Car Insurance - Solution

Here the variables Opinion and Driving Length are ordered categorical variables, while Company, Ownership, Usage are binary variables. Owner Age, Car Age and Long Distance Trips are scaled measurement variables.

It may be of interest to start tabulating Opinion versus Company (CI or not)

Tabulated statistics: Company; Opinion

```
Rows: Company   Columns: Opinion

        1   2   3   4  5  All

1      15  16  22   4  3   60
2       9  24  14  11  2   60
All    24  40  36  15  5  120

Cell Contents:      Count

Pearson Chi-Square = 8.344; DF = 4; P-Value = 0.080
```

We see that the non-CI customers (2) in the sample have a slightly more favourable attitude than the CI-customers (1). However, a chi-square test of independence between Opinion and Company shows that cannot be rejected at the 5% significance level (P=0.08>0.05). Since we are mainly interested in the target customers we may recode Opinion accordingly and tabulate the recoded binary variable.

```
MTB > Code (1:3) 0 (4:5) 1 'Opinion' 'Target'
```

Tabulated statistics: Company; Target Group

```
Rows: Company   Columns: Target Group

           0      1     All

1         53      7      60
       88.33  11.67  100.00

2         47     13      60
       78.33  21.67  100.00

All      100     20     120
       83.33  16.67  100.00

Cell Contents:      Count
                    % of Row

Pearson Chi-Square = 2.160; DF = 1; P-Value = 0.142
```

Estimates of proportions of target customers among the two groups and estimate and test for their differences follow:

Test and CI for One Proportion (CI customers)

```
Sample  X   N  Sample p         95% CI
1       7  60  0.116667  (0.048215; 0.225716)
```

Test and CI for One Proportion (non-CI customers)

```
Sample   X   N  Sample p         95% CI
1       13  60  0.216667  (0.120716; 0.341956)
```

Test and CI for Two Proportions

```
Sample   X   N  Sample p
1       13  60  0.216667
2        7  60  0.116667

Difference = p (1) - p (2)
Estimate for difference:  0.1
95% CI for difference:  (-0.0321530; 0.232153)
Test for difference = 0 (vs not = 0):  Z = 1.47  P-Value = 0.142

Fisher's exact test: P-Value = 0.220
```

We see that the proportions are higher for non-CI customers, but the difference is not statistically significant

We may cross-tabulate the binary variable Target Group (recoded Opinion) versus other binary variables, recoded if necessary as

```
MTB > Code (18:39) 1 (40:99) 2  'Owner Age' ‘Group Owner Age’
MTB > Code (0:3) 1 (4:99) 2  'Car Age' ‘Group Car Age’
MTB > Code (1:2) 1 (3) 2  'Driving Length' ‘Group Driving Length’
MTB > Code (0:4) 1 (5:99) 2  'Long distance trips' ‘Group Long Distance’
```

We report here for short the tendency, chi-square statistic and P-value for each of 6 cross-tabulations:

Target Group vs Variable	Chi-square	P-value	Tendency
Ownership	5.627	0.018	Positive
Usage	2.820	0.093	Weakly pos.
Group Owner Age	0.672	0.412	None
Group Car Age	0.812	0.368	None
Group Driving Length	18.034	0.000	Strong positive
Group Long Distance	1.528	0.216	None

This means that customers with driving length more than 30000 km are definitely more likely to favour the supplement to the insurance than those with shorter driving length. Moreover Ownership shows that firms (code 2) are more likely to favour it than private owners (code 1). There is also a slight tendency for respondents for cars with shared usage to favour the supplement.

These findings may help to narrow the target group to car owners (mostly firms) with longer driving lengths. However, the question will be if such groups are sufficiently large to justify the efforts to promote the scheme by targeting them. We have seen that own customers may be less in favour of such a scheme than outside customers. The population of outside customers is typically larger than the population of own customers, but their favour may not necessarily come to our advantage. It is hard to imagine that this offer alone may make sufficiently many to shift insurance company.

We have here jumped together the two groups: own customer or not. This is questionable since the 60+60=120 respondents do not represent a sample from a well-defined population. It would be more reasonable to perform the analysis separately on the two groups as follows:

Tabulated statistics: Group Driving Length; Target Group; Company

Results for Company = 1

```
Rows: Group Driving Length   Columns: Target Group

           0      1     All

1         47      1      48
       97.92   2.08  100.00

2          6      6      12
       50.00  50.00  100.00

All       53      7      60
       88.33  11.67  100.00

Cell Contents:      Count
                    % of Row

Pearson Chi-Square = 21.388; DF = 1; P-Value = 0.000
```

Results for Company = 2

```
Rows: Group Driving Length   Columns: Target Group

           0      1     All

1         34      6      40
       85.00  15.00  100.00

2         13      7      20
       65.00  35.00  100.00

All       47     13      60
       78.33  21.67  100.00

Cell Contents:      Count
                    % of Row

Pearson Chi-Square = 3.142; DF = 1; P-Value = 0.076
```

We see that as many as 50% of the respondents from own company with long driving length fall in the target group, compared with 35% of the outside respondents with long driving length. This may look promising, but the number of respondents behind these frequencies are small, and results are unreliable (may be judged by computing confidence intervals).

There may be covariation between the various explanatory variables that may be disclosed by computing appropriate correlation measures, or by more detailed analysis. This may lead to some modifications of our statements. Instead we may want to consider the explanatory variables jointly in a regression context.
Here is a regression analysis, where Opinion is explained by the original explanatory variables. We see that Driving Length is the only variable that comes out statistically significant at 5% level, while the number of Long Distance Trips may also have some influence, but is likely to be positively correlated with Driving Length. The signs of the non-significant regression coefficients may be interpreted and kept in mind.

Regression Analysis: Opinion versus Company; Ownership; ...

```
The regression equation is
Opinion = 1.59 - 0.023 Company + 0.237 Ownership - 0.011 Usage
          - 0.0069 Owner Age - 0.0151 Car Age + 0.391 Driving Length
          + 0.0481 Long Distance trips


Predictor                Coef  SE Coef      T      P    VIF
Constant               1.5876   0.7890   2.01  0.047
Company               -0.0227   0.1978  -0.11  0.909  1.245
Ownership              0.2369   0.2081   1.14  0.257  1.279
Usage                 -0.0112   0.1984  -0.06  0.955  1.184
Owner Age            -0.00692  0.01009  -0.69  0.494  1.330
Car Age              -0.01513  0.05164  -0.29  0.770  1.162
Driving Length         0.3909   0.1318   2.97  0.004  1.553
Long Distance trips   0.04812  0.02695   1.79  0.077  1.400


S = 0.971181   R-Sq = 23.4%   R-Sq(adj) = 18.6%
```

This analysis may be criticized for violating a number of basic assumptions for regression analysis: (1) The left hand side variable is not scaled measurement variable, just an ordinal variable, (2) the right hand side involves ordered categorical variables and (3) the measurement variables like Owner Age may not affect the response variable linearly throughout its range and (4) technical assumptions like constant variance, normality etc. Nevertheless such an analysis is often done tentatively in practice to gain some preliminary insight.

In light of the objective to focus on the target group of customers, we may instead choose the binary Target Group as left side variable. This leads to categorical (typically logistic) regression and overcomes (1) as well. To overcome (2) and (3) we may replace the questionable variables by their corresponding recoded group variables. The result follows, where we see that Driving length is still the only statistically significant variable.

Binary Logistic Regression: Target Group versus Company; Ownership; ...

```
Link Function: Logit

Response Information

Variable      Value  Count
Target Group  1         20  (Event)
              0        100
              Total    120

Logistic Regression Table
                                                         Odds     95% CI
Predictor                 Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant              -5.02829   2.31466  -2.17  0.030
Company               0.614743  0.591254   1.04  0.298   1.85   0.58   5.89
Ownership             0.783024  0.593110   1.32  0.187   2.19   0.68   7.00
Usage                 0.378205  0.578113   0.65  0.513   1.46   0.47   4.53
Group Owner Age      -0.378234  0.606495  -0.62  0.533   0.69   0.21   2.25
Group Car Age        -0.529119  0.601737  -0.88  0.379   0.59   0.18   1.92
Group Driving Length   1.66923  0.606008   2.75  0.006   5.31   1.62  17.41
Group Long distance  -0.175724  0.606318  -0.29  0.772   0.84   0.26   2.75
```

Note that the signs of the non-significant regression coefficients do not always conform across analysis, and should be interpreted with care. Further analyses leaving out non-significant variables may add to the understanding. Moreover, one should expect that Group Long Distance should pick up some explanatory power when leaving out the significant Group Driving Length. This is not the case, and leaves us with some confusion about what is going on. It seems that more data is needed to say anything about this.