Patrick Fitzgerald
ADS 534 Statistical Modeling
HW 3

(a) Consider the main effect model and the interaction model (which contains both main effects and interactions) that simultaneously assess two categorical factors (SO2 and CAPs). Write down these two regression models, making sure to explicitly define all independent variables. Write out the interpretation of each regression coefficient specified in each of these two models.

The data is already well suited to our needs in that the values of all the independent variables are either 0 or 1. CAP = 1 indicates concentrated air particles while CAP = 0 indicates filtered air. SO2 = 1 indicates an animal with bronchitis and SO2 = 0 indicates a healthy animal. It then follows that the product of 0*0, 0*1, 1*0, and 1*1 can only take on the values of 0 and 1 meaning that CAP*SO2 = 1 indicates an animal with bronchitis that has been exposed to the pollution and CAP*SO2 = 0 are all other possible combinations.

Main Effect Model: $Y = \beta_0 + \beta_1 CAPs + \beta_2 SO2$
Due to the nature of the independent variables CAP and SO2 (that they can take on values either 0 or 1) we expect animals with bronchitis or exposed to pollution to have higher Neutrophil Numerical Density (Nn) by a factor β_1 and/or β_2.

Interaction Model: $Y = \beta_0 + \beta_1 CAPs + \beta_2 SO2 + \beta_3 CAP*SO2$
Our interpretation of the first two independent variables from the previous problem remains true. Namely, that we expect animals with bronchitis or exposed to pollution to have higher Neutrophil Numerical Density (Nn) by a factor β_1 and/or β_2.

B_3 will explain the difference in Nn based on air pollution inhalation between animals with bronchitis and without bronchitis. Namely, that we would expect a difference of $\beta_1 + \beta_3$ in Nn level for an animal with bronchitis that was exposed to pollution over those with bronchitis that were not that were not.
Likewise, β_3 also explains the difference in Nn based on bronchitis between animals that were exposed to pollution and those that weren't. Namely, that we would expect a difference of $\beta_2 + \beta_3$ in Nn level for an animal exposed to pollution that did have bronchitis over those exposed to pollution that did not have bronchitis.

(b) Consider a test of whether there is a difference in the health effects of air pollution inhalation for healthy animals and that for chronic bronchitis animals (i.e. those who received SO2) under the interaction model. What is the null hypothesis corresponding to this test, in terms of the regression coefficients under the interaction model?

From our previously defined interaction model our null hypothesis would be
$H_0: \beta_3 = 0$.

(c) Perform an $\alpha = 0.05$ level test of the null hypothesis in (b). What do you conclude? Can you perform this test under the main effect model? If yes, carry out the test. If not, explain why.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.23815	0.00674	35.33	<.0001
caps	1	0.23770	0.00923	25.76	<.0001
so2	1	0.09774	0.00903	10.82	<.0001
interaction	1	0.12634	0.01286	9.83	<.0001

We can see from the table that for our β_3 term our p-value for the t-test is <0.0001 which means that we reject the null hypothesis (b) and conclude that there is a difference in the effect of CAP on Nn for animals with bronchitis and those without. This test cannot be carried out under the main effect model because there is no way to differentiate the effect of CAP on Nn for those animals with bronchitis and those without as there is no interaction term.

(d) Suppose the investigator neglected to mention that half of the animals received SO2, and you fit a regression model for Nn using CAPs exposure only. That is, you ignore whether the animal is chronic bronchitic or not. Write down the corresponding regression model. Fit both this model and the main effect model that includes CAPs and SO2, but not the interaction, to the data. Based on the results of these two models, do you think that SO2 confounds the association between CAPs and Nn? Explain why?

Regression for just CAP:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.29260	0.00853	34.29	<.0001
caps	1	0.28947	0.01218	23.77	<.0001

Regression for both CAP and SO2:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.20341	0.00667	30.48	<.0001
caps	1	0.30277	0.00747	40.54	<.0001
so2	1	0.16009	0.00747	21.43	<.0001

In order to determine whether or not SO2 confounds the association between CAP and Nn we perform the following basic arithmetic:
(CAP regression coefficient from main effect model – CAP regressions coefficient from CAP alone) / CAP regression coefficient from CAP alone
=> (0.30277 – 0.29847) / 0.28947 = 0.1485473 = ~1.5% < 10% therefore, we conclude that SO2 does not confound the association between CAP and Nn.

(e) Another way to justify whether SO2 confounds the association between CAPs and Nn is by looking at the pairwise association among these three variables directly. Choose appropriate measures and/or tests to investigate these pairwise associations. Based on the results you get, do you think that SO2 confounds the association between CAPs and Nn? Explain why?

In order to determine this we will utilize the two-sample t-test between CAP and Nn.

DF	t Value	Pr > |t|
134	54.63	<.0001

We see our p-value is <0.0001, which indicates that there is a significant association between CAP and Nn.
Next we will perform a two-sample t-test between SO2 and Nn.

DF	t Value	Pr > |t|
141	31.56	<.0001

We see our p-value is <0.0001, which indicates that there is a significant association between SO2 and Nn.
Finally, we will perform a chi-square test between CAP and SO2 since both are categorical variables.

Statistic	DF	Value	Prob
Chi-Square	1	1.8990	0.1682
Likelihood Ratio Chi-Square	1	1.9011	0.1680
Continuity Adj. Chi-Square	1	1.5810	0.2086
Mantel-Haenszel Chi-Square	1	1.8921	0.1690
Phi Coefficient		-0.0831	
Contingency Coefficient		0.0828	
Cramer's V		-0.0831	

We see our p-value is 0.1682, which indicates that there is not a significant association between CAP and SO2.

(f) Under each of the two models in (a) (the main effect model and the interaction model), test whether there is a health effect of air pollution inhalation for healthy animals, and also test whether there is a health effect of air pollution inhalation for chronic bronchitis animals (i.e. those who received SO2).

First, the main effect model:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.20341	0.00667	30.48	<.0001
caps	1	0.30277	0.00747	40.54	<.0001
so2	1	0.16009	0.00747	21.43	<.0001

We cannot differentiate between the effect of pollution on healthy animals vs those with bronchitis as that is not included in our main effect model. We see that the p-value for our t-test for $\beta_1 = 0$ is < 0.0001, which indicates that there is a significant effect on the health for both healthy animals and those with bronchitis.

Next, for the interaction model:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.23815	0.00674	35.33	<.0001
caps	1	0.23770	0.00923	25.76	<.0001
so2	1	0.09774	0.00903	10.82	<.0001
interaction	1	0.12634	0.01286	9.83	<.0001

Using our interaction model we can differentiate between healthy animals and those with bronchitis. We see again that our p-value for the t-test for $\beta_1 = 0$ is < 0.0001, which indicates that there is a significant effect on the health for healthy animals.

Finally, we use the **test** command in SAS to determine our p-value for the F-test for $\beta_1 + \beta_3 = 0$:

Test test1 Results for Dependent Variable nn				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	4.65884	1653.45	<.0001
Denominator	271	0.00282		

We see that the p-value for our F-test is < 0.0001 which indicates that there is a significant effect on health for those animals exposed to pollution with bronchitis.

(a) Fit the regression model Yi =β0 +β1X1i +β2X2i +β3X3i +εi, to the data and state the estimated regression function. What is the interpretation of β^2 here?

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	162.87590	25.77565	6.32	<.0001
X1	X1	1	-1.21032	0.30145	-4.01	0.0007
X2	X2	1	-0.66591	0.82100	-0.81	0.4274
X3	X3	1	-8.61303	12.24125	-0.70	0.4902

Holding the other beta terms (X1 and X3) constant we expect patient satisfaction to increase β_2 units for every 1 unit increase in severity of illness.

(b) Test whether there is a regression relationship here; that is, if the regression as a whole explains variability in the response. Using significance level α = 0.05, state your null and alternative hypotheses, and your conclusions. What does your test imply about β1, β2, and β3?

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_a: at least one $\beta_i \neq 0$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4133.63322	1377.87774	13.01	<.0001
Error	19	2011.58417	105.87285		
Corrected Total	22	6145.21739			

Since our p-value is < 0.0001 we reject the null hypothesis for our α = 0.05 level and conclude that at least on $\beta_i \neq 0$. We cannot determine which term is significant or how many are, however, we know that at least one must be.

(c) Test the null hypothesis that β1 is equal to 0 at the 0.05 level of significance. What do you conclude?

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	162.87590	25.77565	6.32	<.0001	108.92684	216.82496
X1	X1	1	-1.21032	0.30145	-4.01	0.0007	-1.84126	-0.57937
X2	X2	1	-0.66591	0.82100	-0.81	0.4274	-2.38427	1.05246
X3	X3	1	-8.61303	12.24125	-0.70	0.4902	-34.23426	17.00820

Rather than perform a simple linear regression test, I re-used the same code for our multiple linear regression model and we can interpret our p-value for β_1 as 0.0007 which is less than our $\alpha = 0.05$ level. Thus, we reject the null hypothesis that $\beta_1 = 0$ and conclude that X1 has a significant impact on Y.

(d) Obtain 95% confidence interval estimates of β1, β2, and β3. Interpret your results.

From the table provided in the previous problem, we can see that both our X2 and X3 terms include 0 within their 95% CIs thus we can conclude that they do not significantly impact Y (confirmed by p-values > α).
For X1 we can see that the 95% CI is (-1.841126, -0.57937) which means that as the patient's age increases one unit we are 95% confident that patient's satisfaction score decreases 0.57937 to 1.84126 units.

(e) Obtain a 95% confidence interval estimate of mean satisfaction when X1 = 35, X2 = 45 and X3 = 2.2. Interpret your confidence interval.

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	.	71.6003	4.4432	62.3006	80.9001	.

We can see that the 95% CI for our newly observed data is (62.3006, 80.9001) thus we are 95% confident that the mean patient satisfaction for our new data will fall within our confidence interval.