Dan Sun (das225) & Yue Su (yus55)

Assignment 6: Probabilistic Approaches

We have collected data about List of countries by ***alcohol*** consumption per capita, List of countries by ***coffee*** consumption per capita, and List of countries by ***Nobel laureates*** per capita.

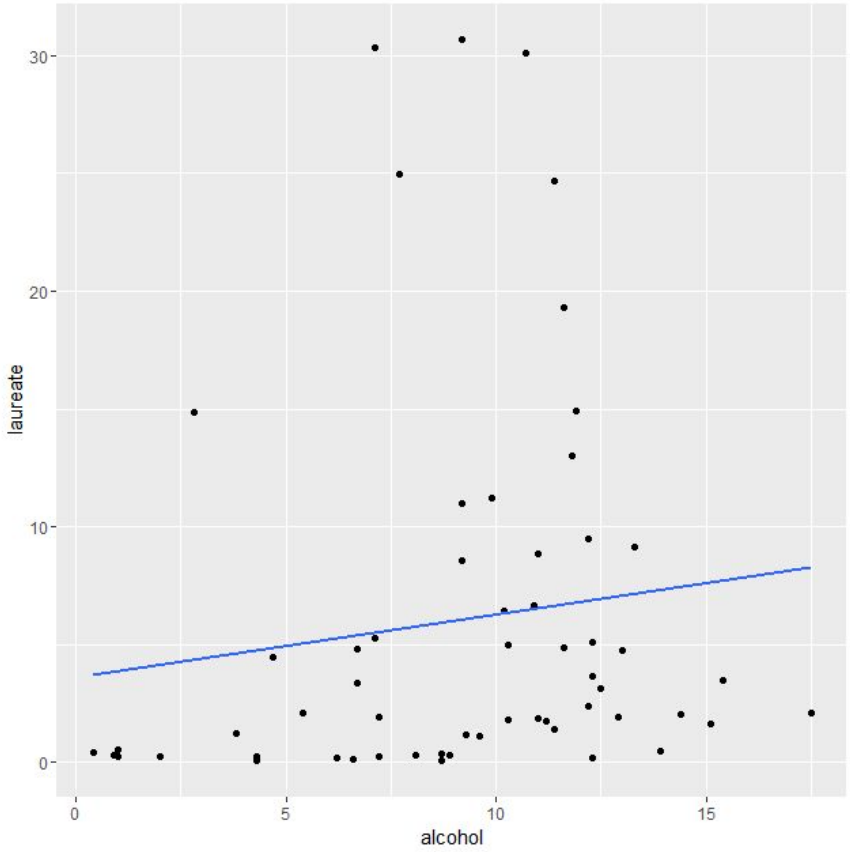
- Organize three dataset into one as a new .csv document

A	B	C	D
ountry	coffee	laureate	alcohol
Finland	12	3.634	12.3
Norway	9.9	24.947	7.7
Iceland	9	30.356	7.1
Denmark	8.7	24.695	11.4
Netherlands	8.4	11.226	9.9
Sweden	8.2	30.667	9.2
Switzerland	7.9	30.125	10.7
Belgium	6.8	8.85	11
Canada	6.5	6.4	10.2
Bosnia and Herzegovin	6.2	5.249	7.1
Austria	6.1	5.006	10.3
Italy	5.9	3.345	6.7
Slovenia	5.8	4.837	11.6
Brazil	5.8	0.048	8.7
Germany	5.5	13.031	11.8
Greece	5.5	1.826	10.3
France	5.4	9.473	12.2
Croatia	5.1	2.358	12.2
Cyprus	4.9	8.581	9.2
Spain	4.5	1.735	11.2
Portugal	4.3	1.932	12.9
United States	4.2	10.97	9.2
Macedonia	4.2	4.811	6.7
Lithuania	4.1	3.474	15.4
Czech Republic	4	4.742	13
Costa Rica	3.8	2.08	5.4
Israel	3.8	14.881	2.8
New Zealand	3.7	6.625	10.9
Algeria	3.5	0.504	1

- Import Data

```
> setwd("F:/【GRADUATE STUDIES】/Data Analytics/Assignment 6")
> nobel <- read.csv("nobel.csv",header=T)
```

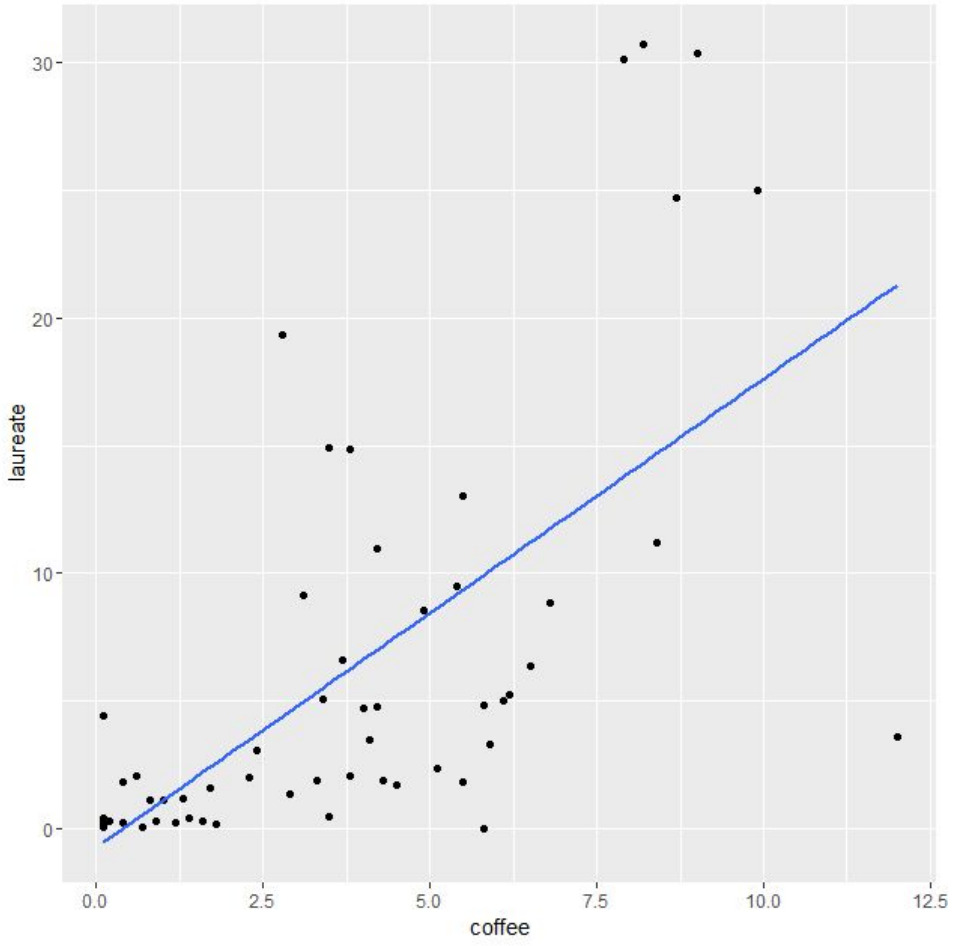
- Use ggplot to draw scatter plot between variables of alcohol and Nobel laureates



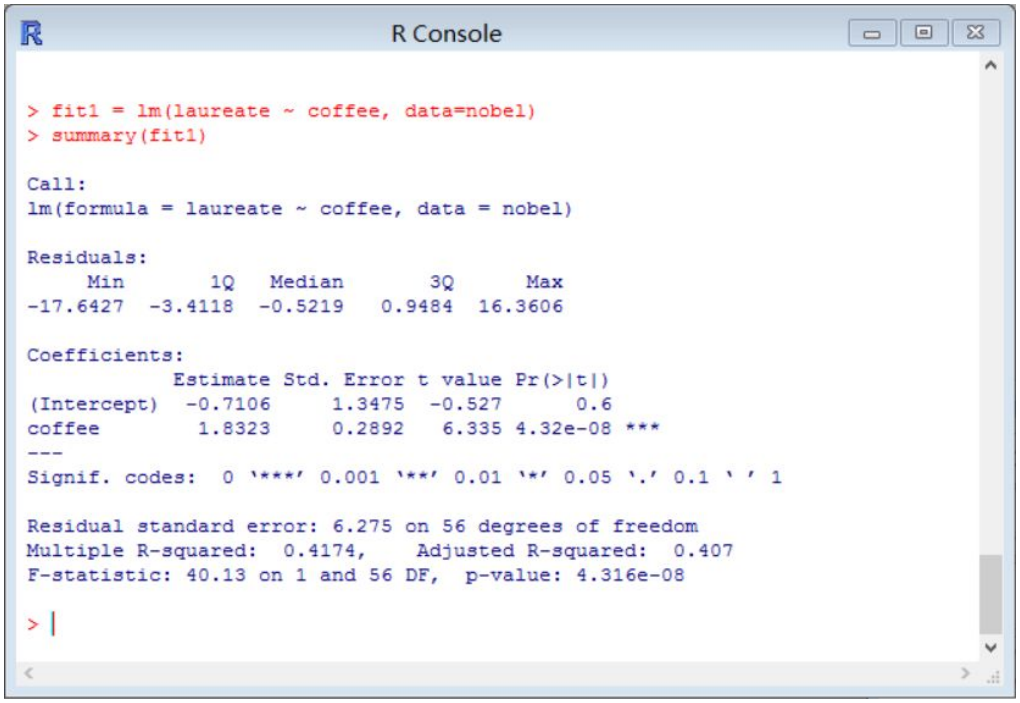
- Create linear model between variables of alcohol and Nobel laureates

```
> fit = lm(laureate ~ alcohol, data=nobel)
> summary(fit)

Call:
lm(formula = laureate ~ alcohol, data = nobel)

Residuals:
   Min     1Q Median     3Q    Max
-6.873 -5.030 -3.392  1.506 24.848

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6164     2.7198   1.330    0.189
alcohol       0.2664     0.2744   0.971    0.336

Residual standard error: 8.154 on 56 degrees of freedom
Multiple R-squared:  0.01655,   Adjusted R-squared:  -0.001012
F-statistic: 0.9424 on 1 and 56 DF,  p-value: 0.3358
```

- Use ggplot to draw scatter plot between variables of coffee and Nobel laureates



- Create linear model between variables of coffee and Nobel laureates



■ Analysis

□ Check the F statistic at the bottom of the summary

The F statistic tells us whether the model is insignificant or significant. Big p-value indicates a high likelihood of insignificance.

As the summary table shown that, alcohol consumption of a country has limited correlation with Nobel laureates. Coffee consumption has more relation with Nobel laureates than that with coffee.

□ Coefficients significant

If a variable's coefficient is zero then the variable is worthless; it adds nothing to the model. The regression coefficient shows that coffee performs more significant than alcohol in the model.

□ Check R-squared near the bottom of the summary.

R-squared is a measure of the model's quality. Bigger is better. The multiple R-squared indicates that the model accounts for coffee as 0.4174 much more than that of alcohol.

Check the residuals

The residual standard error can be thought of as the average error in predicting weight from height using this model. Apparently smaller will be better. Thus coffee consumption is more crucial for the model.

All in all, consumption of coffee in a country has strong correlation with Nobel laureates more than alcohol.