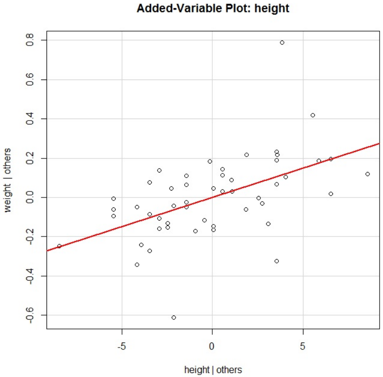
Stat GR 5025 Lecture 9

Jingchen Liu

Department of Statistics
Columbia University

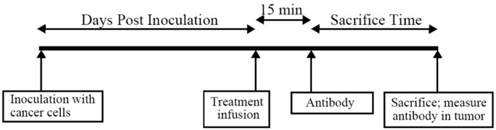
Value-added plot



Outlier detection

Display 11.3 **p. 307**

Time line for blood-brain barrier disruption experiment



Days Post Inoculation

15 min

Sacrifice Time

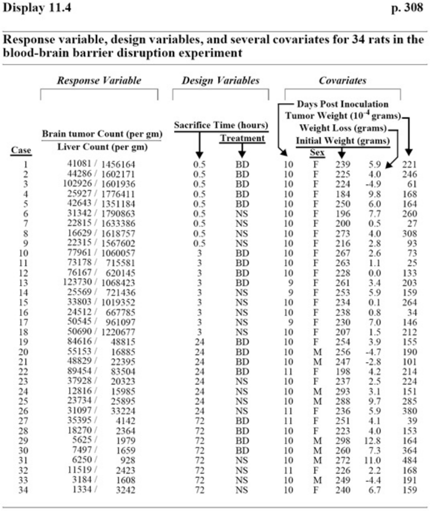
Inoculation with cancer cells

Treatment infusion

Antibody

Sacrifice; measure antibody in tumor

Outlier detection

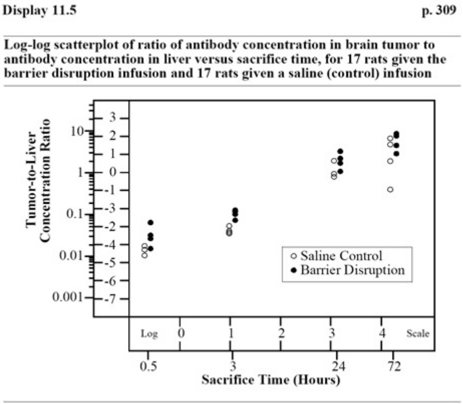


Display 11.4 **p. 308**

Response variable, design variables, and several covariates for 34 rats in the blood-brain barrier disruption experiment

	Response Variable	*Design Variables*		*Covariates*				
Case	Brain tumor Count (per gm) / Liver Count (per gm)	Sacrifice Time (hours)	Treatment	Days Post Inoculation	Sex	Initial Weight (grams)	Weight Loss (grams)	Tumor Weight (10^{-4} grams)
1	41081 / 1456164	0.5	BD	10	F	239	5.9	221
2	44286 / 1602171	0.5	BD	10	F	225	4.0	246
3	102926 / 1601936	0.5	BD	10	F	224	-4.9	61
4	25927 / 1776411	0.5	BD	10	F	184	9.8	168
5	42643 / 1351184	0.5	BD	10	F	250	6.0	164
6	31342 / 1790863	0.5	NS	10	F	196	7.7	260
7	22815 / 1633386	0.5	NS	10	F	200	0.5	27
8	16629 / 1618757	0.5	NS	10	F	273	4.0	308
9	22315 / 1567602	0.5	NS	10	F	216	2.8	93
10	77961 / 1060057	3	BD	10	F	267	2.6	73
11	73178 / 715581	3	BD	10	F	263	1.1	25
12	76167 / 620145	3	BD	10	F	228	0.0	133
13	123730 / 1068423	3	BD	9	F	261	3.4	203
14	25569 / 721436	3	NS	9	F	253	5.9	159
15	33803 / 1019352	3	NS	10	F	234	0.1	264
16	24512 / 667785	3	NS	10	F	238	0.8	34
17	50545 / 961097	3	NS	9	F	230	7.0	146
18	50690 / 1220677	3	NS	10	F	207	1.5	212
19	84616 / 48815	24	BD	10	F	254	3.9	155
20	55153 / 16885	24	BD	10	M	256	-4.7	190
21	48829 / 22395	24	BD	10	M	247	-2.8	101
22	89454 / 83504	24	BD	11	F	198	4.2	214
23	37928 / 20323	24	NS	10	F	237	2.5	224
24	12816 / 15985	24	NS	10	M	293	3.1	151
25	23734 / 25895	24	NS	10	M	288	9.7	285
26	31097 / 33224	24	NS	11	F	236	5.9	380
27	35395 / 4142	72	BD	11	F	251	4.1	39
28	18270 / 2364	72	BD	10	F	223	4.0	153
29	5625 / 1979	72	BD	10	M	298	12.8	164
30	7497 / 1659	72	BD	10	M	260	7.3	364
31	6250 / 928	72	NS	10	M	272	11.0	484
32	11519 / 2423	72	NS	11	F	226	2.2	168
33	3184 / 1608	72	NS	10	M	249	-4.4	191
34	1334 / 3242	72	NS	10	F	240	6.7	159

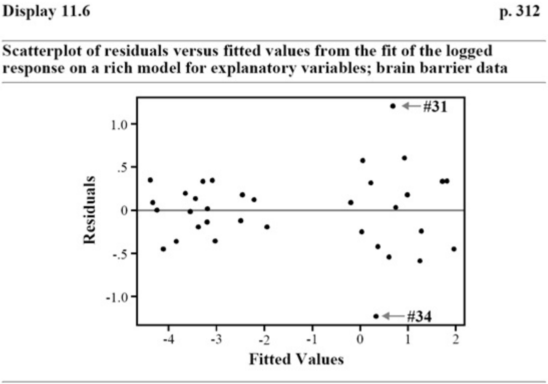
Outlier detection



Display 11.5 **p. 309**

Log-log scatterplot of ratio of antibody concentration in brain tumor to antibody concentration in liver versus sacrifice time, for 17 rats given the barrier disruption infusion and 17 rats given a saline (control) infusion

Outlier detection



Display 11.6 p. 312

Scatterplot of residuals versus fitted values from the fit of the logged response on a rich model for explanatory variables; brain barrier data

Deleted residual

$$d_i = y_i - \hat{y}_{i(i)} = \frac{y_i - \hat{y}_i}{1 - h_i} \quad Var(d_i) = \frac{\sigma^2}{1 - h_i}$$

Studentized residual

- Studentized residual

$$StudRes_i = \frac{d_i}{SE(d_i)}$$

- Another representation

$$StudRes_i = \frac{y_i - \hat{y}_i}{SE(y_i - \hat{y}_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

- About the hat matrix

Studentized residual

- Studentized residual

$$StudRes_i = \frac{d_i}{SE(d_i)}$$

- Another representation

$$StudRes_i = \frac{y_i - \hat{y}_i}{SE(y_i - \hat{y}_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

- About the hat matrix

Studentized residual

- Studentized residual

$$StudRes_i = \frac{d_i}{SE(d_i)}$$

- Another representation

$$StudRes_i = \frac{y_i - \hat{y}_i}{SE(y_i - \hat{y}_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

- About the hat matrix

Leverage

- About leverage
- Simple linear model

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

- Multiple regression: $\mathrm{diag}(X(X^\top X)^{-1}X^\top)$
- Total leverage

Leverage

- About leverage
- Simple linear model

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

- Multiple regression: $\operatorname{diag}(X(X^\top X)^{-1}X^\top)$
- Total leverage

Leverage

- About leverage
- Simple linear model

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

- Multiple regression: $\mathrm{diag}(X(X^\top X)^{-1} X^\top)$
- Total leverage

Leverage

- About leverage
- Simple linear model

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

- Multiple regression: $\mathrm{diag}(X(X^\top X)^{-1}X^\top)$
- Total leverage

Leave-one-out measure

- Cook's distance

$$D_i = \frac{\sum_j (\hat{y}_j - \hat{y}_{j(i)})^2}{p\sigma^2}$$

- Another representation

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p\sigma^2} \frac{h_i}{(1-h_i)^2} = \frac{StudRes_i^2}{p} \frac{h_i}{1-h_i}$$

Model/variable selection

- What is model/variable selection
- Motivation
- Including redundant explanatory variables reduces prediction power
- Large p small n problem

Model/variable selection

- What is model/variable selection
- Motivation
- Including redundant explanatory variables reduces prediction power
- Large p small n problem

Model/variable selection

- What is model/variable selection
- Motivation
- Including redundant explanatory variables reduces prediction power
- Large p small n problem

Model/variable selection

- What is model/variable selection
- Motivation
- Including redundant explanatory variables reduces prediction power
- Large p small n problem

Applications

- Biointormatics
- Wavelet
- Time series analysis
- Neural network, regression trees,...
- Any time you have an alternative model

Applications

- Bioinformatics
- Wavelet
- Time series analysis
- Neural network, regression trees,...
- Any time you have an alternative model

Applications

- Biolnformatics
- Wavelet
- Time series analysis
- Neural network, regression trees,...
- Any time you have an alternative model

Applications

- Bioinformatics
- Wavelet
- Time series analysis
- Neural network, regression trees,...
- Any time you have an alternative model

Applications

- Bioinformatics
- Wavelet
- Time series analysis
- Neural network, regression trees,...
- Any time you have an alternative model

General questions

- What makes a good model?
  - Small prediction error
  - Large R^2
- Criteria for comparing different models

General questions

- What makes a good model?
  - Small prediction error
  - Large R^2
- Criteria for comparing different models

General questions

- What makes a good model?
  - Small prediction error
  - Large R^2
- Criteria for comparing different models

General questions

- What makes a good model?
  - Small prediction error
  - Large R^2
- Criteria for comparing different models

Some examples

$$H_0 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_3 + \beta_4 x_4} + \varepsilon$$

Some examples

$$H_0 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

$$H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$$

Some examples

$$H_0 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_5 x_5} + \varepsilon$$

$$H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_3 + \beta_4 x_4} + \varepsilon$$

Some principles

- What makes a good model?
  - Smaller estimated errors
  - Simpler model/fewer predictors
  - Prediction of *future* observations
- The least squares estimate only considers small fitted errors?

Candidates

- Coefficient of determination, R^2
- Estimated error level, $\hat{\sigma}^2$
- Adjusted R^2

$$\frac{Var(y) - \hat{\sigma}^2}{Var(y)}$$

Algorithmic approach

- Forward selection
- Backward deletion
- Stepwise regression

Forward selection

- Consider variables that are not in the current model, compute the extra-sum-of-squares by adding each variable.
- If the largest extra-sum-of-squares is greater than some value (e.g., 4), then add that variable in; otherwise stop.

Backward deletion

- Consider variables that are in the current model, compute the extra-sum-of-squares by removing each variable.
- If the smallest extra-sum-of-squares is less than some value (e.g., 4), then remove that variable; otherwise stop.

Stepwise regression

- Do one step forward selection and backward deletion alternatively

Pros and cons

- Easy to implement
- Less computation
- In consistency

Likelihood-based criteria

- Akaike information criterion (AIC)

$$n\log(\hat{\sigma}^2) + 2p.$$

Derive AIC.

- Bayesian information criterion

$$n\log(\hat{\sigma}^2) + p\log(n)$$

General form

$$x_1, ..., x_n \sim f(x|\theta)$$

- Akaike information criterion (AIC)

$$-2\log(L(\hat{\theta})) + 2p$$

- Bayesian information criterion

$$-2\log(L(\hat{\theta})) + p\log(n)$$

General form

- Likelihood

$$L(\theta; x_1, ..., x_n)$$

- Maximum likelihood estimator
- Derivation of AIC

Delimma

- Too few variables (missing the true predictor) – bias.
- Too many variables – variance.

Delimma

- Too few variables (missing the true predictor) – bias.
- Too many variables – variance.

Mallows' C_p

- Let $\mu_i = E(y|x_i)$.
- Mean squared error

$$E[(\hat{y}_i - \mu_i)^2|x_i] = E^2(\hat{y}_i - \mu_i|x_i) + Var(\hat{y}_i - \mu_i|x_i)$$

- Total mean squared error

$$\sum_{i=1}^{n} E[(\hat{y}_i - \mu_i)^2|x_i] = \sum_{i=1}^{n} E^2(\hat{y}_i - \mu_i|x_i) + \sum_{i=1}^{n} Var(\hat{y}_i - \mu_i|x_i)$$

Mallows' C_p

- Let $\mu_i = E(y|x_i)$.
- Mean squared error

$$E[(\hat{y}_i - \mu_i)^2|x_i] = E^2(\hat{y}_i - \mu_i|x_i) + Var(\hat{y}_i - \mu_i|x_i)$$

- Total mean squared error

$$\sum_{i=1}^{n} E[(\hat{y}_i - \mu_i)^2|x_i] = \sum_{i=1}^{n} E^2(\hat{y}_i - \mu_i|x_i) + \sum_{i=1}^{n} Var(\hat{y}_i - \mu_i|x_i)$$

Mallows' C_p

- Let $\mu_i = E(y|x_i)$.
- Mean squared error

$$E[(\hat{y}_i - \mu_i)^2|x_i] = E^2(\hat{y}_i - \mu_i|x_i) + Var(\hat{y}_i - \mu_i|x_i)$$

- Total mean squared error

$$\sum_{i=1}^{n} E[(\hat{y}_i - \mu_i)^2|x_i] = \sum_{i=1}^{n} E^2(\hat{y}_i - \mu_i|x_i) + \sum_{i=1}^{n} Var(\hat{y}_i - \mu_i|x_i)$$

Mallows' C_p

$$C_p = \frac{SSE}{\hat{\sigma}_f^2} - (n - 2p) = p + (n - p)\frac{\hat{\sigma}^2 - \hat{\sigma}_f^2}{\hat{\sigma}_f^2}$$

Example

- Response variable: log-survival time
- Covariates: blood clotting score, prognostic index, enzyme function test score, living function test score, age, gender, alcohol use (none, moderate, heavy)
- AIC

Example

Forward

```
Start:  AIC=-75.7
logsurvival ~ 1

          Df Sum of Sq     RSS      AIC
+ enzyme   1    5.4762  7.3316 -103.827
+ liver    1    5.3990  7.4087 -103.262
+ progind  1    2.8285  9.9792  -87.178
+ heavy    1    1.7798 11.0279  -81.782
+ score    1    0.7763 12.0315  -77.079
+ gender   1    0.6897 12.1180  -76.692
<none>                 12.8077  -75.703
+ age      1    0.2691 12.5386  -74.849
+ alcohol  1    0.2052 12.6025  -74.575
```

```
Step:  AIC=-103.83
logsurvival ~ enzyme

          Df Sum of Sq    RSS     AIC
+ progind  1   3.01908 4.3125 -130.48
+ liver    1   2.20187 5.1297 -121.11
+ score    1   1.55061 5.7810 -114.66
+ heavy    1   1.13756 6.1940 -110.93
<none>                 7.3316 -103.83
+ gender   1   0.25854 7.0730 -103.77
+ age      1   0.23877 7.0928 -103.61
+ alcohol  1   0.06498 7.2666 -102.31
```

```
Step:  AIC=-130.48
logsurvival ~ enzyme + progind

          Df Sum of Sq    RSS     AIC
+ heavy    1   1.46961 2.8429 -150.99
+ score    1   1.20395 3.1085 -146.16
+ liver    1   0.69836 3.6141 -138.02
+ alcohol  1   0.22632 4.0862 -131.39
+ age      1   0.16461 4.1479 -130.59
<none>                 4.3125 -130.48
+ gender   1   0.08245 4.2300 -129.53
```

. . .

```
Step:  AIC=-163.83
logsurvival ~ enzyme + progind + heavy + score
                + gender + age

          Df Sum of Sq    RSS     AIC
<none>                 2.0052 -163.83
+ alcohol  1  0.033193 1.9720 -162.74
+ liver    1  0.002284 2.0029 -161.90
```

Example

Backward

```
Start:  AIC=-160.77
logsurvival ~ score + progind + enzyme + liver
              + age + gender + alcohol + heavy

          Df Sum of Sq    RSS     AIC
- liver    1   0.00129 1.9720 -162.74
- alcohol  1   0.03220 2.0029 -161.90
- age      1   0.07354 2.0443 -160.79
<none>                 1.9707 -160.77
- gender   1   0.08415 2.0549 -160.51
- score    1   0.31809 2.2888 -154.69
- heavy    1   0.84573 2.8165 -143.49
- progind  1   2.09045 4.0612 -123.72
- enzyme   1   2.99085 4.9616 -112.91
```

```
Step:  AIC=-162.74
logsurvival ~ score + progind + enzyme
            + age + gender + alcohol + heavy

          Df Sum of Sq    RSS      AIC
- alcohol  1    0.0332 2.0052 -163.834
<none>                 1.9720 -162.736
- age      1    0.0876 2.0596 -162.389
- gender   1    0.0971 2.0691 -162.141
- score    1    0.6267 2.5988 -149.833
- heavy    1    0.8446 2.8166 -145.486
- progind  1    2.6731 4.6451 -118.471
- enzyme   1    5.0986 7.0706  -95.784
```

```
Step:  AIC=-163.83
logsurvival ~ score + progind + enzyme
               + age + gender + heavy

          Df Sum of Sq    RSS      AIC
<none>                 2.0052 -163.834
- age      1    0.0768 2.0820 -163.805
- gender   1    0.0977 2.1029 -163.265
- score    1    0.6282 2.6335 -151.117
- heavy    1    0.9002 2.9055 -145.809
- progind  1    2.7626 4.7678 -119.064
- enzyme   1    5.0801 7.0853  -97.672
```

Example

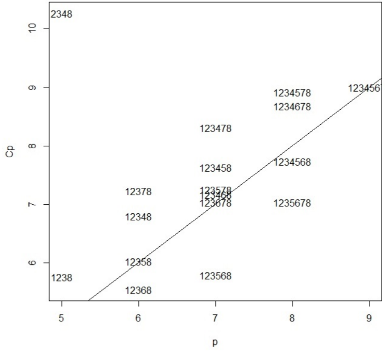


Figure: 1. score, 2. progind, 3. enzyme, 4. liver, 5. age, 6. gender, 7. alcohol, 8. heavy