GROUP 3: ASHWINI SENTHILNATHAN, BRYCE RUTT, GETRUDE SHABIHA, MD. SHAMSUL KHAN

Part B. Regression Model Building

Helpful Instructions:

1. Create an introductory "scenario" of just two to three sentences that describes the data file for your project and why your group is building a regression model to predict Y based on the set of possible independent variables $X_1, X_{2,}..., X_k$

Salary is one of the most important determinants of performance in any work. Using a dataset, we can predict the salary of a player based on various factors. The dataset used for this analysis is the NBA dataset, which shows player number, name, playing position, points earned, bio data among other variables. This group aims at predicting the salary of players based on Rebound_off, Points, Blocks, Steals, Fouls and Turnovers.

2. As you learned in class, use the set of potentially meaningful numerical independent variables in your study to develop a "best" multiple regression model for predicting your numerical dependent variable Y.
    a. Fit a multiple regression model using the set of potentially meaningful numerical independent variables. Using the adjusted r^2 and standard error criteria, determine which numerical independent variables should be included in your regression model.

Multiple regression model was carried out to predict the salary of players based on eight independent variables namely; Rebound_off, Points, Blocks, Steals, assists, Rebound_def, Fouls and Turnovers. All these variables are numerical. The data was cleaned by pairwise deletion of missing values. The regression results are presented below;

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.568165245							
R Square	0.322811746							
Adjusted R Square	0.321513829							
Standard Error	4884584.331							
Observations	4183							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regression	8	4.74731E+16	5.93E+15	248.7152	0			
Residual	4174	9.95882E+16	2.39E+13					
Total	4182	1.47061E+17						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	749591.6282	199086.4338	3.765157	0.000169	359276.2	1139907	359276.2	1139907
Rebound_off	888569.9929	135020.5122	6.581	5.25E-11	623857.9	1153282	623857.9	1153282
Points	330989.3456	22448.12797	14.74463	5.27E-48	286979.1	374999.6	286979.1	374999.6
Assists	110371.2667	68048.29666	1.621955	0.104889	-23039.6	243782.2	-23039.6	243782.2
Blocks	921422.5505	203907.0549	4.518836	6.39E-06	521656.1	1321189	521656.1	1321189
Steals	833595.9787	224427.9773	3.714314	0.000206	393597.6	1273594	393597.6	1273594
Rebound_def	26035.29309	72881.57906	0.357227	0.72094	-116851	168922	-116851	168922
Fouls	-453317.2857	127579.9388	-3.5532	0.000385	-703442	-203193	-703442	-203193
Turnovers	401189.0688	179346.6549	2.236948	0.025342	49574.12	752804	49574.12	752804

The outcome of the regression analysis with the eight variables shows that two variables are not statistically significant since their p-values are greater than 0.05. These include Assists and Rebound_def which have p-values of 0.104889 and 0.72094 respectively. This model has a coefficient of determination (R-squared) of 0.3215. This means that the independent variables explain variation in the dependent variable up to 32.15%.

Dropping the insignificant variables, we plot another model and the results are presented below;

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.567773949							
R Square	0.322367258							
Adjusted R Square	0.321393647							
Standard Error	4885016.92							
Observations	4183							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regression	6	4.74077E+16	7.9E+15	331.105	0			
Residual	4176	9.96535E+16	2.39E+13					
Total	4182	1.47061E+17						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	770921.646	198519.6153	3.883353	0.000105	381717.5	1160126	381717.5	1160126
Rebound_off	867387.0725	117168.4961	7.402904	1.6E-13	637674.5	1097100	637674.5	1097100
Points	335866.4807	21121.37936	15.90173	2.52E-55	294457.3	377275.6	294457.3	377275.6
Blocks	897681.5154	191807.1257	4.680126	2.96E-06	521637.5	1273726	521637.5	1273726
Steals	942236.2903	214434.0067	4.394062	1.14E-05	521831.5	1362641	521831.5	1362641
Fouls	-466865.6815	126713.378	-3.68442	0.000232	-715291	-218440	-715291	-218440
Turnovers	567190.7246	148472.2201	3.820181	0.000135	276106.2	858275.3	276106.2	858275.3

The p-values for each of these variables is below 0.05, hence they are all significant in predicting the dependent variable. The adjusted R-squared for this model is 0.3214, which is less than that of the first model.

From this output we can obtain the general regression equation as;

$$\hat{Y} = 770921.646 + 867387.0725X_1 + 335866.4807X_2 + 897681.5154X_3 + 942236.2903X_4 - 466865.6815X_5 + 567190.7246X_6$$

Where $\hat{Y}$ is predicted salary while the X's represent the various independent variables in the model.

b. Fit a multiple regression model using the set of potentially meaningful numerical and categorical independent variables. Using the adjusted r^2 and standard error criteria, determine which numerical independent variable or variables should be included in your regression model.

GROUP 3: ASHWINI SENTHILNATHAN, BRYCE RUTT, GETRUDE SHABIHA, MD. SHAMSUL KHAN

Multiple regression model was carried out to predict the number of points based on four independent mixed variables which include West Conference, Minutes, Fouls and Height. The regression results are presented below;

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.889943193							
R Square	0.791998887							
Adjusted R Square	0.791804084							
Standard Error	3.041628265							
Observations	4276							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	150452.966	37613.24	4065.636	0			
Residual	4271	39513.16718	9.251503					
Total	4275	189966.1332						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-5.814105667	1.133954528	-5.12728	3.07E-07	-8.037245717	-3.59097	-8.03725	-3.59097
West Conference	0.701099616	0.093205349	7.522096	6.53E-14	0.518368706	0.883831	0.518369	0.883831
Minutes	0.609274919	0.006371246	95.62885	0	0.596783966	0.621766	0.596784	0.621766
Fouls	-0.914780262	0.07937355	-11.525	2.76E-30	-1.070393659	-0.75917	-1.07039	-0.75917
Height	0.038093384	0.014387245	2.647719	0.008133	0.009886908	0.0663	0.009887	0.0663

The model shows an adjusted R-squared of 0.7918 which means that the independent variables explain variation in the dependent variable up to 79.18%. This is a good model. The standard error of the model is equally small and shows that the model does not deviate far from the actual values.

From this output we can obtain the general regression equation as;

$$\hat{Y} = -5.8141 + 0.7011X_1 + 0.6093X_2 - 0.9148X_3 + 0.0381X_4$$

Where $\hat{Y}$ the number of points and the X's are the various independent variables in the model.

c. Select a value for your independent variables in its relevant range:

(1) Predict $\hat{Y}$.

The selected values for the independent variables are:

Rebound_off=2, Points=23, Blocks=2, Steals=3, Fouls=5 and Turnovers=3.5 Replacing these values in the regression equation,

$$\hat{Y} = 770921.646 + 867387.0725(2) + 335866.4807(23) + 897681.5154(3) + 942236.2903(3) - 466865.6815(5) + 567190.7246(3.5)$$

Salary= 14503536

1) Prediction using my values for the mixed model

The selected values for the independent variables are:

West Conference=Yes(1), Minutes=50, Fouls=5 and Height=85

values in the regression equation,

$$\hat{Y} = -5.8141 + 0.7011(1) + 0.6093(50) - 0.9148(5) + 0.0381(85)$$

Points=24.0165

(2) Determine the 95% confidence interval estimate of the average value of Y for all occasions when the independent variable has the particular value you selected.

Confidence interval for the numerical model

	x1	x2	x3	x4	x5	x6	y-hat*
	2	23	2	3	5	3.5	14503536
Alpha	0.05						
z	1.960						
se	4885016.92						
n	4183						
$Confidence\ Interval = \hat{y} \pm z * \frac{se}{\sqrt{n}}$							
Lower CI	**Upper CI**						
14355498.89	14651572.86						

Confidence Interval for the mixed model

GROUP 3: ASHWINI SENTHILNATHAN, BRYCE RUTT, GETRUDE SHABIHA, MD. SHAMSUL KHAN

	x1	x2	x3	x4	yhat*
	1	50	5	85	24.01477625
Alpha	0.05				
z	1.960				
se	3.041628265				
n	4276				
$Confidence\ Interval = \hat{y} \pm z * \frac{se}{\sqrt{n}}$					
Lower CI	**Upper CI**				
23.92360973	24.10594277				

(3) Determine the 95% prediction interval estimate of Y for an individual occasion when the independent variable has the particular value you selected

Prediction interval for the numerical model

	x1	x2	x3	x4	x5	x6	y-hat*				
	2	23	2	3	5	3.5	14503536				
Alpha	0.05							$Prediction\ Interval = \hat{y} \pm z * se$			
z	1.960							**Lower PI**	**Upper PI**		
se	4885016.92							4929079	24077993		
n	4183										

Prediction interval for the mixed model

	x1	x2	x3	x4	yhat*			
	1	50	5	85	24.01477625			
						$Prediction\ Interval = \hat{y} \pm z * se$		
Alpha	0.05					**Lower PI**	**Upper PI**	
z	1.960					18.0532944	29.9762581	
se	3.041628265							
n	4276							