Regression Example

Dave Lorenz

August 14, 2013

This examples demonstrates a couple of linear regression functions in the `USGSwsStats` package. The example uses the Haan (1977) dataset C14 from Helsel and Hirsch (2002) to replicate the analysis done in section 11.6 in their book. Please note that there are only 13 observations in the dataset and most practitioners would prefer more observations for this kind of multiple regression analysis; Harrell (2001) has some very good guidance in chapter 4 of his book.

```
> # Load the USGSwsStats package
> library(USGSwsStats)
> # Create the dataset
> Haan1977 <- data.frame(
+ ROFF=c(17.38, 14.62, 15.48, 14.72, 18.37, 17.01, 18.2, 18.95, 13.94, 18.64,
+  17.25, 17.48, 13.16),
+ PCIP=c(44.37, 44.09, 41.25, 45.5, 46.09, 49.12, 44.03, 48.71, 44.43, 47.72,
+  48.38, 49, 47.03),
+ AREA=c(2.21, 2.53, 5.63, 1.55, 5.15, 2.14, 5.34, 7.47, 2.1, 3.89, 0.67,
+  0.85, 1.72),
+ SLOPE=c(50, 7, 19, 6, 16, 26, 7, 11, 5, 18, 21, 23, 5),
+ LEN=c(2.38, 2.55, 3.11, 1.84, 4.14, 1.92, 4.73, 4.24, 2, 2.1, 1.15, 1.27,
+  1.93),
+ PERIM=c(7.93, 7.65, 11.61, 5.31, 11.35, 5.89, 12.59, 12.33, 6.81, 9.87,
+  3.93, 3.79, 5.19),
+ DI=c(0.91, 1.23, 2.11, 0.94, 1.63, 1.41, 1.3, 2.35, 1.19, 1.65, 0.62, 0.83,
+  0.99),
+ Rs=c(0.38, 0.48, 0.57, 0.49, 0.39, 0.71, 0.27, 0.52, 0.53, 0.6, 0.48, 0.61,
+  0.52),
+ FREQ=c(1.36, 2.37, 2.31, 3.87, 3.3, 1.87, 0.94, 1.2, 4.76, 3.08, 2.99,
+  3.53, 2.33),
+ Rr=c(332, 55, 77, 68, 68, 230, 44, 72, 40, 115, 352, 300, 39)
+ )
```

1 Subset Selection

The `allReg` function creates a data frame that contains candidate models and selection criteria. Note that the name of the response variable is taken from the column name for the `y` argument if it is rectangular, and from the argument name if it is not. Two general approaches for specifying the `x` and `y` arguments. The first uses the `with` function and is most useful when a subset of columns are explanatory variables, but it is probably more clear in identifying the variables. The second, which is commented out, requires less typing and can be useful when most columns are explanatory variables as in this case.

```
> # Create the allReg output dataset
> HaanSub <- with(Haan1977, allReg(cbind(PCIP, AREA, SLOPE, LEN, PERIM,
+                 DI, Rs, FREQ, Rr), ROFF))
> # An alternative call, note the use of the drop argument
> HaanSub <- allReg(Haan1977[, -1], Haan1977[, 1, drop=FALSE])
> # What are the "best" 5 models by Cp
> head(HaanSub[order(HaanSub$Cp),])
```

```
                            model.formula nvars    stderr       R2    adjr2
13    ROFF ~ PCIP + PERIM + DI + FREQ + Rr     5 0.5157295 95.85716 92.89799
10           ROFF ~ PCIP + PERIM + DI + Rr     4 0.6201892 93.15309 89.72964
11         ROFF ~ PCIP + AREA + PERIM + Rr     4 0.6271707 92.99807 89.49710
12         ROFF ~ PCIP + PERIM + FREQ + Rr     4 0.6418703 92.66600 88.99900
7                ROFF ~ PCIP + PERIM + Rr     3 0.6880745 90.51866 87.35822
14 ROFF ~ PCIP + AREA + SLOPE + PERIM + Rr     5 0.5853334 94.66345 90.85162
         Cp    press
13 2.895526 6.907075
10 3.438170 7.956472
11 3.583937 7.665134
12 3.896182 9.514826
7  3.915331 9.968029
14 4.017979 7.249822
```

Helsel and Hirsch (2002) state "Based on Cp, the best model would be the 5 variable model having PCIP, PERIM, DI, FREQ and Rr as explanatory variables—the same model as selected by `allReg`. Remember that there is no guarantee that stepwise procedures regularly select the lowest Cp or PRESS models. The advantage of using an overall statistic like Cp is that options are given to the scientist to select what is best. If the scientist decided AREA must be in the model, the lowest CP model containing AREA (the same four-variable model) could be selected. Cp and PRESS allow model choice to be based on multiple criteria such as prediction quality (PRESS), low VIF, cost, etc."

To select a good model, Helsel and Hirsch (2002) describe several criteria in section 11.7. Those criteria need to be considered in addition to the assumptions of linear regression (section 9.1.1) and regression diagnostics (sections 9.5 and 11.5).

The output from `allReg` can be used to evaluate any of the selected models, by using the `as.formula` function on the contents of the `model.formula` column as in the following example.

```
lm(as.formula(HaanSub[13, "model.formula"]), data=Haan1977)
```

2 Model Diagnostics

The `multReg` function is designed to assist the user by performing many of the model diagnostic tests and plots suggested by Helsel and Hirsch (2002). This section will discuss the use of the `multReg` function to perform the selected model diagnostics and set up diagnostic plots for the model that Haan (1977) used.

The code below specifies the model, created the regression model and prints the diagnostic tests.

```
> # Create the regression model
> Haan.lm <- lm(ROFF ~ PCIP + PERIM + Rr, data=Haan1977)
> # Create the diagnostic object and print it.
> Haan.reg <- multReg(Haan.lm)
> print(Haan.reg)

Call:
lm(formula = ROFF ~ PCIP + PERIM + Rr, data = Haan1977)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0050 -0.4747  0.0203  0.5085  0.8361

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.64420    4.44080   -2.17  0.05795
PCIP         0.42957    0.09302    4.62  0.00126
PERIM        0.61658    0.07485    8.24  1.8e-05
Rr           0.01042    0.00201    5.19  0.00057

Residual standard error: 0.688 on 9 degrees of freedom
Multiple R-squared:  0.905,        Adjusted R-squared:  0.874
F-statistic: 28.6 on 3 and 9 DF,  p-value: 6.18e-05

press: 9.97
  AIC: 32.4
  BIC: 35.2

Anova Table (Type II tests)

Response: ROFF
          Sum Sq Df F value  Pr(>F)
PCIP        10.1  1    21.3 0.00126
PERIM       32.1  1    67.9 1.8e-05
Rr          12.8  1    26.9 0.00057
Residuals    4.3  9
```

```
Variance inflation factors
    PCIP    PERIM       Rr
1.297794 1.455248 1.460949

Test criteria
leverage   cooksD    dfits
   0.923    0.939    1.109
        Observations exceeding at least one test criterion
   ROFF yhat resids stnd.res stud.res leverage cooksD dfits
8  18.9 19.6 -0.683     -1.4    -1.50    0.500  0.492 -1.50
13 13.2 14.2 -1.005     -1.8    -2.12    0.342  0.421 -1.53
```

The printed results are comprised of several sections. The first section is the regression summary, consisting of the call, residual statistics, the coefficient table (without the significance stars), and statistics of the overall fit; the next section is the analysis of variance (ANOVA) table, which is most useful for assessing the overall significance of complex terms such as first- and second-order polynomials (`quadratic`) or sine and cosine transforms (`fourier`); the third section is a listing of the variance inflation factors (VIFs); and the last section shows the selected test criteria and the observations that exceed at least one of those criteria.

The following sections highlight selected diagnostic plots. When using the `plot` function in an interactive session, it is not necessary to specify which plot to create nor to set up a graphics device. The only call that would be necessary would be `plot(Haan.reg)`. Note that plot number 4 cannot be shown because it describes serial correlation and these data are not collected at specific points in time.

3 Response vs. Fitted

The first diagnostic plot is response vs. fitted. The second is residuals vs fitted and is not shown. The basic difference is that the deviation shown by the smoothed line is exaggerated in the second plot! Each observation is plotted, the dashed line is the 1:1 fit and the solid line is a loess smooth (function `loess.smooth`) using the "symmetric" option for the `family` argument. The regression equation with the residual standard error.

```
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("regplot01", 5, 5)
> plot(Haan.reg, which=1, set.up=FALSE)
> # Required call to close PDF output graphics
>
> graphics.off()
```

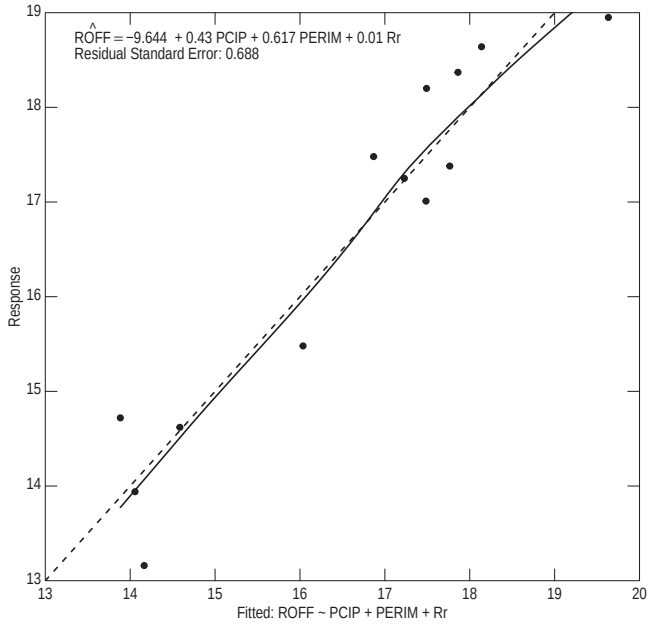

Figure 1. The residual vs. fitted diagnostic plot.

4 Scale-Location Plot

The third diagnostic plot is the scale-location plot, which plots the square root of the residuals vs. the fitted values. It is useful for diagnosing heteroscedasticity and is described by Cleveland (1993). Each observation is plotted, the dashed line is the theoretical mean, assuming a normal distribution, and the solid line is a loess smooth (function `loess.smooth` using the "symmetric" option for the `family` argument. Wooding's test for heteroscedasticity is also shown—it is a straightforward interpretation of the data, simply the results of the Spearman correlation of the data that are shown. The null hypothesis is that the residuals are homoscedastic.

```
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("regplot02", 5, 5)
> plot(Haan.reg, which=3, set.up=FALSE)
> # Required call to close PDF output graphics
> graphics.off()
```

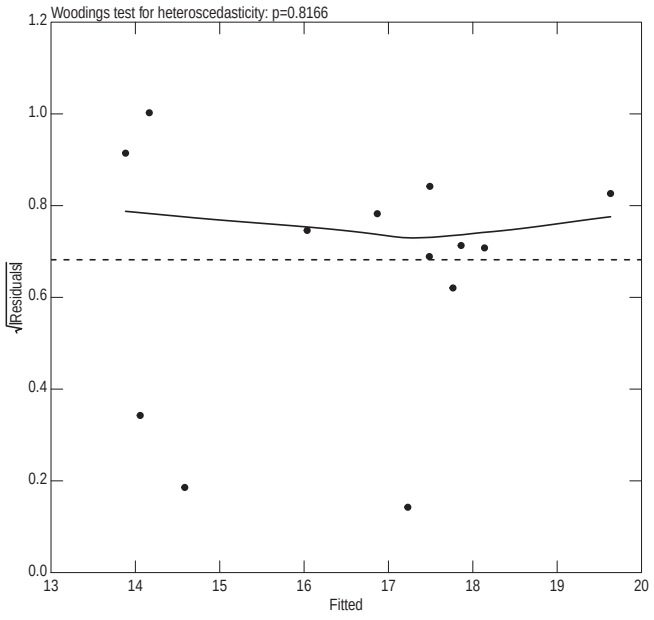

Figure 2. The scale-location diagnostic plot.

5 Probability Plot

The fifth diagnostic plot tests for the normality of the residuals. Each observation is plotted, the solid line is the theoretical fit, assuming a normal distribution. The PPCC test for normality is also shown. The null hypothesis is that the residuals are from a normal distribution.

```
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("regplot03", 5, 5)
> plot(Haan.reg, which=5, set.up=FALSE)
> # Required call to close PDF output graphics
> graphics.off()
```

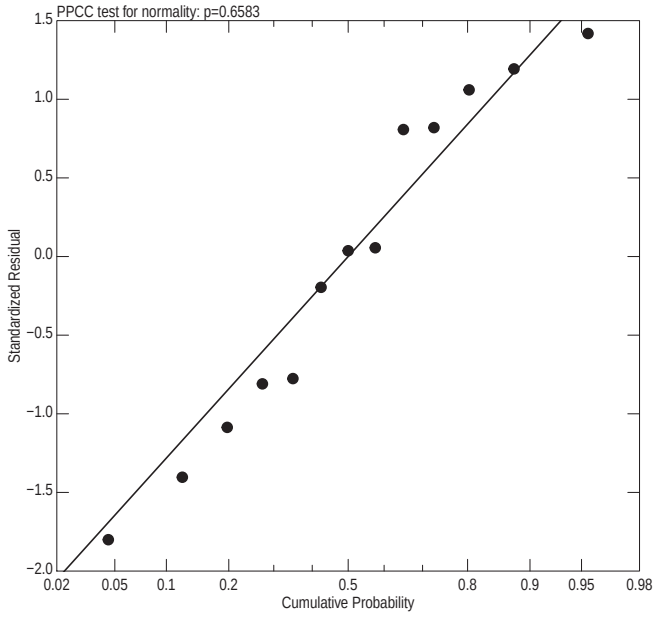

Figure 3. The normal probability diagnostic plot.

6 Influence Plot

The sixth diagnostic plot shows the approximate influence of each observation identified as exceeding one of the test criteria. Each observation is plotted, the solid line is the actual fit. Each identified observation is plotted in a different color and the fitted line with that observation removed is plotted in the same color. The seventh diagnostic plot is a plot of the studentized residual vs. the fitted value and is not shown in this vignette. Note that the label for observation number 8 is not shown in this example because it would be outside the range of the plot area.

```
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("regplot04", 5, 5)
> plot(Haan.reg, which=6, set.up=FALSE)
> # Required call to close PDF output graphics
> graphics.off()
```

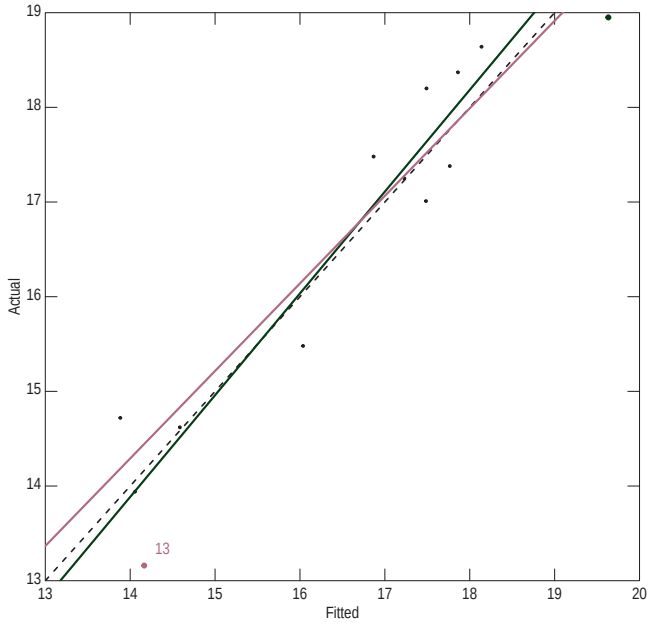

Figure 4. The influence diagnostic plot.

7 Residual Dependence Plot

The eighth diagnostic plot is actually a series of plots, one for each explanatory variable. But, a single explanatory variable can be selected instead of the series, as is shown in this example. Each observation is plotted, the dashed line is 0, the expected value of the residual for each observation and the solid line is a loess smooth (function `loess.smooth` using the "symmetric" option for the `family` argument. The results for a second order polynomial fit is also shown; it is the attained p-value of the squared explanatory variable added to the model.

```
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("regplot05", 5, 5)
> plot(Haan.reg, which="PERIM", set.up=FALSE)
> # Required call to close PDF output graphics
> graphics.off()
```

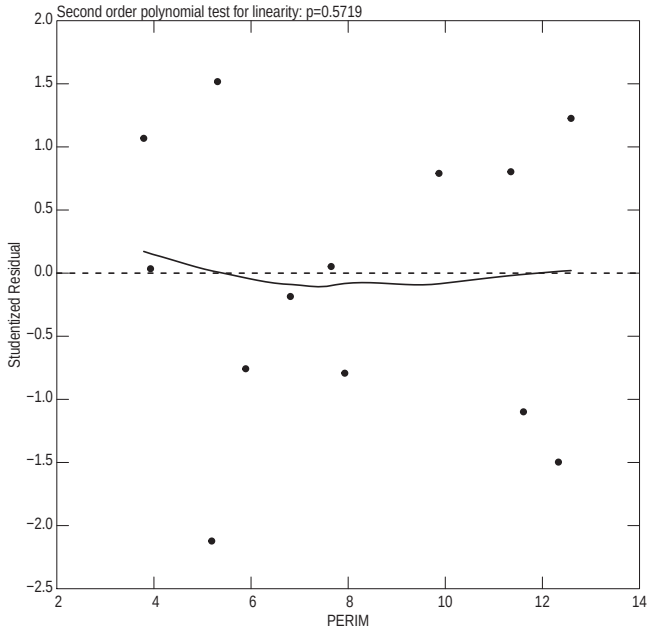

Figure 5. The residual dependence diagnostic plot.

References

[1] Cleveland, W.S., 1993, Visualizing data: Summit, New Jersey, Hobart Press, 360 p.

[2] Haan, C. T., 1977. Statistical methods in hydrology: Iowa State University Press, Ames, Iowa, 378 p.

[3] Harrell, F.E., Jr., 2001, Regression modeling strategies with applications to linear models, logistic regression and survival analysis: New York, N.Y., Springer, 568 p.

[4] Helsel, D.R. and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.