IS621_hw1_completed

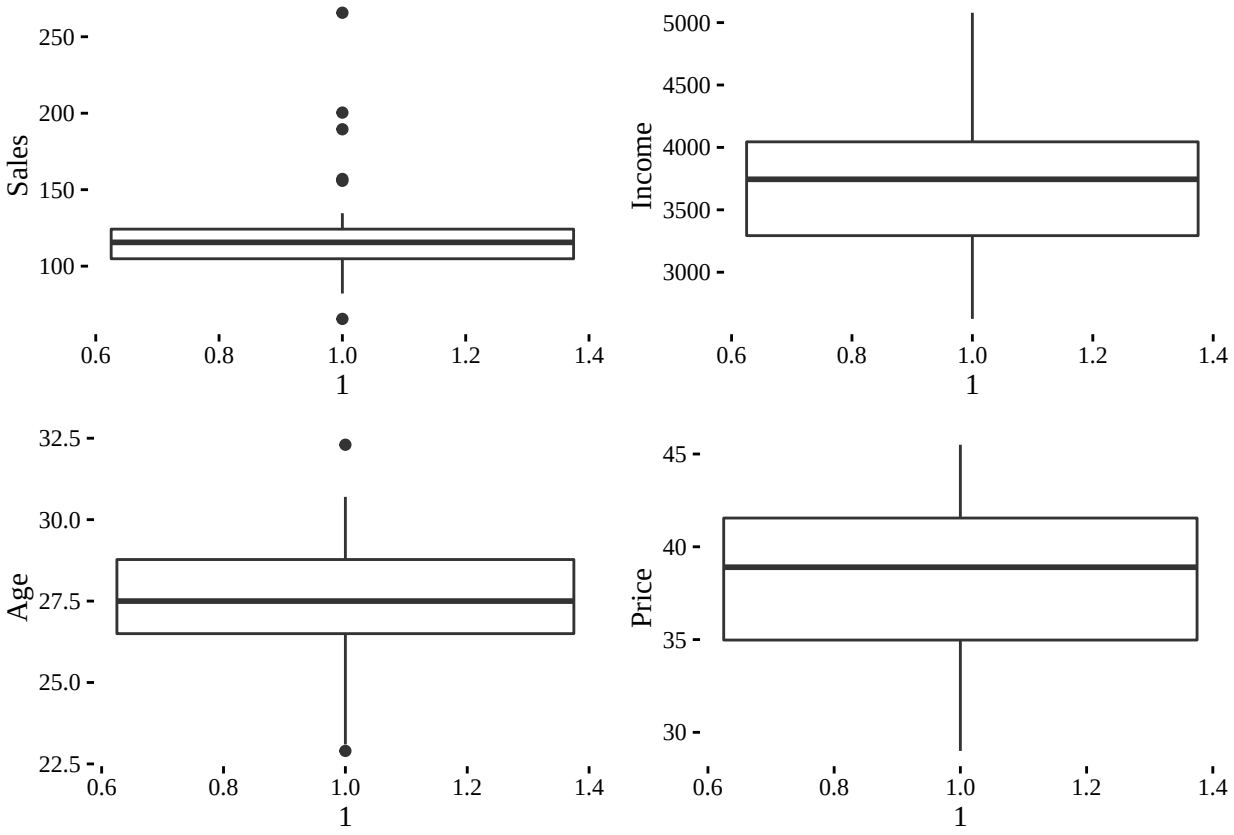
Charley Ferrari

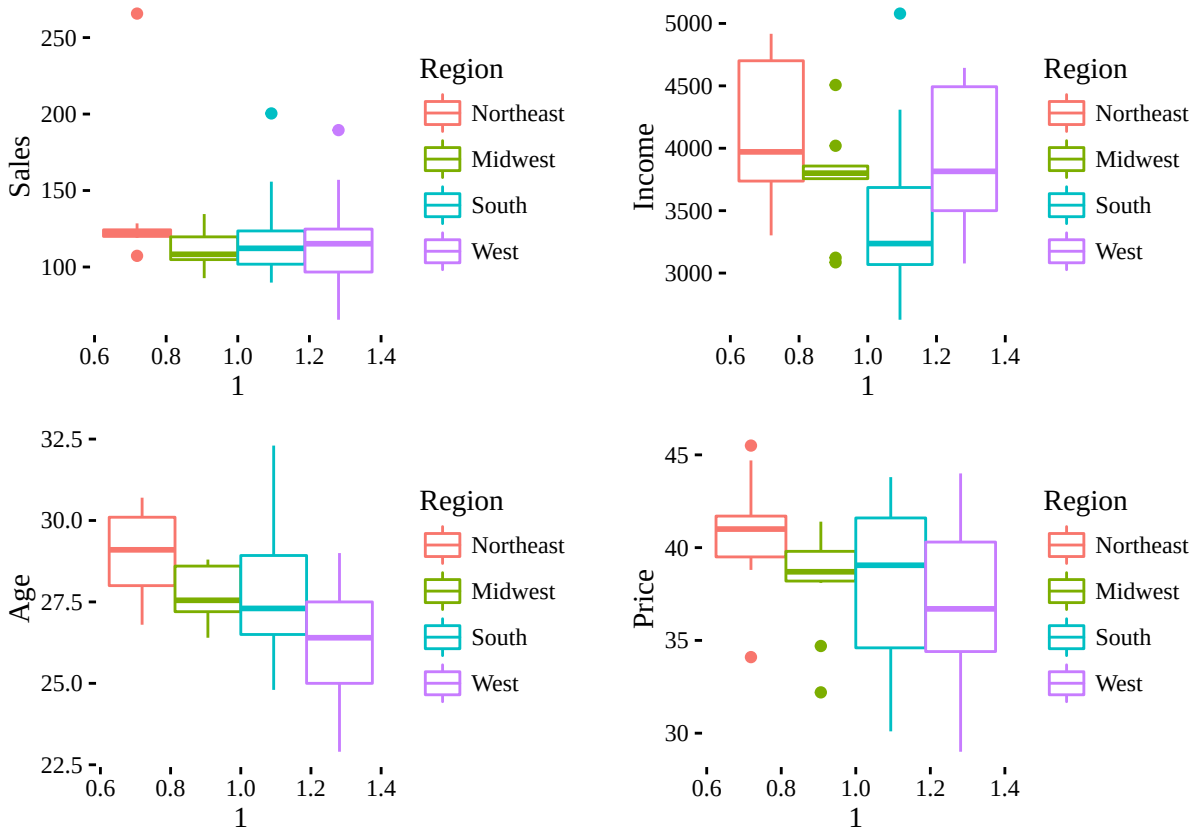
February 9, 2016

Data Exploration and Data Transformation

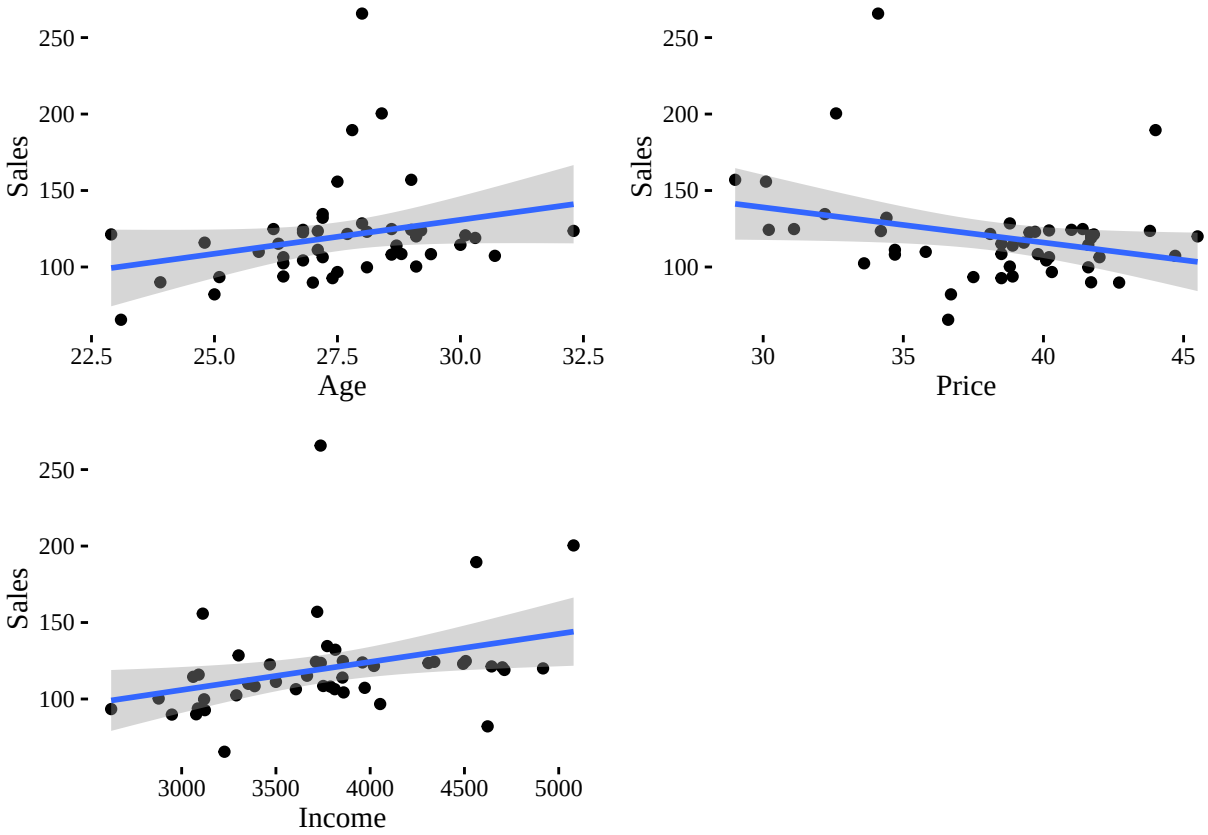
The cigarettes data set consists of five variables: State, Age, Income, Price, and Sales. All of these variables except for State are continuous. Since there is one categorical observation per state however, this is not useful for analysis. In order to test if there are any geographical effects in this 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.



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

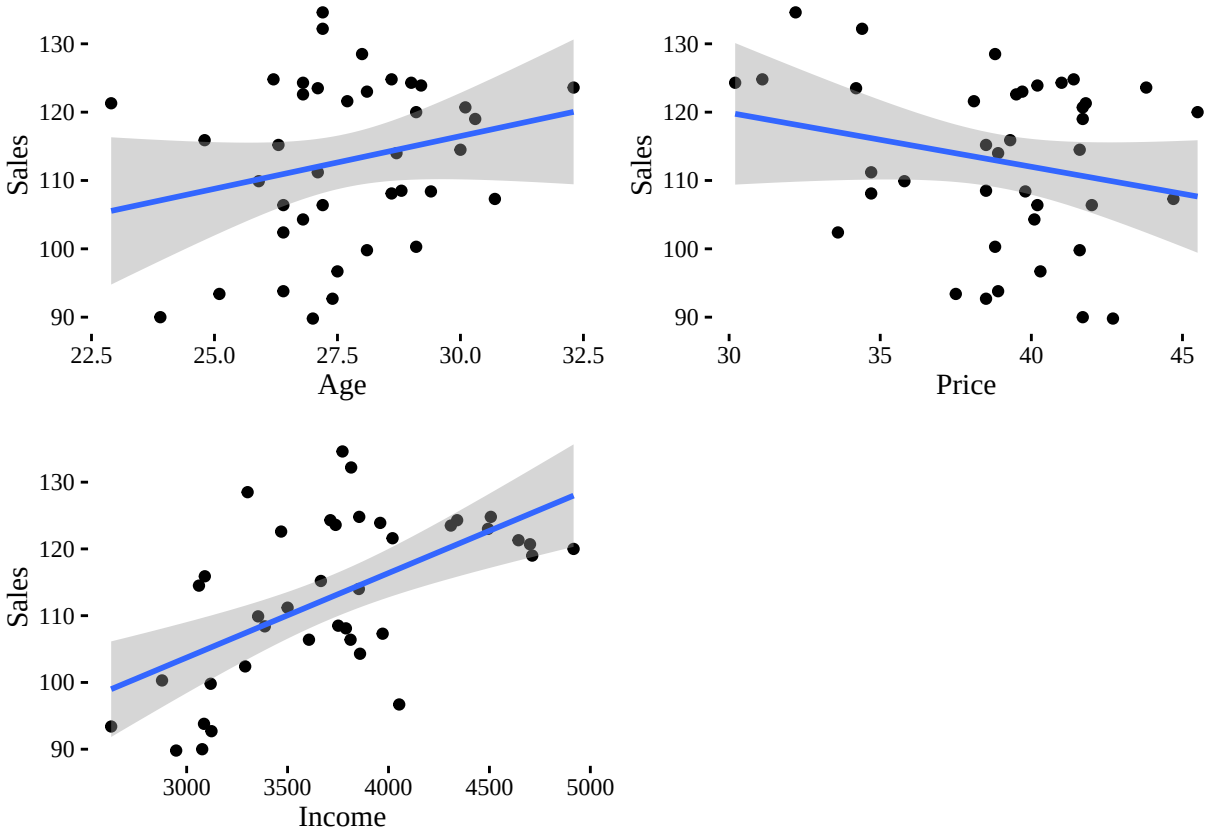
State	Age	Income	Price	Sales
DC	28.4	5079	32.6	200.4
KY	27.5	3112	30.1	155.8
NV	27.8	4563	44.0	189.5
NH	28.0	3737	34.1	265.7
OR	29.0	3719	29.0	157.0
UT	23.1	3227	36.6	65.5

A few of these can 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

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.)

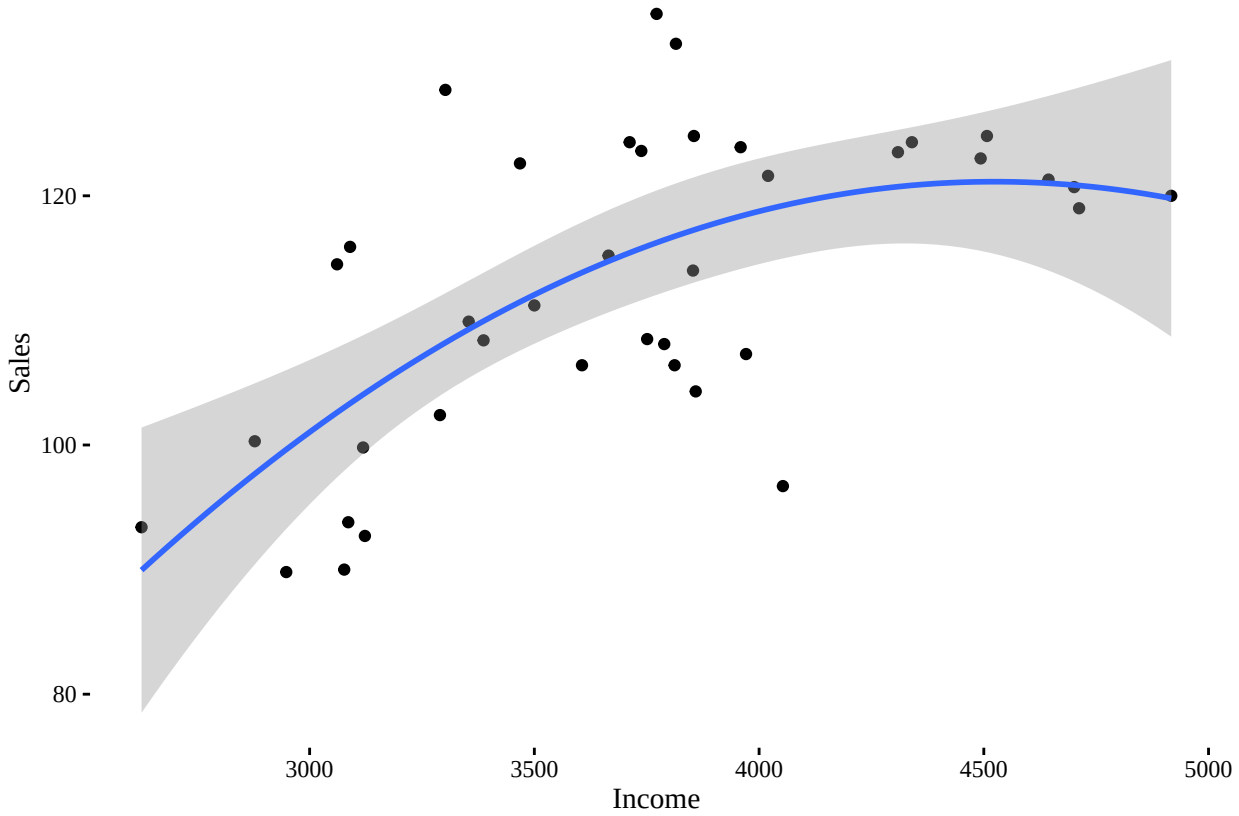
I will revisit these outliers, but for now after removing them the scatter plots show clearer relationships:



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.

I'm making the choice to include both Income and $Income^2$. When I run a model with just poly(Income,2) and Sales, I get the following equation: $Sales = 119.95 + 73.551 \times Income - 8.964 \times Income^2$. In several test models, $Income^2$ shows low significance. To fit the justification above, both of these variables need to be included. Income's positive coefficient describes the effect of having more disposable income, while $Income^2$'s negative coefficient describes the tapering off of this effect (or potential negative corellation.)

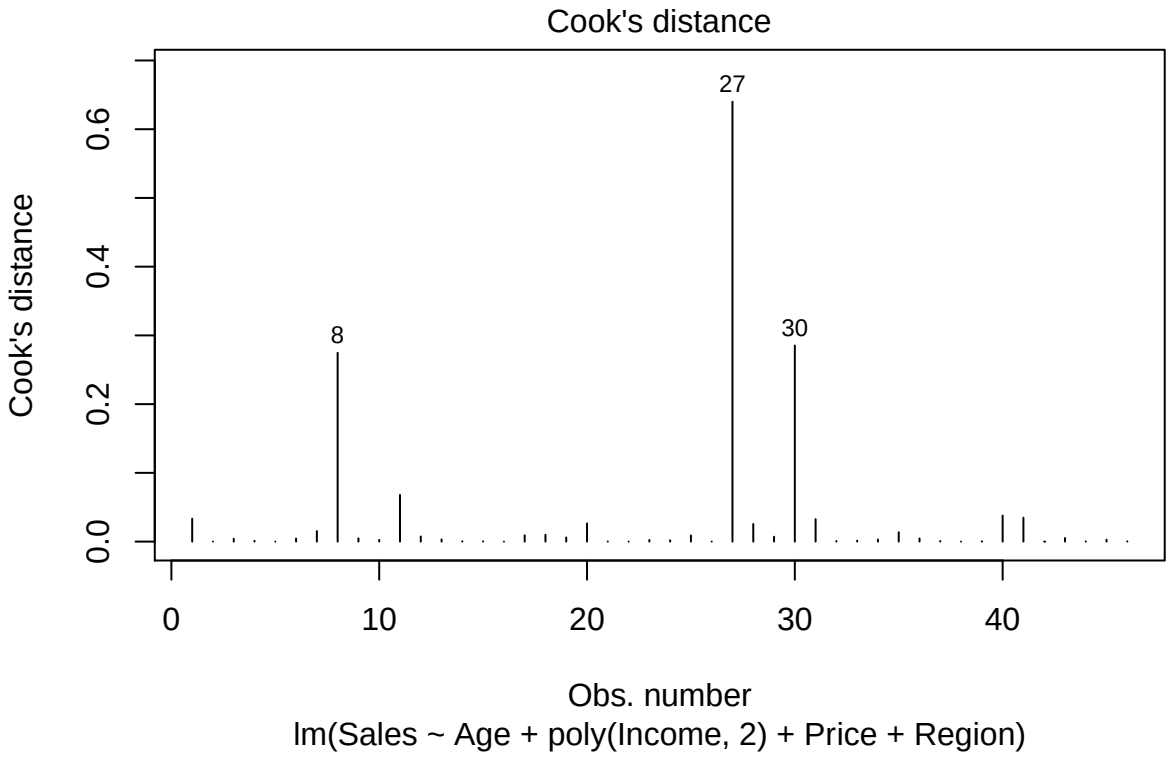
Lastly, I'll look at the correlation between my predictor variables.

	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.

Now that we have a base model, we can test the cook's distance of our outliers. Using poly(Income,2) to include both Income and $Income^2$, as well as Region as a categorical variable, I'll create a model to try to predict Sales.

Below is a graph of the Cook's distances, and a table of the top 10 cook distances for this model:



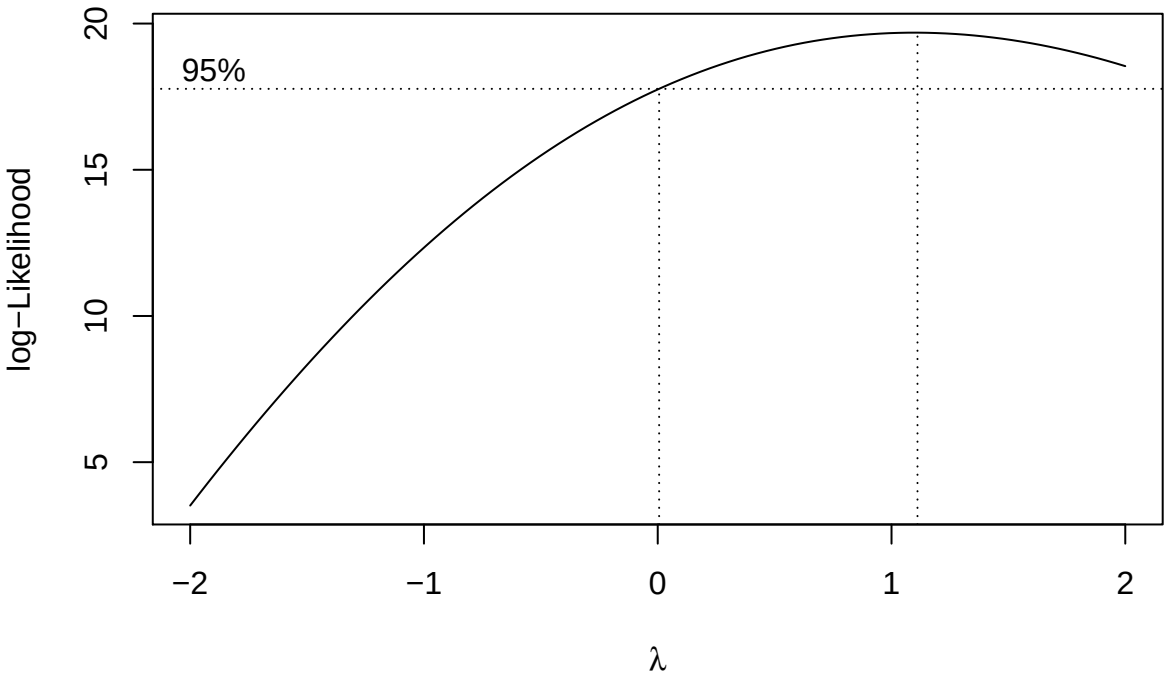
State	Age	Income	Price	Sales	Region	Cooks.Distance
NH	28.0	3737	34.1	265.7	Northeast	0.6400453
NV	27.8	4563	44.0	189.5	West	0.2853320
DC	28.4	5079	32.6	200.4	South	0.2746808
HI	25.0	4623	36.7	82.1	West	0.0679632
UT	23.1	3227	36.6	65.5	West	0.0378684
VA	26.8	3712	30.2	124.3	South	0.0347405
AK	22.9	4644	41.8	121.3	West	0.0333348
NY	30.3	4712	41.7	119.0	Northeast	0.0326281
MD	27.1	4309	34.2	123.5	South	0.0264615
NJ	30.1	4701	41.7	120.7	Northeast	0.0255985

The table and graph suggest removing the top 4 states: New Hampshire, Nevada, Wsashington DC, and Hawaii. This is interesting because Hawaii was within 1.5IQR in my first analysis of outliers, while other outliers such as Utah had less extreme Cook's Distances. This changes the relative Cook's Distances:

State	Age	Income	Price	Sales	Region	Cooks.Distance
AK	22.9	4644	41.8	121.3	West	0.3339042
UT	23.1	3227	36.6	65.5	West	0.2422717
KY	27.5	3112	30.1	155.8	South	0.1920800
OR	29.0	3719	29.0	157.0	West	0.1520540
VA	26.8	3712	30.2	124.3	South	0.0856215
MD	27.1	4309	34.2	123.5	South	0.0731025
ME	28.0	3302	38.8	128.5	Northeast	0.0700159
LA	24.8	3090	39.3	115.9	South	0.0481287
VT	26.8	3468	39.5	122.6	Northeast	0.0459474
NY	30.3	4712	41.7	119.0	Northeast	0.0357764

Now Alaska is topping the list. However, Alaska sales are within the IQR of our set of cigarette sales. Since we have a smaller data set, I believe it's better to stick with the four outliers we removed based on our first analysis of Cook's Distance.

Lastly, I'll run a Box-Cox on my model to see if I need to transform my dependent variable:



We have a wide 95% confidence interval centered very close to 1, suggesting no transormation is needed. Because the interval included 0 and 2 however, I tested squaring and taking the natural logarithm of Sales. When I inserted $\frac{Sales^2-1}{2}$ as my dependent variable, I ended up with an adjusted R-squared of 0.481, with log(Sales) I ended up with an adjusted R-squared of 0.4711, which both failed to beat my base model's adjusted R-squared of 0.4869.

To summarize, my base model, including all variables, is specified as:

$$Sales\tilde{A}ge + Income + Income^2 + Price + Region$$

Region is a categorical variable, which is expressed through the dummies RegionMidwest, RegionSouth, and RegionWest (leaving out the Northeast.)

Build Models

For the purposes of this section, I will remove the categorical variable Region from my model (according to the p-values I calculated, none of the Region dummy variables were significant anyway.) I will keep Income as a polynomial however, and either keep both Income and $Income^2$ or remove them both. In several test models, I have found that $Income^2$ has a low significance, while Income has a high significance. If I include only $Income^2$ however, I end up with a positive income coefficient. I'm treating these as a singular transformation to account for the behavior I described above.

Below is a summary of my base model:

```
##
## Call:
## lm(formula = Sales ~ Age + poly(Income, 2) + Price, data = filteredcigarettesregion)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.982  -7.018  -1.734   6.772  28.300
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      116.8950    31.6732   3.691 0.000716 ***
## Age                3.3773     1.0791   3.130 0.003407 **
## poly(Income, 2)1  44.1614    13.0800   3.376 0.001739 **
## poly(Income, 2)2   4.8153    13.4448   0.358 0.722263
## Price             -2.5039     0.5461  -4.585 5.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.29 on 37 degrees of freedom
## Multiple R-squared:  0.5211, Adjusted R-squared:  0.4693
## F-statistic: 10.06 on 4 and 37 DF,  p-value: 1.295e-05
```

Accounting for my transformation of Income, all variables are highly significant, with p-values less than 0.01. It would still be helpful to compare the Adjusted R-squares of each of the 7 iterations of variable combinations:

Age	Income	Price	Adjusted.R
TRUE	TRUE	TRUE	0.4692825
FALSE	TRUE	TRUE	0.3464452
TRUE	FALSE	TRUE	0.3406615
TRUE	TRUE	FALSE	0.1896485
TRUE	FALSE	FALSE	0.1037836
FALSE	TRUE	FALSE	0.1557890
FALSE	FALSE	TRUE	0.1284038

I checked for corellation earlier, but now that we have removed outliers, it would make sense to revisit our corellation matrix:

	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

Now that I'm fairly convinced that the base model is best, I could calculate the variance inflation for each variable:

	Age	poly(Income, 2)1	poly(Income, 2)2	Price
vif.model.	1.19398	1.133307	1.197416	1.307523

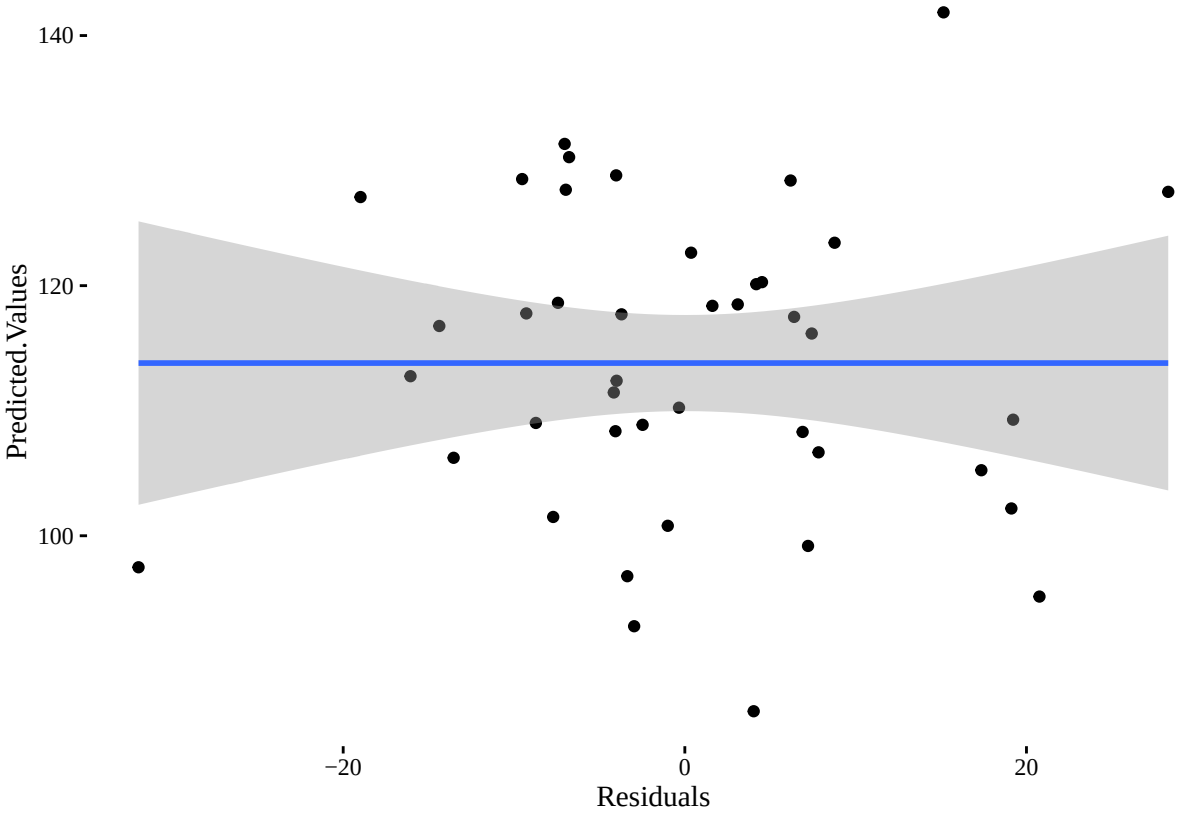
Using the rule of thumb cutoff of ~10, it's pretty clear I still don't have to worry about multi-collinearity issues.

Select Models

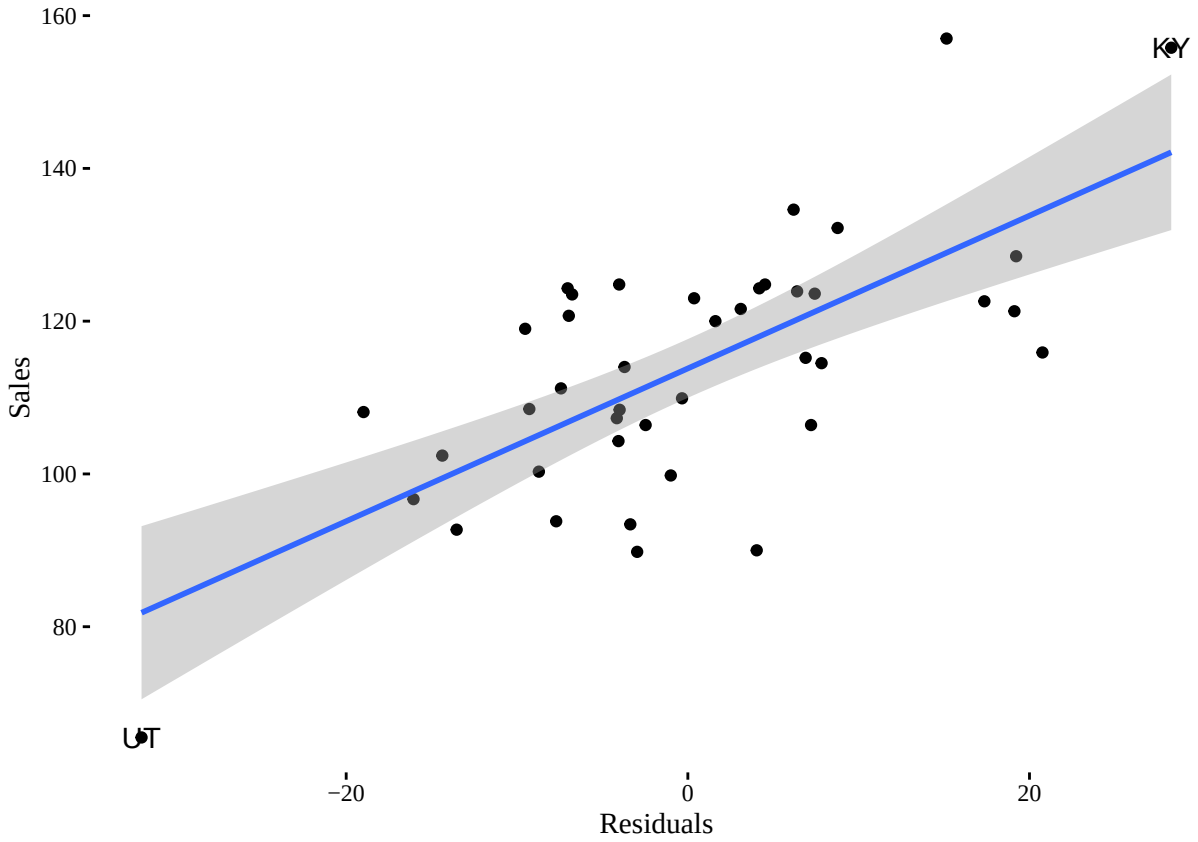
I am basing my model selection mainly on the significance of my variables, and the Adjusted R-squared. Both of these measures seem to be in sync. My base model including all variables showed all variables as significant, and the systematic removal of variables resulted in reductions in my Adjusted R-squared. Models with two variables performed better than models with one variable.

One last sanity check is to plot the residuals with our predicted values, to make sure there are no patterns in the data.

The plot shows no clear pattern, suggesting a sound model choice.



However, when I plot the residuals versus the actual values, the outliers exert more “bad leverage” on the fit:



It would appear that Kentucky and Utah, two outliers identified using my 1.5IQR cutoff, are causing this pattern. As predicted in my explanation above, Utah is being over estimated based on my theory that the high Mormon population causes a reduction in cigarette consumption, while Kentucky is being under estimated due to the population of neighboring states crossing borders to buy cigarettes.

In order to improve this model, it would probably make sense to perform more research on these two effects. While Utah's Mormon character is truly unique, it might make sense to use church attendence, or other religious measures as something that can apply to all states. There could also be a measure of how a state's Price compares to neighboring states, and whether or not there is an anomaly.