Regression
Lesson 1b

Kevin Zollicoffer

10/14/2013

Introduction

Regression assignment 1b using R.

The complete source for this assigment is available on Github:

`https://github.com/zollie/PASS-Regression-Assignment1b`

Problem 2.3

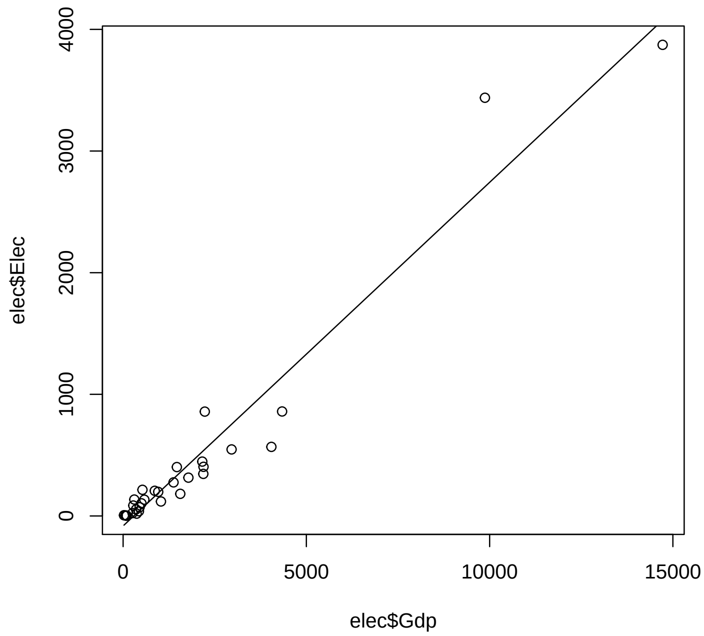
```
> elec <- read.csv("~/R/PASS/Regression/Assignment1b/electricity.csv")
```

a

GDP should be the predictor vairable with Electricity should be the response vairable. b_1 would be positive under the claim that electricity consumption increases in response to increases in GDP.

b

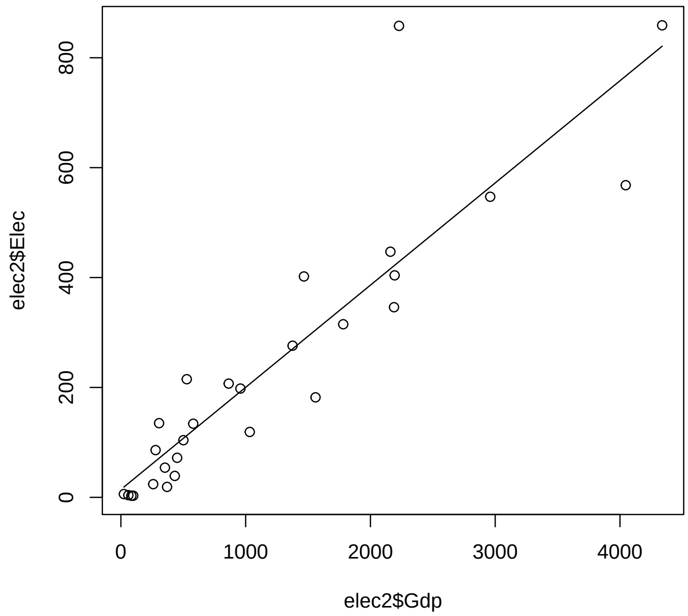
```
> plot(elec$Gdp, elec$Elec)
> model <- lm(elec$Elec ~ elec$Gdp)
> lines(sort(elec$Gdp), fitted(model)[order(elec$Gdp)])
```



There is a clearly a positive relationship between GDP and electrictiy consumption in a country. There are 2 outliers skewing the results to the right, and perhaps overestimating the slope of the resultant regression line. This increases the standard error of the model. It also bunches the non-outlier data points toward the lower left of the model inhibiting interpretation of the model.

c

```
> max2 <- order(elec$Gdp,decreasing=T)[1:2]
> elec2 <- elec[-max2,]
> plot(elec2$Gdp, elec2$Elec)
> model2 <- lm(elec2$Elec ~ elec2$Gdp)
> lines(sort(elec2$Gdp), fitted(model2)[order(elec2$Gdp)])
```



With the 2 outliers removed, the standard error is apprently decreased and the obersvations appear more tightly correlated about the regression line (taking into account the scale of the X/Y axes).

d

```
> options(scipen=999) # disable scientific notation
> summary(model2)

Call:
lm(formula = elec2$Elec ~ elec2$Gdp)

Residuals:
    Min      1Q  Median      3Q     Max
-198.69  -32.61  -18.01   22.78  429.21

Coefficients:
            Estimate Std. Error t value        Pr(>|t|)
(Intercept) 14.27579   29.11331    0.49           0.628
elec2$Gdp    0.18596    0.01752   10.62 0.0000000000601 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

```
Residual standard error: 107 on 26 degrees of freedom
Multiple R-squared:  0.8126,        Adjusted R-squared:  0.8054
F-statistic: 112.7 on 1 and 26 DF,  p-value: 0.00000000006009
```

Hypotheis Test

$H_0 = b_1 = 0$
$H_a = b_1 > 0$
$b_1\ t - stat = 10.62$
$b_1\ p - value = .00000000601$

t-distribution upper tail signifigance level for 5% (1-.05) confidence and 26 degrees of freedom = 1.706

Hypothesis Test Result

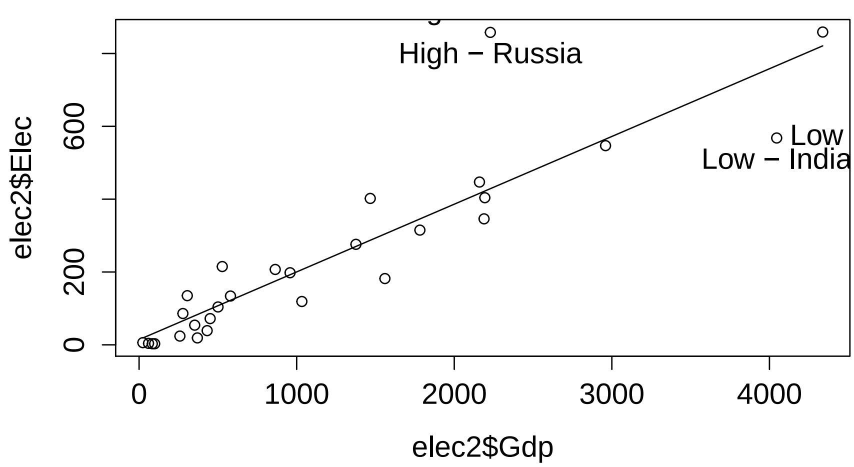
$b_1\ t - stat = 10.62 > 1.706 \therefore reject\ H_0$

alternatively

$b_1\ p - value = .00000000601 < .005 \therefore reject\ H_0$

An elec2$Gdp slope of zero seems inplausible. The sample data favor a positive slope at 5% confidence level.

e



2.5

```
> cars2 <- read.csv("~/R/PASS/Regression/Assignment1b/cars2.csv")
```

a

```
> cars2[["Cgphm"]] <- 100/cars2$Cmpg
> mean(cars2$Cgphm)
```

```
[1] 4.613156
```

b

Regression using Eng as the predictor

```
> model_eng <- lm(Cgphm ~ Eng, data=cars2)
> summary(model_eng)
```

```
Call:
lm(formula = Cgphm ~ Eng, data = cars2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.61401 -0.22593 -0.04419  0.15520  1.32962 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)    
(Intercept)   2.5894     0.1026   25.24 <0.0000000000000002 ***
Eng           0.8183     0.0397   20.61 <0.0000000000000002 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.3351 on 125 degrees of freedom
Multiple R-squared:  0.7726,	Adjusted R-squared:  0.7708 
F-statistic: 424.8 on 1 and 125 DF,  p-value: < 0.00000000000000022
```

Regression using Vol as the predictor

```
> model_vol <- lm(Cgphm ~ Vol, data=cars2)
> summary(model_vol)
```

```
Call:
lm(formula = Cgphm ~ Vol, data = cars2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2039 -0.4521 -0.1067  0.3734  2.3482 
```

```
Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)   1.8760     0.5337   3.515    0.000613 ***
Vol           2.5010     0.4849   5.157 0.000000953 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.6382 on 125 degrees of freedom
Multiple R-squared:  0.1755,        Adjusted R-squared:  0.1689
F-statistic:  26.6 on 1 and 125 DF,  p-value: 0.0000009527
```

Residual standard error evaluation

For the linear regression model using Eng as the predictor $s = .3351$

For the linear regression model using Vol as the predictor $s = .6382$

$.3351 < .6382 \therefore$ when considering s, the model using predictor Eng is preferable

Coefficient of determination - $\mathbf{R^2}$

For the linear regression model using Eng as the predictor $R^2 = .7726$

For the linear regression model using Vol as the predictor $R^2 = .1755$

$.7726 > .1755 \therefore$ when considering R^2, the model using predictor Eng is preferable. Moreover, .1755 is significantly < 1, therefore the model using Vol as the predictor is highly questionable.

The p-value of b_1

For the linear regression model using Eng as the predictor the p-value of Eng =0.0000000000000002

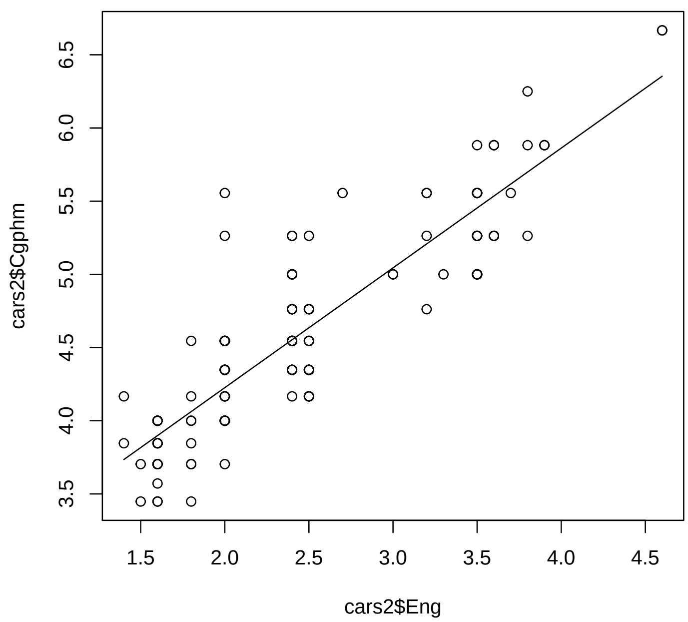
For the linear regression model using Vol as the predictor the p-value of Vol=0.0000009527

$.0000000000000002 < .0000009527 \therefore$ when considering the statistical significance of b_1, the model using predictor Eng is preferable.

Visual Interpretation

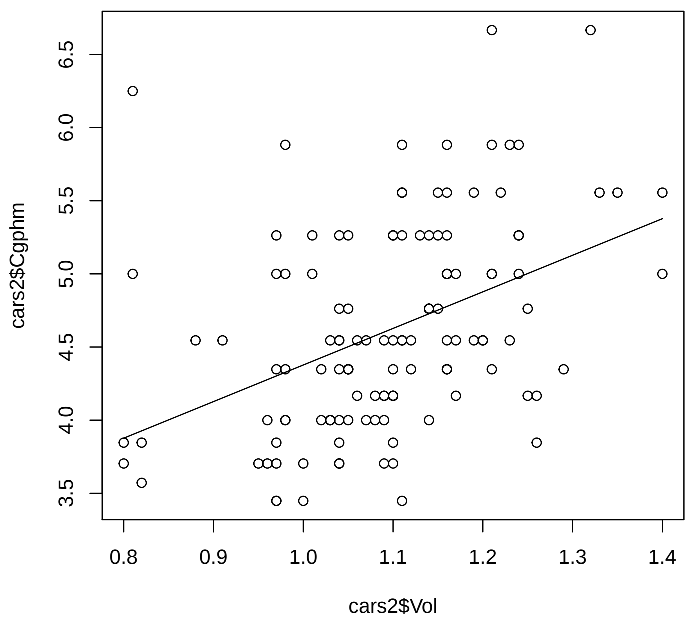
Eng

```
> plot(cars2$Eng, cars2$Cgphm)
> lines(sort(cars2$Eng), fitted(model_eng)[order (cars2$Eng)])
```



Vol

```
> plot(cars2$Vol, cars2$Cgphm)
> lines(sort(cars2$Vol), fitted(model_vol)[order (cars2$Vol)])
```



Visually the plot of Eng is more tightly correlated around the regression line and the slope of the regression line exhibits more lift. Using Eng as the predictor is preferable over using Vol as the predicotr for this linear regression excercise.

C

Using Eng as the predictor for the cars2 data was reccomended. As shown above $s = .3351$ for this model.

It can be shown that approximately 95% of the observed Y-values lie within approximately $\pm$ 2s, therefore it can be said that with 95% confidence our future predictions of Y using this linear regression model will fall within $\pm 2s$. That is, we have a 95% confidence interval of $((X).8183 \pm .6702)$ given an observation of Eng=X.