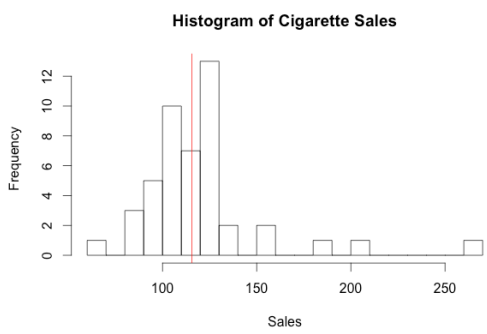
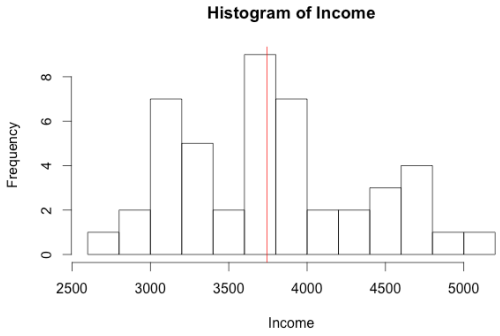
David Stern
DATA 621: Business Analytics and Data Mining
MSc Data Analytics, CUNY SPS

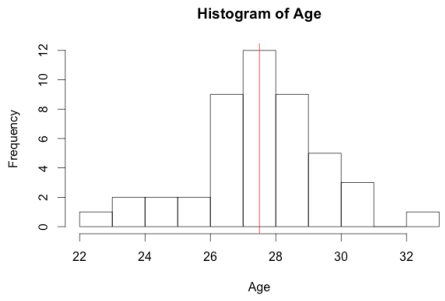
Data Exploration

Our dataset consists of four variables describing cigarette usage in each of the 50 states and Washington D.C in 1970. The variables include the median age, per capita incomes, average sales price and per capita sales for each state. The data is split into training and evaluation sets with sales data removed from the latter.

Here we have a histogram of the cigarette sales. The vertical red line indicates the median, 115.55. We see that the distribution is slightly skewed to the right as the long tail towards higher prices brings the mean sales price to 119.95. The distribution of the Income seems to be somewhat uneven. There is some concentration around the median, $3744.50, and little skew with the mean of $3765.70. The ages are similarly distributed with a mean and median of 27.5. We can see immediately from the x-axis, however, that this data may not be very helpful in building a model. The range of ages is just under ten years, since the data point is merely the median age for each state. Last we have the price distribution, with a mean and median around 38 cents/pack and a range across states from a floor of 29 cents to a max of 46 cents.





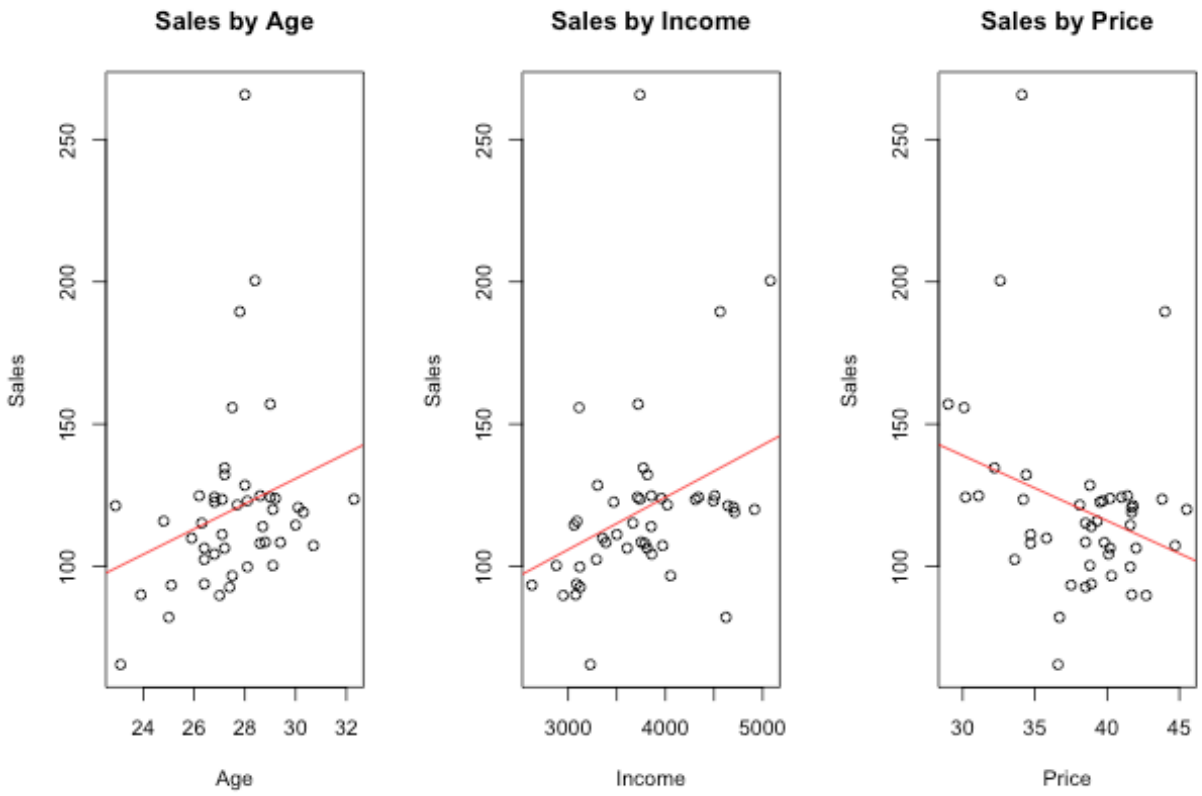


Data Preparation

The data has been split into a training set with 46 observations and the evaluation set, which includes five states without data on per capita sales. Both sets are well formatted and there did not appear to be any missing or obviously aberrant data. No changes were made to either set.

To see if any transformations were needed to the independent variables, I fit each predictor individually to the sales data and checked the assumptions for a linear regression model.

Linearity and Independence of Errors: Each of the variables demonstrates a weak linear relationship. In each of the plots, there seem to be a few very large residuals, but they do not appear to demonstrate any particular pattern. We will assume these errors are independent.

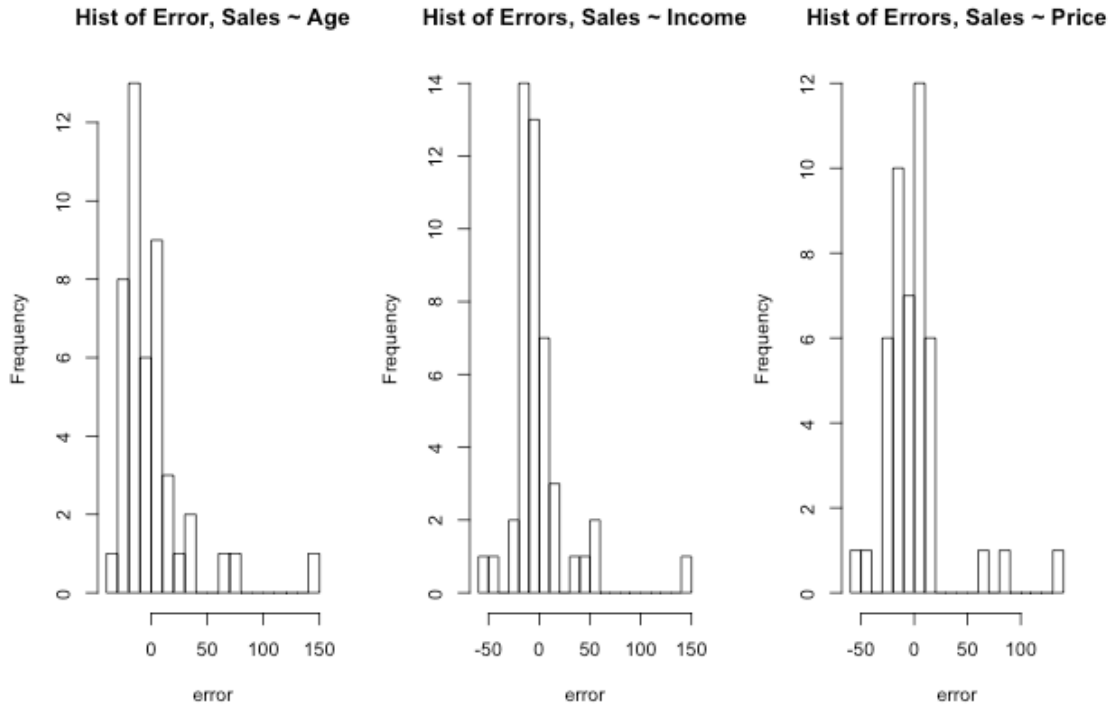


The following correlation matrix will give us an early hint at whether we can expect multicollinearity. (We will later use variance inflation factors (VIFs) to detect multicollinearity in our fitted models.) Here we include the target variable, sales, but we are looking to see whether there is a high degree of correlation between the predictive variables. Both age and income have weak positive correlations sales: 0.26 and 0.34, respectively. Price has a negative correlation, as we would expect for a demand curve, at -0.29. The absolute value of the values are below 0.40, so we should not expect to see too much of a collinear effect in a multiple regression model.

	Age	*Income*	*Price*	*Sales*
Age	1	0.248959	0.237582	0.259996
Income	0.248959	1	0.146552	0.338436
Price	0.237582	0.146552	1	-0.289126
Sales	0.259996	0.338436	-0.289126	1

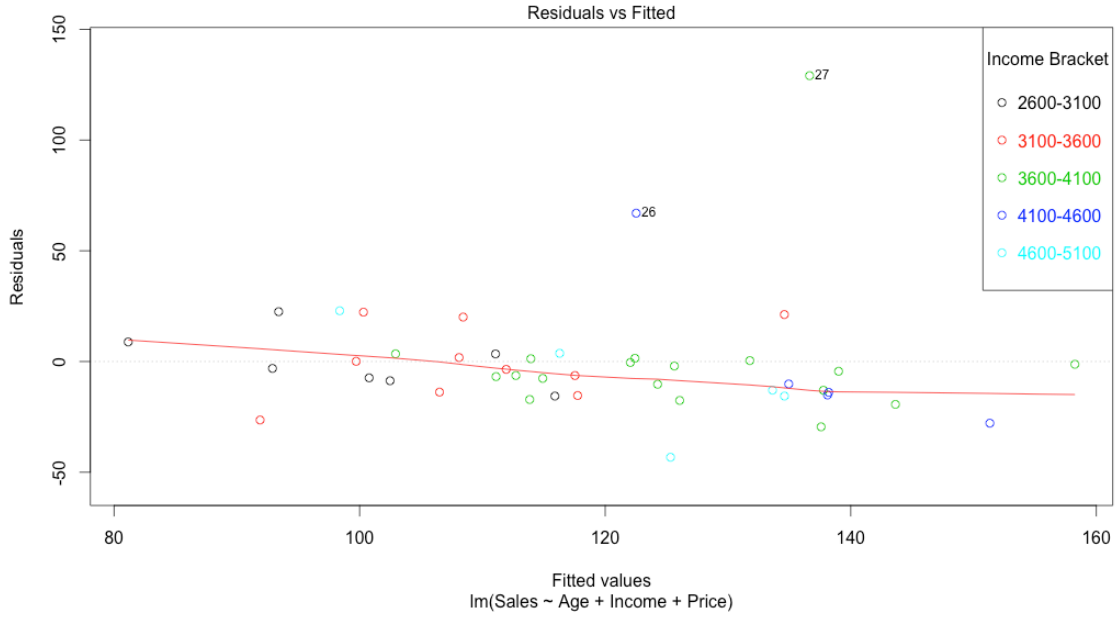
Normality of Error Distribution: The three independent variables all seemed to be normally distributed with a mean near or at zero.

``````


``````

Homoscedasticity: As we noted above, there does not appear to be a pattern in the size of the residuals. We will confirm this by examining a plot of the square of the residuals for each. Here we see that a slope of zero means the data is homoscedastic.

After fitting the model, I wanted to transform the Income data from continuous to categorical brackets to see if there was any pattern in the plot of the residuals versus the fitted values for the model. I did so by taking the range of income by state, approximately 2500, and bucketing the data into five uniform $500 income brackets. In the following plot, we see that there isn't a pattern in the distribution of residuals by income bracket, and that it does appear to be random.



Building Models

Before attempting to build a multiple linear regression model, I built three simple linear regression models for each of the independent variables – age, income, and price, as predictors of Sales. Using the summary function for linear regression models in R.

model	model p-value	f-stat	R-squared	predictor p-value	RSE
Age	0.08099	3.19	0.0676	0.08099	31.64
Price	0.05132	4.014	0.08359	0.05132	31.36
Income	0.02142	5.692	0.1145	0.02142	30.83

Here we see that the predictors do not have much predictive power individually. Income has the greatest R-squared value and is the only of the three to have a p-value below an acceptable level of significance (0.05). Here the p-values for the predictors are the same as the p-values for the model.

Now we will examine the various combinations of variables. In the following four combinations of the predictor variables, we see the p-values for each of the models begin to drop and the R-squared goodness of fit begin to rise. The model with the highest R-squared value is the multiple regression model that simply includes all three variables. In the models with two-predictors, we see that their may be some interactions between the variables that increases the fit of the model. This is best demonstrated by the *Age + Income* model, with a significant p-value for the model, but p-values for the predictors that exceed 0.05.

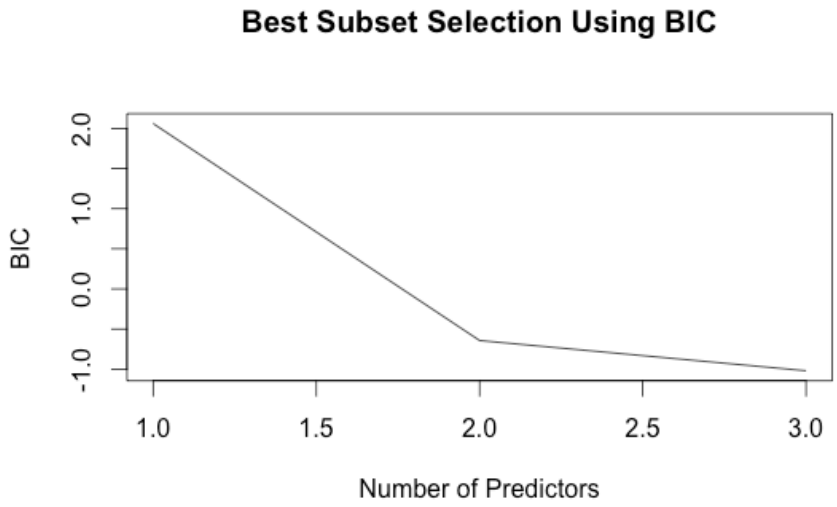
model	**model p-value**	**f-stat**	**R-squared**	**predictor p > 0.05**	**RSE**
Age + Income	0.03238	3.719	0.1475	Age (0.2044), Income(0.0511)	30.6
Price + Income	0.0345	6.487	0.23118		29.05
Age + Price	0.008683	5.311	0.1981		29.68
Age + Price + Income	0.001744	5.969	0.2989	Age (0.05139)	28.08

Select Models

Of these seven models, I believe the best fit to be the full regression model using all three predictor variables. Although the age variable is slightly greater than 0.05, it's R-squared value is about 25% greater than the model that includes price and income but excludes age. This model also has the lowest residual standard error (RSE) of the seven combinations of predictors. The calculated VIFs for the full regression model also returned very low values for each of the predictor variables, so we can conclude that there is very little effect of multicollinearity in this model.

VIF *lm(Sales ~ Age + Price + Income)*		
Age	**Income**	**Price**
1.115210	1.075358	1.069049

Here I also used Bayesian information criterion (BIC) to determine which subset of predictor variables will return the best model. The plot below demonstrates that the best subset – the one that returns the lowest BIC – is the one with all three predictors.



I attempted to fit several variations of the model based on interactions with the three predictors, but none seemed to be a better model. The model with all possible interactions, *lm(Sales~Age*Income*Price,cig)*, had the highest R^2 value at 0.39 but the model seem highly biased. Each of the predictors had a p-value greater than 0.05 and VIFs in the thousands. The model itself returned a BIC of 461 (likely because our number of observations in the dataset is too small to be fit by a subset with a large number of predictors and interactions terms.. Based on these results, along with models that transformed age and took the ratio of price to income, I determined that the fit of the basic *lm(Sales ~ Age + Price + Income)* could not be improved. The coefficients for this model are as follows:

Regression Model Coefficient Estimates

Intercept	**Age**	**Income***	**Price***
46.59	4.67	0.02	-3.12

The coefficients in this model do appear to make sense. The inverse relationship between price and sales does reflect the nature of a demand curve. The intercept must be positive because we would expect the minimum possible value of sales to be zero. The fact that the intercept is not actually zero should not be troubling since we could not have a data point with zero values for each predictor variable (zero is not in the domain for our age values, for instance.) Using these coefficients we can predict the sales of the five states in the evaluation set:

DE	**MO**	**MI**	**SC**	**NC**
120.35	133.62	118.59	106.02	134.31

Summary

After evaluating the results of these results, I would decide to keep the present model. The mean squared error for the model is 719.84 and a standard error of 384.10. Although the R^2 value is relatively low at 0.2989, the p-value and f-statistic do indicate that the model is statistically significant. A plot of the distribution of the residuals appears to be normal (give the small sample size) and centered around a mean of 0.

Appendix:

My data analysis was performed using R. The packages and code used for my plots and analysis can be found here:

```
library(ggplot2)
library(car)
library(leaps)
library(coefplot)

cig <- read.csv("cigarette-training-data.csv",header=T)
eval <- read.csv("cigarette-evaluation-data.csv",header=T)

# correlation matrix

cor(cig[-1])

# histograms and summary stats

hist(cig$Sales, breaks=20, xlab="Sales", main="Histogram of Cigarette Sales")
abline(v=115.55,col="red")
summary(cig$Sales)
hist(cig$Income, breaks=10, xlab="Income", main="Histogram of Income")
abline(v=3744.5,col="red")
summary(cig$Income)
hist(cig$Age, breaks=12, xlab="Age", main="Histogram of Age")
abline(v=27.5,col="red")
summary(cig$Age)
hist(cig$Price, breaks=12, xlab="Price", main="Histogram of Price")
abline(v=38.9,col="red")
summary(cig$Price)

cor(cig$Age,cig$Sales)
cor(cig$Income,cig$Sales)
cor(cig$Price,cig$Sales)

fitAge = lm(Sales~Age,cig)
fitIncome = lm(Sales~Income,cig)
fitPrice = lm(Sales~Price,cig)

# Simple Plots

par(mfrow=c(1,3))
plot(Sales~Age,cig,main="Sales by Age")
abline(a=-2.216,b=4.435,col="red")
```

```
plot(Sales~Income,cig,main="Sales by Income")
abline(a=50.86239,b=0.01835,col="red")
plot(Sales~Price,cig,main="Sales by Price")
abline(a=207.919,b=-2.298,col="red")

# Error Distributions

par(mfrow=c(1,3))
hist(summary(fitAge)$residuals,xlab="error",main="Hist of Error, Sales ~
Age",breaks=20)
hist(summary(fitIncome)$residuals,xlab="error",main="Hist of Errors, Sales ~
Income",breaks=20)
hist(summary(fitPrice)$residuals,xlab="error",main="Hist of Errors, Sales ~
Price",breaks=20)

# Plots of Square of residuals

par(mfrow=c(1,3))
plot(summary(fitAge)$residuals^2,ylab="Residual Squared",main="Sales ~ Age")
plot(summary(fitIncome)$residuals^2,ylab="Residual Squared",main="Sales ~
Income")
plot(summary(fitPrice)$residuals^2,ylab="Residual Squared",main="Sales ~ Price")

# Bucket Income data

cig$Income[cig$Income <= 3100] <- "2600-3100"
cig$Income[(cig$Income <= 3600) & (cig$Income >= 3100)] <- "3100-3600"
cig$Income[(cig$Income <= 4100) & (cig$Income >= 3600)] <- "3600-4100"
cig$Income[(cig$Income <= 4600) & (cig$Income >= 4100)] <- "4100-4600"
cig$Income[(cig$Income <= 5100) & (cig$Income >= 4600)] <- "4600-5100"

fitAll = lm(Sales~Age+Income+Price,cig)

plot(fitAll, which=1, col=as.numeric(factor(fitAll$model$Income)))
legend("topright", legend=levels(factor(fitAll$model$Income)),
      pch=1, col=as.numeric(factor(levels(factor(fitAll$model$Income)))),
      text.col= as.numeric(factor(levels(factor(fitAll$model$Income)))),
      title="Income Bracket")

plot(fit1, which=2)

# Reload data so Income is numeric and continous

cig <- read.csv("cigarette-training-data.csv",header=T)

# Fit different models, "A" for Age, "I" for Income, "P" for Price in fit titles
```

```
fitAI= lm(Sales~Age+Income,cig)
fitAP= lm(Sales~Age+Price,cig)
fitPI= lm(Sales~Income+Price,cig)
fitA = fitAge
fitI = fitIncome
fitP = fitPrice
fitAPI = fitAll

# Try other variations:

fit1 = lm(formula = Sales ~ Age:Income + Price, data = cig)  # 1
fit2 = lm(formula = Sales ~ I(Price/Income), data = cig) # 3
fit3 = lm(formula = Sales ~ I(Price/Income) + I(Age^2), data = cig) #2
fitAllInteractions = lm(Sales~Age*Income*Price,cig) # all interactions

summary(fitAllInteractions) # repeat for all above
coefplot(fitAllInteractions)
BIC(fitAllInteractions)
vif(fitAllInteractions)

vif(fitAPI)

# BIC Plot

full <- regsubsets(Sales ~ Income+Age+Price, data = cig, nvmax = 10)
par(new=True)
plot(summary(full)$bic, xlab = "Number of Predictors", ylab = "BIC", type = "l",
    main = "Best Subset Selection Using BIC")


# Coefficients for full model:

fitAPI$coefficients

# Predict values of Evaluation Set

predict(fitAPI,eval)

mean((cig$Sales - predict(fitAPI, cig))^2) # Mean Predictor Error (test MSE) =
719.84
sd((cig$Sales - predict(fitAPI, cig))^2)/sqrt(46) # Standard Error = 384.1

hist(summary(fitAPI)$residuals,xlab="residual",main="Residual Plot, Sales ~ Age +
Price + Income",breaks=20)
```