Multiple Linear and Logistic Regression
Objectives

- Review the basics of Linear Regression
- Understand the use of Multiple Linear Regression versus Simple Linear Regression
- Understand the statistics that are included in the output from SPSS and how they should be interpreted
- Become familiar with Linear Regression Diagnostics, including model fit and residuals
- Understand Logistic Regression
- Compare the use of Linear and Logistic Regression
- Review an example of Multiple Logistic Regression
- Understand the statistics that are included in the output from SPSS and how they should be interpreted
Simple Linear Regression

- Forms a line of best fit using 2 variables
- Unlike correlation it is directional, therefore it has both an independent and a dependent variable
- The independent variable is also called a predictor or explanatory variable
- The dependent variable is also called the outcome or response variable
- The coefficient of the independent variables represents the impact of that variable on the outcome. If it is significant then the variable has a significant impact on the result.
- The statistics to interpret include the $r^2$ of the model, as well as the coefficient for the independent variable and the p value (significance) associated with it.
Multiple Regression

- Whereas Simple Linear Regression has 1 dependent and 1 independent variable, multiple linear regression has 1 dependent and multiple independent variables.
- Things get more complicated when there is more than one independent variable but the theory stays the same. The equation becomes $y = a + bx_1 + bx_2 + bx_3 \ldots$ to as many independent variables as you have – you will also get that many coefficients in your output.
- In the text you are introduced to the concept of confounders and ways to accommodate them including stratification. Multiple regressions are another way to do this. Actually, one of the major uses of multiple linear regression is to investigate the relationship of a specific independent variable to the dependent variable after adjusting for potential confounders.
- The other use of Multiple regression is to create a model that will allow you to predict the dependent variables based on the independent ones.
Interpretation of SPSS Output, 2
Independent Variables

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.828a</td>
<td>.686</td>
<td>.608</td>
<td>1.27272</td>
</tr>
</tbody>
</table>

This output tells you the model fit. The r of .828 suggests that the correlation is strong. The r² of .686 means that approximately 69% of the variance of the dependent variable is explained by the two independent ones.

<table>
<thead>
<tr>
<th>Model</th>
<th>Un-Standardized Coefficients</th>
<th>95.0% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>3.0</td>
<td>2.667</td>
</tr>
<tr>
<td>X3</td>
<td>.451</td>
<td>3.217</td>
</tr>
<tr>
<td>X4</td>
<td>-.099</td>
<td>-.705</td>
</tr>
</tbody>
</table>

The model represented here is:

\[ y = 3 + .451 \times X3 - .099 \times X4 \]
Why Include a Non-Significant Independent Variable?

• In the previous slide the variable X4 is not significantly related to the dependent variable, so why is it important?

• While the relationship between X4 and the dependent variable is not significant the presence of X4 may be very important to the model for 2 reasons.
  – First it may have some influence on the relationship between X3 and Y1 (the dependent variable in this case)
  – Second it may improve the model fit represented by the $r^2$. 
When is a Second Independent Variable a Confounder?

To decide whether one variable is a confounder to the relationship between 2 others you need to consider the following:

– Does the significance of the variable X3 change when we add X4? If the p value changes but the decision of whether it is significant does not change, it is not a confounder. So a change from .001 to .049 would not in itself suggest confounding, but a change from .049 to .062 would.

– Does the relationship change directions? To answer this we look at the coefficient for X3 and see whether it is negative or positive. Since it is positive both with and without X4 in the model we can say that the addition of X4 did not change the direction of the relationship between X3 and Y1.

– Since neither the significance nor the direction of the relationship between X3 and Y1 changed when X4 was added, X4 it is not a confounder of that relationship.
Residuals

• There is one other part of the regression output that you need to consider.
• These are the residuals. These are the values you get when you take the difference between the predicted y value (based on the independent variables) and the observed y value.
• Regression attempts to minimize these residuals to make the best fit – but it cannot eliminate them.
• An investigation of the residuals can tell you some things about your model – such as point out outliers, which might have had excess effect on the model, suggest transformations which could improve the model, or even make you question the validity of the model itself.
An Example of Residuals

• Here is a list of residuals such as you might see from an output

• Note it lists the number of each observation, the predicted y, the difference between the predicted y and the observed y (residuals), and the standard residuals – which is the residual divided by the standard error from the top of the output (in this example 7.41)

• Looking at this you can see that there are several observations which may have had an excessive influence on the line – of particular concern is observation 13 with a residual of -11.95. You could try running the regression again after omitting that observation and see how it affects your line

<table>
<thead>
<tr>
<th>Observation</th>
<th>Predicted Y</th>
<th>Residuals</th>
<th>Standard Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>146.9518</td>
<td>1.048227</td>
<td>0.163218</td>
</tr>
<tr>
<td>2</td>
<td>153.9651</td>
<td>1.96508</td>
<td>-0.30598</td>
</tr>
<tr>
<td>3</td>
<td>189.0316</td>
<td>0.968387</td>
<td>0.150786</td>
</tr>
<tr>
<td>4</td>
<td>99.61195</td>
<td>-8.61195</td>
<td>-1.34095</td>
</tr>
<tr>
<td>5</td>
<td>41.75217</td>
<td>9.247828</td>
<td>1.439962</td>
</tr>
<tr>
<td>6</td>
<td>129.4185</td>
<td>-6.41851</td>
<td>-0.99941</td>
</tr>
<tr>
<td>7</td>
<td>199.5516</td>
<td>3.55157</td>
<td>-0.55301</td>
</tr>
<tr>
<td>8</td>
<td>113.6386</td>
<td>-2.63857</td>
<td>-0.41085</td>
</tr>
<tr>
<td>9</td>
<td>150.4584</td>
<td>-1.541574</td>
<td>0.240036</td>
</tr>
<tr>
<td>10</td>
<td>176.7583</td>
<td>-0.75833</td>
<td>-0.11808</td>
</tr>
<tr>
<td>11</td>
<td>129.4185</td>
<td>-7.41851</td>
<td>-1.15512</td>
</tr>
<tr>
<td>12</td>
<td>206.5649</td>
<td>-1.43512</td>
<td>0.22346</td>
</tr>
<tr>
<td>13</td>
<td>146.9518</td>
<td>-11.9518</td>
<td>-1.86099</td>
</tr>
<tr>
<td>14</td>
<td>169.745</td>
<td>8.25498</td>
<td>1.285368</td>
</tr>
<tr>
<td>15</td>
<td>117.1452</td>
<td>-8.854781</td>
<td>1.378762</td>
</tr>
<tr>
<td>16</td>
<td>185.525</td>
<td>3.47504</td>
<td>0.541092</td>
</tr>
</tbody>
</table>
Warnings about Multiple Regression

• You can use this equation of a line to predict the dependent variable based on one or more x variables as long as you stay within the range of your line. If you extend your line beyond the known x values (in either direction) and try to predict the corresponding y value, you can run into trouble because in nature few things are that predictable.

• For example, if x ranges from 25 to 125 you can plug in a value of 35 for x and predict a value of 62.7 for y. If, however, you plug in a value of 200 for x and predict a value of 351.43 for y, you will probably get erroneous results.
Other Considerations

- Statisticians often use residuals and graphical representations of the model and of residuals for diagnostics and to determine if transformations are necessary.
- Transformations are adjustments made to the variables that do not affect the overall relationship but allow you to get a better line.
- The most commonly used transformation is the Logit, which gives you logistic regression, which you will read about as well in this lecture.
- Other common transformations are squaring the x and later taking the square root of the y (to get the real value), adding a number to the x and then later subtracting it from the y, or multiplying the x by a value and then later dividing the y by that same value.
- The proper choice of transformation can often be suggested by looking at the graph of residuals but it takes a bit of training to get good at doing this.
Logistic Regression

• Now that you have a basic understanding of linear regression, you can learn a bit about logistic regression as well.

• Many of the applications in medicine will require you to predict variables that are not normally distributed. If the dependent variable can only be yes or no disease, for example, you cannot expect to use predictor variables to “fit” a line. Let’s say you set yes as 1 and no as 2, then for every x value you will have a y value of 1 or 2. A scatter plot would look like this:

You cannot “fit” a line to this scatter plot.
Logistic Regression

- SO… some statisticians, way back, determined that you could transform the y variable into something that would give you a line.

- In this case they used a form of log transformation (models using the log transformation are in general referred to as log-linear models), called the logit.

- The logit is the log of the odds ratio (log $y/1-y$) or expressed simply in terms of $y$: $y = (\exp x)/(1 + \exp x)$ (fortunately you do not need to know this formula or how to use it to understand the output of logistic regression).

- This transformation has some helpful properties. It yields values for $y$ between 0 and 1 – corresponding to odds or probability of disease. Further, interpretation of results is simplified because the coefficients ($b$) of the independent variables ($x_1, x_2, x_3$) can be transformed into odds ratios by exponentiation (again do not worry that you need to be able to explain this right now).
Linear versus Logistic Regression

Linear
- Dependent or Outcome variable is continuous so can form a line with the independent variables
- Has 1 dependent variable and 1 or more independent variables
- Can be used to adjust for confounders or to model an outcome
- The coefficient of the independent variables are interpreted as incremental increases in the outcome variable similar to correlation

Logistic
- Dependent or Outcome variable is dichotomous so cannot form a line with the independent variables unless it is converted using the Logit transformation
- Has 1 dependent variable and 1 or more independent variables
- Can be used to adjust for confounders or to model an outcome
- The coefficient of the independent variables are interpreted as odds ratios in relation to the dependent variable
An Example of Logistic Regression

Let’s say you want to predict whether a person will develop a disease (y) based on a set of risk factors (x₁ and x₂). We use logistic regression to fit a line with the equation

\[ y = 1 + \exp(2.3 + .045x_1 + .408x_2) \]

In this equation \( e^{2.3} \) represents the intercept which in logistic regression has no meaning and can be ignored, \( e^{0.045} \) yields the odds of disease associated with factor \( x_1 \) and \( e^{0.408} \) represents the odds of disease associated with factor \( x_2 \). So people with factor \( x_1 \) have \( e^{0.045} \) or 1.04 times the risk of disease, while those with factor \( x_2 \) have \( e^{0.408} \) or 1.50 times the risk of disease.
Interpreting Results

• Most statistical packages will give you the output of a logistic regression with the estimated coefficient, the estimated standard deviation, and the estimated odds ratio. Some will also calculate the confidence limits for you.

• As with all odds ratios, interpretation is the odds of disease y given the presence of factor x is … and significance is determined by either a p value less than 0.05 or a confidence interval that does not include the number 1.

• A negative coefficient will yield an estimated odds ratio of less than 1 which will imply that the person with that factor actually has a reduced chance of disease, while a positive coefficient and an odds ratio greater than one implies increased risk.
When to Use Logistic Regression

- The true utility of logistic regression is the ability to calculate odds ratios in the presence of other factors without having to stratify. The 2x2 table and Chi-Square does not allow you to adjust the odds ratio based on other factors – we say it yields a crude odds ratio.

- If you want to adjust for age or sex, for example, you can stratify, but if there are too many factors you need to adjust for, then stratification becomes impractical and you need to use a modeling technique like linear or logistic regression to get adjusted odds ratios.

- But interpretation can be complicated by interactions, unmet assumptions, models with a poor fit, and other issues.
Another Example

A Logistic Regression Model is:

\[ \text{Outcome (Heart disease yes or no)} = 1 + \exp(a + b_1(\text{smoking}) + b_2(\text{family history}) + b_3(\text{age category}) + b_4(\text{cholesterol level})) \]

Where smoking yes=1 (referent) and no=2; family history 1=yes (referent) and no=2; age categories 1=under or equal to 30 (referent); 2=31-45; 3=46-60; 4=over 60; and cholesterol level is the actual number minus 120 (not categorical)

If the results were reported as the following (fictional values):

- \( a = .431 \)
- \( b_1 = -.233 \)
- \( b_2 = -.345 \)
- \( b_3 = (#2 \text{ vs. } #1 = .22, #3 \text{ vs. } #1 = .33 \text{ and } #4 \text{ vs. } #1 = .44) \)
- \( b_4 = .02 \)

The interpretation of these would be:

\( b_1 \): The odds ratio of heart disease for non-smoker vs. smokers= .79, or non-smokers have a lower risk of developing heart disease

\( b_2 \): The odds ratio of heart disease for people with no family history vs. family history= .71, or non-history have a lower risk of developing heart disease

\( b_3 \): The odds ratio of heart disease for cat 2 age vs. <=30=1.25, cat 3 v<=30=1.39, and cat 4 v<=30=1.55, or risk of heart disease increases with age category

\( b_4 \): The odds ratio of heart disease associated with cholesterol level is 1.02, or for each increase of 1 unit over 120, the risk of heart disease increases 1.02.

***Note with real values you would also have confidence intervals, which are interpreted as significant if they do not include the number 1, for example the odds ratio of 1.39 would be significant if the 95% CI was (1.25, 1.49), while the odds ratio of 0.79 would not if its 95% CI was (0.34, 1.04)***
Logistic Regression in Excel

• The Data Analysis Package we have been using does not have the option of performing logistic regression
• Most other software designed to do statistical analysis offers this
• But…that still does not mean you should use it indiscriminately-keep that in mind when you read other’s research
Goodness of Fit

- In Linear Regression you examine the residuals to see how closely the predicted (y)s match the actual ones, you also look at the $r^2$ to estimate the model fit.
- In Logistic Regression – the Goodness of Fit also examines the residuals to determine if the model really reflects the data.
- Goodness of Fit uses the Chi Square test statistic that you have used before, remember that with Goodness of Fit you do not want to reject the null hypothesis – you want the test statistic to be small and the p value to be large— in fact the smaller the better.
- The less difference there is between the expected and observed result (a small Chi Square) the better the model is at predicting the outcome or dependent variable.
Interpretation of SPSS Output, 2
Independent Variables

In this example we test the relationship between the independent variable “alcs” and bladder cancer after adjusting for income level. The referent value for bladder cancer is 1 or No, while the referent value for alcohol (alcs) is 2, or less than 4 drinks per month.

This output tells you that the Model is significant with a p value related to the Chi Square of .001

This output tells you that the Chi Square Goodness of Fit is not significant since the significance associated with it is .155 (greater than .05), therefore the model is a reasonably good fit although it may not do a good job of explaining all the variance of the dependent model.
This output tells you that the coefficient (B) of the independent variable Inc is .255. This converts to an odds ratio that suggests with increase in income level the odds of bladder cancer increases 1.291. This is not significant at .071. It is also not significant based on the confidence interval. The coefficient (B) of the independent variable alcs = 1 is .328 with reference to the variable alcs=2. This converts to an odds ratio of 2.904 with a significance of .001. This means that a person who fits the category of alcohol consumption represented by the number 1 is 2.904 times as likely to have bladder cancer as the person represented by alcs=2. This is a significant finding and would lead you to reject the null hypothesis of no association between alcohol consumption and bladder cancer.