Linear Regression

Veerasak Kritsanapraphan

8/3/2017

Stepwise Logistic Regression

Import packages necessary first.

```
library(MASS)
library(plyr)
library(ggplot2)
library(knitr)
```

Prepare data

```
crime <- read.table("crime_simple.txt", sep="\t", header = TRUE)
```

Run Linear regression

```
# Assign more meaningful variable names
colnames(crime) <- c("crime.per.million", "young.males", "is.south", "average.ed",
                     "exp.per.cap.1960", "exp.per.cap.1959", "labour.part",
                     "male.per.fem", "population", "nonwhite",
                     "unemp.youth", "unemp.adult", "median.assets", "num.low.salary")

# Convert is.south to a factor
# Divide average.ed by 10 so that the variable is actually average education
# Convert median assets to 1000's of dollars instead of 10's
crime <- transform(crime, is.south = as.factor(is.south),
                          average.ed = average.ed / 10,
                          median.assets = median.assets / 100)

# Fit model
crime.lm <- lm(crime.per.million ~ ., data = crime)
# Remove 1959 expenditure and youth unemployment
#crime.lm2 <- update(crime.lm, . ~ . - exp.per.cap.1959 - unemp.youth)
crime.lm2 <- lm(crime.per.million ~ young.males + average.ed + unemp.adult + num.low.salary, data = crim
summary(crime.lm)
```

```
##
## Call:
## lm(formula = crime.per.million ~ ., data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.884 -11.923  -1.135  13.495  50.560
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6.918e+02  1.559e+02  -4.438 9.56e-05 ***
## young.males       1.040e+00  4.227e-01   2.460  0.01931 *
## is.south1        -8.308e+00  1.491e+01  -0.557  0.58117
## average.ed        1.802e+01  6.497e+00   2.773  0.00906 **
## exp.per.cap.1960  1.608e+00  1.059e+00   1.519  0.13836   
## exp.per.cap.1959 -6.673e-01  1.149e+00  -0.581  0.56529   
## labour.part      -4.103e-02  1.535e-01  -0.267  0.79087   
## male.per.fem      1.648e-01  2.099e-01   0.785  0.43806   
## population       -4.128e-02  1.295e-01  -0.319  0.75196   
## nonwhite          7.175e-03  6.387e-02   0.112  0.91124   
## unemp.youth      -6.017e-01  4.372e-01  -1.376  0.17798   
## unemp.adult       1.792e+00  8.561e-01   2.093  0.04407 * 
## median.assets     1.374e+01  1.058e+01   1.298  0.20332   
## num.low.salary    7.929e-01  2.351e-01   3.373  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 21.94 on 33 degrees of freedom
## Multiple R-squared:  0.7692, Adjusted R-squared:  0.6783 
## F-statistic: 8.462 on 13 and 33 DF,  p-value: 3.686e-07
```

```
summary(crime.lm2)
```

```
## 
## Call:
## lm(formula = crime.per.million ~ young.males + average.ed + unemp.adult + 
##     num.low.salary, data = crime)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -47.279 -25.068  -4.437  16.835  91.654 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)   
## (Intercept)    -349.1583   155.0258  -2.252   0.0296 * 
## young.males       0.7675     0.5870   1.307   0.1982   
## average.ed       22.9954     7.8909   2.914   0.0057 **
## unemp.adult       1.7367     0.7065   2.458   0.0182 * 
## num.low.salary    0.1618     0.2275   0.711   0.4809   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 35.52 on 42 degrees of freedom
## Multiple R-squared:  0.2298, Adjusted R-squared:  0.1564 
## F-statistic: 3.133 on 4 and 42 DF,  p-value: 0.02422
```

Here’s a comparison of the regression models (with and without the collinearity problem).

```
kable(summary(crime.lm)$coef,
      digits = c(3, 3, 3, 4), format = 'markdown')
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-691.838	155.888	-4.438	0.0001
young.males	1.040	0.423	2.460	0.0193
is.south1	-8.308	14.912	-0.557	0.5812
average.ed	18.016	6.497	2.773	0.0091
exp.per.cap.1960	1.608	1.059	1.519	0.1384
exp.per.cap.1959	-0.667	1.149	-0.581	0.5653
labour.part	-0.041	0.153	-0.267	0.7909

	Estimate	Std. Error	t value	Pr(>|t|)
male.per.fem	0.165	0.210	0.785	0.4381
population	-0.041	0.130	-0.319	0.7520
nonwhite	0.007	0.064	0.112	0.9112
unemp.youth	-0.602	0.437	-1.376	0.1780
unemp.adult	1.792	0.856	2.093	0.0441
median.assets	13.736	10.583	1.298	0.2033
num.low.salary	0.793	0.235	3.373	0.0019

```r
crime.lm.summary2 <- summary(crime.lm2)
kable(crime.lm.summary2$coef,
      digits = c(3, 3, 3, 4), format = 'markdown')
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-349.158	155.026	-2.252	0.0296
young.males	0.767	0.587	1.307	0.1982
average.ed	22.995	7.891	2.914	0.0057
unemp.adult	1.737	0.707	2.458	0.0182
num.low.salary	0.162	0.227	0.711	0.4809

Stepwise Regression

```r
backwards = step(crime.lm) # Backwards selection is the default
```

```
## Start:  AIC=301.66
## crime.per.million ~ young.males + is.south + average.ed + exp.per.cap.1960 +
##     exp.per.cap.1959 + labour.part + male.per.fem + population +
##     nonwhite + unemp.youth + unemp.adult + median.assets + num.low.salary
##
##                    Df Sum of Sq   RSS    AIC
## - nonwhite          1       6.1 15885 299.68
## - labour.part       1      34.4 15913 299.76
## - population        1      48.9 15928 299.81
## - is.south          1     149.4 16028 300.10
## - exp.per.cap.1959  1     162.3 16041 300.14
## - male.per.fem      1     296.5 16175 300.53
## <none>                          15879 301.66
## - median.assets     1     810.6 16689 302.00
## - unemp.youth       1     911.5 16790 302.29
## - exp.per.cap.1960  1    1109.8 16988 302.84
## - unemp.adult       1    2108.8 17988 305.52
## - young.males       1    2911.6 18790 307.57
## - average.ed        1    3700.5 19579 309.51
## - num.low.salary    1    5474.2 21353 313.58
##
## Step:  AIC=299.68
## crime.per.million ~ young.males + is.south + average.ed + exp.per.cap.1960 +
##     exp.per.cap.1959 + labour.part + male.per.fem + population +
##     unemp.youth + unemp.adult + median.assets + num.low.salary
##
##                    Df Sum of Sq   RSS    AIC
## - labour.part       1      28.7 15913 297.76
## - population        1      48.6 15933 297.82
## - exp.per.cap.1959  1     156.3 16041 298.14
## - is.south          1     158.0 16043 298.14
## - male.per.fem      1     294.1 16179 298.54
## <none>                          15885 299.68
## - median.assets     1     820.2 16705 300.05
## - unemp.youth       1     913.1 16798 300.31
## - exp.per.cap.1960  1    1104.3 16989 300.84
## - unemp.adult       1    2107.1 17992 303.53
## - young.males       1    3365.8 19250 306.71
## - average.ed        1    3757.1 19642 307.66
## - num.low.salary    1    5503.6 21388 311.66
##
## Step:  AIC=297.76
## crime.per.million ~ young.males + is.south + average.ed + exp.per.cap.1960 +
##     exp.per.cap.1959 + male.per.fem + population + unemp.youth +
##     unemp.adult + median.assets + num.low.salary
##
##                    Df Sum of Sq   RSS    AIC
## - population        1      62.2 15976 295.95
## - is.south          1     129.4 16043 296.14
## - exp.per.cap.1959  1     134.8 16048 296.16
## - male.per.fem      1     276.8 16190 296.57
## <none>                          15913 297.76
## - median.assets     1     801.9 16715 298.07
## - unemp.youth       1     941.8 16855 298.47
## - exp.per.cap.1960  1    1075.9 16989 298.84
## - unemp.adult       1    2088.5 18002 301.56
## - young.males       1    3407.9 19321 304.88
## - average.ed        1    3895.3 19809 306.06
## - num.low.salary    1    5621.3 21535 309.98
##
## Step:  AIC=295.95
## crime.per.million ~ young.males + is.south + average.ed + exp.per.cap.1960 +
##     exp.per.cap.1959 + male.per.fem + unemp.youth + unemp.adult +
##     median.assets + num.low.salary
##
##                    Df Sum of Sq   RSS    AIC
## - is.south          1     104.4 16080 294.25
## - exp.per.cap.1959  1     123.3 16099 294.31
## - male.per.fem      1     533.8 16509 295.49
## <none>                          15976 295.95
## - median.assets     1     748.7 16724 296.10
## - unemp.youth       1     997.7 16973 296.80
## - exp.per.cap.1960  1    1021.3 16997 296.86
## - unemp.adult       1    2082.3 18058 299.71
## - young.males       1    3425.9 19402 303.08
## - average.ed        1    3887.6 19863 304.19
## - num.low.salary    1    5896.9 21873 308.71
##
## Step:  AIC=294.25
## crime.per.million ~ young.males + average.ed + exp.per.cap.1960 +
##     exp.per.cap.1959 + male.per.fem + unemp.youth + unemp.adult +
##     median.assets + num.low.salary
```

```
##
##                    Df Sum of Sq   RSS    AIC
## - exp.per.cap.1959  1     171.5 16252 292.75
## - male.per.fem      1     563.4 16643 293.87
## <none>                          16080 294.25
## - median.assets     1     734.7 16815 294.35
## - unemp.youth       1     906.0 16986 294.83
## - exp.per.cap.1960  1    1162.0 17242 295.53
## - unemp.adult       1    1978.0 18058 297.71
## - young.males       1    3354.5 19434 301.16
## - average.ed        1    4139.1 20219 303.02
## - num.low.salary    1    6094.8 22175 307.36
##
## Step:  AIC=292.75
## crime.per.million ~ young.males + average.ed + exp.per.cap.1960 +
##     male.per.fem + unemp.youth + unemp.adult + median.assets +
##     num.low.salary
##
##                    Df Sum of Sq   RSS    AIC
## - male.per.fem      1     691.0 16942 292.71
## <none>                          16252 292.75
## - median.assets     1     759.0 17010 292.90
## - unemp.youth       1     921.8 17173 293.35
## - unemp.adult       1    2018.1 18270 296.25
## - young.males       1    3323.1 19574 299.50
## - average.ed        1    4005.1 20256 301.11
## - num.low.salary    1    6402.7 22654 306.36
## - exp.per.cap.1960  1   11818.8 28070 316.44
##
## Step:  AIC=292.71
## crime.per.million ~ young.males + average.ed + exp.per.cap.1960 +
##     unemp.youth + unemp.adult + median.assets + num.low.salary
##
##                    Df Sum of Sq   RSS    AIC
## - unemp.youth       1     408.6 17351 291.83
## <none>                          16942 292.71
## - median.assets     1    1016.9 17959 293.45
## - unemp.adult       1    1548.6 18491 294.82
## - young.males       1    4511.6 21454 301.81
## - average.ed        1    6430.6 23373 305.83
## - num.low.salary    1    8147.7 25090 309.16
## - exp.per.cap.1960  1   12019.6 28962 315.91
##
## Step:  AIC=291.83
## crime.per.million ~ young.males + average.ed + exp.per.cap.1960 +
##     unemp.adult + median.assets + num.low.salary
##
##                    Df Sum of Sq   RSS    AIC
## <none>                          17351 291.83
## - median.assets     1    1252.6 18604 293.11
## - unemp.adult       1    1628.7 18980 294.05
## - young.males       1    4461.0 21812 300.58
## - average.ed        1    6214.7 23566 304.22
## - num.low.salary    1    8932.3 26283 309.35
## - exp.per.cap.1960  1   15596.5 32948 319.97
```

```r
formula(backwards)
```

```
## crime.per.million ~ young.males + average.ed + exp.per.cap.1960 + 
##     unemp.adult + median.assets + num.low.salary
```

```r
summary(backwards)
```

```
## 
## Call:
## lm(formula = crime.per.million ~ young.males + average.ed + exp.per.cap.1960 + 
##     unemp.adult + median.assets + num.low.salary, data = crime)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -38.306 -10.209  -1.313   9.919  54.544 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -618.5028   108.2456  -5.714 1.19e-06 ***
## young.males         1.1252     0.3509   3.207 0.002640 ** 
## average.ed         18.1786     4.8027   3.785 0.000505 ***
## exp.per.cap.1960    1.0507     0.1752   5.996 4.78e-07 ***
## unemp.adult         0.8282     0.4274   1.938 0.059743 .  
## median.assets      15.9565     9.3900   1.699 0.097028 .  
## num.low.salary      0.8236     0.1815   4.538 5.10e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 20.83 on 40 degrees of freedom
## Multiple R-squared:  0.7478, Adjusted R-squared:   0.71 
## F-statistic: 19.77 on 6 and 40 DF,  p-value: 1.441e-10
```