22S:172
Instructor: Cowles
Lab session 7

July 6, 2005

1 Background

The Federal Trade Commission annually rates varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide content. The US Surgeon General considers each of these substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke.

The dataset "cigarettes.dat" on the course web page contains measurements of weight as well as tar, nicotine, and carbon monoxide (CO) content for 25 brands of cigarettes. The data were taken from Mendenhall and Sincich (1992). The original source of the data is the Federal Trade Commission.

2 Preliminary analysis

Download the dataset into the `c:\temp*`directory.

Use the following data step to read this dataset into SAS;

```
data cigs ;
infile 'c:\temp\cigarettes.dat' ;
input  brand $16. tar nic wgt co1 ;
run ;
```

Bring up Insight to examine this dataset.

Our goal is to predict *co1*. We will first try two simple linear regressio ns. First, create scatter plots of *co1* vs. *nic* and of *co1* vs. tar. For each one, from the "Analyze" menu, select "Scatter plot." Designate the appropriate variables as X and Y, and click "OK" to produce the scatter plot.

Do the relationships between *co1* and each of the other two variables look fairly linear? Are there any outlying points?

Go back to the "Analyze" menu and select "Fit X Y." Specify the appropriate two

variables as X and Y, and click "OK" to fit each of these simple linear regression models.

1. Regress *co1* on *nic*
2. Regress *co1* on *tar*

What do the coefficients mean? Are the results more or less what you expected? Write the regression equation that you obtain from each model.

Are the slopes in the two models significantly different from zero at some reaso nable significance level?

3 Multiple linear regression

If *nic* and *tar* are useful predictors one at a time, perhaps they will do even better if used together. First, let's create a 3-dimensional plot of *co1*, *nic*, and *tar*. In Insight, choose "Analysis," "Rotating plot." Then make *co1* the Y variable and *nic* and *tar* the X and Z variables.

Note that, whereas a perfect linear relationship between a response variable an d a single predictor would look like a straight *line* in two dimensions, a perfect linear relationship between a response variable and two predictor variables would look like a *plane* in 3 dimensions.

Use the arrows to rotate the 3D plot in different directions. See whether you c an find a view that looks *roughly* like a flat plane.

What else do you notice about the plot?

Now fit the multiple regression model regressing *co1* on both *tar* and *nic*. Make *co1* the Y variable and both *tar* and *nic* X variables. Write the regression equation and comment on whether and how it differed from what you expected.

Perhaps the outlying point is the problem. Identify it by clicking on it in the rotating plot, which will highlight it and mark it on the data list. Then use "Edit," "Observations" to eliminate it from calculations. Note that we would not wish to do such deletion in a real analysis unless we had a good subject-matter reason to believe that the data was wrong or the observation was expected to be different from the others.

Rerun the regression. Write the resulting regression equation and comment.

Why might the coefficient for *nic* not be significantly different from zero in this model even though it was highly significant when *nic* was the only predictor?

Let's investigate the relationship between *nic* and *tar*. Create a scatter plot of *nic* vs. *tar*. Also, return to the Program Editor window and compute the correlation between *nic* and *tar*.

```
proc corr ;
 var nic tar ;
 where brand ne 'BullDurham' ;  * exclude obs ;
 run ;
```

What was the correlation?

Nic and *tar* are nearly "collinear." Is this good or bad with respect to their usefulness together as predictors in multiple regression?

When two or more candidate predictor variables are very highly correlated, then only one of them should be included in the model. The decision as to which to i nclude and which to omit is a subject-matter question, not a statistical questio n.

4 Further remarks

The data included in this dataset were obtained via smoking machines. Some studies (Davis et al. 1990, Coultas et al. 1993) suggest that measurements of the tar, nicotine, and carbon monoxide in cigarettes that come from smoking machines are not representative of actual human exposure from smoking.

Credit where credit is due: The basic outline for this lab was taken from:

McIntyre, Lauren. (1994) "Using Cigarette Data for an Introduction to Multiple Regression." *Journal of Statistics Education*.

5 Plots and analyses to be carried out *before* fitting a regression model

The remainder of this lab will deal with systematic steps in regression analysis.

Carry out item 2 in the list below in the program editor before going into Insight.

```
proc corr ;
var co1 tar nic wgt ;
run ;
```

Note that we hope to see high correlation between the response variable and each predictor. However, if there is extremely high correlation between 2 predictors (absolute value ¿ .95 or so), we probably should include only one of the pair in our regression model.

1. Univariate data summarization using proc univariate or Insight Distribution
2. Pairwise correlation analysis using proc corr or Insight Distribution
3. Univariate plotting using Insight
   (a) From the Insight pull-down menus, choose Analyze / Distribution. Enter the response variable as Y and click OK. Then do the same thing for each predictor variable one at a time. Notice from the plots whether the distributions appear skewed or symmetric; whether there are any outliers; whether there are any values that seem unreasonable.
   - According to the boxplots, for which variables are there outliers?
4. Scatterplot matrix using Insight
   (a) Choose Analyze / Multivariate

(b) Enter the response variable and each of the predictor variables in one long list under "Y"

(c) Then choose "Output" and check off Univariate, Corr, and Scatterplot matrix

(d) Click OK on the main window

(e) Are there any pairs of candidate predictor variables that are so highly correlated that only one should be included in the model?

(f) We will choose to retain the one that is either of more scientific interest or else easier to get data for.

6 Plots and diagnostics to carry out after fitting the regression model

1. Use "Fit" in Insight to regress *co1* on *nic* and *wgt*.
2. Note the plot of residuals versus predicted values that is automatically produced.
   (a) Do you see any possible outliers?
   (b) Do you see evidence of heteroscedasticity?
3. Notice the "Tolerance" and "Variance Inflation Factor" columns in the table of parameter estimates. Recall that predictors with high values of "VIF" are highly correlated with one or more other predictors.
4. Choose the "Graphs" menu that appears on the output window. Check "Residual Normal QQ" and "Partial leverage."
   (a) The "Residual Normal QQ" plot enables you to check whether the residuals resemble draws from a normal distribution.
   (b) "Partial leverage" or "added-variable" plots enable us to check whether the relationship between the response variable and each predictor is linear *assuming that the other predictors are already in the model*. It also lets us check for outliers with respect to each individual predictor, and to see whether each predictor is useful after the others are already in the model.
      i. If you click on a point in any plot, it will be highlighted in all the plots. Try this with an outlying point. Is the same observation an outlier in all the added-variable plots?
5. Choose the "Vars" menu that appears on the output window. Click on "Standardized residual" and "Cook's distance." These two variables will now appear as additional columns in the table.

6. Now use "Analyze" "Distribution" to make plots of the standardized residuals and Cook's distance. Click on any extreme outlying points to identify them.
   (a) An observation with an unusually large standardized residual is not well described by the model.
   (b) An observation with an unusually large Cook's distance is *influential*. The inference is likely to change depending on whether it is included in the dataset or deleted.
7. Let's see what happens if we delete the influential point. First make a note of the coefficients and the p-values for their individual t-tests and of R-squared, in the original fitted model. Then click on observation 3 in the spreadsheet, then click "Edit", "Observations", "Exclude from calculations."