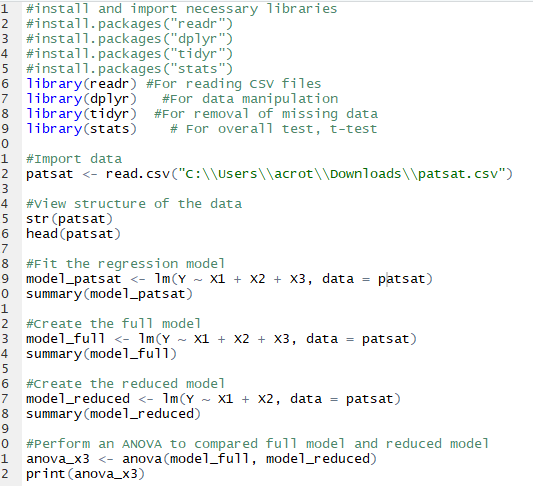
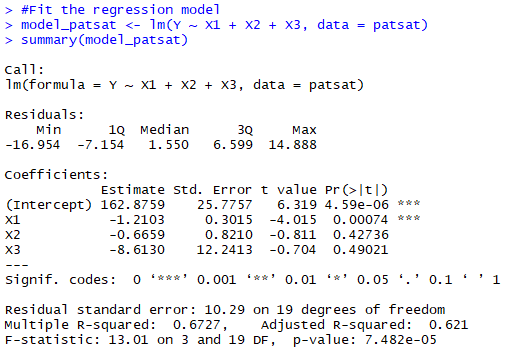
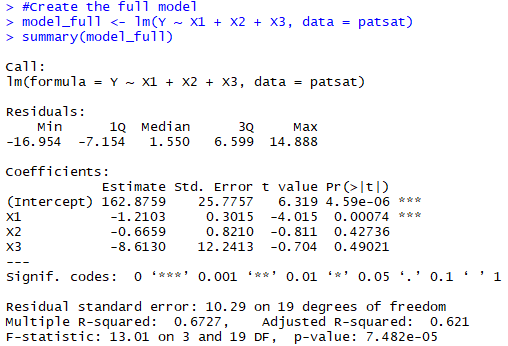
1. A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X1, in years), severity of illness (X2, an index) and anxiety level (X3, an index). The administrator randomly selected 23 patients and collected the data in patsat.sas7bdat, where larger values of Y, X2, and X3 are, respectively, associated with more satisfaction, increased severity of illness and more anxiety. Consider the multiple linear regression model Yi = β0 + β1X1i + β2X2i + β3X3i + εi, i = 1, 2, …, n.
   1. Test whether X3 can be dropped from the regression model given that X1 and X2 are retained. Use the F test statistic and level of significance 0.05. State the null and alternative hypotheses, degrees of freedom of test statistic, and conclusion. What is the p-value of the test?

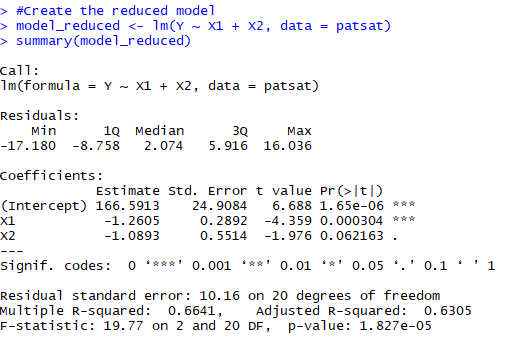
Null hypothesis: H0 : β3 = 0

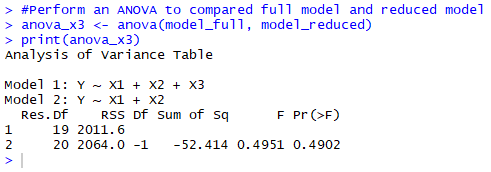
Alternate hypotheses: Ha : β3 ≠ 0









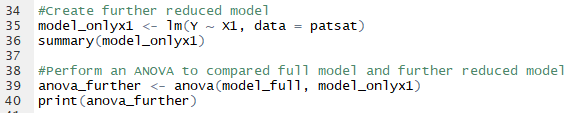


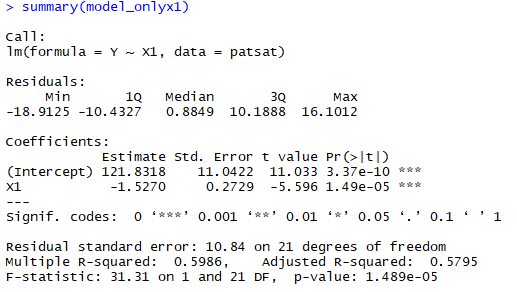
The F statistic is 0.4951 with a p-value of 0.4902. The test statistic degrees of freedom has a numerator of n-1 = 1, and denominator of 19. Given the p-value is above the specified level of significance 0.05, we fail to reject the null hypothesis and must conclude that β3 = 0. If β1 and β2 are still included in the model, then the model will still be valid without β3 present.

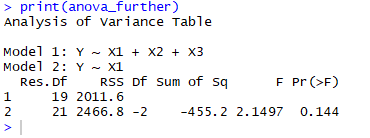
* 1. Test whether X2, X3 can be dropped from the regression model given that X1 is retained. Perform an appropriate test at level of significance 0.05. State the name of the test, the null and alternative hypotheses, degrees of freedom of test statistic, and conclusion. What is the p-value of the test?

Null hypothesis: H0 : β2 = β3 = 0

Alternate hypotheses: Ha : at least one of β2 or β3 ≠ 0





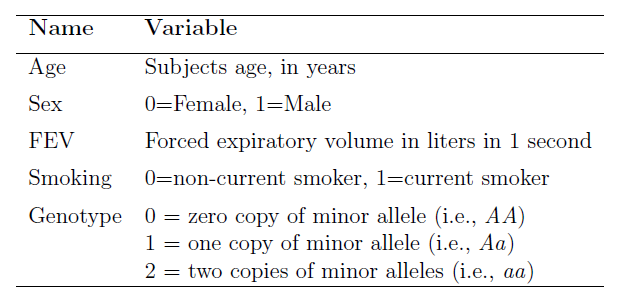


An ANOVA was used to determine the F statistic is 2.1497 with a p-value of 0.144. The test statistic degrees of freedom has a numerator of n-1 = 1, and denominator of 19. Given the p-value is above the specified level of significance 0.05, we fail to reject the null hypothesis and must conclude that β2 = β3 = 0. If β1 is included in the model, then the model will still be valid without β2 and β3 present.

1. In a study for 800 respiratory disease patients, the investigators were interested in assessing the impacts of environmental and genetic factors on a patient's lung function, which is measured by volume of air expelled in 1 second in liters (FEV). The potential predictors include age, sex, smoking status, and genotype of certain locus with two alleles (i.e., AA, Aa, or aa; where A is major allele and a is minor allele). The code sheet for these variables are as follows:

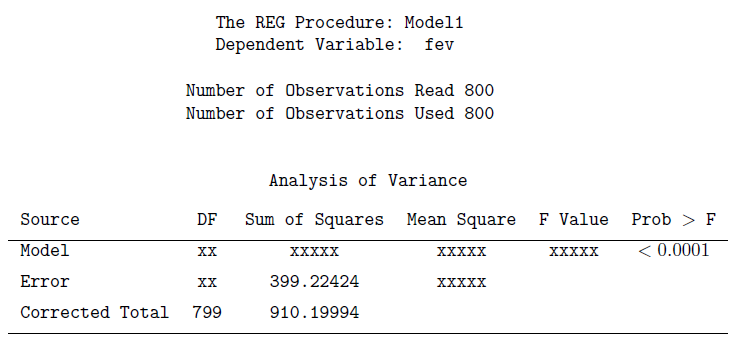
First, we fit a multiple linear regression model with all the n = 800 observations, using Age, Sex and Smoking status as predictors:

FEVi = β0 + β1Agei + β2Sexi + β3Smokingi + εi, i = 1, 2, …, n. (1)



and we obtain the following SAS output. A few items have been removed from the SAS

output and replaced by clusters of x's.



Use the above information answering parts (a) to (e).

* 1. What are the values of sum of squares of total (SST), sum of squares of regression (SSR), sum of squares of error (SSE) for regression model (1)?

SST = SSR + SSE = 510.9757 + 399.22424 = 910.19994. It is also the value listed under Sum of Squares for the Corrected Total source.

SSR = SST - SSE = 910.19994 - 399.22424 = 510.9757

SSE = (1 - R-squared) \* SST. The SSE is listed by the value listed under Sum of Squares for the Error source, or 399.22424.

* 1. What are values of mean squares of regression (MSR), mean squares of error (MSE) for regression model (1)?

MSR = SSR / DFmodel = 510.9757 / 3 = 170.32523

MSE = SSE / DFerror = SSE / (n - DFmodel - 1) = 399.22424 / (800 - 3 - 1) = 399.22424 / 796 = 0.50154

* 1. What is the value of the F statistic in this analysis? What are numerator and denominator degrees of freedom of the F statistic?

F = MSR / MSE = 170.32523 / 0.50154 = 339.60448

DFnumerator = number of independent variables = 3

DFdenominator = (n - DFmodel - 1) = (800 - 3 - 1) = 796

* 1. What is the null hypothesis and alternative hypothesis for this F test in the table? What can you conclude based on the test results?

H0: β1 = β2 = β3 = 0

Ha : at least one of β1, β2, β3 ≠ 0

Based on the test results, the p-value is <0.0001, which is below the standard level of significance of 0.05. This would give us evidence to reject the null hypothesis and conclude that at least one of β1, β2, or β3 has a statistically significant relationship with FEV.

* 1. Calculate the coefficient of determination R2 and adjusted coefficient of determination R2adj for the fitted regression model (1). Interpret R2.

R2 = (SSR / SST) = (510.9757 / 910.19994) = 0.56139

R2adj = 1 - ([1 - R2] \* [n - 1] / [n - DFmodel - 1]) = 1 - ([1 - 0.56139] \* [800 - 1] / [800 - 3 - 1]) = 1 - (0.43861 \* 799 / 796) = 1 - 0.44026 = 0.55974

The R2 value of 0.56139 means that the model can explain 56.139% of the variability in FEV due to age, sex, and smoking status. The R2 value can indicate the goodness of fit of the model with higher values. At 0.55974, the model has a moderate fit, meaning the models somewhat explain the dependent variable’s variability.

Next, we add the genotype variable into the regression model (1) and fit a regression model (2). We treat the genotype variable as a categorical variable, and use Genotype=0 as the reference level. Answer parts (f) to (g).

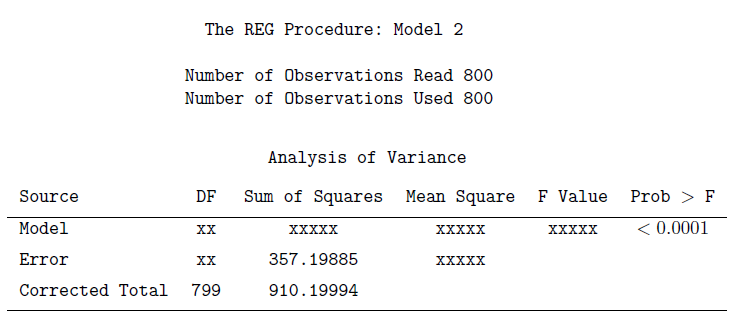
* 1. Create and define appropriate indicator variables to represent Genotype? How many indicator variables are needed?

Since genotype = 0 is the reference level, it is not included as an indicator variable, so we need two indicator variables to cover the other two levels. The indicator variables can be G1 for genotype = 1 and G2 for genotype = 2. Since the alleles can either be AA, Aa, or aa, but not more than one type at a time, they will be binary, taking on a value of 1 if present and 0 if not present.

* 1. Write out the expression for the regression model (2)? (Don't use matrix form.)

FEVi = β0 + β1Agei + β2Sexi + β3Smokingi + β4G1 + β5G2 + εi, i = 1, 2, …, n.

We obtain the following SAS output. A few items have been removed from the SAS output and replaced by clusters of x's.



Use the above output answering parts (h) to (k).

* 1. What is the value of the F statistic in this analysis? What are numerator and denominator degrees of freedom of the F statistic?

SSR = SST - SSE = 910.19994 - 357.19885 = 553.00109

MSR = SSR / DFmodel = 553.00109 / 5 = 110.60022

MSE = SSE / DFerror = SSE / (n - DFmodel - 1) = 357.19885 / (800 - 5 - 1) = 357.19885 / 794 = 0.44987

F = MSR / MSE = 110.60022 / 0.44987 = 245.84929

DFnumerator = number of independent variables = 5

DFdenominator = (n - DFmodel - 1) = (800 - 5 - 1) = 794

* 1. Calculate the coefficient of determination R2 and adjusted coefficient of determination R2adj for the fitted regression model (2).

R2 = (SSR / SST) = (553.00109 / 910.19994) = 0.60756

R2adj = 1 - ([1 - R2] \* [n - 1] / [n - DFmodel - 1]) = 1 - ([1 - 0.60756] \* [800 - 1] / [800 - 5 - 1]) = 1 - (0.39244 \* 799 / 794) = 1 - 0.39491 = 0.60509

* 1. Is there an improvement in R2 after adding Genotype into the model? Is there an improvement in R2adj after adding Genotype into the model? Which one is helpful for you to determine whether model (2) is an improvement of model (1)?

The R2 value is higher after adding Genotype in the model. This means that the model can explain more variation in FEV due to the included variables, so it is an improvement. The R2adj value is also higher, indicating the model overall has a better fit after adding genotype. Since R2adj calculates goodness of fit while R2 just checks for explanation of variability, R2adj is more helpful for determining whether model 2 is an improvement over model 1.

* 1. Using the information provided here for model (1) and model (2), conduct a hypothesis test to determine whether adding Genotype significantly accounts for variability in FEV above that explained by Age, Sex and Smoking status. What hypothesis test will you use? What's the value of the test statistic? What's its degrees of freedom?

H0 : β4 = β5 = 0

Ha : at least one of β4, β5 ≠ 0

A partial F-test is the appropriate test for performing a hypothesis test on binary indicator variables.

F = ([{SSR2 - SSR1} / 2] / MSE2) = ([{553.00109 - 510.9757} / 2] / 0.44987) = 46.70837

DFnumerator = number of independent variables = 2

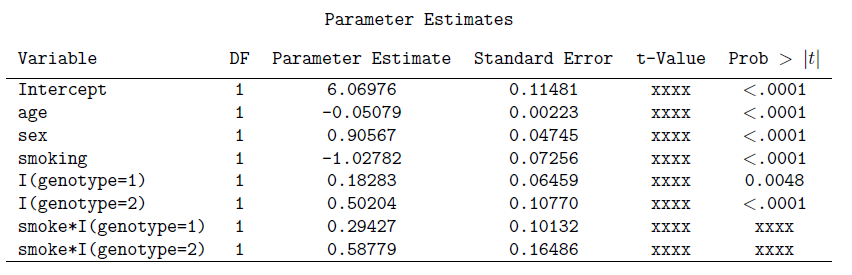
DFdenominator = (n - DFmodel - 1) = (800 - 2 - 1) = 797

Next, we fit a multiple linear regression model by adding interactions between smoking status and genotype into a regression model (2). We call this regression model with interactions “model (3)”.

* 1. Write out the expression for the regression model (3)? (Don't use matrix form.)

FEVi = β0 + β1Agei + β2Sexi + β3Smokingi + β4G1 + β5G2 + β6SmokingiG1 + β7SmokingiG2 + εi, i = 1, 2, …, n.

The table below summarizes the model (3) fitting results from SAS. A few items have been removed from the SAS output and replaced by clusters of x's.



* 1. Calculate the t-test statistic values for the two interaction terms smoke \* I (genotype = 1) and smoke \* I (genotype = 2). What are the degrees of freedom for the two t-test statistics, respectively? What's the p-values? Interpret the effects of smoke \* I (genotype = 1) and smoke \* I (genotype = 2), respectively.

tβ6 = PEβ6 / SEβ6 = 0.29427 / 0.10132 = 2.90436

tβ7 = PEβ7 / SEβ7 = 0.58779 / 0.16486 = 3.56539

DFβ6 = n - DFmodel - 1 = 800 - 7 - 1 = 792

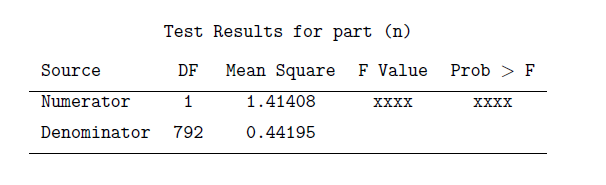
DFβ7 = n - DFmodel - 1 = 800 - 7 - 1 = 792

Using a t-table, the p-value for β6 is 1 - 0.998 = 0.002. The p-value for β7 is 1 - 1 = 0 (likely smaller than 0.001). Given an acceptance level of 0.05, the effect of smoking on G1 indicates that there will be an increase of FEV of 2.90436 when compared to the reference level G0. This relationship is statistically significant as the p-value is below 0.05. The effect of smoking on G2 indicates that there will be an increase of FEV of 3.56539 when compared to the reference level G0. This relationship is statistically significant as the p-value is below 0.05.

* 1. Suppose we are interested in assessing whether there is a difference in the effects of smoking on FEV for patients of genotype aa versus that for patients of genotype Aa under the interaction model (3). What is the null hypothesis corresponding to this test, using the notations you defined in part (l)?

H0 : β6 AND β7 = 0

* 1. Using the SAS output of “test results for part (n)” below, perform the null hypothesis in (n). What do you conclude?



F = MSN / MSD = 1.41408 / 0.44195 = 3.19964

p = 1 - 0.999 = 0.001

Based on the p-value, we can reject the null hypothesis that the values are both equal to 0. Therefore, smoking has a statistically significant relationship with at least one of β6 or β7.