Jessica Miron

BME 598: Applied Programming

In Class Activity 10.2

Executive Summary

Tager et al. collected data on children's pulmonary function in the presence and absence of smoking. Previous work was done with this dataset that showed smoking caused a significant difference in forced expiratory volume (FEV). The goal of the analysis is to create a model that can predict FEV based on other data such as age, height, sex, and smoking. This model will help to show how different variables impact FEV. The steps taken are shown below.

1. Load Python packages and dataset
2. Make a pairsplot to show which variables have relationships to each other
3. Create a model that predicts FEV from smoking and analyze the results
4. Create a model that predicts FEV from smoking and age and analyze the results
5. Create a model that predicts FEV from smoking, age, height, and height squared and analyze the results
6. Create a model that predicts FEV from smoking, age, and their interactions and analyze the results
7. Create a model that predicts FEV from all the previously stated variables and interactions and analyze the results
8. Create a model that predicts FEV from a new combination of variables and analyze the results
9. Complete a biological interpretation on the relationships between variables and which model performed the best.

Biological Interpretation

After completing the above steps, one pairsplot and six predictive models were created. The pairsplot shown in Figure 1 shows that there are five variables in the dataset: two categorical and three quantitative. Of the quantitative variables, "age" and "FEV" has a positive relationship, "age" and "ht" has a positive relationship, and "ht" and "FEV" has a positive quadratic relationship. These relationships helped to guide which variables and interactions were to be used in the models.

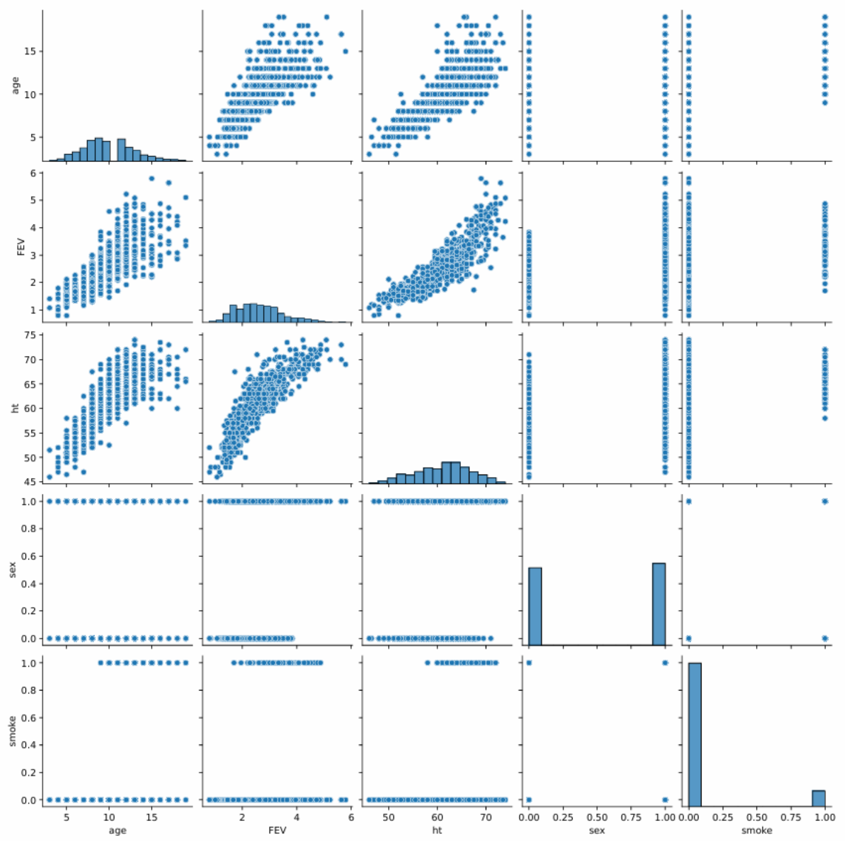
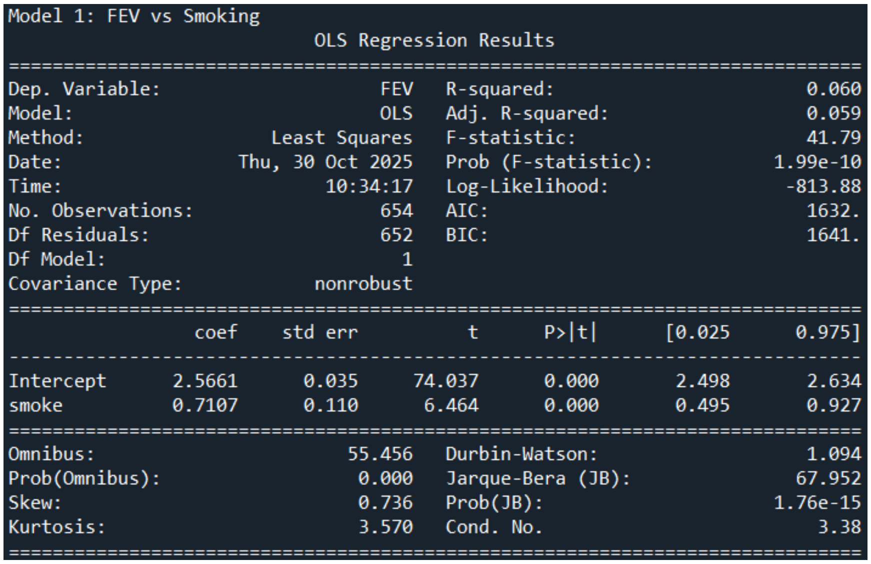


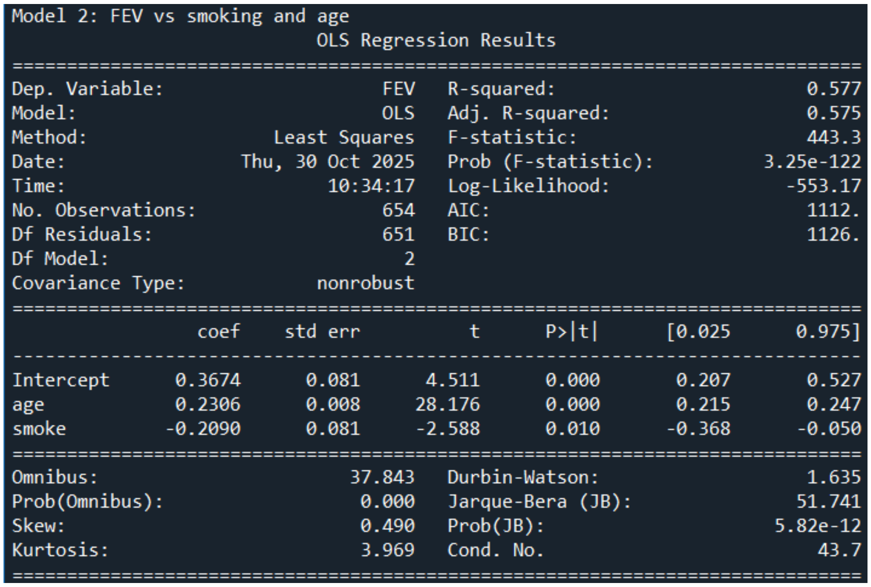
Figure 1: Pairsplot of the five variables in the fev dataset. The variables “sex” and “smoke” are categorical variables and “age”, “FEV”, and “ht” are quantitative variables.

Model 1 attempts to predict FEV from whether each participant smoked or not. Figure 2 shows the OLS regression results of the model. The only variable included in the model is “smoke” which shows a significant positive relationship with FEV with a coefficient of 0.7107 liters/s and p-value of 0.000. FEV should decrease with smoking, not increase like observed in the model [1]. It then makes sense that the adjusted R^2 value of the model is 0.059 and the model has an AIC of 1632 and BIC of 1641. This means that the model performed poorly and requires more information to accurately predict FEV. Model 2 attempts to predict FEV using whether the participant smoked and their age. Figure 3 shows the OLS regression results of the model. Two variables are included in the model “age” and “smoke”. Age has a significant positive relationship with FEV with a coefficient of 0.2306 liters/s/year and a p-value of 0.000. This makes sense as the participants are children and will get a larger lung volume as they grow. Smoke has a significant negative relationship with FEV with a coefficient of -0.2090 liters/s and a p-value of 0.010. This improves from Model 1 as the relationship is now negative. The model performs much better than Model 1 with an adjusted R^2 of 0.575, an AIC value of 1112, and a BIC value of 1126. Model 3 attempts to predict FEV from whether a participant smoked, their age, their height, and their height squared. Figure 4 shows the OLS regression results of the model. From the pairsplot, a positive quadratic relationship was observed between height and FEV leading to the squared term in the model. Four variables are in the model of which two have positive relationships and two have negative relationships. All four are significant. From the coefficients, height has the greatest effect on FEV with a coefficient of -0.3069 liters/s/in. The model is best performing so far with an adjusted R^2 of 0.790, an AIC of 653.6, and a BIC of 676.0.



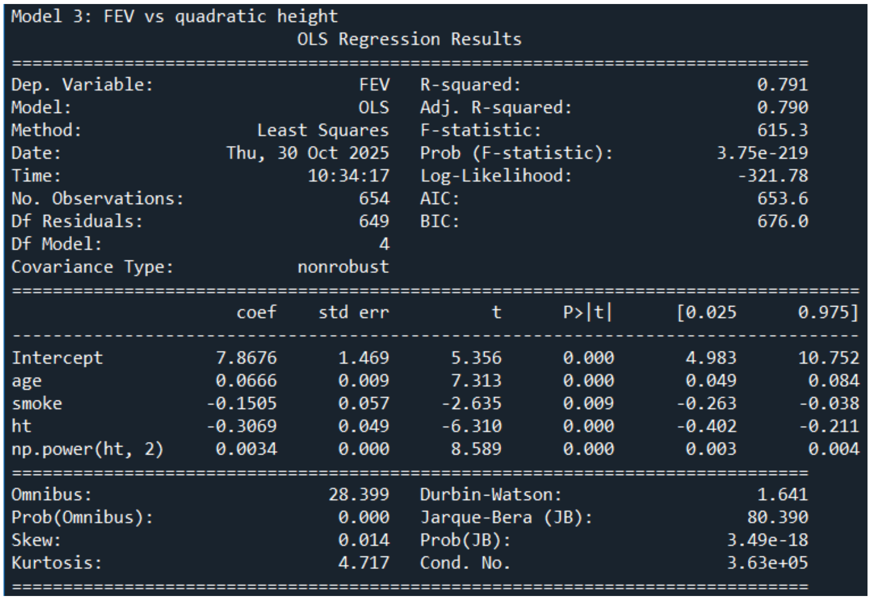
```
Model 1: FEV vs Smoking
                            OLS Regression Results
==============================================================================
Dep. Variable:                    FEV   R-squared:                       0.060
Model:                            OLS   Adj. R-squared:                  0.059
Method:                 Least Squares   F-statistic:                     41.79
Date:                Thu, 30 Oct 2025   Prob (F-statistic):           1.99e-10
Time:                        10:34:17   Log-Likelihood:                -813.88
No. Observations:                 654   AIC:                             1632.
Df Residuals:                     652   BIC:                             1641.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.5661      0.035     74.037      0.000       2.498       2.634
smoke          0.7107      0.110      6.464      0.000       0.495       0.927
==============================================================================
Omnibus:                       55.456   Durbin-Watson:                   1.094
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               67.952
Skew:                           0.736   Prob(JB):                     1.76e-15
Kurtosis:                       3.570   Cond. No.                         3.38
==============================================================================
```

Figure 2: OLS Regression Results for Model 1: FEV vs Smoking. Shows adjusted R^2 of 0.059, AIC of 1632, BIC of 1641, and the smoke variable with a coefficient of 0.7107 and a p-value of 0.000.



```
Model 2: FEV vs smoking and age
                            OLS Regression Results
==============================================================================
Dep. Variable:                    FEV   R-squared:                       0.577
Model:                            OLS   Adj. R-squared:                  0.575
Method:                 Least Squares   F-statistic:                     443.3
Date:                Thu, 30 Oct 2025   Prob (F-statistic):          3.25e-122
Time:                        10:34:17   Log-Likelihood:                -553.17
No. Observations:                 654   AIC:                             1112.
Df Residuals:                     651   BIC:                             1126.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.3674      0.081      4.511      0.000       0.207       0.527
age            0.2306      0.008     28.176      0.000       0.215       0.247
smoke         -0.2090      0.081     -2.588      0.010      -0.368      -0.050
==============================================================================
Omnibus:                       37.843   Durbin-Watson:                   1.635
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               51.741
Skew:                           0.490   Prob(JB):                     5.82e-12
Kurtosis:                       3.969   Cond. No.                         43.7
==============================================================================
```

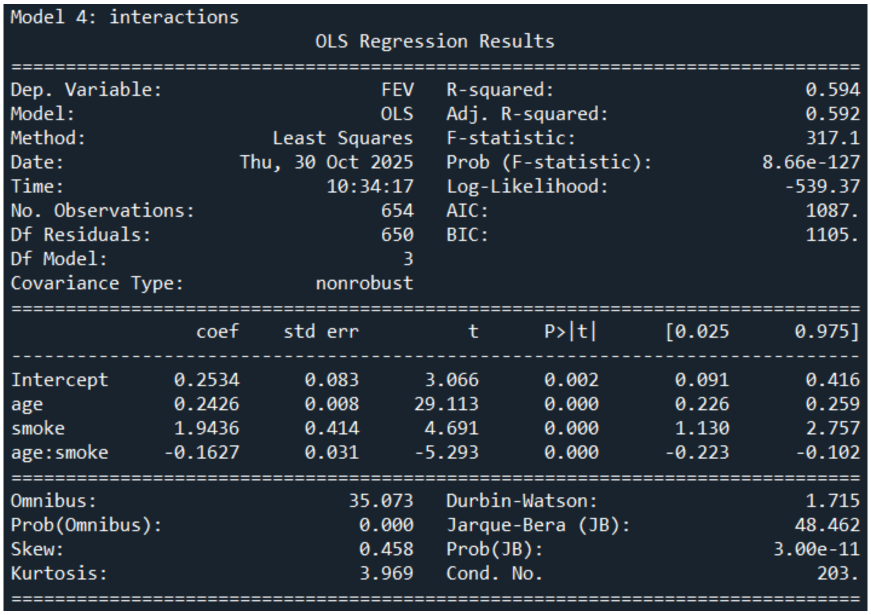
Figure 3: OLS Regression Results for Model 2: FEV vs smoking and age. Shows adjusted R^2 of 0.575, AIC of 1112, BIC of 1126. Age variable has a coefficient of 0.2306 liters/s/year and p-value of 0.000. Smoke variable has a coefficient of -0.2090 liters/s and p-value of 0.010.



```
Model 3: FEV vs quadratic height
                            OLS Regression Results
==============================================================================
Dep. Variable:                    FEV   R-squared:                       0.791
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     615.3
Date:                Thu, 30 Oct 2025   Prob (F-statistic):         3.75e-219
Time:                        10:34:17   Log-Likelihood:                -321.78
No. Observations:                 654   AIC:                             653.6
Df Residuals:                     649   BIC:                             676.0
Df Model:                           4
Covariance Type:            nonrobust
=====================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept           7.8676      1.469      5.356      0.000       4.983      10.752
age                 0.0666      0.009      7.313      0.000       0.049       0.084
smoke              -0.1505      0.057     -2.635      0.009      -0.263      -0.038
ht                 -0.3069      0.049     -6.310      0.000      -0.402      -0.211
np.power(ht, 2)     0.0034      0.000      8.589      0.000       0.003       0.004
==============================================================================
Omnibus:                       28.399   Durbin-Watson:                   1.641
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               80.390
Skew:                           0.014   Prob(JB):                     3.49e-18
Kurtosis:                       4.717   Cond. No.                     3.63e+05
==============================================================================
```

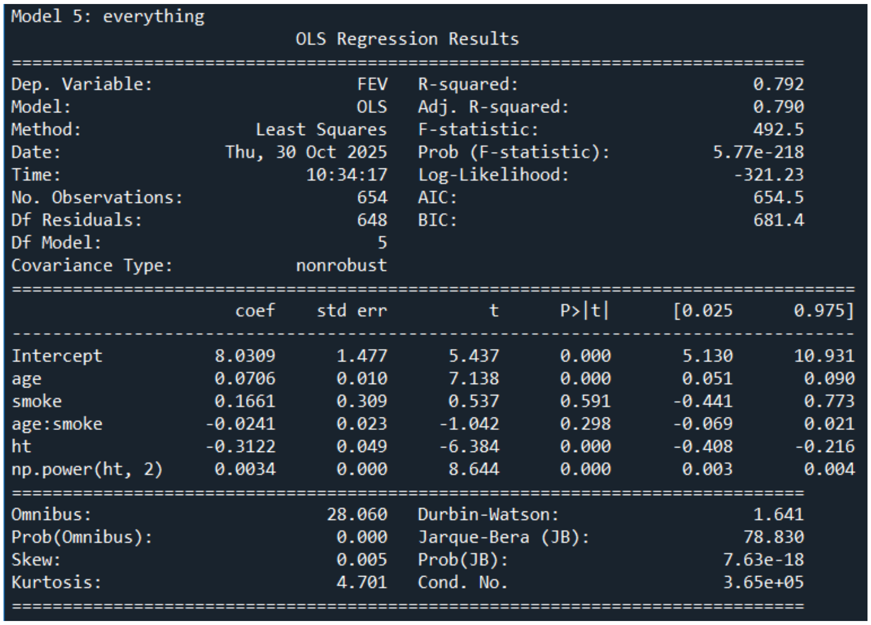
Figure 4: OLS Regression Results for Model 3: FEV vs quadratic height. Shows adjusted R^2 of 0.790, AIC of 653.6, and BIC of 676.0. Age has a coefficient 0.0666 liters/s/year and a p-value of 0.000. Smoke has a coefficient of -0.1505 liters/s and a p-value of 0.009. Height has a coefficient of -0.3069 liters/s/in and a p-value of 0.000. Height squared has a coefficient of 0.0034 liters/s/in^2 and a p-value of 0.000.

Model 4 attempts to predict FEV from whether a participant smokes, their age, and the interaction of smoking and age. Figure 5 shows the OLS regression results of the model. Of the three variables in the model, two have positive relationships and one has a negative relationship but all of them are significant. Smoking has a large positive coefficient of 1.9436 liters/s which, as in Model 1, shows that the model is most likely not performing as well. This then matches the adjusted R^2 of 0.592, an AIC of 1087, and a BIC of 1105. This means the model is better than Model 1, comparable to Model 2, and worse than Model 3. Model 5 attempts to predict FEV with all of the variables used in the other models. Figure 6 shows the OLS regression results. There are five variables in the model. Three have positive relationships and two have negative relationships. Two variables, smoke and the interaction of age and smoke, are not significant. This follows the pattern of as more variables are added the smoke variable becomes less significant in predicting the FEV. The adjusted R^2 value is 0.790, the AIC is 654.5, and the BIC is 681.4. This makes the model comparable to Model 3 despite the added variables. Model 6 attempts to predict FEV using age, height, sex and their interactions and smoke. Figure 7 shows the OLS regression results of the model. There are eight variables in the model. Five have positive relationships and three have negative relationships. Three variables (age, the interaction of age and height, and smoke) are not significant. Age and height are closely related which could lead to age and the interaction being not significant. Sex shows the greatest effect with a coefficient of 3.3760 liters/s. Other studies have shown that males have significantly higher FEV compared to females which matches with what the model describes [2]. This model has an adjusted R^2 of 0.801, an AIC of 620.8, and a BIC of 661.2. These values mean the model is the best performing with the highest R^2 and lowest AIC and BIC.



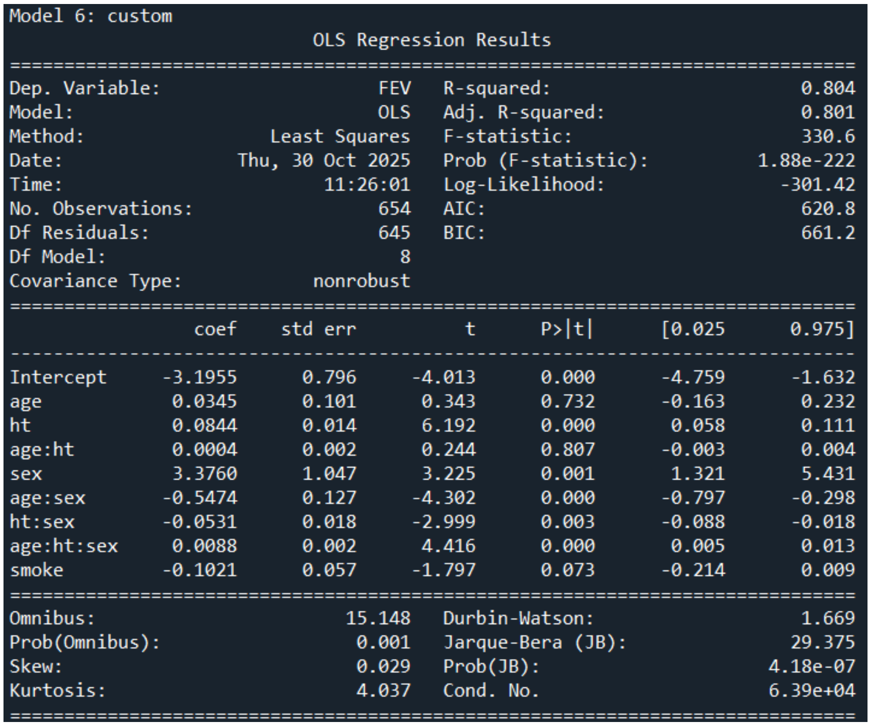
```
Model 4: interactions
                            OLS Regression Results
==============================================================================
Dep. Variable:                    FEV   R-squared:                       0.594
Model:                            OLS   Adj. R-squared:                  0.592
Method:                 Least Squares   F-statistic:                     317.1
Date:                Thu, 30 Oct 2025   Prob (F-statistic):          8.66e-127
Time:                        10:34:17   Log-Likelihood:                -539.37
No. Observations:                 654   AIC:                             1087.
Df Residuals:                     650   BIC:                             1105.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.2534      0.083      3.066      0.002       0.091       0.416
age            0.2426      0.008     29.113      0.000       0.226       0.259
smoke          1.9436      0.414      4.691      0.000       1.130       2.757
age:smoke     -0.1627      0.031     -5.293      0.000      -0.223      -0.102
==============================================================================
Omnibus:                       35.073   Durbin-Watson:                   1.715
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.462
Skew:                           0.458   Prob(JB):                     3.00e-11
Kurtosis:                       3.969   Cond. No.                         203.
==============================================================================
```

Figure 5: OLS Regression Results for Model 4: interactions. Shows adjusted R^2 of 0.592, AIC of 1087, and BIC of 1105. Age has a coefficient of 0.2426 liters/s/year and a p-value of 0.000. Smoke has a coefficient of 1.9436 liters/s and a p-value of 0.000. The interaction of age and smoke has a coefficient of -0.1627 liters/s and a p-value of 0.000.



```
Model 5: everything
                            OLS Regression Results
==============================================================================
Dep. Variable:                    FEV   R-squared:                       0.792
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     492.5
Date:                Thu, 30 Oct 2025   Prob (F-statistic):          5.77e-218
Time:                        10:34:17   Log-Likelihood:                -321.23
No. Observations:                 654   AIC:                             654.5
Df Residuals:                     648   BIC:                             681.4
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             8.0309      1.477      5.437      0.000       5.130      10.931
age                   0.0706      0.010      7.138      0.000       0.051       0.090
smoke                 0.1661      0.309      0.537      0.591      -0.441       0.773
age:smoke            -0.0241      0.023     -1.042      0.298      -0.069       0.021
ht                   -0.3122      0.049     -6.384      0.000      -0.408      -0.216
np.power(ht, 2)       0.0034      0.000      8.644      0.000       0.003       0.004
==============================================================================
Omnibus:                       28.060   Durbin-Watson:                   1.641
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               78.830
Skew:                           0.005   Prob(JB):                     7.63e-18
Kurtosis:                       4.701   Cond. No.                     3.65e+05
==============================================================================
```

Figure 6: OLS Regression Results for Model 5: everything. Shows adjusted R^2 of 0.790, AIC of 654.5, and BIC of 681.4. Age has a coefficient of 0.0706 liters/s/year and a p-value of 0.000. Smoke has a coefficient of 0.1661 liters/s and a p-value of 0.591. The interaction of age and smoke has a coefficient of -0.0241 liters/s and a p-value of 0.298. Height has a coefficient of -0.3122 liters/s/in and a p-value of 0.000. Height squared has a coefficient of 0.0034 liters/s/in^2 and a p-value of 0.000.



```
Model 6: custom
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    FEV   R-squared:                       0.804
Model:                            OLS   Adj. R-squared:                  0.801
Method:                 Least Squares   F-statistic:                     330.6
Date:                Thu, 30 Oct 2025   Prob (F-statistic):          1.88e-222
Time:                        11:26:01   Log-Likelihood:                -301.42
No. Observations:                 654   AIC:                             620.8
Df Residuals:                     645   BIC:                             661.2
Df Model:                           8                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.1955      0.796     -4.013      0.000      -4.759      -1.632
age            0.0345      0.101      0.343      0.732      -0.163       0.232
ht             0.0844      0.014      6.192      0.000       0.058       0.111
age:ht         0.0004      0.002      0.244      0.807      -0.003       0.004
sex            3.3760      1.047      3.225      0.001       1.321       5.431
age:sex       -0.5474      0.127     -4.302      0.000      -0.797      -0.298
ht:sex        -0.0531      0.018     -2.999      0.003      -0.088      -0.018
age:ht:sex     0.0088      0.002      4.416      0.000       0.005       0.013
smoke         -0.1021      0.057     -1.797      0.073      -0.214       0.009
==============================================================================
Omnibus:                       15.148   Durbin-Watson:                   1.669
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               29.375
Skew:                           0.029   Prob(JB):                     4.18e-07
Kurtosis:                       4.037   Cond. No.                     6.39e+04
==============================================================================
```

Figure 7: OLS Regression Results for Model 6: custom. Shows an adjusted R^2 of 0.801, AIC of 620.8, and BIC of 661.2. Age has a coefficient of 0.0345 liters/s/year and a p-value of 0.732. Height has a coefficient of 0.0844 liters/s/in and a p-value of 0.000. The interaction of age and height has a coefficient of 0.0004 liters/s and a p-value of 0.807. Sex has a coefficient of 3.3760 liters/s and a p-value of 0.001. The interaction of age and sex has a coefficient of -0.5474 liters/s and a p-value of 0.000. The interaction of height and sex has a coefficient of -0.0531 liters/s and a p-value of 0.003. The interaction of age, height, and sex has a coefficient of 0.0088 liters/s and a p-value of 0.000. Smoke has a coefficient of -0.1021 liter/s and a p-value of 0.073.

After analyzing all the models, they can be ranked from best to worst. Model 6, the custom model, performed the best most likely because of the addition of sex. Model 3 and Model 5 were the next best with very comparable performances though Model 3 had lower AIC and BIC making it slightly better despite the identical adjusted R^2 values. Model 4 was the next best performing model followed closely by Model 2 with adjusted R^2 values of 0.592 and 0.575 respectively. Model 1 was the worst performing model which makes sense as smoking is a categorical variable and doesn’t have the necessary information to encode for FEV. Since Model 6 is the best performing model, it is the model I would choose to use further. The model has the greatest adjusted R^2 and the lowest AIC and BIC of all the models and uses all the available data to encode for FEV. The coefficients and significance of variables also match with the interpretation of the variables which are expected.

References

[1] Trakas, Nikolaos et al. “Association between smoking cessation and alterations in forced expiratory volume in one second (FEV1). A Follow-Up Study from a Greek Tobacco Cessation Clinic.” *Addiction & health* vol. 14,2 (2022): 87-95. doi:10.22122/AHJ.2022.196722.1244

[2] Heraganahally SS, Howarth T, Sorger L, Ben Saad H (2022) Sex differences in pulmonary function parameters among Indigenous Australians with and without chronic airway disease. PLOS ONE 17(2): e0263744. https://doi.org/10.1371/journal.pone.0263744