**Report of Findings**

The goal of the survey was to develop an equation that could provide a reasonable estimation of Carbon monoxide content (y) in cigarettes from values of three other variables: tar (X1), nicotine (X2) and mass (X3). The data used to construct the model was taken from 25 popular brands of cigarettes. Values for the four variables (tar, nicotine, mass and CO- content) were collected for each brand. CO- content for all 25 brands was then plotted against each individual X variable to get an idea of their relationship. The plots for tar and nicotine showed positive, linear trends, while the plot of mass against CO- showed no discernible pattern. A linear regression analysis, which is a statistical technique for fitting a model to data, was then performed on each predictor variable (X) with CO- content. The resulting models confirmed the relationships between the three predictor variables with y that were proposed when examining the scatterplots – positive linear relationships between tar and nicotine with y, and no useful relationship between mass and y. The models for the first two are: CO- content = 2.7433 + 0.8010(tar), and CO- content = 1.6647 + 12.3954(nicotine).

The criteria used in quantifying the utility of the models was investigated next. There are several things statisticians look for when assessing how well a regression model fits the collected data. The first is how well each predictor variable contributes to the model. The first two independent variables contributed significant information to their respective models; mass on the other hand proved insignificant. Another important metric is how well the overall model fits the data. Because these models consist of only one predictor variable, the significance of each model as a whole is equal to that of the predictor variable. A third measure of model utility is how much variation that is inherent in the data set our model accounts for. No model will ever be perfect; often when one obtains the same value of an independent variable, the corresponding dependent variable will assume a slightly different value. This could be due to random chance or experimental error. The metric associated with this called a correlation coefficient. The correlation coefficients for X1 and X2 were 0.9167 and 0.8573, respectively. This means, for example, the linear model we fit using X1 and y accounts for 91.67% of the variation in our sample data. A perfect relationship [theoretically] would be 1.0. Unsurprisingly, X3 had a low coefficient (0.2153).

Typically, when multiple independent variables (X’s) are known to be related to a dependent variable (y), one would like to construct a model that incorporates *all* of the independent variables. Because mass proved insufficient in predicting values of y, tar and nicotine were used in the full model. A regression analysis was again performed, but this time with both independent variables. The resulting model was: CO- content = 3.0896 +.9625(tar) – 2.6463(nicotine). An immediate problem arose when first examining this model: nicotine is now negatively related with CO- content. From the previous regression analysis of nicotine with CO- content, the relationship was positive. Furthermore, previous studies have also confirmed that, as nicotine concentrations increase in cigarettes, so too does CO- content. This, along with other statistical measures, suggested that the two independent variables tar and nicotine are correlated themselves. Because of this, it is not possible to have one model with both variables. Rather, we will have to settle with two separate models relating each to CO- content.

As previously mentioned, a statistical model is not a perfect relationship between two variables. This is because random chance and experimental limitations will always be present in a data set as variations in our y values. The goal of a regression analysis is to choose the best model; one that minimizes the error of prediction of y. The precise definition of an error, or residual, is the difference between the predicted values of y at any given point, with what is actually observed at that point. It is imperative then that the correct model is fit to a given set of data. Because of the random error, it can sometimes be difficult to see if the relationship between two variables is indeed linear, and not slightly curved. This issue and related nuances can be detected by plotting residuals against the x and y values of the data. When a mispecified model is fit to a data set, plotting the residuals against the x values can reveal the true nature of the relationship between x and y. Because our residual plot showed rather chaotic behavior when plotted against each independent variable, it is safe to assume the model has been correctly specified.

Another assumption of regression analysis is that the residuals are normally distributed. This assumption is rather technical, but it’s implications should seem apparent. Normal distributions are symmetric about their average, and have constant variance. This is exactly what we would expect from our random error, assuming we have fit the correct model. The value of any given y on the fitted regression line can be thought of as the average value y assumes for a given x. Of course, the true regression model has random error, which is why we have residuals in the first place. One would expect if the line is a “best fit” that most of the errors would be close to the line, the average value of y, with a few that deviate relatively further. We would also expect the probability of obtaining a residual some distance above our expected value to be the same as that of getting one of equal magnitude below. We can test this assumption by examining a plot of the residuals against our predicted y values. Such a plot is called a Normal probability plot. Fortunately, the plots for both models indicate this assumption to be true.