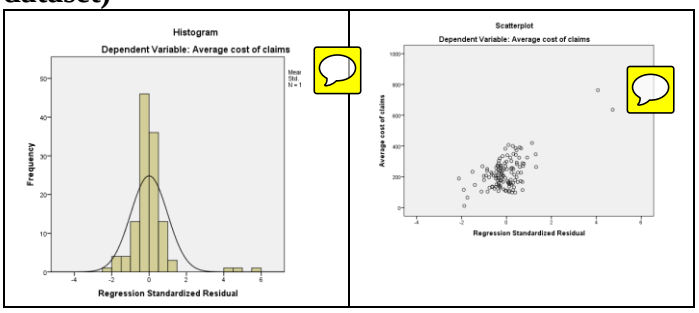
Bus Adm 216: Linear Regression Activity (Car insurance claims)

Dataset – variable description:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Policyholder age	128	1	8	4.50	2.300
Vehicle group	128	1	4	2.50	1.122
Vehicle age	128	1	4	2.50	1.122
Average cost of claims	123	11	850	231.14	117.048
Number of claims	128	0	434	69.86	91.852
Valid N (listwise)	123				

Model 1a - Average cost of Claims (using entire dataset)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.684[a]	.467	.454	86.484

a. Predictors: (Constant), Vehicle age, Policyholder age, Vehicle group

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	781375.850	3	260458.617	34.823	.000[b]
	Residual	890050.801	119	7479.418		
	Total	1671426.650	122			

a. Dependent Variable: Average cost of claims

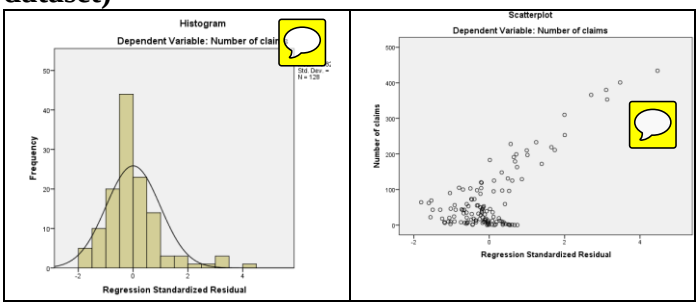
b. Predictors: (Constant), Vehicle age, Policyholder age, Vehicle group

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	321.335	29.822		10.775	.000
	Policyholder age	-11.773	3.462	-.228	-3.400	.001
	Vehicle group	38.662	7.074	.367	5.465	.000
	Vehicle age	-53.382	7.074	-.506	-7.546	.000

a. Dependent Variable: Average cost of claims

Model 1b - Number of Claims (using entire dataset)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.641[a]	.411	.397	71.340

a. Predictors: (Constant), Vehicle age, Vehicle group, Policyholder age

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	440379.262	3	146793.087	28.843	.000[b]
	Residual	631092.207	124	5089.453		
	Total	1071471.469	127			

a. Dependent Variable: Number of claims

b. Predictors: (Constant), Vehicle age, Vehicle group, Policyholder age

Coefficients[a]

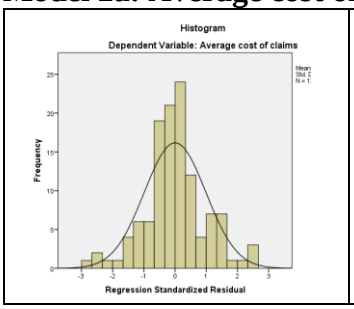
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	127.263	24.305		5.236	.000
	Policyholder age	14.289	2.752	.358	5.192	.000
	Vehicle group	-5.500	5.640	-.067	-.975	.331
	Vehicle age	-43.181	5.640	-.528	-7.656	.000

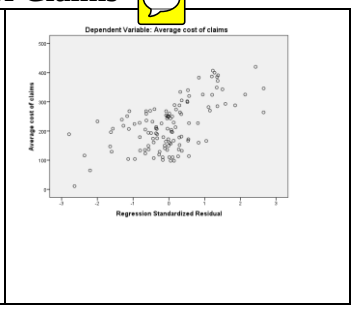
a. Dependent Variable: Number of claims

Gb

With three possibly outlier observations deleted (claim cost is now set to be less than 600).

Model 2a: Average cost of Claims





Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.808[a]	.653	.644	49.596

a. Predictors: (Constant), Vehicle age, Policyholder age, Vehicle group

b. Dependent Variable: Average cost of claims

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	536148.315	3	178716.105	72.656	.000[b]
	Residual	285333.010	116	2459.767		
	Total	821481.325	119			

a. Dependent Variable: Average cost of claims

b. Predictors: (Constant), Vehicle age, Policyholder age, Vehicle group

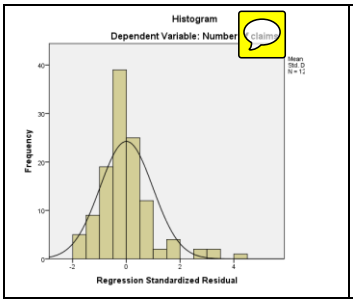
Coefficients[a]

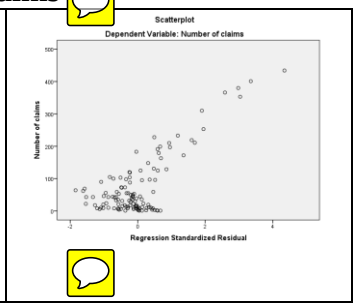
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	316.711	17.262		18.348	.000
	Policyholder age	-4.651	2.040	-.126	-2.279	.024
	Vehicle group	23.175	4.177	.306	5.548	.000
	Vehicle age	-54.126	4.119	-.721	-13.139	.000

a. Dependent Variable: Average cost of claims

With three possibly outlier observations deleted (claim cost is now set to be less than 600).

Model 2b: Number of Claims





Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.634[a]	.402	.387	72.885

a. Predictors: (Constant), Vehicle age, Policyholder age, Vehicle group

b. Dependent Variable: Number of claims

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	414484.784	3	138161.595	26.008	.000[b]
	Residual	616221.082	116	5312.251		
	Total	1030705.867	119			

a. Dependent Variable: Number of claims

b. Predictors: (Constant), Vehicle age, Policyholder age, Vehicle group

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	132.994	25.367		5.243	.000
	Policyholder age	14.658	2.998	.353	4.888	.000
	Vehicle group	-6.136	6.138	-.072	-1.000	.320
	Vehicle age	-45.826	6.054	-.545	-7.570	.000

a. Dependent Variable: Number of claims

Questions:

Comment on the regression assumptions in all the four models.

Write the regression equations from models 2a and 2b.

Comment whether deleting the three outlier observations helped predict average cost of claims better or not?

Interpret b1, b2, and b3 in model 2a.

Interpret coefficient of determination 2a and 2b.

Which one is better – 1a or 2a? Why?

Which one is better – 1b or 2b? Why?

Using 2a, compute estimated cost of claims if the age group is 6, vehicle group is 2 and the vehicle age group is 4.

Using 2b, compute estimated number of claims if the age group is 6, vehicle group is 2 and the vehicle age group is 4.

Interpret f test in 2a and 2b.

Find the p value of f test in 2a. Use F table to find the p values here.

Find the p value of all the t values in 2a. Use T table to find the p values here.

What are the sample sizes in 1a and 2a?

Why do you think vehicle group is significantly related to cost of claims but is not related to the number of claims?