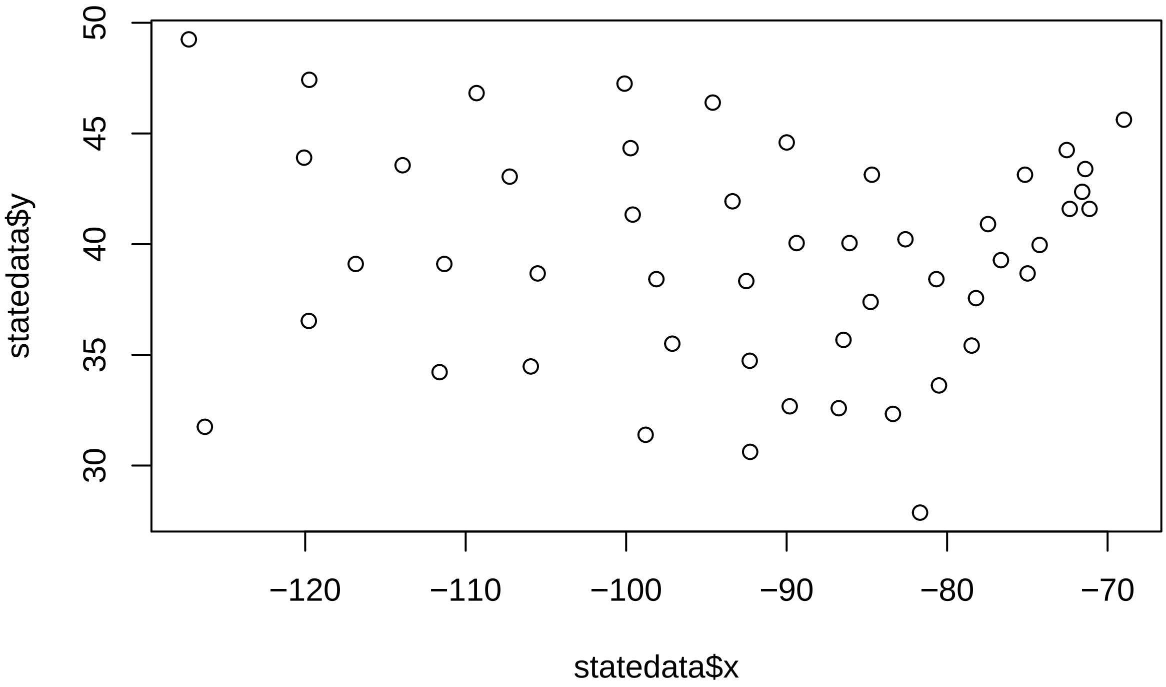
Unit 2 - State Data

```
data(state)
statedata = cbind(data.frame(state.x77), state.abb, state.area, state.center,
                  state.division, state.name, state.region)
str(statedata)
```

```
## 'data.frame':    50 obs. of  15 variables:
##  $ Population    : num  3615 365 2212 2110 21198 ...
##  $ Income        : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy    : num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life.Exp      : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder        : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS.Grad       : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost         : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area          : num  50708 566432 113417 51945 156361 ...
##  $ state.abb     : Factor w/ 50 levels "AK","AL","AR",..: 2 1 4 3 5 6 7 8 9 10 ...
##  $ state.area    : num  51609 589757 113909 53104 158693 ...
##  $ x             : num  -86.8 -127.2 -111.6 -92.3 -119.8 ...
##  $ y             : num  32.6 49.2 34.2 34.7 36.5 ...
##  $ state.division: Factor w/ 9 levels "New England",..: 4 9 8 5 9 8 1 3 3 3 ...
##  $ state.name    : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ state.region  : Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...
```

Problem 1.1

```
plot(statedata$x, statedata$y)
```



statedata$x

statedata$x

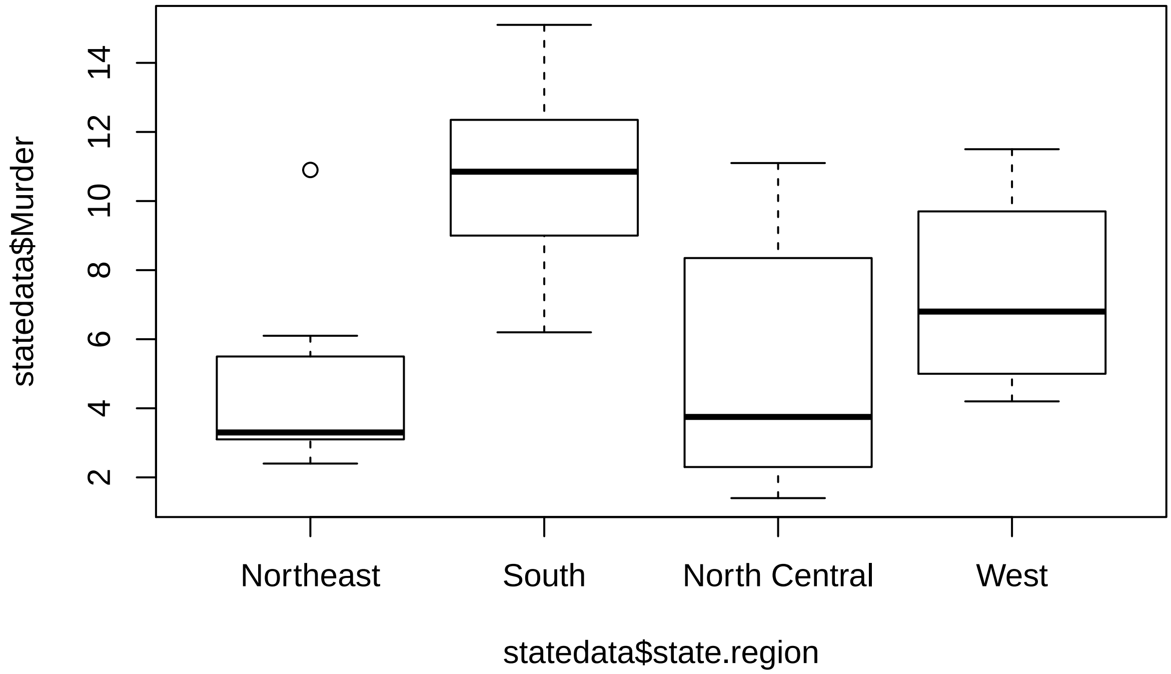
```r
tapply(statedata$HS.Grad, statedata$state.region, mean)
```

```
##     Northeast         South North Central          West 
##      53.96667      44.34375      54.51667      62.00000
```

West

Problem 1.3

```r
boxplot(statedata$Murder ~ statedata$state.region)
```



South

Problem 1.4

```
ne_df <- subset(statedata, state.region == "Northeast")
sort(tapply(ne_df$Murder, ne_df$state.name, mean), decreasing = T)
```

```
##      New York  Pennsylvania       Vermont    New Jersey Massachusetts 
##          10.9           6.1           5.5           5.2           3.3 
## New Hampshire   Connecticut         Maine  Rhode Island 
##           3.3           3.1           2.7           2.4
```

New York

Problem 2.1

```
model <- lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
              Frost + Area, statedata)
summary(model)
```

```
##
```

```
## Call:
## lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +
##     HS.Grad + Frost + Area, data = statedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## Population   5.180e-05  2.919e-05   1.775   0.0832 .
## Income      -2.180e-05  2.444e-04  -0.089   0.9293
## Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
## Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
## Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
## Area        -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

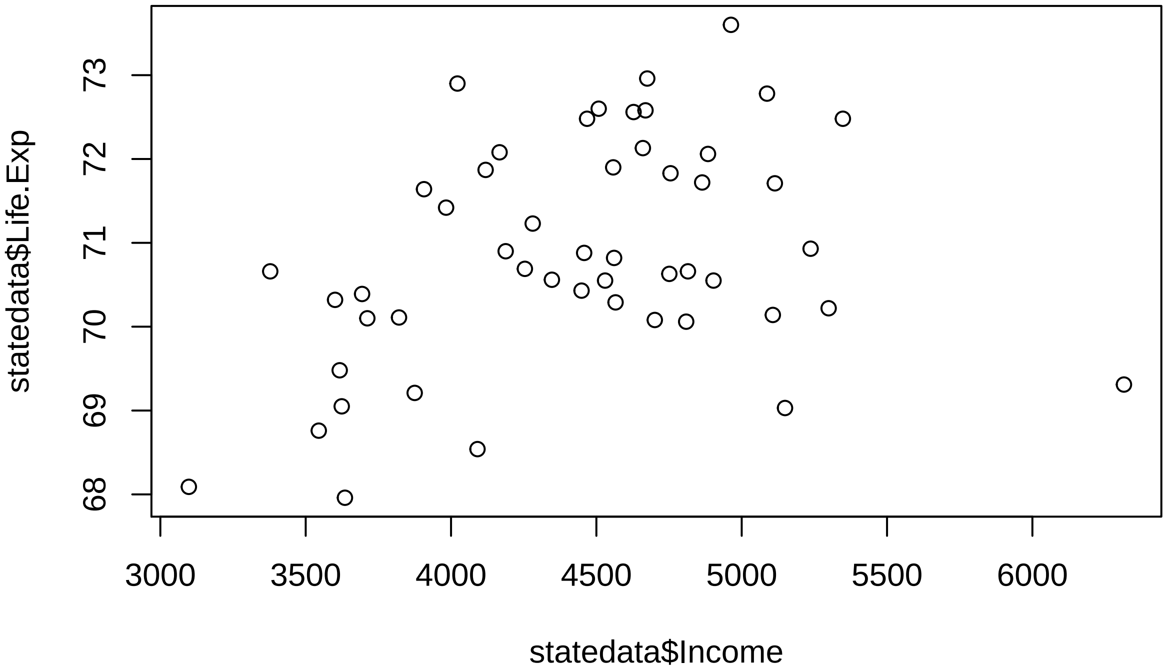
-0.0000218

Problem 2.2

For a one unit increase in income, predicted life expectancy decreases by $|x|$

Problem 2.3

```r
plot(statedata$Income, statedata$Life.Exp)
```



Life expectancy is somewhat positively correlated with income.

Problem 2.4

Multicollinearity
Although income is an insignificant variable in the model, this does not mean that there is no association between income and life expectancy. However, in the presence of all of the other variables, income does not add statistically significant explanatory power to the model. This means that multicollinearity is probably the issue.

Problem 3.1

```r
model <- lm(Life.Exp ~ Population + Murder + HS.Grad + Frost, statedata)
summary(model)
```

```
##
## Call:
## lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
##     data = statedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## Population   5.014e-05  2.512e-05   1.996  0.05201 .
## Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
## Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

Problem 3.2

We expect the “Multiple R-squared” value of the simplified model to be slightly worse than that of the initial model. It can’t be better than the “Multiple R-squared” value of the initial model.
When we remove insignificant variables, the “Multiple R-squared” will always be worse, but only slightly worse. This is due to the nature of a linear regression model. It is always possible for the regression model to make a coefficient zero, which would be the same as removing the variable from the model. The fact that the coefficient is not zero in the intial model means it must be helping the R-squared value, even if it is only a very small improvement. So when we force the variable to be removed, it will decrease the R-squared a little bit. However, this small decrease is worth it to have a simpler model.
On the contrary, when we remove insignificant variables, the “Adjusted R-squred” will frequently be better. This value accounts for the complexity of the model, and thus tends to increase as insignificant variables are removed, and decrease as insignificant variables are added.

Problem 3.3

```r
pred <- predict(model)
sort(pred)[1]
```

```
##  Alabama
## 68.48112
```

```r
statedata$state.name[which.min(statedata$Life.Exp)]
```

```
## [1] South Carolina
## 50 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

Problem 3.4

```r
sort(pred, decreasing = T)[1]
```

```
## Washington
##   72.68272
```

```
statedata$state.name[which.max(statedata$Life.Exp)]
```

```
## [1] Hawaii
## 50 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

Problem 3.5

```
re <- abs(statedata$Life.Exp - pred)
sort(re)[1]
```

```
##    Indiana
## 0.02158526
```

```
sort(re, decreasing = T)[1]
```

```
##   Hawaii
## 1.506831
```