

MATH 143: Introduction to Probability and Statistics

Homework for Multiple Regression (a portion of Problem Set 21)

It is very difficult to weigh bears in the wild, since **they don't willingly step on a scale**. Even if one anesthetizes them, they're a bit heavy to put on a scale. Thus, it would be nice to know of some alternate way of estimating a bear's weight. There is a dataset containing information about a sample of bears. The following variables were measured:

Variable name	Variable description	Units
age	age of bear	months
sex	gender of bear	1=male, 2=female
headlen	length of head	inches
headwid	width of head	inches
neck	neck girth	inches
length	length of bear	inches
chest	chest girth	inches
weight	weight of bear	pounds

We want to determine which combination of factors does the best job of predicting bear weight.

1. Look at the correlation coefficient matrix below. For each combination of variables, there are two numbers. The top number is the sample correlation coefficient (r), and the bottom number is the P -value associated with the null and alternative hypotheses:

H_0 : (true) correlation between these variables is 0

H_a : (true) correlation between these variables is nonzero.

	age	sex	headlen	headwid	neck	length	chest
age	1.0000						
sex	0.1696	1.0000					
headlen	0.2201		1.0000				
headwid	0.7112	-0.1756	0.0000	1.0000			
neck	0.6614	-0.2509	0.7535	0.0000	1.0000		
length	0.7166	-0.2814	0.8848	0.8188	0.0000	1.0000	
chest	0.0000	0.0393	0.0000	0.0000	0.0000	0.8674	1.0000
weight	0.7188	-0.1032	0.9200	0.7354	0.8674	0.0000	0.0000
	0.0000	0.4577	0.0000	0.0000	0.0000	0.8894	0.0000
	0.7125	-0.1678	0.8625	0.7785	0.9348	0.8894	1.0000
	0.0000	0.2251	0.0000	0.0000	0.0000	0.0000	0.0000
	0.7490	-0.1896	0.8342	0.7835	0.9341	0.8644	0.9631
	0.0000	0.1698	0.0000	0.0000	0.0000	0.0000	0.0000

- (a) Which predictors are significantly correlated with *weight*, and therefore might end up in our final model?

- (b) Which predictor is *most highly correlated* with bear weight?
- (c) Multicollinearity is a problem in this data set. Because of all the significant correlations between the predictors, we may not need all the predictors in your answer to part (a) in the final model to predict bear weight. Circle (in the correlation coefficient matrix above) the significant correlations between the *predictors*.

2. At right is the output from multiple regression on the bear dataset produced by StatCrunch (similar to CrunchIt). It includes *all* predictors.

Multiple linear regression results
 Dependent Variable: weight
 Independent Variable(s): neck, age, length, sex, chest, headlen, headwid
Parameter estimates:

Variable	Estimate	Std. Err.	Tstat	P-value
Intercept	-179.55136	42.87694	-4.1875978	0.0001
neck	5.0796175	2.6756816	1.8984387	0.0639
age	0.5608353	0.22175312	2.5290978	0.0149
length	0.7204891	1.1521529	0.62534153	0.5348
sex	-13.392128	11.324815	-1.1825472	0.2431
chest	9.204655	1.4261256	6.4543085	<0.0001
headlen	-9.150646	5.585313	-1.6383408	0.1082
headwid	-0.039021138	4.9576187	-0.007870944	0.9938

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	7	744379.75	106339.97	116.73563	<0.0001
Error	46	41903.562	910.947		
Total	53	786283.3			

Root MSE: 30.181898
 R-squared: 0.9467
 R-squared (adjusted): 0.9386

- (a) What are the null and alternative hypotheses for the *model utility test*?
- (b) What are the test statistic and *P*-value associated with the model utility test here?
- (c) What do you conclude?
- (d) Look at the coefficient of each predictor. Some are positive, some negative.
- i. What does the positive sign by *age* tell us?
 - ii. What does the negative sign by *sex* tell us?
- (e) Circle (on the StatCrunch output) those predictors which are statistically significant.
- (f) Based on the *P*-values, which variable do you think will be added first in the stepwise models? What would be the first predictor *eliminated* in a backwards stepwise process?
- (g) What are the adjusted *R*-squared and standard error of the regression line (Root MSE) values?

3. The raw bear data set is found at <http://www.calvin.edu/~scofield/data/comma/bears.csv>. Start StatCrunch (or CrunchIt) and load this data. (You will notice that there is a column indicating the month in which data was collected for each bear. We did not include this variable in the regression, and you may continue to ignore it.) As a warmup to the problems below, use the software to reproduce the regression output displayed in the previous problem.
- Do a backward stepwise analysis of the model, until the only predictors which remain are significant at the $\alpha = 0.05$ level. What predictors remain? What is the adjusted R -squared value? (If you are using CrunchIt, you'll only be able to give the R -squared value.) Have we lost much in the way of predictive value with this reduced collection of predictors?
 - Are you surprised at any of the predictors that were removed? Are you surprised at any that remain? Explain why or why not, in light of Question 1 (c).
 - Write the final model.
 - What does the model say should be the weight of a bear that has a 5-foot chest, a 2.5-foot neck, and is 3 years old?
4. Repeat the regression that produced your final model, this time asking the software to “save residuals”. You should have noticed by now that **multiple regression** does not offer the same list of optional plots—ones to help check for flagrant violations of the *conditions for inference on regression*—as are offered under **simple regression**. Having saved the residuals, you can still produce some useful plots for checking these assumptions.
- Produce a histogram of residuals and include it with what you hand in. What shape would be consistent with the assumptions for regression inference? Is that what you see?
 - Produce a scatterplot of residuals vs. fitted values (to be handed in). To do so, you must first produce another column that contains the fitted values. Do so using the **Transform data** option off the **Data** menu.

- (c) What do you look to find in the plot from part (b) that would be consistent with the *conditions for regression inference*? Is this, indeed, what you see?