HW 8

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. The data set FEV.csv in Blackboard contains determinations of FEV for 654 children ages 3 through 19 who were seen in the Childhood Respiratory Disease (CRD) Study in East Boston, Massachusetts. These data are part of a longitudinal study to follow the change in pulmonary function over time in children. Variables in the data set or the participant ID number, Age (in years), FEV (in liters), Height (in inches), a binary Sex indicator (0 = female/1 = male), and Smoking status (0 = noncurrent smoker/1 = current smoker).

Consider all variables, Age, Height, Sex, and Smoking status, simultaneously in a multiple regression model.

To assess the association between FEV all four variables, I will examine the **multiple linear regression** of FEV on these variables, $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ where initially $y_i =$ FEV for individual i and $x_{i1} =$ Age, $x_{i2} =$ Height, $x_{i3} =$ is the Sex indicator (taking values of 0 for female and 1 for male), and $x_{i4} =$ Smoking status indicator taking the values of 0 for noncurrent smoker and 1 for current smoker. All for individual i = 1,…,654.

1. I first checked the assumptions of the multiple regression model using untransformed FEV and then, because there were definitely issues, I assessed the model using the log transformation of FEV. I discuss both here. These plots are given on the next pages. **If you jumped directly to assessing log(FEV)** because we know from HW 7 we will be using log(FEV) **– that is OK!**

i. Independent observations - We would need to know more from the study design to really assess this. We'll assume it holds.

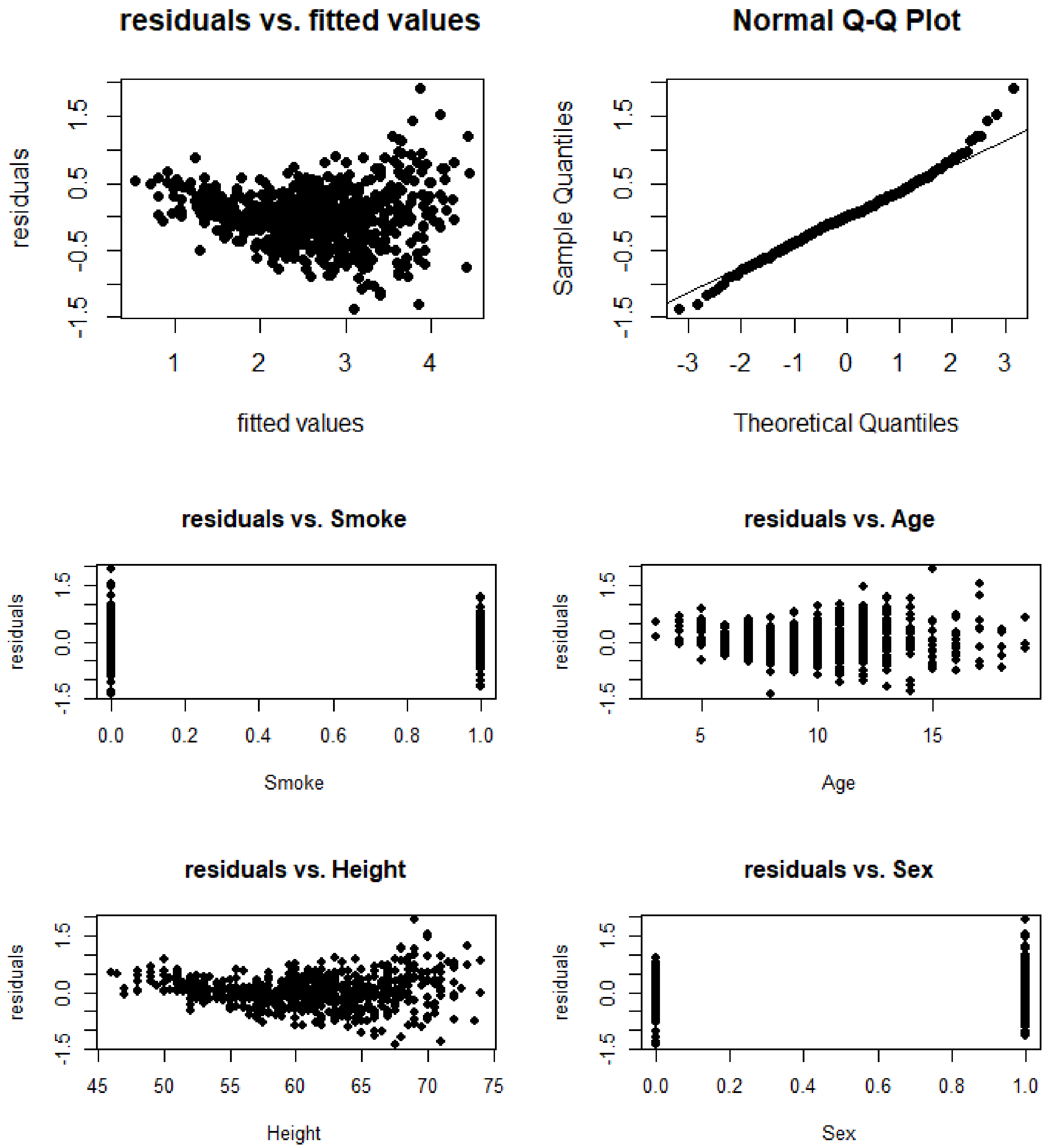
ii. The residual/errors follow a normal distribution - The QQ plot below shows some deviations from normality, particularly in the upper tail of the distribution. The QQ plot from the model using log transformed FEV indicates fewer deviations in the upper tail but possible some in the lower. The transformation doesn't appear to have caused issues. Moreover, with n = 654 I am not worried about small deviations from normality.

iii. With equal variances - From the scatterplot of the (studentized) residuals vs. predicted values using the untransformed FEV, there are huge issues with this. I notice two things, slight curvature (almost like a parabola – this is a separate issues) but also the *spread* of the residuals increases as you move from left to right on this plot. This is also evident in each of the residuals vs. x_{ij} plots, differences in spread of the residuals with Age and Height and between smoking and sex groups. The log transformed residuals vs. predicted values indicates that the spread of the residuals remains reasonably constant. This also holds in the residuals vs. x_{ij} plots, each of which suggests the equal variances assumption is reasonably met.

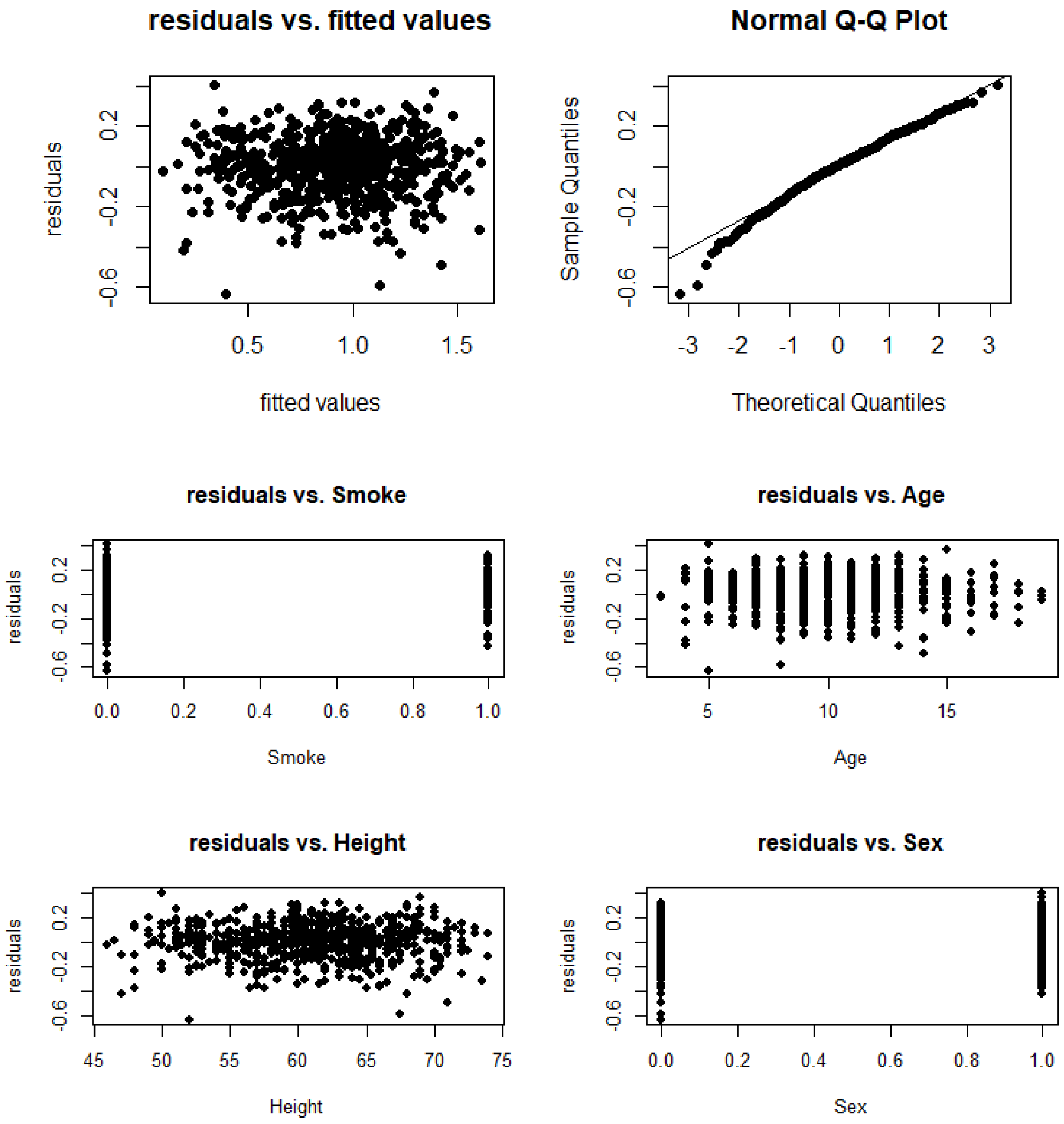
iv. There is a linear association between FEV and each of Age and Height (the only quant covariates in the model) – OK, knowing we're going to use the log transformed FEV because of the equal variances assumption, it's not worth commenting on FEV so I'll focus on the residuals form the model using y = ln(FEV). Examination of residuals vs. predicted, residuals vs. Age, and residuals vs. Height plots each indicate that this assumption is met. Unlike the residual plots for FEV versus these values, there are no trends evident in these.

v. There are no outliers influencing the results. - There are at least two, maybe three, observations with low residuals. We'll explore outliers more in the next problem using our diagnostic tools.

Residual plots from model using *untransformed* FEV:



Residual plots from model using *log transformed* FEV – huge improvements:

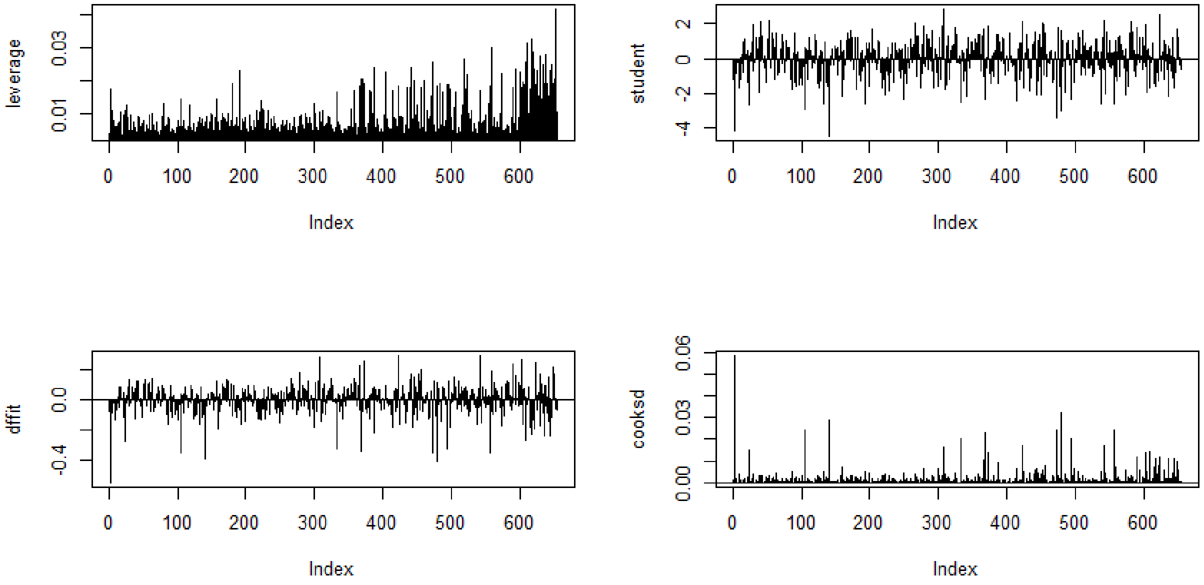


2. There are two observations with studentized residuals that are less than -4.0 (these are the two lowest observations in all the residual plots above). These two observations correspond to the following two individuals:

   Row 2, ID =451, FEV = 1.724 liters, log(FEV) = 0.545, Age = 8, Height = 67.5 inches, female noncurrent smoker
   Row 140, ID = 33351, FEV = 0.791 liters, log(FEV) = -0.234, Age = 5, Height = 52 inches, female noncurrent smoker

   Row 2 observations also has the highest (in magnitude) DFFIT and Cooks Distance.

   Row 652, (ID = 73751, FEV = 2.853, Age = 18, Height = 60, female noncurrent smoker) has the largest leverage – in fact many observations past the 550th row have high leverages (and several other observations). A closer look at the leverages reveals that all individuals with Smoke = 1 have high leverages (and about a dozen with Smoke = 0). As the leverage seems to be primarily highlighting the current smokers, I will focus only on those observations with extreme residuals, DFFIT values, and Cooks Distance (i.e. row 2 and 140).



3. Based on the residual analysis and diagnostics, I will focus the interpretation on the full data, log transformed FEV model. The model fit summary is on the next page. The test of the regression effect tests the equal means reduced model vs. the full multiple linear regression model (with log transformed FEV) - Ho: the equal means (one-sample) model , $y_i = \beta_0 + \varepsilon_i$ vs. Ha: , $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$. From the output on the next page, F-statistic = 694.6 on 4 and 649 DF giving p-value: < 2.2e-16. There is a lot of evidence against Ho in favor of a regression effect of age, height, sex, and smoking status.

4. From the model fit summary we see that R^2 = 0.8106 thus 81.1% of the variation in ln(FEV) can be explain by the multiple regression with Age, Height, Sex, and Smoking status.

5. The coefficient table below gives the following estimates for the explanatory variables in the model. The Age coefficient estimate is $\hat{\beta}_1 = 0.023$ (with 95% confidence interval of 0.017 to 0.030) with p-value for a test of the Age effect (given Height, Sex, and Smoking status are in the model) is p < 0.001. So Age is a significant predictor, after adjusting for the other explanatory variables. Likewise, Height has coefficient estimate of $\hat{\beta}_2 = 0.043$ (with 95% confidence interval of 0.039 to 0.046) and is significantly associated with ln(FEV) after adjusting for Age, Sex, and Smoking status (p-value < 0.001). The Sex coefficient estimate is $\hat{\beta}_3 = 0.029$ with 95% confidence interval of 0.006 to 0.052 which is also significant (p-value = 0.013) as is Smoking status, with coefficient estimate $\hat{\beta}_4 = -0.046$ and 95% confidence interval of -0.087 to -0.005 and p-value = 0.028.

6. The 95% confidence interval for β_4 is -0.087 to -0.005. Thus, we are 95% confident that, adjusting for Age, Height, and Sex, current smokers (status = 1) have lower average ln(FEV) by 0.087 to 0.005 log liters than noncurrent smokers (status = 0). Notice, that in HW 7, current smokers had higher average ln(FEV) values (i.e. higher pulmonary function) but after adjusting for Age, Height, and Sex, we estimate their pulmonary function is *lower*, on average.

   Note: we can exponentiate to get inferences about the median FEV: medianFEV of current smokers = exp(0.087) = 0.9165604 to exp(0.005) = 0.9950048 times the medianFEV of nonsmokers. (You can take the reciprocal to estimate that the median FEV of nonsmokers is estimated to be 1.005 to 1.09 (.5% to 9%) higher than the median FEV of smokers, with 95% confidence.

7. A sensitivity analysis suggests these two observations are not influencing the results of the analysis, especially the interpretation of the association of average logFEV (median FEV) with smoking status. (Output w/o these observations is on the last page).

Model fit output from model using *log transformed* FEV:

```
> fit <- lm(log(FEV)~Age+Hgt+Sex+Smoke,data=data)
> summary(fit)

Call:
lm(formula = log(FEV) ~ Age + Hgt + Sex + Smoke, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63278 -0.08657  0.01146  0.09540  0.40701 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.943998   0.078639 -24.721  < 2e-16 ***
Age          0.023387   0.003348   6.984  7.1e-12 ***
Hgt          0.042796   0.001679  25.489  < 2e-16 ***
Sex          0.029319   0.011719   2.502   0.0126 *  
Smoke       -0.046068   0.020910  -2.203   0.0279 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.1455 on 649 degrees of freedom
Multiple R-squared:  0.8106,  Adjusted R-squared:  0.8095
F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16>
confint(fit)
                   2.5 %       97.5 %
(Intercept) -2.098414941 -1.789581413
Age          0.016812109  0.029962319
Hgt          0.039498923  0.046092655
Sex          0.006308481  0.052330236
Smoke       -0.087127344 -0.005007728

> exp(confint(fit))
                2.5 %    97.5 %
(Intercept) 0.1226507 0.1670301
Age         1.0169542 1.0304157
Hgt         1.0402894 1.0471714
Sex         1.0063284 1.0537237
Smoke       0.9165604 0.9950048
```

Model fit output from model using *log transformed* FEV with rows 2 and 140 temporarily held out:

```
Call:
lm(formula = log(FEV) ~ Age + Hgt + Sex + Smoke, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.49439 -0.08539  0.01135  0.09072  0.40642 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.965881   0.076987 -25.535  < 2e-16 ***
Age          0.021491   0.003279   6.554 1.14e-10 ***
Hgt          0.043532   0.001646  26.450  < 2e-16 ***
Sex          0.024454   0.011439   2.138   0.0329 *  
Smoke       -0.045157   0.020358  -2.218   0.0269 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1416 on 647 degrees of freedom
Multiple R-squared:  0.8174,    Adjusted R-squared:  0.8163 
F-statistic:   724 on 4 and 647 DF,  p-value: < 2.2e-16

> confint(fit)
                   2.5 %       97.5 %
(Intercept) -2.117056068 -1.814706725
Age          0.015052346  0.027930045
Hgt          0.040300287  0.046763805
Sex          0.001991257  0.046916675
Smoke       -0.085133458 -0.005180823
```

```
> exp(confint(fit))
                2.5 %    97.5 %
(Intercept) 0.1203855 0.1628857
Age         1.0151662 1.0283237
Hgt         1.0411234 1.0478745
Sex         1.0019932 1.0480347
Smoke       0.9183897 0.9948326
```