ENTO/ RNR 613 Multiple regression -- Model Checking and Refinement

Least squares regression analysis is not resistant to outliers. A few observations that lie away from the multivariate average (of the Xs and Ys) can strongly influence the outcome of the analysis.

It is important to consider transformations and outliers before searching for an inferential model. When in doubt, special tools may be useful to flag outliers: *leverage* or *studentized residual* values, and *Cook's distance*.

When a few observations have distant values for the explanatory variables, it may help to *omit the distant cases* and restrict the statistical inferences to a reduced range of explanatory variables.

Selection of an inferential model depends on the *question asked* and on the *patterns* of the data. Especially when many explanatory variables are used (i.e. complex models), there is not necessarily a "best" model. There are only models that are *adequate* to answer the question of interest.

Selecting a tentative model

1. The model must contain parameters whose values answer the questions of interest

2. It should include potentially confounding variables

3. It should take into consideration the relationships indicated by initial graphical inspection of the data

4. The number of explanatory variables included in the initial model depends on sample size:
   a) with large samples, one can fit a *rich model* including interaction and quadratic terms without overfitting the data (i.e. without explaining real outliers)
   b) with small samples, fitting several *simple models* may be needed, remembering that some observations may appear like outliers in simple models because real effects are not taken into account.

Example: Blood-Brain Barrier

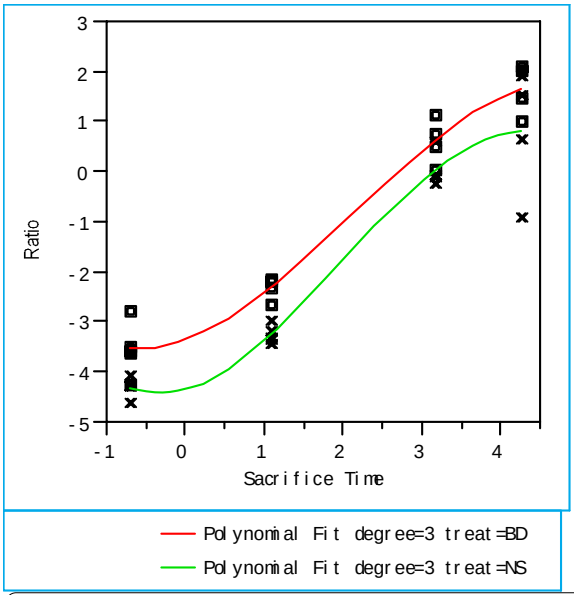
Rats were inoculated with cancer cells to induce brain tumors.

About 10 days later (range 9-11) they were given a saline solution (control) or a solution of concentrated sugar to disrupt the brain membrane barrier.

Fifteen minutes later, the rats got a fixed dose of an antibody supposed to attack the brain tumor.

Finally, rats were sacrificed after 0.5, 3, 24, or 72 hours, and the amount of antibody in the brain and liver was quantified.

The questions asked are: does the barrier disruption solution increases permeability of the brain membrane to the therapeutic antibody, and if yes does time after injection of the antibody affects quantity of antibody reaching the brain?



Squares = Barrier disruption

Crosses = Saline control

Response = Anti Brain / Anti liver

Choosing an initial model:

The coded scatterplot indicates some curvature in the response (log [brain A/liver A]) to the 2 treatments as a function of sacrifice time (log transformed). To avoid mismodeling the effect of sacrifice time at the start, sacrifice time is treated as a *factor with 4 levels* (categorical variable).

The effect of the treatment appears greater for the shorter sacrifice time than the larger ones. Thus the sacrifice time × treatment *interaction* term is included in the initial model.

Finally, that experiment involved 2 types of explanatory variables:

1- *Design variables* which are manipulated by the researcher (Sacrifice time, Treatment)
2- *Covariates* that measure characteristics of the subjects that are not controllable but that nonetheless may affect the response.

<<Display 11.4 in Sleuth>>

The initial regression model includes those covariates, because *including important covariates in a model yield higher resolution*:

$$\mu\{antibody\ ratio \mid SAC, TREAT, DAYS, FEM, weight, loss, tumor\} =$$
$$SAC + TREAT + (SAC\ x\ TREAT) + DAYS + FEM + weight + loss + tumor$$

Where SAC is sacrifice time (4 levels: categorical variable)
TREAT is treatment (2 levels: categorical)
DAYS is days after inoculation (3 levels: categorical)
FEM is sex (2 levels: categorical)
Weight, *loss*, and *tumor* are initial weight, weight loss, and tumor weight (continuous covariates)

Another example: Alcohol Metabolism

Women and men categorized as alcoholic or not, received ethanol orally on one day and intravenously on another (order determined at random). The difference in blood alcohol concentration between the 2 treatments provides a measure of activity of alcohol-degrading enzymes in the stomach: such a difference is called first-pass metabolism. Alcohol dehydrogenase (AD) activity was also measured directly in samples from stomachs.

Do level of first-pass metabolism differ between men and women?

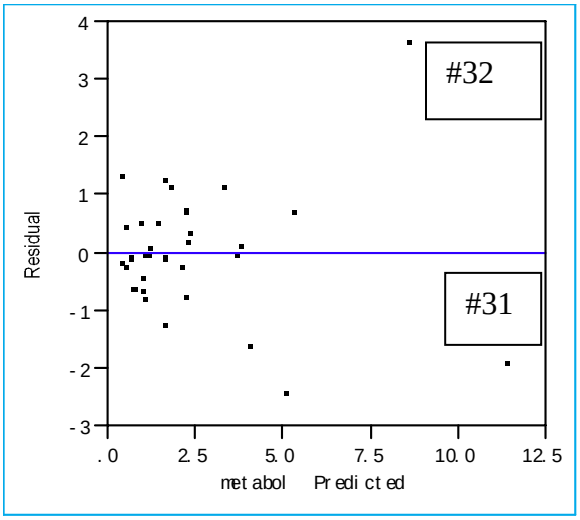
Can the difference be explained by difference in AD levels in males and females?

<<Display 11.2>>

In the scatterplot, note the 2 isolated observations with extreme values of X. Those unusually "distant" observations could have much influence on regression results.

There are 32 observations: the "saturated model" includes 7 terms.

$$\mu\{metabolism \mid gast, FEM, ALCO\} = FEM + ALCO + gast + gast\ x\ FEM$$
$$+ FEM\ x\ ALCO + gast\ x\ ALCO + gast\ x\ FEM\ x\ ALCO$$



What is the influence of those outliers?

A strategy for dealing with Influential Observations

Least squares regression is *not resistant to outliers*. It is unwise to state conclusions that depend on one or two data points.

When a few outliers are influential, 2 options are available:

1- Use a resistant statistical procedure (Permutation of Rank sum test)
2- Examine whether including the observations results in qualitative changes in conclusions from the tests. If the observations have *distant values for a given explanatory variable*, restrict the analysis to a *smaller range of the data*. Mention that the rare observations were not considered in the analysis.

<< Display 11.8, Sleuth>>

Results with and without observation #31 and # 32, Alcohol Metabolism Study:

With all observations

Term	Estimate	Std Error	t Ratio	P
Intercept	-1.65966	0.999646	-1.66	0.1099
gastric	2.5141572	0.34337	7.32	<.0001
female	1.4657195	1.332553	1.10	0.2823
alcohol	2.5521036	1.945986	1.31	0.2021
g*f	**-1.673438**	**0.6202**	**-2.70**	**0.0126**
f*a	-2.251711	4.3937	-0.51	0.6130
g*a	-1.458742	1.052858	-1.39	0.1786
g*f*a	1.1986678	2.997826	0.40	0.6928

Without #31 and #32

Term	Estimate	Std Error	t Ratio	P
Intercept	-0.679714	1.309056	-0.52	0.6088
gastric	1.9212411	0.608182	3.16	0.0046
female	0.4857728	1.466535	0.33	0.7436
alcohol	1.5721569	1.811896	0.87	0.3949
g*f	**-1.080522**	**0.72115**	**-1.50**	**0.1483**
f*a	-1.271765	3.466858	-0.37	0.7172
g*a	-0.865825	0.963107	-0.90	0.3784
g*f*a	0.6057516	2.315807	0.26	0.7961

The distant cases #31 and #32 are two males. The slope for males is *much greater* with those 2 cases considered than without. There may be 2 reasons for this:
1- The slope is really greater for males than for females
2- The relationship between gastric metabolism and enzyme concentration is not linear.

A prudent strategy is to exclude the 2 distant cases and restrict the analysis to a range of gastric AD activity lower than 3.

Case-Influences Statistics

When the residual plot from fitting a good inferential model does not suggest any problems, there is generally *no need* to examine case influence statistics.

Leverage for flagging cases with unusual explanatory variable value

The leverage (h) of an observation is a measure of the distance between its explanatory variable value and the average of the values of the explanatory variable used in the model.

One formula is (see Sleuth p. 303 for the other):

$$h_i = \frac{1}{(n-1)}\left[\frac{X_i - \overline{X}}{s_x}\right] + \frac{1}{n}$$

i.e. h is a distance of Xi from $\overline{X}$ in units of standard deviation. It varies between 1/n and 1.

A case with high leverage *may have* strong influence on the statistical inferences. A case with high leverage is likely the only observation in the "region". Because the shape of the regression surface is determined by the method of Least squares, the residual of a case with high leverage must be small, which implies that such cases act as a magnet on the estimated regression surface.

<< Display 11.10 in Sleuth>>

A case with high leverage *will not influence the regression surface* if the value of the observation falls close to the regression surface.

The average value of h is p/n, where p is the number of regression coefficients in the model. Some statisticians use twice that value as a lower threshold to flag influential values.

Studentized Residuals for flagging Outliers

A *studentized residual* is a residual divided by its estimated standard deviation. It is a useful influence statistic because not all residuals have equal variability, so visual inspection of the residuals may not always provide a correct evaluation of outliers. The higher the leverage, the lower the expected spread of the residuals: *SD (Residual$_i$)*=

$\sigma\sqrt{(1-h_i)}$. This is why the usual residual plot may fail to direct attention to "distant" outliers.

The studentized residual:

$$studentres_i = \frac{res_i}{\hat{\sigma}\sqrt{1-h_i}}$$

put all residuals on a common scale. Roughly 95 % of those residuals are expected to be between –2 and 2.

Cook's Distance for flagging influential cases

Cook's distance measures the overall influence of an observation: the effect that omitting a case has on *all* the estimated regression coefficients.

One way of calculating this statistic shows that a case with a large Cook's D is influential because it has a large studentized residual, a large leverage, or both:

$$D_i = \frac{1}{p}(studres_i)^2\left(\frac{h_i}{1-h_i}\right)$$

Some statisticians use as a rough guideline that a value of D_i larger than 1 indicates a large influence.

The Alcohol Metabolism example:

<<Display 11.11 Sleuth>>

(Cook's D, Leverage, and Studentized Residuals can be saved after an analysis from the Fit Model platform, then plotted using the Graph platform)

We thus *eliminate case #31 and #32*. From the saturated model above, none of the coefficients involving Alcoholism were different from zero. We can assess formally whether any terms containing ALCO should be retained in the inferential model.

The Extra Sum of square test for comparing the following full and reduced model:

Full:

$$\mu\{metabolism \mid gast, FEM, ALCO\} = FEM + ALCO + gast + gast\ x\ FEM + FEM\ x\ ALCO + gast\ x\ ALCO + gast\ x\ FEM\ x\ ALCO$$

Reduced:

$$\mu\{metabolism \mid gast, FEM\} = FEM + gast + gast\ x\ FEM$$

is $F_{4,22} = \dfrac{(20.22 - 19.48)/4}{0.88}$ = 0.21, which yields a p-value of 0.93. The data indicate no effect of alcoholism on first-pass metabolism.

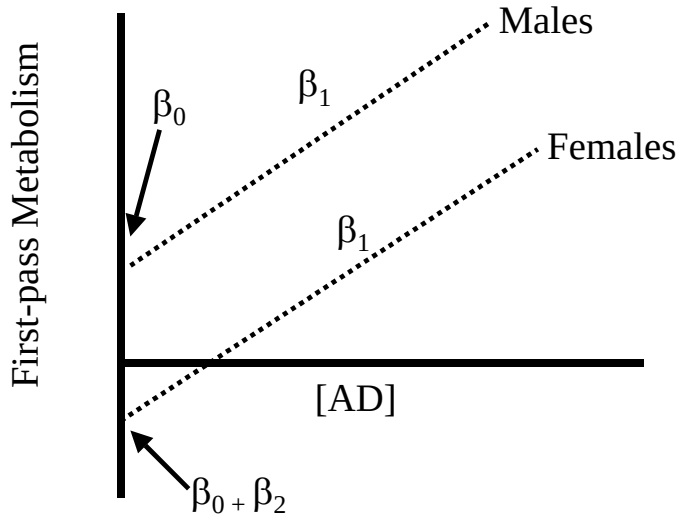
For the reduced model we obtain:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0695235	0.801948	0.09	0.9316
gastric	1.5654339	0.40739	3.84	0.0007
female	-0.266793	0.993244	-0.27	0.7904
gastric*female	-0.728486	0.53937	-1.35	0.1885

So we have three possible choices for going further in selecting an inferential model:

a) $\mu\{metabolism \mid gast, fem\} = \beta_o + \beta_1 gast + \beta_2 fem$ (parallel lines model)



Summary of Fit

RSquare	0.649988
RSquare Adj	0.624061
Root Mean Square Error	0.895307
Mean of Response	1.856667
Observations (or Sum Wgts)	30

Parameter Estimates

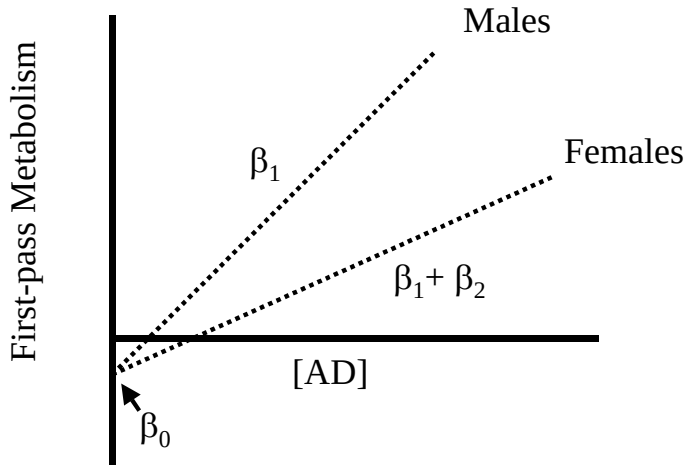
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8452994	0.568125	1.49	0.1484
gastric	1.1498396	0.271036	4.24	0.0002
female	-1.527551	0.344523	-4.43	0.0001

This model indicates that the difference in first-pass metabolism between males and females is 1.52 units, for a given level of AD activity (two-sided p-value = 0.0001). The F-statistic for comparing this model to the previous one is 1.82 with 1, 27 df (p-value > 0.05), so this smaller model is adequate, at least for prediction.

The intercept is allowed to differ from zero for both males and females; this would be acceptable if we knew that other factors than AD contribute to first-pass metabolism. The fact that we obtain an intercept for males that is not different from zero (two-tailed p-value = 0.15), and the possibly negative intercept for females, suggests that a better inferential model could be obtained.

b) $\mu\{metabolism \mid gast,\ fem\} = \beta_o + \beta_1\, gast + \beta_2 (gast\ x\ fem)$

(common intercept, different slope model)



Summary of Fit

RSquare	0.672028
RSquare Adj	0.647734
Root Mean Square Error	0.866661
Mean of Response	1.856667
Observations (or Sum Wgts)	30

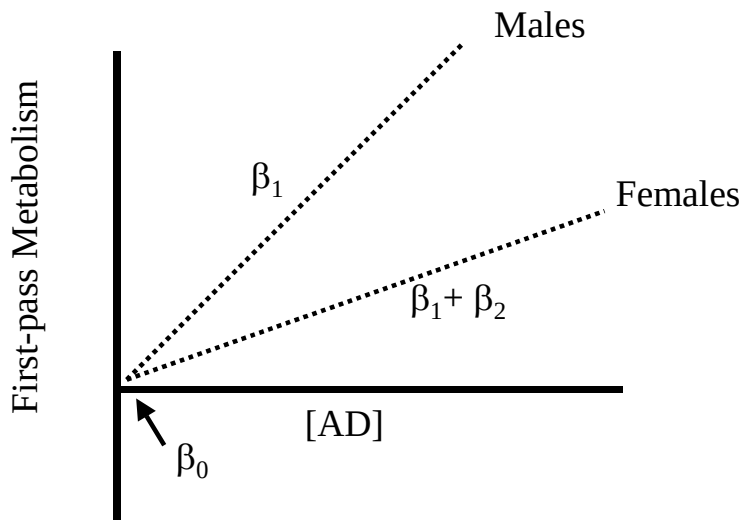
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.104399	0.464948	-0.22	0.8240
gastric	1.6492157	0.257522	6.40	<.0001
female*gastric	-0.864645	0.181103	-4.77	<.0001

This model states that first-pass metabolism is directly proportional to gastric activity (two-sided p-value < 0.001), but the slope of the relationship differs for males and females (two-sided p-value < 0.0001). The F-statistic to compare this model to our previous (full) model is 1.74 with 1, 27 df (p-value > 0.05), so this smaller model is adequate, at least for prediction.

We allowed the common intercepts to be different from zero, but there is no evidence that this is so (two-sided p-value = 0.82). To obtain a simple answer to our questions, a third model seems indicated.

c) $\mu\{metabolism \mid gast, fem\} = \beta_1 gast + \beta_2 (gast\ x\ fem)$ (common intercept forced through the origin, different slopes)



The logic here is that without any AD there should not be any first-pass metabolic activity. This thinking could have arisen from previous knowledge on alcohol digestion. The pattern of the data observed in the previous analysis also provides an empirical argument for forcing the origin through 0.

To force the regression line through the origin, tick the box "no intercept" in the JMP Fit Model platform.

Summary of Fit

RSquare	?
RSquare Adj	?
Root Mean Square Error	0.851838
Mean of Response	1.856667
Observations (or Sum Wgts)	30

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Zeroed	0	0	?	?
gastric		1.5989247	0.12492	12.80	<.0001
female*gastric		-0.873232	0.173991	-5.02	<.0001

Here again, first-pass metabolism is directly proportional to AD activity (two-sided p-value < 0.0001), and the slope for males and females differs (two-sided p-value < 0.0001). The extra-sum-of-squares F-test for comparing this model to the previous full model yields a value of 0.56 with 2, 28 df (P > 0.05), so this model fits.

A simple interpretation of this model can be obtained, noting that the mean first-pass metabolism for males divided by the mean first-pass metabolism for females is:

$$\frac{\beta_1 gast}{\beta_1 gast + \beta_2 gast} = \frac{\beta_1}{\beta_1 + \beta_2}$$

which is estimated to be 2.20. For a given level of gastric AD activity, the mean first-pass alcohol metabolism for men is estimated to be 2.20 times as large as the mean first-pass alcohol metabolism for women.