DNSC 6279 – Quiz 5

Name:

NetId/Email:

A regression analysis was conducted by a junior data scientist to determine the relationship between the amounts a hospital charges for a medical service (AVE_ave_provider_charge), the amount a hospital is reimbursed by Medicare (AVE_ave_medicare_payment), and the number of services a hospital provides (AVE_num_service).

The model formula was specified as:

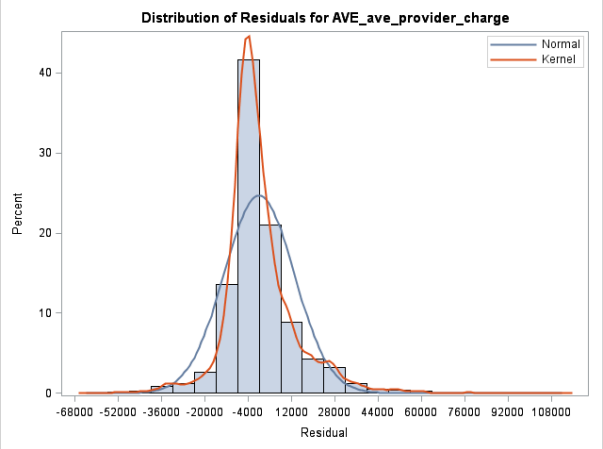
AVE_ave_provider_charge ~ AVE_ave_medicare_payment + AVE_num_service

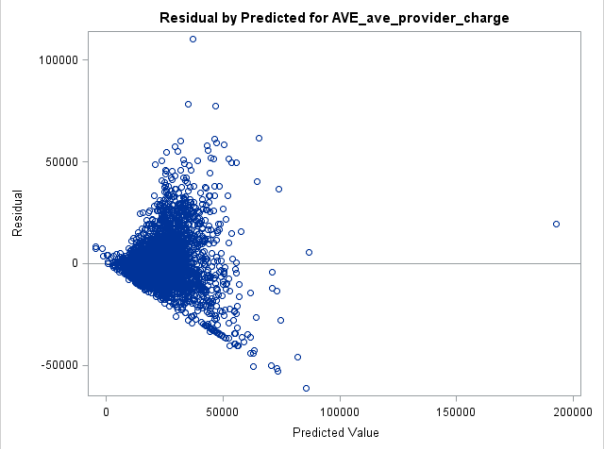
Among many other tables and plots, the following information was provided by the statistical software package after training the traditional regression model:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.85E+11	1.92E+11	1148.9	<.0001
Error	3334	5.58E+11	167376011		
Corrected Total	3336	9.43E+11			

Root MSE	**12937**	**R-Square**	**0.408**
Dependent Mean	24721	Adj R-Sq	0.4076
Coeff Var	52.33355		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1219.43	598.38	-2.04	0.0416	0
AVE_ave_medicare_payment	Average Medicare Payment	1	3.83	0.08	47.88	<.0001	1.02
AVE_num_service	Number of Services	1	-5.84	1.17	-4.96	<.0001	1.02





DNSC 6279 – Quiz 5

Name:

NetId/Email:

1.) **(2 pts.)** State the exact interpretation of the presented standard R-Square statistic.

The trained linear model explains 40.8% of the variance in the response variable, AVE_ave_provider_charge.

2.) **(2 pts.)** State the exact interpretation of the presented parameter estimate for AVE_ave_medicare_payment.

Holding the average number of procedures (AVE_num_service) constant, a 1-unit increase in the average medicare payments received by a hospital (AVE_ave_medicare_payment) will result in the average amount charged by a provider (AVE_ave_provider_charge) increasing by 3.83.

3.) **(3 pts.)** As you may have noticed, there is a serious problem with this regression analysis. Given the information provided what is the technical term that describes this problem?

Heteroscedasticity and/or nonnormally distributed residuals.

4.) **(3 pts.)** The presented output states that the parameter describing the linear relationship between the target variable and both input variables is statistically different from zero at the default alpha=0.05 level for the parameter ***t***-tests. Given the problem identified in 3 above, will the ***t***-tests remain unbiased?

No. The standard errors are biased when heteroscedasticity is present. This in turn leads to bias in test statistics and confidence intervals.

Nonnnormally distributed residuals can indicate that your model is highly biased or not specified correctly (e.g. trying to model nonlinear phenomena with a linear model), rendering the entire model questionable.