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

$$\hat{\epsilon} = Y - X\hat{\beta}.$$

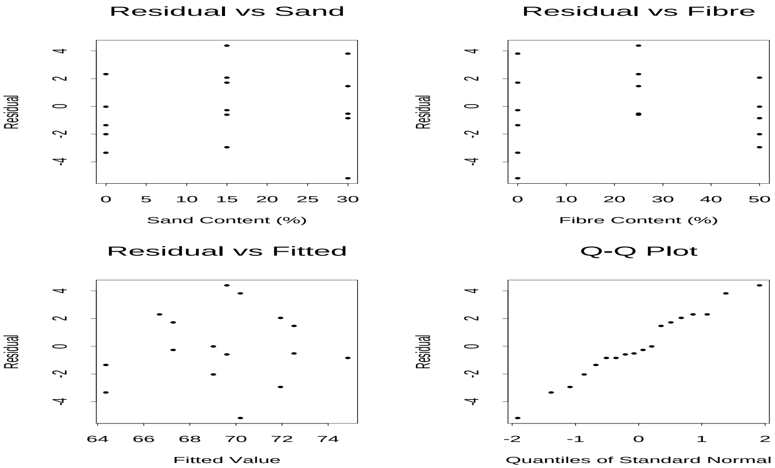
- Examine residual plots to assess quality of model.
- Plot residuals $\hat{\epsilon}_i$ against each x_i, i.e. against S_i and F_i.
- Plot residuals against other covariates, particularly those deleted from model.
- Plot residuals against $\hat{\mu}_i$ = fitted value.
- Plot residuals squared against 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:



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

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

- 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 β_4 to test $H_o : \beta_4 = 0$ in second model.
- We find

$$\begin{aligned}\hat\beta_4 &= -0.00373 \\ \hat\sigma_{\hat\beta_4} &= 0.001995 \\ t &= \frac{-0.00373}{0.001995} = -1.87\end{aligned}$$

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 iith 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 σ with the obvious estimate and find that

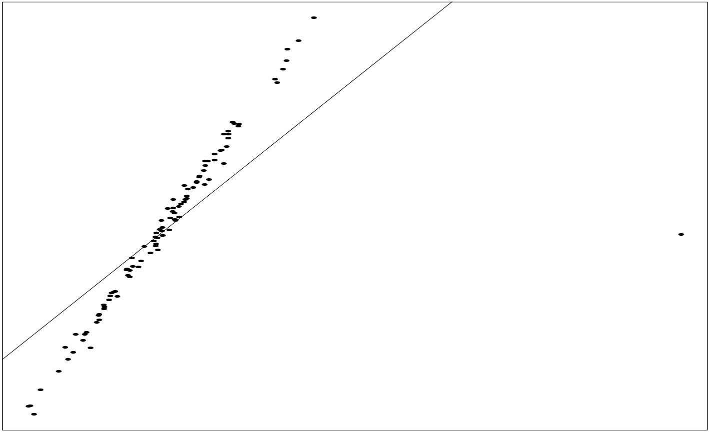
$$\frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \sim N(0, 1)$$

  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 σ 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;
```

Linear Fit Output

OBS	YEAR	COST	CODE	LEVERAGE	FITTED	RESID	ISR	PRESS	ESR
1	1971	45.13	-4.5	0.34545	42.5196	2.6104	0.36998	3.9881	0.34909
2	1972	51.71	-3.5	0.24848	48.8713	2.8387	0.37550	3.7773	0.35438
3	1973	60.17	-2.5	0.17576	55.2229	4.9471	0.62485	6.0020	0.59930
4	1974	64.83	-1.5	0.12727	61.5745	3.2555	0.39960	3.7302	0.37758
5	1975	65.24	-0.5	0.10303	67.9262	-2.6862	-0.32524	-2.9947	-0.30626
6	1976	65.17	0.5	0.10303	74.2778	-9.1078	-1.10275	-10.1540	-1.12017
7	1977	67.65	1.5	0.12727	80.6295	-12.9795	-1.59320	-14.8723	-1.80365
8	1978	79.80	2.5	0.17576	86.9811	-7.1811	-0.90702	-8.7124	-0.89574
9	1979	96.13	3.5	0.24848	93.3327	2.7973	0.37001	3.7222	0.34912
10	1980	115.19	4.5	0.34545	99.6844	15.5056	2.19772	23.6892	3.26579

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

OBS	YEAR	COST	CODE	LEVERAGE	FITTED	RESID	ISR	PRESS	ESR
1	1971	45.13	-4.5	0.82378	43.972	1.15814	1.21745	6.57198	1.28077
2	1972	51.71	-3.5	0.30163	54.404	-2.69386	-1.42251	-3.85737	-1.59512
3	1973	60.17	-2.5	0.32611	60.029	0.14061	0.07559	0.20865	0.06903
4	1974	64.83	-1.5	0.30746	62.651	2.17852	1.15521	3.14570	1.19591
5	1975	65.24	-0.5	0.24103	64.073	1.16683	0.59104	1.53738	0.55597
6	1976	65.17	0.5	0.24103	66.098	-0.92750	-0.46981	-1.22205	-0.43699
7	1977	67.65	1.5	0.30746	70.528	-2.87752	-1.52587	-4.15503	-1.78061
8	1978	79.80	2.5	0.32611	79.166	0.63372	0.34066	0.94039	0.31403
9	1979	96.13	3.5	0.30163	93.817	2.31320	1.22150	3.31229	1.28644
10	1980	115.19	4.5	0.82378	116.282	-1.09214	-1.14807	-6.19746	-1.18642

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

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

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 σ
- So we suggest **Externally Studentized Residuals** which are also called **Case Deleted Residuals**:

$$\frac{\hat{\epsilon}_{i(i)}}{\text{est'd SE not using case } i} = \frac{Y_i - \hat{Y}_{i(i)}}{\text{Case } i \text{ deleted SE of numerator}}$$

Computing Externally Standardized Residuals

- Apparent problem: If $n = 100$ do I have to run SAS 100 times? NO.
- **FACT 1**:
$$Y_i - \hat{Y}_{i(i)} = \frac{\hat{\epsilon}_i}{1 - h_{ii}}$$
- Recall jargon: h_{ii} is the **leverage** of point i.
- If h_{ii} is large then
$$\left| \frac{\hat{\epsilon}_i}{1 - h_{ii}} \right| >> |\hat{\epsilon}_i|$$
and point i influences the fit strongly.
- **FACT 2**:
$$\mathrm{Var}\left(\frac{\hat{\epsilon}_i}{1 - h_{ii}} \right) = \frac{\sigma^2}{1 - h_{ii}} \left(= \frac{\sigma^2 (1 - h_{ii})}{(1 - h_{ii})^2} \right)$$

Externally Standardized Residuals Continued

- The Externally Standardized Residual is

$$\frac{\hat{\epsilon}_i/(1-h_{ii})}{\sqrt{\mathrm{MSE}_{(i)}/(1-h_{ii})}} = \frac{\hat{\epsilon}_i}{\sqrt{\mathrm{MSE}_{(i)}(1-h_{ii})}}$$

  where

$$\mathrm{MSE}_{(i)} = \text{ estimate of } \sigma^2 \text{ not using data point } i$$

- Fact:

$$\mathrm{MSE} = \frac{(n-p-1)\mathrm{MSE}_{(i)} + \hat{\epsilon}_i^2/(1-h_{ii})}{n-p}$$

  so the externally studentized residual is

$$\hat{\epsilon}_i\sqrt{\frac{n-p-1}{\mathrm{ESS}(1-h_{ii})-\hat{\epsilon}_i^2}}$$

Distribution Theory of Externally Standardized Residuals

1. $\hat\epsilon_{(i)}/\sqrt{\mathrm{Var}(\hat\epsilon_i)} \sim N(0,1)$
2. $$\frac{(n-p-1)\mathrm{MSE}_{(i)}}{\sigma^2} \sim \chi^2_{n-p-1}$$
3. These two are independent.
4. SO:

$$t_i = \frac{(n-p-1)\mathrm{MSE}_{(i)}}{\sigma^2} \sim \chi^2_{n-p-1}$$
$$\sim t_{n-p-1}$$

Example: Insurance Data

Cubic Fit:

Year	$\hat{\epsilon}_i$	Internally Studentized	PRESS	Externally Studentized	Leverage
1975	1.17	0.59	1.54	0.56	0.24
1980	-1.09	-1.15	-6.20	-1.19	0.82

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

Quintic Fit

Year	$\hat{\epsilon}_i$	Internally Studentized	PRESS	Externally Studentized	Leverage
1978	0.82	1.79	1.43	3.48	0.43
1980	0.08	1.02	4.79	1.03	0.98

- 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 α 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.