Chapter 12

Chicago Insurance Redlining - a complete example

In a study of insurance availability in Chicago, the U.S. Commission on Civil Rights attempted to examine charges by several community organizations that insurance companies were redlining their neighborhoods, i.e. canceling policies or refusing to insure or renew. First the Illinois Department of Insurance provided the number of cancellations, non-renewals, new policies, and renewals of homeowners and residential fire insurance policies by ZIP code for the months of December 1977 through February 1978. The companies that provided this information account for more than 70% of the homeowners insurance policies written in the City of Chicago. The department also supplied the number of FAIR plan policies written an renewed in Chicago by zip code for the months of December 1977 through May 1978. Since most FAIR plan policyholders secure such coverage only after they have been rejected by the voluntary market, rather than as a result of a preference for that type of insurance, the distribution of FAIR plan policies is another measure of insurance availability in the voluntary market.

Secondly, the Chicago Police Department provided crime data, by beat, on all thefts for the year 1975. Most Insurance companies claim to base their underwriting activities on loss data from the preceding years, i.e. a 2-3 year lag seems reasonable for analysis purposes. the Chicago Fire Department provided similar data on fires occurring during 1975. These fire and theft data were organized by zip code.

Finally the US Bureau of the census supplied data on racial composition, income and age and value of residential units for each ZIP code in Chicago. To adjust for these differences in the populations size associated with different ZIP code areas, the theft data were expressed as incidents per 1,000 population and the fire and insurance data as incidents per 100 housing units.

The variables are

`race` racial composition in percent minority

`fire` fires per 100 housing units

`theft` theft per 1000 population

`age` percent of housing units built before 1939

`volact` new homeowner policies plus renewals minus cancellations and non renewals per 100 housing units

`involact` new FAIR plan policies and renewals per 100 housing units

income median family income

The data comes from the book by Andrews and Herzberg (1985). We choose the involuntary market activity variable (the number getting FAIR plan insurance) as the response since this seems to be the best measure of those who are denied insurance by others. It is not a perfect measure because some who are denied insurance may give up and others still may not try at all for that reason. The voluntary market activity variable is not as relevant.

Furthermore, we do not know the race of those denied insurance. We only know the racial composition in the corresponding zip code. This is an important difficulty and brings up the following topic:

Ecological Correlation

When data is collected at the group level, we may observe a correlation between two variables. The ecological fallacy is concluding that the same correlation holds at the individual level. For example, in countries with higher fat intakes in the diet, higher rates of breast cancer have been observed. Does this imply that individuals with high fat intakes are at a higher risk of breast cancer? Not necessarily. Relationships seen in observational data are subject to confounding but even if this is allowed for, bias is caused by aggregating data. We consider an example taken from US demographic data:

```
> data(eco)
> plot(income ~ usborn, data=eco, xlab="Proportion US born",
  ylab="Mean Annual Income")
```

In the first panel of Figure 12.1, we see the relationship between 1998 per capita income dollars from all sources and the proportion of legal state residents born in the United States in 1990 for each of the 50 states plus the District of Columbia. We can see a clear negative correlation.

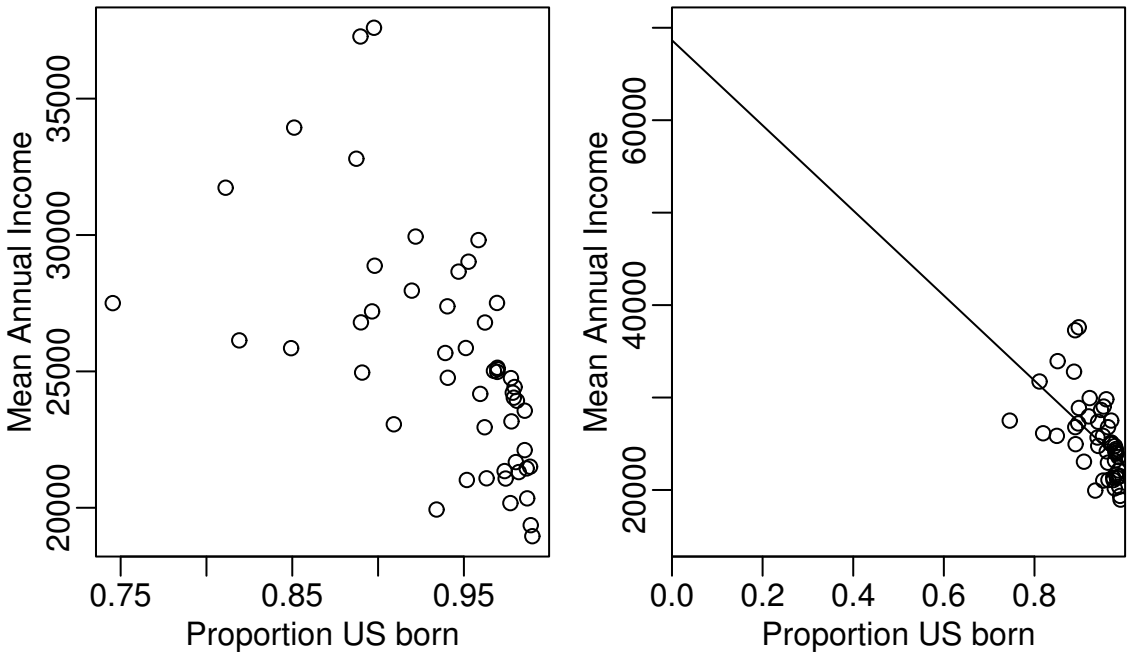

Figure 12.1: 1998 annual per capita income and proportion US born for 50 states plus DC. Plot on the right is the same data as on the left but with an extended scale and the least squares fit shown

We can fit a regression line and show the fitted line on an extended range:

```
> g <- lm(income ~ usborn, eco)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    68642       8739    7.85  3.2e-10
usborn        -46019       9279   -4.96  8.9e-06

Residual standard error: 3490 on 49 degrees of freedom
Multiple R-Squared: 0.334,      Adjusted R-squared: 0.321
F-statistic: 24.6 on 1 and 49 degrees of freedom, p-value: 8.89e-06

> plot(income ~ usborn, data=eco, xlab="Proportion US born",
  ylab="Mean Annual Income",xlim=c(0,1),ylim=c(15000,70000),xaxs="i")
> abline(g$coef)
```

We see that there is a clear statistical significant relationship between per capita annual income and the proportion who are US born. What does this say about the average annual income of people who are US born and those who are naturalized citizens? If we substitute, `usborn=1` into the regression equation, we get 68642-46019=$22,623, while if we put `usborn=0`, we get $68,642. This suggests that on average, naturalized citizens are three times wealthier than US born citizens. In truth, information US Bureau of the Census indicates that US born citizens have an average income just slightly larger than naturalized citizens. What went wrong with our analysis?

The ecological inference from the aggregate data to the individuals requires an assumption of constancy. Explicitly, the assumption would be that the incomes of the native-born do not depend on the proportion of native born within the state (and similarly for naturalized citizens). This assumption is unreasonable for this data because immigrants are naturally attracted to wealthier states.

This is also relevent to the analysis of the Chicago insurance data since we have only aggregate data. We must keep in mind that the results for the aggregated data may not hold true at the individual level.

We will focus on the relationship between race and the response although similar analyses might be done for the income variable.

Start by reading the data in and examining it:

```
> data(chicago)
> chicago
      race fire theft  age volact involact income
60626 10.0  6.2    29 60.4    5.3      0.0  11744
60640 22.2  9.5    44 76.5    3.1      0.1   9323
etc.
60645  3.1  4.9    27 46.6   10.9      0.0  13731
```

Rescale the income variable and omit volact

```
> ch <- data.frame(chicago[,1:4],involact=chicago[,6],income=chicago[,7]/1000)
> ch
      race fire theft  age involact income
60626 10.0  6.2    29 60.4      0.0 11.744
60640 22.2  9.5    44 76.5      0.1  9.323
etc.
60645  3.1  4.9    27 46.6     0.0       13.731
```

Summarize:

```
> summary(ch)
      race             fire           theft             age
 Min.   : 1.00   Min.   : 2.00   Min.   :  3.0   Min.   : 2.0
 1st Qu.: 3.75   1st Qu.: 5.65   1st Qu.: 22.0   1st Qu.:48.6
 Median :24.50   Median :10.40   Median : 29.0   Median :65.0
 Mean   :35.00   Mean   :12.30   Mean   : 32.4   Mean   :60.3
 3rd Qu.:57.60   3rd Qu.:16.10   3rd Qu.: 38.0   3rd Qu.:77.3
 Max.   :99.70   Max.   :39.70   Max.   :147.0   Max.   :90.1
    involact         income
 Min.   :0.000   Min.   : 5.58
 1st Qu.:0.000   1st Qu.: 8.45
 Median :0.400   Median :10.70
 Mean   :0.615   Mean   :10.70
 3rd Qu.:0.900   3rd Qu.:12.00
 Max.   :2.200   Max.   :21.50
```

We see that there is a wide range in the `race` variable with some zip codes being almost entirely minority or non-minority. This is good for our analysis since it will reduce the variation in the regression coefficient for race, allowing us to assess this effect more accurately. If all the zip codes were homogenous, we would never be able to discover an effect from this aggregated data. We also note some skewness in the `theft` and `income` variables. The response `involact` has a large number of zeroes. This is not good for the assumptions of the linear model but we have little choice but to proceed.

Now make some graphical summaries:

```
> par(mfrow=c(2,3))
> for(i in 1:6) hist(ch[,i],main=names(ch)[i])
> for(i in 1:6) boxplot(ch[,i],main=names(ch)[i])
> pairs(ch)
```

Only the boxplots are shown in Figure 12.

An examination of the data using xgobi would also be worthwhile.

Now look at the relationship between involact and race:

```
> summary(lm(involact ~ race,data=ch))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12922    0.09661    1.34     0.19
race         0.01388    0.00203    6.84  1.8e-08

Residual standard error: 0.449 on 45 degrees of freedom
Multiple R-Squared: 0.509,      Adjusted R-squared: 0.499
F-statistic: 46.7 on 1 and 45 degrees of freedom,       p-value: 1.78e-08
```

We can clearly see that homeowners in zip codes with high % minority are being denied insurance at higher rate than other zip codes. That is not in doubt. However, can the insurance companies claim that the discrepancy is due to greater risks in some zip-codes? For example, we see that % minority is correlated

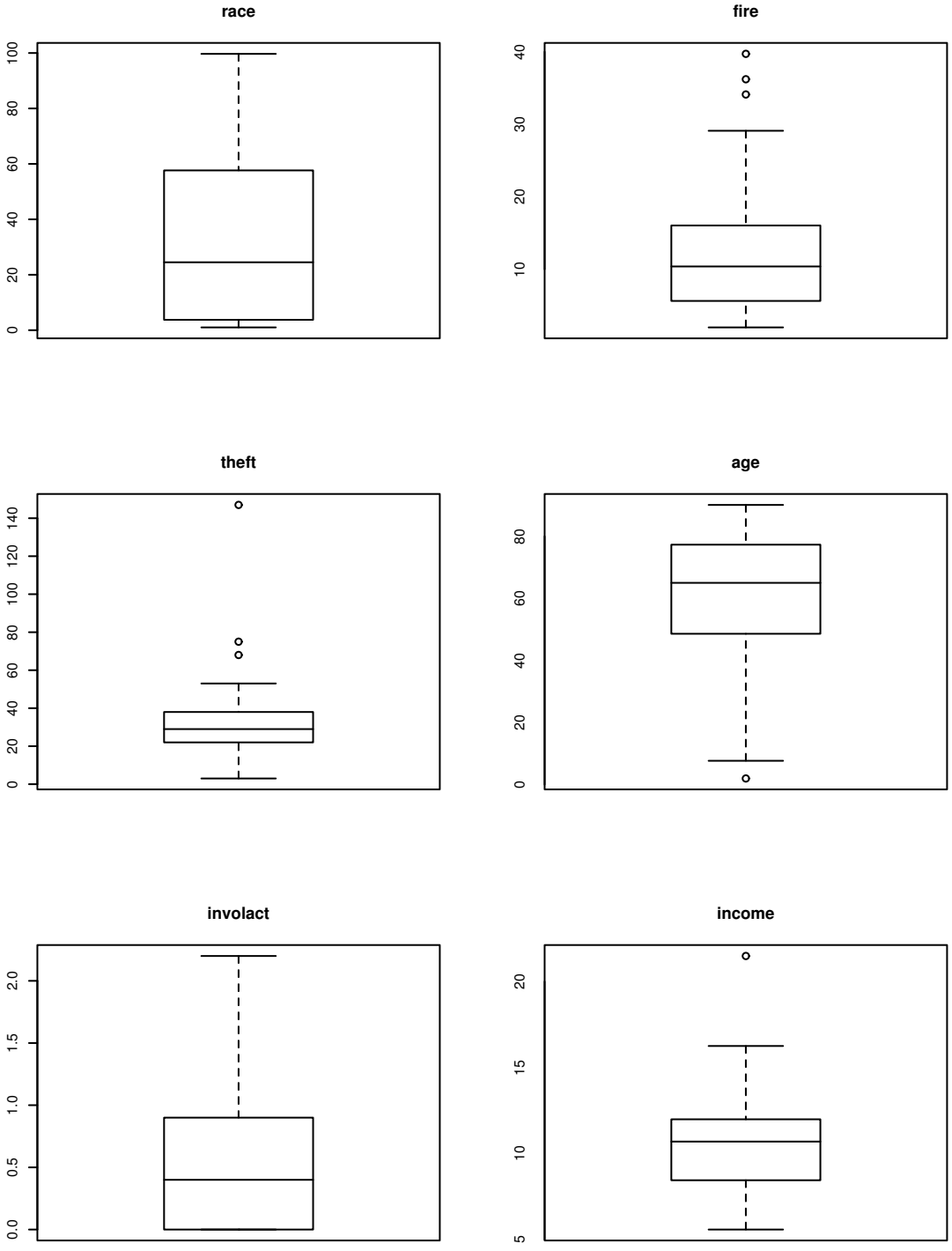

Figure 12.2: Boxplots of the Chicago Insurance data

with the fire rate from the plots. The insurance companies could say that they were denying insurance in neighborhoods where they had sustained large fire-related losses and any discriminatory effect was a by-product of (presumably) legitimate business practice. What can regression analysis tell us about this claim?

The question of which variables should also be included in the regression so that their effect may be adjusted for is difficult. Statistically, we can do it, but the important question is whether it should be done at all. For example, it is known that the incomes of women in the US are generally lower than those of men. However, if one adjusts for various factors such as type of job and length of service, this gender difference is reduced or can even disappear. The controversy is not statistical but political - should these factors be used to make the adjustment?

In this example, suppose that if the effect of adjusting for income differences was to remove the race effect? This would pose an interesting but non-statistical question. I have chosen to include the `income` variable here just to see what happens.

I use log(income) partly because of skewness in this variable but also because income is better considered on a multiplicative rather than additive scale. In other words, $1,000 is worth a lot more to a poor person than a millionaire because $1,000 is a much greater fraction of the poor person's wealth.

We start with the full model:

```
> g <- lm(involact ~ race + fire + theft + age + log(income), data = ch)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.18554    1.10025   -1.08  0.28755
race         0.00950    0.00249    3.82  0.00045
fire         0.03986    0.00877    4.55  4.8e-05
theft       -0.01029    0.00282   -3.65  0.00073
age          0.00834    0.00274    3.04  0.00413
log(income)  0.34576    0.40012    0.86  0.39254

Residual standard error: 0.335 on 41 degrees of freedom
Multiple R-Squared: 0.752,      Adjusted R-squared: 0.721
F-statistic: 24.8 on 5 and 41 degrees of freedom,       p-value: 2.01e-11
```

Before we start making any conclusions, we should check the model assumptions.

```
> plot(g$fit,g$res,xlab="Fitted",ylab="Residuals",
  main="Residual-Fitted plot")
> abline(h=0)
> qqnorm(g$res)
```

These two diagnostic plots are shown in Figure 12.

The diagonal streak in the residual-fitted plot is caused by the large number of zero response values in the data. When $y =$, the residual $\hat{\varepsilon} = \hat{y} = -x^T\hat{\beta}$, hence the line. Turning a blind eye to this feature, we see no particular problem. The Q-Q plot looks fine too.

Now let's look at influence - what happens if points are excluded? We'll use a function `qqnorml()` that I wrote that labels the points in a Q-Q plot with the case numbers Plot not shown but cases 6 and 24 seem to stick out.

```
> gi <- lm.influence(g)
> for(i in 1:5) qqnorml(gi$coef[,i+1],main=names(ch)[-5][i])
```

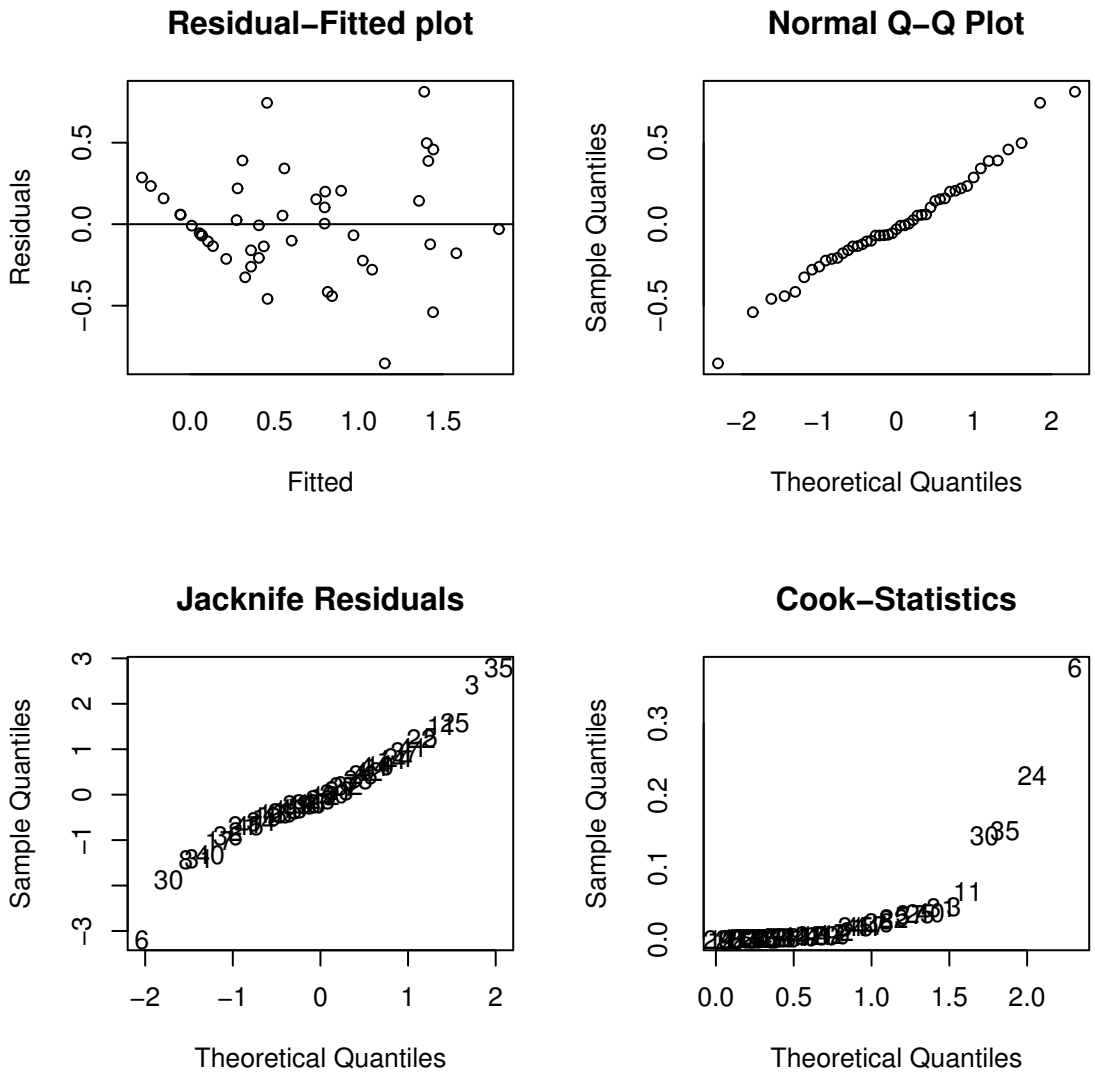

Figure 12.3: Diagnostic plots of the Chicago Insurance data

Check out the jacknife residuals:

```
> qqnorml(rstudent(g),main="Jacknife Residuals")
> qt(0.05/(2*47),47-6-1)
[1] -3.529468
```

Nothing too extreme - now look at the Cook statistics using the `halfnorm()` function that I wrote:

```
> halfnorm(cooks.distance(g),main="Cook-Statistics")
```

Cases 6 and 24 stick out again. Let's take a look at these two cases:

```
> ch[c(6,24),]
      race fire theft  age involact income
```

```
60610 54.0 34.1    68 52.6      0.3  8.231
60607 50.2 39.7   147 83.0      0.9  7.459
```

These are high theft and fire zip codes. See what happens when we exclude these points:

```
> g <- lm(involact ~ race + fire + theft + age + log(income),ch,
              subset=(1:47)[-c(6,24)])
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.57674    1.08005   -0.53    0.596
race         0.00705    0.00270    2.62    0.013
fire         0.04965    0.00857    5.79    1e-06
theft       -0.00643    0.00435   -1.48    0.147
age          0.00517    0.00289    1.79    0.082
log(income)  0.11570    0.40111    0.29    0.775

Residual standard error: 0.303 on 39 degrees of freedom
Multiple R-Squared: 0.804,      Adjusted R-squared: 0.779
F-statistic:   32 on 5 and 39 degrees of freedom,       p-value: 8.19e-13
```

`theft` and `age` are no longer significant at the 5% level. We now address the question of transformations - because the response has some zero values and for interpretational reasons we will not try to transform it. Similarly, since the race variable is the primary predictor of interest we won't try transforming it either so as to avoid interpretation difficulties. We try fitting a polynomial model with quadratic terms in each of the predictors:

```
> g2 <- lm(involact ~ race + poly(fire,2) + poly(theft,2) + poly(age,2)
       + poly(log(income),2), ch, subset=(1:47)[-c(6,24)])
> anova(g,g2)
Analysis of Variance Table

Model 1: involact ~ race + fire + theft + age + log(income)
Model 2: involact ~ race + poly(fire, 2) + poly(theft, 2) + poly(age,
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     39       3.59
2     35       3.20  4   0.38    1.04    0.4
```

Seems that we can do without the quadratic terms. A check of the partial residual plots also reveals no need to transform. We now move on to variable selection. We are not so much interested in picking one model here because we are mostly interested in the dependency of involact on the race variable. So $\hat{\beta}_1$ is the thing we want to focus on. The problem is that collinearity with the other variables may cause $\hat{\beta}_1$ to vary substantially depending on what other variables are in the model. We address this question here. `leaps()` is bit picky about its input format so I need to form the x and y explicitly:

```
> y <- ch$inv[cooks.distance(g) < 0.2]
> x <- cbind(ch[,1:4],linc=log(ch[,6]))
> x <- x[cooks.distance(g) < 0.2,]
```

Removing all points with Cook's Statistics greater than 0.2 takes out cases 6 and 24.
We make the Cp plot.

```
> library(leaps)
> a <- leaps(x,y)
> Cpplot(a)
```

See Figure 12.

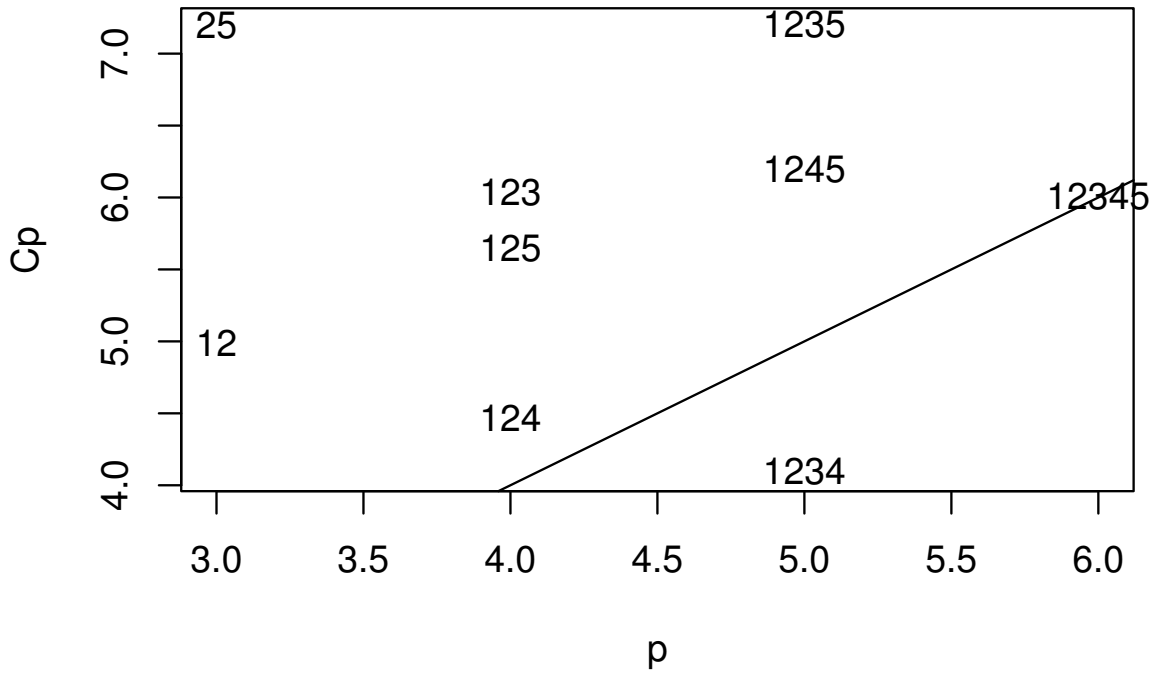

Figure 12.4: Cp plot of the Chicago Insurance data

The best model seems to be this one:

```
> g <- lm(involact ~ race + fire + theft + age, ch, subset=(1:47)[-c(6,24)])
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26787    0.13967   -1.92   0.0623
race         0.00649    0.00184    3.53   0.0011
fire         0.04906    0.00823    5.96  5.3e-07
theft       -0.00581    0.00373   -1.56   0.1271
age          0.00469    0.00233    2.01   0.0514

Residual standard error: 0.3 on 40 degrees of freedom
Multiple R-Squared: 0.804,      Adjusted R-squared: 0.784
F-statistic: 40.9 on 4 and 40 degrees of freedom,       p-value: 1.24e-13
```

The fire rate is also significant and actually has higher t-statistics. Thus, we have verified that there is a positive relationship between involact and race while controlling for a selection of the other variables.

How robust is the conclusion? Would other analysts have come to the same conclusion? One alternative model is

```
> galt <- lm(involact ~ race+fire+log(income),ch,subset=(1:47)[-c(6,24)])
> summary(galt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.75326    0.83588    0.90    0.373
race         0.00421    0.00228    1.85    0.072
fire         0.05102    0.00845    6.04  3.8e-07
log(income) -0.36238    0.31916   -1.14    0.263

Residual standard error: 0.309 on 41 degrees of freedom
Multiple R-Squared: 0.786,      Adjusted R-squared: 0.77
F-statistic: 50.1 on 3 and 41 degrees of freedom,       p-value: 8.87e-14
```

In this model, we see that `race` is not statistically significant. The previous model did fit slightly better but it is important that there exists a reasonable model in which `race` is not significant since although the evidence seems fairly strong in favor of a `race` effect, it is not entirely conclusive. Interestingly enough, if `log(income)` is dropped:

```
> galt <- lm(involact ~ race+fire,ch,subset=(1:47)[-c(6,24)])
> summary(galt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.19132    0.08152   -2.35   0.0237
race         0.00571    0.00186    3.08   0.0037
fire         0.05466    0.00784    6.97  1.6e-08

Residual standard error: 0.31 on 42 degrees of freedom
Multiple R-Squared: 0.779,      Adjusted R-squared: 0.769
F-statistic: 74.1 on 2 and 42 degrees of freedom,       p-value: 1.7e-14
```

we find `race` again becomes significant which raises again the question of whether income should be adjusted for since it makes all the difference here.

We now return to the two left-out cases. Observe the difference in the fit when the two are re-included on the best model. The quantities may change but the qualitative message is the same. It is better to include all points if possible, especially in a legal case like this where excluding points might lead to criticism and suspicion of the results.

```
> g <- lm(involact ~ race + fire + theft + age, data=ch)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24312    0.14505   -1.68  0.10116
race         0.00810    0.00189    4.30  0.00010
fire         0.03665    0.00792    4.63  3.5e-05
theft       -0.00959    0.00269   -3.57  0.00092
age          0.00721    0.00241    2.99  0.00460
```

```
Residual standard error: 0.334 on 42 degrees of freedom
Multiple R-Squared: 0.747,      Adjusted R-squared: 0.723
F-statistic:   31 on 4 and 42 degrees of freedom,       p-value: 4.8e-12
```

The main message of the data is not changed - we should check the diagnostics. I found no trouble. (Adding back in the two points to the `race+fire+log(income)` model made `race` significant again. So it looks like there is some good evidence that zip codes with high minority populations are being "redlined" - that is improperly denied insurance. While there is evidence that some of the relationship between race and involact can be explained by the fire rate, there is still a component that cannot be attributed to the other variables.

However, there is some doubt due to the response not being a perfect measure of people being denied insurance. It is an aggregate measure which raises the problem of ecological correlations. We have implicitly assumed that the probability that a minority homeowner would obtain a FAIR plan after adjusting for the effect of the other covariates is constant across zip-codes. This is unlikely to be true. If the truth is simply variation about some constant, then our conclusions will still be reasonable but if this probability varies in a systematic way, then our conclusions may be off the mark. It would be a very good idea to obtain some individual level data.

Another point to be considered is the size of the effect. The largest value of the response is only 2.2% and most cases are much smaller. Even assuming the worst, the number of people affected is small.

There is also the problem of a potential latent variable that might be the true cause of the observed relationship, but it is difficult to see what that variable might be. Nevertheless, this always casts a shadow of doubt on our conclusions.

There are some special difficulties in presenting this during a court case. With scientific enquiries, there is always room for uncertainty and subtlety in presenting the results, but this is much more difficult in the court room. The jury may know no statistics and lawyers are clever at twisting words. A statistician giving evidence as an expert witness would do well to keep the message simple.

Another issue that arises in cases of this nature is how much the data should be aggregated. For example, I divided the data using a zip code map of Chicago into north and south. Fit the model to the south of Chicago:

```
> data(chiczip)
> g <- lm(involact ~ race + fire + theft +age, subset=(chiczip == "s"), ch)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23441    0.23774   -0.99    0.338
race         0.00595    0.00328    1.81    0.087
fire         0.04839    0.01689    2.87    0.011
theft       -0.00664    0.00844   -0.79    0.442
age          0.00501    0.00505    0.99    0.335

Residual standard error: 0.351 on 17 degrees of freedom
Multiple R-Squared: 0.743,      Adjusted R-squared: 0.683
F-statistic: 12.3 on 4 and 17 degrees of freedom,       p-value: 6.97e-05
```

and now to the north.

```
> g <- lm(involact ~ race + fire + theft +age, subset=(chiczip == "n"), ch)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31857    0.22702   -1.40    0.176
race         0.01256    0.00448    2.81    0.011
fire         0.02313    0.01398    1.65    0.114
theft       -0.00758    0.00366   -2.07    0.052
age          0.00820    0.00346    2.37    0.028

Residual standard error: 0.343 on 20 degrees of freedom
Multiple R-Squared: 0.756,      Adjusted R-squared: 0.707
F-statistic: 15.5 on 4 and 20 degrees of freedom,       p-value: 6.52e-06
```

What differences do you see? By dividing the data into smaller and smaller subsets it is possible to dilute the significance of any predictor. On the other hand it is important not to aggregate all data without regard to whether it is reasonable. Clearly a judgment has to be made and this often a point of contention in legal cases.

After all this analysis, the reader may be feeling somewhat dissatisfied. It seems we are unable to come to any truly definite conclusions and everything we say has been hedged with "ifs" and "buts". Winston Churchill once said

> Indeed, it has been said that democracy is the worst form of Government except all those other forms that have been tried from time to time.

We might say the same thing about Statistics in relation to how it helps us reason in the face of uncertainty.