Chapter 11 - Model checking and refinement

November 29, 2017

Setting:

- You have data: a response $Y$ and explanatory variables $X_1, X_2, \ldots$
- We want to build an MLR model of the mean of a response ($Y$) as a function of predictors and/or factors $X_1, X_2, \ldots$
- Model-building efforts are wasted if the analyst fails to detect problems with non-constant variance, non-linearities, non-normality and outliers early on. So postpone detailed model fitting until after outliers and transformations have been thoroughly considered.
- Start with an initial tentative model that includes:
  1. parameters that can be used to answer the planned research question(s) of interest;
  2. potential confounding variables;
  3. features (such as transforms) that capture important relationships found in the initial graphical analysis (i.e., scatterplots, matrix plots, Trellis plots).
- The tentative model should be fairly complex and perhaps include terms that may not be used for the final model. But be careful not to overfit to the data which especially worrisome for small sample sizes.
- The tentative model is an MLR model:

$$
\mu\{Y|X_1, X_2, \ldots\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \ldots
$$

Using the above notation implies that any factors have already been encoded as Dummy (indicator) variable(s) $X_j$ that either take on 0 or a 1 as described in Chapter 9.
- There are $\tilde{p}$ parameters: $\beta_0, \beta_1, \beta_2, \ldots, \beta_{\tilde{p}-1}$. Be wary, your book uses the notation $p$ that that is defined differently in Chapter 11 compared to Chapter 10!

11.1 Alcohol Metabolism in Men and Women

When men and women of the same size and drinking history consume equal amounts of alcohol, other studies have suggested that women have a higher blood alcohol content. In the study presented here, the researchers wanted to determine whether women metabolize alcohol differently in their stomachs differently than men. Does this “first pass” stomach metabolism differ between the sexes? Can the differences be explained by a difference in gastric enzyme activity? Is there an effect due to alcoholism?

```r
library(Sleuth3)
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")
# Make R arrange categories as in Display 11.9
case1101$Sex=factor(case1101$Sex,levels=c("Male","Female"))
case1101$Alcohol=factor(case1101$Alcohol,levels=c("Non-alcoholic","Alcoholic"))
summary(case1101)
```

<table>
<thead>
<tr>
<th>Subject</th>
<th>Metabol</th>
<th>Gastric</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>1.00</td>
<td>0.100</td>
<td>Male :14</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>8.75</td>
<td>0.600</td>
<td>Female:18</td>
</tr>
</tbody>
</table>
## Median :16.50 Median : 1.700 Median :1.600
## Mean :16.50 Mean : 2.422 Mean :1.863
## 3rd Qu.:24.25 3rd Qu.: 2.925 3rd Qu.:2.200
## Max. :32.00 Max. :12.300 Max. :5.200
## Alcohol
## Non-alcoholic:24
## Alcoholic : 8
##
## dim(case1101)  # n=32
##
## [1] 32 5
## # The unusual data far from the rest
## case1101[c(31,32),]
##
## # Subject Metabol Gastric Sex Alcohol
## # 31 31 9.5 5.2 Male Non-alcoholic
## # 32 32 12.3 4.1 Male Non-alcoholic
##
## # Trellis plot
## pch.index=numeric(32)
pch.index[case1101$Sex=="Female"&case1101$Alcohol=="Non-alcoholic"]=17  # filled tri's
pch.index[case1101$Sex=="Female"&case1101$Alcohol=="Alcoholic"]=19  # filled circ's
pch.index[case1101$Sex=="Male"&case1101$Alcohol=="Non-alcoholic"]=2  # open tri's
pch.index[case1101$Sex=="Male"&case1101$Alcohol=="Alcoholic"]=1  # open circ's
coplot(Metabol ~ Gastric|Sex*Alcohol, data=case1101, pch=pch.index)
11.2 Residual plots

Your text considers a tentative MLR model of the alcohol data that includes all 2-way interactions and the 3-way interaction ($p = 8$ parameters) that allows for four separate regression lines with different intercepts and slopes:

$$
\mu \{\text{Metabol} | \text{Gastric, Sex, Alcohol} \} = \beta_0 + \beta_1 \text{Gastric} + \beta_2 \text{Dummy}_F(\text{Sex}) + \beta_3 \text{Dummy}_A(\text{Alcohol}) \\
+ \beta_4 \text{Gastric} \times \text{Dummy}_F(\text{Sex}) + \beta_5 \text{Gastric} \times \text{Dummy}_A(\text{Alcohol}) \\
+ \beta_6 \text{Dummy}_F(\text{Sex}) \times \text{Dummy}_A(\text{Alcohol}) \\
+ \beta_7 \text{Gastric} \times \text{Dummy}_F(\text{Sex}) \times \text{Dummy}_A(\text{Alcohol})
$$

```r
m <- lm(Metabol ~ Gastric*Sex*Alcohol, data=case1101)
summary(m)  # compare with Display 11.7
```

## Call:
## lm(formula = Metabol ~ Gastric * Sex * Alcohol, data = case1101)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -8.500  0.2500  0.7500  1.0000  2.5000

```r
# matrix plot
#pairs(case1101)
```
# Coefficients:

| Term                           | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------------------------|----------|------------|---------|---------|
| (Intercept)                   | -1.6597  | 0.9996     | -1.660  | 0.1099  |
| Gastric                       | 2.5142   | 0.3434     | 7.322   | 1.46e-07|
| SexFemale                     | 1.4657   | 1.3326     | 1.100   | 0.2823  |
| AlcoholAlcoholic              | 2.5521   | 1.9460     | 1.311   | 0.2021  |
| Gastric:SexFemale             | -1.6734  | 0.6202     | -2.698  | 0.0126  |
| Gastric:AlcoholAlcoholic      | -1.4587  | 1.0529     | -1.386  | 0.1786  |
| SexFemale:AlcoholAlcoholic    | -2.2517  | 4.3937     | -0.512  | 0.6130  |
| Gastric:SexFemale:AlcoholAlcoholic | 1.1987 | 2.9978     | 0.400   | 0.6928  |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.254 on 24 degrees of freedom
Multiple R-squared: 0.8277, Adjusted R-squared: 0.7774
F-statistic: 16.47 on 7 and 24 DF, p-value: 9.354e-08

# Plot the data and the model fit

```r
plot(Metabol ~ Gastric, pch=pch.index, data=case1101, ylim=c(-1,13), xlim=c(0,6))
pred.m=predict(m, se.fit=TRUE)
lines(case1101$Gastric[pch.index==17],pred.m$fit[pch.index==17],lty=1)
lines(case1101$Gastric[pch.index==19],pred.m$fit[pch.index==19],lty=2)
lines(case1101$Gastric[pch.index==2],pred.m$fit[pch.index==2],lty=3)
lines(case1101$Gastric[pch.index==1],pred.m$fit[pch.index==1],lty=4)
legend("topleft",legend=c("Non-alcoholic female","Alcoholic female","Non-alcoholic male","Alcoholic male"),
pch=c(17,19,2,1),lty=1:4)
```
# Plot residuals
# diagANOVAM(m)
par(mfrow=c(1,3))
plot(m,which=c(1,2))  # Two of R’s built-in residual plots includes case #'s
plot(case1101$Gastric,resid(m))  # plot residuals vs the single predictor
11.3 Dealing with Influential observations

We identified the data for Subjects 31 and 32 as unusual in our original plots of the data. The residual plots above show that Subject 32 is an outlier because the observed Metabolism is far from the predicted Metabolism (i.e., has a large residual). Subject 31, while maybe not an outlier, potentially can highly influence the regression fit because the observed Gastric activity is far from the rest of the data (i.e., it has an unusually high value of an explanatory variable). What to do with these data for Subjects 31 and 32? Display 3.6 (p. 69) and Display 11.8 (p. 321) and Chapter 3 notes offer guidance:

- Examine unusual data and outliers for recording errors or other contamination.
- Check whether a standard transformation resolves the problem - an unusual value on the original scale may be fine on a log scale.
- If neither of the first two steps resolve the problem, determine whether the unusual data or outliers are influential i.e., if they influence the fitted model and affect conclusions. One way to do this is to fit the model with and without the unusual data and compare the results. If removing the unusual data does not substantially change the fitted model then the unusual data are not influential.
- If influential data have unusual explanatory values then you may be justified in removing them. In this case, report the exclusion and limit your scope of inference to the appropriate range of explanatory variables.
- If the influential data does not have an unusual explanatory value and there is no clear reason to exclude it then keep it in the data set.

Let’s determine the influence of Subjects 31 and 32 by analyzing the alcohol data after removing Subjects 31 and 32:
`case1101[-c(31,32),]`  # Look at this data frame to convince yourself that
# Subjects 31 and 32 have been removed, so now n = 30

```r
m2 <- lm(Metabol ~ Gastric*Sex*Alcohol, data=case1101[-c(31,32),])
summary(m2)
```

```verbatim
## Call:
## lm(formula = Metabol ~ Gastric * Sex * Alcohol, data = case1101[-c(31,
## 32),])
##
## Residuals:
##    Min     1Q   Median     3Q    Max
## -1.80764 -0.57006 -0.04664  0.49762  1.40024
##
## Coefficients:
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   -0.6797     1.3091  -0.519   0.60878
## Gastric       1.9212     0.6082   3.159  0.004549 **
## SexFemale     0.4858     1.4665   0.331   0.74360
## AlcoholAlcoholic 1.5722     1.8119   0.868   0.39493
## Gastric:SexFemale -1.0805     0.7211  -1.498   0.14826
## Gastric:AlcoholAlcoholic -0.8658     0.9631  -0.899   0.37839
## SexFemale:AlcoholAlcoholic 1.2718     3.4669   0.367   0.71725
## Gastric:SexFemale:AlcoholAlcoholic 0.6058     2.3158   0.262   0.79608
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9411 on 22 degrees of freedom
## Multiple R-squared: 0.6849, Adjusted R-squared: 0.5846
## F-statistic: 6.831 on 7 and 22 DF,  p-value: 0.0002262
```

```r
# diagANOVAm2
par(mfrow=c(1,3))
plot(m2,which=c(1,2))  # Two of R's built-in residual plots includes case #'s
plot(case1101$Gastric[-c(31,32)], resid(m2))  # plot residuals vs the single predictor
```
Not surprisingly, the residual plots look better when the unusual data for Subjects 31 and 32 are omitted. However, there are substantial changes in the size of the estimated regression coefficients. In particular, the slope of the line for Non-alcoholic males decreased substantially (the plot below shows the updated model fit) and there is a change in our conclusion regarding the statistical significance of the interaction term for $\text{Gastric} \times \text{Sex}$. Hence the data for Subjects 31 and 32 are influential.

```r
# Model fit after removing Subjects 31 and 32
par(mfrow=c(1,1))
plot(Metabol ~ Gastric, pch=pch.index, data=case1101, ylim=c(-1,13), xlim=c(0,6))
pred.m2=predict(m2, se.fit=TRUE)
lines(case1101$Gastric[pch.index==17], pred.m2$fit[pch.index==17], lty=1)
lines(case1101$Gastric[pch.index==19], pred.m2$fit[pch.index==19], lty=2)
lines(case1101$Gastric[pch.index==2], pred.m2$fit[pch.index==2], lty=3)
lines(case1101$Gastric[pch.index==1], pred.m2$fit[pch.index==1], lty=4)
legend("topleft", legend=c("Non-alcoholic female","Alcoholic female","Non-alcoholic male", "Alcoholic male"), pch=c(17,19,2,1), lty=1:4)
```
11.4 Case influence statistics

Identifying influential data (or observations or cases) by refitting models with and without the unusual cases as we did above can become logistically daunting when there are many explanatory variables and a LARGE data set. Case influence statistics are numerical measures of the influence of each individual case are without having to explicitly refit models.

11.4.1 Leverage

The leverage of a case is the distance between a case’s explanatory variable values and the average of the explanatory variable values in the entire data set. The leverage of the $i$th case is:

$$h_i = \left[ \frac{SE(\hat{\mu}_i(Y|X_1, X_2, ...))}{\hat{\sigma}} \right]^2.$$ 

Some useful facts:

- $\frac{1}{n} < h_i < 1$; for the alcohol example $\frac{1}{32} = 0.0313 < h_i < 1$
- The average of all the leverages is $\bar{h} = \frac{\bar{h}}{n}$; for the alcohol example $\bar{h} = \frac{8}{32} = \frac{1}{4}$
- Your text suggests that the $i$th case has a “large” leverage if $h_i > 2\bar{h} = \frac{2\bar{h}}{n}$
- The $i$th case with a large leverage $h_i$ is not necessarily influential, but does have the potential to be.
- $SD(\text{Residual}_i) = \sigma \sqrt{1 - h_i}$
# Calculate leverages by hand
se.fits <- pred.m$se.fit
sig.hat <- summary(m)$sigma
leverages <- (se.fits/sig.hat)^2

# Or let R calculate leverages
leverages <- as.numeric(lm.influence(m)$hat)

# leverages # view leverage for all 32 cases
mean(leverages)

## [1] 0.25

# Are Subjects 31 and 32 identified by large leverages?
hist(leverages) # Histogram 'center' = mean(h) = 0.25

![Histogram of leverages](image.png)

plot(leverages,xlab="Case")
lines(c(0,32),c(2*8/32,2*8/32),lty=2) # Points above \(\frac{2p}{n}\) have "high" leverage
text(32,0.45,labels="\(\frac{2p}{n}\")
11.4.2 Studentized residuals

An (internally) Studentized Residual for the $i$th case is a residual divided by its estimated SD,

$$
\text{studres}_i = \frac{\text{Residual}_i}{\text{SD(Residual)}_i} = \frac{\text{Residual}_i}{\hat{\sigma}\sqrt{1 - h_i}}.
$$

It is common to investigate cases with Studentized residuals greater in magnitude than 2 or 3.

```
# Are Subjects 31 and 32 identified by large studentized residuals?
#rstandard(m)  # View studentized residuals for all 32 cases
#hist(rstandard(m))
plot(rstandard(m),xlab="Case")
lines(c(0,32),c(-2,-2),lty=2)
lines(c(0,32),c(2,2),lty=2)
```
11.4.3 Cook’s Distance

Cook’s Distance measures the influence of the $i$th case by determining the effect of omitting that case on the fitted values:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})}{\hat{\sigma}^2} = \frac{1}{\hat{\rho}} (\text{studres}_i)^2 \left( \frac{h_i}{1 - h_i} \right)$$

where

- $\hat{Y}_j$ is the $j$th fitted value using all observations
- $\hat{Y}_{j(i)}$ is the $j$th fitted value excluding observation $i$
- $\hat{\rho}$ is the number of regression coefficients $\beta_0, \beta_1, ... \beta_{\hat{\rho} - 1}$.
- $\hat{\sigma}^2$ is the estimate of $\sigma^2$ from the model with all observations

The formula shows that a large Cook’s Distance can arise because the observation has either a large residual, high leverage, or both.

Your book suggests that a cutoff for Cook’s Distances is close to or greater than 1.

```r
# Are Subjects 31 and 32 identified by large Cook's distances?
cooks.distance(m)  # View studentized residuals for all 32 cases
hist(cooks.distance(m))
plot(cooks.distance(m), xlab="Case")
lines(c(0,32),c(1,1), lty=2)
```
11.4.4 Using case influence statistics

- Start with scatterplots, matrix plots and Trellis plots of the data. These may alert you to unusual data to assist you in determining a tentative model.

- Fit a tentative model and inspect the residual plots. If the residual plots do not identify unusual data, then generally you do not have to examine case influence statistics at all.

- Unusual data include cases with large residuals (i.e., outliers) and cases with large values of the explanatory variable(s).

- If unusual data are identified, then it is appropriate to consider case influence statistics. No one case influence statistic is “better” than the others in determining influence, so they should be considered jointly. The best way to do this is graphically.

```r
par(mfrow=c(2,2))
plot(m, which=c(1,2,4,5))
```

```r
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```r
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```
QUESTIONS:

1. What do you conclude about cases 31 and 32? Leave them in or exclude them for the final statistical analysis?

2. Based on the above analysis, the textbook removes cases 31 and 32 from the data set. State the scope of inference for any conclusions that we find based on this updated data set.

3. Go back and check the model fit to the data set without cases 31 and 32. Does the model fit the updated data well? Is it necessary to check case influence statistics for this new data set?
11.5 Refining the model

After dealing with potential problems the next goal is to find a good fitting final model that includes important variables but not nonessential variables.

My first step is to drop the 3-way interaction that is not statistically significant:

```r
# Succinct notation to include all 2-way interactions without higher order interactions
m3 = lm(Metabol ~ (Gastric + Sex + Alcohol)^2, data=case1101[-c(31,32),])
summary(m3)
```

```text
## Call:
## lm(formula = Metabol ~ (Gastric + Sex + Alcohol)^2, data = case1101[-c(31,32),])
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.78138 -0.53691 -0.05874  0.49371  1.43068
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5932     1.2406  -0.478  0.63708
## Gastric      1.8795     0.5748   3.270  0.00337 **
## SexFemale   -0.3723     1.3723  -0.271  0.78857
## AlcoholAlcoholic  1.3861    1.6323   0.849  0.40455
## Gastric:SexFemale -1.0218     0.6713  -1.522  0.14160
## Gastric:AlcoholAlcoholic -0.7611    0.8579  -0.887  0.38423
## SexFemale:AlcoholAlcoholic -0.3973    0.8991  -0.442  0.66270
## ---
## Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 1
## Residual standard error: 0.9218 on 23 degrees of freedom
## Multiple R-squared: 0.6839, Adjusted R-squared: 0.6015
## F-statistic: 8.295 on 6 and 23 DF, p-value: 7.513e-05

#anova(m3,m2)
```

The previous output suggests that all of the 2-way interactions with Alcohol may be dropped from the model.

```r
m4 = lm(Metabol ~ Gastric*Sex + Alcohol, data=case1101[-c(31,32),])
#summary(m4)
anova(m4,m2)
```

```text
## Analysis of Variance Table
## Res.Df RSS  Df Sum of Sq      F Pr(>F)
## 1 25 20.223
## 2 22 19.483  3 0.73955 0.2784 0.8404
```

One might argue to keep Alcohol in the model because it is a potential confounder. Using model m4 as the final model would allow us to make conclusions re: the effect of the gastric enzyme after accounting for known differences in metabolism by sex and drinking history. However, using this model seems to require that we need to specify subjects’ sex and Alcohol history in our conclusions.
If we want to make conclusions pooled across all Alcohol histories, and because alcoholism is not of primary concern, the textbook argues to drop Alcohol history entirely from the final model.

```r
m5 = lm(Metabol ~ Gastric*Sex, data=case1101[-c(31,32),])
#anova(m5,m2)
summary(m5)  # Compare with Display 11.13
```

## Call:
## lm(formula = Metabol ~ Gastric * Sex, data = case1101[-c(31, 32), ])
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -1.59619 -0.60249 -0.04076  0.47590  1.64726
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.06952   0.80195   0.087   0.93158
## Gastric                 1.56543   0.40739   3.843  0.000704 ***
## SexFemale               -0.26679   0.99324  -0.269   0.79035
## Gastric:SexFemale       -0.72849   0.53937  -1.351   0.18845
## ---
## Signif. codes:  < 0.001 ***  0.001 **  0.05 *  1
##
## Residual standard error: 0.8819 on 26 degrees of freedom
## Multiple R-squared:  0.6729, Adjusted R-squared:  0.6352
## F-statistic: 17.83 on 3 and 26 DF,  p-value: 1.711e-06

#diagANOVA(m5)

Based on the above output, you may be tempted to drop the interaction with Sex from the model as well.

```r
m6 = lm(Metabol ~ Gastric + Sex, data=case1101[-c(31,32),])
anova(m6,m2)
```

## Analysis of Variance Table
##
## Model 1: Metabol ~ Gastric + Sex
## Model 2: Metabol ~ Gastric * Sex * Alcohol
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 27 21.642
## 2 22 19.483  5  2.1591 0.4876 0.7818

summary(m6)

```r
## Call:
## lm(formula = Metabol ~ Gastric + Sex, data = case1101[-c(31, 32), ])
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -1.81012 -0.49381 -0.05505  0.52307  2.03515
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```
## (Intercept)  0.8453  0.5681  1.488  0.148371
## Gastric     1.1498  0.2710  4.242  0.000232 ***
## SexFemale   -1.5276  0.3445 -4.434  0.000139 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8953 on 27 degrees of freedom
## Multiple R-squared: 0.65, Adjusted R-squared: 0.6241
## F-statistic: 25.07 on 2 and 27 DF,  p-value: 7e-07

# diagANOVA(m6)
par(mfrow=c(1,3))
plot(m6, which=c(1,2))  # Two of R's built-in residual plots includes case #'s
plot(case1101$Gastric[-c(31,32)], resid(m6))  # plot residuals vs the single predictor

Even though the residual plot fails to identify unusual data, let's check the case influence statistics. Note that $n = 30$ and $\hat{p} = 3$ in this scenario, so “large” leverages are $h_i > 2\hat{p}/n = 1/5 = 0.2$.

par(mfrow=c(1,2))
plot(m6, which=c(4,5))
This output suggests that we retain Sex in the model. More importantly, the planned research questions require that we retain Sex in the model as in model $m_6$ (this model is different than the final model obtained by your textbook):

$$
\mu \{\text{Metabol} | \text{Gastric, Sex}\} = \beta_0 + \beta_1 \text{Gastric} + \beta_2 \text{DummyF(Sex)}.
$$

ANSWER the research questions; be clear which model you are using for inference:

1. *Does the “first pass” stomach metabolism differ between the sexes?*

2. *Can the differences be explained by a difference in gastric enzyme activity?*

3. *Is there an effect due to alcoholism?*