NHH, Bergen

Operating Expenses - Solution

Note: Software suitable for flexible statistical analysis may not be the best for presentation purposes. However, output which is unsuitable for presentation as is, may be edited to make it readable without any accompanying text. We have done some slight modifications in our output, but further improvements could be made before presentation.

(1)
After recoding of Driving Length (1= ≤ 20 000, 2= >20 000) and Age (1 = ≤2, 2 = >2), it may be of interest to see how the cars are distributed in the different groups. We get.

Tabulated statistics: Count Driving Group; Age Group; Car Type

Car Type = 1
Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	8	10	18
2	6	1	7
All	14	11	25

Cell Contents: Count

Car Type = 2
Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	35	28	63
2	20	17	37
All	55	45	100

Cell Contents: Count

Car Type = 3
Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	67	50	117
2	53	30	83
All	120	80	200

Cell Contents: Count

We see that there are 25 sedans, 100 station wagons and 200 vans in the sample. If the dividing lines between the groups lead to a very skew distribution, they may be modified. Here we are comfortable with the ones chosen. We also see that the driving length patterns do not vary much with age, except for a possible interaction effect of more frequent use of newer vans.

(2)
We want to compute the mean and standard deviation of the costs for each age group and each group of driving length. This can be done by making separate tables for each of the two category variable as follows:

Tabulated statistics O-Cost

Rows: Age Group

	O-Cost Mean	O-Cost StDev
1	34253	17166
2	39938	18538
All	36632	17946

Rows: Driving Group

	O-Cost Mean	O-Cost StDev
1	27500	12612
2	50870	15624
All	36632	17946

Tabulated statistics R-Cost

Rows: Age Group

	R-Cost Mean	R-Cost StDev
1	7895	6786
2	11357	7537
All	9344	7302

Rows: Driving Group

	R-Cost Mean	R-Cost StDev
1	7470	6934
2	12264	6914
All	9344	7302

We see that both costs tend on average to increase from the low age group to the high, and from the low driving length group to the high. However, these one-dimensional tables may hide information about combined effects age and driving length. The standard deviation for each variable does not seem to deviate much between groups, except that the operating costs naturally vary less in the group of low diving lengths.

In order to see if there may be hidden combined effects, we tabulate the average and standard deviation of each cost type in a 2 by 2 layout for each combined category of Age and Driving Group.

Tabulated statistics: O-Cost vs. Driving Group; Age Group

Rows: Driving Group
Columns: Age Group

	1	2	All
1	24439	31326	27500
	11129	*13355*	*12612*
2	47919	55726	50870
	14566	*16235*	*15624*
All	34253	39938	36632
	17166	*18538*	*17946*

Cell Contents:
O-Cost : Mean
O-Cost : *Standard deviation*

Tabulated statistics: R-Cost vs. Driving Group; Age Group

Rows: Driving Group
Columns: Age Group

	1	2	All
1	5764	9602	7470
	6509	*6894*	*6934*
2	10862	14572	12264
	6038	*7674*	*6914*
All	7895	11357	9344
	6786	*7537*	*7302*

Cell Contents:
R-Cost : Mean
R-Cost : *Standard deviation*

We see clearly that the combination high age and high driving length gives higher costs of both types. It is of interest to investigate whether the effects are just adding, or goes beyond that (i.e. a so called interaction effect).

So far the three car categories are lumped together. There may be differences in costs between the car categories. Some software provides the opportunity to compute descriptive statistics for multi-way category data in a compact manner. Here we present a table with mean and standard deviation in two three-way layouts, one for each cost type. In fact we could have combined the counts above and other statistics, for instance the median, in the same layout as well.

Tabulated statistics: O-Cost for Driving Group; Age Group; Car Type

Car Type = 1

Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	24800	19708	21971
	10151	*13068*	*11816*
2	39577	47770	40747
	8231	*	*8127*
All	31133	22259	27228
	11795	*15010*	*13764*

Cell Contents:
O-Cost: Mean
O-Cost: *Standard deviation*

Car Type = 2

Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	23914	30996	27061
	8165	*8806*	*9105*
2	36402	44169	39970
	7642	*10827*	*9916*
All	28455	35972	31838
	9964	*11486*	*11266*

Cell Contents:
O-Cost: Mean
O-Cost: *Standard deviation*

Car Type = 3

Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	24670	33835	28587
	12612	*14441*	*14118*
2	53209	62541	56582
	14241	*15297*	*15222*
All	37275	44600	40205
	19478	*20269*	*20072*

Cell Contents:
O-Cost: Mean
O-Cost: *Standard deviation*

We see that the operating costs for Car type=1 come out favourable compared to the other car types in the low driving length group, and that the operating costs of Car type=3 come out unfavourable to the other car types in the high driving length group, and particularly so if the car also is in the high age group. Here we clearly see non-additive (interaction) effects. We also see that the standard deviations become inflated.

Tabulated statistics: R-Cost for Driving Group; Age Group; Car Type

Car Type = 1

Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	9202	7339	8167
	9298	*6358*	*7610*
2	13034	18438	13806
	5123	*	*5103*
All	10844	8348	9746
	7779	6898	7363

Cell Contents:
R-Cost: Mean
R-Cost: *Standard deviation*

Car Type = 2

Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	5730	9444	7381
	5177	*8065*	*6818*
2	8743	11613	10062
	7039	*7350*	*7230*
All	6826	10263	8373
	6037	*7790*	*7057*

Cell Contents:
R-Cost: Mean
R-Cost: *Standard deviation*

Car Type = 3

Rows: Driving Length Group
Columns: Age Group

	1	2	All
1	5372	10144	7411
	6738	*6305*	*6946*
2	11416	16120	13116
	5609	*7584*	*6743*
All	8041	12385	9779
	6929	*7365*	*7403*

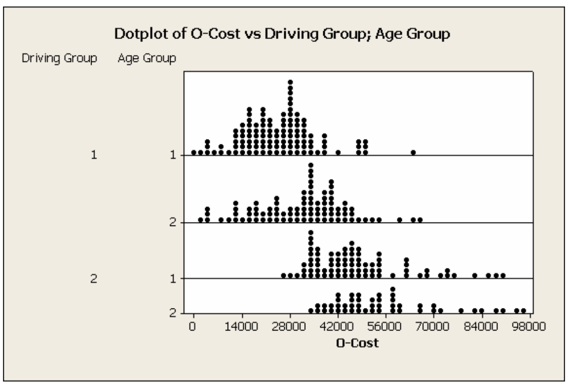
Cell Contents:
R-Cost: Mean
R-Cost: *Standard deviation*

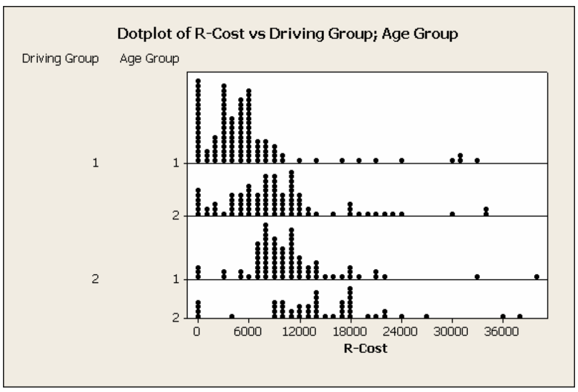
We see that the repair and maintenance tend to increase with age and driving length, but does not seem to vary much with car type. However, the combinations high age and high driving length come out unfavourably for car type 1 and 3 compared to car type 2. Note, however, that there is only one car of type 1 in this group. Standard deviations are very similar throughout.

Note: We could alternatively display the result of both cost factors within the same table. However, this may not be the best way to present the results.

(3)

We may illustrate the data in dotplots for grouped data as follows:

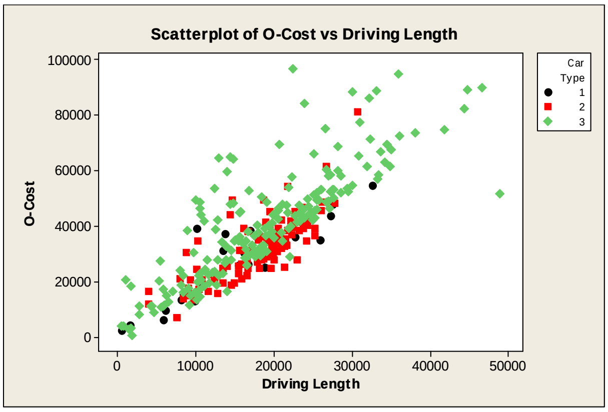


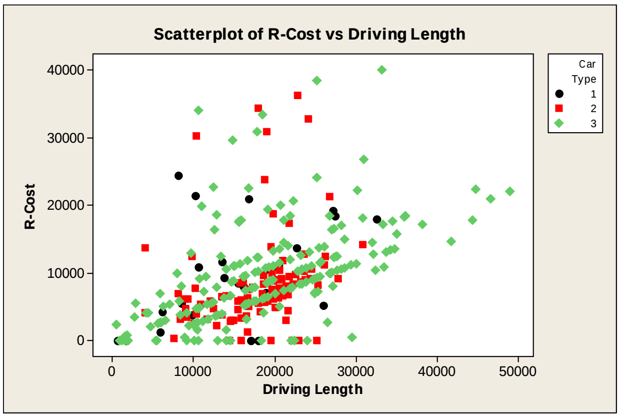


We see the main features commented upon above, but also that repair and maintenance costs have not occurred at all for some cars.

(4)

Here follows scatterplots for the two cost types versus Driving length. The three car types sedan (1), station wagon (2) and pick-up van (3) are marked with different symbols (and color)





For operating costs we see a clear linear tendency, but there is a clear lower limit to the downside cost for a given driving length, due to the fuel cost. However, note one strange outlier on the right side of the plot. The upside costs are more varying, with one outlier for a middle driving length at the top of the plot. For repair and maintenance costs there are also a linear tendency, except for some cars without costs and some with very high costs, probably due to special circumstances.

(5)

The correlations asked for follows

Correlations: O-Cost; Driving Length; Age; R-Cost

	O-Cost	Driving Length	Age
Driving Length	0.824		
Age	0.180	-0.108	
R-Cost	0.526	0.464	0.323

We see that the correlation between the two cost types is moderately high, just above 0.5. For O-Cost, the correlation with Driving Length is high, and with Age low. For R-Cost the correlations with Driving Length and Age are both moderate. The correlation between Driving Length and Age is negative, but small. If we look at the correlations for sedans only (see below) we see that this correlation is more negative. This means that (at least the sedans) are likely to be used less the older they are. This may possibly affect some analyses, where older cars may come out with favourably low costs, unless we take their driving length into account as well. We may see this in the two-way tabulation above and in the correlations below

Correlations: O-Cost; Driving Length; Age; R-Cost for Sedan

	O-Cost	Driving Length	Age
Driving Length	0.889		
Age	-0.007	-0.307	
R-Cost	0.515	0.426	0.192

(6)

We want to explain the O-Cost and R-Cost by Driving Length, Age and Car Type by linear regression. We have exposed the danger of having explaining the costs with one variable at a time, and go for a multiple regression. Car Type is categorical, and can be represented by three indicators. Taking sedan as base category, the other two is specified in the regression. Here is the output:

Regression Analysis: O-Cost versus Driving Length; Age; ...

```
The regression equation is
O-Cost = - 7793 + 1.79 Driving Length + 2685 Age
                + 1118 Car type 2 + 8157 Car type 3


Predictor           Coef  SE Coef      T      P
Constant           -7793     2007  -3.88  0.000
Driving Length  1.78744  0,05501  32.49  0.000
Age               2685.0    247,9  10.83  0.000
Car type 2          1118     1848   0.60  0.546
Car type 3          8157     1759   4.64  0.000


S = 8241.02   R-Sq = 79.2%   R-Sq(adj) = 78.9%
```

We see that we have explained 79.2% of the variation in O-Cost by the specified variables. Both Driving Length and Age have positive regression coefficients and are clearly statistical significant. The coefficients for Car type 2 and 3 are positive as well, but only type 3 is significant. This says that pick-up vans definitely has higher expected O-costs than sedans, but not necessarily so for station wagons. The regression coefficient of Driving Length represents the expected additional cost per increase by one unit Driving Length, regardless of Age and Car type, and the regression coefficient of Age represents the expected additional cost per increase by one year, regardless of Driving Length and Car type. The regression coefficients for Car type represents the additional expected cost compared to the base category (sedan).

Regression Analysis: R-Cost versus Driving Length; Age; ...

```
The regression equation is
R-Cost = - 1118 + 0,439 Driving Length + 1479 Age
         - 1993 Car type 2 - 852 Car type 3


Predictor           Coef  SE Coef      T      P
Constant           -1118     1427  -0.78  0.434
Driving Length  0.43906  0.03912  11.22  0.000
Age               1478.9    176.3   8.39  0.000
Car type 2         -1993     1315  -1.52  0.130
Car type 3          -852     1251  -0.68  0.496


S = 5861.47   R-Sq = 36.4%   R-Sq(adj) = 35.6%
```

We see that we have explained 36.4% of the variation in R-Cost by the specified variables. Both Driving Length and Age have positive regression coefficients and are clearly statistical significant. The coefficients for Car type 2 and 3 are negative, but none of them is significant. Nevertheless, this may be an interesting observation which may be given an explanation. We may now simplify the model by removing the insignificant Car type variables, thus giving a prediction formula with just two predictor variables. However, in practice this will not matter much, and we may just as well leave it as it is.

For both regression analyses it may be useful to perform an analysis of the residuals. This may tell whether the standard assumptions for inference in regression are fulfilled and whether the regression model may be improved. We have already seen from our plots that we are not likely to have strict linearity, homoscedasticity and normality. In the given context we are not that worried, since our purpose is not to do exact statistical inferences. However, revealed model inadequacies may sometimes lead to better understanding and models. A residual analysis here hardly reveals anything new, which cannot be inferred from the scatterplots above. It would clearly be of interest to be able to explain the many outlying R-Costs. Most likely, the R-Cost are mainly of two kinds: Regularly scheduled services with occasional minor repairs and accidental major repairs, the latter occurring more or less at random not depending on driving length and age or anything else observable. It may not be feasible to bring the explanatory power for R-Cost up to the level to that of O-Cost.