IS621_hw1_final

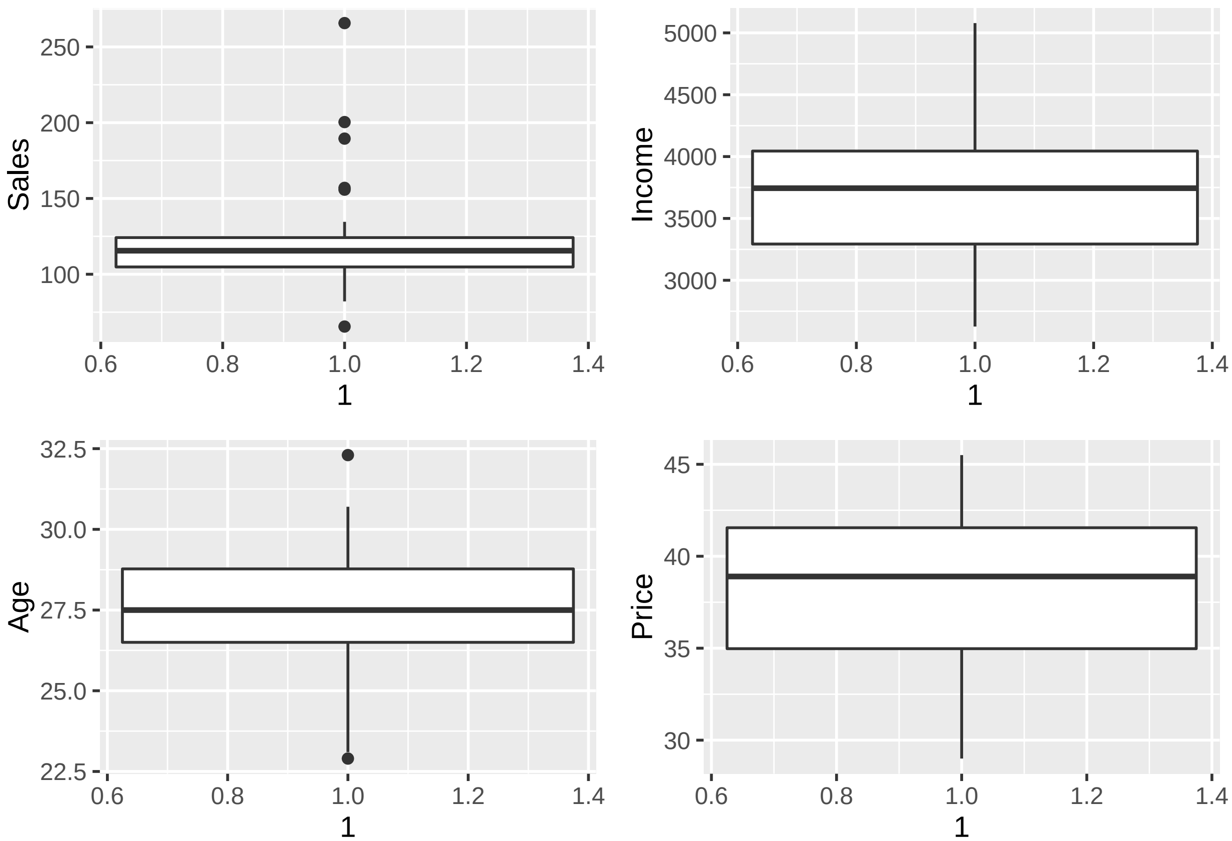
Charley Ferrari

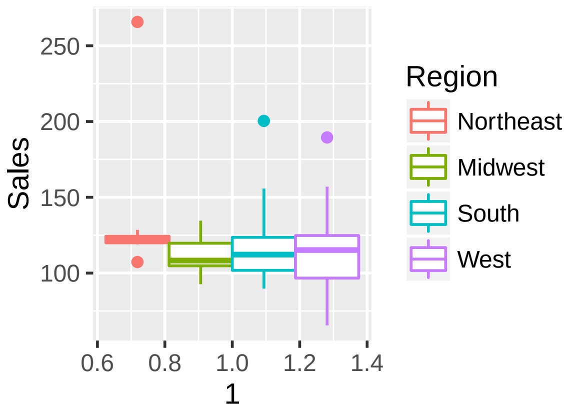
February 5, 2016

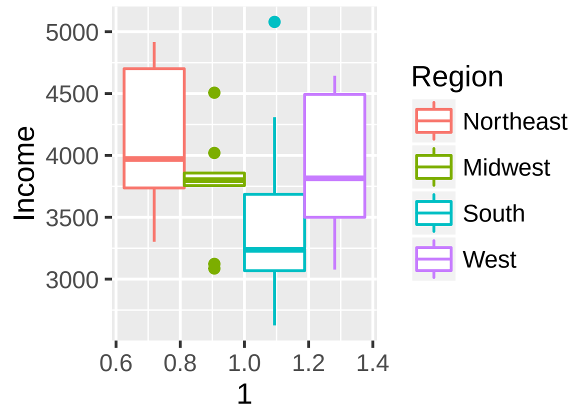
Data Exploration

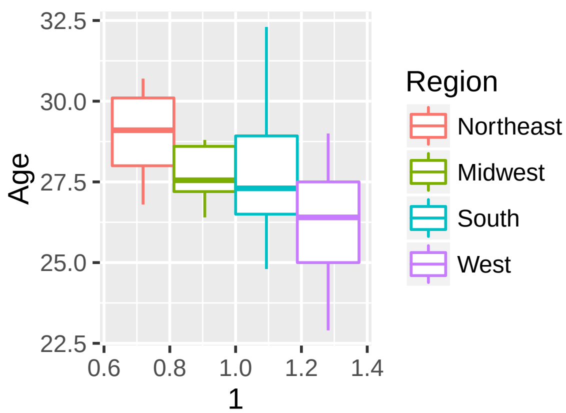
The cigarettes data set consists of five variables: State, Age, Income, Price, and Sales. ALl of these variables are continuous except for State which is categorical. Since there is one observation per state however, this is not useful for analysis. In order to test if there are any geographical effects in this data, I divided the states into the four major census regions: Northeast, Midwest, South, and West. With multiple observations of each factor, I can generate three dummy variables to see if region has any effect.

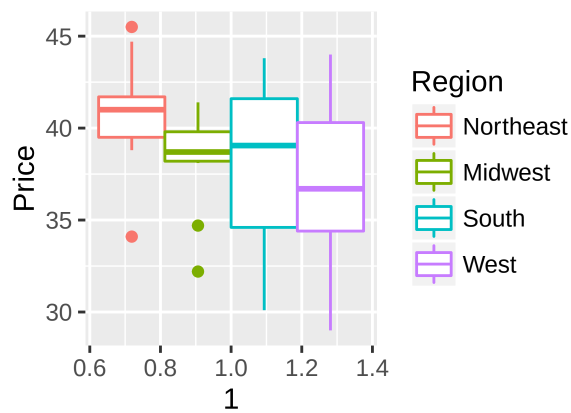
First, I will look at a few boxplots to get a handle on the numerical variables:



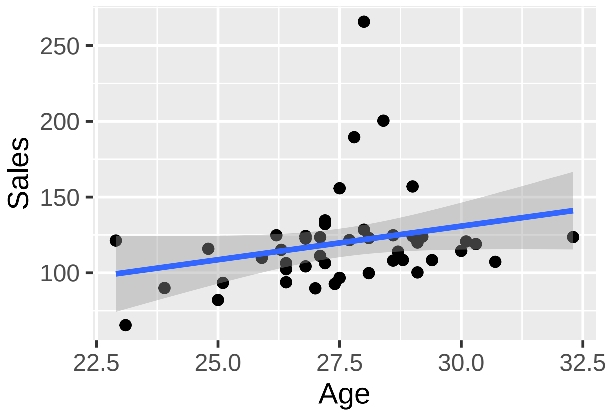


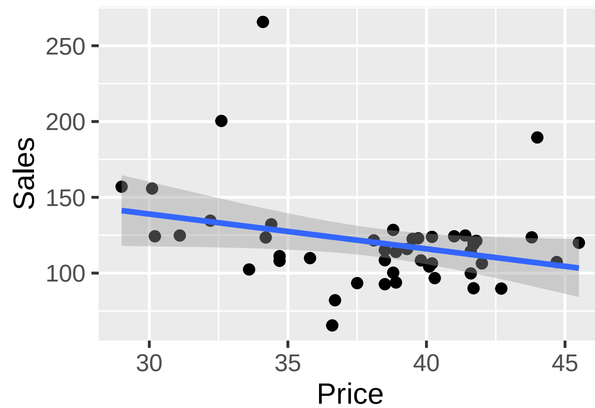


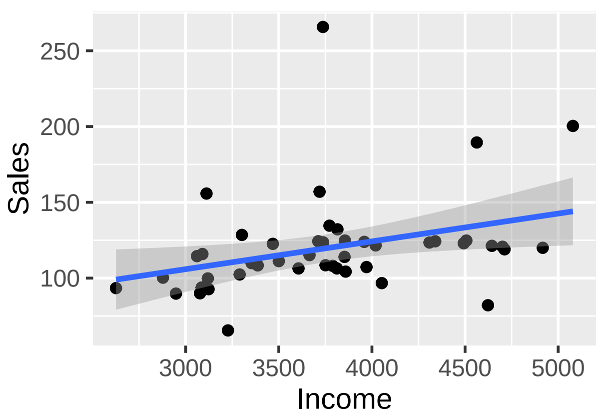




Sales looks to have the most outliers, and they appear to be spread out across multiple regions (suggesting these outliers aren't a regional effect.) Since this is the variable we're predicting, I'll analyze a few scatter plots of our predictor variables versus sales.

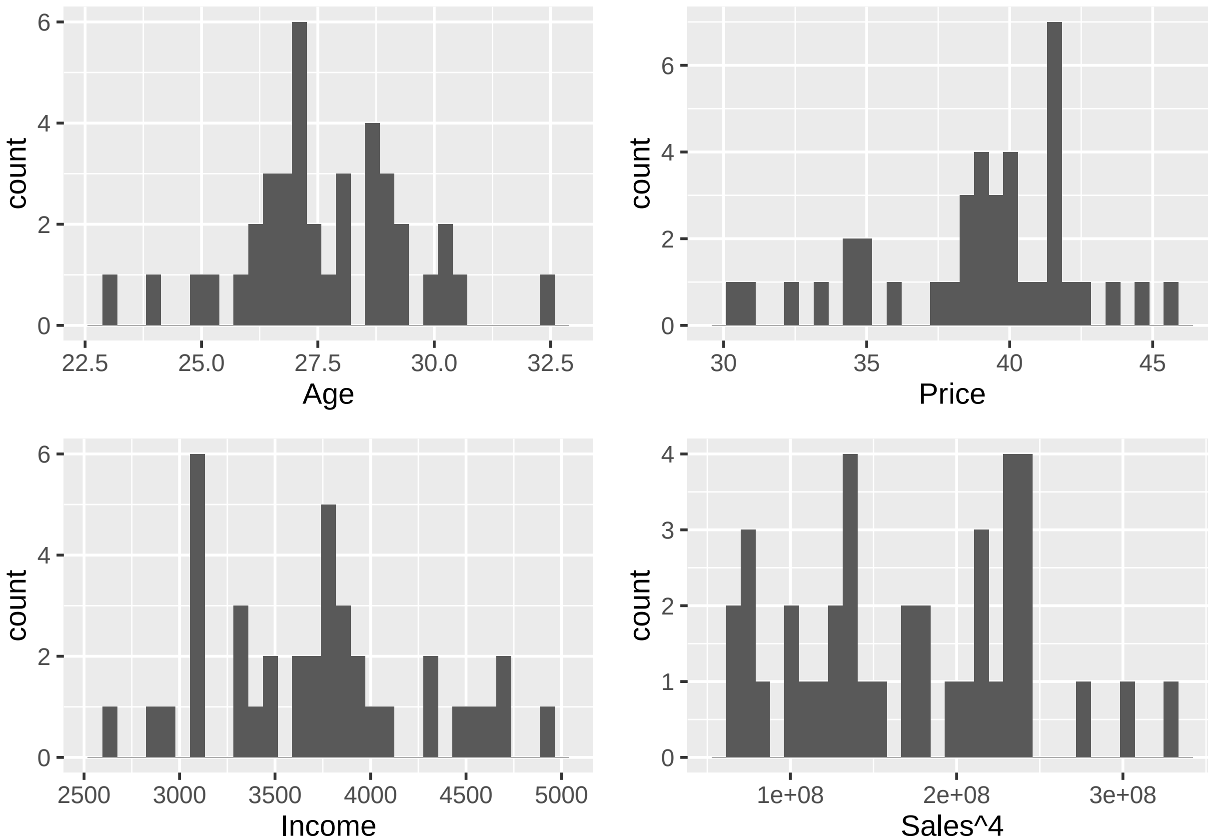






Lets try to make some histograms?

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There seems to be no pattern in the outliers indicating non-linearity, so it might make sense to remove these outliers. Below is a table of the outliers:

State	Age	Income	Price	Sales
DC	28.4	5079	32.6	200.4
KY	27.5	3112	30.1	155.8
NV	27.8	4563	44.0	189.5
NH	28.0	3737	34.1	265.7
OR	29.0	3719	29.0	157.0
UT	23.1	3227	36.6	65.5

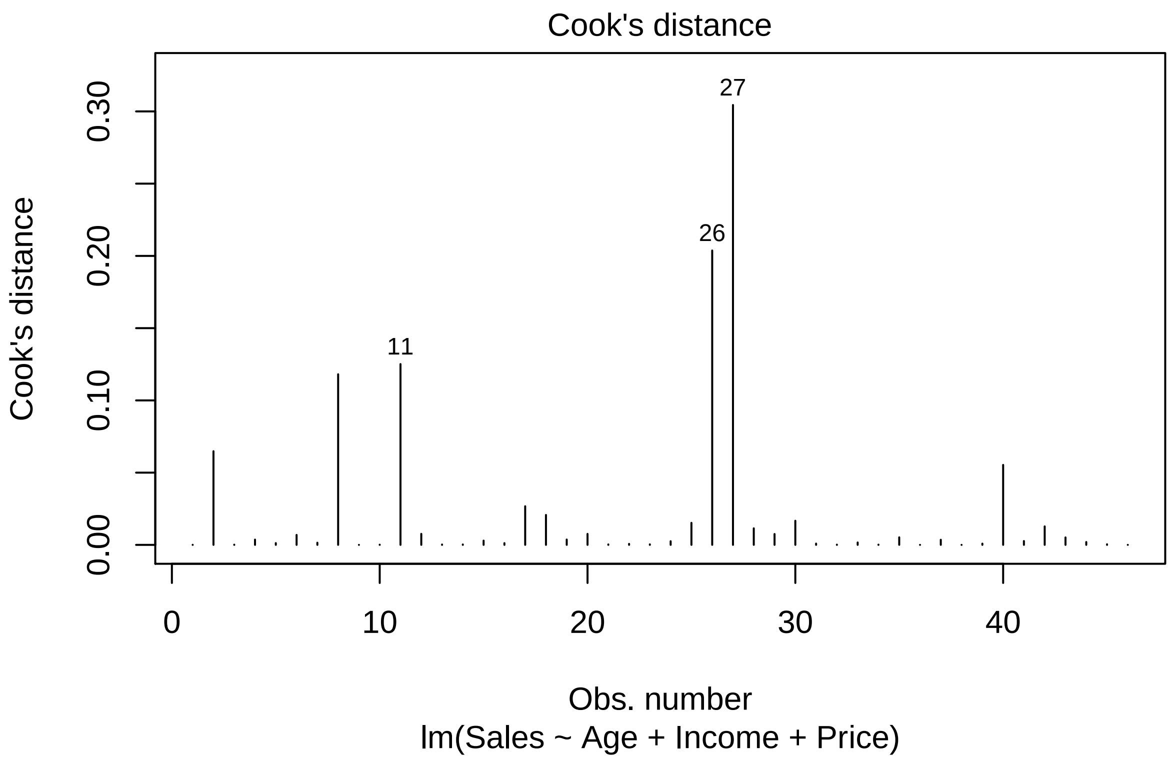
A few of the outliers could be explained as quirky examples. DC is a city, and is perhaps not comparable to other states. Utah has a strong Mormon influence which could explain their low Sales, and Nevada could be influenced by Las Vegas' 'Sin City' status. New Hampshire was the most perplexing, but looking at a table of the Northeast gives a potential solution:

State	Age	Income	Price	Sales
CT	29.1	4917	45.5	120.0
MA	29.0	4340	41.0	124.3
ME	28.0	3302	38.8	128.5
NH	28.0	3737	34.1	265.7
NJ	30.1	4701	41.7	120.7
NY	30.3	4712	41.7	119.0
PA	30.7	3971	44.7	107.3
RI	29.2	3959	40.2	123.9
VT	26.8	3468	39.5	122.6

State	Age	Income	Price	Sales

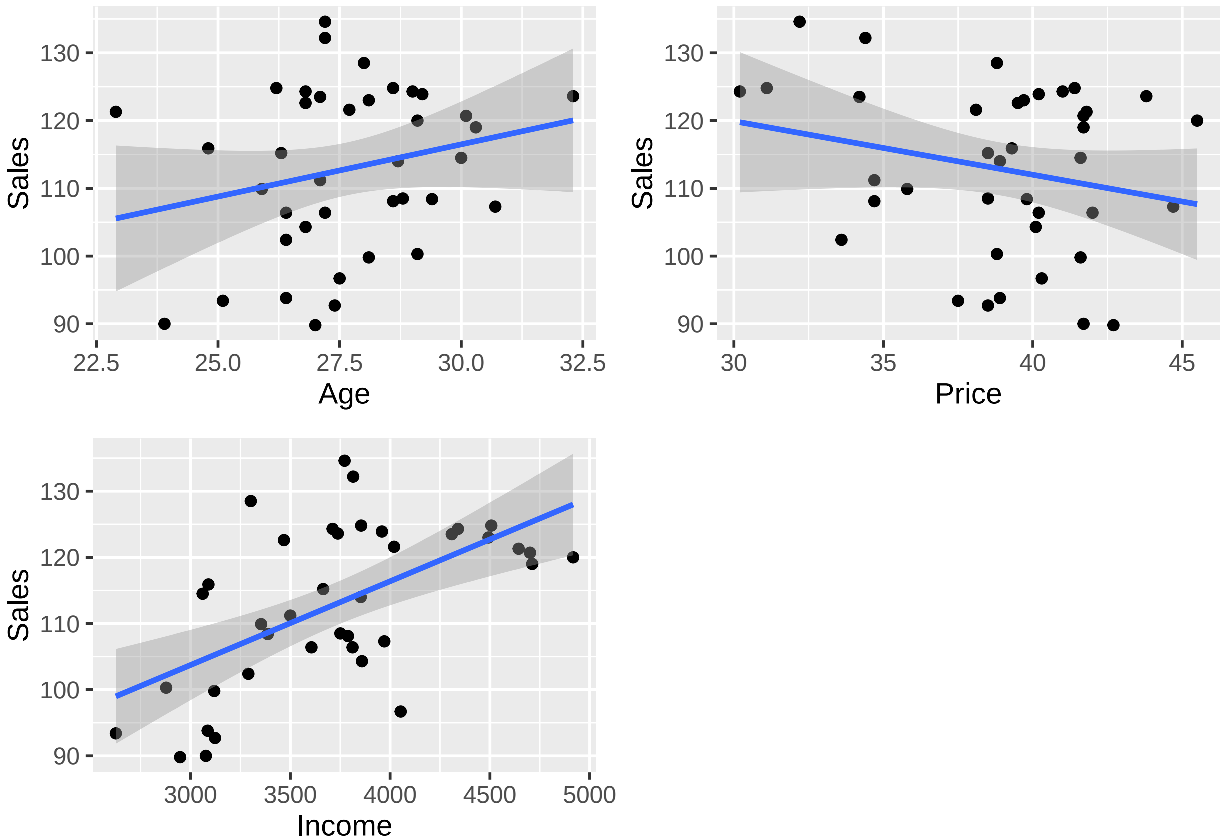
The Northeastern states are smaller, which mean more people can cross state borders to get goods for a lower price. New Hampshire cigarettes are substantially cheaper than those in Massachusetts, which has a comparatively higher population. Similar border effects could be occurring in Oregon and Kentucky (which also have comparatively lower prices.)

As one further test for these outliers, lets look at the Cook's distance in a straight linear model fit for Income, Price, and Age prediciting Sales:



Using the rule of thumb of $\frac{4}{n-2}$ suggests we could keep a few of the outliers, but given how much they differ from the Cook distance of the rest of the datapoints, I believe removing all outliers is the safest decision.

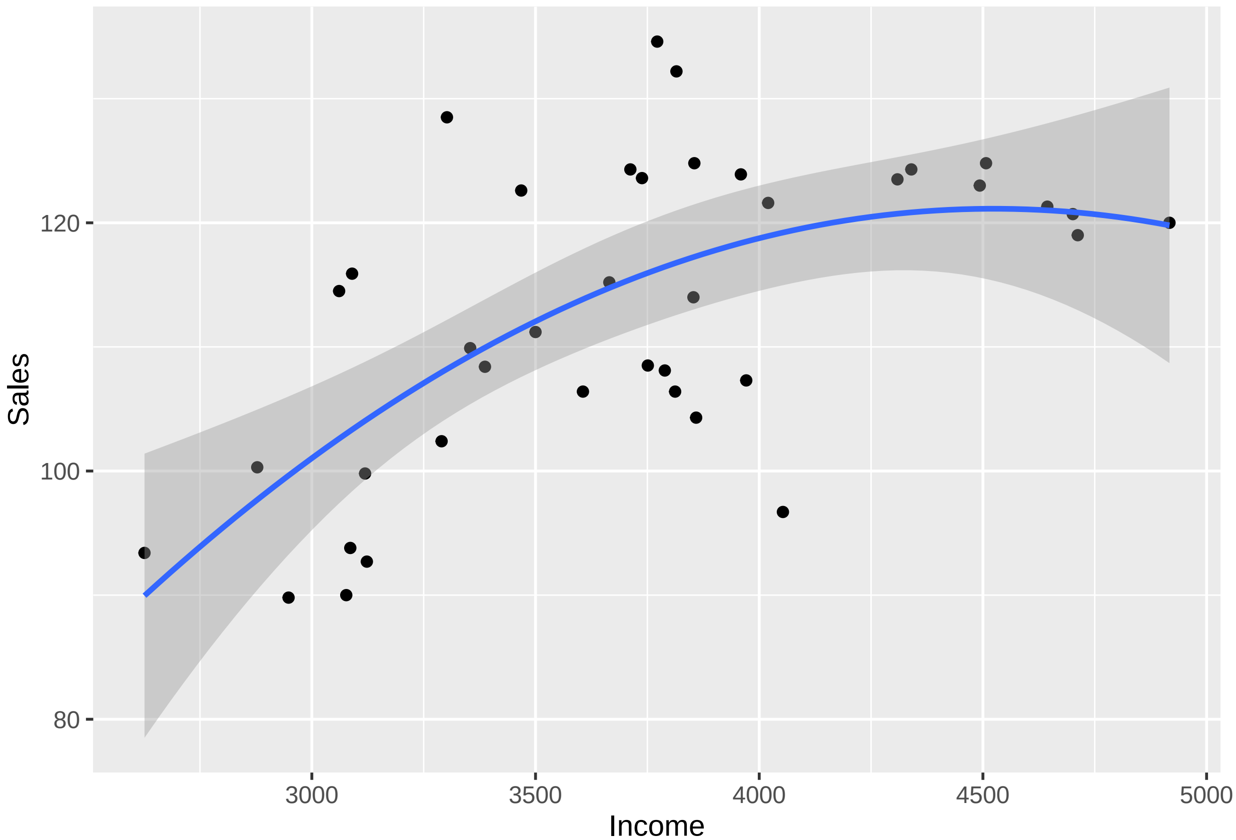
Looking at the scatter plots once we do so shows clearer relationships:



With these outliers removed, we can see some possible non-linearity in the Sales vs Income scatter plot.

This makes intuitive sense. In a lower income range, higher incomes may lead to more cigarette sales simply because more people have disposable income to spend on cigarettes. As income increases past this threshold, the relationship may not be as strong (even if one is a heavy smoker, there is an upper limit to cigarette consumption.)

To capture some of this, lets plot a best fit line for a model including Income and $Income^2$:

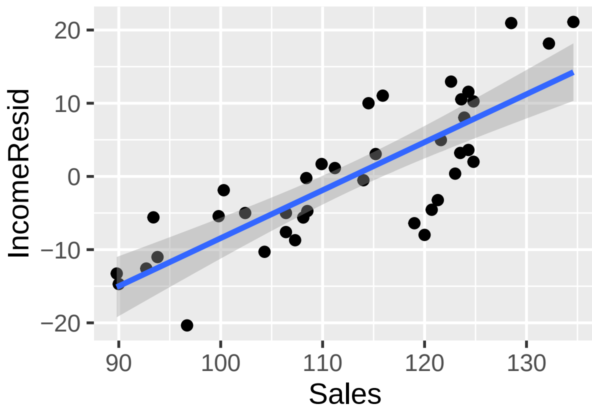


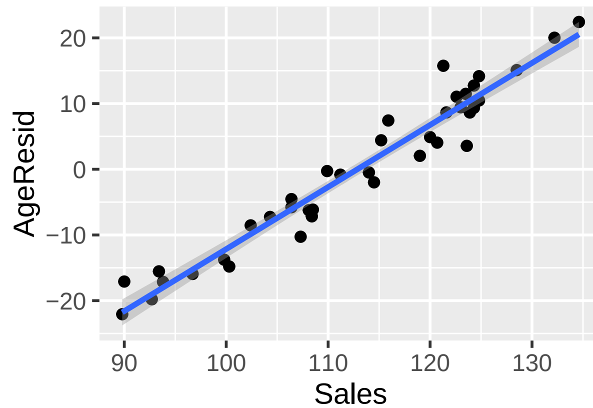
This also gives us a better adjusted R-squared: 0.2488 vs 0.1678 when just looking at income.

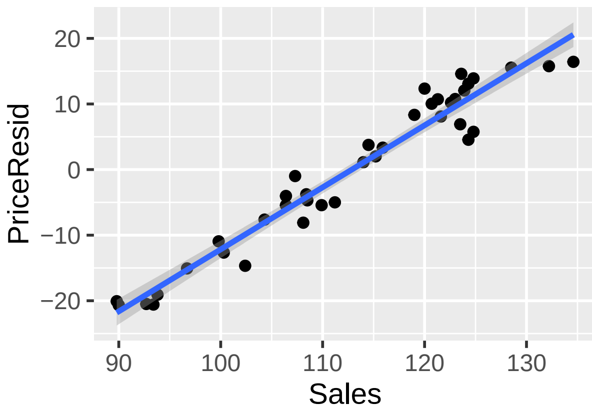
Lastly, I'll look at the correlation between my predictor variables.

	Age	Income	Price	Sales
Age	1.0000000	0.2837023	0.3417938	0.2357681
Income	0.2837023	1.0000000	0.1886738	0.5868890
Price	0.3417938	0.1886738	1.0000000	-0.2327881
Sales	0.2357681	0.5868890	-0.2327881	1.0000000

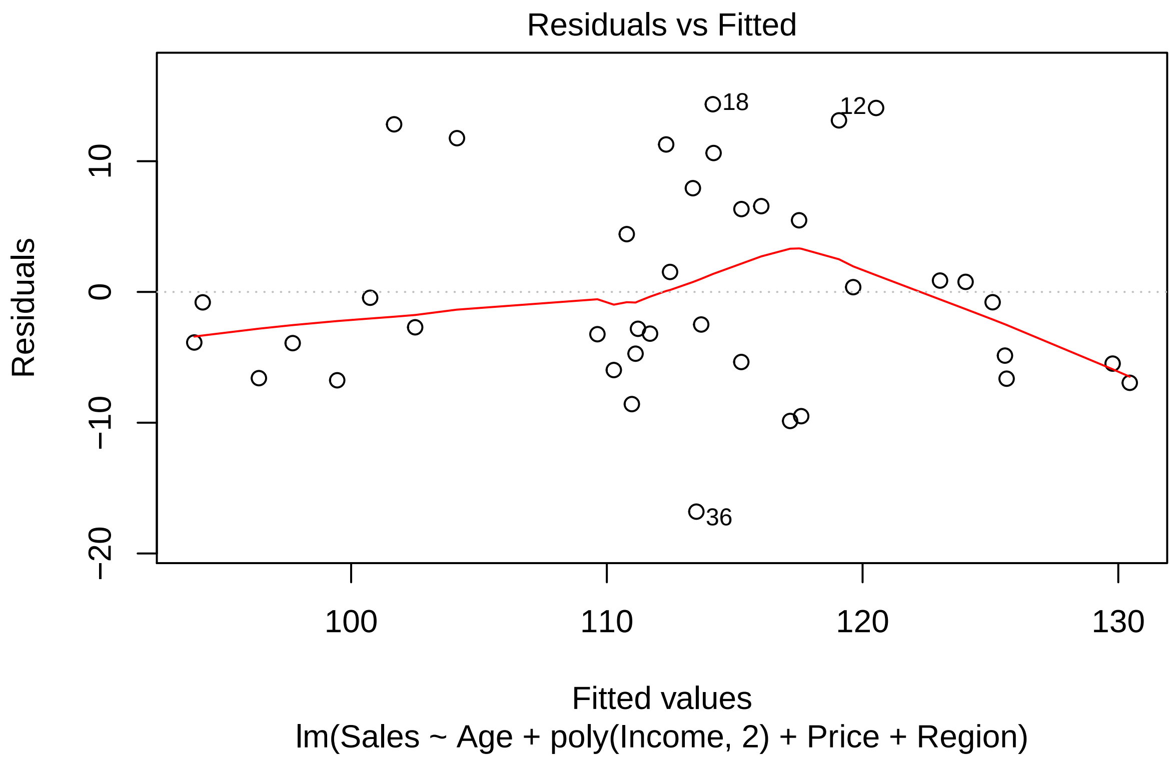
Nothing seems abnormally correlated, but it might be worth it to keep an eye on Price and Age.

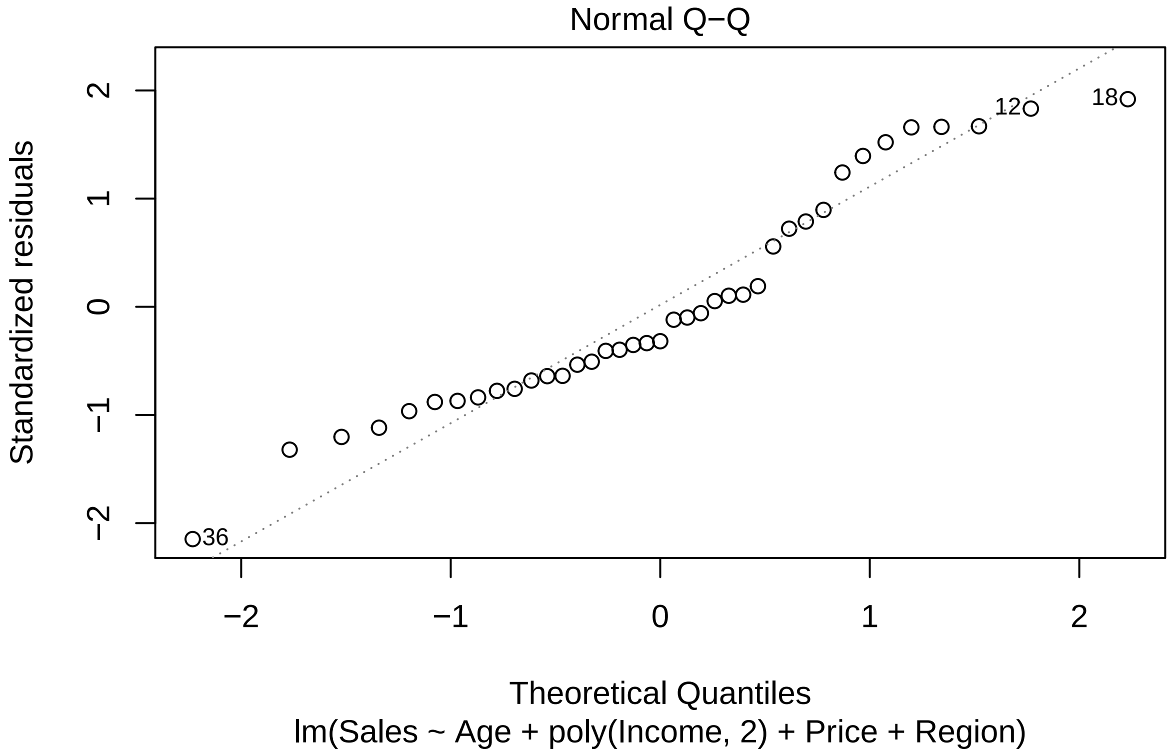


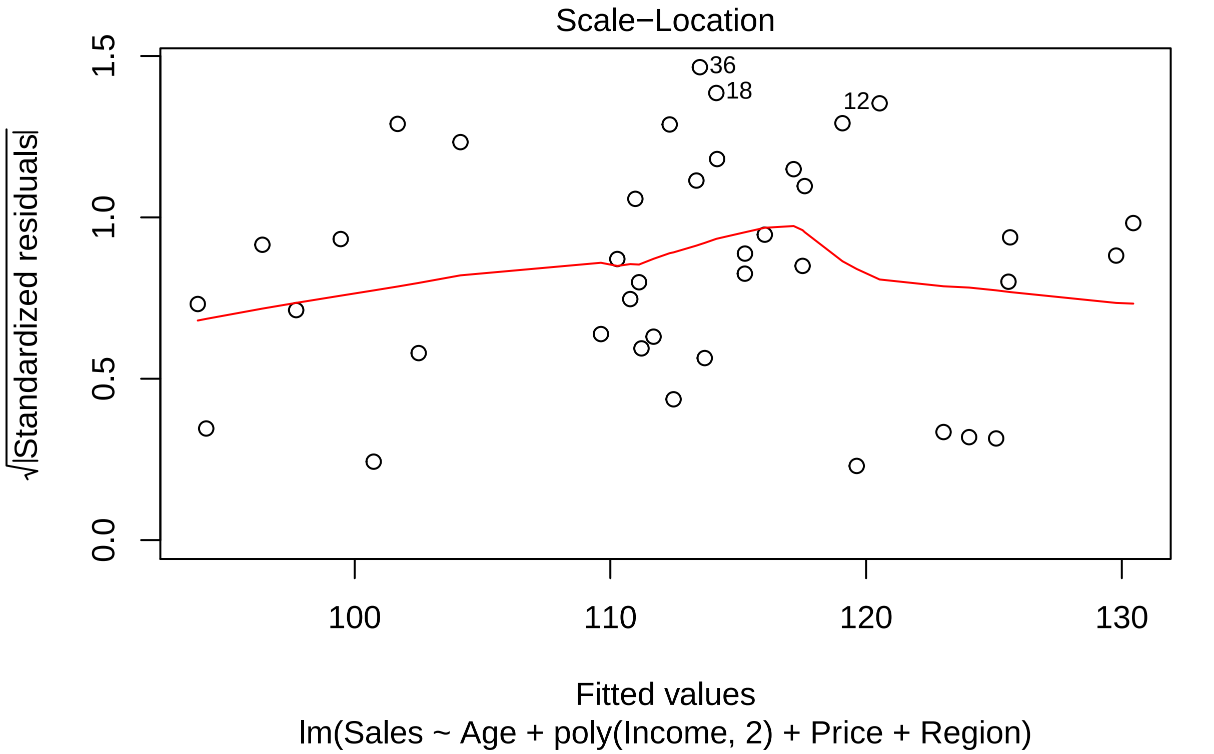


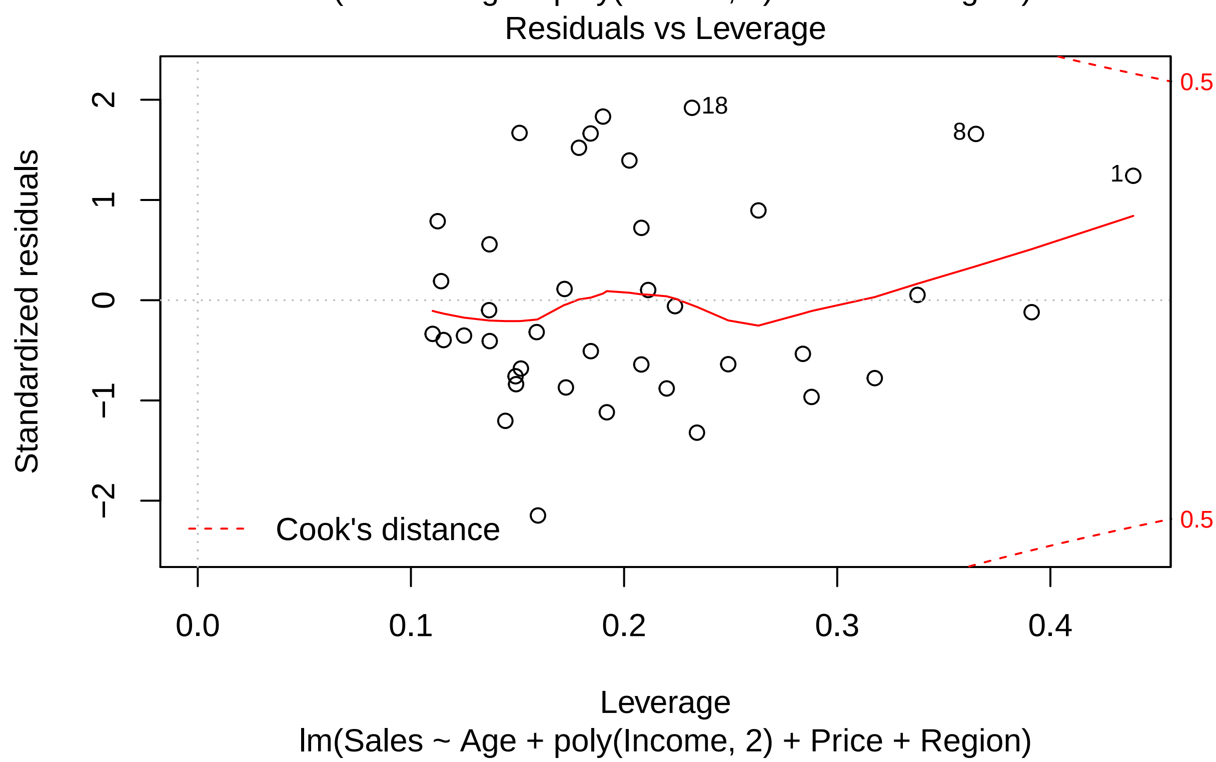



```
##
## Call:
## lm(formula = Sales ~ Age + poly(Income, 2) + Price + Region,
##     data = filteredcigarettesregion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.800  -5.418  -2.490   5.909  14.357
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      166.6254    29.8005   5.591 3.94e-06 ***
## Age                0.2941     0.9637   0.305  0.76226
## poly(Income, 2)1  47.8108    10.6166   4.503 8.87e-05 ***
## poly(Income, 2)2 -16.0757    10.1200  -1.589  0.12232
## Price             -1.4308     0.4669  -3.064  0.00449 **
## RegionMidwest    -10.9924     4.6339  -2.372  0.02407 *
## RegionSouth       -3.6209     4.7377  -0.764  0.45049
## RegionWest        -9.8849     5.2049  -1.899  0.06688 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.534 on 31 degrees of freedom
## Multiple R-squared:  0.6015, Adjusted R-squared:  0.5115
## F-statistic: 6.684 on 7 and 31 DF,  p-value: 7.389e-05
```


Residuals vs Fitted
Residuals
Fitted values
lm(Sales ~ Age + poly(Income, 2) + Price + Region)
18
12
36

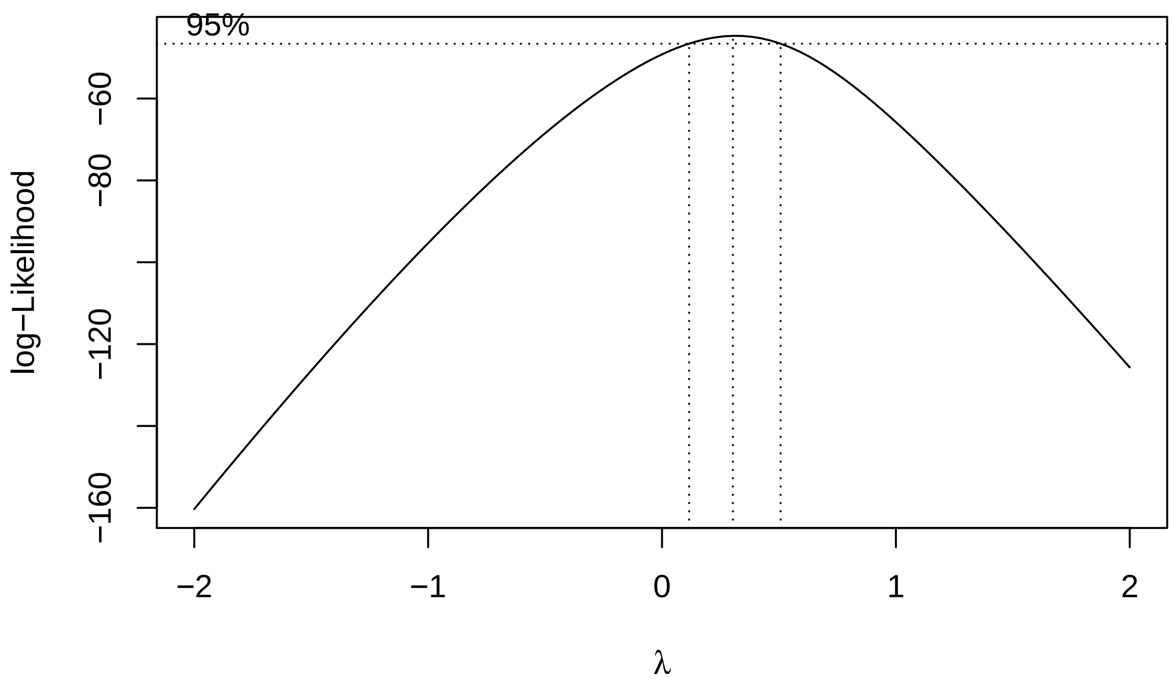

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ Age + poly(Income, 2) + Price + Region)
12
18
36






```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gala)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Area        -0.023938   0.022422  -1.068 0.296318
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```




```
##
## Call:
## lm(formula = I(Species^(1/3)) ~ Area + Elevation + Nearest +
##     Scruz + Adjacent, data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54306 -0.47863 -0.08499  0.56349  1.83283
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.2479224  0.3052013   7.365 1.32e-07 ***
## Area        -0.0007349  0.0003573  -2.057  0.05070 .
## Elevation    0.0054510  0.0008551   6.375 1.37e-06 ***
```

```
## Nearest      0.0118152  0.0167965   0.703  0.48855
## Scruz       -0.0045951  0.0034322  -1.339  0.19317
## Adjacent    -0.0010597  0.0002820  -3.757  0.00097 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9716 on 24 degrees of freedom
## Multiple R-squared:  0.7543, Adjusted R-squared:  0.7032
## F-statistic: 14.74 on 5 and 24 DF,  p-value: 1.192e-06
```

```
##
## Call:
## lm(formula = Species13 ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54306 -0.47863 -0.08499  0.56349  1.83283
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.2479224  0.3052013   7.365 1.32e-07 ***
## Area        -0.0007349  0.0003573  -2.057  0.05070 .
## Elevation    0.0054510  0.0008551   6.375 1.37e-06 ***
## Nearest      0.0118152  0.0167965   0.703  0.48855
## Scruz       -0.0045951  0.0034322  -1.339  0.19317
## Adjacent    -0.0010597  0.0002820  -3.757  0.00097 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9716 on 24 degrees of freedom
## Multiple R-squared:  0.7543, Adjusted R-squared:  0.7032
## F-statistic: 14.74 on 5 and 24 DF,  p-value: 1.192e-06
```