Population vs. Drinking Data

Original Dataset:

Urban population (percentage)=V3
Late births (reciprocal * 100)=V4
Wine consumption per capita=V5
Liquor consumption per capita=V6
Cirrhosis death rate=V7

1	1	44	33.2	5	30	41.2
2	1	43	33.8	4	41	31.7
3	1	48	40.6	3	38	39.4
4	1	52	39.2	7	48	57.5
5	1	71	45.5	11	53	74.8
6	1	44	37.5	9	65	59.8
7	1	57	44.2	6	73	54.3
8	1	34	31.9	3	32	47.9
9	1	70	45.6	12	56	77.2
10	1	54	45.9	7	57	56.6
11	1	70	43.7	14	43	80.9
12	1	65	32.1	12	33	34.3
13	1	36	36.9	10	48	53.1
14	1	47	38.9	10	69	55.4
15	1	63	47.6	14	54	57.8
16	1	35	33.0	9	47	62.8
17	1	50	38.9	7	68	67.3
18	1	55	35.7	18	47	56.7
19	1	33	31.2	6	27	37.6
20	1	81	53.8	31	79	129.9
21	1	63	42.5	13	59	70.3
22	1	78	53.3	20	97	104.2
23	1	63	47.0	19	95	83.6
24	1	65	44.9	10	81	66.0
25	1	45	35.6	4	26	52.3
26	1	78	50.5	16	76	86.9
27	1	60	42.3	9	37	66.6
28	1	52	43.8	6	46	40.1
29	1	37	33.2	6	40	55.7
30	1	55	36.0	21	76	58.1
31	1	69	47.6	15	70	74.3
32	1	84	50.0	17	66	98.1
33	1	54	43.8	7	63	40.7
34	1	61	45.0	13	59	66.7
35	1	47	42.2	8	55	48.0
36	1	57	53.0	28	149	122.5
37	1	87	51.6	23	77	92.1
38	1	50	31.9	22	43	76.0
39	1	85	56.1	23	74	97.5
40	1	27	31.5	7	56	33.8
41	1	84	50.0	16	63	90.5
42	1	37	32.4	2	41	29.7
43	1	33	36.1	6	59	28.0
44	1	44	35.3	3	32	51.6
45	1	63	39.3	8	40	55.7
46	1	58	43.8	13	57	55.5

We decided to run a simple regression comparing each independent variable with the dependent variable instead of running a multiple regression (which is also a viable option) because we wanted to show, in depth, how each variable correlates to each other. The cirrhosis death rate was the dependent variable in these tests and the other variable in each case was the independent variable. Following each simple regression, we plotted the dependent versus the independent variable with a regression line.

Simple Regression Tests

Urban Population vs. Cirrhosis Death Rate

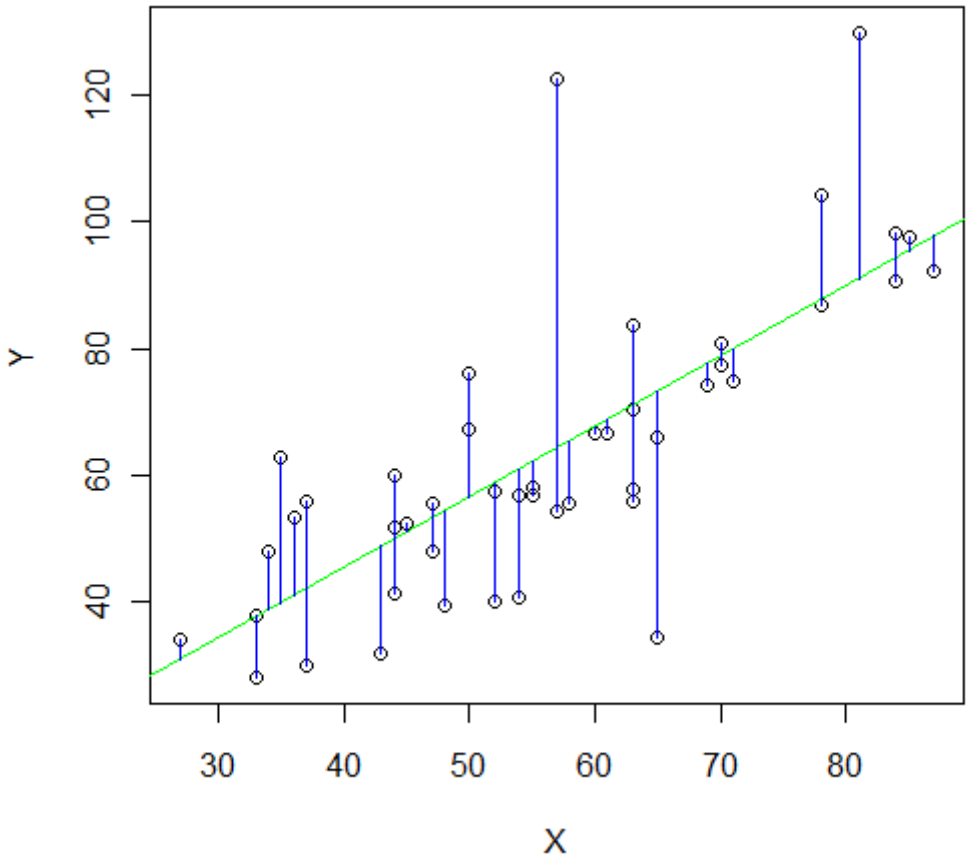
V3 vs. V7

```
> setwd("c:/Rdata")
> regression=read.table("regression.txt")
> attach(regression)
> regression
> X=V3
> Y=V7
```

```
> LM=lm(Y~X)
> LM
> Yhat=fitted(LM)
> e=residuals(LM)
> RESULTS=data.frame(X,Y,Yhat,e)
> plot(X,Y)
> abline(LM,col="green")
> segments(X,Yhat,X,Y,col="blue")
> anova(LM)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1  13883 13882.7  56.253 2.128e-09 ***
Residuals 44  10859   246.8
> summary(LM)
Call:
lm(formula = Y ~ X)
Residuals:
    Min      1Q  Median      3Q     Max
-38.941  -8.274  -1.429   3.486  58.182
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7407     8.6815   0.085    0.932
X             1.1154     0.1487   7.500 2.13e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 15.71 on 44 degrees of freedom
Multiple R-squared: 0.5611,     Adjusted R-squared: 0.5511
F-statistic: 56.25 on 1 and 44 DF,  p-value: 2.128e-09
> cor.test(X,Y,method="pearson",alternative="two.sided",conf.level=.95)

        Pearson's product-moment correlation
data:  X and Y
t = 7.5002, df = 44, p-value = 2.128e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5862612 0.8537254
sample estimates:
    cor 0.749074
```

Based on the correlation value, we determine that the correlation between these two variable is significant. The closer the value is to 1 (or -1), the higher the correlation and the closer the value to 0, the lower the correlation. In this case, the value is 0.75, which is a pretty high correlation. Thus we can conclude that there is a significant degree of correlation between Urban Population and Cirrhosis Death Rate.



Late Births vs. Cirrhosis Death rate

V4 vs. V7

```
> setwd("c:/Rdata")
> regression=read.table("regression.txt")
> attach(regression)
> regression
> Y=V7
> X=V4
> LM=lm(Y~X)
> Yhat=fitted(LM)
> e=residuals(LM)
> RESULTS=data.frame(X,Y,Yhat,e)
> RESULTS
> anova(LM)
```

Analysis of Variance Table

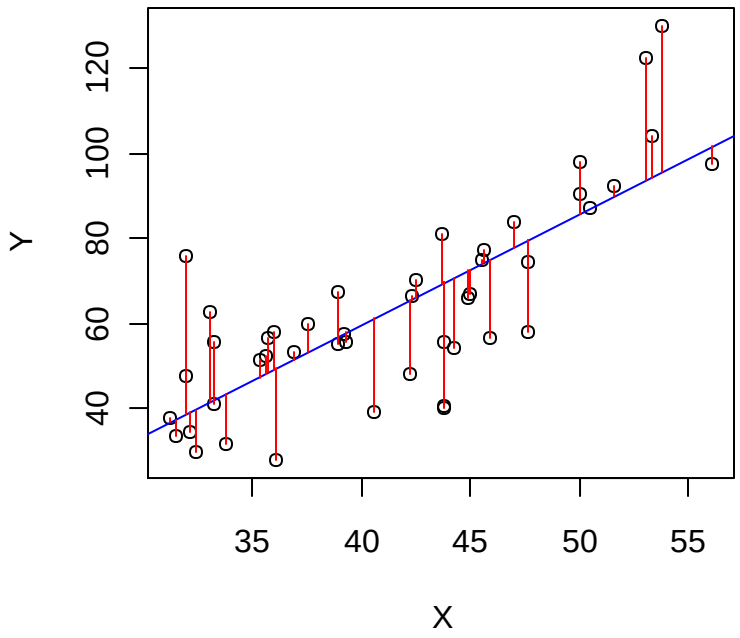
Response: Y

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 15158.0 15158.0  69.594 1.308e-10 ***
Residuals 44  9583.4   217.8
> summary(LM)
Call:
lm(formula = Y ~ X)
Residuals:
    Min      1Q  Median      3Q     Max
-29.448  -6.304   0.851   7.858  37.456
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.5682    13.1349  -3.393  0.00147 **
X             2.6054     0.3123   8.342 1.31e-10 ***

Residual standard error: 14.76 on 44 degrees of freedom
Multiple R-squared: 0.6127,     Adjusted R-squared: 0.6039
F-statistic: 69.59 on 1 and 44 DF,  p-value: 1.308e-10

> cor.test(X,Y, method="pearson",alternative="two.sided",conf.level=0.95)
        Pearson's product-moment correlation
data:  X and Y
t = 8.3423, df = 44, p-value = 1.308e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6372169 0.8743497
sample estimates:
      cor 0.7827244
```

Again, the correlation value is 0.78, which is pretty high and close to 1, meaning these variables are very closely correlated. Thus, we can conclude that there is a significant degree of correlation between Late Births vs. Cirrhosis Death rate.



Wine consumption per capita vs. Cirrhosis death rate

V5 vs. V7:

```
> setwd("c:/Rdata")
> regression=read.table("regression.txt")
> attach(regression)
> regression
> X=V4
> Y=V7
> LM=lm(Y~X)
> LM
> Yhat=fitted(LM)
> e=residuals(LM)
> RESULTS=data.frame(X,Y,Yhat,e)
> plot(X,Y)
> abline(LM,col="blue")
> segments(X,Yhat,X,Y,col="red")
> anova(LM)
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 17649.7 17649.7  109.51 1.616e-13 ***
Residuals 44  7091.7   161.2
```

> summary(LM)

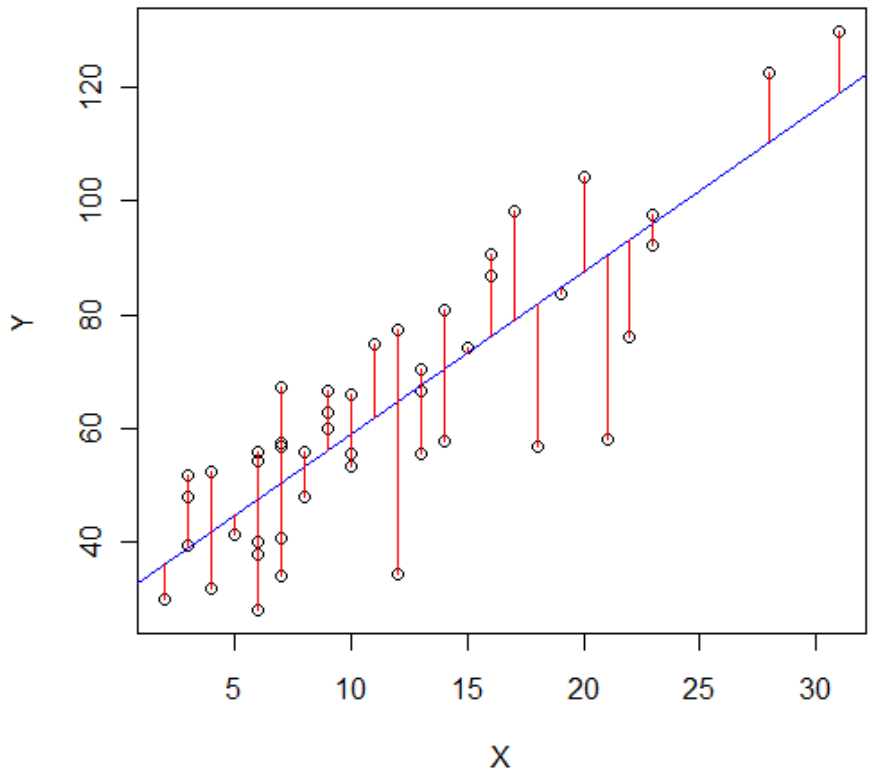
```
Call:
lm(formula = Y ~ X)
Residuals:
   Min     1Q  Median     3Q    Max
-32.331  -7.143   1.908  10.508  19.116
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.3347     3.6803   8.243 1.81e-10 ***
X             2.8617     0.2735  10.465 1.62e-13 ***

Residual standard error: 12.7 on 44 degrees of freedom
Multiple R-squared: 0.7134,     Adjusted R-squared: 0.7069
F-statistic: 109.5 on 1 and 44 DF,  p-value: 1.616e-13
```

> cor.test(X,Y,method="pearson",alternative="two.sided",conf.level=.95)

```
        Pearson's product-moment correlation
data:  X and Y
t = 10.4646, df = 44, p-value = 1.616e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7343734 0.9114353
sample estimates:
      cor
0.8446112
```

In this case, the correlation value is 0.84, which is very and thus, a very high correlation between the two variables.



Liquor Consumption per capita vs. Cirrhosis Death Rate

V6 vs. V7

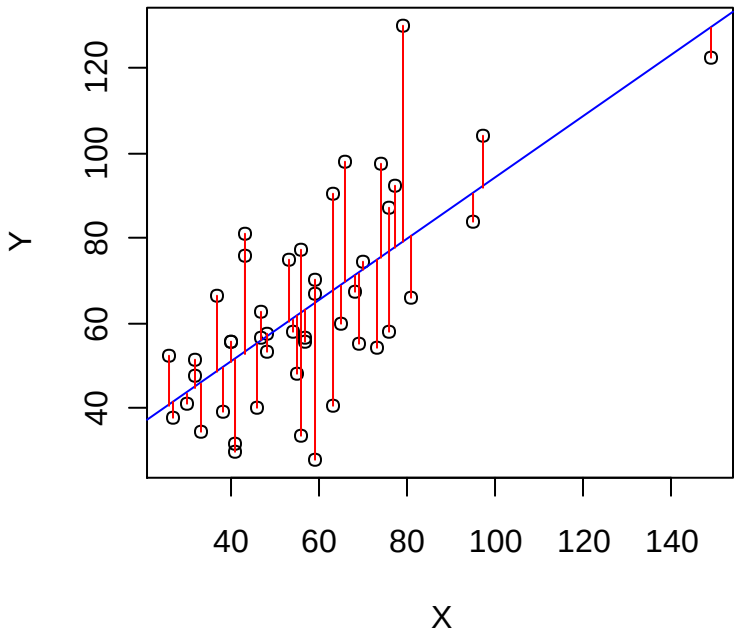
```
> setwd("c:/Rdata")
> regression=read.table("regression.txt")
> attach(regression)
> regression
> X=V6
> LM=lm(Y~X)
> Yhat=fitted(LM)
> e=residuals(LM)
> RESULTS=data.frame(X,Y,Yhat,e)
> anova(LM)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1  11507 11506.8  38.256 1.803e-07 ***
Residuals 44  13235   300.8
> summary(LM)
Call:
lm(formula = Y ~ X)
Residuals:
    Min      1Q  Median      3Q     Max
-36.577 -11.127  -0.821  11.179  50.878
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.9649     7.1847   3.057  0.00379 **
X             0.7222     0.1168   6.185 1.80e-07 ***

Residual standard error: 17.34 on 44 degrees of freedom
Multiple R-squared: 0.4651,     Adjusted R-squared: 0.4529
F-statistic: 38.26 on 1 and 44 DF,  p-value: 1.803e-07

> cor.test(X,Y, method="pearson",alternative="two.sided",conf.level=0.95)
        Pearson's product-moment correlation
data:  X and Y
t = 6.1851, df = 44, p-value = 1.803e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4883526 0.8115928
sample estimates:
      cor 0.6819694
```

```
> plot(X,Y)
> abline(LM,col="blue")
> segments(X,Yhat,X,Y,col="red")
```

The correlation value for these two variable is 0.68, which is still significant, and is also very well correlated. Thus, we can conclude that there is a significant correlation between Liquor Consumption per capita vs. Cirrhosis Death Rate.



When comparing the correlation values, it is surprising that the liquor consumption per capita was less correlated to cirrhosis death rates than the other variables. Perhaps there are other underlying factors that cause the other variables to be more correlated to cirrhosis death rate than the liquor consumption per capita.

Multiple Regression

It is also possible to use multiple regressions to analyze data such as this. In this case, there are multiple independent variables and one dependent variable. To use multiple regressions, one must assume the following:

1) Multiple Linear Regression depends on specifying in advance which variable is considered 'dependent' and which others 'independent'. This decision matters as changing roles for Y versus X's usually produces a different result.

2) $Y_1, Y_2, Y_3, \ldots, Y_n$ (dependent variable) is a random sample

3) k Vectors of *fixed* Independent Variables:

- $X_{1,1}, X_{1,2}, X_{1,3}, \ldots, X_{1,n}$ (independent variable) with each value of $X_{1,i}$ matched to Y_i
- $X_{2,1}, X_{2,2}, X_{2,3}, \ldots, X_{2,n}$ (independent variable) with each value of $X_{2,i}$ matched to Y_i

…

- $X_{k,1}, X_{k,2}, X_{k,3}, \ldots, X_{k,n}$ (independent variable) with each value of $X_{k,i}$ matched to Y_i

```
> Y=V7
> X1=V3
> X2=V4
> X3=V5
> X4=V6
> LM=lm(Y~X1+X2+X3+X4)
> LM
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)
Coefficients:
(Intercept)         X1         X2         X3         X4
  -13.96310    0.09829    1.14838    1.85786    0.04817
> summary(LM)
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)
Residuals:
    Min      1Q  Median      3Q     Max
-18.8723  -6.7803   0.1507   7.3252  16.4419
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.96310   11.40035  -1.225   0.2276
X1            0.09829    0.24407   0.403   0.6893
X2            1.14838    0.58300   1.970   0.0556 .
X3            1.85786    0.40096   4.634 3.61e-05 ***
X4            0.04817    0.13336   0.361   0.7198

Residual standard error: 10.61 on 41 degrees of freedom
Multiple R-squared: 0.8136,     Adjusted R-squared: 0.7954
F-statistic: 44.75 on 4 and 41 DF,  p-value: 1.951e-14
> anova(LM)
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
```

X1 1 13882.7 13882.7 123.4387 6.108e-14 ***
X2 1 1954.0 1954.0 17.3739 0.0001546 ***
X3 1 4278.9 4278.9 38.0463 2.494e-07 ***
X4 1 14.7 14.7 0.1305 0.7197931
Residuals 41 4611.1 112.5
One can compare the output from the different simple regression to the output for multiple regression.

Choosing an Optimal Model

Which model should we choose as the optimal model? We can use the AIC method to determine which model to use as the optimal model. Whatever model has the lowest AIC, we can choose that as the optimal model.

```
> Y=V7
> X1=V3
> X2=V4
> X3=V5
> X4=V6
> FM=lm(Y~X1+X2+X3+X4)
> n=length(Y)
> I=diag(n)
> J=matrix(nrow=n,ncol=n,1)
> X=model.matrix(FM)
> p=ncol(X)
> H=X%*%solve(t(X)%*%X)%*%t(X)
> SSR=t(Y)%*%(H-((1/n)*J))%*%Y
> SSE=t(Y)%*%(I-H)%*%Y
> SSTO=t(Y)%*%(I-(1/n)*J)%*%Y
> SSTO
         [,1]
[1,] 24741.35
> rbind(SSR,SSE,SSTO)
          [,1]
[1,] 20130.239
[2,]  4611.109
[3,] 24741.348
> AIC=n*log(SSE[1])-n*log(n)+2*p
> AIC
[1] 221.9488
> extractAIC(FM)
[1]   5.0000 221.9488…The same as our starting AIC below.
> drop1(FM)
Single term deletions
```

Model:
Y ~ X1 + X2 + X3 + X4
Df Sum of Sq RSS AIC
<none> 4611.1 221.95
X1 1 18.24 4629.3 220.13
X2 1 436.38 5047.5 224.11
X3 1 2414.63 7025.7 239.32
X4 1 14.67 4625.8 220.09
> require(car)
> Anova(FM)
Model:
Y ~ X1 + X2 + X3 + X4
Df Sum of Sq RSS AIC
<none> 4611.1 221.95
X1 1 18.24 4629.3 220.13
X2 1 436.38 5047.5 224.11
X3 1 2414.63 7025.7 239.32
X4 1 14.67 4625.8 220.09…Lowest AIC, so this is the model we choose
Anova Table (Type II tests)
Response: Y
Sum Sq Df F value Pr(>F)
X1 18.2 1 0.1622 0.68926
X2 436.4 1 3.8801 0.05564 .
X3 2414.6 1 21.4698 3.615e-05 ***
X4 14.7 1 0.1305 0.71979
Residuals 4611.1 41

Automated "backwards" Stepwise Regression:
> step(FM,direction="backward")
Start: AIC=221.95
Y ~ X1 + X2 + X3 + X4
Df Sum of Sq RSS AIC
- X4 1 14.67 4625.8 220.09…Lowest AIC
- X1 1 18.24 4629.3 220.13
<none> 4611.1 221.95
- X2 1 436.38 5047.5 224.11
- X3 1 2414.63 7025.7 239.32
Step: AIC=220.09
Y ~ X1 + X2 + X3
Df Sum of Sq RSS AIC
- X1 1 6.3 4632.1 218.16…Lowest AIC
<none> 4625.8 220.09

- X2 1 1046.8 5672.6 227.48
- X3 1 4278.9 8904.7 248.22

Step: AIC=218.16

Y ~ X2 + X3

	Df	Sum of Sq	RSS	AIC
<none>			4632.1	218.16…Lowest AIC
- X2	1	2459.6	7091.7	235.75
- X3	1	4951.3	9583.4	249.60

Call:

lm(formula = Y ~ X2 + X3)

Coefficients:

(Intercept)	X2	X3
-16.001	1.366	1.972

Conclusion:

The final model and most optimal model, in this case is Y~X2+X3. X2 and X3 are V4 and V5 respectively, or late births and wine consumption per capita respectively.

There is a variety of options and functions one may carry out in terms of linear regression models and data sets. The shown functions are just a mere sample of one might do.

Maheen Kibria and Ryan Dean