STAT 614 - HW 8 – Yunting Chiu

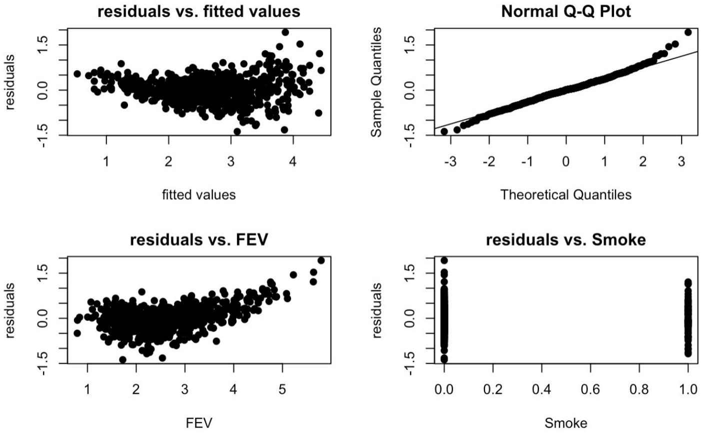
Due: Thursday, December 3, 2020 in Blackboard by 11:59pm.
Instructions: Please type your solutions in a separate document and upload the document in Blackboard as a pdf. I will not be collecting syntax for this assignment. You will need concepts fromChapters 9 through 12 on the multiple linear regression model and the results from HW 7.

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 are 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 = non-smoker/1 = current smoker).
Consider all variables, Age, Height, Sex, and Smoking status, simultaneously in a multiple regression model.

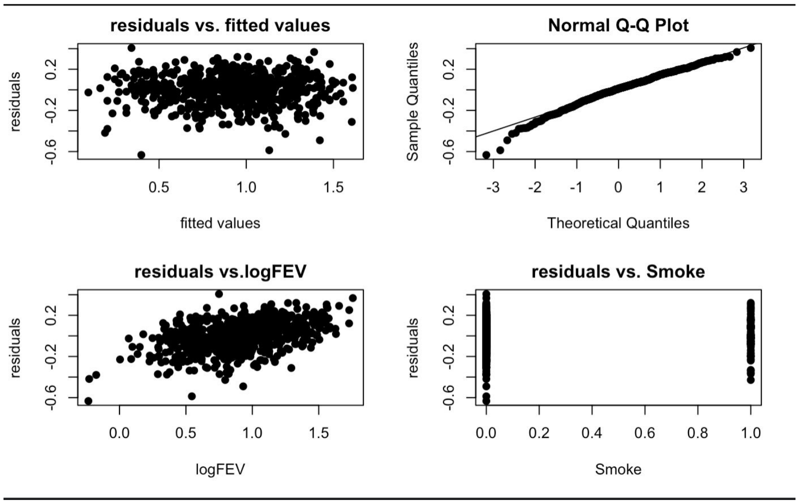
1. Assess the assumptions of the model and make any adjustments. Be sure to look at all residual plots. Make any necessary adjustments.

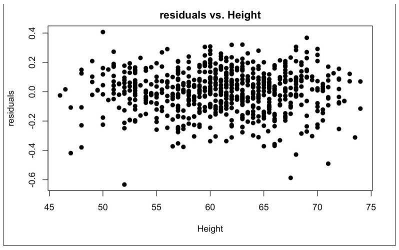
The assumptions of the multiple regression model were assessed using our traditional residual analysis:
- Independent observations: met
- Normally distribution
- Equal variances
- No influential outliers
- Linear association between (mean) y and each explanatory variable (x1 and x2)

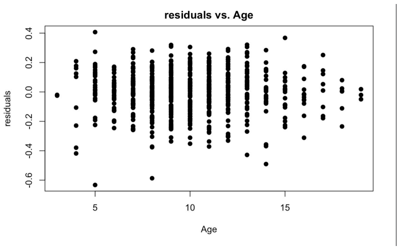
The residual plots do not look like equal variances, so we try to transform FEV.

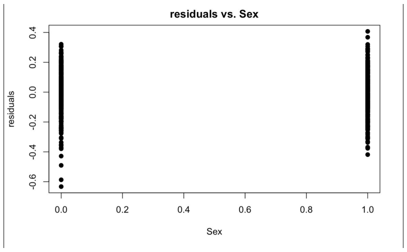


After we take a log of FEV, the residues are randomly distributed along the center line of zero, without a consistent pattern. The residues therefore have a constant variance, roughly normally distributed, and are independent, which met the assumptions of multiple regression model.



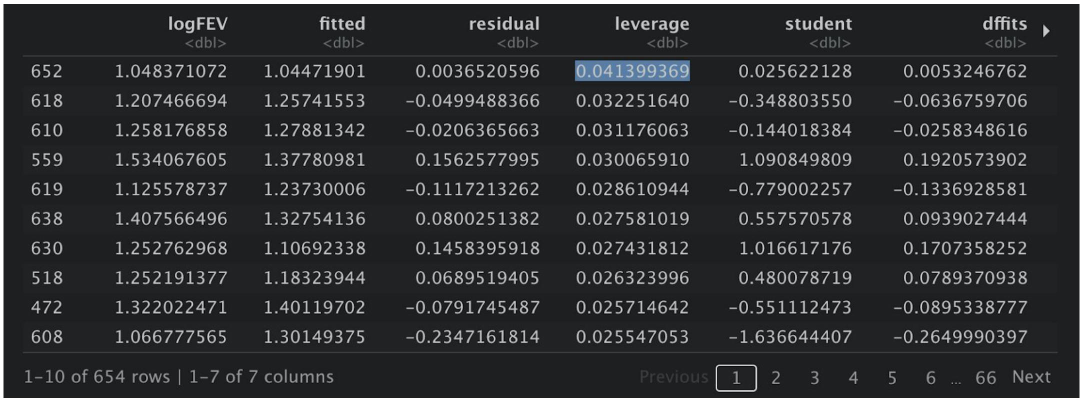






2. Use the diagnostic tools to identify potential influential observations. Which observations are flagged as potentially being influential? (Note: you'll deal with these in 7 below).

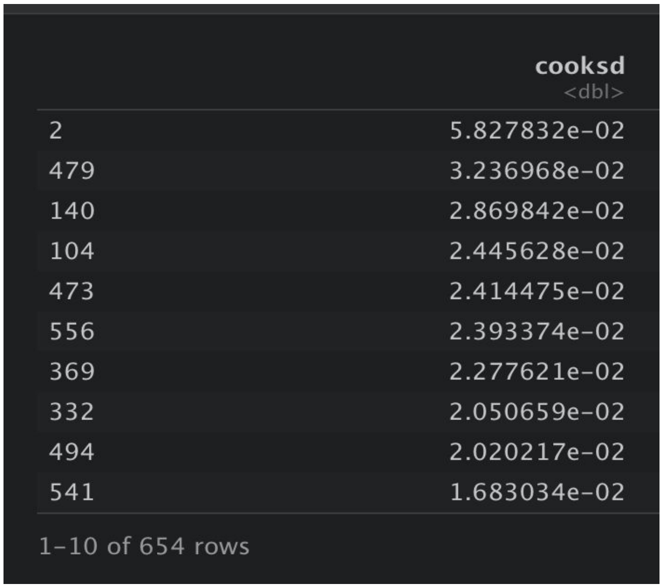
Observation 652 has the largest leverage which at 0.041399369 is above the average leverage of 2(k+1)/n = 2(4+1)/654 = 0.015, for this dataset. This observation's observed logFEV (y652 = 1.048371072 is higher than predicted (ŷ 652 = 1.04471901) from the model, for this particular FEV and all variables (without ID) combination.



	logFEV <dbl>	fitted <dbl>	residual <dbl>	leverage <dbl>	student <dbl>	dffits <dbl>
652	1.048371072	1.04471901	0.0036520596	0.041399369	0.025622128	0.0053246762
618	1.207466694	1.25741553	-0.0499488366	0.032251640	-0.348803550	-0.0636759706
610	1.258176858	1.27881342	-0.0206365663	0.031176063	-0.144018384	-0.0258348616
559	1.534067605	1.37780981	0.1562577995	0.030065910	1.090849809	0.1920573902
619	1.125578737	1.23730006	-0.1117213262	0.028610944	-0.779002257	-0.1336928581
638	1.407566496	1.32754136	0.0800251382	0.027581019	0.557570578	0.0939027444
630	1.252762968	1.10692338	0.1458395918	0.027431812	1.016617176	0.1707358252
518	1.252191377	1.18323944	0.0689519405	0.026323996	0.480078719	0.0789370938
472	1.322022471	1.40119702	-0.0791745487	0.025714642	-0.551112473	-0.0895338777
608	1.066777565	1.30149375	-0.2347161814	0.025547053	-1.636644407	-0.2649990397

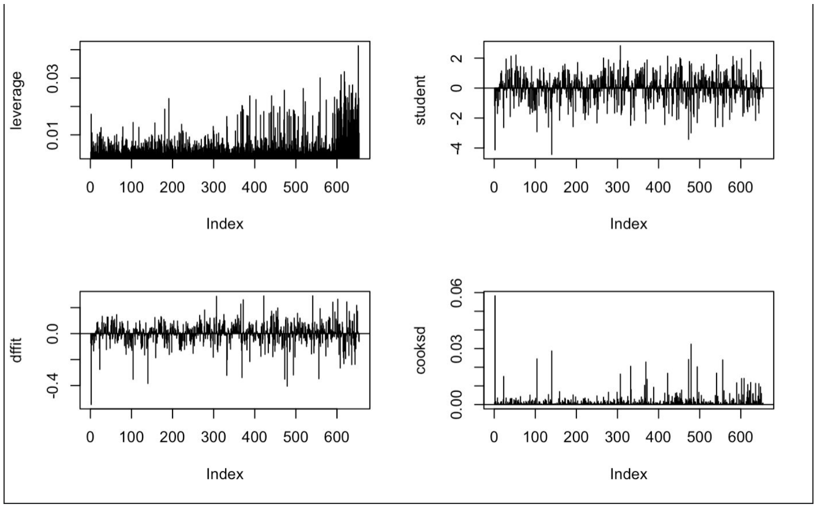
1-10 of 654 rows | 1-7 of 7 columns

Observation 2 has the highest Cock's Distance. We can find the descending order of cooksd's values below. I'll remove these two observations in question 7.



	cooksd <dbl>
2	5.827832e-02
479	3.236968e-02
140	2.869842e-02
104	2.445628e-02
473	2.414475e-02
556	2.393374e-02
369	2.277621e-02
332	2.050659e-02
494	2.020217e-02
541	1.683034e-02

1-10 of 654 rows



3. Is there evidence of a regression effect? Write the appropriate null and alternative hypotheses. Give the test statistic, p-value, and conclusions of the test.

Ho: equal means model vs. Ha: at least one of β1, β2, β3, β4 is not 0.
The result of ANOVA table indicates all x-variables are statistically significant so that we can see the model has separate means.

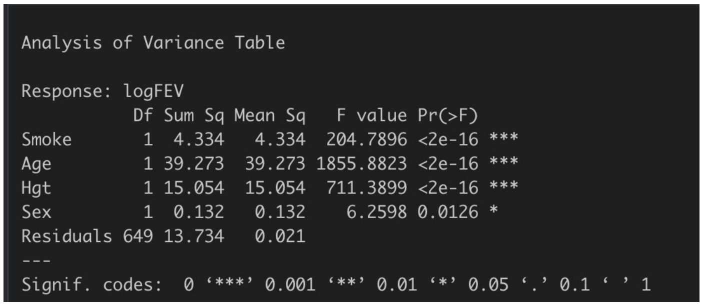
$H_0: y_i = \beta_0 + \varepsilon_i$

vs.

$H_a:$

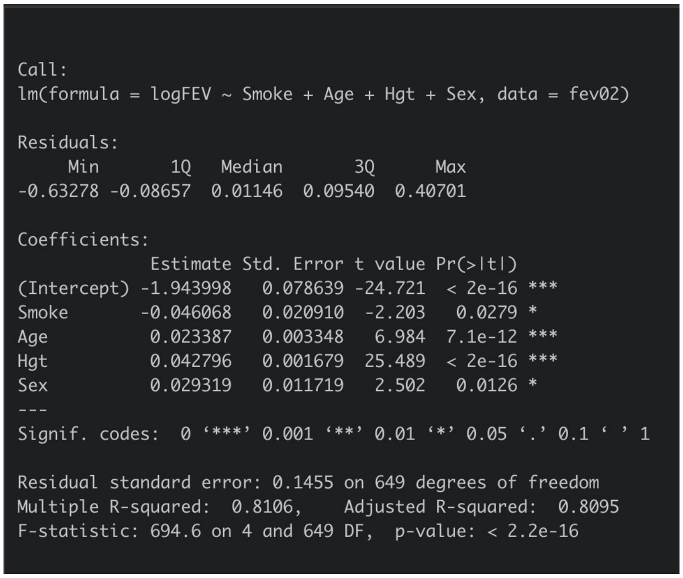
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2)$



```
Analysis of Variance Table

Response: logFEV
           Df Sum Sq Mean Sq   F value Pr(>F)
Smoke       1  4.334   4.334  204.7896 <2e-16 ***
Age         1 39.273  39.273 1855.8823 <2e-16 ***
Hgt         1 15.054  15.054  711.3899 <2e-16 ***
Sex         1  0.132   0.132    6.2598 0.0126 *
Residuals 649 13.734   0.021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The estimated mean log of forced expiratory volume from the fitted regression is ŷ = -1.943998 -0.046068 Smoke + 0.023387 Age + 0.042796 Height + 0.029319 Sex. The corresponding standards errors for each coefficient estimate are 0.078639 for the intercept, others have shown below. P-value = P(F >= 694.6) which is approximately < 2.2e-16 where F has a F(4, 649) distribution.



```
Call:
lm(formula = logFEV ~ Smoke + Age + Hgt + Sex, data = fev02)

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 ***
Smoke       -0.046068   0.020910  -2.203   0.0279 *  
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 *  
---
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
```

4. Give and interpret the coefficient of determination (R^2). What is the Adjusted-R^2 value?

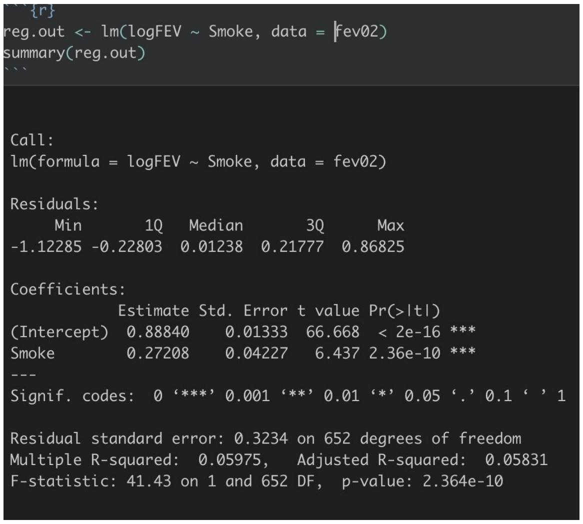
R-squared measures the proportion of the variation in the dependent variable (logFEV) explained by the independent variables (Smoke + Age + Hgt + Sex) for a linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables (totally 4) in the model. R^2 indicates that 81% of the variation in the outcome has been explained just by predicting the outcome using the covariates included in the model, that the model is a good fit for the data. Adjusted-R^2 value is 0.81. Because the adjusted-R^2 increases only if the increase in R^2 is greater than one would expect from chance alone, so we estimate this model has 81 % increased when Smoke , Age , Hgt, Sex variables are significant and affects logFEV.

Note: $R^2 = 1 - (SSE/SSTotal)$. Adjusted $R^2 = 1 - (MSE/MSTotal)$
Reference: https://www.britannica.com/science/coefficient-of-determination

5. Which of the four explanatory variables have “significant” associations with FEV, after adjusting for the other variables?
Smoking status, Age, Height, and Sex have significant level with logFEV. That is, all of the x-variables have significant associations with logFEV (p-values < 0.05).

6. Find and interpret the 95% confidence interval for the coefficient of **Smoking status**. This is the adjusted estimate (adjusting for Age, Height, and Sex). How does this adjusted estimate and CI compare to the unadjusted analysis from HW7?

Recall HW7:
simple linear regression of Smoking status vs . multiple linear regression Smoking status



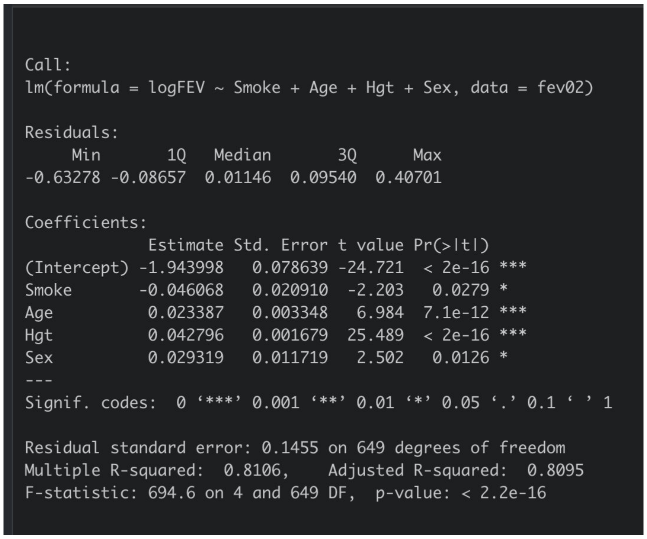
~~~
```{r}
reg.out <- lm(logFEV ~ Smoke, data = fev02)
summary(reg.out)
~~~

Call:
lm(formula = logFEV ~ Smoke, data = fev02)

Residuals:
     Min   1Q   Median   3Q   Max
-1.12285 -0.22803 0.01238 0.21777 0.86825

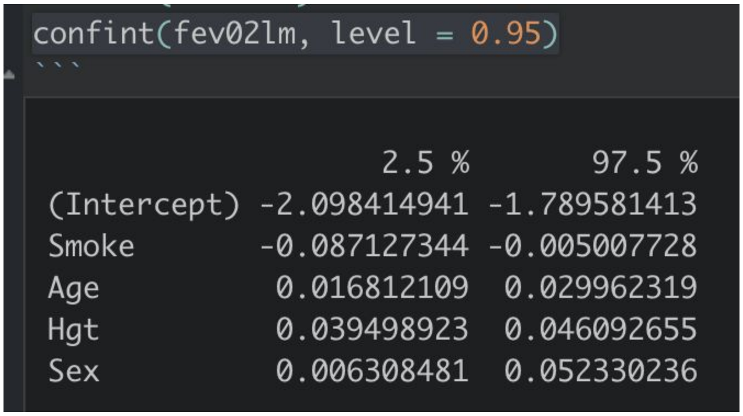
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.88840   0.01333 66.668 < 2e-16 ***
Smoke   0.27208   0.04227   6.437 2.36e-10 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3234 on 652 degrees of freedom
Multiple R-squared: 0.05975,	Adjusted R-squared: 0.05831
F-statistic: 41.43 on 1 and 652 DF, p-value: 2.364e-10
```



```
Call:
lm(formula = logFEV ~ Smoke + Age + Hgt + Sex, data = fev02)

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

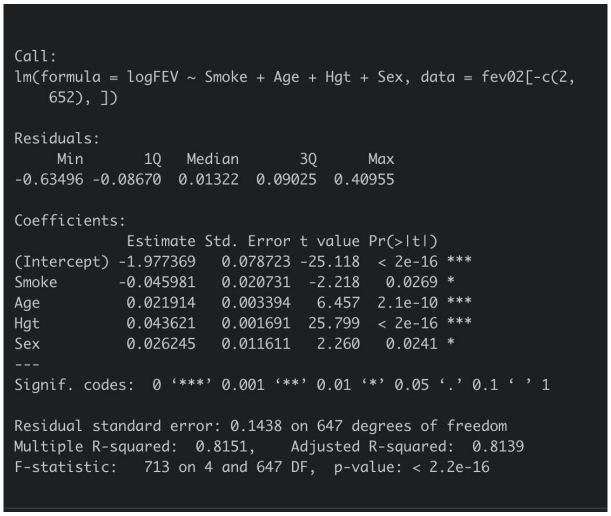
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.943998   0.078639 -24.721  < 2e-16 ***
Smoke       -0.046068   0.020910  -2.203   0.0279 *  
Age          0.023387   0.003348   6.984  7.1e-12 ***
Hgt          0.042796   0.001679  25.489  < 2e-16 ***
Sex          0.029319   0.011719   2.502   0.0126 *  
---
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(fev02lm, level = 0.95)
```

                   2.5 %   97.5 %
(Intercept) -2.098414941 -1.789581413
Smoke   -0.087127344 -0.005007728
Age   0.016812109 0.029962319
Hgt   0.039498923 0.046092655
Sex   0.006308481 0.052330236
```

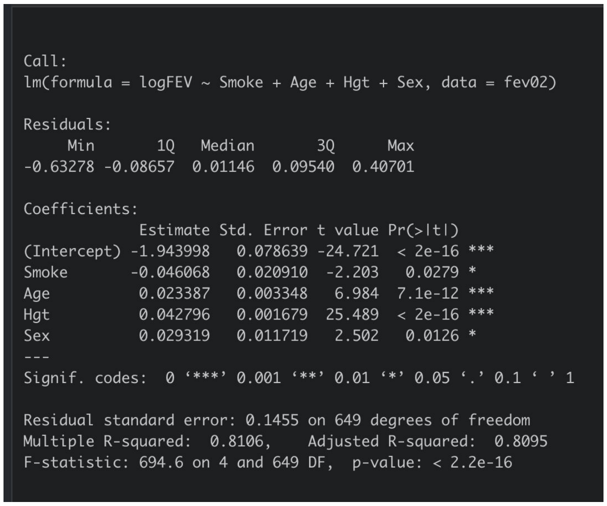
In the simple regression case (one variable plus the intercept), for every one level increase in Smoke (0 for non-smokers and 1 for current smokers), the model predicts an increase of 0.27208 FEV / liters. By contrast, in the multiple linear regression of Smoking status, for every one level increase in Smoke (0 for non-smokers and 1 for current smokers), the model predicts an decrease of 0.046068 FEV / liters. With a high adjusted-R² value of MLR model, we have in favor of the result of MLR model with Smoking status.

7. Temporarily hold out any outliers you identified in (2). Do any of the results in (3) – (6) change when holding out the outliers? That is, were the outliers influencing the conclusions? If so, discuss the differences.

MLR model of holding out observations 2 and 652 vs. MLR full model



```
Call:
lm(formula = logFEV ~ Smoke + Age + Hgt + Sex, data = fev02[-c(2,
    652), ])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63496 -0.08670  0.01322  0.09025  0.40955 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.977369   0.078723 -25.118  < 2e-16 ***
Smoke       -0.045981   0.020731  -2.218   0.0269 *  
Age          0.021914   0.003394   6.457  2.1e-10 ***
Hgt          0.043621   0.001691  25.799  < 2e-16 ***
Sex          0.026245   0.011611   2.260   0.0241 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1438 on 647 degrees of freedom
Multiple R-squared:  0.8151,	Adjusted R-squared:  0.8139 
F-statistic:   713 on 4 and 647 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = logFEV ~ Smoke + Age + Hgt + Sex, data = fev02)

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 ***
Smoke       -0.046068   0.020910  -2.203   0.0279 *  
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 *  
---
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
```

After we removed the potentially influential outliers, the results in (3) – (6) did not change. In other words, these two observations have not influenced the conclusions, the values of R² and adjusted-R² are still similar. In 95 % confidence intervals, the revised model does not change the overall interpretation of the impact of four x-variables. For instance, the slope of Smoking status decreases from -0.0459 to -0.0460 in the model without observation 2 and 652 but still indicates a negative association.