Chapter 12

Insurance Redlining — A Complete Example

In this chapter, we present a relatively complete data analysis. The example is interesting because it illustrates several of the ambiguities and difficulties encountered in statistical practice.

Insurance redlining refers to the practice of refusing to issue insurance to certain types of people or within some geographic area. The name comes from the act of drawing a red line around an area on a map. Now few would quibble with an insurance company refusing to sell auto insurance to a frequent drunk driver, but other forms of discrimination would be unacceptable.

In the late 1970s, the US Commission on Civil Rights examined charges by several Chicago community organizations that insurance companies were redlining their neighborhoods. Because comprehensive information about individuals being refused homeowners insurance was not available, the number of FAIR plan policies written and renewed in Chicago by zip code for the months of December 1977 through May 1978 was recorded. The FAIR plan was offered by the city of Chicago as a default policy to homeowners who had been rejected by the voluntary market. Information on other variables that might affect insurance writing such as fire and theft rates was also collected at the zip code level. The variables are:

`race` racial composition in percentage of minority

`fire` fires per 100 housing units

`theft` thefts per 1000 population

`age` percentage of housing units built before 1939

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

`income` median family income in thousands of dollars

`side` north or south side of Chicago

The data come from Andrews and Herzberg (1985) where more details of the variables and the background are provided.

12.1 Ecological Correlation

Notice that we do not know the races of those denied insurance. We only know the racial composition in the corresponding zip code. This is an important difficulty that needs to be considered before starting the analysis.

When data are 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, package="faraway")
> 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 (D.C.). We can see a clear negative correlation. We can fit a regression line and show the

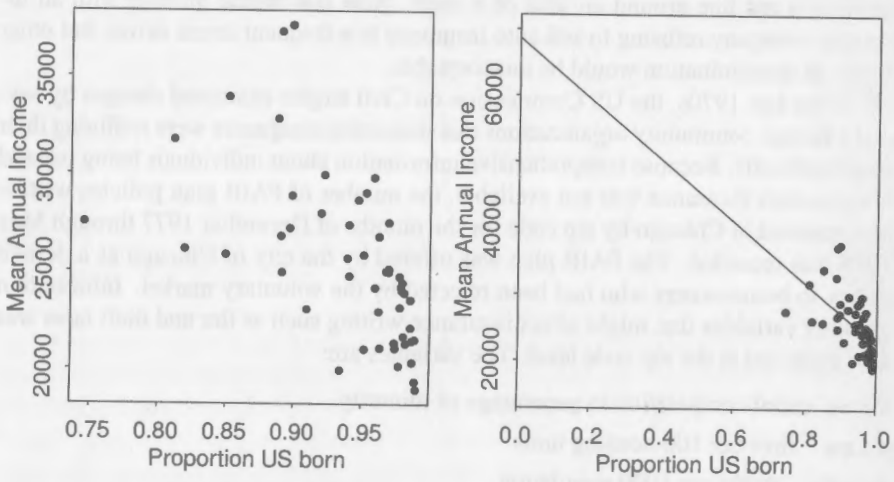

Figure 12.1 1998 annual per capita income and proportion US born for 50 states plus D.C. The plot on the right shows the same data as on the left, but with an extended scale and the least squares fit shown.

fitted line on an extended range:

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

n = 51, p = 2, Residual SE = 3489.541, R-Squared = 0.33
> plot(income ~ usborn, data=eco, xlab="Proportion US born", ylab="
    Mean Annual Income",xlim=c(0,1),ylim=c(15000,70000),xaxs="i")
> abline(coef(lmod))
```

We see that there is a clear statistically significant relationship between the 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 earn three times more than US born citizens. In truth, information from the 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 these data because immigrants are naturally attracted to wealthier states.

This assumption is also relevant 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.

12.2 Initial Data Analysis

Start by reading the data in and examining it:

```
> data(chredlin, package="faraway")
> head(chredlin)
      race fire theft  age involact income side
60626 10.0  6.2    29 60.4      0.0 11.744    n
60640 22.2  9.5    44 76.5      0.1  9.323    n
60613 19.6 10.5    36 73.5      1.2  9.948    n
60657 17.3  7.7    37 66.9      0.5 10.656    n
60614 24.5  8.6    53 81.4      0.7  9.730    n
60610 54.0 34.1    68 52.6      0.3  8.231    n
```

Summarize:

```
> summary(chredlin)
      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   :34.99   Mean   :12.28   Mean   : 32.4   Mean   :60.3
 3rd Qu.:57.65   3rd Qu.:16.05   3rd Qu.: 38.0   3rd Qu.:77.3
 Max.   :99.70   Max.   :39.70   Max.   :147.0   Max.   :90.1
    involact         income       side
 Min.   :0.000   Min.   : 5.58   n:25
 1st Qu.:0.000   1st Qu.: 8.45   s:22
 Median :0.400   Median :10.69
 Mean   :0.615   Mean   :10.70
 3rd Qu.:0.900   3rd Qu.:11.99
 Max.   :2.200   Max.   :21.48
```

We see that there is a wide range in the `race` variable, with some zip codes 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 homogeneous, we would never be able to discover an effect from these aggregated data. We also note some skewness in the theft and income variables. The response involact has a large number of zeros. This is not good for the assumptions of the linear model but we have little choice but to proceed. We will not use the information about north versus south side until later. Now make some graphical summaries:

```
> require(ggplot2)
> ggplot(chredlin,aes(race,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(fire,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(theft,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(age,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(income,involact)) + geom_point() +stat_smooth(
    method="lm")
> ggplot(chredlin,aes(side,involact)) + geom_point(position = position
    _jitter(width = .2,height=0))
```

The plots are seen in Figure 12.2. We have superimposed a linear fit to each pair of variables with a 95% confidence band shown in grey. Strong relationships can be seen in several of the plots. We can see some outlier and influential points. We can also see that the fitted line sometimes goes below zero which is problematic since observed values of the response cannot be negative. Jittering has been added in the final plot to avoid overplotting of symbols. Let's focus on the relationship between involact and race:

```
> sumary(lm(involact ~ race,chredlin))
            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

n = 47, p = 2, Residual SE = 0.449, R-Squared = 0.51
```

We can clearly see that homeowners in zip codes with a high percentage of minorities are taking the default FAIR plan insurance at a 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? The insurance companies could claim that they were denying insurance in neighborhoods where they had sustained large fire-related losses and any discriminatory effect was a by-product of legitimate business practice. We plot some of the variables involved by this question in Figure 12.3.

```
> ggplot(chredlin,aes(race,fire)) + geom_point() +stat_smooth(method="
    lm")
> ggplot(chredlin,aes(race,theft)) + geom_point() +stat_smooth(method
    ="lm")
```

We can see that there is indeed a relationship between the fire rate and the percentage of minorities. We also see that there is large outlier that may have a disproportionate effect on the relationship between the theft rate and the percentage of minorities.

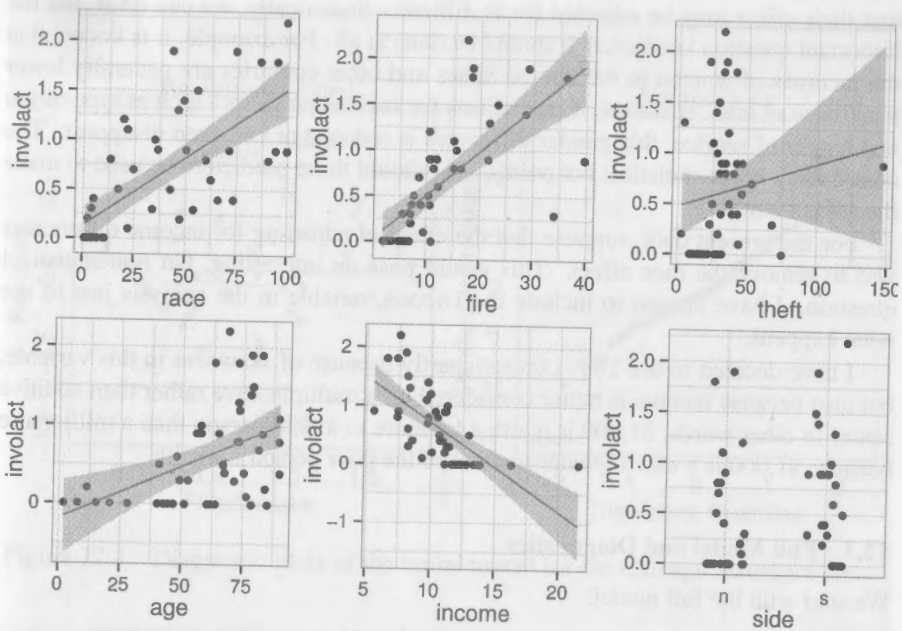

Figure 12.2 Plots of the Chicago insurance data.

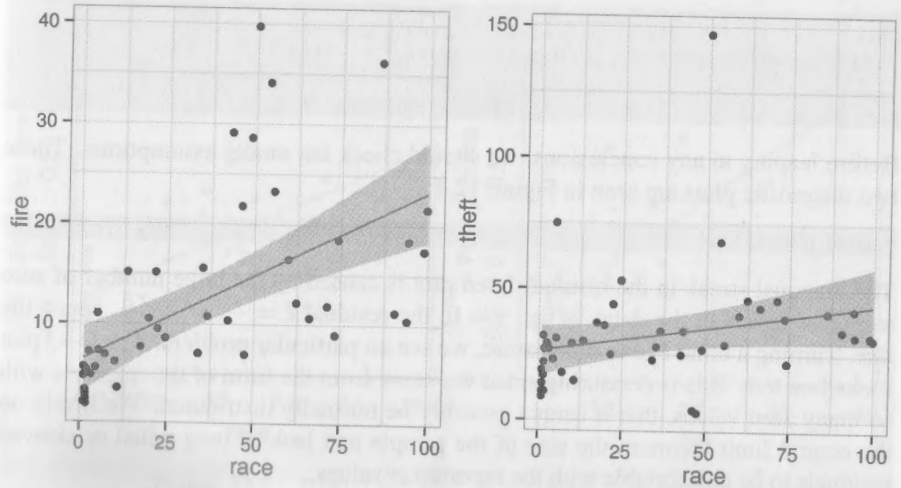

Figure 12.3 Relationship between fire, theft and race in the Chicago data.

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 United States and other countries are generally lower than those of men. However, if one adjusts for various predictors 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 predictors be used to make the adjustment?

For the present data, suppose that 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 in the analysis just to see what happens.

I have decided to 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.

12.3 Full Model and Diagnostics

We start with the full model:

```
> lmod <- lm(involact ~ race + fire + theft + age + log(income),
    chredlin)
> sumary(lmod)
             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 0.000048
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

n = 47, p = 6, Residual SE = 0.335, R-Squared = 0.75
```

Before leaping to any conclusions, we should check the model assumptions. These two diagnostic plots are seen in Figure 12.4:

```
> plot(lmod,1:2)
```

The diagonal streak in the residual-fitted plot is caused by the large number of zero response values in the data. When $y = 0$, 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. This is reassuring since we know from the form of the response with so many zero values, that it cannot possibly be normally distributed. We'll rely on the central limit theorem, the size of the sample and lack of long-tailed or skewed residuals to be comfortable with the reported p-values.

We now look for transformations. We try some partial residual plots as seen in Figure 12.5:

```
> termplot(lmod, partial.resid=TRUE, terms=1:2)
```


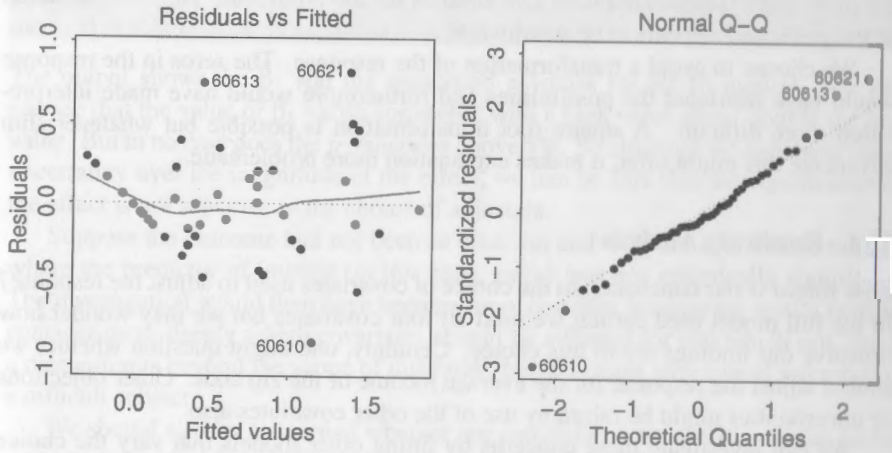

Figure 12.4 Diagnostic plots of the initial model for the Chicago insurance data.

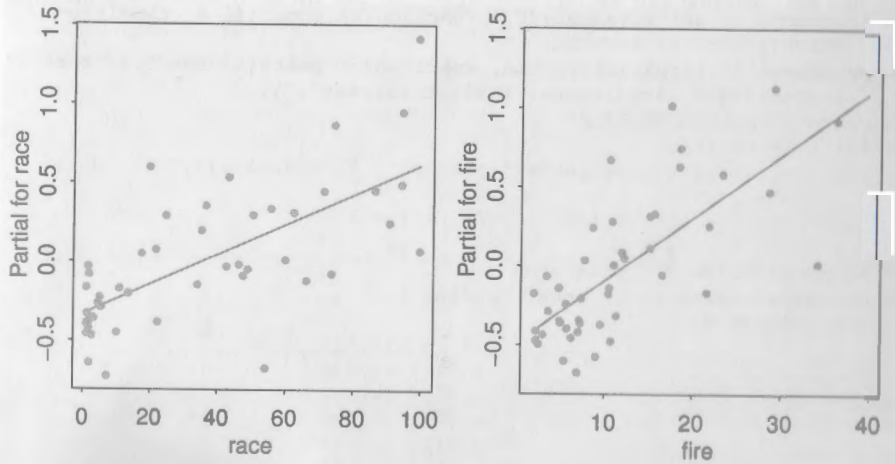

Figure 12.5 Partial residual plots for `race` and `fire`.

These plots indicate no need to transform. It would have been inconvenient to transform the race variable since that would have made interpretation more difficult. Fortunately, we do not need to worry about this. We examined the other partial residual plots and experimented with polynomials for the predictors. No transformation of the predictors appears to be worthwhile.

We choose to avoid a transformation of the response. The zeros in the response would have restricted the possibilities and furthermore would have made interpretation more difficult. A square root transformation is possible but whatever slim advantage this might offer, it makes explanation more problematic.

12.4 Sensitivity Analysis

How robust is our conclusion to the choice of covariates used to adjust the response? In the full model used earlier, we used all four covariates but we may wonder how sensitive our findings are to this choice. Certainly, one might question whether we should adjust the response for the average income of the zip code. Other objections or uncertainties might be raised by use of the other covariates also.

We can investigate these concerns by fitting other models that vary the choice of adjusting covariates. In this example, there are four such covariates and so there are only 16 possible combinations in which they may be added to the model. It is practical to fit and examine all these models.

The mechanism for creating all 16 models is rather complex and you may wish to skip to the output. The first line creates all subsets of (1,2,3,4). The second line creates the predictor part of the model formulae by pasting together the chosen variables. We then iterate over all 16 models, saving the terms of interest for the race variable:

```
> listcombo <- unlist(sapply(0:4,function(x) combn(4, x, simplify=
    FALSE)),recursive=FALSE)
> predterms <- lapply(listcombo, function(x) paste(c("race",c("fire","
    theft","age","log(income)")[x]),collapse="+"))
> coefm <- matrix(NA,16,2)
> for(i in 1:16){
   lmi <- lm(as.formula(paste("involact ~ ",predterms[[i]])), data=
       chredlin)
   coefm[i,] <- summary(lmi)$coef[2,c(1,4)]
}
> rownames(coefm) <- predterms
> colnames(coefm) <- c("beta","pvalue")
> round(coefm,4)
                                beta pvalue
race                          0.0139 0.0000
race+fire                     0.0089 0.0002
race+theft                    0.0141 0.0000
race+age                      0.0123 0.0000
race+log(income)              0.0082 0.0087
race+fire+theft               0.0082 0.0002
race+fire+age                 0.0089 0.0001
race+fire+log(income)         0.0070 0.0160
race+theft+age                0.0128 0.0000
race+theft+log(income)        0.0084 0.0083
race+age+log(income)          0.0099 0.0017
race+fire+theft+age           0.0081 0.0001
race+fire+theft+log(income)   0.0073 0.0078
race+fire+age+log(income)     0.0085 0.0041
race+theft+age+log(income)    0.0106 0.0010
race+fire+theft+age+log(income) 0.0095 0.0004
```


The output shows the $\hat{\beta}_1$ and the associated p-values for all 16 models. We can see see that the value of $\hat{\beta}_1$ varies somewhat with a high value about double the low value. But in no case does the p-value rise above 5%. So although we may have some uncertainty over the magnitude of the effect, we can be sure that the significance of the effect is not sensitive to the choice of adjusters.

Suppose the outcome had not been so clear cut and we were able to find models where the predictor of interest (in this case, race) was not statistically significant. The investigation would then have become more complex because we would need to consider more deeply which covariates should be adjusted for and which not. Such a discussion is beyond the scope of this book, but illustrates why causal inference is a difficult subject.

We should also be concerned whether our conclusions are sensitive to the inclusion or exclusion of a small number of cases. Influence diagnostics are useful for this purpose. We start with a plot of the differences in the coefficient caused by the removal of one point. These can be seen for the race variable in Figure 12.6.

```
> diags <- data.frame(lm.influence(lmod)$coef)
> ggplot(diags,aes(row.names(diags),race)) +geom_linerange(aes(ymax=0,
    ymin=race)) +theme(axis.text.x=element_text(angle=90)) + xlab("
    ZIP") +geom_hline(yint=0)
```

The ggplot function requires the data in the form of a data frame. We extract the relevant component from the the lm.influence call for this purpose. We can see that the largest reduction is about 0.001 which would be insufficient to change the statistical significance of this term.

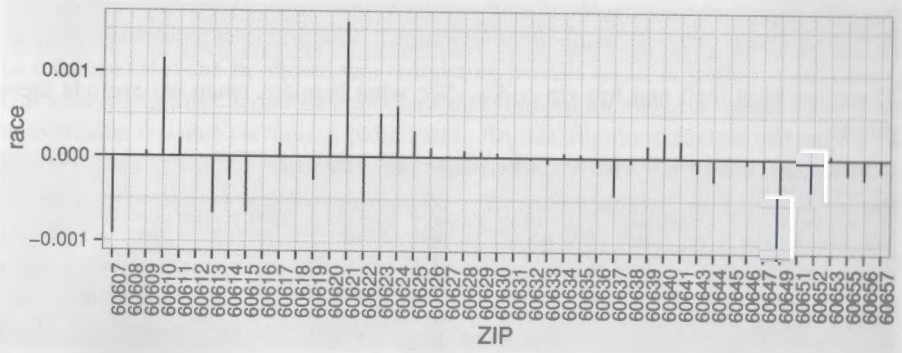

Figure 12.6 Leave-out-one change in coefficient values for $\hat{\beta}_{race}$.

It is also worth considering the influence on the adjustment covariates We plot the leave-out-one differences in $\hat{\beta}$ for theft and fire:

```
> ggplot(diags,aes(x=fire,y=theft))+geom_text(label=row.names(diags))
```

Let's also take a look the standardized residuals and leverage which can be conveniently constructed using the default `plot` function for a linear model object:

```
> plot(lmod,5)
```

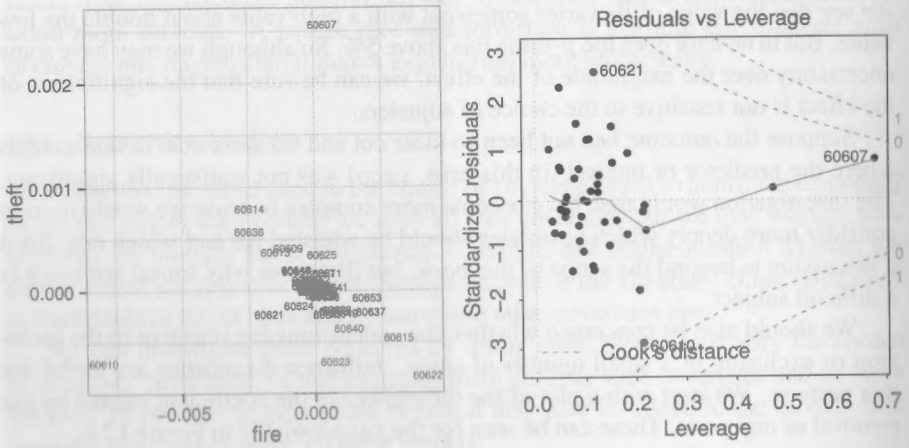

Figure 12.7 Plot of the leave-out one coefficient differences is shown on the left. Plot of the standardized residuals against the leverages is shown on the right

See Figure 12.7 where zip codes 60607 and 60610 stick out. It is worth looking at other leave-out-one coefficient plots also. We also notice that there is no standardized residual extreme enough to call an outlier. Let's take a look at the two cases:

```
> chredlin[c("60607","60610"),]
      race fire theft  age involact income side
60607 50.2 39.7   147 83.0      0.9  7.459    n
60610 54.0 34.1    68 52.6      0.3  8.231    n
```

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

```
> match(c("60607", "60610"),row.names(chredlin))
[1] 24  6
> lmode <- lm(involact ~ race + fire + theft + age + log(income),
    chredlin,subset=-c(6,24))
> sumary(lmode)
             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

n = 45, p = 6, Residual SE = 0.303, R-Squared = 0.8
```


The predictors `theft` and `age` are no longer significant at the 5% level. The coefficient for `race` is reduced compared to the full data fit but remains statistically significant.

So we have verified that our conclusions are also robust to the exclusion of one or perhaps two cases from the data. This is reassuring since a conclusion based on the accuracy of measurement for a single case would be a cause for concern. If this problem did occur, we would need to be particularly sure of these measurements. In some situations, one might wish to drop such influential cases but this would require strong arguments that such points were in some way exceptional. In any case, it would be very important to disclose this choice in the analysis.

Now if we try very hard to poke a hole in our result, we can find this model where two cases have been dropped:

```
> modalt <- lm(involact ~ race+fire+log(income),chredlin,  subset=-c
    (6,24))
> sumary(modalt)
             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

n = 45, p = 4, Residual SE = 0.309, R-Squared = 0.79
```

In this model, `race` no longer meets the threshold for significance. However, there is no compelling reason to advocate for this model against the large weight of other alternatives we have considered.

This illustrates a wider problem with regression modeling in that the data usually do not unequivocally suggest one particular model. It is easy for independent analysts to apply similar methods but in different orders and in somewhat different ways resulting in different model choices. See Faraway (1994) for some examples. For this reason, the good analyst explores the data thoroughly and considers multiple models. One might settle on one final model but confidence in the conclusions will be enhanced if it can be shown that competing models result in similar conclusions. Our analysis in this chapter demonstrates this concern for alternatives but there is an unavoidable reliance on human judgement. An unscrupulous analyst can explore a large number of models but report only the one that favors a particular conclusion.

A related concept is *model uncertainty*. We surely do not know the true model for this data and somehow our conclusions should reflect this. The regresssion summary outputs provide standard errors and p-values that express our uncertainty about the parameters of the model but they do not reflect the uncertainty about the model itself. This means that we will tend to be more confident about our inferences than is justified. There are several possible ways to mitigate this problem. One simple approach is data splitting as used in the running example on the `meatspec` data in Chapter 11. Another idea is to bootstrap the whole data analysis as demonstrated by Faraway (1992). Alternatively, it may be possible to use *model averaging* as in Raftery, Madigan, and Hoeting (1997).

12.5 Discussion

There is some ambiguity in the conclusion here. These reservations have several sources. There is some doubt because the response is not a perfect measure of people being denied insurance. It is an aggregate measure that raises the problem of ecological correlations. We have implicitly assumed that the probability 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 a 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.

We have demonstrated statistical significance for the effect of race on the response. But statistical significance is not the same as practical significance. The largest value of the response is only 2.2% and most other values are much smaller. Using our preferred models, the predicted difference between 0% minority and 100% minority is about 1%. So while we may be confident that some people are affected, there may not be so many of them. We would need to know more about predictors like insurance renewal rates to say much more but the general point is that the size of the p-value does not tell you much about the practical size of the effect.

There is also the problem of a potential latent variable that might be the true cause of the observed relationship. Someone with first-hand knowledge of the insurance business might propose one. This possibility always casts a shadow of doubt on our conclusions.

Another issue that arises in cases of this nature is how much the data should be aggregated. For example, suppose we fit separate models to the two halves of the city. Fit the model to the south of Chicago:

```
> lmod <- lm(involact ~ race+fire+theft+age, subset=(side == "s"),
    chredlin)
> sumary(lmod)
            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

n = 22, p = 5, Residual SE = 0.351, R-Squared = 0.74
```

and now to the north:

```
> lmod <- lm(involact ~ race+fire+theft+age, subset=(side == "n"),
    chredlin)
> sumary(lmod)
            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

n = 25, p = 5, Residual SE = 0.343, R-Squared = 0.76
```


We see that race is significant in the north, but not in the south. 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 can be a point of contention in legal cases.

There are some special difficulties in presenting this during a court case. With scientific inquiries, there is always room for uncertainty and subtlety in presenting the results, particularly if the subject matter is not contentious. In an adversarial proceeding, it is difficult to present statistical evidence when the outcome is not clear-cut, as in this example. There are particular difficulties in explaining such evidence to non-mathematically trained people.

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 about statistics with respect to how it helps us reason in the face of uncertainty. It is not entirely satisfying but the alternatives are worse.

Exercises

In all the following questions, a full answer requires you to perform a complete analysis of the data including an initial data anaysis, regression diagnostics, a search for possible transformations and a consideration of model selection. A report on your analysis needs to be selective in its content. You should include enough information for the steps leading to your selection of model to be clear and reproducible by the reader. But you should not include everything you tried. Dead ends can be reported in passing but do not need to be described in full detail unless they contain some message of interest. Above all your analysis should have a clear statement of the conclusion of your analysis.

1. Reliable records of temperature taken using thermometers are only available back to the 1850s, but it would be interesting to estimate global temperatures in the pre-industrial era. It is possible to obtain various *proxy* measures of temperature. Trees grow faster in warmer years so the width of tree rings (seen in tree cross-sections) provides some evidence of past temperatures. Other natural sources of proxies include coral and ice cores. Such information can go back for a thousand years or more. The dataset globwarm contains information on eight proxy measures and northern hemisphere temperatures back to 1856. Build a model and predict temperatures back to 1000 AD. State the uncertainty in your predictions. Comment on your findings. (Note: that this data has been modified and simplified for the purposes of this exercise — see the R help page for the data to find out about the source).
2. The happy dataset contains data from 39 MBA students on predictors affecting