



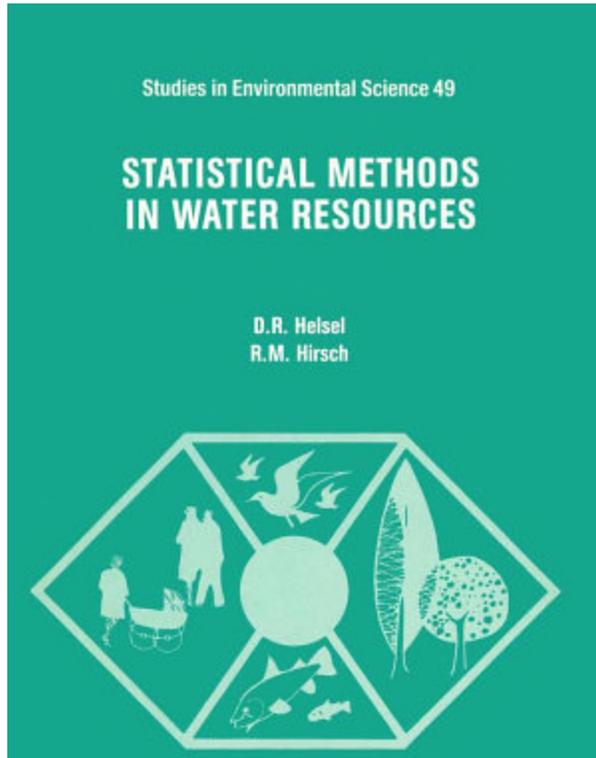
Statistical Methods in Water Resources: New Edition, New Methods, Now Using R

10th National Monitoring Conference

Karen R. Ryberg¹, Robert M. Hirsch¹, Dennis Helsel²,
Ed Gilroy², and Stacey Archfield¹

¹U.S. Geological Survey, ²Practical Stats

Statistical Methods in Water Resources

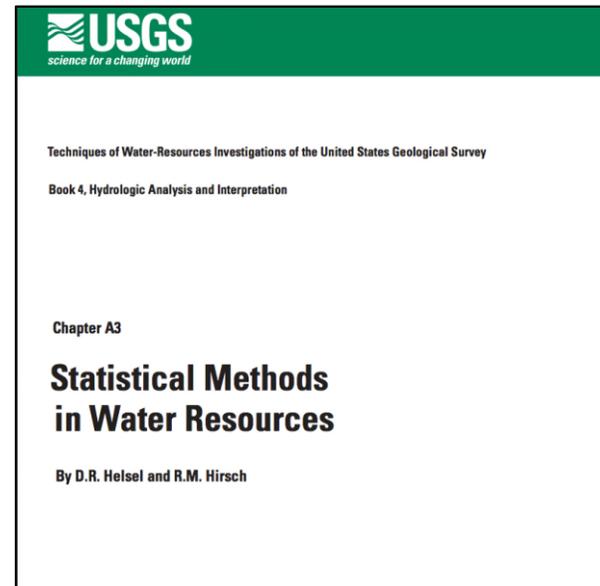


[BOOK] **Statistical methods in water resources**

DR Helsel, [RM Hirsch](#) - 1992 - [books.google.com](#)

Data on water quality and other environmental issues are being collected at an ever-increasing rate. In the past, however, the techniques used by scientists to interpret this data have not progressed as quickly. This is a book of modern statistical methods for analysis ...

[Cited by 2564](#) [Related articles](#) [All 9 versions](#) [Cite](#) [Save](#) [More](#)



[BOOK] **Statistical methods in water resources**

DR Helsel, [RM Hirsch](#) - 2002 - [cala.ca](#)

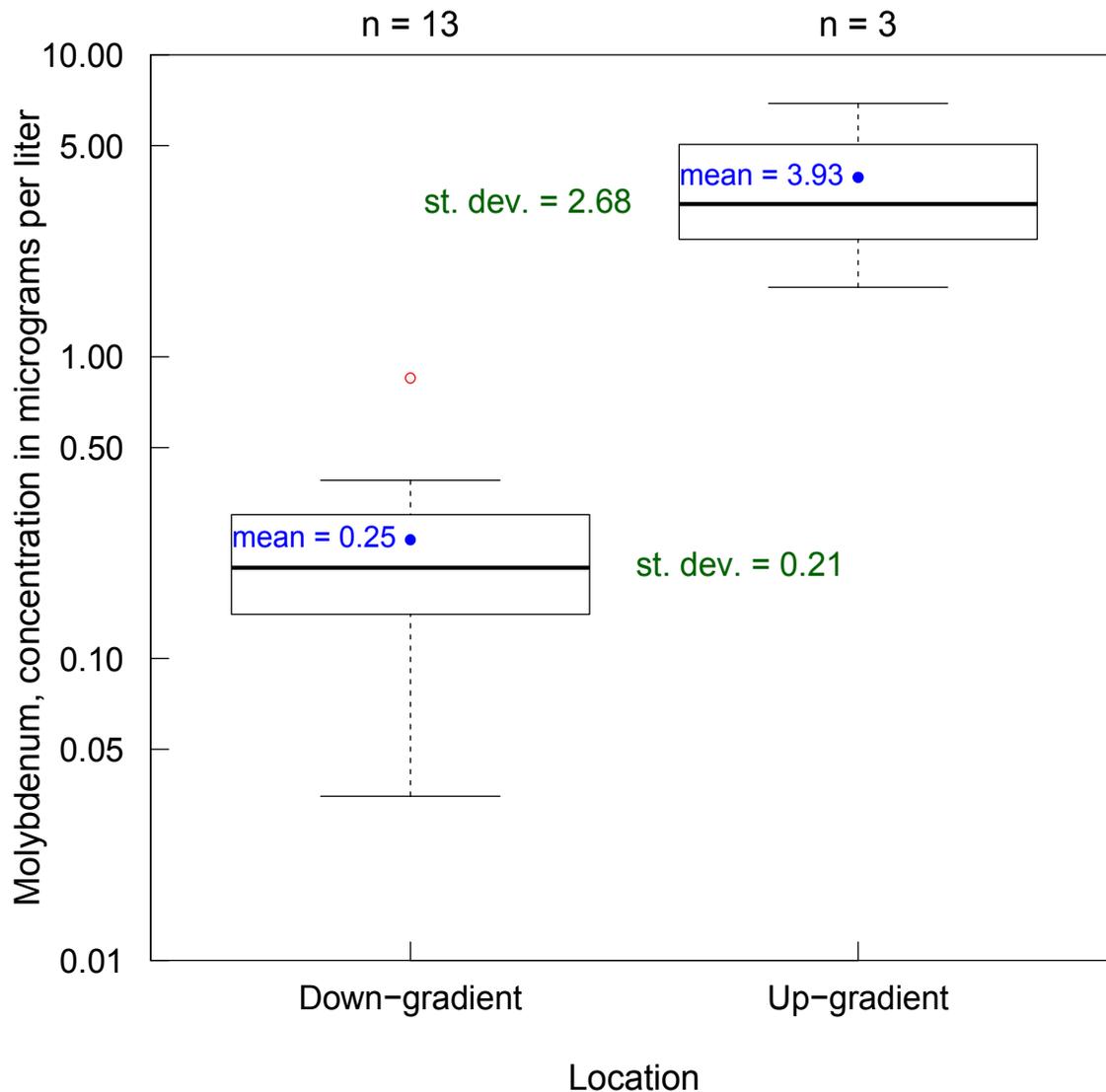
This book began as class notes for a course we teach on applied statistical methods to hydrologists of the Water Resources Division, US Geological Survey (USGS). It reflects our attempts to teach statistical methods which are appropriate for analysis of water resources ...

[Cited by 1135](#) [Related articles](#) [All 9 versions](#) [Cite](#) [Save](#) [More](#)

Almost 25 Years Later, It's Time for an Update!

- Additional techniques and examples using R.
- Computer-intensive methods (bootstrapping and permutation tests) now improve upon & replace past dependence *t*-tests & ANOVA.
- New chapter on sampling design addresses questions such as “How many observations do I need?”
- Trends chapter updated to include the WRTDS (Weighted Regressions on Time, Discharge, and Season) method.
- Much of the original content remains, but with updated graphics & updated guidance on the use of statistical techniques (including some discussion of *p*-values).
- R code used to generate figures & examples provided for download.

Permutation Tests Have Large Advantages in Power Over Normal-Theory Tests



t-test results in $p=0.14$;
means not significantly
different.

Permutation test $p=0.005$
for the same data;
means are significantly
different.

No assumption of
normality required.

Use the `permTS` function
in the `perm` package as
the 'go-to' two-sample test
for difference in means.

Bootstrapping

- The bootstrap is a resampling method which can be used for statistical inference in place of standard methods, such as those that require distributional assumptions.
- It is a computer intensive method; therefore, not included in the original Statistical Methods in Water Resources text.

Bootstrapping

- The upper one-sided confidence bound on the mean (UCL95) is used in many regulatory situations.
- Normal-theory t -interval is unrealistic for skewed data, especially with small sample sizes.
- Transforming with logs produces a UCL on the median (geometric mean), not the mean.
- Instead, bootstrap the interval (the **boot** package). No distributional assumptions made.

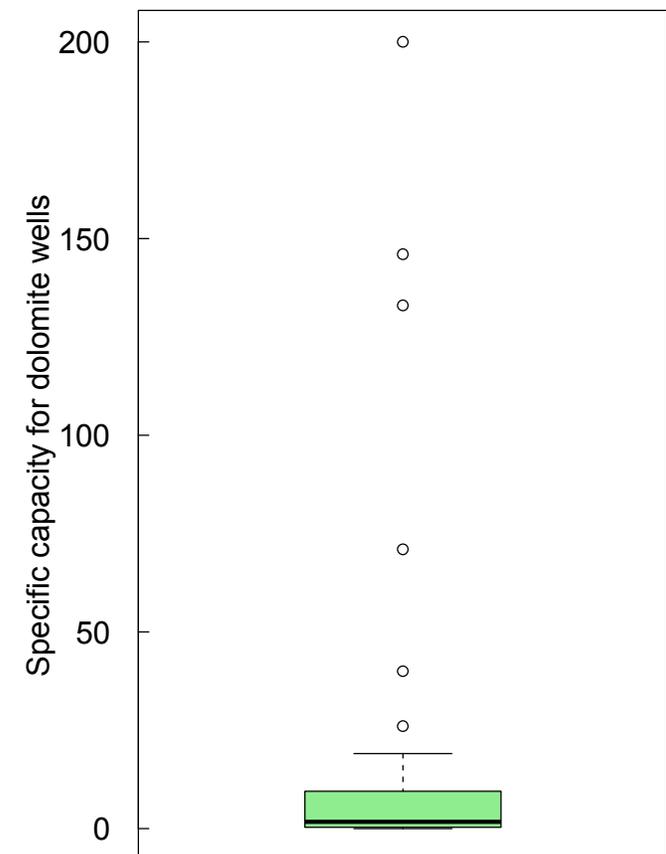
UCL95

t -interval

bootstrap

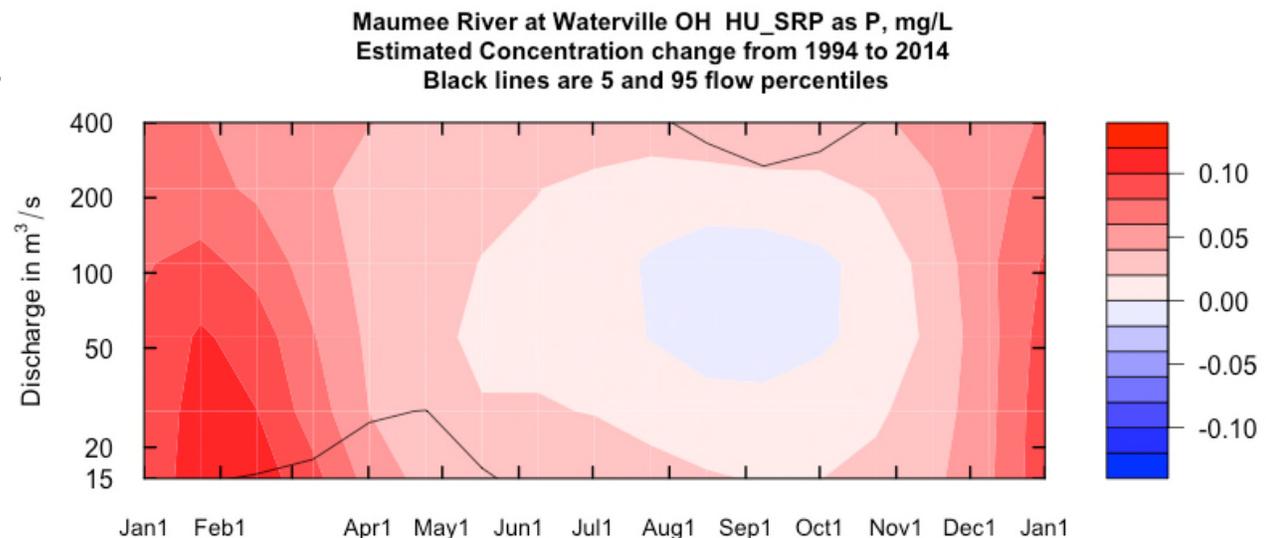
24.82

25.21



WRTDS

- Added the flux calculation and trend analysis method Weighted Regressions on Time, Discharge, and Season (WRTDS, written in R)
- WRTDS was the subject of an earlier session:
D5—Using WRTDS to Determine Long and Short Term Trends
- Soluble reactive phosphorus concentration example:
flexible depiction of change using smoothing concepts



Distribution-free Methods for Sampling Design

New Chapter Discusses

- what is needed to estimate sample size for group tests,
- methods for both parametric & nonparametric tests,
- and nonparametric methods not found in other applied books

Includes R code for computing sample size & power

What's Staying the Same?

- Content still at a level that practitioners can understand & use with a limited background in statistics.
- Examples provided are still simple enough to be worked through “by-hand.”
- Textbook will continue to be freely available online.

R Code for Chapter 11, Multiple Linear Regression

```
# provided the .RData file is saved to the current working  
# directory  
# it may be loaded for use in some of the examples and  
# exercises.
```

```
load("Chapter11.RData")
```

**Electronic rather
than
tabular text
delivery of data.**



<u>Obs. #</u>	<u>DE</u>	<u>DN</u>	<u>D</u>	<u>C</u>	<u>hi</u>	<u>DFFITs</u>
1	1	1	4.2122	30.9812	0.289433	-0.30866
2	2	1	8.0671	33.1540	0.160670	-0.01365
3	3	1	10.7503	37.1772	0.164776	0.63801
4	4	1	11.9187	35.3864	0.241083	-0.04715
5	1	2	11.2197	35.9388	0.170226	0.42264
6	2	2	12.3710	31.9702	0.086198	-0.51043
7	3	2	12.9976	34.9144	0.087354	-0.19810
8	4	2	15.0709	36.5436	0.165040	-0.19591
9	1	3	12.9886	38.3574	0.147528	0.53418
10	2	3	18.3469	39.8291	0.117550	0.45879
11	3	3	20.0328	40.0678	0.121758	0.28961
12	4	3	20.5083	37.4143	0.163195	-0.47616
13	1	4	17.6537	35.3238	0.165025	-0.59508
14	2	4	17.5484	34.7647	0.105025	-0.77690
15	3	4	23.7468	40.7207	0.151517	0.06278
16	4	4	13.1110	42.3420	0.805951	4.58558
17	1	5	20.5215	41.0219	0.243468	0.38314
18	2	5	23.6314	40.6483	0.165337	-0.08027
19	3	5	24.1979	42.8845	0.160233	0.17958
20	4	5	28.5071	43.7115	0.288632	0.09397

Table 11.1 Data and diagnostics for Example 1

```
#####  
Figure 11.1  
#####  
  
# view a subset of the data  
head(Chap11Ex1)  
  
# color code observation 16  
Chap11Ex1$ptpch <- 1; Chap11Ex1$ptcol <- 1  
Chap11Ex1$ptpch[16] <- 2; Chap11Ex1$ptcol[16] <- 2  
  
# use a pairs plot to plot the variables against each other  
pairs(Chap11Ex1[,1:4], pch=Chap11Ex1$ptpch,  
      col=Chap11Ex1$ptcol, las=1)
```

Code provided for figures



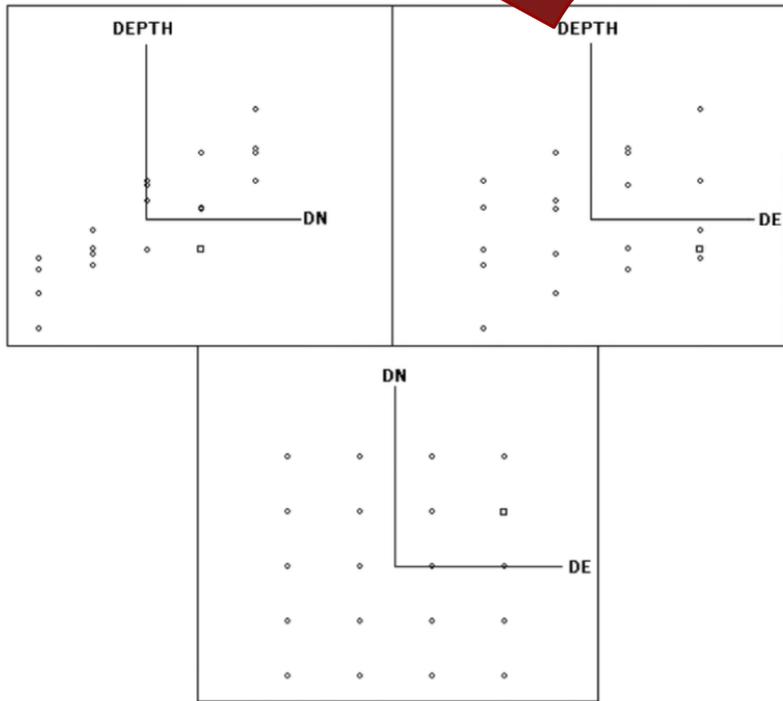
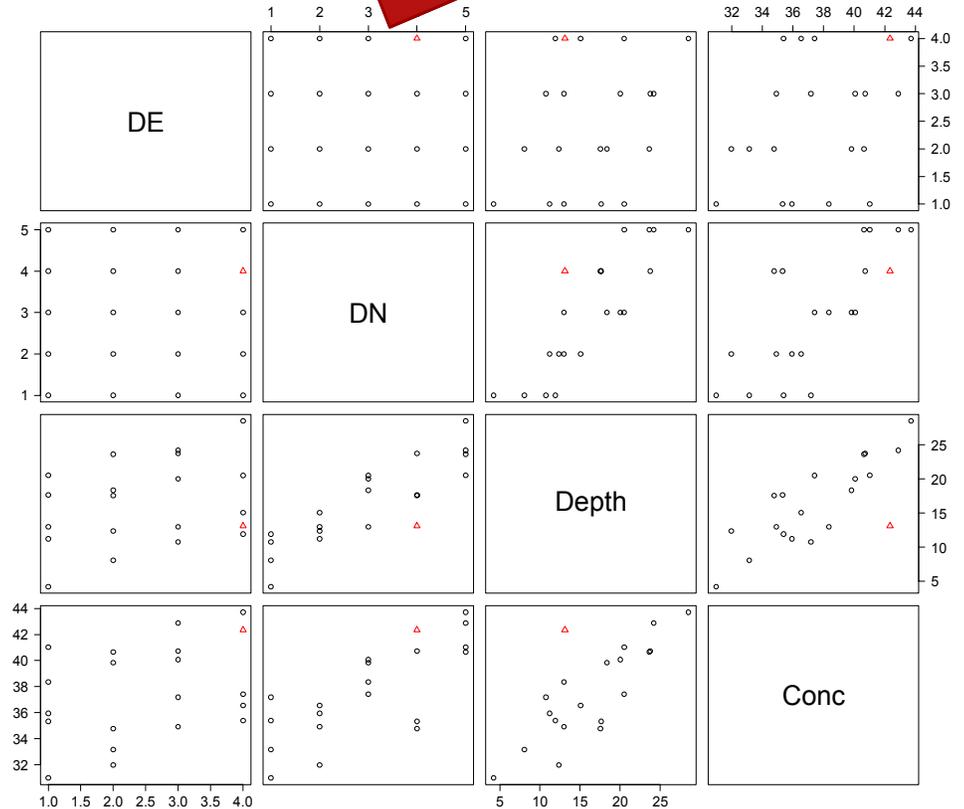
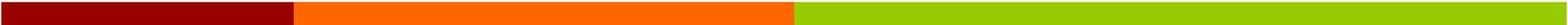


Figure 11.1 Scatterplot matrix for the 3 explanatory variables (obs. 16 is shown as a square)

**Data + technique +
plus code =
Reproducible results** ✓





```
#####
```

```
# Example 1 model
```

```
#####
```

```
mod1 <- lm(Conc ~ DE + DN + Depth, data=Chap11Ex1)
summary(mod1)
```

```
Chap11Ex1.correct <- Chap11Ex1
Chap11Ex1.correct[16,"D"] <- 23.111
```

```
mod2 <- lm(Conc ~ DE + DN + D, data=Chap11Ex1.correct)
summary(mod2)
```

```
mod3 <- lm(Conc ~ D, data=Chap11Ex1.correct)
summary(mod3)
```

Conc = 28.9 + 0.991 DE + 1.60 DN + 0.091 D

n = 20 s = 2.14 R² = 0.71

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β ₀	28.909	1.582	18.28	0.000
Slopes β _k				
DE	0.991	0.520	1.90	0.075
DN	1.596	0.751	2.13	0.049
D	0.091	0.186	0.49	0.632

Table 11.2 Regression statistics for Example 1

One outlier has had a severe detrimental effect on the regression coefficients and model structure. Points of high leverage and influence should always be examined before accepting a regression model, to determine if they represent errors. Suppose that the "typographical error" was detected and corrected. Table 11.3 shows that the resulting regression relationship is drastically changed:

C = 29.2 - 0.419 DE - 0.82 DN + 0.710 D

n = 20 s = 1.91 R² = 0.77

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β ₀	29.168	1.387	21.03	0.000
Slopes β _k				
DE	-0.419	0.833	-0.50	0.622
DN	-0.816	1.340	-0.61	0.551
D	0.710	0.339	2.10	0.052

Table 11.3 Regression statistics for the corrected Example 1 data

**Data + code + software
interaction + text explanation =
Increased opportunity for
interactive learning**



```
> mod1 <- lm(Conc ~ DE + DN + Depth, data=Chap11Ex1)
> summary(mod1)

Call:
lm(formula = Conc ~ DE + DN + Depth, data = Chap11Ex1)

Residuals:
    Min       1Q   Median       3Q      Max
-4.101 -1.006  0.106  1.645  2.726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.90857   1.58151  18.279 3.81e-12 ***
DE           0.99058   0.52033   1.904  0.0751 .
DN           1.59599   0.75055   2.126  0.0494 *
Depth        0.09069   0.18572   0.488  0.6319
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.144 on 16 degrees of freedom
Multiple R-squared:  0.7112,    Adjusted R-squared:  0.657
F-statistic: 13.13 on 3 and 16 DF,  p-value: 0.0001393

>
> Chap11Ex1.correct <- Chap11Ex1
> Chap11Ex1.correct[16,"D"] <- 23.111
>
> mod2 <- lm(Conc ~ DE + DN + Depth, data=Chap11Ex1.correct)
> summary(mod2)

Call:
lm(formula = Conc ~ DE + DN + Depth, data = Chap11Ex1.correct)

Residuals:
    Min       1Q   Median       3Q      Max
-4.101 -1.006  0.106  1.645  2.726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.90857   1.58151  18.279 3.81e-12 ***
DE           0.99058   0.52033   1.904  0.0751 .
DN           1.59599   0.75055   2.126  0.0494 *
Depth        0.09069   0.18572   0.488  0.6319
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.144 on 16 degrees of freedom
Multiple R-squared:  0.7112,    Adjusted R-squared:  0.657
F-statistic: 13.13 on 3 and 16 DF,  p-value: 0.0001393
```

Workflow

By providing

descriptions of the assumptions of statistical methods, guidelines, interpretation, and code that allows readers to reproduce the tests and graphics,

We hope

to contribute to improved workflow in statistical analyses of hydrologic data.

Parting Comments

- Draft is in preparation.
- The plan is to publish it on-line as a USGS Techniques and Methods Report and as hardcopy.
- We will be looking for reviewers! For individual chapters as well as the complete text.
- Please share your thoughts on needs in hydrologic statistics if you have any particular concerns.

Two of us are at the conference if you would like to share your thoughts in person.

Bob Hirsch, rhirsch@usgs.gov

Karen Ryberg, kryberg@usgs.gov

