Patrick Fitzgerald
ADS 534 Statistical Modeling
Lab # 3

1.1 Overall Test

We will be asking the question whether BMI or Age is useful in explaining the variability of LDL.

$$LDL_i = \beta_0 + \beta_1 BMI_i + \beta_2 Age_i + \varepsilon_i$$

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

- What is alternative hypothesis H_1?
- H_1: β_1 or $\beta_2 \neq 0$.
- Perform the test in SAS and interpret the results.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	22013	11007	7.74	0.0004
Error	2744	3903361	1422.50781		
Corrected Total	2746	3925375			

- Since our p-value (0.0004) < α (0.05) we reject the null and conclude that at least one of β_1 or $\beta_2 \neq 0$. We cannot conclude from this test which of the β-terms is non-zero merely that at least one is.

1.2 Tests of individual regression coefficients

Once we have determined that at least one of the regressors is important, we can then test individual coefficients.

Given the model

$$E(LDL_i) = \beta_0 + \beta_1 BMI_i + \beta_2 Age_i$$

We want to test H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$, given that Age_i is in the model.

1. t test

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	151.44273	8.77445	17.26	<.0001
BMI	BMI (kg/m^2)	1	0.36651	0.13196	2.78	0.0055
AGE	age in years	1	-0.25302	0.10970	-2.31	0.0212

Since our p-value (0.0055) < α (0.05) we reject the null and conclude that BMI is significantly associated with LDL.

2. Partial F test by comparing SSR from two models: linear regression models with and without BMI.

First the linear regression model with BMI:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	22013	11007	7.74	0.0004
Error	2744	3903361	1422.50781		
Corrected Total	2746	3925375			

- Now the linear regression model without BMI:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	11039	11039	7.74	0.0054
Error	2745	3914336	1425.98754		
Corrected Total	2746	3925375			

- F = (SSR(BMI, Age) – SSR(Age))/MSE = (22013 – 11039)/1422.50781
- F = 7.714545 => p-value ~= 0.0055 which is the same as seen in the t-test. This means that since our p-value (0.0055) < α (0.05) we reject the null and conclude that BMI is significantly associated with LDL.
- 3. Partial F test using the test statement in SAS.

Test test1 Results for Dependent Variable LDL				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	10974	7.71	0.0055
Denominator	2744	1422.50781		

- This confirms our previous results. Our F-statistic value (7.71) is the same as we calculated in (2) and our p-value (0.0055) is the same as we calculated in (1) and (2).

1.3 Tests for groups of predictors

Often, it is of interest to determine if, collectively, a group of predictors significantly contribute to the variability in Y given another group of predictors are in the model.

Given the model

$$E(LDL_i) = \beta_0 + \beta_1 statins_i + \beta_2 BMI_i + \beta_3 statins_i \times BMI_i + \beta_4 Age_i + \beta_5 Smoking_i + \beta_6 Drinkany_i + \beta_7 Nonwhite_i$$

Where there are two terms associated with BMI. We would like to know if BMI is significantly associated with LDL levels, given the model that this association differs by statin use?

We want to test H_0: $\beta_2=\beta_3=0$ vs H_1: at least one of β_2, $\beta_3 \neq 0$, given that other predictors are in the model. We will do it in different ways.

- 1. Partial F test by comparing SSR from two models: linear regression models with and without BMI.
- First we look at the full model including the interaction term (BMI*Statins):

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	216681	30954	22.85	<.0001
Error	2737	3707501	1354.58568		
Corrected Total	2744	3924182			

- Next we look at the reduced model (excluding BMI and BMI*Statins):

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	198227	39645	29.14	<.0001
Error	2739	3725955	1360.33418		
Corrected Total	2744	3924182			

- Now we will compare the F-statistics to determine the effect of BMI on LDL.
- F = ((SSR(Full Model) – SSR(Reduced Model))/r)/MSE

- F = ((216681 – 198227)/2)/1354.58568
- F = 6.811677 => p-value ~= 0.0011. This means that since our p-value (0.0011) < α (0.05) we reject the null and conclude that BMI is significantly associated with LDL.
- 2. Partial F test using the test statement in SAS.

Test test2 Results for Dependent Variable LDL				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	9227.15306	6.81	0.0011
Denominator	2737	1354.58568		

- We see the same p-value as calculated in (1) thus we have the same conclusion that BMI is significantly associated with LDL.

What if we want to test whether BMI is significantly associated with LDL levels for those people receiving statin? In other words, we want to test H_0: $\beta_2+\beta_3$ =0 vs H_1: $\beta_2+\beta_3 \neq 0$, given that other predictors are in the model. Notice the differences in H_0 from the above.

- 1. Partial F test using the test statement in SAS.

Test test3 Results for Dependent Variable LDL				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	399.33073	0.29	0.5872
Denominator	2737	1354.58568		

- This means that since our p-value (0.5872) > α (0.05) we fail to reject the null and conclude that BMI is not significantly associated with LDL for people receiving Statins.