MATH 204 - ASSIGNMENT 2: SOLUTIONS

Although the original data indicate a non-linear relationship of FEV with height, and potentially complicated modelling, a log transformation of the response, yields a fairly simple linear relationship (see Figure 1), and an exploratory fit of a model involving height and height squared indicates that there is no need for the squared term.

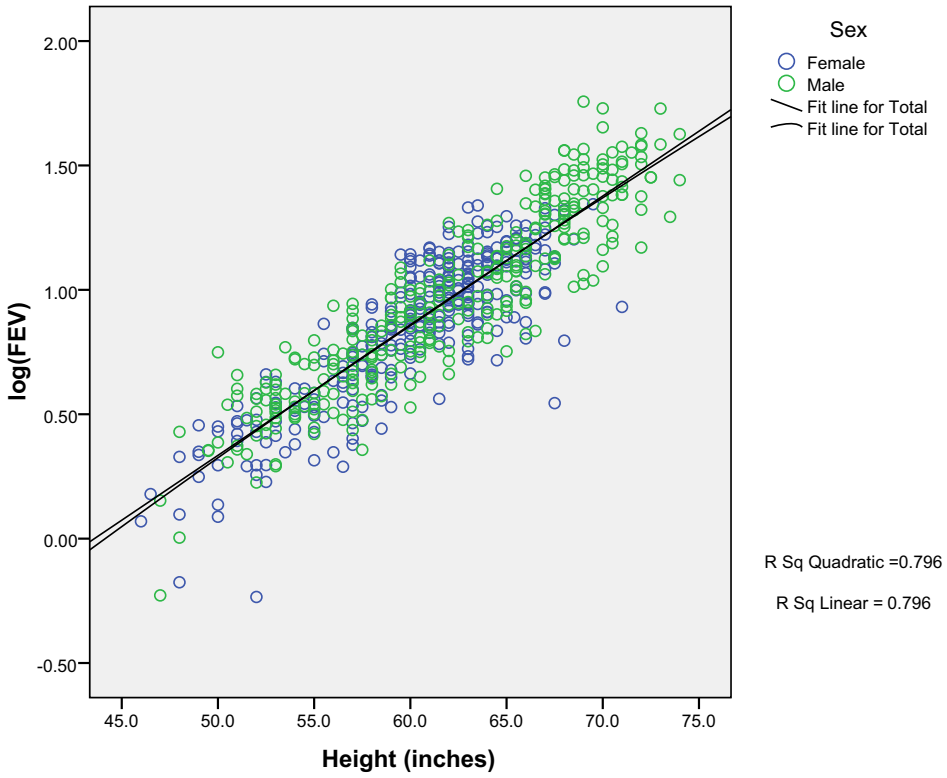


Figure 1: $\log(FEV)$ vs Height

The log transformation is also variance-stabilizing; in the original plot it was evident that the variability for larger response measurements was higher, whereas in the plot above the residual variance seems to be constant.

Given the plot above, it appears that a simple model will be adequate. Hence we try stepwise selection, beginning with the main effects only model

$$\mathrm{M}_0 \ : \ \text{Height} + \text{Age} + \text{Sex} + \text{Smoke}$$

At the first stage we try to remove one main effect. The table below contains the residual sums of squares for these (reduced) models. In this analysis, for all models, we have $\mathrm{SSE}_C = 13.734, n-k-1 = 649, k-g =$

Model	Terms	SSE_R	F	$F_{0.05}$	Significant ?
M_1	Age+ Sex+ Smoke	27.482	649.662	3.86	YES
M_2	Height+ Sex+ Smoke	14.766	48.77	3.86	YES
M_3	Height+ Age+ Smoke	13.866	6.23	3.86	YES
M_4	Height+ Age+ Sex	13.836	4.82	3.86	YES

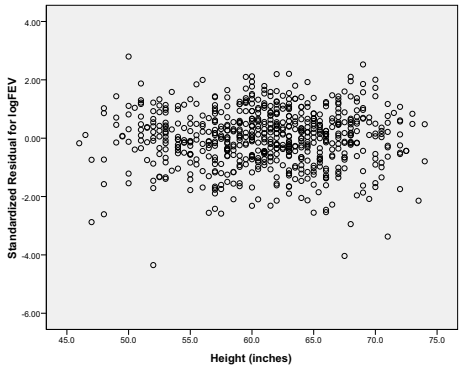
1 and F_α is the $1-\alpha$ quantile of the Fisher$(k-g, n-k-1) \equiv$ Fisher$(1, 649)$ distribution. Clearly all results are significant at $\alpha = 0.05$, indicating that the model M_0 cannot be simplified without a significant depreciation in fit.

For the next stage, we try to extend the model by including two-way interactions. There are $4 \times 3/2 = 6$ possible two-way interactions, and we add them in turn to M_0.

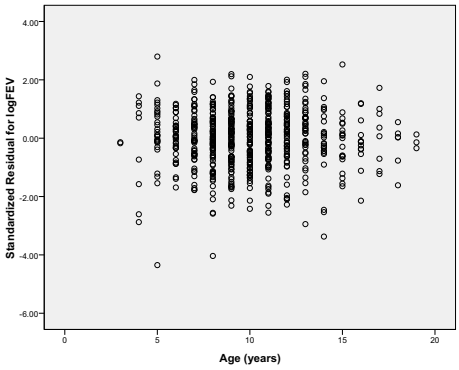
Model	Terms	SSE_C	F	$F_{0.05}$	Significant ?
M_5	M_0 + Height.Age	13.732	0.094	3.86	NO
M_6	M_0 + Height.Sex	13.727	0.330	3.86	NO
M_7	M_0 + Height.Smoke	13.732	0.094	3.86	NO
M_8	M_0 + Age.Sex	13.730	0.189	3.86	NO
M_9	M_0 + Age.Smoke	13.693	1.940	3.86	NO
M_{10}	M_0 + Sex.Smoke	13.730	0.189	3.86	NO

In this analysis, for all models, we have $SSE_R = 13.734, n-k-1 = 648, k-g = 1$. Thus it is not apparently useful to add terms to the model. Hence M_0 seems to be the most appropriate model; with an $R^2 = 0.811$, Adj. $R^2 = 0.809$, it seems that the global model fit is quite good.

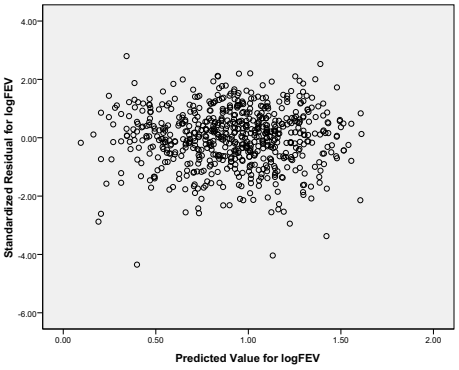
The only things remaining to be checked are the residuals. We plot the standardized residuals versus Height, Age, Predicted and Observed response.



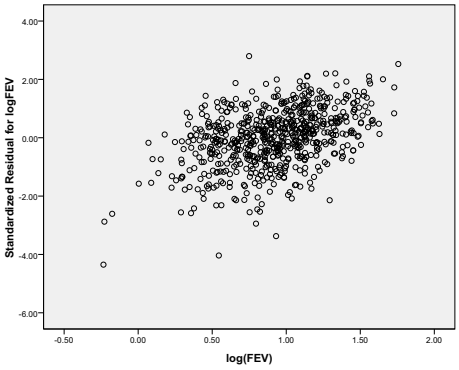
(a) Standardized Residuals vs Height



(b) Standardized Residuals vs Age



(c) Standardized Residuals vs Predicted log(FEV)



(d) Standardized Residuals vs Observed log(FEV)

Figure 2: Residual Plots

These plots are generally satisfactory; the slight positive correlation in (d) is acceptable. There are potentially a couple of outliers that might be omitted. Histograms and P-P plots for the standardized residuals indicate that the normality assumption is valid.

Note that it possible to get reasonable fit to the original scale data, using main effects and Height squared; the R^2 value is 0.794, and the residual plot indicates zero mean, but non-constant variance residuals. Hence the fit not perfect, but adequate for prediction purposes.