**STAT 614 - HW 8 – LAST ONE!!!!**

By Sihyuan Han

**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 from Chapters 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.

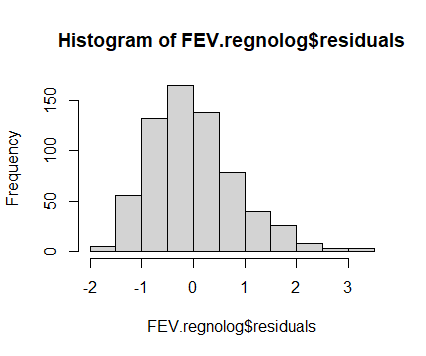
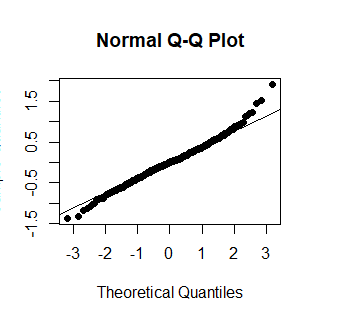
1. Assess the assumptions of the model and make any adjustments. Be sure to look at *all* residual plots. Make any necessary adjustments.

**Ans**:

Assumptions:

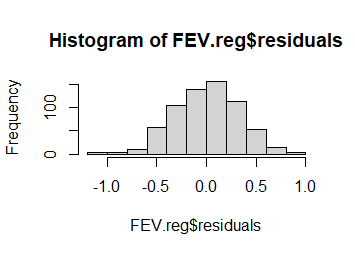
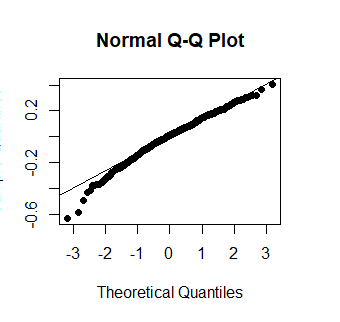
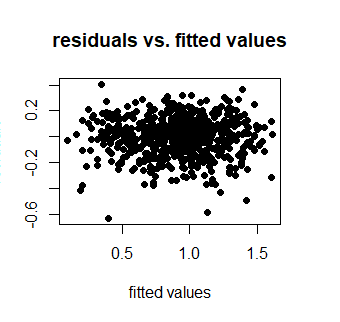
1. Normally distributed
2. Equal variances
3. Influential outliers
4. Independents
5. Linear association

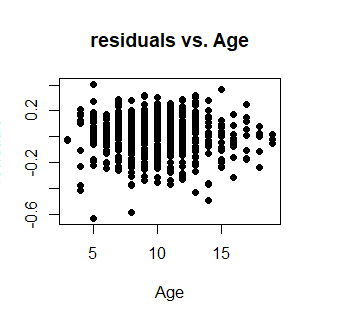
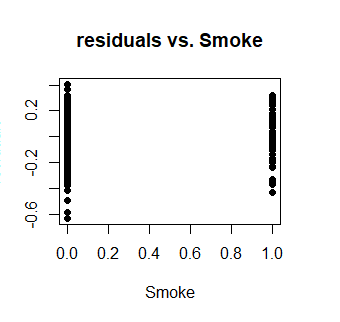
Before log FEV

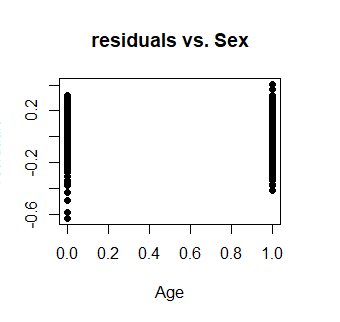
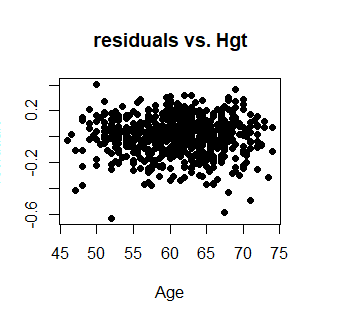


After log FEV

The following QQ plot shows more linear compare to the one above. Residual plots are show in below as well.

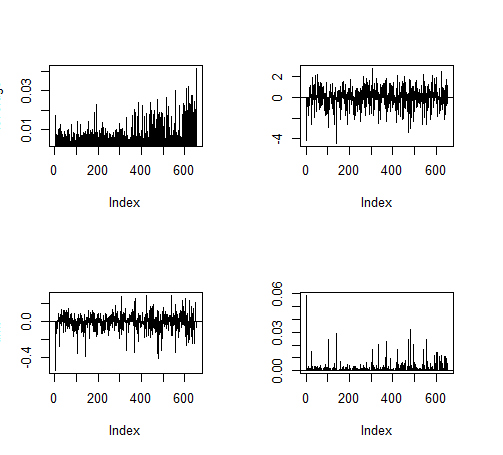
 

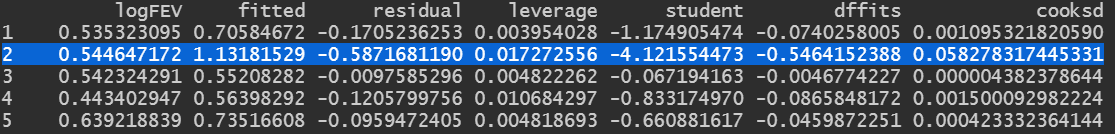


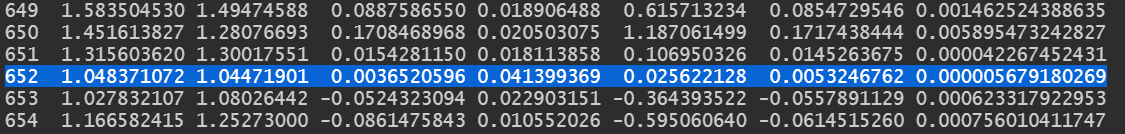


1. Use the diagnostic tools to identify potential influential observations. Which observations are flagged as potentially being influential? (Note: you’ll deal with these in 7 below).

**Ans**:

leverage, student, dfs and cooksd

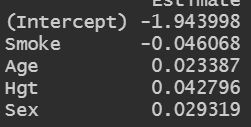
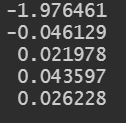
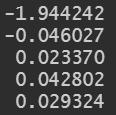
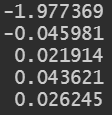




Observations 2 and 652 are highlighted as having extreme values of these diagnostics. Observation 2 has the highest Cook's Distance. Observation 652 has the largest leverage which at 0.041399369 is above the average leverage of 2(k+1)/n = 2(4+1)/654 = 0.01529052, for this dataset. This observation’s observed logFEV (y652 = 1.048371072 is higher than predicted (yhat652 = 1.04471901) from the model for this combination (Smoke, Age, Hgt, Sex).

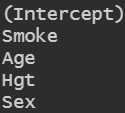
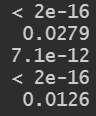
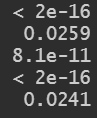
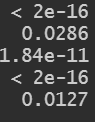
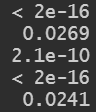
Model term coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | full | no2 | no652 | no2 & 652 |

Model term p-value:

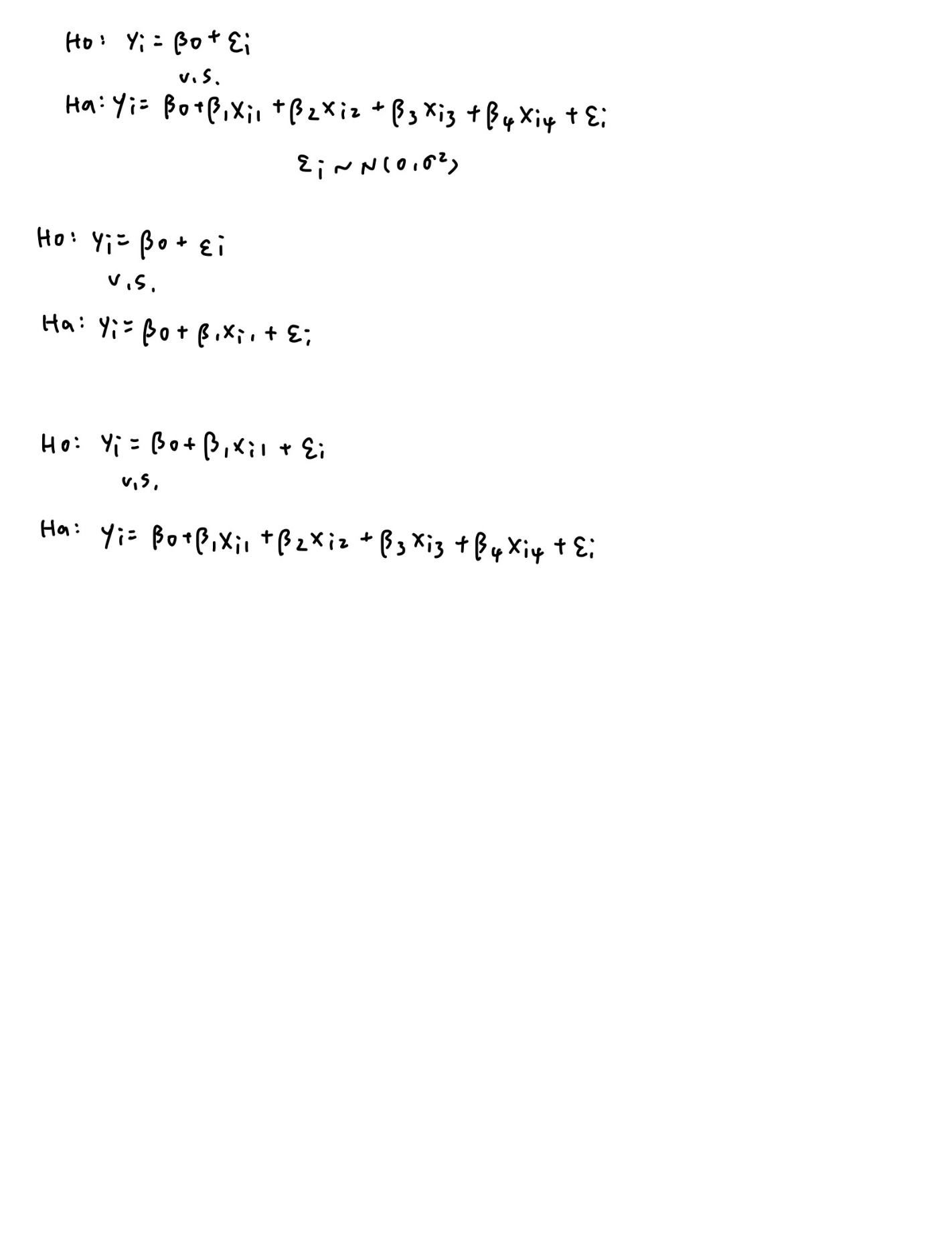
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | full | no2 | no652 | no2 & 652 |

Holding out observation 19 has little impact on the overall fit and interpretation. Holding out 17 alone, or in conjunction with 19, does not change the overall interpretation of the impact of lsize or bodwt. The slope of lsize decreases from 0.007 to 0.004 in the model without 17 but still indicates a positive association. We do see the significance of this effect change substantially, from p = 0.04751 to p = 0.3066.

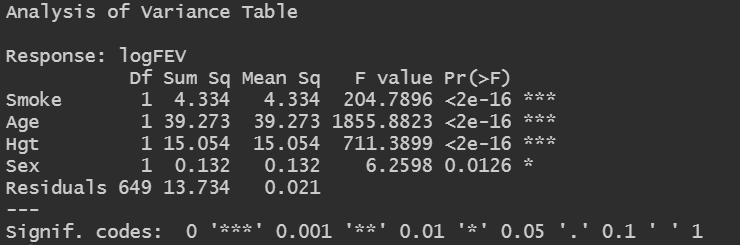
1. Is there evidence of a regression effect? Write the appropriate null and alternative hypotheses. Give the test statistic, p-value, and conclusions of the test.

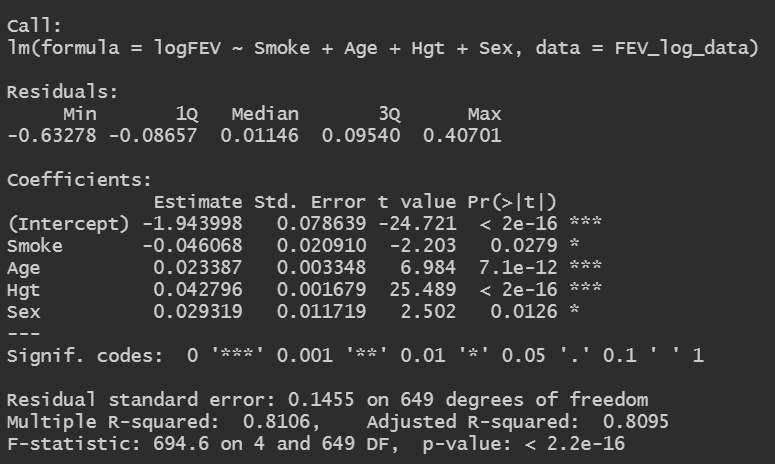
**Ans**:



H0: β1 = β2 = 0

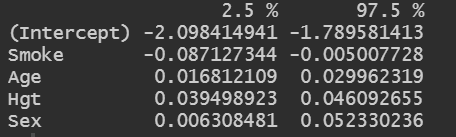
Ha: at least one of β1, β2 is not 0





The estimated mean logFEV from the fitted regression is: yhat = -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, and other data are also shown in the plot above.

The following output shows the CI:

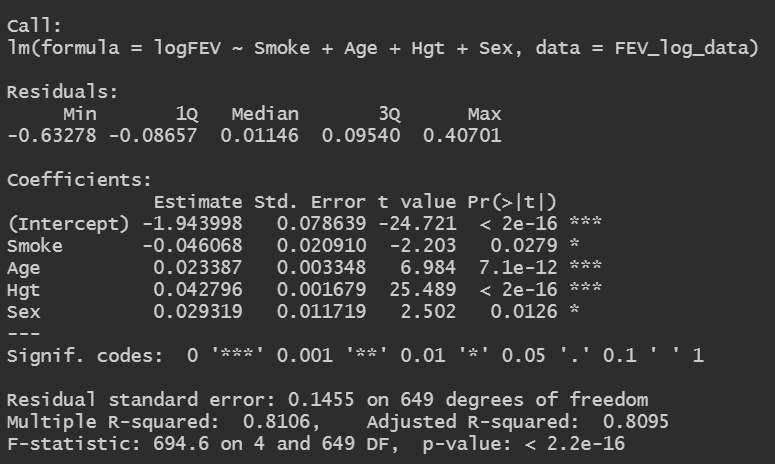


1. Give and interpret the coefficient of determination (R2). What is the Adjusted-R2 value?

**Ans**: Coefficient of determination (R2) is the portion of the total variability in y (FEV in this study) explained by the multiple regression model. R2 = 1 - (SSE/SSTotal). Adjusted R2 = 1 - (MSE/MSTotal). R2 should never decrease when adding more variables to the model. In this case, Adjusted R2 = 0.8095, which is approximately 81%.

1. Which of the four explanatory variables have “significant” associations with FEV, after adjusting for the other variables?

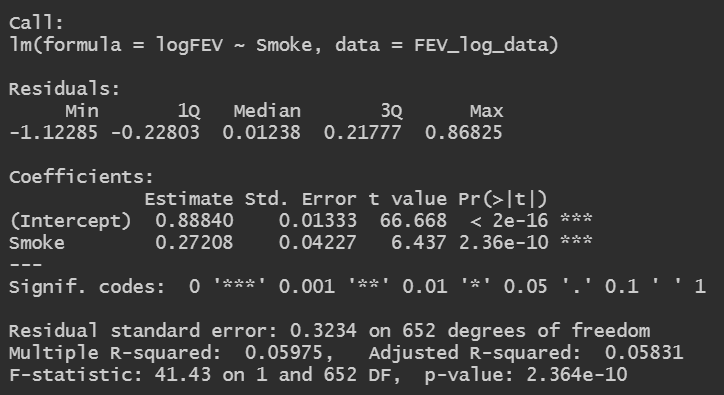
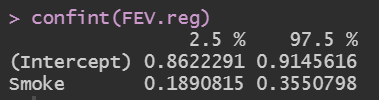
**Ans**: All of the four explanatory variables have associations with FEV because we can see from the following plot that p-values are all < 0.05. Especially for Age and Hgt which have significant associations with FEV after log of FEV.



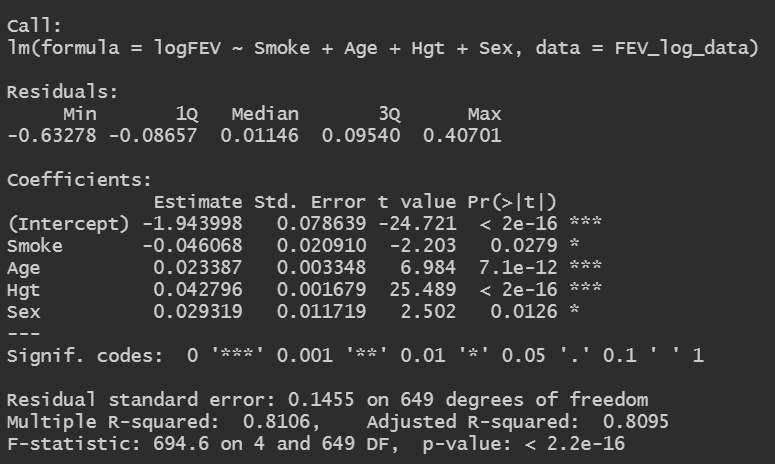
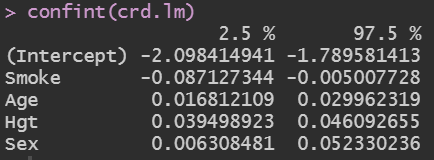
1. Find and interpret the 95% confidence interval for the coefficient of **Smoking** status. This is the adjusted estimate (adjusting for Age, Height, and Sex). How does this adjusted estimate and CI compare to the unadjusted analysis from HW7?

**Ans**:

From hw7:

hw8:

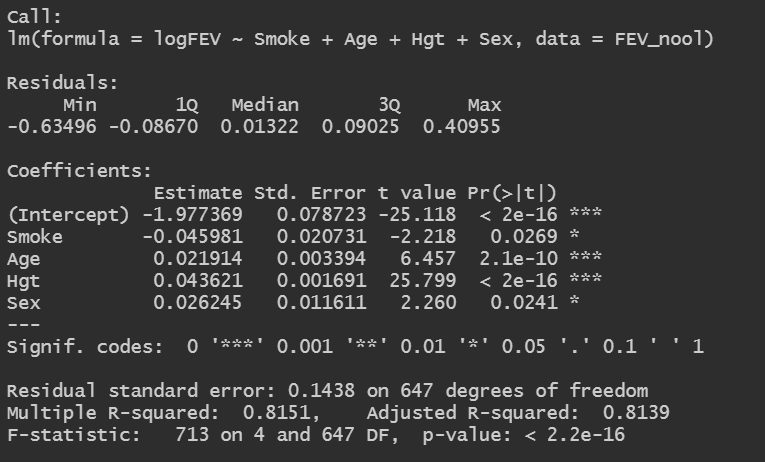
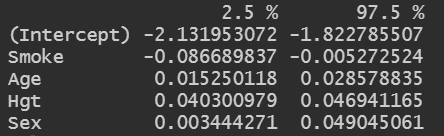
 

As we can see from the data above, Adjusted R2 changes from 0.05831 to 0.8095, which is better because Adjusted R2 should be larger!

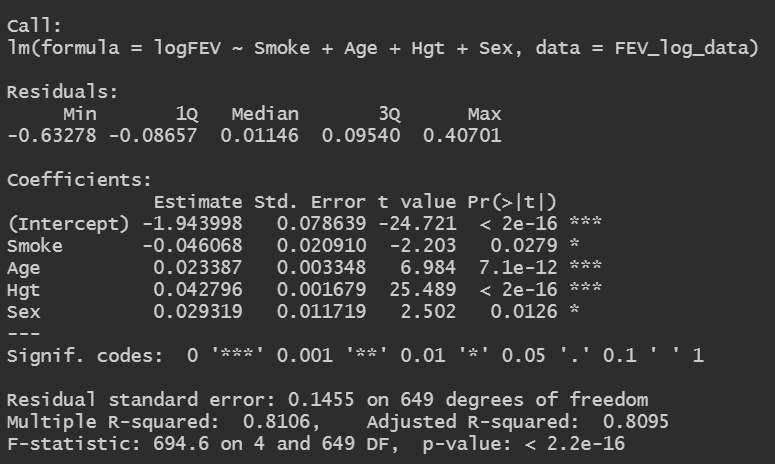
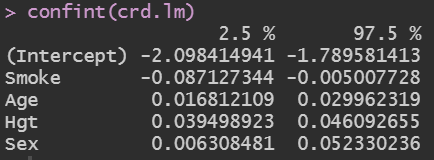
1. Temporarily hold out any outliers you identified in (2). Do any of the results in (3) – (6) change when holding out the outliers? That is, were the outliers influencing the conclusions? If so, discuss the differences.

**Ans**:

Remove outliers (row 2 and 652)

Original:

There are no significant changes after removing outliers. The R2, adjusted R2 and 95% CI do not change significantly.