# Chapter 11: A Complete Example

Gunwoong Park

Lecture Note

University of Seoul

# Insurance Redlining

Insurance redlining: practice of refusing to issue insurance to certain types of people or within certain geographical areas.

**Question of interest:** Which variables influence denial of insurance? E.g., using fire rates is fine, but using race is illegal.

- Data: Chicago, 1977–1978, $n = 47$, $p = 6$.
- FAIR: offered as a default policy to homeowners who were rejected by the voluntary market.
- Do not have information about individuals. All variables are measured at zip code level.

# Chicago Insurance Data: Variables

- Response: involact: new FAIR plan policies and renewals per 100 housing units
- race: minority percentage
- fire: fires per 100 housing units
- theft: theft per 1000 population
- age: percent of housing units build before 1939
- income: median family income in 1000 dollars
- side: North or South side of Chicago

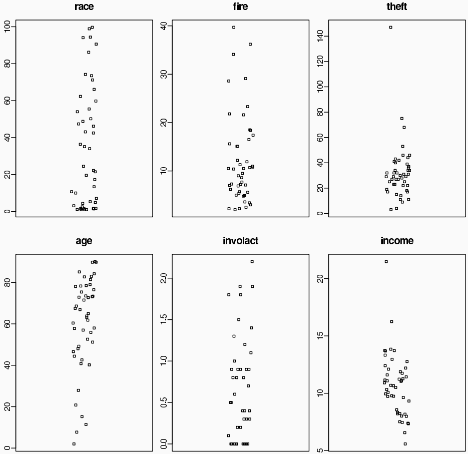
# Initial Data Analysis

```
> library(faraway)
> data(chredlin)
> ## Initial data analysis
> summary(chredlin)
race            fire            theft
Min.   : 1.00   Min.   : 2.00   Min.   :  3.00
1st Qu.: 3.75   1st Qu.: 5.65   1st Qu.: 22.00
Median :24.50   Median :10.40   Median : 29.00
Mean   :34.99   Mean   :12.28   Mean   : 32.36
3rd Qu.:57.65   3rd Qu.:16.05   3rd Qu.: 38.00
Max.   :99.70   Max.   :39.70   Max.   :147.00

age         involact          income
Min.   : 2.00   Min.   :0.0000   Min.   : 5.583
1st Qu.:48.60   1st Qu.:0.0000   1st Qu.: 8.447
Median :65.00   Median :0.4000   Median :10.694
Mean   :60.33   Mean   :0.6149   Mean   :10.696
3rd Qu.:77.30   3rd Qu.:0.9000   3rd Qu.:11.989
Max.   :90.10   Max.   :2.2000   Max.   :21.480
## Wide range of race
## Many involact values equal to zero
```

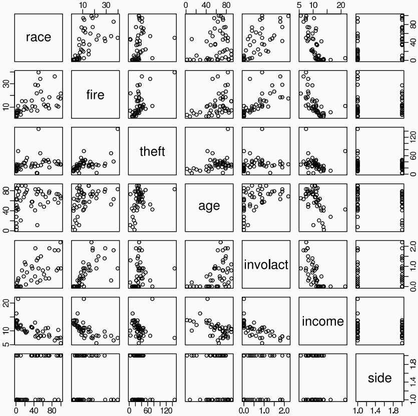
# Initial Data Analysis

```
# Make plots
> par(mfrow=c(2,3))
> for (i in 1:6)
+   stripchart(chredlin[,i],
+   main=names(chredlin)[i],
+   vertical=T, method="jitter")
> par(mfrow=c(1,1))
> pairs(chredlin)
## ‘‘theft’’ and ‘‘income’’ are skewed
```

# Strip plots

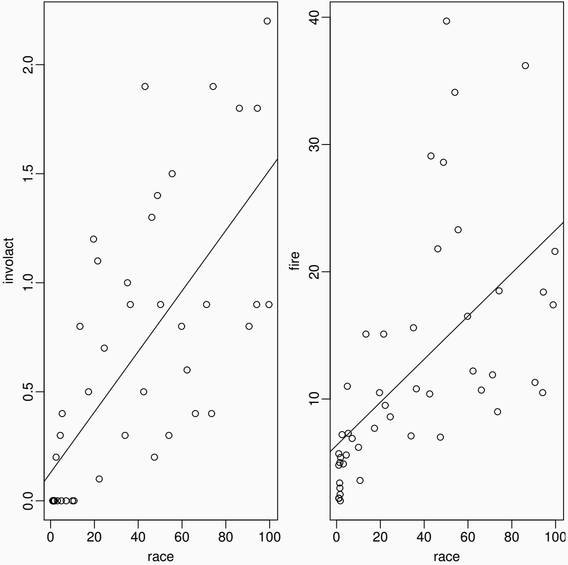


# Pairwise scatterplots



```
## Quick check of ‘‘involact’’ and ‘‘race’’
> summary(lm(involact ~ race, chredlin))

Coefficients:
Estimate Std.Error t value Pr(>|t|)
(Intercept) 0.129218  0.096611   1.338    0.188
race        0.013882  0.002031   6.836 1.78e-08

Residual standard error: 0.4488 on 45 degrees of freedom
Multiple R-Squared: 0.5094, Adjusted R-squared: 0.4985
F-statistic: 46.73 on 1 and 45 DF,  p-value: 1.784e-08

## Effect of fire
> par(mfrow=c(1,2))
> plot(involact ~ race, chredlin)
> abline(lm(involact ~ race, chredlin))
> plot(fire ~ race, chredlin)
> abline(lm(fire ~ race, chredlin))
```

# Fire and race plots



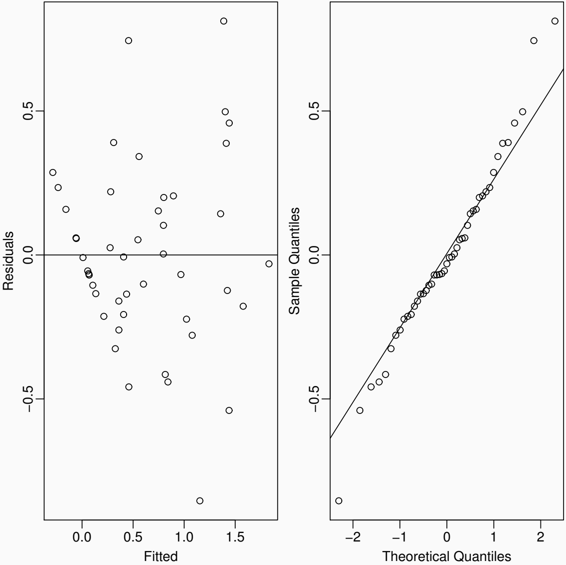
# Initial Model and Diagnostics

```
## log(income): skewed, typical transform
> g <- lm(involact ~ race + fire + theft + age
+ log(income), chredlin)
> summary(g)
Coefficients:
Estimate Std.Error t value Pr(>|t|)
Intercept -1.185540  1.100255  -1.078 0.287550
race       0.009502  0.002490   3.817 0.000449
fire       0.039856  0.008766   4.547 4.76e-05
theft     -0.010295  0.002818  -3.653 0.000728
age        0.008336  0.002744   3.038 0.004134
log(income)0.345762  0.400123   0.864 0.392540

Residual standard error: 0.3345 on 41 degrees of freedom
Multiple R-Squared: 0.7517  Adjusted R-squared: 0.7214
F-statistic: 24.83 on 5 and 41 DF,  p-value: 2.009e-11
```

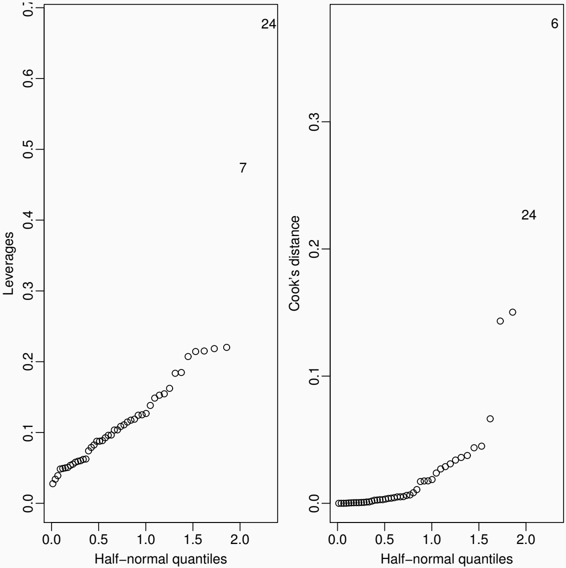
# Initial Model and Diagnostics

```
## Diagnostics
> plot(fitted(g), residuals(g), xlab="Fitted",
ylab="Residuals")
> abline(h=0)
> qqnorm(residuals(g))
> qqline(residuals(g))
> ## Influence
> halfnorm(lm.influence(g)$hat, ylab="Leverages")
> halfnorm(cooks.distance(g), ylab="Cook’s distance")
```

# Diagnostics



# Influential Points



```
> chredlin[c(6, 7, 24), ]
race fire theft  age involact income side
60610 54.0 34.1    68 52.6      0.3  8.231    n
60611  4.9 11.0    75 42.6      0.0 21.480    n
60607 50.2 39.7   147 83.0      0.9  7.459    n
> ## Check for outliers
> range(rstudent(g))
[1] -3.184960  2.792884
> 2*pt(-3.18,df=41)*47
[1] 0.1317647
```

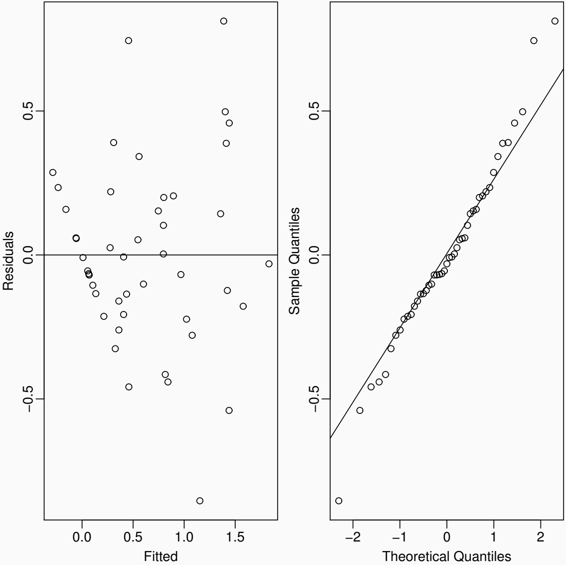
```
## Remove suspicious points
> g <- lm(involact ~ race + fire + theft + age
+ log(income), chredlin, subset=-c(6, 7, 24))
> summary(g)
Coefficients:
Estimate Std.Error t value Pr(>|t|)
Intercept -0.874752  1.241626  -0.705   0.4854
race       0.007105  0.002724   2.608   0.0129
fire       0.051394  0.009327   5.510 2.67e-06
theft     -0.005030  0.005205  -0.966   0.3400
age        0.004987  0.002946   1.693   0.0986
log(income)0.223032  0.457970   0.487   0.6291

Residual standard error: 0.3062 on 38 degrees of freedom
Multiple R-Squared: 0.8012,  Adjusted R-squared: 0.775
F-statistic: 30.62 on 5 and 38 DF,  p-value: 2.352e-12
```

# Transformations

- Choose not to attempt to tranform response (many zeros, interpretation important)
- Consider predictor transforms

```
## Partial residual plot
> prplot(g, 1)
> prplot(g, 2)
```

# Partial residual plots



# Variable Selection

```
## Sensitive to outliers
## select without the influential points
> chreduc <- chredlin[-c(6, 7, 24), ]
> library(leaps)
> b <- regsubsets(involact ~ race + fire +
theft + age + log(income), force.in=1,
data=chreduc)
> (rs <- summary(b))
Subset selection object
Forced in Forced out
race             TRUE      FALSE
fire            FALSE      FALSE
theft           FALSE      FALSE
age             FALSE      FALSE
log(income)     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
race fire theft age log(income)
2  ( 1 ) "*"  "*"  " "   " " " "
3  ( 1 ) "*"  "*"  " "   "*" " "
4  ( 1 ) "*"  "*"  "*"   "*" " "
5  ( 1 ) "*"  "*"  "*"   "*" "*"
```

# Variable Selection

```
> rs$adj
[1] 0.7751279 0.7802010 0.7793875 0.7749863
> g <- lm(involact ~ race + fire + age,
chredlin, subset=-c(6, 7, 24))
> summary(g)
Coefficients:
Estimate Std.Error t value Pr(>|t|)
Intercept -0.324716  0.131520  -2.469   0.0179
race       0.005453  0.001844   2.957   0.0052
fire       0.051235  0.008404   6.096 3.46e-07
age        0.003158  0.002264   1.395   0.1707

Residual standard error: 0.3026 on 40 degrees of freedom
Multiple R-Squared: 0.7955, Adjusted R-squared: 0.7802
F-statistic: 51.88 on 3 and 40 DF,  p-value: 7.521e-14
```

# Variable Selection

```
## With all the data
> g <- lm(involact ~ race + fire + age,  chredlin)
> summary(g)
Coefficients:
Estimate Std.Error t value Pr(>|t|)
Intercept -0.354962  0.159754  -2.222 0.031601
race       0.008866  0.002114   4.194 0.000134
fire       0.023296  0.007868   2.961 0.004978
age        0.006194  0.002697   2.296 0.026582

Residual standard error: 0.3762 on 43 degrees of freedom
Multiple R-Squared: 0.6707, Adjusted R-squared: 0.6477
F-statistic: 29.19 on 3 and 43 DF,  p-value: 1.877e-10
```

# Variable Selection

```
## Do North and South side separately
> g <- lm(involact ~ race + fire + age,
subset=(side=='s'), chredlin)
> summary(g)
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.337126   0.196564  -1.715  0.10349
race         0.004750   0.002872   1.654  0.11556
fire         0.054363   0.014923   3.643  0.00186 **
age          0.003254   0.004478   0.727  0.47676
Residual standard error: 0.3471 on 18 degrees of freedom
Multiple R-Squared: 0.7341, Adjusted R-squared: 0.6897
F-statistic: 16.56 on 3 and 18 DF,  p-value: 2.045e-05

> g <- lm(involact ~ race + fire + age,
subset=(side=='n'), chredlin)
> summary(g)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.386699   0.241539  -1.601  0.12432
race         0.015692   0.004532   3.463  0.00233 **
fire         0.004099   0.011316   0.362  0.72077
age          0.007555   0.003706   2.039  0.05429 .
Residual standard error: 0.3687 on 21 degrees of freedom
Multiple R-Squared: 0.7032, Adjusted R-squared: 0.6608
F-statistic: 16.59 on 3 and 21 DF,  p-value: 9.334e-06
```

# Variable Selection

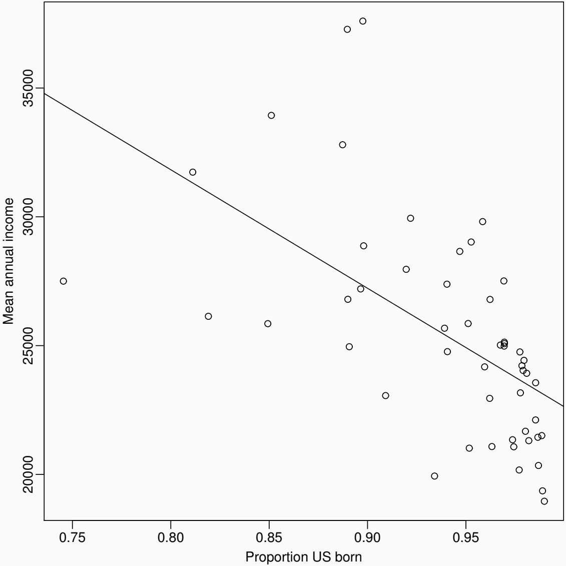
```
> summary(race[side == 's'])
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00   34.28   48.10   49.80   72.92   99.70

> summary(race[side == 'n'])
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    1.80   10.00   21.95   24.50   94.40
# The difference is NOT due to more uniform race
# values on one  side
```

# Ecological Correlation

- When variables are aggregated, correlations appear stronger than at the individual level
- Example: state-level data on income and % US-born legal residents (1990)

```
> data(eco)
> summary(eco$usborn)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.7454  0.9144  0.9621  0.9403  0.9781  0.9899
> summary(eco$income)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18960   21890   24960   25370   27440   37600
> g <- lm(income ~ usborn, eco)
> summary(g)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    68642       8739   7.855 3.19e-10 ***
usborn        -46019       9279  -4.959 8.89e-06 ***
Residual standard error: 3490 on 49 degrees of freedom
Multiple R-Squared: 0.3342,     Adjusted R-squared: 0.3206
F-statistic:  24.6 on 1 and 49 DF,  p-value: 8.891e-06

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

# US Demographic Example Continued



# Ecological Fallacy

Hypothetical state with `usborn = 1`:

Hypothetical state with `usborn = 0`:

- Census data: US-born citizens have slightly higher average income than naturalized citizens.
- Why does the ecological regression lie? – There are more immigrants in wealthier states.