Multiple Regression

Now we're going to look at the rest of the data that we collected about the weight lifters. We will still have one response (y) variable, clean, but we will have several predictor (x) variables, age, body, and snatch. We're not going to use total because it's just the sum of snatch and clean.

Data

The heaviest weights (in kg) that men who weigh more than 105 kg were able to lift are given in the table.

Data Dictionary

Age
: The age the competitor will be on their birthday in 2004.

Body
: The weight (kg) of the competitor

Snatch
: The maximum weight (kg) lifted during the three attempts at a snatch lift

Clean
: The maximum weight (kg) lifted during the three attempts at a clean and jerk lift

Total
: The total weight (kg) lifted by the competitor

Age	Body	Snatch	Clean	Total
26	163.0	210.0	262.5	472.5
30	140.7	205.0	250.0	455.0
22	161.3	207.5	240.0	447.5
27	118.4	200.0	240.0	440.0
23	125.1	195.0	242.5	437.5
31	140.4	190.0	240.0	430.0
32	158.9	192.5	237.5	430.0
22	136.9	202.5	225.0	427.5
32	145.3	187.5	232.5	420.0
27	124.3	190.0	225.0	415.0
20	142.7	185.0	220.0	405.0
29	127.7	170.0	215.0	385.0
23	134.3	160.0	210.0	370.0
18	137.7	155.0	192.5	347.5

Regression Model

If there are k predictor variables, then the regression equation model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$.

The x_1, x_2, ..., x_k represent the k predictor variables. Those parameters are the same as before, β_0 is the y-intercept or constant, β_1 is the coefficient on the first predictor variable, β_2 is the coefficient on the second

predictor variable, and so on. ε is the error term or the residual that can't be explained by the model. Those parameters are estimated by b_0, b_1, b_2, ..., b_k.

This gives us a regression equation used for prediction of $y = b_0 + b_1x_1 + b_2x_2 + ...+ b_kx_k$.

Basically, everything we did with simple linear regression will just be extended to involve k predictor variables instead of just one.

Regression Analysis Explained

Round 1: All Predictor Variables Included

Minitab was used to perform the regression analysis. This is not really something you want to try by hand.

Response Variable: clean
Predictor Variables: age, body, snatch

Regression Equation

```
The regression equation is
clean = 32.9 + 1.03 age + 0.106 body + 0.828 snatch
```

There's the regression equation. You can use it for estimation purposes, but you really should look further down the page to see if the equation is a good predictor or not.

Table of Coefficients

```
Predictor     Coef  SE Coef     T      P
Constant     32.88    28.33  1.16  0.273
age         1.0257   0.4809  2.13  0.059
body        0.1057   0.1624  0.65  0.530
snatch      0.8279   0.1371  6.04  0.000
```

Notice how the coefficients column (labeled "Coef") are again the coefficients that you find in the regression equation. The constant 32.88 is b_0, the coefficient on age is $b_1 = 1.0257$, and so on.

Also notice that we have four test statistics and four p-values. That means that there were four hypothesis tests going on and four null hypotheses. The null hypothesis in each case is that the population parameter for that particular coefficient (or constant) is zero. If the coefficient is zero, then that variable drops out of the model and it doesn't contribute significantly to the model.

Here's a summary of the table of coefficients. We're making our decision at an $\alpha = 0.05$ level of significance, so if the p-value < 0.05, we'll reject the null hypothesis and retain it otherwise.

Predictor	P	Null Hyp.	Decision	Conclusion
Constant	0.273	$\beta_0 = 0$	Retain H_0	The constant appears to be zero. Even so, we leave it in the model.
age	0.059	$\beta_1 = 0$	Retain H_0	Age does not significantly contribute to the ability to perform the clean & jerk

body	0.530	$\beta_2 = 0$	Retain H_0	Body weight does not significantly contribute to the ability to perform the clean & jerk
snatch	0.000	$\beta_3 = 0$	Reject H_0	The weight one is able to snatch does significantly contribute to the ability to perform the clean & jerk

A note about the T test statistics. They are once again the coefficient divided by the standard error of the coefficient, but this time that don't have n-2 degrees of freedom. If you remember what we wrote during simple linear regression, the df for each of these tests was actually the sample size minus the number of parameters being estimated. Well, in this case, we have four (4) parameters we're estimating, the constant and the three coefficients. Since our sample size was n = 14, our df = 14 - 4 = 10 for these tests.

A further note - don't just blindly get rid of every variable that doesn't appear to contribute to the model. This will be explained later, but there are correlations between variables that don't show themselves here.

Analysis of Variance

This is why we're really here, but if we take what we learned in simple linear regression and apply it, it's not that difficult to understand.

```
Source           DF      SS      MS      F      P
Regression        3  3558.0  1186.0  20.20  0.000
Residual Error   10   587.1    58.7
Total            13  4145.1
```

Notice how the total line is exactly the same as it was for the simple linear regression? That's because the response variable, clean, is still the same. All that has happened is that the amount of variation due to each source has changed.

Here's the table we saw with simple linear regression with the comments specific to simple linear regression removed. The same instructions work here with multiple regression.

Source	SS	df
Regression (Explained)	Sum the squares of the explained deviations $\sum(\hat{y} - \bar{y})^2$	# of parameters - 1 # of predictor variables (k)
Residual / Error (Unexplained)	Sum the squares of the unexplained deviations $\sum(y - \hat{y})^2$	sample size - # of parameters n - k - 1
Total	Sum the squares of the deviations from the mean $\sum(y - \bar{y})^2$	sample size - 1 n - 1

The df(Regression) is one less than the number of parameters being estimated. There are k predictor variables and so there are k parameters for the coefficients on those variables. There is always one additional parameter for the constant so there are k+1 parameters. But the df is one less than the number of parameters, so there are k+1 - 1 = k degrees of freedom. That is, the df(Regression) = # of predictor variables.

The df(Residual) is the sample size minus the number of parameters being estimated, so it becomes df(Residual) = n - (k+1) or df(Residual) = n - k - 1. It's often easier just to use subtraction once you know the total and the regression degrees of freedom.

The df(Total) is still one less than the sample size as it was before. df(Total) = n - 1.

The table still works like all ANOVA tables. A variance is a variation divided by degrees of freedom, that is MS = SS / df. The F test statistic is the ratio of two sample variances with the denominator always being the error variance. So F = MS(Regression) / MS(Residual).

Even the hypothesis test here is an extension of simple linear regression. There, the null hypothesis was H_0: $\beta_1 = 0$ versus the alternative hypothesis H_1: $\beta_1 \neq 0$.

In multiple regression, the hypotheses read like this:

H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$
H_1: At least one β is not zero

The null hypothesis claims that there is no significant correlation at all. That is, all of the coefficients are zero and none of the variables belong in the model.

The alternative hypothesis is not that every variable belongs in the model but that at least one of the variables belongs in the model. If you remember back to probability, the complement of "none" is "at least one" and that's what we're seeing here.

In this case, because our p-value is 0.000, we would reject that there is no correlation at all and say that we do have a good model for prediction.

Summary Line

```
S = 7.66222   R-Sq = 85.8%   R-Sq(adj) = 81.6%
```

Recall that all the values on the summary line (plus some other useful ones) can be computed from the ANOVA table.

First, the MS(Total) is not given in the table, but we need it for other things. MS(Total) = SS(Total) / df(Total), it is not simply the sum of the other two MS values. MS(Total) = 4145.1 / 13 = 318.85. This is the value of the sample variance for the response variable clean. That is, $s^2 = 318.85$ and the sample standard deviation would be the square root of 318.85 or s = 17.86.

The value labeled S = 7.66222 is actually s_e, the standard error of the estimate, and is the square root of the error variance, MS(Residual). The square root of 58.7 is 7.66159, but the difference is due to rounding errors.

The R-Sq is the multiple R^2 and is R^2 = (SS(Total) - SS(Residual)) / SS(Total).

R^2 = (4145.1 - 587.1) / 4145.1 = 0.858 = 85.8%

The R-Sq(adj) is the adjuster R^2 and is Adj-R^2 = (MS(Total) - MS(Residual)) / MS(Total).

Adj-R^2 = (318.85 - 58.7) / 318.85 = 0.816 = 81.6%

R-Squared vs Adjusted R-Squared

There is a problem with the R^2 for multiple regression. Yes, it is still the percent of the total variation that can be explained by the regression equation, but the largest value of R^2 will always occur when all of the predictor

variables are included, even if those predictor variables don't significantly contribute to the model. R^2 will only go down (or stay the same) as variables are removed, but never increase.

The Adjusted-R^2 uses the variances instead of the variations. That means that it takes into consideration the sample size and the number of predictor variables. The value of the adjusted-R^2 can actually increase with fewer variables or smaller sample sizes. You should always look at the adjusted-R^2 when comparing models with different sample sizes or number of predictor variables, not the R^2. If you have a tie for two models that have the same adjusted-R^2, then take the one with the fewer variables as it's a simpler model.

Regression Analysis Repeated

Round 2: Remove a Predictor Variable

Do you remember earlier in this document when it appeared that neither age (p-value = 0.059) or body weight (p-value = 0.530) belonged in the model? Well now it's time to remove some variables.

We don't want to remove all the variables at once, though, because there might be some correlation between the predictor variables, so we'll pick the one that contributes the least to the model. This is the one with the largest p-value, so we'll get rid of body weight first.

Here are the results from Minitab.

Response Variable: clean
Predictor Variables: age, snatch

```
The regression equation is
clean = 42.2 + 1.02 age + 0.857 snatch

Predictor    Coef  SE Coef     T      P
Constant    42.23    23.77  1.78  0.103
age        1.0223   0.4682  2.18  0.052
snatch     0.8571   0.1262  6.79  0.000

S = 7.45893   R-Sq = 85.2%   R-Sq(adj) = 82.6%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  3533.1  1766.5  31.75  0.000
Residual Error  11   612.0    55.6
Total           13  4145.1
```

Notice there are now 2 regression df in the ANOVA because we have two predictor variables. Also notice that the p-value on age is only marginally above the significance level so we may want to use it.

But the thing I want to look at here is the values of R-Sq and R-Sq(adj).

Model	Variables	R-Sq	R-Sq(adj)
1	age, body, snatch	85.8%	81.6%
2	age, snatch	85.2%	82.6%

Notice that the R^2 has gone down but the Adjusted-R^2 has actually gone up from when we included all three variables. That is, we have a better model with only two variables than we did with three. That means that the model is easier to work with since there's not as much information to keep track of or substitute into the equation to make a prediction.

Round 3: Eliminating Another Variable

We said that the p-value for the age was slightly above 0.05, so we could say that age doesn't contribute greatly to the model. Let's throw it out and see how things are affected. At this point, we'll be back to the simple linear regression that we did earlier since we only have one predictor variable.

```
The regression equation is
clean = 54.6 + 0.931 snatch

Predictor    Coef  SE Coef     T      P
Constant    54.61    26.47  2.06  0.061
snatch     0.9313   0.1393  6.69  0.000

S = 8.55032   R-Sq = 78.8%   R-Sq(adj) = 77.1%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        1  3267.8  3267.8  44.70  0.000
Residual Error   12   877.3    73.1
Total            13  4145.1
```

Here is the summary table again

Model	Variables	R-Sq	R-Sq(adj)
1	age, body, snatch	85.8%	81.6%
2	age, snatch	85.2%	82.6%
3	snatch	78.8%	77.1%

Wow! Notice the big drops in both the R^2 and Adjusted-R^2. For that reason, we're going to stick with the two variable model and use a competitor's age and the weight they can snatch to predict how much they can lift in the clean and jerk.

Last modified August 29, 2004 11:21 AM
Return to ICTCM 2004 Short Course page
Return to James Jones homepage