Multiple Regression and Regression Model Adequacy

Joseph J. Luczkovich, PhD

February 14, 2014

Introduction

Regression is a technique to mathematically model the linear association between two or more variables, the predictor (or independent) variable and the response (or dependent variable). Linear regression results in a least-squares fit of the variation in the response variable to the predictor variable. But sometimes, we have multiple possible predictor variables with a single response variable. For example, we might also have measured temperature and fertilizer application rate in addition to the worminess and apples yield. **Multiple linear regression** is a method for adding additional predictor variables in the model. It is analogous to doing a two- or three-way (or even more factors) ANOVA, in which one adds additional factors to be considered as predictors. One key difference between ANOVA and regression is that ANOVA uses **categorical predictors**, and regression does not - it uses **continuous predictors**. **Analysis of covariance** is like a combination of ANOVA and regressions, it uses both categorical and continuous variables as predictors. Both ANOVA and regression typically use continuous response variables (but not always - we will get to logistic regression later). In addition, however, multiple regression is often done using non-experimental designs, that is, data collected are often based on observational procedures, allowing the predictor and response variables to vary naturally, and not under experimental (manipulated) procedures. The linear regression approach to do ANOVA is called General Linear Models, with factors specified as "dummy" variables. For now, we will focus on regression and multiple regression.

Regression's Pitfalls

"Regression analysis is widely used, and unfortunately, it is frequently misused" This is a quote from Montgomery et.al. (2012) [3]. They make a good point - regression can be mis-used, mis-applied, and give the an investigator a wrong impression - just like any statistical procedure. Here are some things to be aware of in Regression and Multiple Regression:

- Extrapolation beyond range of data

- Outliers and Influential Points
- Predictors not evenly distributed - clustering of X's
- Non-linearity of data (do you need transformations?)
- Data are non-normal (use non-parametric regression) -[2]
- Small sample size
- Overfitting of the data - too many predictors (multiple regression)
- Autocorrelated predictors (multiple regression)

In the sections below, I will try to illuminate these pitfalls with examples. But first, let me introduce Multiple Linear Regression.

Multiple Linear Regression

Multiple regression is used when there is a single response variable, but many possible predictors. The goal is to find the model that best predicts the response variable (explains the greatest % of variation) with the fewest predictor variables. Some variables (and combinations of variables) are better predictors than others. So, we need to find a way of determining which possible combination of variables is the best set of predictors.

Basics of Multiple Regression

- Similar to basic linear regression, with more β's (one term for each predictor x variable)
- You can add as many βs as you want, up to k, the number of predictors you have available
- Each β gets its own coefficient in the model output, each gets its own line in the F-table
- Think of each predictor as a factor in an ANOVA (but the df is different. Why?)
- 1 df used for each predictor

The general multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...\beta_k x_k + \epsilon$$

0.0.1 Hypotheses and Assumptions

Remember the words "Never assume - you make an Ass of U and Me". We make assumptions when using any model or statistical procedure, so we always run the risk of making silly statements. Try not to do that by knowing what you are assuming.

- H_a is still that each $\beta \neq 0$, null H_0 is $\beta = 0$
- Assume that predictors are independent (no **multi-collinearity**, uncorrelated)
- Assume that effects of predictors are additive
- Assume that there is a normal distribution of errors underlying the observations
- Assume linear relationship between each x and y, this can be tested (plot it, use correlation, use linear regression)
- Assume Equality of variances across x range
- If you violate any of these, transformations may be needed (log or others)

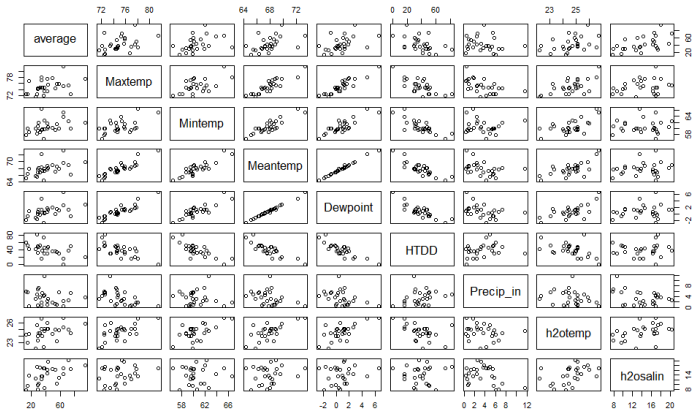
0.0.2 Multiple Regression Example in R

We will use the NC Division of Marine Fisheries (NCDMF) brown shrimp *Farfante penaeus* catch data, obtained from NCDMF Program 120, and NOAA weather data (Cape Hatteras Station[1]). Program 120 data are Collected with trawls at 72 stations in Pamlico Sound and Core Sound. These stations have been sampled with trawls replicated in May and June of each year from 1978-2004 (1 trawl pulled per station, 75m length). The dependent variable is average catch in May and June at each station (number of shrimp/trawl) over 27 years. Catches and other data obtained in each year are the observations, so there are 27 data records in this analysis, one per year. The independent variables we may use as predictors are: May mean air temperature at Cape Hatteras (°F used by NOAA Weather), May dew point at Cape Hatteras, May heating degree days at Cape Hatteras[1], May precipitation at Cape Hatteras, bottom salinity at time of collection, water bottom temperature (°C used by NCDMF) at time of collection. Of course, we could include even more predictors (June weather and climate data, weather data from other stations, fisher harvest rates in the previous year, and predator abundances at each shrimp life stage), but these data are harder to come by and integrate. First ask: "What is the correlation among the predictor variables?"

[1]Monthly heating degree days (HTDD) in May of each year. HTDD are defined as the sum of the differences in ambient temperature from 65 °F for a given period of time. 65 °F was chosen by NOAA as a base temperature at which buildings and homes are commonly maintained. For example, temperatures of 64, 65, 60, 58, 57,63, and 64 °F would have differences of 1,0,5,7,8,2,1. A sum of these = 24 HTDD for that week.

```
> load("C:/Users/luczkovichj/Dropbox/CRM7008/Lecture5/shrimp2.Rdata")
> cor1<-cor(as.matrix(shrimp2[1:27,73:81]))
> write.csv(cor1,file="cor1.csv")
> #this writes a csv file with the output in the local directory
> cor1

             average    Maxtemp     Mintemp   Meantemp   Dewpoint        HTDD
average    1.0000000  0.4043793  0.41252773  0.4599688  0.4646784 -0.46922628
Maxtemp    0.4043793  1.0000000  0.56930319  0.8885641  0.9097669 -0.70898887
Mintemp    0.4125277  0.5693032  1.00000000  0.8828876  0.8510661 -0.77812805
Meantemp   0.4599688  0.8885641  0.88288757  1.0000000  0.9944996 -0.84102417
Dewpoint   0.4646784  0.9097669  0.85106608  0.9944996  1.0000000 -0.83892557
HTDD      -0.4692263 -0.7089889 -0.77812805 -0.8410242 -0.8389256  1.00000000
Precip_in -0.1678119 -0.5272537 -0.09795523 -0.3572977 -0.3910884  0.26893994
h2otemp    0.4013902  0.6080413  0.58842782  0.6727228  0.6652060 -0.48568510
h2osalin   0.5434998  0.2551023  0.03548443  0.1610938  0.1696446  0.03181778
            Precip_in    h2otemp    h2osalin
average   -0.16781194  0.4013902  0.54349982
Maxtemp   -0.52725367  0.6080413  0.25510235
Mintemp   -0.09795523  0.5884278  0.03548443
Meantemp  -0.35729768  0.6727228  0.16109377
Dewpoint  -0.39108841  0.6652060  0.16964460
HTDD       0.26893994 -0.4856851  0.03181778
Precip_in  1.00000000 -0.2206994 -0.37416193
h2otemp   -0.22069936  1.0000000  0.41060420
h2osalin  -0.37416193  0.4106042  1.00000000

> pairs(shrimp2[1:27,73:81])
```

The output is a matrix of Pearson correlations for each variable versus the others. Do you see any highly correlated variables? If so, they are something to consider for multi-collinearity in our model (we will deal with this later). Another way to examine the data from this is to make a scatterplot matrix, with multiple plots of each variable versus the other. Use the **pairs()** command to do this.



This scatterplot matrix can be easily scanned for highly correlated variables. We will want to come back here and look at this plot some more, but we can see that Meantemp and dewpoint are very highly correlated (r = 0.99). Why? Now let's do the multiple regression model, using the linear model function **lm()**:

```
> shrimp2<-data.frame(shrimp2)
> attach(shrimp2)
> fit.shrimp<-lm(shrimp2$average~shrimp2$Meantemp+shrimp2$Dewpoint
+ +shrimp2$HTDD+shrimp2$Precip_in+shrimp2$h2otemp+shrimp2$h2osalin)
> summary(fit.shrimp)

Call:
lm(formula = shrimp2$average ~ shrimp2$Meantemp + shrimp2$Dewpoint +
    shrimp2$HTDD + shrimp2$Precip_in + shrimp2$h2otemp + shrimp2$h2osalin)

Residuals:
    Min      1Q  Median      3Q     Max
-17.548 -10.376  -1.890   7.823  26.207

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       528.1059   965.3394   0.547 0.590385
shrimp2$Meantemp   -6.8963    14.5333  -0.475 0.640271
shrimp2$Dewpoint    7.2862    14.9662   0.487 0.631659
shrimp2$HTDD       -0.6768     0.3101  -2.182 0.041175 *
shrimp2$Precip_in   1.9453     1.3153   1.479 0.154727
shrimp2$h2otemp    -2.3889     3.7813  -0.632 0.534685
shrimp2$h2osalin    3.8467     0.9745   3.948 0.000795 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 14.26 on 20 degrees of freedom
Multiple R-squared:  0.5866,        Adjusted R-squared:  0.4626
F-statistic:  4.73 on 6 and 20 DF,  p-value: 0.00376
```

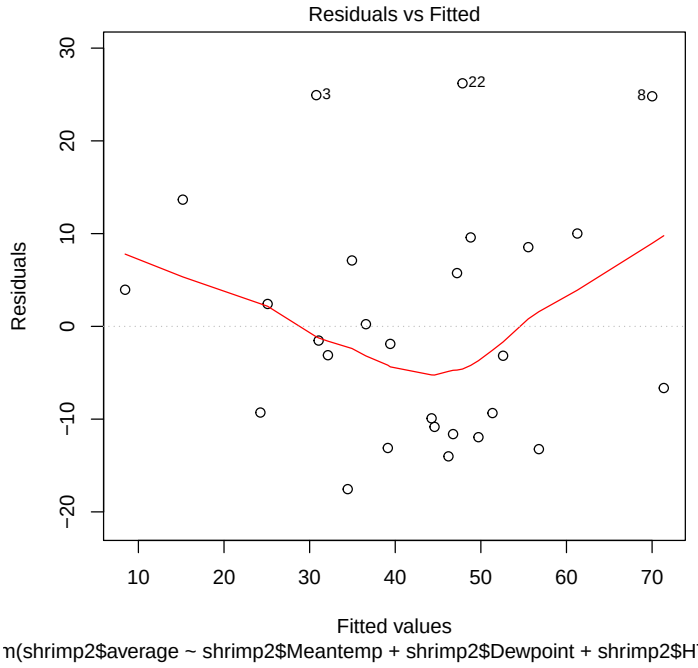
Let us interpret this output. First, after a listing of the model, a list of the residuals summary statistics is printed, giving the minimum, first quartile, median, 3rd quartile, and maximum of the residuals. The Min and Max are the largest and smallest residual values that are from the fitted line. More on this later. Next is what is analogous to an F-table, called "Coefficients". Each line of the table has a β term from the model, with its associated Estimate (β coefficient), standard error of the estimate, t-value, and p-value of getting a lower or higher t. The statistically significant predictors are given *'s to show which are largest in t-values. Here we see that HTDD and h2osalin are significant, salinity much more so than heating degree days. Recall that HTDD is highly negatively correlated with air temperature, so HTDD may be just a surrogate for temperature. At the bottom of the output, notice the global measures of this model's fit, the multiple R^2 and adjusted R^2, the overall F-statistic, df, and p-value for the model. These provide an estimate of the goodness of fit of the multiple regression model. Adjusted R^2 is a way of comparing models with different numbers of predictors. It is adjusted to account for the number of variables used. Additional predictor variables will always explain more variation, so Multiple R^2 will always increase, but sometimes not by much. Adjusted R^2 should be used to compare after adding or dropping a variable, sometimes it will go down if you add a variable, or go up if you get rid of a variable.

Here is the regression equation for this model, predicting average shrimp catch, y:

$$y = 528.106 + (-6.896)(meantemp) + (7.2862)(Dewpoint)$$
$$+(0.6768)(HTDD) + (1.9453)(Precip.in) + (-2.3889)(h2otemp)$$
$$+(3.8467)(h2osalin)$$

The output of a **lm()** command is called a **model output object** (called fit.shrimp here). It stores a lot of the information in a R-object that can be re-used, plotted or summarized. The name of the variables in the object (such as beta coefficients, residuals, fitted values) that can be re-used in plots and further analyses are given by a **names()** command:

```
> names(fit.shrimp)

 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

> attach(fit.shrimp)

> plot(fit.shrimp)
```



ת(shrimp2$average ~ shrimp2$Meantemp + shrimp2$Dewpoint + shrimp2$HTDD

Here we just plotted the model output object with **plot(fit.shrimp)**, and it gave us pre-programmed diagnostic plots. The first plot you will see is printed above is called a **residual plot**, with the **fitted or predicted y values** (so-called $\hat{y}$) on the horizontal axis and the **residuals** (e_i) on the vertical axis. There should be no observable pattern in your residual plot (it should look like a shotgun blast). There is a dashed line that is horizontal at $e_i = 0$. That line is the multiple regression line with all six predictors. The points are the fitted or model predictions of y values, here the shrimp catch. Any points not falling on this imaginary horizontal line are positive and negative residuals and are thus shrimp catches that are not well explained by this model. It plots the residuals (points falling off of the multiple regression line) versus the fitted values or predicted y values, $\hat{y}$ shows that there are some extreme values, records 3, 8 and 22 have large residuals. These correspond to years 1980, 1985, and 1999. Some thing happened in those years to make the model a bad predictor of shrimp catch. This is a default plot for a regression model, and useful in determining **model adequacy**. Better fitting models will have small residuals, close to the 0 line, with some random variation around the line. The red curve in the plot is a fit of a line to the residuals. It should be horizontal and un-curved, approximating the residual = 0 line. It may show a slope (high or low residuals at high $\hat{y}$) or could be funnel-shaped, increasing in variability as $\hat{y}$ increases, "U" or parabola-shaped (high residuals at low and high $\hat{y}$'s, with negative residuals at intermediate $\hat{y}$'s). If you keep hitting return, while doing this interactively,

you will get Normal QQ plots, Scale-location plots, and residuals vs. leverage plots (see explanations below). Type help(plot.lm) to read about these plots.

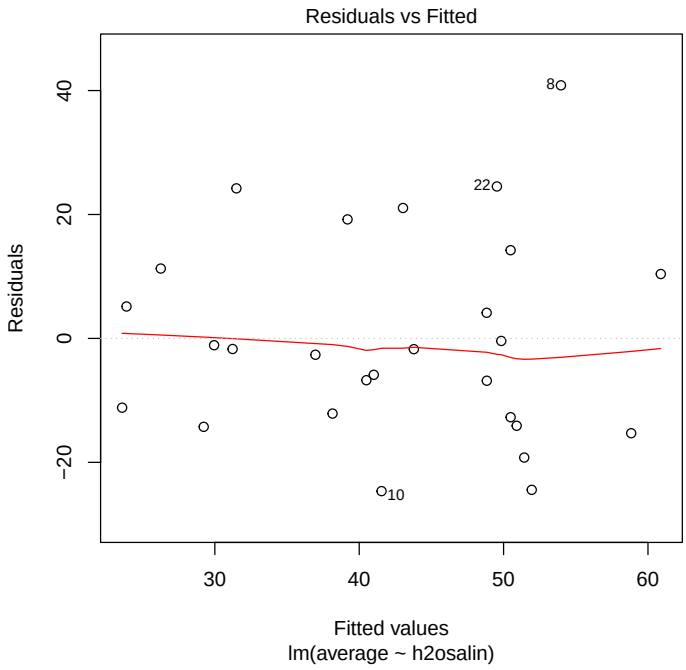
Let's run this regression again on just h2osalin, a single predictor variable (simple linear regression). How will the results differ?

```
> fit.shrimp2<-lm(average~h2osalin)
> summary(fit.shrimp2)

Call:
lm(formula = average ~ h2osalin)

Residuals:
    Min      1Q  Median      3Q     Max
-24.646 -12.424  -1.734  10.834  40.848

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7240    13.2773   0.055  0.95695
h2osalin      2.9488     0.9108   3.237  0.00339 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.66 on 25 degrees of freedom
Multiple R-squared:  0.2954,        Adjusted R-squared:  0.2672
F-statistic: 10.48 on 1 and 25 DF,  p-value: 0.00339
```

The adjusted R^2 value went down! This is because we have eliminated other explanatory variables, each of which explains some of the variation. But the predictor variable h2osalin is still significant, as it was before with all six predictors. Note that the t-value is lower and p-value is higher with just one (albeit important) explanatory variable. Adding more variables back to the model will improve this. But which predcitors one should we add? More on this later.

Let's run the residual plot again on this simple single-predictor model, this time getting all the possible plots in the plot.lm arsenal:

```
> plot(fit.shrimp2, which = c(1:5))
```



You will see only the first one on this printed page, but if you try it on your own, you will see five plots. They are:

- residuals versus fitted plot
- Normal QQ plot
- Scale-location plot - $\sqrt{e_i}$ versus fitted values $\hat{y}$ this is similar to the first plot, but the square root of the residuals is plotted
- Cook's distance for each observation - another kind of residual, measuring influence of an observation, which shows the impact on the least squares fit of the β's after deleting that observation from the model. The bigger the Cook's distance, the bigger the influence of that observation on each β
- standardized residuals (normalized with mean $e_i = 0$, sd $= 1$) versus leverage - leverage is another measure of influence of each point when it is deleted.

Each one of these diagnostic plots are useful for looking at the overall fit, detecting outliers and influential points. We will come back to use these plots as we try different models, in order to assess their adequacy.

Model Adequacy Measures

But how do we know if the model we have created is any good? How can we assess the model's adequacy? Can we explain as much variation with fewer predictors? If so, we have less to measure to make good predictions. We could remove each predictor variable one at a time. We could only include predictors with significant t-values in the full model. Or we could add predictors one by one see where non-significant ones appear. There are automated methods to do this (stepwise regression). We will try these methods next.

Some common methods to assess model adequacy include:

- Examination of R^2
- Examination of Residuals
- Examination of Outliers and Influential Points
- Examination of Akaike's Information Criterion (AIC)

We will explore these in the next section...

References

[1] Noaa cape hatteras monthly climate data, http://www.ncdc.noaa.gov/cdo-web/datasets/ghcndms/stations/ghcnd:usw00093729.

[2] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

[3] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.