Homework3

Tung Nguyen

2/17/2020

```
knitr::opts_chunk$set(echo=FALSE, comment=NA)
```

Problem 2.1

a. Find the least square line for the data.

```
Call:
lm(formula = Int ~ Gdp, data = data2.1)

Residuals:
    Min      1Q  Median      3Q     Max
-61.742 -11.914  -3.276   9.417  63.644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.36278    1.71827   7.195 1.09e-11 ***
Gdp          1.36093    0.07975  17.065  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.14 on 210 degrees of freedom
Multiple R-squared:  0.581, Adjusted R-squared:  0.579
F-statistic: 291.2 on 1 and 210 DF,  p-value: < 2.2e-16
```

The least square line for the data, based on the output, is:

$\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 12.363 + 1.36\text{X}$

b. Interpret the estimates of the slop and the intercept in the context of the problem

Since the intercept is 12.363, when GDP equals 0, about 12.363% of the population of that country uses the Internet.

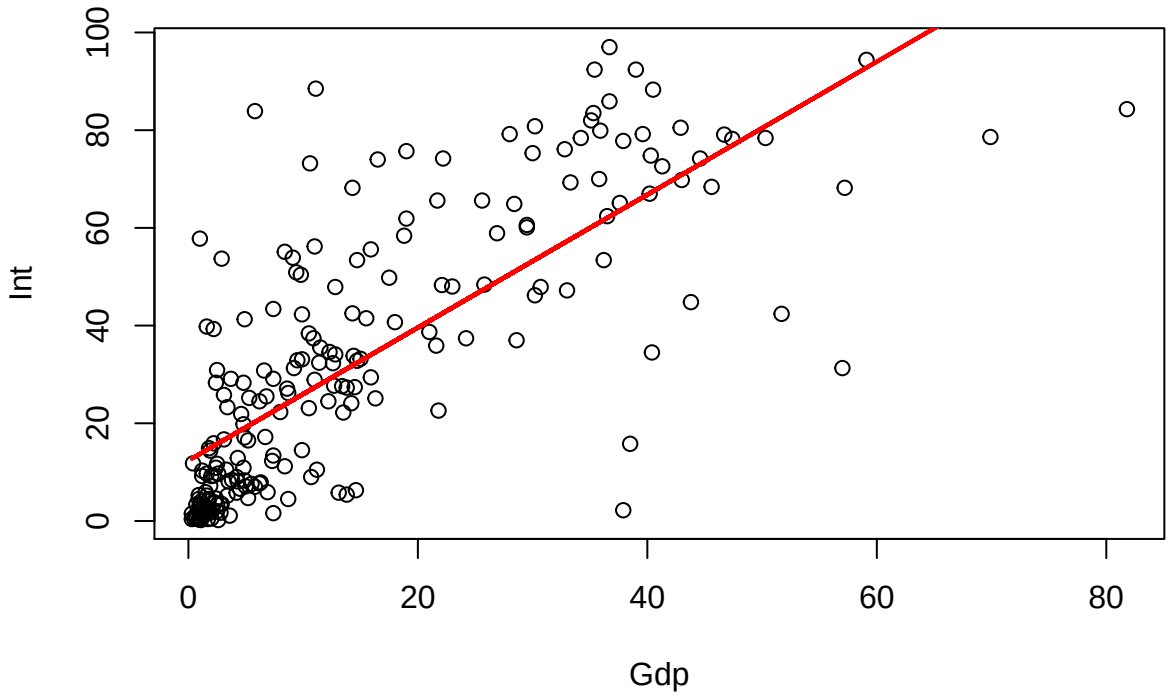
Since GDP = 1.36, when GDP increases by 1000 $USD, the percentage of the corresponding population increases by 1.36%.

c. Predict the percentage of the Internet users if GDP per capita is US$20,000.

```
       1
39.58131
```

Accordingly, when Gdp = 20, Int ~ 39.581.

d. Draw a scatterplot with Int on the vertical axis and Gdp on the horizontal axis. Add the least squares line to the plot



e. Based on the scaterplot, do you think it is appropriate to use this simple linear regression model on this problem, or is the model potentially misleading.

The model is potentially misleading. As the plot suggests, when Gdp =0, the internet usage is around 12% of the population. This result,indeed, does not make sense.

Problem 2.5

Exploring data

```
                  Name Cmpg Eng  Vol
1    Acura TL 2wd 3.5L   18 3.5 1.11
2           Audi A3 2L   22 2.0 1.09
3           Audi A4 2L   22 2.0 1.03
4 Audi A5 Cabriolet 2L   22 2.0 0.91
5         Audi A6 3.2L   21 3.2 1.14
6  Buick Lacrosse 2.4L   19 2.4 1.16
```

a. Transform city miles per gallon into "city gallons per hundred miles". In other words, create another variable called Cgphm = 100/Cmpg.

Here are the first five rows of the transformed data

```
                  Name Cmpg Eng  Vol    Cgphm
1    Acura TL 2wd 3.5L   18 3.5 1.11 5.555556
2           Audi A3 2L   22 2.0 1.09 4.545455
3           Audi A4 2L   22 2.0 1.03 4.545455
4 Audi A5 Cabriolet 2L   22 2.0 0.91 4.545455
5         Audi A6 3.2L   21 3.2 1.14 4.761905
6  Buick Lacrosse 2.4L   19 2.4 1.16 5.263158
```

The mean of the newly transformed variable is

```
[1] 4.613156
```

b. Predicting Cgphm using Eng or Vol

Predicting Cgphm using Eng

```
Call:
lm(formula = Cgphm ~ Eng, data = data2.5)

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   <2e-16 ***
Eng           0.8183     0.0397   20.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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: < 2.2e-16
```

Predicting Cgphm using Vol

```
Call:
lm(formula = Cgphm ~ Vol, data = data2.5)

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 9.53e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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: 9.527e-07
```

I would recommend using the model with predictor Eng. Althought both model output p-value $<$ 0.001, the R-squared for the Eng model is 0.7726 while that of the Vol model is only 0.1755. The result means the the Eng model captures more variability of the response than the Vol model.

C. Report the regression standard error (s) for the model you recommended in part (b). Say something about the predictive ability of your model.

```
Analysis of Variance Table

Response: Cgphm
           Df Sum Sq Mean Sq F value    Pr(>F)
Eng         1 47.704  47.704  424.76 < 2.2e-16 ***
Residuals 125 14.039   0.112
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The regression standard deviation s is: 6.906823 0.3351249
```

The five number summary for Cgphm is:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.448   4.000   4.545   4.613   5.000   6.667
```

Since s = 6.906 $>$ Max = 6.667, the model errors seem to be pretty big. Hence, its predictive ability is not reliable.

2.7

a. 95% confidence interval for the regression slope.

```
                2.5 %    97.5 %
(Intercept) 2.3863697 2.7924758
Eng         0.7396795 0.8968317
```

The 95% confidence interval for the slop β_1 is