The Statistical Sleuth in R: Chapter 12

Kate Aloisio Ruobing Zhang Nicholas J. Horton*

July 25, 2024

Contents

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Second Edition of the *Statistical Sleuth* (2002) by Fred Ramsey and Dan Schafer. More information about the book can be found at `http://www.proaxis.com/~panorama/home.htm`. This file as well as the associated knitr reproducible analysis source file can be found at `http://www.amherst.edu/~nhorton/sleuth`.

This work leverages initiatives undertaken by Project MOSAIC (`http://www.mosaic-web.org`), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the mosaic package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignette (`http://cran.r-project.org/web/packages/mosaic/vignettes/MinimalR.pdf`).

To use a package within R, it must be installed (one time), and loaded (each session). The package can be installed using the following command:

*Department of Mathematics, Amherst College, nhorton@amherst.edu

```
> install.packages('mosaic')                # note the quotation marks
```

Once this is installed, it can be loaded by running the command:

```
> require(mosaic)
```

This needs to be done once per session.

In addition the data files for the *Sleuth* case studies can be accessed by installing the `Sleuth2` package.

```
> install.packages('Sleuth2')               # note the quotation marks
```

```
> require(Sleuth2)
```

We also set some options to improve legibility of graphs and output.

```
> trellis.par.set(theme=col.mosaic())  # get a better color scheme for lattice
> options(digits=4)
```

The specific goal of this document is to demonstrate how to calculate the quantities described in Chapter 12: Strategies for Variable Selection using R.

2 State Average SAT Scores

What variables are associated with state average SAT scores? This is the question addressed in case study 12.1 in the *Sleuth*.

2.1 Summary statistics

We begin by reading the data and summarizing the variables.

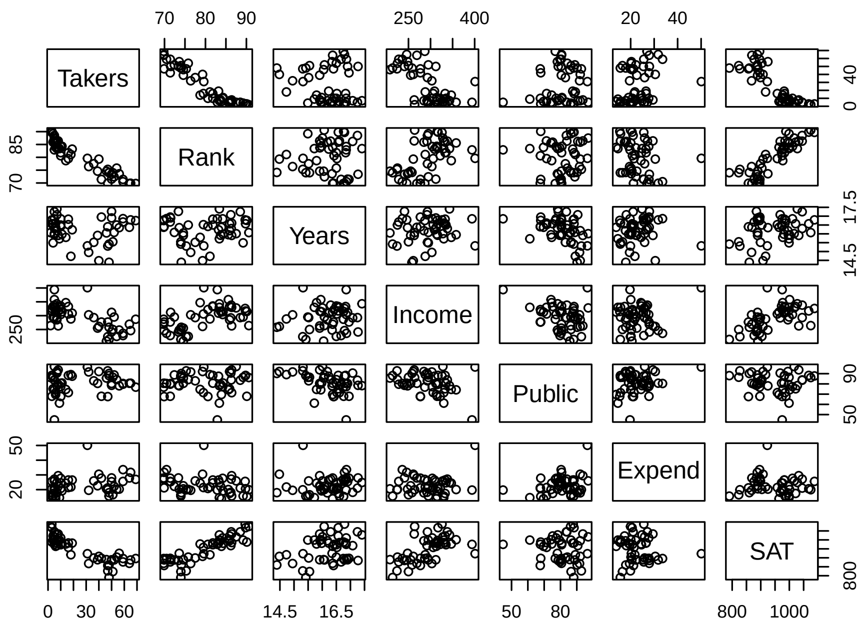
```
> summary(case1201)

    State                SAT           Takers          Income         Years     
 Length:50          Min.   : 790   Min.   : 2.00   Min.   :208   Min.   :14.4  
 Class :character   1st Qu.: 889   1st Qu.: 6.25   1st Qu.:262   1st Qu.:15.9  
 Mode  :character   Median : 966   Median :16.00   Median :295   Median :16.4  
                    Mean   : 948   Mean   :26.22   Mean   :294   Mean   :16.2  
                    3rd Qu.: 998   3rd Qu.:47.75   3rd Qu.:325   3rd Qu.:16.8  
                    Max.   :1088   Max.   :69.00   Max.   :401   Max.   :17.4  
     Public         Expend          Rank     
 Min.   :44.8   Min.   :13.8   Min.   :69.8  
 1st Qu.:76.9   1st Qu.:19.6   1st Qu.:74.0  
 Median :80.8   Median :21.6   Median :80.8  
 Mean   :81.2   Mean   :23.0   Mean   :80.0  
 3rd Qu.:88.2   3rd Qu.:26.4   3rd Qu.:85.8  
 Max.   :97.0   Max.   :50.1   Max.   :90.6  
```

The data are shown on page 340 (display 12.1). A total of 50 state average SAT scores are included in this data.

2.2 Dealing with Many Explanatory Variables

The following graph is presented as Display 12.4, page 348.

```
> pairs(~ Takers+Rank+Years+Income+Public+Expend+SAT, data=case1201)
```

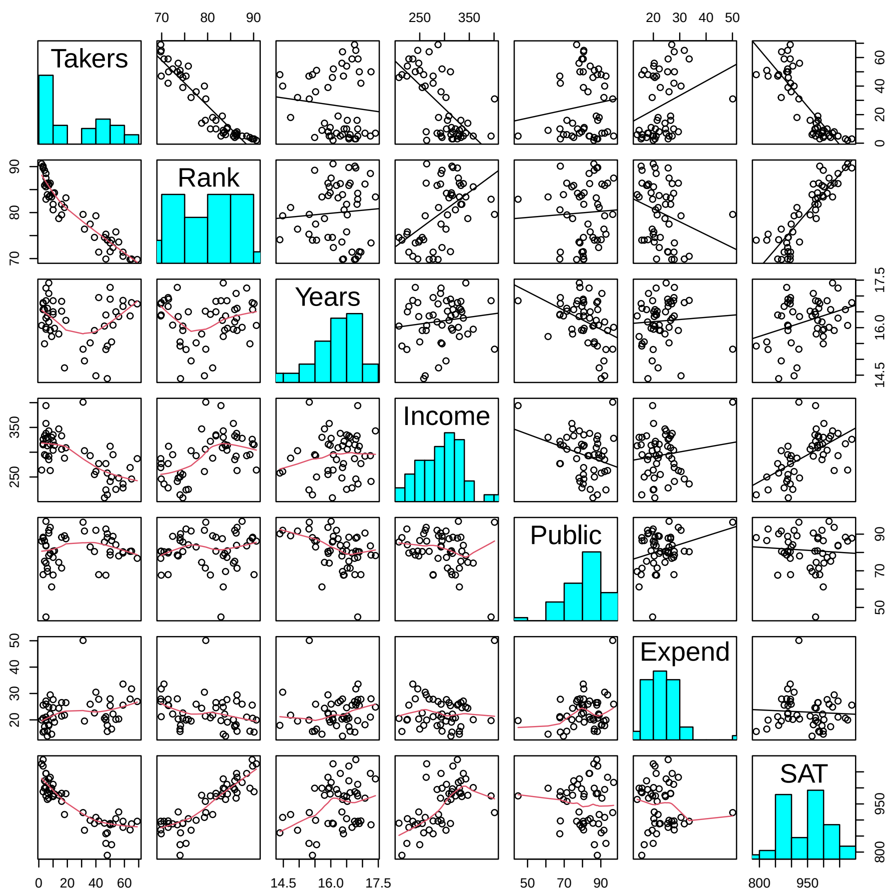


We can get a fancier graph using following code:

```
>   panel.hist = function(x, ...)
+   {
+     usr = par("usr"); on.exit(par(usr))
+     par(usr = c(usr[1:2], 0, 1.5) )
+     h = hist(x, plot=FALSE)
+     breaks = h$breaks; nB = length(breaks)
+     y = h$counts; y = y/max(y)
+     rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
+   }
>
> panel.lm = function(x, y, col=par("col"), bg=NA,
+                     pch=par("pch"), cex=1, col.lm="red", ...)
+ {
+   points(x, y, pch=pch, col=col, bg=bg, cex=cex)
+   ok = is.finite(x) & is.finite(y)
```

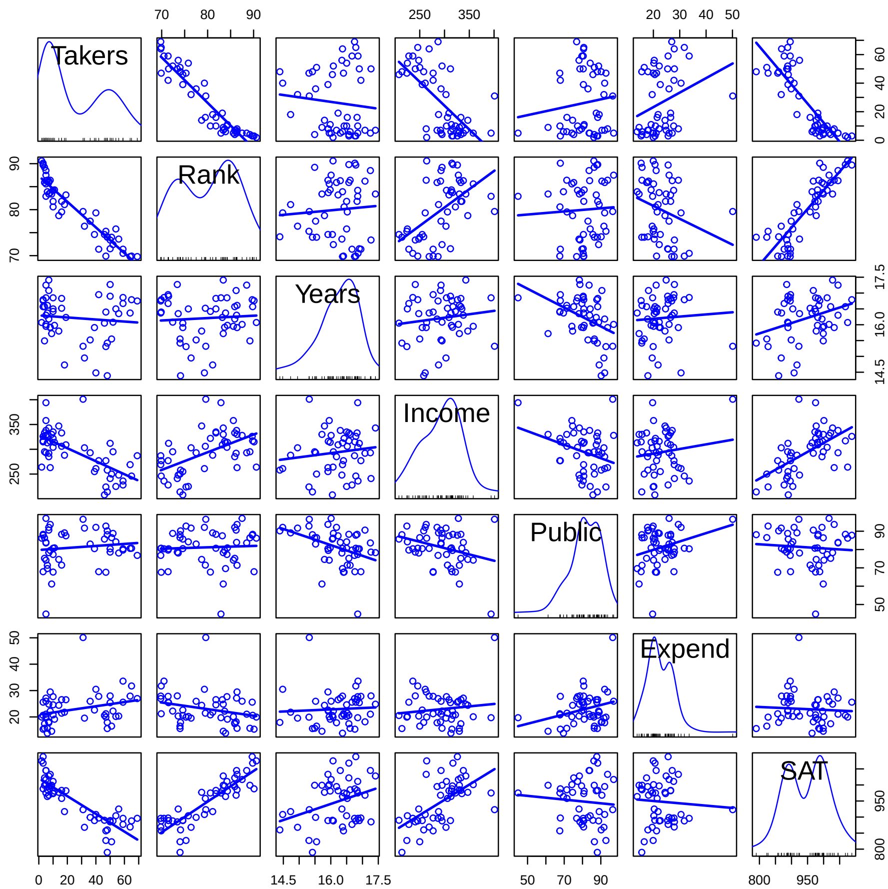
```
+   if (any(ok))
+     abline(lm(y[ok] ~ x[ok]))
+ }
```

```
> pairs(~ Takers+Rank+Years+Income+Public+Expend+SAT,
+       lower.panel=panel.smooth, diag.panel=panel.hist,
+       upper.panel=panel.lm, data=case1201)

Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
```



An alternative graph can be generated using the `car` package.

```
> require(car)
> scatterplotMatrix(~ Takers+Rank+Years+Income+Public+Expend+SAT,
+   diagonal="histogram", smooth=FALSE, data=case1201)

Warning in applyDefaults(diagonal, defaults = list(method = "adaptiveDensity"), :  unnamed
diag arguments, will be ignored
```



Based on the scatterplot, we choose the logarithm of percentage of SAT takers and median class rank to fit our first model (page 349):

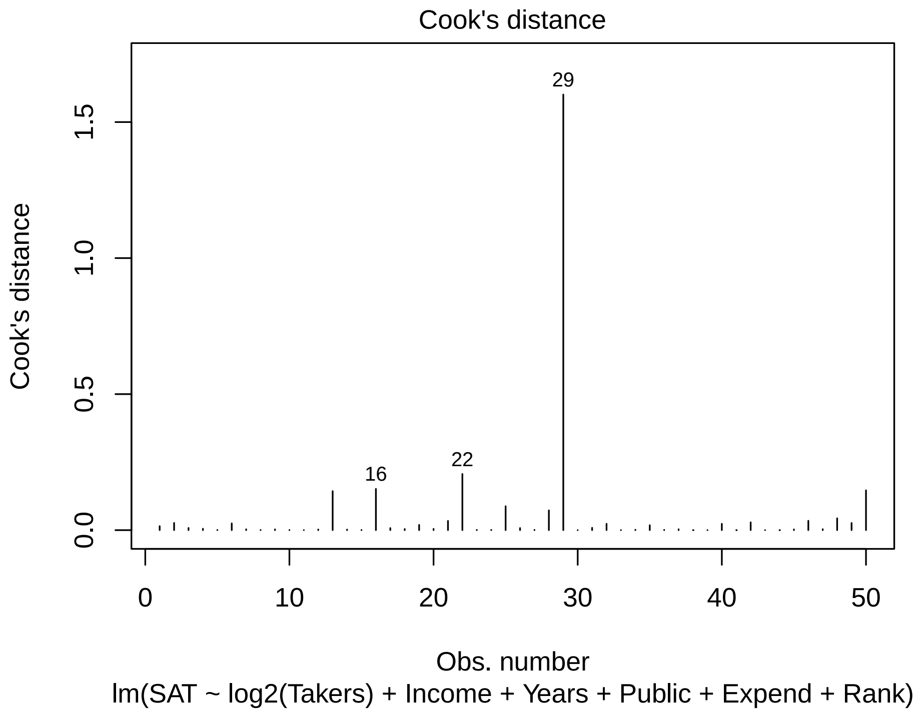
```
> lm1 = lm(SAT ~ Rank+log(Takers), data=case1201)
> summary(lm1)


Call:
lm(formula = SAT ~ Rank + log(Takers), data = case1201)

Residuals:
   Min      1Q  Median      3Q     Max
```

```
-94.46 -17.31   5.32  22.82  48.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   882.08     224.13    3.94  0.00027
Rank            2.40       2.33    1.03  0.30898
log(Takers)   -45.19      14.06   -3.21  0.00236

Residual standard error: 31.1 on 47 degrees of freedom
Multiple R-squared:  0.815, Adjusted R-squared:  0.807
F-statistic:  103 on 2 and 47 DF,  p-value: <2e-16
```

From the regression output, we observe that these two variables can explain 81.5% of the variation.

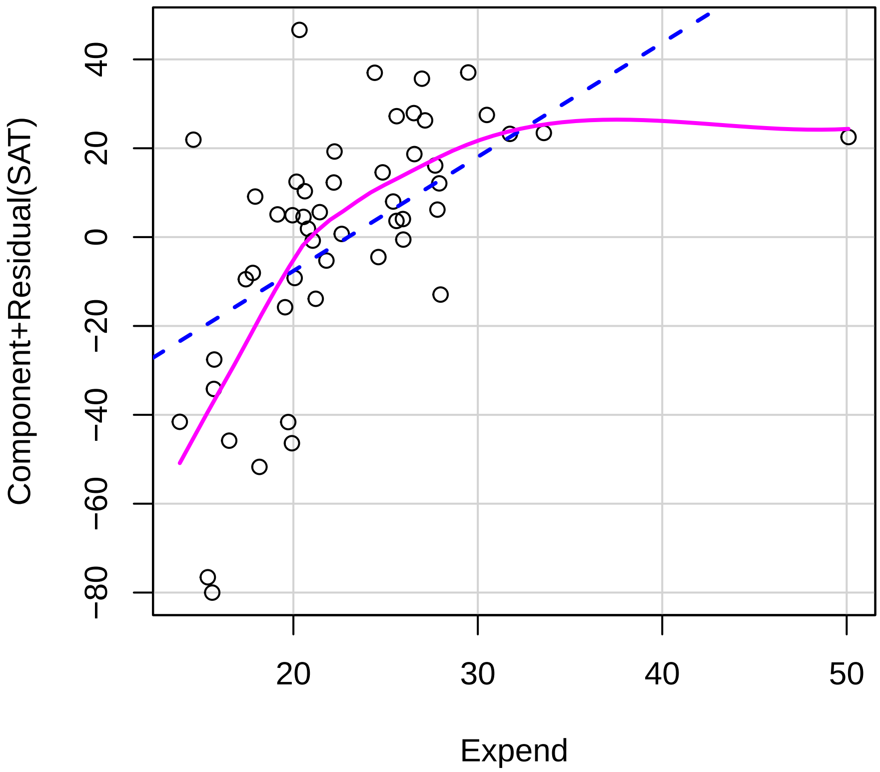
Next we fit a linear regression model using all variables and create the partial residual plot presented on page 349 as Display 12.5:

```
> lm2 = lm(SAT ~ log2(Takers)+Income+Years+Public+Expend+Rank, data=case1201)
> summary(lm2)


Call:
lm(formula = SAT ~ log2(Takers) + Income + Years + Public + Expend +
    Rank, data = case1201)

Residuals:
   Min     1Q Median     3Q    Max
-61.11  -8.60   2.86  14.77  53.40

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  407.5399   282.7633    1.44   0.1567
log2(Takers) -26.6429    11.0572   -2.41   0.0203
Income        -0.0359     0.1301   -0.28   0.7841
Years         17.2181     6.3201    2.72   0.0093
Public        -0.1130     0.5624   -0.20   0.8417
Expend         2.5669     0.8064    3.18   0.0027
Rank           4.1143     2.5017    1.64   0.1073

Residual standard error: 24.9 on 43 degrees of freedom
Multiple R-squared:  0.892, Adjusted R-squared:  0.877
F-statistic: 59.2 on 6 and 43 DF,  p-value: <2e-16

> plot(lm2, which=4)
```

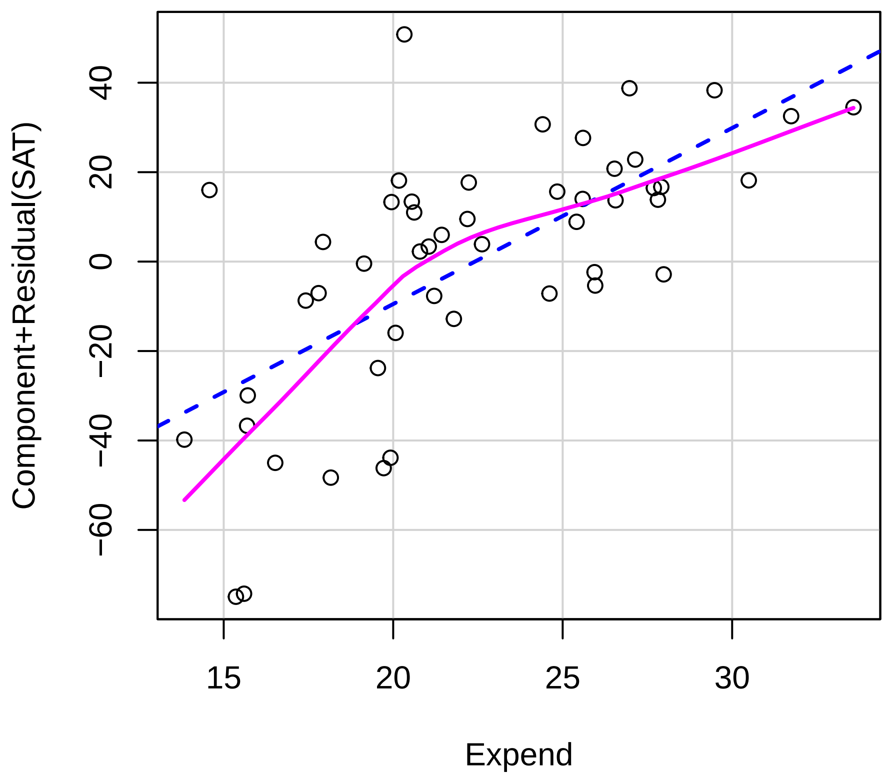


According to the Cook's distance plot, obs 29 (Alaska) seems to be an influential outlier. We may consider removing this observation from the dataset.

```
> case1201r = case1201[-c(29),]
> lm3 = lm(SAT ~ log2(Takers) + Income+ Years + Public + Expend + Rank, data=case1201r)
> anova(lm3)

Analysis of Variance Table

Response: SAT
             Df Sum Sq Mean Sq F value  Pr(>F)
log2(Takers)  1 199007  199007  390.63 < 2e-16
Income        1    785     785    1.54  0.2214
Years         1   5910    5910   11.60  0.0015
Public        1   5086    5086    9.98  0.0029
Expend        1  10513   10513   20.64 4.6e-05
Rank          1   2679    2679    5.26  0.0269
Residuals    42  21397     509

> crPlots(lm2, term = ~ Expend) # with Alaska
> crPlots(lm3, term = ~ Expend) # without Alaska
```



Component+Residual(SAT)
40
20
0
−20
−40
−60
−80
20
30
40
50
Expend



The difference between these two slopes indicates that Alaska is an influential observation. We decide to remove it from the original dataset.

2.3 Sequential Variable Selection

The book uses F-statistics as the criterion and present the procedures of forward selection and backward elimination on page 351-353. The forward selection requires fitting 16 of 64 possible models. The final model uses Expenditure and `log(Takers)` to predict SAT. The backward elimination method needs to fit 3 models, and the final model uses Year, Expenditure, Rank and log(Takers) to predict SAT.

To the best of our knowledge, there is no built-in mechanism to undertake the procedure using the F statistic as criterion. Instead, we demonstrate how to use AIC (another criterion) and select the final model. Note that we choose log(Taker) as our preliminary predictor for forward selection, because it has the largest F-value when we fit `lm3`.

```
> # Forward Selection
> lm4 = lm(SAT ~ log2(Takers), data=case1201r)
> stepAIC(lm4, scope=list(upper=lm3, lower=~1), direction="forward",
+   trace=FALSE)$anova
```

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
SAT ~ log2(Takers)

Final Model:
SAT ~ log2(Takers) + Expend + Years + Rank


      Step Df Deviance Resid. Df Resid. Dev   AIC
1                             47      46369 339.8
2 + Expend  1    20523         46      25846 313.1
3  + Years  1     1248         45      24598 312.7
4   + Rank  1     2676         44      21922 309.1

> # Backward Elimination
> stepAIC(lm3, direction="backward", trace=FALSE)$anova

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
SAT ~ log2(Takers) + Income + Years + Public + Expend + Rank

Final Model:
SAT ~ log2(Takers) + Years + Expend + Rank


      Step Df Deviance Resid. Df Resid. Dev   AIC
1                             42      21397 311.9
2 - Public  1     20.0         43      21417 309.9
3 - Income  1    505.4         44      21922 309.1

> # Stepwise Regression
> stepAIC(lm3, direction="both", trace=FALSE)$anova

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
SAT ~ log2(Takers) + Income + Years + Public + Expend + Rank

Final Model:
SAT ~ log2(Takers) + Years + Expend + Rank
```

```
      Step Df Deviance Resid. Df Resid. Dev   AIC
1                             42      21397 311.9
2 - Public  1     20.0        43      21417 309.9
3 - Income  1    505.4        44      21922 309.1
```

The final model includes log(Takers), Expenditure, Years and Rank.

```
> lm5 = lm(SAT ~ log2(Takers) + Expend + Years + Rank, data=case1201r)
> summary(lm5)


Call:
lm(formula = SAT ~ log2(Takers) + Expend + Years + Rank, data = case1201r)

Residuals:
   Min     1Q Median     3Q    Max
-52.30  -9.92   0.60  11.88  59.20

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   399.115    232.372    1.72   0.0929
log2(Takers)  -26.409      8.259   -3.20   0.0026
Expend          3.996      0.764    5.23  4.5e-06
Years          13.147      5.478    2.40   0.0207
Rank            4.400      1.899    2.32   0.0252

Residual standard error: 22.3 on 44 degrees of freedom
Multiple R-squared:  0.911, Adjusted R-squared:  0.903
F-statistic:  112 on 4 and 44 DF,  p-value: <2e-16
```

The final model can explain 91.1% percent or the variation of SAT. All of the explanatory variables are statistically significant at the $\alpha = .05$ level.

2.4 Model Selection Among All Subsets

The Cp-statistic can be an useful criterion to select model among all subsets. We'll give an example about how to calculate this statistic for one model, which includes log(Takers), Expenditure, Years and Rank.

```
> sigma5 = summary(lm5)$sigma^2 # sigma-squared of chosen model
> sigma3 = summary(lm3)$sigma^2 # sigma-squared of full model
> n = 49 # sample size
> p = 4+1 # number of coefficients in model
> Cp=(n-p)*sigma5/sigma3+(2*p-n)
> Cp
```

```
[1] 4.031
```

The Cp statistic for this fitting model is 4.0312.

Alternatively, the Cp statistic can be calculated using the following command:

```
> require(leaps)
> explanatory = with(case1201r, cbind(log(Takers), Income, Years, Public, Expend, Rank))
> with(case1201r, leaps(explanatory, SAT, method="Cp"))$which[27,]

    1     2     3     4     5     6
 TRUE FALSE  TRUE FALSE  TRUE  TRUE
```

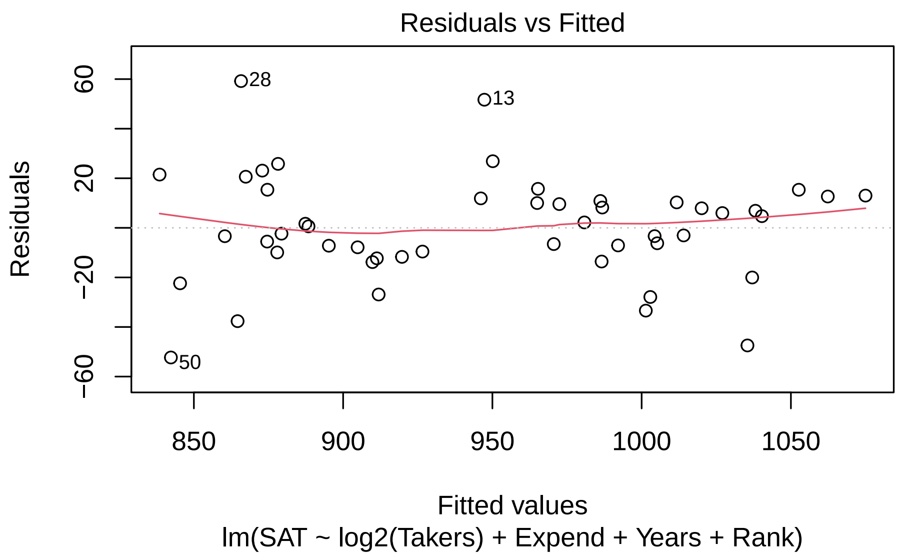
This means that the 27th fitting model includes log(Takers), Years and Expend.

```
> with(case1201r, leaps(explanatory, SAT, method="Cp"))$Cp[27]

[1] 4.031
```

The Cp statistic for this model is 4.0312. This will be the "tyer" point on Display 12.9, page 357.

We use the following code to generate the graph presented as Display 12.14 on page 364.

```
> plot(lm5, which=1)
```



From the scatterplot, we see that obs 28 (New Hampshire) has the largest residual, while obs 50 (Sorth Carolina) has the smallest.

2.5 Contribution of Expend

Display 12.13 (page 363) shows the contribution of Expend to the model.

```
> lm7 = lm(SAT ~ Expend, data=case1201r)
> summary(lm7)


Call:
lm(formula = SAT ~ Expend, data = case1201r)

Residuals:
   Min     1Q Median     3Q    Max
-162.5  -57.7   17.0   46.6  141.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  961.724     49.888   19.28   <2e-16
Expend        -0.592      2.178   -0.27     0.79

Residual standard error: 72.2 on 47 degrees of freedom
Multiple R-squared:  0.00157, Adjusted R-squared:  -0.0197
F-statistic: 0.074 on 1 and 47 DF,  p-value: 0.787

> lm8 = lm(SAT ~ Income + Expend, data=case1201r)
> summary(lm8)


Call:
lm(formula = SAT ~ Income + Expend, data = case1201r)

Residuals:
   Min     1Q Median     3Q    Max
-91.15 -38.41  -2.58  27.29 159.52

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  604.682     73.209    8.26  1.2e-10
Income         1.127      0.196    5.73  7.2e-07
Expend         0.672      1.695    0.40     0.69

Residual standard error: 55.7 on 46 degrees of freedom
Multiple R-squared:  0.418, Adjusted R-squared:  0.392
F-statistic: 16.5 on 2 and 46 DF,  p-value: 3.95e-06
```

3 Sex Discrimination in Employment

Do females receive lower starting salaries than similarly qualified and similarly experience males and did females receive smaller pay increases than males? These are the questions explored in case 12.2 in the *Sleuth*.

3.1 Summary Statistics

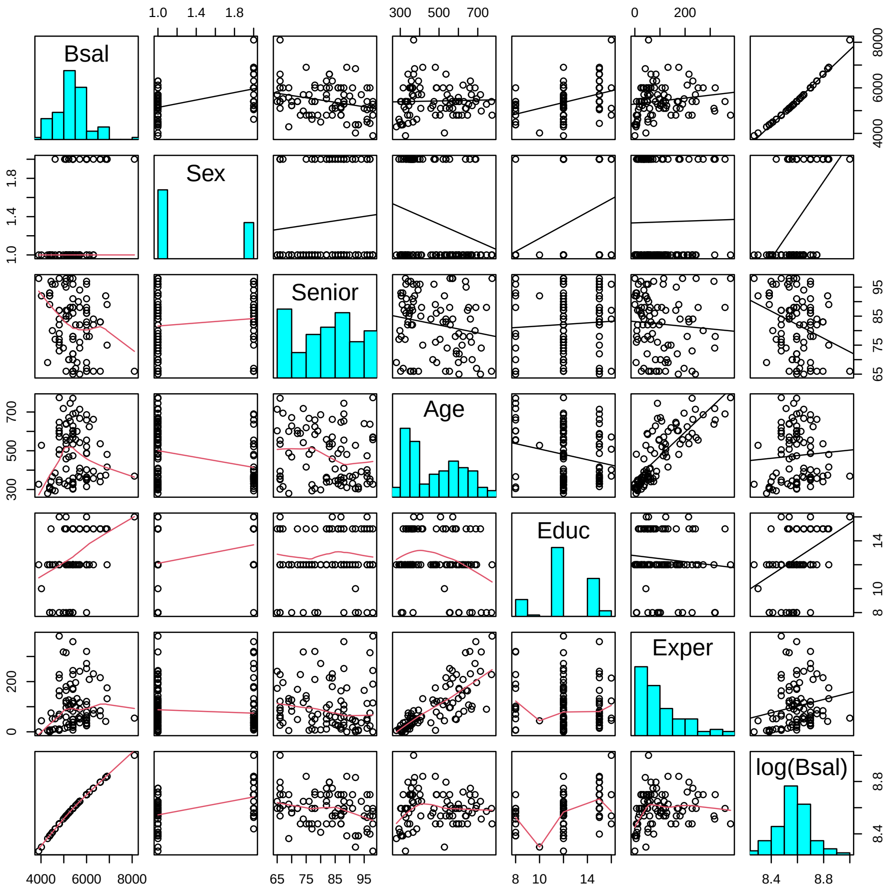
We begin by summarizing the data.

```
> summary(case1202)

      Bsal           Sal77            Sex          Senior           Age
 Min.   :3900   Min.   : 7860   Female:61   Min.   :65.0   Min.   :280
 1st Qu.:4980   1st Qu.: 9000   Male  :32   1st Qu.:74.0   1st Qu.:349
 Median :5400   Median :10020               Median :84.0   Median :468
 Mean   :5420   Mean   :10393               Mean   :82.3   Mean   :474
 3rd Qu.:6000   3rd Qu.:11220               3rd Qu.:90.0   3rd Qu.:590
 Max.   :8100   Max.   :16320               Max.   :98.0   Max.   :774
      Educ           Exper
 Min.   : 8.0   Min.   :  0.0
 1st Qu.:12.0   1st Qu.: 35.5
 Median :12.0   Median : 70.0
 Mean   :12.5   Mean   :100.9
 3rd Qu.:15.0   3rd Qu.:144.0
 Max.   :16.0   Max.   :381.0
```

The data is shown on page 343-344 as display 12.3. A total of 93 employee salaries are included: 61 females and 32 males.

Next we present a full graphical display for the variables within the dataset and the log of the beginning salary variable.

```
> pairs(~ Bsal+Sex+Senior+Age+Educ+Exper+log(Bsal),
+       lower.panel=panel.smooth, diag.panel=panel.hist,
+       upper.panel=panel.lm, data=case1202)

Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
Warning in par(usr):  argument 1 does not name a graphical parameter
```



Through these scatterplots it appears that beginning salary should be on the log scale and the starting model without the effects of gender will be a saturated second-order model with 14 variables including Seniority, Age, Education, Experience, as main effects, quadratic terms, and their full interactions.

3.2 Model Selection

To determine the best subset of these variables we first compared Cp statistics. Display 12.11 shows the Cp statistics for models that meet 'good practice' and have small Cp values. We will demonstrate how to calculate the Cp statistics for the two models with the lowest Cp statistics discussed in "Identifying Good Subset Models" on pages 359-360.

The first model includes Seniority, Age, Education, Experience, and the interactions between Seniority and Education, Age and Education, and Age and Experience. The second model includes Seniority, Age, Education, Experience, and the interactions between Age and Education and Age and Experience.

```
> require(leaps)
> explanatory1 = with(case1202, cbind(Senior, Age, Educ, Exper, Senior*Educ, Age*Educ, Age*Expe
> # First model (saexnck)
```

```
> with(case1202, leaps(explanatory1, log(Bsal), method="Cp"))$which[55,]

   1    2    3    4    5    6    7
TRUE TRUE TRUE TRUE TRUE TRUE TRUE

> with(case1202, leaps(explanatory1, log(Bsal), method="Cp"))$Cp[55]

[1] 8

> # second model (saexck)
> with(case1202, leaps(explanatory1, log(Bsal), method="Cp"))$which[49,]

    1     2     3     4     5     6     7
 TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE

> with(case1202, leaps(explanatory1, log(Bsal), method="Cp"))$Cp[49]

[1] 8.124
```

This first model has a Cp statistic of 8. Compared to the second model with a Cp statistic of 8.12.

We can also compare models using the BIC, we will next compare the second model with a thrid model defined as *saexyc* = Seniority + Age + Education + Experience + Experience^2 + Age*Education.

```
> BIC(lm(log(Bsal) ~ Senior+Age+Educ+Exper+Age*Educ+Age*Exper, data=case1202))

[1] -140.2

> BIC(lm(log(Bsal) ~ Senior+Age+Educ+Exper+(Exper)^2+Age*Educ, data=case1202))

[1] -131.3
```

Thus our final model is the second model, summarized below.

```
> lm1 = lm(log(Bsal) ~ Senior + Age + Educ + Exper + Age*Educ + Age*Exper, data=case1202)
> summary(lm1)


Call:
lm(formula = log(Bsal) ~ Senior + Age + Educ + Exper + Age *
    Educ + Age * Exper, data = case1202)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2817 -0.0476  0.0132  0.0605  0.2341
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.89e+00   2.45e-01   32.21  < 2e-16
Senior      -3.15e-03   1.04e-03   -3.04  0.00313
Age          1.24e-03   4.02e-04    3.09  0.00270
Educ         7.20e-02   1.67e-02    4.31  4.3e-05
Exper        2.86e-03   6.67e-04    4.28  4.8e-05
Age:Educ    -1.02e-04   3.15e-05   -3.25  0.00166
Age:Exper   -3.72e-06   1.02e-06   -3.65  0.00044

Residual standard error: 0.0974 on 86 degrees of freedom
Multiple R-squared:  0.469, Adjusted R-squared:  0.431
F-statistic: 12.6 on 6 and 86 DF,  p-value: 3.58e-10
```

3.3 Evaluating the Sex Effect

After selecting the model *saexck* = Seniority + Age + Education + Experience + Age*Education + Age*Experience we can add the sex indicator variable as summarized on page 360.

```
> lm2 = lm(log(Bsal) ~ Senior + Age + Educ + Exper + Age*Educ + Age*Exper + Sex, data=case1202)
> summary(lm2)


Call:
lm(formula = log(Bsal) ~ Senior + Age + Educ + Exper + Age *
    Educ + Age * Exper + Sex, data = case1202)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17822 -0.05197 -0.00203  0.05301  0.20466

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.16e+00   2.21e-01   36.99  < 2e-16
Senior      -3.48e-03   9.09e-04   -3.83  0.00024
Age          9.15e-04   3.57e-04    2.56  0.01218
Educ         4.23e-02   1.57e-02    2.70  0.00836
Exper        2.18e-03   5.98e-04    3.65  0.00045
SexMale      1.20e-01   2.29e-02    5.22  1.3e-06
Age:Educ    -5.46e-05   2.91e-05   -1.88  0.06402
Age:Exper   -3.23e-06   8.96e-07   -3.61  0.00052

Residual standard error: 0.0853 on 85 degrees of freedom
Multiple R-squared:  0.598, Adjusted R-squared:  0.564
F-statistic:   18 on 7 and 85 DF,  p-value: 1.79e-14
```

In contrast to the book, our reference group is Male, therefore the median male salary is estimated to be 1.13 times as large as the median female salary, adjusted for the other variables.