BIOS 663 HW 4

Yue Jiang

March 24, 2015

Exercise 1.

In this exercise, we assess the full model which contains 102 observations, 17 predictors, and an intercept. Note that exploratory analysis of the dataset shows that there are no missing observations in any of the predictors, and there do not appear to be any systematic errors with data entry (that is, all data appear plausible for the predictors).

The full model is $\mathbf{y}_{102\times1} = \mathbf{X}_{102\times18}\beta_{18\times1} + \epsilon_{102\times1}$, where $\mathbf{y}_{102\times1}$ is the vector of 102 responses, $\mathbf{X}_{102\times18}$ is the full-rank design matrix which contains a column of all 1's (for our intercept) and the values of our 17 predictors, $\beta_{18\times1}$ is the vector of parameters, and $\epsilon_{102\times1}$ is a the vector of errors, assumed to be normally distributed with mean 0 and common variance σ^2.

Regarding our model assumptions, existence is satisfied given that we have a finite number of observations, and independence is assumed satisfied given that the investigators appropriately recruited their study population. As well, we assume a linear relationship between our response and predictors, which is supported by examining our residual plot.

However, homogeneity of variance is not satisfied; in the residual plot there is evidence of a fanning pattern, with larger values of our response having a larger variance. As well, the Q-Q plot suggests that normality of residuals may not be satisfied as well, with evidence of a heavy tailed distribution. Performing a log transformation of our response may be appropriate, as it helps alleviate variance homogeneity concerns. After performing the log transformation, we find that the variance of the residuals is more homogeneous, but the Q-Q plot of the log-transformed model still suggests lack of normality of residuals (albeit better than the original model).

Note that there was one outlying value which was a very influential observation. However, for the analyses that follow, we leave the observation in since there is no compelling reason to remove it from the dataset. Still, it is important to note.

Exercise 2.

To check for collinearity in the full model, we perform an eigenanalysis for a more in-depth exploration. For the full model, the condition number was 69.1 (adjusting for intercept), which suggests collinearity concerns. As well, VIFs were all quite large, ranging from 37.5 to 307.7, supporting our results from the eigen-analysis.

To perform adjustments to the maximum model, we examined results from eigen-analysis on the full model, removing highly collinear effects sequentially until the condition number fell below 30. At the end, we removed calories*LDL, calories*BMI, and calories*age in years.

Exercise 3.

Here, we conduct backward elimination on our full model on the log-transformed response. Our criterion will be significance level, and we will stop when all candidate effects for removal at an individual step are significant at the 0.10 level. As well, we force hierarchy on the model such that if we include an interaction effect in our model, we must include the main effects as well. The following table summarizes each step, showing the predictor removed, the significance level of that predictor, and the AIC for each step. Note as well that our final also has our optimal value of our AIC, in addition to satisfying the selection criteria regarding significance levels of predictors.

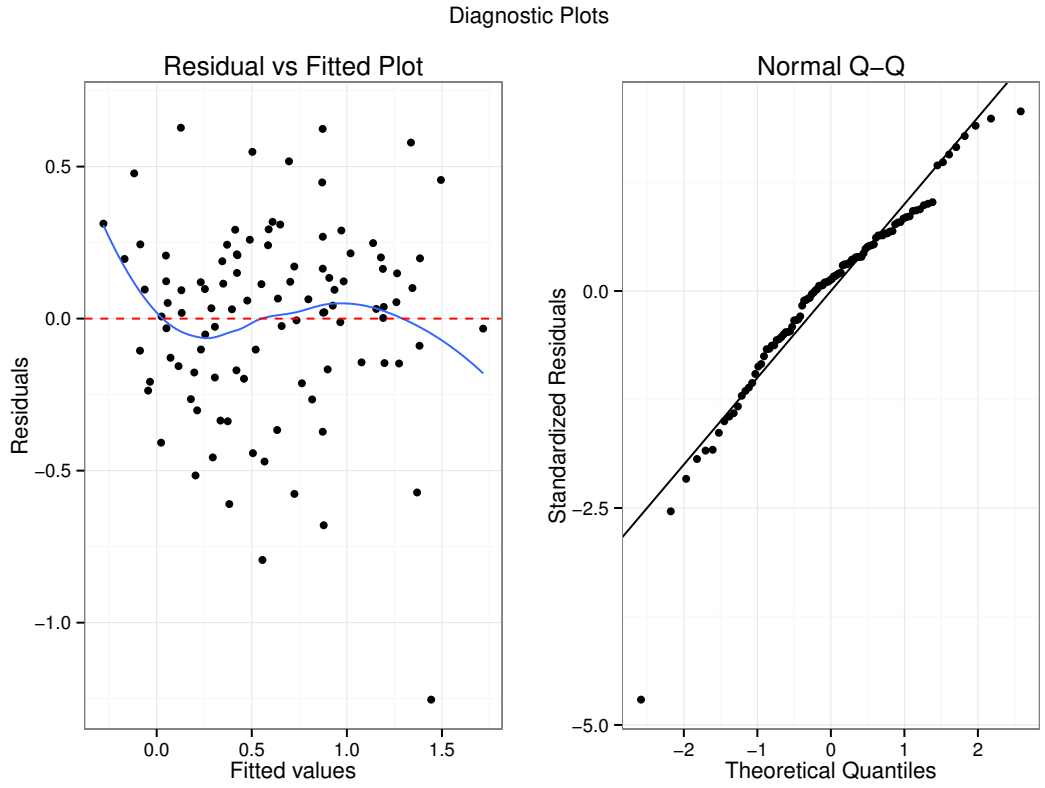
Step	Effect Removed	R^2	AIC	p-value
1	calories*apoa1	0.6904	-104.8604	0.9636
2	calories*pai	0.6904	-106.8379	0.8915
3	calories*fibrin	0.6902	-108.7879	0.8376
4	fibrin	0.6902	-110.7871	0.9799
5	calories*lpa	0.6900	-112.7159	0.8048
6	calories*bmi	0.6898	-114.6561	0.8197
7	bmi	0.9704	-116.5494	0.7596
8	calories*ageyrs	0.6882	-118.1277	0.5408
9	pai	0.6854	-119.2282	0.3690
10	calories*ldl	0.6793	-117.3022	0.1814
11	calories*apob	0.6773	-120.6344	0.4494

The following table shows estimated slopes for the final model:

Parameter	β	Estimate
Intercept	β_0	0.7811
ldl	β_1	-0.0200
lpa	β_2	-0.0050
apoa1	β_3	-0.0096
apob	β_4	0.0256
calories	β_5	0.0002
ageyrs	β_6	0.0107

Exercise 4.

As for model assumptions in our final model, we again assume existence and independence. Given our transformation, we assume a linear model in parameters in predicting the *log* response. The residual plot and Q-Q plots are given on the following page. The residuals show better homogeneity of variance, satisfying our assumption, but the normal Q-Q plot still shows slight evidence of non-normal residuals.



Exercise 5.

As we have log transformed the response, the interpretation of our coefficients is different. For the main effects, the straightforward interpretation is that a one-unit increase in x_i corresponds to a β_i increase in log TG/HDL ratio, controlling for other covariates. Translating that into our original units, we have that each additional unit increase in x_i *multiplies* our expected value of TG/HDL ratio by e^{β_i}, controlling for other covariates. For our intercept, we have that the expected value of log TG/HDL ratio is β_0 given that all covariates are 0, so in our original units, we have that the expected value of the TG/HDL ratio is e^{β_0} given that all covariates are 0.

Our interpretations are thus as follows:

β_0: Given that all our covariates are 0, we would expect a TG/HDL ratio of $e^{0.7811} = 2.184$.
β_1: Holding other covariates constant, each mg/dL increase in LDL multiplies our expected TG/HDL ratio by $e^{-0.0200} = 0.980$.
β_2: Holding other covariates constant, each mg/dL increase in Lp(a) multiplies our expected TG/HDL ratio by $e^{-0.0050} = 0.995$.
β_3: Holding other covariates constant, each mg/dL increase in Apo-A1 multiplies our expected TG/HDL ratio by $e^{-0.0096} = 0.990$.
β_4: Holding other covariates constant, each mg/dL increase in Apo-B multiplies our expected TG/HDL ratio by $e^{0.0256} = 1.026$.
β_5: Holding other covariates constant, additional unit increase in calories consumed per day multiplies our expected TG/HDL ratio by $e^{0.0002} = 1.0002$.
β_6: Holding other covariates constant, each additional year of age multiplies our expected TG/HDL ratio by $e^{0.0107} = 1.011$.