Patrick Fitzgerald
ADS 534 Statistical Modeling
Lab # 2

What is the multiple linear regression model in matrix form?

$$Y = \beta X + \varepsilon$$

What are each of the pieces of the model representing?

Y represents the nx1 column matrix of all possible outcomes from all the interactions between the nxj matrix X and the two nx1 matrices that represent the coefficient (or effect) that a given row vector, X_{ij}, has based on the acquired data. It is important to note that the first column of matrix X is comprised only of 1s to preserve the case where only β_0 is affecting the outcome.

What is the least squares estimate for $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ in matrix form?

Using matrix algebra it is possible to derive the following least squares estimate for β:

From the first problem we know $Y = \beta X + \varepsilon \Rightarrow \varepsilon = Y - X\beta$

then, $e'e = (Y - X\beta)'(Y - X\beta) = (Y' - \beta'X')(Y - X\beta)$

- ⇨ $Y'Y - 2\beta'X'Y + \beta'X'X\beta$ using partial derivatives with respect to β it can be shown that $X'X\beta = X'Y$
- ⇨ $B = (X'X)^{-1}X'Y$

multiple linear regression with categorical predictors

We will being by considering the impact of the new variable in the data set, FEV_2 on PEmax.

- Create binary indicator variables to represent FEV_2, using level 1 of FEV_2 as the reference level. How many binary indicator variables do you need?

  We require 2 indicator variables: using level 1 as the indicator variable we will need to account for the effect of the other 2 possible outcomes for FEV2 (2 and 3).

- Write the multiple linear regression model for prediction PEmax from FEV_2, using level 1 of FEV_2 as the reference level.

  - $PE_{max} = \beta_0 + \beta_1 I_{FEV2=2} + \beta_2 I_{FEV2=3}$

- Where $I_{FEV2=2}$ = { 1 if FEV2 = 2 and 0 otherwise

- And $I_{FEV2=3}$ = { 1 if FEV2 = 3 and 0 otherwise

- We will now fit this model in SAS.

The REG Procedure
Model: MODEL1
Dependent Variable: PEmax

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	22093	11046	51.27	<.0001
Error	22	4739.87013	215.44864		
Corrected Total	24	26833			

Root MSE	14.67817	R-Square	0.8234
Dependent Mean	109.12000	Adj R-Sq	0.8073
Coeff Var	13.45140		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	78.57143	5.54783	14.16	<.0001	67.06594	90.07692
I2	1	20.97403	7.09680	2.96	0.0073	6.25616	35.69190
I3	1	76.14286	7.84581	9.70	<.0001	59.87164	92.41407

- Interpret the regression coefficients in this model?
- From our parameter estimates we can see the change in PE_{max} associated with FEV2 changing from our reference level (1) to FEV2 = 2 is 20.97.
- We can also see that the difference between PE_{max} between someone with our reference level (1) and FEV2 = 3 is 76.14.

confounding

. Calculate Pearson correlation coefficient for continuous variables Age and PEmax. Is r significantly different from 0? Is there association between Age and PEmax?

Pearson Correlation Coefficients, N = 25 Prob > |r| under H0: Rho=0	
	Age
PEmax	0.61347 0.0011

.

We can see r = 0.613 which is significantly different from 0 meaning that there is an association between age and PE_{max}.

. Investigate the association between FEV_2 and PEmax. Notice that FEV_2 is categorical variable with 3 levels and PEmax is continuous. What test should we use?

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10.58744	10.58744	71.36	<.0001
Error	23	3.41256	0.14837		
Corrected Total	24	14.00000			

.

ANOVA is the appropriate test for a categorical variable and a continuous variable. We see a low p-value (<.0.0001) which suggests there is a significant difference between the means of the levels of FEV2 base don the PE_{max} of the subject.

. Investigate the association between FEV_2 and Age. What test should we use?

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.40219	4.40219	10.55	0.0035
Error	23	9.59781	0.41730		
Corrected Total	24	14.00000			

ANOVA is the appropriate test for a categorical variable and a continuous variable. We see a low p-value (0.0.0035) which suggests there is a significant difference between the means of the levels of FEV2 with respect to age of the subject.

. Assuming that there is no causal relationship between Age and FEV_2, do we think that FEV_2 is a confounder of the relationship between Age and PEmax? Why?

Yes, we conclude that FEV2 is a confounder of the relationship between age and PE_{max} because of the results shown in the previous two problems – namely, that FEV2 is associated with both age and PE_{max}.

We can also compare the unadjusted β for Age with the adjusted β for Age after controlling for FEV_2 to see if FEV_2 confounds the association between Age and PEmax. Usually, we conclude that FEV_2 is a confounder when we see a change in β of 10% or more.

- To begin, we fit simple linear regression model with Age alone.

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	50.40792	16.65711	3.03	0.0060
Age	1	4.05470	1.08835	3.73	0.0011

- We can see the β term associated with age is 4.05.
- Then we fit the multiple linear regression model with both Age and FEV_2 included.

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	69.38266	10.09836	6.87	<.0001
Age	1	0.79409	0.73049	1.09	0.2893
I2	1	19.47866	7.20030	2.71	0.0133
I3	1	70.24390	9.51313	7.38	<.0001

We can see that the β term associated with age when adjusting for FEV2 is 0.794.

. Assuming that there is no causal relationship between Age and FEV_2, do we think that FEV_2 is a confounder of the relationship between Age and PEmax, after looking at the output from the two above models? Why?

Yes, we do conclude that FEV2 is a confounder of the relationship between age and PEmax due to the significant difference between the simple linear regression β-term we found (4.05) and the adjusted β-term we found (0.794).

. What is the expected (or average) PEmax score from someone who is Age 16 and has FEV_2 score of 1? FEV_2 score of 2? FEV_2 score of 3?

. FEV2 = 1

. $PE_{max} = 69.38266 + 0.79409(16) = 82.0881$

. FEV2 = 2

. $PE_{max} = 69.38266 + 19.47866(1) + 0.79409(16) = 101.5668$

. FEV2 = 3

. $PE_{max} = 69.38266 + 70.24390(1) + 0.79409(16) = 152.332$

interactions

. Using PEmax as response variable, write out the full model for Age and each level of FEV_2, as well as interaction terms between Age and FEV_2.

. $PE_{max} = \beta_0 + \beta_1 I_{FEV2=2} + \beta_2 I_{FEV2=3} + \beta_3 Age_i + \beta_4 I_{FEV2=2} Age_i + \beta_5 I_{FEV2=3} Age_i$

. We will look at this relationship graphically. What do you notice from the plot?

.

. We will now fit the model with the interaction terms.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	78.49469	13.76348	5.70	<.0001	49.68739	107.30200
Age	1	0.00663	1.12647	0.01	0.9954	-2.35110	2.36436
I2	1	22.33166	17.47867	1.28	0.2168	-14.25163	58.91494
I3	1	-43.41004	33.50679	-1.30	0.2106	-113.54056	26.72048
age_fev2_2	1	-0.10183	1.35701	-0.08	0.9410	-2.94209	2.73842
age_fev2_3	1	6.28966	1.94938	3.23	0.0044	2.20957	10.36976

.

. What do you conclude from the model?

Two of the terms associated with age (β_3 and β_4 from our model) are effectively 0 (0.00663 and -0.10183) meaning that they have little effect. However, β_5 is significant and this will become obvious when we calculate PE_{max} for someone age 16 in the next problem.

. What is the expected PEmax score from someone who is Age 16 and has FEV_2 score of 1? FEV_2 score of 2? FEV_2 score of 3?

. FEV2 = 1

. $PE_{max} = 78.49469 + 0.00663(16) = 78.59508$

- FEV2 = 2

    - $PE_{max} = (78.49469 + 22.33166) + (0.00663 - 0.10183)(16) = 99.30315$

- FEV2 = 3

    - $PE_{max} = (78.49469 - 43.41004) + (0.00663 + 6.28966)(16) = 135.8253$

  - How does this compare to your previous estimate?

  While there is some change in all three values, the most notable change in PE_{max} comes from FEV2 = 3.