Multiple Regression Handout

The purpose of this handout is to take you through all the pieces of a multiple regression printout, explaining in detail what each one means. I will use the STATA printout from our class example of heart rate after jogging as the basis for this handout but I have changed a few variables to simplify the interpretations

The variables are as follows:

Y = Heart Rate in beats per minute
X_1 = Distance run in miles
X_2 = Average speed during the run in miles per hour
X_3 = Humidity as a percentage
X_4 = Altitude in thousands of feet above sea level
X_5 = Temperature in degrees Farenheit
X_6 = Gender: $X_6 = 1$ for males and 0 for females
X_7 = Fitness: $X_7 = 1$ if the person exercises regularly and 0 if not
X_8, X_9 = Terrain (flat, hilly or on sand): These are coded as $X_8 = X_9 = 0$ for flat, $X_8 = 1$ and $X_9 = 0$ for hilly and $X_8 = 0, X_9 = 1$ for on sand.

Note: $\bar{Y} = 100, n = 62, p = 9$ where n is the sample size and p is the number of predictors.

The STATA printout for the multiple regression is as follows:

```
. reg HRate Distance Time Speed Weight RestHRate Gender Fitness Hilly Sand
      Source |       SS      df       MS              Number of obs =      62
-------------+------------------------------          F(  9,    52) = 160.889
       Model |    23168       9     2574.22           Prob > F      =  0.0000
    Residual |      832      52       16.00           R-squared     =   0.965
-------------+------------------------------          Adj R-squared =   0.959
       Total |    24000      61      393.44           Root MSE      =   4.000

------------------------------------------------------------------------------
       HRate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Distance |       3.0       1.80      1.67    0.101      -0.62       6.62
        Time |     -0.05       0.04     -1.25    0.217      -0.11       0.03
    Humidity |       0.0       1.30      0.00    1.000      -2.61       2.61
 Temperature |       0.1       0.04      2.50    0.016       0.02       0.18
    Altitude |       1.0       0.25      4.00    0.000       0.50       1.50
      Gender |      -4.0        1.6     -2.50    0.016      -7.22      -0.78
     Fitness |     -10.0        3.6     -2.80    0.004     -17.24      -2.76
       Hilly |       5.0        1.2      4.17    0.000       2.59       7.41
        Sand |       5.5        1.0      5.50    0.000       3.49       7.51
       _cons |      70.0       10.0      7.00    0.000      50.10      90.10
------------------------------------------------------------------------------
```

Interpreting Regression Coefficients

Intercept: β_0 for the population, b_0 for the sample. The intercept tells us the average value of Y when **all the X's are 0.** Here $b_0 = 70$ in the row labeled "_cons" gives us the average heart rate of a woman ($X_6 = 0$) who does not exercise regularly ($X_7 = 0$) who has run 0 miles ($X_1 = 0$) in 0 minutes ($X_2 = 0$) at 0 humidity ($X_3 = 0$) at a temperature of 0 F ($X_4 = 0$) at sea level ($X_5 = 0$) on flat ground ($X_8 = X_9 = 0$). In other words the **resting heart rate** of a female couch potato under these atmospheric and geographical conditions is 70 beats per minute. With my adjusted list of variables this value actually makes sense! The corresponding confidence interval for β_0 is [50.10, 90.10] which means that we are 95% sure that the average resting heart rate of a woman under these conditions is somewhere between 50.1 and 90.1 beats per minute. This interval lies entirely above 0, meaning we are sure that such a person has a positive heart rate (fortunate or she would be dead!) and this is confirmed by the p-value for the correpsonding t-test which is 0 (more on that below).

Slope for X_1: β_1 for the population, b_1 for the sample. The slope tells us the average change in Y associated with a one unit change in X_1, **assuming all the other variables are held fixed**. Here $b_1 = 3$ means that all else equal, each extra mile run is associated with the heart rate being 3 beats per minute higher. Another reasonable way to think about this is that if I had two people of the same gender and exercise habits running in the same conditions (temperature, altitude, humidity, terrain) at the same speed, but one ran a mile further than the other, then on average I would expect the one who ran further to have a heart rate 3 beats per minute higher at the end of the run. The corresponding confidence interval of [-.62, 6.62] means that all else equal, each additional mile run is associated with anywhere from a .62 beats per minute decrease to a 6.62 beats per minute increase in hear rate. Since this interval overlaps 0 it is possible that after taking all the other variables into account, the distance run provides no additional information about exercising heart rate. The p-value of the corresponding t-test at .101 is larger than our significance level, $\alpha = .05$, which reinforces this idea. (more on this below.)

Slope for X_2: The estimated slope for the time variable is $b_2 = -.05$ which means all elese equal, each additional minute run is associated with a heart rate that is lower on average by .05 beats per minute, or more intuitively, every extra hour of jogging (60 minutes) is associated with a decrease in heart rate of 3 beats per minute (.05*60). This is surprising. We would expect that the longer you jogged, the higher your heart rate would be. However we note that the confidence interval for β_2 includes 0 so we can not be sure that there is association with time after the other variables are included in the model. Most likely this has to do with a connection between the distance and time variables. The longer you run, the further you will go and it may be that once we know the distance run, the time of the run provides no further information.

Slopes for X_3-X_5: The coefficient of the humidity variable, $b_3 = 0$ means that humidity appears to make no contribution to this model after the other factors have been taken into account. This could be because humidity is related to some of the other factors or because it really has no impact on heart rate. The coefficient $b_4 = .1$ for temperature means that all else equal, a 1 degree increase in temperature is associated with an increase of .1 beats per minute in heart rate, or equivalently that a 10 degree increase in temperature is associated with a 1 beat per minute increase in heart rate. The corresponding confidence interval suggests that the effect could in fact be anywhere from .02 to .18 beats per minute. The coefficient $b_5 = 1$ means that all else equal, an increase in altitude of 1000 feet (a 1 unit change in X_5) is associated with an increase in heart rate of 1 beat per minute while the confidence interval implies that the effect could be as little as .5 beats per minute to as much as 1.5 beats per minute per 1000 feet of elevation. Be careful of your units! The confidence intervals for these two latter coefficients lie entirely above 0, meaning that even after adjusting for the other factors, temperature and alttitude have a positive relationship with heart rate.

Slopes for X_6-X_9: The last four variables in this model are indicators or dummy variables. An indicator variable, as its name suggests, "indicates" whether a subject has a particular characteristic of interest. Usually the indicator is 1 if the subject has the characteristic and 0 if they do not, though that choice is very

arbitrary. We are told that the gender variable, $X_6 = 1$ for males and $X_6 = 0$ for females. For an indicator variable it does not really make sense to talk about a slope. You can't have a "1 unit change in gender." What you can do is talk about the difference between males and females. With an indicator variable, the coefficient tells you the average difference in Y between people who have the characteristic (X = 1) and those who do not (X = 0). For gender in this model we have $b_6 = -4$ meaning all else equal (same distance and time run, humidity, temperature, altitude, fitness and terrain), a man ($X_6 = 1$) will have a heart rate that is 4 beats per minute slower (the negative sign) than a woman ($X_6 = 0$). The confidence interval of [-7.22, -.78] says we are 95% sure that on average a man has an exercising heart rate somewhere between .78 and 7.22 beats per minute slower than an otherwise equivalent woman. Since the entire interval is below 0, we are confident that the men do indeed have the slower heart rates on average. Similarly, the coefficient $b_7 = -10$ means that a person who exercises regularly will have a heart rate on average 10 beats per minute slower after exercise than a person who does not exercise regularly will have under the same conditions. Once again the entire interval is below 0, meaning that all else equal, regular exercise is associated with a lower exercising heart rate with the amount of the decrease being anywhere from 17.24 to 2.76 beats per minute from the confidence interval. The last measure of interest in this model is the terrain on which the race was run. We have three categories of terrain: flat, hilly and sandy. When you have a qualitative variable with more than 2 categories you use multiple indicators to describe it. One group serves as the **reference** and has all the indicators equal to 0 and the other categories are compared to that reference. Here flat terrain, $X_8 = X_9 = 0$ is the reference category and hilly ($X_8 = 1, X_9 = 0$) and sandy ($X_8 = 0, X_9 = 1$) are compared to it with the coefficients of X_8 and X_9 giving the respective differences. Specifically, $b_8 = 5.0$ means all else equal, a person running on flat terrain will have an exercising heart rate 5 beats per minute higher than someone running on flat terrain, while $b_9 = 5.5$ means someone running on sandy terrain will have a heart rate of 5.5 beats per minute higehr than someone who does a comparable run on flat terrain. Of course we can deduce from this that someone running on sand will on average have a heart rate .5 beats per minute faster than someone running on hilly terrain. The corresponding confidence intervals give the possible ranges for these differences. Note that the CIs for β_8 and β_9 overlap quite a bit meaning that running on hilly or sandy terrain may have an equal impact on heart rate compared to flat terrain–in other words those two surfaces may not be significantly different from each other, although to be sure we would need to test that directly.

Measuring Model Performance

As in simple linear regression, to evaluate how good a job the model is doing we divide the total variability in Y into a piece that can be attributed to our X variables and a piece that can not. Here that amounts to saying that some of the differences in exercising heart rate can be attributed to distance and time run, atmospheric conditions such as humidity and termperature, geographical conditions such as altitude and type of terrain, and to personal characteristics including gender and basic fitness. However there may be other factors which we have not included that will cause a model based on these variables not to make perfect predictions. This division of the variability into two components results in an ANOVA table which in turn provides our usual measures for comparing regression models, R^2, R^2_{adj} and RMSE. These values can be interpreted as follows.

SST: The "Sum of Squares Total" is a measure of the total variability in Y and is just the variance in Y without dividing by n-1. It serves as our reference point for deciding how well a regression model is fitting. You can think if it as the total (squared) error we would make in predicting the Y values in our sample if we ignored the information in our X variables and just always guessed $\bar{Y}$. The raw number is hard to interpret as anything other than a reference point, especially as it gets bigger as the sample size increases. Here SST = 24,000 is the total variability in the exercising heart rates of the people in our jogging sample. Some if it is due to the 9 predictor variables and some of it is due to other factors that we did not measure.

SSR: The "Sum of Squares for Regression" is the variability in Y that is explained by the X variables as a group. Here SSR = 23,168 represents the total variability in exercising heart rate that is explained by distance and time run, the current humidity, temperature, altitude and terrain, and the gender and fitness of the runners in our sample. The actual numeric value is hard to interpret without comparing it to SST which is how we get R^2 (see below). Here our value of 23,168 represents a huge proportion of the total variability of 24,000 so we have explained quite a lot of the variability we started with. Note that this does not say WHICH of these variables are useful or how much each one explains. For this we must look at the individual t-tests and also examine simple linear regressions of Y on each of the X variables to see how much each variable can explain on its own.

SSE: The "Sum of Squared Errors" is the variability in Y that is NOT explained by any of the X's. It is the error that still remains even after taking advantage of what the X's tell you about Y. The number is hard to interpret except in reference to SST. Here SSE = 832 is the variability in exercising heart rate that is due to something other than distance, time, humidity, temperature, altitude, gender, fitness or terrain. This number is relatively small compared to our SST of 24,000 meaning that we have identified most of the important factors.

MSR: The "Mean Squares for Regression" gives the average variability in Y explained **per predictor variable**. Sepcifically, $MSR = SSR/p$ where p is the number of predictors. Higher obviously is better as it suggests that each variable contributes a substantial amount to the model on average. Note however that some variables may contribute more than others–there is no need for them all to contribute the same amount! Here MSR = 2574.22 which means that each of our 9 variables explains about this much of the total variability of 24,000 that exists in our joggers' heart rates which sounds pretty good.

MSE: The "Mean Squared Error" is the average squared error **per data point** that we make when we use all the X's to predict Y. Sepcifically $MSE = SSE/(n-p-1)$ where n is the sample size and p is the number of predictors. You can think of MSE as the variance of the points about the estimated regression surface. You square root it to get the RMSE which is a more useful number. Here MSE=16 is the average squared error we make in predicting heart rate using our list of 9 predictors.

MST: The "Mean Squares Total" is just the variance of the Y's, in this case the variance of the exercising heart rates in our sample.

R^2 : $R^2 = SSR/SST$ gives the percentage of variability in Y that is explained by **all** the X variables as a group. In our example $R^2 = .965$ meaning that distance, time, humidity, temperature, altitude, gender, fitness and terrain colelctively explain 96.5% of the variability in exercising heart rate. This is an exceedingly high percentage. In this sense our model is doing a very good job. However, R^2 does not take into account the size of your sample and the number of variables you are using to explain Y. If the number of variables is high compared to the the number of data points, R^2 (which always increases as you add more variables) can produce quite inaccurate estimates of how much the model will explain for new data points which is really the measure of interest. Thus we get the following improved measure:

R^2_{adj}: This has the same interpretation as R^2 except that in multiple regression it tends to be much more accurate because it penalizes you for including too many variables that do not significantly improve the model. Specifically $R^2_{adj} = 1 - (SSE/SST) * (n - 1/n - p - 1)$. So if p goes up a lot but SSE does not correspondingly go down a lot R^2_{adj} can actually decrease as you add more variables. (Note that this is different from saying R^2_{adj} is lower than R^2 which is always true. To actually see R^2_{adj} increase or decrease you would need to fit separate models with different numbers of variables.) Here $R^2_{adj} = .950 = 95.9\%$ which is very similar to R^2, suggesting that we do not have a major problem with **overfitting**, the term for adding lots of useless variables to the model.

RMSE = MSE/(n-p-1): This number is called "root mean squared error" and tells you the average error **per data point** that you make when you use **all** the X's to predict Y. Here $RMSE = 4.0$ means that on average when we predict a person's exercising heart rate using distance, time, humidity, temperature, altitude, gender, fitness and terrain we are typically off by about 4 beats per minute. We were told that the average exercising heart rate for people in this data set was $\bar{Y} = 100$ so our average error is about 4%. The model seems to make quite accurate predictions. Note that you don't have to compare to $\bar{Y}$ specifically–sometimes it is useful to look at the minimum and maximum Y to get an idea of best-case/worst-case scenarios for instance–but it is critical to use the context of the problem to determine how good the predictions are.

F: F = MSR/MSE is the ratio of explained to unexplained variability. The bigger it is, the better a job the model **as a whole** is doing and the more sure we wuill be that there is a statistically significant relationship between Y and **at least one of the X variables.** On an intuitive level, F = 1 represents a balance between explained an unexplained variability so we are looking for values much bigger than 1. To really fell whether your F statistic is big you look at the corresponding p-value (Prob > F) on the STATA printout). A small p-value means your F counts as big (see below for details). Here our F = 160.889 which is much bigger than 1 and the corresponding p-value of 0 is very small so it looks like the model as explained much more of the variability in heart rates than it hasn't explained. We can make this into a more formal test if we make some assumptions.

Inference In Multiple Regression

The summary measures above derive naturally from the least squares formulation of teh regresison line and do not implicitly make any assumptions about the distributions of the various pieces of the model. However, if we do make some assumptions we can actually obtain confidence intervals for the various model parameters, and also perform hypothesis tests about the model as a whole and the contributions of the individual variables. Specifically we assume that the error terms on our model are independent, normally distributed, mean 0 and have constant variance for all values of X. Under these assumptions the least squares estimates $b_0, b_1, \ldots, b_p, MSE$ are unbiased estimates of $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ and the esitmated slope and intercept are normally distributed with standard errors we can compute using matrix algebra. From this we can get the following forms of inference.

Overall F test: The first thing we would like to determine is whether the model as a whole is useful. Recall that our multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon$$

If all the β's are 0 then none of the X's will make a contribution to predicting Y and the model will not be useful. We therefore want to test (in the context of the current example):

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_9 = 0$–none of distance, time, humidity, temperature, altitude, gender, fitness or terrain helps to explain exercising heart rate. The model as a whole is not useful.
H_A : At least one of $\beta_1, \ldots, \beta_9 \neq 0$–at least one of distance, time, etc. IS useful for explaining heart rate after jogging. The model as a whole IS useful.

As discussed above, a reasonable test statistic for looking at this set of hypotheses is $F_{obs} = MSR/MSE$. The bigger this is, the more of the variability in heart rates in our sample has been explained by the predictors and the more convinced we are that the model as a whole is useful. Here we have F = 160.889. Under the null hypothesis, F_{obs} has an F distribution with p and n-p-1 degrees of freedom and the corresponding p-value is

$$P(F_{9,52} \geq 160.889) = 0.0000$$

so we reject the null hypothesis and conclude (not surprisingly!) that at least one of these 9 factors does have something to do with heart rate. Of course this test does not tell us WHICH variable or variables is useful. For that we need the following.

Individual t-tests: To determine whether a variable makes a useful contribution to a multiple regression model you use a t-test. Note that this test does NOT ask whether the variable **by itself** is related to Y. It instead asks whether the variable in question provides any information about Y **that was not provided by the other variables in the model.** Since your predictor variables can be related to each other it is quite possible that some of them provide the same information about Y and that as a result a variable that by itself is related to Y may not be significant as part of the larger model. Formally the hypotheses work as follows. I use X_1, the distance variable, as an example in this model. The others work similarly:

$H_0 : \beta_1 = 0$–after accounting for time run, humidity, temperature, altitude, gender, fitness and terrain, knowing the distance a person has run does not provide any additional information about exercising heart rate. X_1 is not worth adding to a model tha already contains these other factors.
$H_A : \beta_1 \neq 0$–distance run provides additional information about heart rate beyond what is explained by time, humidity, etc. It is worth adding to the model.

Our test statistic is

$$t_{obs} = \frac{b_1 - 0}{s_{b_1}} = 1.67$$

Under the null hypothesis t_{obs} has a t distribution with n-p-1 and the two-sided p-value (which is what is give on the STATA regression printout next to the t statistic) is

$$p - value = 2P(t_{52} \geq 1.67) = .101$$

Since this p-value os greater than $\alpha = .05$ we fail to reject the null hypothesis. We do not have sufficient evidence to conclude that the distance run provides additional infromation about heart rate than what is explained by the other 8 variables. It may not be a useful addition to this model.

Of the 9 variables in the current model, all except distance, time and humidity appear to be significant with p-values $< .05$. However we can't simply decide to remove these variables from the model all at once because their p-values depend on the presence of ALL the other variables. For instance it is quite possible that distance and time of the run are providing the same information and if we removed one of them the other would be significant.

Confidence Intervals for $\beta_j's$: The STATA printout also gives confidence intervals for the intercept and each of the slopes in the regression model. These are in the last column of the parameter estimates table. If they were not given we could get them from the parameter estimates themselves and their standard errors which are given in the columns labeled "Coef." for coefficients and "STd. Err." for standard error. Specifically a confidence interval for a β in a regresison model is given by

$$b_j \pm t_{\alpha/2,n-p-1} s_{b_j}$$

Here for a 95% confidence interval we would need $t_{.025,52} = 2.01$ so the confidence interval for β_1 for instance is

$$3.0 \pm 2.01(1.8) = [-.62, 6.62]$$

The other intervals are obtained analagously and their interpretations have been given above along with the interpretations of the intercept and slopes.