Checking Assumptions

Merlise Clyde

STA721 Linear Models

Duke University

November 20, 2017

Linear Model

Linear Model:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

Assumptions:

$$\begin{aligned} \boldsymbol{\mu} \in C(\mathbf{X}) \quad &\Leftrightarrow \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \\ \boldsymbol{\epsilon} \quad &\sim \quad \mathsf{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n) \end{aligned}$$

Focus on

- Wrong mean for a case or cases
- Wrong distribution for $\boldsymbol{\epsilon}$
- Cases that influence the mean

If $\mu_i \neq \mathbf{x}_i^T \boldsymbol{\beta}$ then expected value of $e_i = Y_i - \hat{Y}_i$ is not zero;

Standardized residuals

standardized residuals

$$r_i = e_i/\sqrt{\sigma^2(1 - h_{ii})}$$

- Under correct model have mean 0 and scale 1
- Use estimate of σ^2
- The distribution of r_i is not a t
- if h_{ii} (leverage) is close to 1, then $\hat{Y}_i$ is close to Y_i so e_i is approximately 0
- Variance is also almost 0, so standardize value may not flag "outliers"

Predicted Residuals

Estimates without Case (i):

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(i)} &= (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^T \mathbf{Y}_{(i)} \\ &= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i e_i}{1 - h_{ii}}\end{aligned}$$

Predicted residual

$$e_{(i)} = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(i)} = \frac{e_i}{1 - h_{ii}}$$

with variance

$$\text{var}(e_{(i)}) = \frac{\sigma^2}{1 - h_{ii}}$$

Standardized predicted residual is

$$\frac{e_{(i)}}{\sqrt{\text{var}(e_{(i)})}} = \frac{e_i/(1 - h_{ii})}{\sigma/\sqrt{1 - h_{ii}}} = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}}$$

these are the same as standardized residual!

External estimate of σ^2

Estimate $\hat{\sigma}^2_{(i)}$ using data with case i deleted

$$
\begin{aligned}
\mathrm{SSE}_{(i)} &= \mathrm{SSE} - \frac{e_i^2}{1 - h_{ii}} \\
\hat{\sigma}^2_{(i)} = \mathrm{MSE}_{(i)} &= \frac{\mathrm{SSE}_{(i)}}{n - p - 1}
\end{aligned}
$$

Externally Standardized residuals

$$
t_i = \frac{e_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1 - h_{ii})}} = \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1 - h_{ii})}} = r_i \left(\frac{n - p - 1}{n - p - r_i^2} \right)^{1/2}
$$

May still miss extreme points with high leverage, but will pick up unusual y_is

Distribution of Externally Studentized Residual

$$t_i = \frac{e_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1 - h_{ii})}} = \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1 - h_{ii})}} \sim \mathsf{St}(n - p - 1)$$

Outlier Test

H_0: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ versus H_a: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha_i$

- Show that t-test for testing H_0: $\alpha_i = 0$ is equal to t_i
- if p-value is small declare the ith case to be an outlier: $E[Y_i]$ not given by $\mathbf{X}\beta$ but $\mathbf{X}\beta + \delta_i \alpha_i$
- Can extend to include multiple δ_i and δ_j to test that case i and j are both outliers
- Extreme case $\boldsymbol{\mu} = \mathbf{X}\beta + \mathbf{I}_n \boldsymbol{\alpha}$ all points have their own mean!
- Control for multiple testing

Multiple Testing

look to see if the max $|t_i|$ is larger than expected under the null

- Need distribution of the max of Student t random variables (simulate)
- Conservative is Bonferroni Correction: For n tests of size α the probability of falsely labeling at least one case as an outlier is no greater than $n\alpha$; e.g. with 21 cases and $\alpha = 0.05$, the probability is no greater than 1.05!
- Adjust $\alpha^* = \alpha/n$ so that the probability of falsely labeling at least one point an outlier is α $\alpha/n = .00238$ with $n = 21$

Cook's Distance

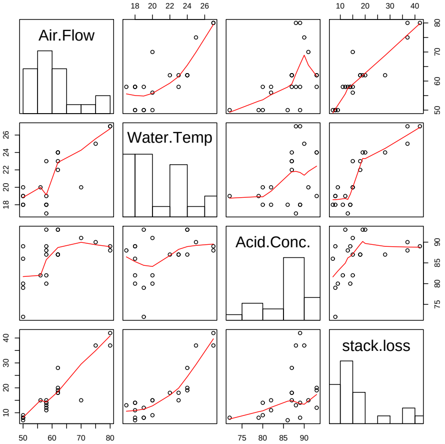
Cook's Distance measure of how much predictions change with ith case deleted

$$
\begin{aligned}
D_i &= \frac{\|\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}}\|^2}{p\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} \\
&= \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}
\end{aligned}
$$

Flag cases where $D_i > 1$ or large relative to other cases

Influential Cases

Stackloss Data



Case 21

- Leverage 0.285 (compare to $p/n = .19$)
- p-value t_{21} is 0.0042
- Bonferroni adjusted p-value is 0.0024
- Cooks' Distance .69
- Other points? Masking?
- Refit without Case 21

Other analyses have suggested that cases (1, 2, 3, 4, 21) are outliers

Multiple Outliers

- Hoeting, Madigan and Raftery (in various permutations) consider the problem of simultaneous variable selection and outlier identification.
- This is implemented in the library(BMA) in the function MC3.REG. This has the advantage that more than 2 points may be considered as outliers at the same time.
- The function uses a Markov chain to identify both important variables and potential outliers, but is coded in Fortran so should run reasonably quickly.
- Can also use BAS or other variable selection programs

Using BAS

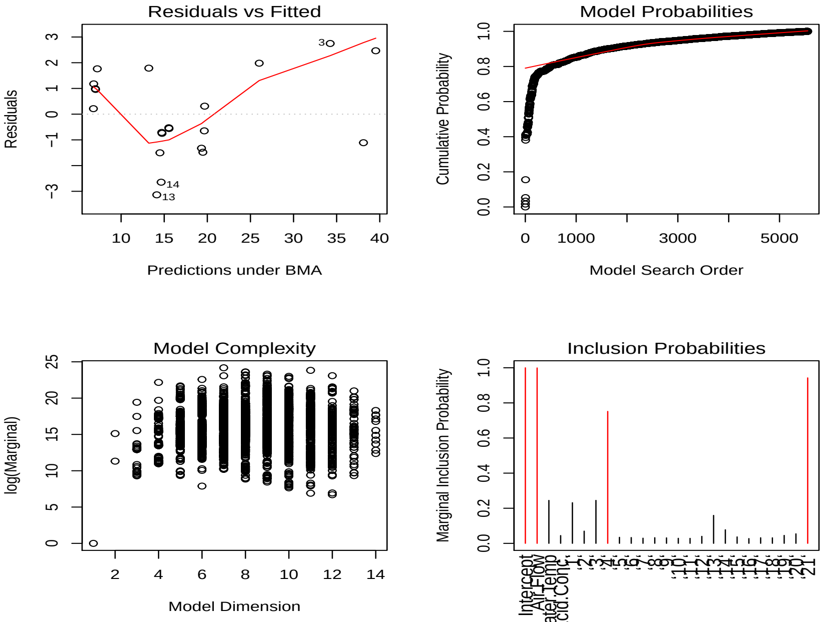
```
library(MASS)
data(stackloss)
n = nrow(stackloss)
stack.out = cbind(stackloss, diag(n))

library(BAS)
BAS.stack = bas.lm(stack.loss ~ ., data=stack.out,
                   prior="hyper-g-n", a=3,
                   modelprior=tr.beta.binomial(1, 1,15) ,
                   method="MCMC", MCMC.it=200000)
```

Output

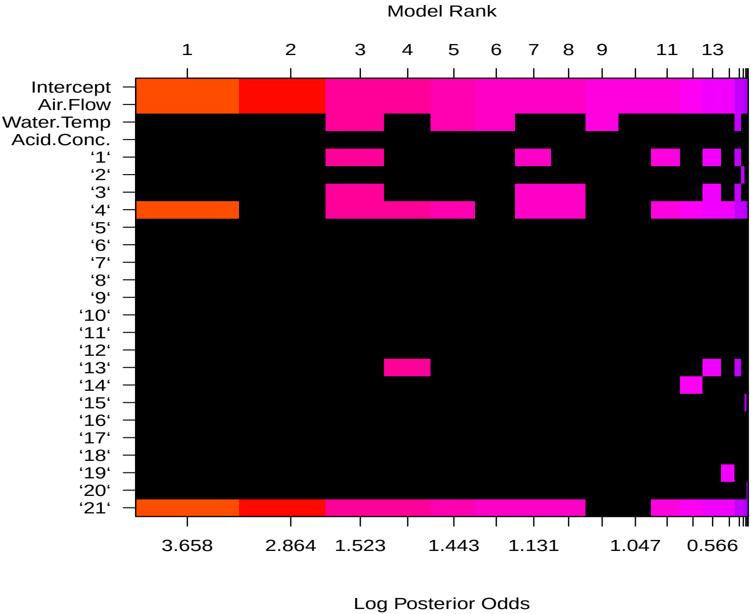
	P(B != 0 | Y)	model 1	model 2	model 3	model 4	model 5
Intercept	1.00	1.00	1.00	1.00	1.00	1.00
Air.Flow	1.00	1.00	1.00	1.00	1.00	1.00
Water.Temp	0.24	0.00	0.00	1.00	0.00	1.00
Acid.Conc.	0.04	0.00	0.00	0.00	0.00	0.00
‘1‘	0.23	0.00	0.00	1.00	0.00	0.00
‘2‘	0.07	0.00	0.00	0.00	0.00	0.00
‘3‘	0.24	0.00	0.00	1.00	0.00	0.00
‘4‘	0.75	1.00	0.00	1.00	1.00	1.00
‘5‘	0.03	0.00	0.00	0.00	0.00	0.00
‘6‘	0.03	0.00	0.00	0.00	0.00	0.00
‘7‘	0.03	0.00	0.00	0.00	0.00	0.00
‘8‘	0.03	0.00	0.00	0.00	0.00	0.00
‘9‘	0.03	0.00	0.00	0.00	0.00	0.00
‘10‘	0.03	0.00	0.00	0.00	0.00	0.00
‘11‘	0.03	0.00	0.00	0.00	0.00	0.00
‘12‘	0.04	0.00	0.00	0.00	0.00	0.00
‘13‘	0.16	0.00	0.00	0.00	1.00	0.00
‘14‘	0.08	0.00	0.00	0.00	0.00	0.00
‘15‘	0.04	0.00	0.00	0.00	0.00	0.00
‘16‘	0.03	0.00	0.00	0.00	0.00	0.00
‘17‘	0.03	0.00	0.00	0.00	0.00	0.00
‘18‘	0.03	0.00	0.00	0.00	0.00	0.00
‘19‘	0.05	0.00	0.00	0.00	0.00	0.00
‘20‘	0.05	0.00	0.00	0.00	0.00	0.00
‘21‘	0.94	1.00	1.00	1.00	1.00	1.00
BF		0.13	0.01	1.00	0.08	0.07
PostProbs		0.23	0.10	0.03	0.03	0.02
R2		0.96	0.93	0.99	0.97	0.97
dim		4.00	3.00	7.00	5.00	5.00
logmarg		22.17	19.43	24.18	21.68	21.57

BAS



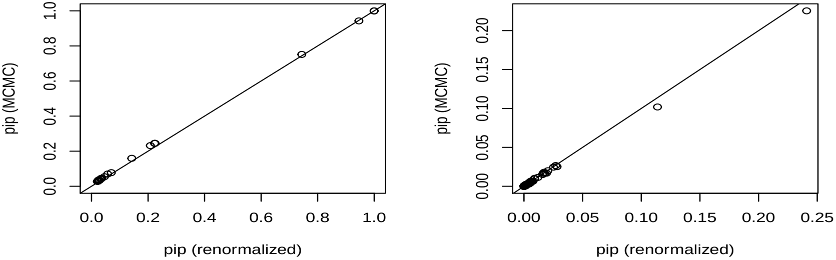
Residuals vs Fitted
Residuals
Predictions under BMA
Model Probabilities
Cumulative Probability
Model Search Order
Model Complexity
log(Marginal)
Model Dimension
Inclusion Probabilities
Marginal Inclusion Probability

BAS Model Space



duke.eps

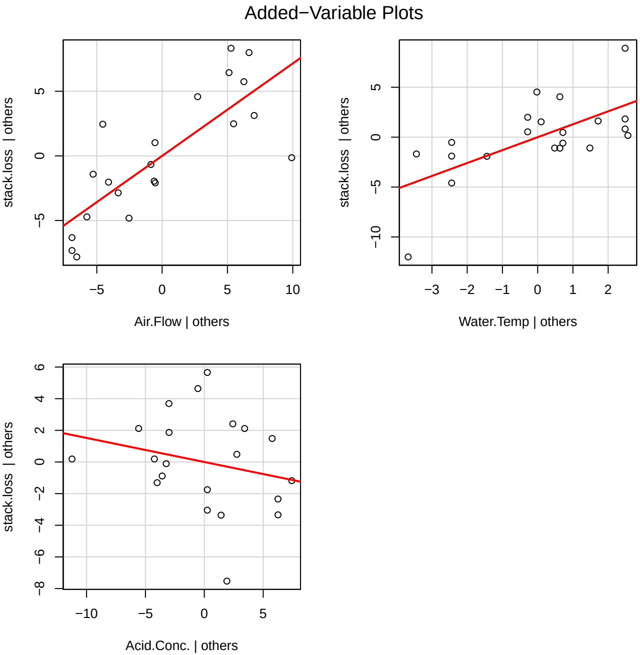
Diagnostics



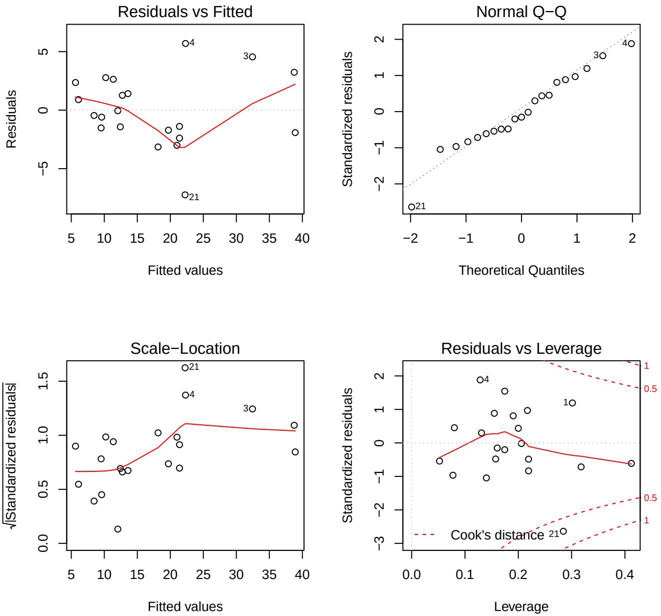
To Remove or Not?

- For suspicious cases, check data sources for errors
- Check that points are not outliers because of wrong mean function or distributional assumptions
- Investigate need for transformations (use EDA at several stages)
- Influential cases - report results with and without cases (results may change - are differences meaningful?)
- Outlier test - suggests alternative population for the case(s); if not influential may in keep analysis, but will inflate $\hat{\sigma}^2$ and interval estimates
- Document how you handle any case deletions - reproducibility!
- Robust Regression Methods

Stackloss Added Variable Plot



Stackloss Data Again



duke.eps