36-401 Class Notes and Exercise for 10/24/00

Howard Seltman

This dataset contains stream sediment yield and other measurements for 19 mountain basins mostly in the Southern Alps of New Zealand. Interest centers on learning about the factors that influence the amount of sediment ending up in downstream lakes.

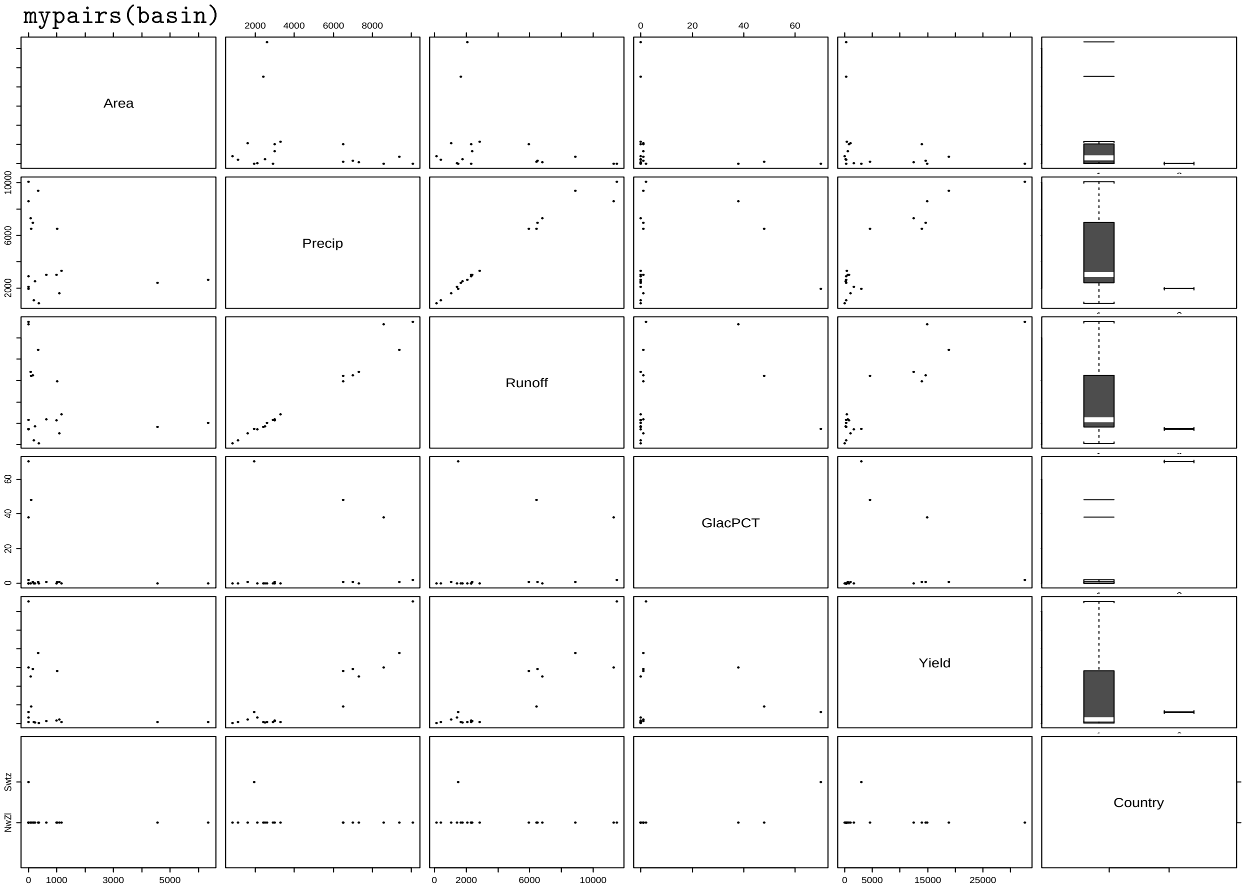
```
Contains data from Hicks et al. (1990)     Obs: 19
Basin     Basin
Area      Drainage Area (km^2)
Precip    Annual Precipitation (mm)
Runoff    Annual Runoff (mm)
Glacier   Percent Glacierized
Yield     Annual Sediment Yield (t/km^2)
Country   Country

basin_read.table("basins.dat",T)
dim(basin)  # 19 4
basin
#                       Area Precip Runoff GlacPCT Yield Country
#               Ivory    2.1   8600  11300      38 14900 NewZeal
#               Cropp    2.9  10070  11500       2 32600 NewZeal
# Upper Waitangitoana   77.0   7300   6800       0 12500 NewZeal
#            Hokitika  352.0   9400   8870       1 18800 NewZeal
#               Haast 1020.0   6500   5970       1 14000 NewZeal
# Little Hopwood Burn   17.0   2100   1430       0  1640 NewZeal
#            Shotover 1088.0   1600   1060       1  1120 NewZeal
#               Arrow  199.0   1100    410       0   270 NewZeal
#         Manuherikia  370.0    830    130       0    44 NewZeal
#             Karamea 1160.0   3300   2860       0   360 NewZeal
#            Buller A 4560.0   2400   1660       0   300 NewZeal
#            Buller B 6350.0   2600   2070       0   300 NewZeal
#         Inangahua A  234.0   2500   1750       0   190 NewZeal
#         Inangahua B 1000.0   3000   2300       0   800 NewZeal
#                Grey  642.0   3000   2360       1   610 NewZeal
#      Butchers Creek    4.0   2900   2320       0   300 NewZeal
#             Cleddau  155.0   7000   6510       1 14600 NewZeal
#              Hooker  103.0   6500   6430      48  4600 NewZeal
#     Tsidjiore Nouve    4.8   1947   1480      70  3010   Switz
```

What problems can you see in the data?

What kind of variable is Country?

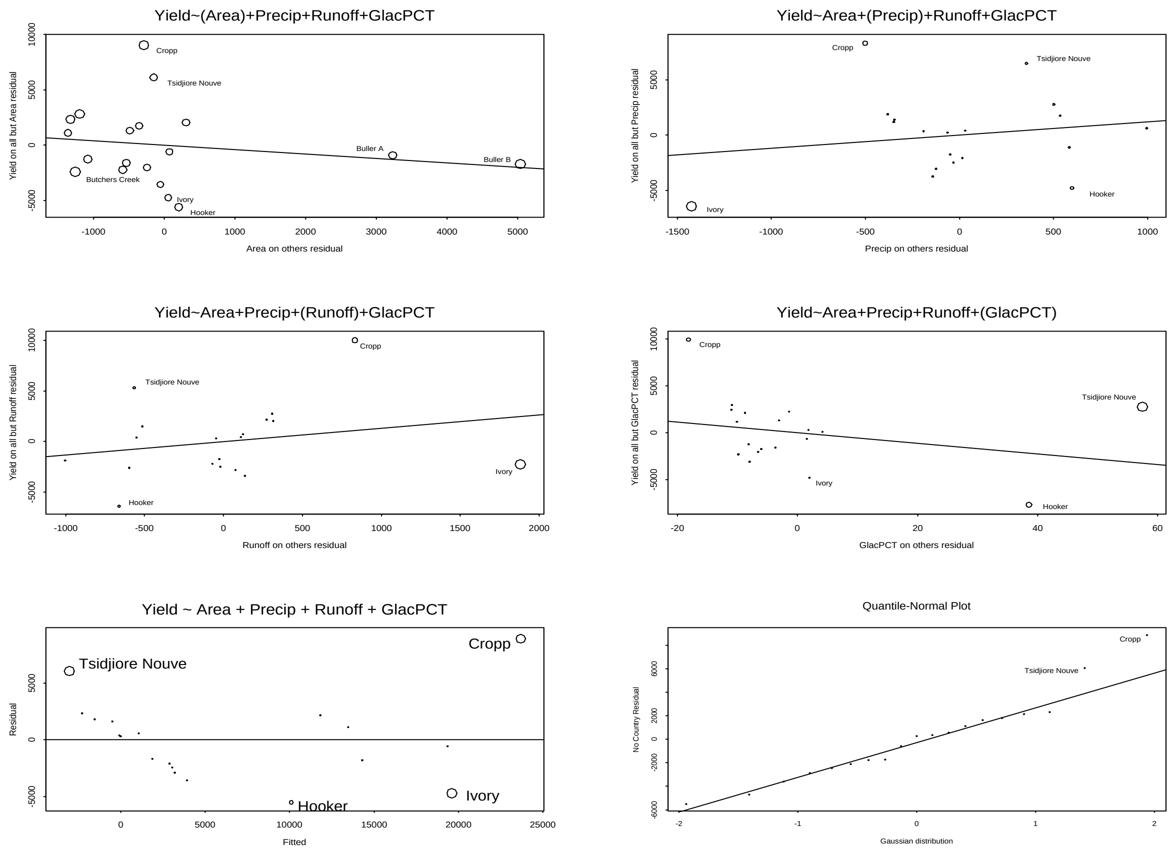
What alternate outcome can you think of?

```
mypairs(basin)
```



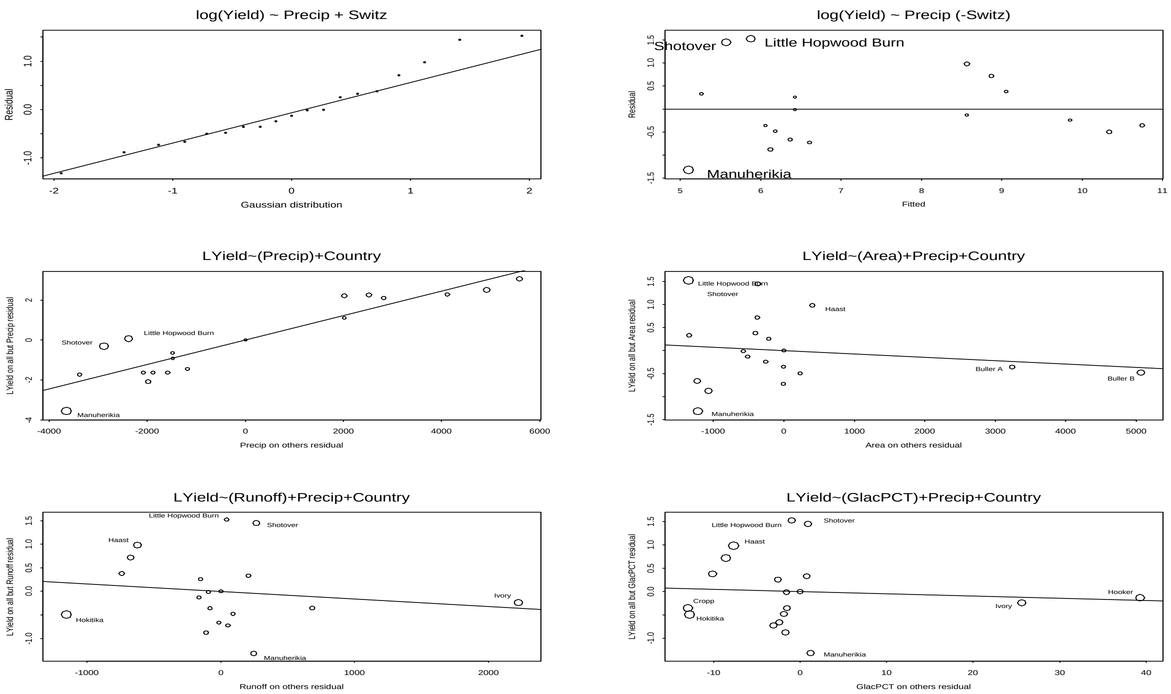

```
full_lm(Yield~Area+Precip+Runoff+GlacPCT+Country, basin)
summary(full)

# Coefficients:    Value Std. Error    t value   Pr(>|t|)
# (Intercept)  6688.2221  3220.7741     2.0766     0.0582
#        Area    -0.3321     0.4271    -0.7776     0.4507
#      Precip    -0.1610     1.3157    -0.1223     0.9045
#      Runoff     2.8338     1.1354     2.4959     0.0268
#     GlacPCT  -247.1119    62.3663    -3.9623     0.0016
#     Country  9740.5499  2564.7974     3.7978     0.0022
#
# Residual standard error: 2860 on 13 degrees of freedom
# Multiple R-Squared: 0.9276
#
# Correlation of Coefficients:
#         (Intercept)    Area  Precip  Runoff GlacPCT
#    Area -0.1972
#  Precip -0.6432      0.0590
#  Runoff  0.6446     -0.0108 -0.9807
# GlacPCT -0.7978      0.0911  0.4246 -0.4923
# Country  0.8176      0.0399 -0.2714  0.3529 -0.8019
```

```
inf_influence(full)
Error in .Fortran(if(!cmplx) "dbksl" else "zbksl",: subroutine dbksl: 6 Inf
        value(s) in argument 4

par(mfrow=c(3,2), mai=c(0.6,0.6,0.2,0.2))
partreg(1:4, "Yield", "Area", basin, cex=0.6, tcex=0.7, prop=T, id=T)
partreg(1:4, "Yield", "Precip", basin, cex=0.6, tcex=0.7, prop=T, id=T)
partreg(1:4, "Yield", "Runoff", basin, cex=0.6, tcex=0.7, prop=T, id=T)
partreg(1:4, "Yield", "GlacPCT", basin, cex=0.6, tcex=0.7, prop=T, id=T)
CookPlot(lm(Yield~Area+Precip+Runoff+GlacPCT, basin), cex=0.6, tcex=0.7, id=T)
qqn(lm(Yield~Area+Precip+Runoff+GlacPCT, basin)$res, ylab="No Country Residual",
  labels=row.names(basin))
ps("basinfull")
```



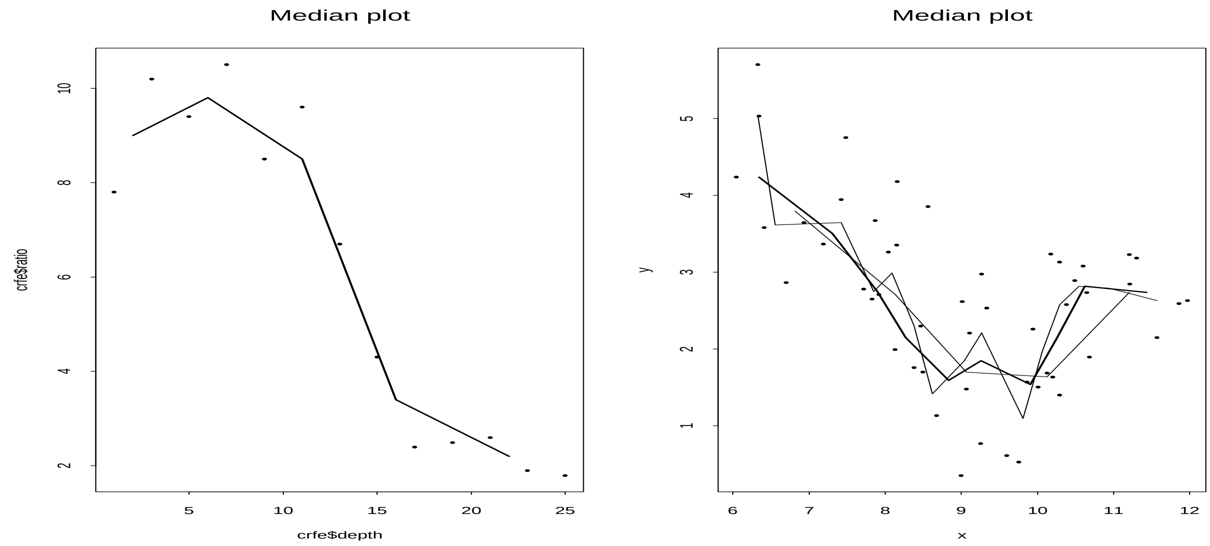

```
basin$LRunoff_log(basin$Runoff)
basin$LPrecip_log(basin$Precip)
basin$LYield_log(basin$Yield)
basin$LArea_log(basin$Area)
sw_stepwise(basin[,c(1:4,6:8,10)], basin$LYield, meth="exh", plot=F)
sum.step(sw,basin,"LYield")
#          BIC   R2a Area Precip Runoff GlacPCT Country LRunoff LPrecip LArea
# 2(#1) -25.9 0.827    0      1      0       0       1       0       0     0
# 3(#1) -26.8 0.829    0      1      0       0       1       1       0     0
# 3(#3) -27.2 0.820    1      1      0       0       1       0       0     0
# 3(#2) -27.2 0.821    0      1      0       0       1       0       1     0
# 2(#2) -27.7 0.790    0      1      0       1       0       0       0     0
# 1(#1) -27.7 0.767    0      1      0       0       0       0       0     0
```

```
BIC1_lm(LYield~Precip+Country, basin)
BIC1$contrast
# $Country:
# NewZeal   -1
#   Switz    1
summary(BIC1)
# Coefficients: Value Std. Error t value Pr(>|t|)
# (Intercept)  5.7074  0.4660    12.2478  0.0000
#      Precip  0.0006  0.0001     9.3507  0.0000
#     Country  1.1130  0.4241     2.6241  0.0184
#
# Residual standard error: 0.8098 on 16 degrees of freedom
# Multiple R-Squared: 0.8463

BIC1_lm(LYield~Precip+C(Country,treatment), basin)
BIC1$contrast
# $"C(Country, treatment)":
#         Switz
# NewZeal     0
#   Switz     1
# summary(BIC1)
# Coefficients:           Value Std. Error t value Pr(>|t|)
#           (Intercept)  4.5944  0.3496    13.1425  0.0000
#                Precip  0.0006  0.0001     9.3507  0.0000
# C(Country, treatment)  2.2260  0.8483     2.6241  0.0184
```

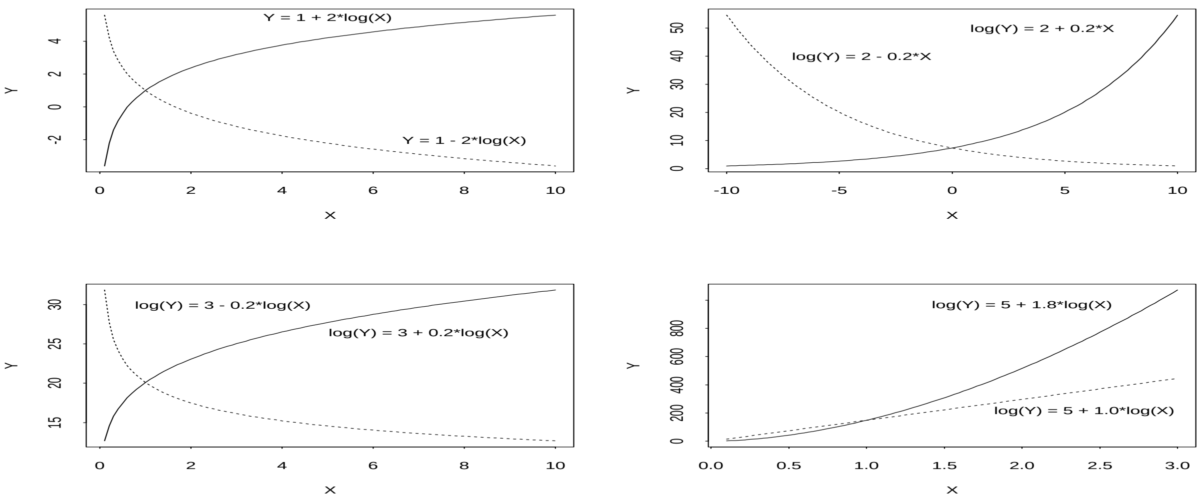


RwG Chapter 5: Fitting Curves

- Exploratory Band Regression: non-parametric guidance for parametric model fitting



- Some Curvilinear Regression Models



Multiplicative error: $Y_i = \alpha e^{\beta X_i} \epsilon_i \;\Rightarrow\; \log(Y_i) = \alpha^* + \beta X_i + \epsilon_i^*$
$Y_i = \alpha X_i^{\beta} \epsilon_i \;\Rightarrow\; \log(Y_i) = \alpha^* + \beta \log(X_i) + \epsilon_i^*$

Breakout Session for 10/24/00

1. Exactly what effect does a "single dummy" variable with n-1 zeros and one one have on a regression? Consider RSS, slope, and R^2. What can you say about the geometry of a problem with one continuous variable and one "single dummy" variable?

2. Use the results

   ```
   > predict(lm(LYield~Precip,basin,subset=basin$Country!="Switz"),basin[19,])
    Tsidjiore Nouve
           5.783697
   ```

   to verify the meaning of the "single dummy" slope.

3. What benefit is there to including a "single case dummy" variable as opposed to dropping the case?

4. Exactly what commands could you use to set a single case variable for case number 134 in the wages data set?