Patrick Fitzgerald
ADS 534 Statistical Modeling
HW 4

[This is a follow-up of HW3, Problem 2. Use SAS software to do it.] A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index) and anxiety level (X_3, 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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i, \; i=1,2,\ldots,n.$$

(a) Test whether X_3 can be dropped from the regression model given that X_1 and X_2 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?

H_0: $\beta_3 = 0$ and H_a: $\beta_3 \neq 0$

The degree of freedom of the numerator is 1 and the degrees of freedom for the denominator are 19. The F-statistic and p-value are calculated:

Test test Results for Dependent Variable Y				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	52.41373	0.50	0.4902
Denominator	19	105.87285		

Therefore, we fail to reject the null that $\beta_3 = 0$ given that X1 and X2 are in our model we can drop X3.

(b) Test whether X_2, X_3 can be dropped from the regression model given that X_1 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?

H_0: $\beta_3 = 0$ and H_a: $\beta_3 \neq 0$

The degrees of freedom of the numerator are 2 and the degrees of freedom for the denominator are 19. The F-statistic and p-value are calculated:

Test test Results for Dependent Variable Y				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	227.59869	2.15	0.1440
Denominator	19	105.87285		

Therefore, we fail to reject the null that $\beta_2 = \beta_3 = 0$ given that X1 is in our model we can drop X2 and X3.

In a study for 800 respiratory disease patients, the investigators were interested in assessing the impacts of environmental and genetic factors on patients' 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:

$$FEV_i = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 Smoking_i + \varepsilon_i,\ i=1,2,...,n. \quad (1)\ 1$$

Name Variable

Age Subjects age, in years Sex 0=Female, 1=Male FEV Forced expiratory volume in liters in 1 second Smoking 0=non-current smoker, 1=current smoker

Genotype 0 = zero copy of minor allele (i.e., AA) 1 = one copy of minor allele (i.e., Aa)

2 = two copies of minor alleles (i.e., aa) 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 provided information answering parts (a) to (e).

(a) What are 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 = 910.19994, SSE = 399.22424,
SSR = SST – SSE = 910.19994 – 399.22424 = 510.97571

(b) What are values of mean squares of regression (MSR), mean squares of error (MSE) for regression model (1)?
MSR = SSR / p = 510.97571 / 3 = 170.3252
MSE = SSE / (n-p-1) = 399.22424 / (800-3-1) = 0.501538

(c) 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 = 170.3252 / 0.501538 = 339.6058
The numerator degrees of freedom for the F-statistic are 3 and the denominator 796.
p-value = $P(F_{3,\ 796} > 339.6058) = < 0.0001$

(d) What is the null hypothesis and alternative hypothesis for this F test in the table? What can you conclude based on the test results?

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_a: at least one $\beta_i \neq 0$ for I = 1, 2, 3

Since we calculated our p-value as (< 0.0001) we reject the null and say that at least one of Age, Sex, and Smoking has significant impact on FEV.

(e) Calculate the coefficient of determination R^2 and adjusted coefficient of determination R^2 for the fitted regression model (1). Interpret R^2. Adj

$R^2 = 1 - SSE/SST = 0.5613884$ which means that ~56% of the variation in FEV is explained by the linear regression model.

R^2 adjusted = $1 - ((n-1)/(n-p-1))*(SSE/SST)$

$= 1 - (799/796)*(399.2242/910.19994) = 0.5597354.$

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

(f) Create and define appropriate indicator variables to represent Genotype? How many indicator variables are needed?

Genotype has 3 levels -> 0, 1, and 2 so each must be accounted for. If we choose to make Genotype = 0 by our reference level then we need two indicator variables for Genotype = 1 and Genotype = 2.

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

$FEV = \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 Smoking + \beta_4(Genotype=1) + \beta_5(Genotype=2) + \varepsilon_i$

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).

(h) What is the value of the F statistic in this analysis? What are numerator and denominator degrees of freedom of the F statistic?

$F = MSR / MSE = 102.1951 / 0.5034354 = 245.5382$

The numerator degrees of freedom for the F-statistic are 5 and the denominator 794.

p-value = $P(F_{5, 794} > 245.5382) = < 0.0001$

(i) Calculate the coefficient of determination R^2 and adjusted coefficient of determination R^2 for the fitted regression model (2). adj

$R^2 = 0.60756$ and $R^2adj = 0.6050856$

(j) Is there an improvement in R^2 after adding Genotype into the model? Is there an improvement in R^2 after adding Genotype into the model? Which one is helpful for you to determine whether model (2) is an improvement of model (1)?

There is a noticeable improvement to the model with the addition of Genotype. Previously, we say ~56% explanation in the variation of FEV based on our linear regression model, now we see ~61% explanation in the variation of FEV based on our linear regression model with Genotype included.

The important thing to note is that R^2adj increased with the inclusion of Genotype as R^2 will increase with each added variable regardless of whether it actually impacts the accuracy of the model itself, however R^2adj will only increase if the model is actually improved by the addition of new variables.

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

We will utilize a partial F test to determine the impact of the a group of predictors:

$$H_0: \beta_4 = \beta_5 = 0$$

$$H_a: \beta_4 \neq 0 \text{ or } \beta_5 \neq 0$$

$F = (((SSR(Full) – SSR(Reduced))/r) / MSE(Full)$

$F = ((553.0011 – 510.9757)/2) / 0.501538$

$F = 46.64927$

The numerator degrees of freedom for the F-statistic are 2 and the denominator 794.

p-value = $P(F_{2, 794} > 46.64927) = < 0.0001$

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

(l) Write out the expression for regression model (3)? (Don't use matrix form.)
FEV = $\beta_0 + \beta_1$Age + β_2Sex + β_3Smoking + β_4(Genotype=1) + β_5(Genotype=2) + β_6Smoking*(Genotype=1) + β_7Smoking*(Genotype=2) + ε_i

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.

(m) 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 are the p-values? Interpret the effects of smoke*I(genotype=1) and smoke*I(genotype=2), respectively.
From the table we can see the Parameter Estimate for each interaction term and the Standard Error for each interaction term. To calculate the t-test statistic we simply must divide the parameter estimate by the standard error and thus we see the t-test statistic for Smoking*(Genotype=1) = 2.904362 and for Smoking*(Genotype=2) = 3.565389
The degrees of freedom for each interaction term can be calculated as (n-p-1) = 800-7-1 = 792.
The p-values based on these t-test statistics and dof are 0.0038 and 0.0004, respectively.
From the calculated results we can conclude that, when age and sex are constant, the effect of Smoking*(Genotype=1) means people who smoke and have genotype Aa are expected to have FEV levels 0.29427 units higher than those who smoke and have genotype AA.
Similarly, from the calculated results we can conclude that, when age and sex are constant, the effect of Smoking*(Genotype=2) means people who smoke and have genotype aa are expected to have FEV levels 0.58779 units higher than those who smoke and have genotype AA.

(n) 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)?
Using the notation from (l) we know that those with genotype aa fall under

$\beta_3 + \beta_7$ and those with genotype Aa fall under $\beta_6 + \beta_7$ so it follows that we are really only interested in β_6 and β_7. Specifically, we are interested in whether or not $\beta_6 = \beta_7$ which is algebraically equivalent to $\beta_6 - \beta_7 = 0$ (our null hypothesis). Thus, H_a: $\beta_6 - \beta_7 \neq 0$.

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

Using the partial F test we see that

F = MSR/MSE = 1.41408 / 0.44195 = 3.199638

From this we can calculate our p-value = 0.074 which given an α-level=0.05 we conclude that there is no statistically significant difference between the effects of smoking on FEV for patients of genotypes aa versus Aa given our interaction model.