# ANCOVA Example

Dave Lorenz

August 14, 2013

This examples demonstrates the `ancovaReg` function in the `USGSwsStats` package. The example uses the UraniumTDS dataset C16 from Helsel and Hirsch (2002). The dataset is available from the `USGSwsData` pacakge.

```
> # Load the USGSwsStats and USGSwsData packages
> library(USGSwsStats)
> library(USGSwsData)
> # Get the dataset
> data(UraniumTDS)
> head(UraniumTDS)

      TDS Uranium  HCO3
1  682.65  0.9315 < 50%
2  819.12  1.9380 < 50%
3  303.76  0.2919 < 50%
4 1151.40 11.9042 < 50%
5  582.42  1.5674 < 50%
6 1043.39  2.0623 < 50%
```

# 1 Build the Model

The `ancovaReg` function performs a series of diagnostic tests on an ANCOVA regression model created by `lm`. It can also find the best subset of explanatory variables and does so by default.

This vignette assumes that best relation between Uranium and TDS uses the log transform for both variables. That was established prior to setting up the regression.

```
> # Create the ANCOVA model
> UTDS.anc <- lm(log(Uranium) ~ HCO3*log(TDS), data=UraniumTDS)
> # Perform the diagnostics after seleting the "best" subset
> # The trace can be instructive, but is not necessary.
> UTDS.best <- ancovaReg(UTDS.anc, trace=TRUE)

Stepwise elimintation of terms from original model
Start:  AIC=-23.14
log(Uranium) ~ HCO3 * log(TDS)

                Df Sum of Sq    RSS     AIC
- HCO3:log(TDS)  1   0.17919 21.861 -24.778
<none>                       21.681 -23.140

Step:  AIC=-24.78
log(Uranium) ~ HCO3 + log(TDS)

           Df Sum of Sq    RSS     AIC
<none>                  21.861 -24.778
- log(TDS)  1    20.171 42.031   1.986
- HCO3      1    26.176 48.037   7.862
```

There is only 1 step in the selection process for this model. The AIC of the full model is -23.14. Step 1 indicates that by dropping the interaction term, the AIC can be reduced to -24.78. Removal of any other terms results in an increase in AIC.

## 2 Model Diagnostics

The `ancovaReg` function is designed to assist the user by performing many of the model diagnostic tests and plots suggested by Helsel and Hirsch (2002) for any regression. The output is similar to the `multReg` function described in the `regression` vignette.

```
> # Print the final ANCOVA model
> print(UTDS.best)

Original model
Anova Table (Type II tests)

Response: log(Uranium)
              Sum Sq Df F value  Pr(>F)
HCO3           26.18  1   48.29 2.2e-08
log(TDS)       20.17  1   37.21 3.4e-07
HCO3:log(TDS)   0.18  1    0.33    0.57
Residuals      21.68 40



Final model

Call:
lm(formula = log(Uranium) ~ HCO3 + log(TDS), data = UraniumTDS)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6675 -0.3483  0.0532  0.4993  1.4656

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.807      2.355   -5.86  6.8e-07
HCO3> 50%      2.575      0.367    7.01  1.6e-08
log(TDS)       2.157      0.351    6.15  2.6e-07

Residual standard error: 0.73 on 41 degrees of freedom
Multiple R-squared:  0.549,        Adjusted R-squared:  0.527
F-statistic:   25 on 2 and 41 DF,  p-value: 8.13e-08


Variance inflation factors
HCO3> 50%  log(TDS)
  2.78055   2.78055

Test criteria
```

```
leverage   cooksD    dfits
   0.205    0.853    0.522
        Observations exceeding at least one test criterion
   log.Uranium   yhat resids stnd.res stud.res leverage cooksD  dfits
3       -1.231 -1.478  0.247    0.395    0.391   0.2679  0.019  0.236
18      -1.915 -0.293 -1.622   -2.325   -2.465   0.0876  0.173 -0.764
21      -0.275  1.392 -1.667   -2.369   -2.519   0.0708  0.143 -0.695
```

The printed results are comprised of several sections. The first section is the type II sum of squares ANOVA table of the original (full) model. The type II sum of squares reports the marginal significance of each term—that is, it computes the sum of squares for the model excluding that term and all higher-order terms that include that term. For this model, it is clear that the interaction term (change in slope in the relation between Uranium and TDS) does not contribute to the overall fit. The second section is the summary of the final model that shows the coefficients for each variable. For cases where there are more than 2 groups in the factor variable, the coefficients can be useful in assessing their individual effect. The third section consists of the variance inflation factors. For this example, they are both very reasonable. The last section lists any observations that might be stronglu affecting the regression. One observation (3) has large leverage, but none have strong influnce based on Cook's D.

The following sections highlights 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(UTDS.best)`. 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. the partial fitted. The partial fitted drops the effect of the grouping variable so that the user can see its effect. The second is residuals vs fitted and shows the overall fit of the regression.

```
> # 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("ancplot01", 5, 5)
> plot(UTDS.best, which=1, set.up=FALSE)
> # Required call to close PDF output graphics
>
> graphics.off()
```

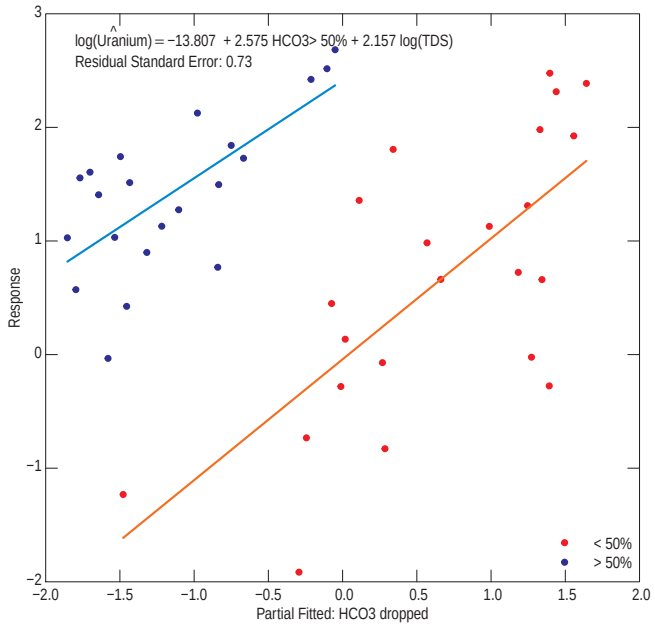


**Figure 1.** The response vs. partial fitted diagnostic plot.

```
> # 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("ancplot02", 5, 5)
```

```
> plot(UTDS.best, which=2, set.up=FALSE)
> # Required call to close PDF output graphics
>
> graphics.off()
```

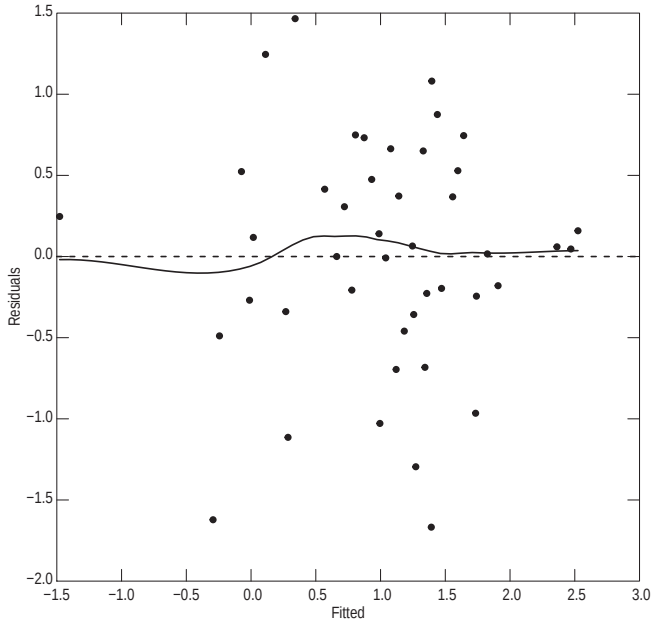


**Figure 2.** The residual vs. fitted diagnostic plot.

# 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("ancplot03", 5, 5)
> plot(UTDS.best, which=3, set.up=FALSE)
> # Required call to close PDF output graphics
> graphics.off()
```

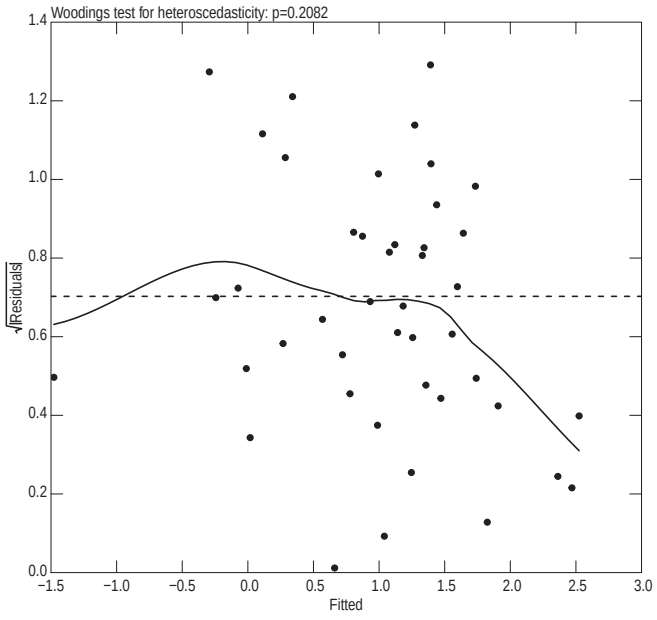


**Figure 3.** The scale-location diagnostic plot.

# 5 Probability Plot

The fifth diagnostic plot is a series of 2 plots, the first is a probability plot and test for the normality of the residuals, the second is a boxplot that shows the distribution for each value of the grouping variable. For the vignette, only the box plot is shown.

```
> # 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("ancplot04", 5, 5)
> plot(UTDS.best, which=5, set.up=FALSE)
> # Required call to close PDF output graphics
> graphics.off()
```

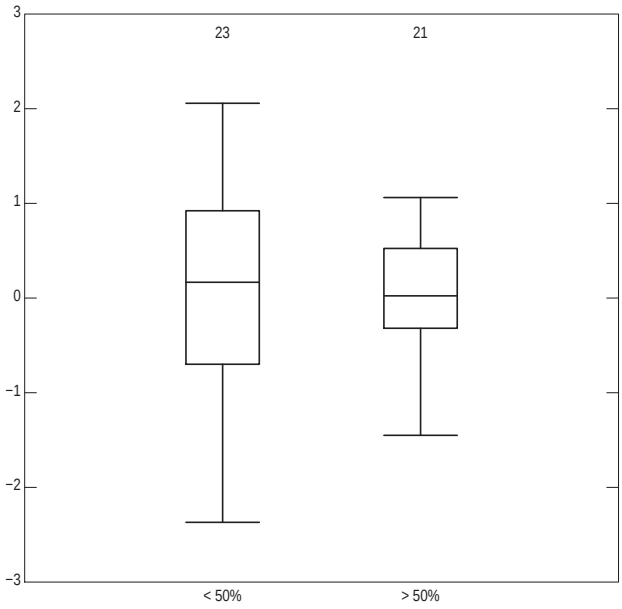


**Figure 4.** The box plots of the residuals for each value in the group.

# 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("ancplot05", 5, 5)
> plot(UTDS.best, which=6, set.up=FALSE)
> # Required call to close PDF output graphics
> graphics.off()
```

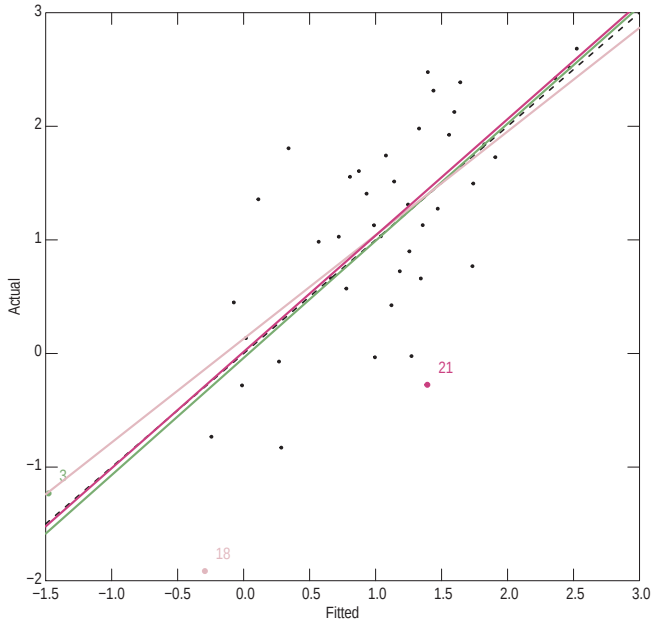


**Figure 5.** The influence diagnostic plot.

# References

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

[2] 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.