit strongly.

- **FACT 2:**

  $$\text{Var} \left( \frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right) = \frac{\sigma^2 (1 - h_{ii})}{(1 - h_{ii})^2}$$
Externally Standardized Residuals Continued

- The Externally Standardized Residual is

\[
\frac{\hat{\epsilon}_i/(1 - h_{ii})}{\sqrt{\text{MSE}(i)/(1 - h_{ii})}} = \frac{\hat{\epsilon}_i}{\sqrt{\text{MSE}(i)(1 - h_{ii})}}
\]

where

\[
\text{MSE}(i) = \text{estimate of } \sigma^2 \text{ not using data point } i
\]

- Fact:

\[
\text{MSE} = \frac{(n - p - 1)\text{MSE}(i) + \hat{\epsilon}_i^2/(1 - h_{ii})}{n - p}
\]

so the externally studentized residual is

\[
\hat{\epsilon}_i \sqrt{\frac{n - p - 1}{\text{ESS}(1 - h_{ii}) - \hat{\epsilon}_i^2}}
\]
Distribution Theory of Externally Standardized Residuals

1. \( \hat{\epsilon}_{(i)}/\sqrt{\text{Var}(\hat{\epsilon}_i)} \sim N(0, 1) \)

2. \( \frac{(n - p - 1)\text{MSE}_i}{\sigma^2} \sim \chi^2_{n-p-1} \)

3. These two are independent.

4. SO:

\[
t_i = \frac{(n - p - 1)\text{MSE}_i}{\sigma^2} \sim \chi^2_{n-p-1} \\
\sim t_{n-p-1}
\]
Example: Insurance Data

Cubic Fit:

<table>
<thead>
<tr>
<th>Year</th>
<th>$\hat{\epsilon}_i$</th>
<th>Internally Studentized PRESS</th>
<th>Externally Studentized</th>
<th>Leverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td>1.17</td>
<td>0.59</td>
<td>1.54</td>
<td>0.56</td>
</tr>
<tr>
<td>1980</td>
<td>-1.09</td>
<td>-1.15</td>
<td>-6.20</td>
<td>-1.19</td>
</tr>
</tbody>
</table>

- Note the influence of the leverage.
- Note that edge observations (1980) have large leverage.
## Quintic Fit

<table>
<thead>
<tr>
<th>Year</th>
<th>$\hat{\epsilon}_i$</th>
<th>Internally Studentized</th>
<th>PRESS</th>
<th>Externally Studentized</th>
<th>Leverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>0.82</td>
<td>1.79</td>
<td>1.43</td>
<td>3.48</td>
<td>0.43</td>
</tr>
<tr>
<td>1980</td>
<td>0.08</td>
<td>1.02</td>
<td>4.79</td>
<td>1.03</td>
<td>0.98</td>
</tr>
</tbody>
</table>

- Notice 1978 residual is unacceptably large.
- Notice 1980 leverage is huge.
Formal assessment of Externally Standardized Residuals

1. Each residual has a $t_{n-p-1}$ distribution.
2. For example, for the quintic, $t_{10-7,0.025} = 3.18$ is the critical point for a 5% level test.
3. But there are 10 residuals to look at.
4. This leads to a multiple comparisons problem.
5. The simplest multiple comparisons procedure is the Bonferroni method: divide $\alpha$ by the number of tests to be done, 10 in our case giving $0.025/10 = 0.0025$.
6. The corresponding critical point is
   
   $$t_{3,0.0025} = 7.45$$

so none of the residuals are significant.
7. For the cubic model

   $$t_{5,0.0025} = 4.77$$

and again all the residuals are judged ok.