Stat GR 5205 Lecture 8

Jingchen Liu

Department of Statistics
Columbia University

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

ANOVA table

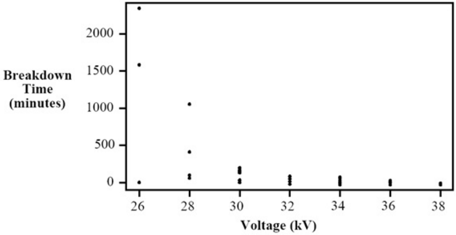
$SST = SSR + SSE$

$R^2 = \frac{SSE}{SST}$

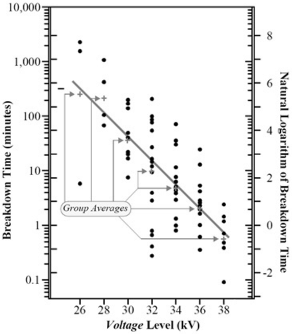
Multiple correlation

source	sums of sq	d.f.	mean sum of sq	F-stat	p-value
Regression	SST	p-1	SST/(p-1)		
Residual	SSE	n-p	SSE/(n-p)		
Total	SST				

Lack-of-fit test



Lack-of-fit test



Lack-of-fit test

source	sum of sq	d.f.	mean sq	F-stat	p-value
regression	190	1	190	78	< 0.0001
residual	180	74	2.4		
total	370	75			

source	sum of sq	d.f.	mean sq	F-stat	p-value
between group	196	6	33	13	< 0.0001
residual	174	69	2.5		
total	370	75			

Lack-of-fit test

$$F - statisitc = \frac{(196 - 190)/5}{174/69} = 0.48$$

Multicollinearity – orthogonality

$$X = (x_{ij})_{n \times p}$$

- Standardized covariates

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} = 0, \quad SD(x_j) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) = 1$$

- Uncorrelated covariates,

$$\sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

Multicollinearity – orthogonality

$$X = (x_{ij})_{n\times p}$$

- Standardized covariates

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \quad SD(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = 1$$

- Uncorrelated covariates,

$$\sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

Multicollinearity – orthogonality

- Uncorrelated covariates

$$X^\top X = I_{p\times p}$$

- The least-square estimate

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = X^\top Y$$

- Individual estimated coefficient

$$\hat{\beta}_j = \sum_{i=1}^{n} x_{ij} y_i$$

Multicollinearity – orthogonality

- Uncorrelated covariates

$$X^\top X = I_{p\times p}$$

- The least-square estimate

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = X^\top Y$$

- Individual estimated coefficient

$$\hat{\beta}_j = \sum_{i=1}^{n} x_{ij} y_i$$

Multicollinearity – orthogonality

- Uncorrelated covariates

$$X^\top X = I_{p\times p}$$

- The least-square estimate

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = X^\top Y$$

- Individual estimated coefficient

$$\hat{\beta}_j = \sum_{i=1}^{n} x_{ij} y_i$$

Multicollinearity – orthogonality

- Variance

$$Var(\beta_j) = \sigma^2 \sum_{i=1}^{n} x_{ij}^2 = \sigma^2$$

- Covariance

$$Cov(\beta_{j_1}, \beta_{j_2}) = \sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

Multicollinearity – orthogonality

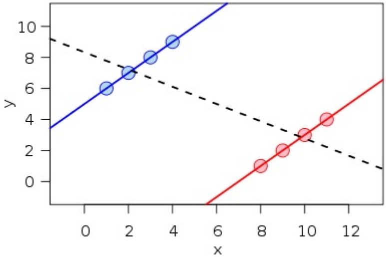
- Variance

$$Var(\beta_j) = \sigma^2 \sum_{i=1}^{n} x_{ij}^2 = \sigma^2$$

- Covariance

$$Cov(\beta_{j_1}, \beta_{j_2}) = \sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

A paradox due to multicollinearity



Quantification of multicollinearity

$$y, x_1, x_2, \ldots, x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1}$

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Quantification of multicollinearity

$$y, x_1, x_2, \ldots, x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1}$

Quantification of multicollinearity

$$y, x_1, x_2, \ldots, x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1}$

Quantification of multicollinearity

$$y, x_1, x_2, \ldots, x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1}$

Quantification of multicollinearity

- Variance inflation factor (VIF)

$$VIF(\beta_1) = \frac{Var(\tilde{\beta}_1)}{Var(\hat{\beta}_1)}$$

- A representation

$$VIF(\beta_1) = \frac{1}{1 - R_1^2}$$

where R_1^2 is the coefficient of determination of x_1 on x_2, $x_3, \ldots, x_p$.

Quantification of multicollinearity

- Variance inflation factor (VIF)

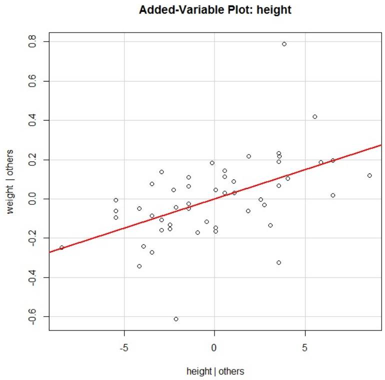
$$VIF(\beta_1) = \frac{Var(\tilde{\beta}_1)}{Var(\hat{\beta}_1)}$$

- A representation

$$VIF(\beta_1) = \frac{1}{1 - R_1^2}$$

where R_1^2 is the coefficient of determination of x_1 on x_2, $x_3, \dots, x_p$.

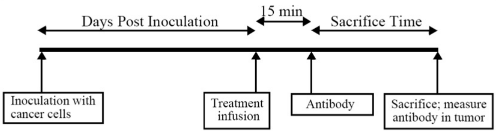
Value-added plot



Outlier detection

Display 11.3 **p. 307**

Time line for blood-brain barrier disruption experiment



Outlier detection

Display 11.4 p. 308

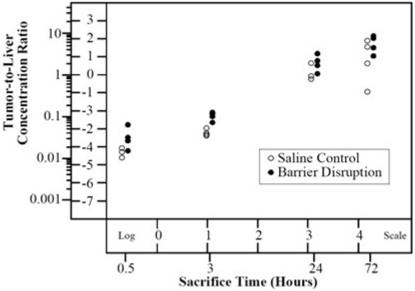
Response variable, design variables, and several covariates for 34 rats in the blood-brain barrier disruption experiment

	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**

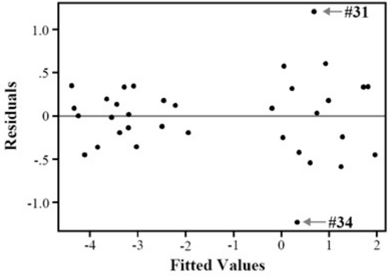
Log-log scatterplot of ratio of antibody concentration in brain tumor to antibody concentration in liver versus sacrifice time, for 17 rats given the barrier disruption infusion and 17 rats given a saline (control) infusion



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
- 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
- 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
- Total leverage

Leverage

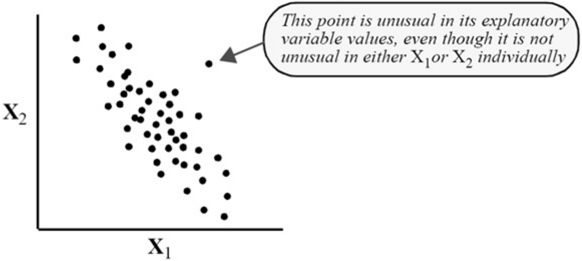
- About leverage
- Simple linear model

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

- Multiple regression
- Total leverage

High leverage

An illustration of what is meant by "far from the average" of multiple explanatory variables when they are correlated



Leave-one-out measure

- DIFFITS

$$DIFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_i}}$$

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

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

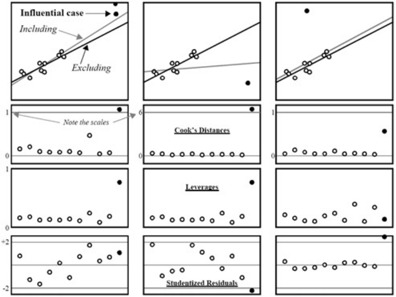
Leave-one-out measure

- DFBETAS

$$DFBETAS_{k(i)} = \frac{\beta_k - \beta_{k(i)}}{\sigma\sqrt{c_k}}$$

Influential points

Three examples of influential cases in simple linear regression. The top row shows regression lines with and without the influential case included. The next three rows show the resulting case influence statistic plots: Cook's distances, leverages, and Studentized residuals. The horizontal axes for the case statistic plots show the case numbers (=11 for the influential case).



A. High leverage and mild departure changes the slope so that the residual is small. Cook's distance identifies the offending case.

B. High leverage and huge departure drastically pulls the line away from all observations. Cook's distance identifies the case.

C. Low leverage does not allow the large departure to alter the slope, so it ends up with a big residual. Cook's distance shows a mild problem.