HW - Linear Regression

Shyam BV

November 13, 2016

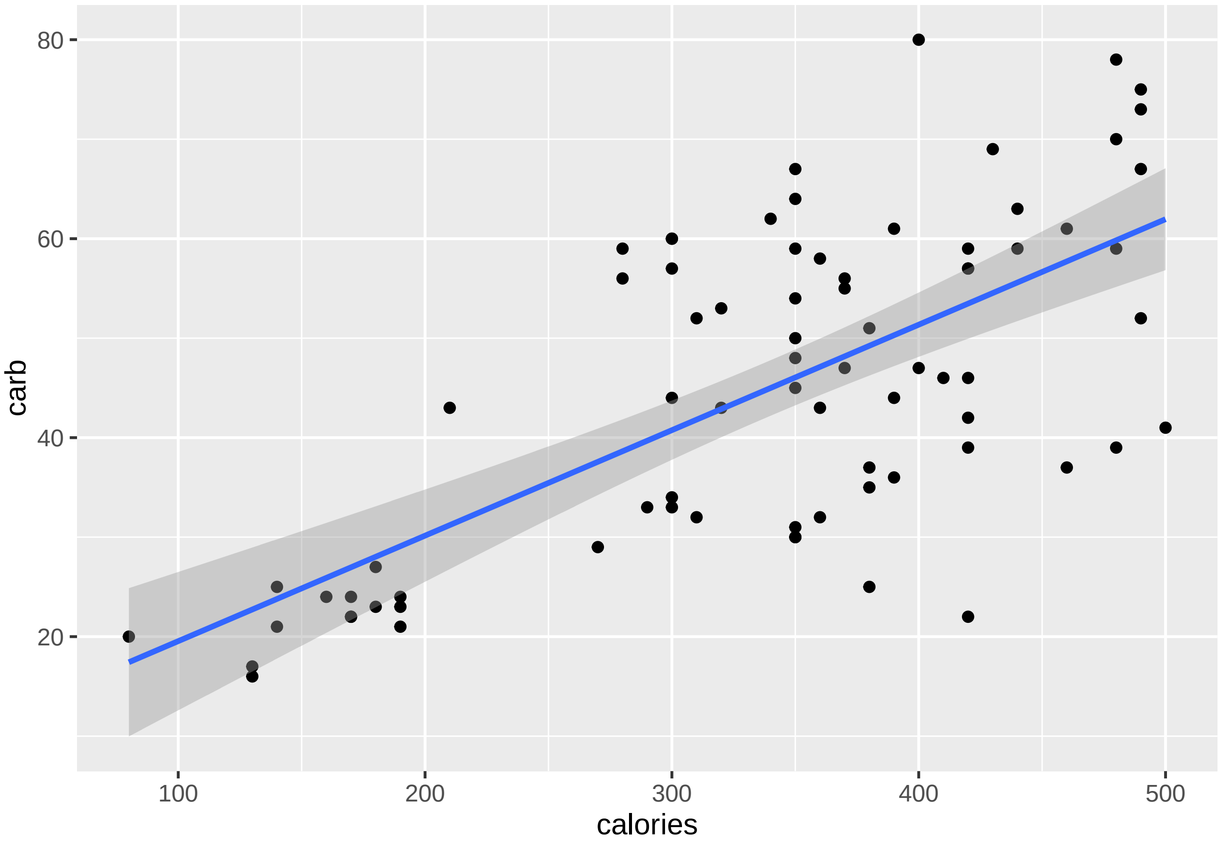
7.24 Nutrition at Starbucks, Part I.

The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.

```
starbucks <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from
```

(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

```
ggplot(starbucks,aes(x=calories,y=carb)) + geom_point() + geom_smooth(method="lm")
```



```
cor(starbucks$calories,starbucks$carb)
```

```
## [1] 0.674999
```

The relationship between the two variables is positive and linear. Although there are some outliers, relationship is linear. On a initial look from residual plot, it seems there is a formation of two clouds in the distribution.

The residual plot from the histogram is nearly normal.

(b) In this scenario, what are the explanatory and response variables?

In this scenario, the explanatory variable is `Calories` and the response variable is `Carb`.

(c) Why might we want to fit a regression line to these data?

Because we want to see the historical trend and predict the future outcomes. If the the calories increases, what might be the point estimate of `Carb` level.

(d) Do these data meet the conditions required for fitting a least squares line?

Below are conditions for fitting a least squares line.

1. Linearity: The data should show linear trend. The plots show that the data is linear between two variables.
2. Nearly normal residuals: From the residual plot, we can see the points are surrounded by the line. Although we see many outliers and influential points, on overall look, it looks nearly normal.
3. Constant Variability: From the plot, the variablity is more in the lower section and high in the upper section of plot. But there is no drastic changes in the variability. So it can be accepted.
4. Independent Observations: These observations are independent of each other data points.

7.26 Body measurements, Part III.

Exercise 7.15 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.

```r
bdim <- read.delim("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from%20

mean(bdim$sho.gi)
```

```
## [1] 108.1951
```

```r
sd(bdim$sho.gi)
```

```
## [1] 10.37483
```

```r
mean(bdim$hgt)
```

```
## [1] 171.1438
```

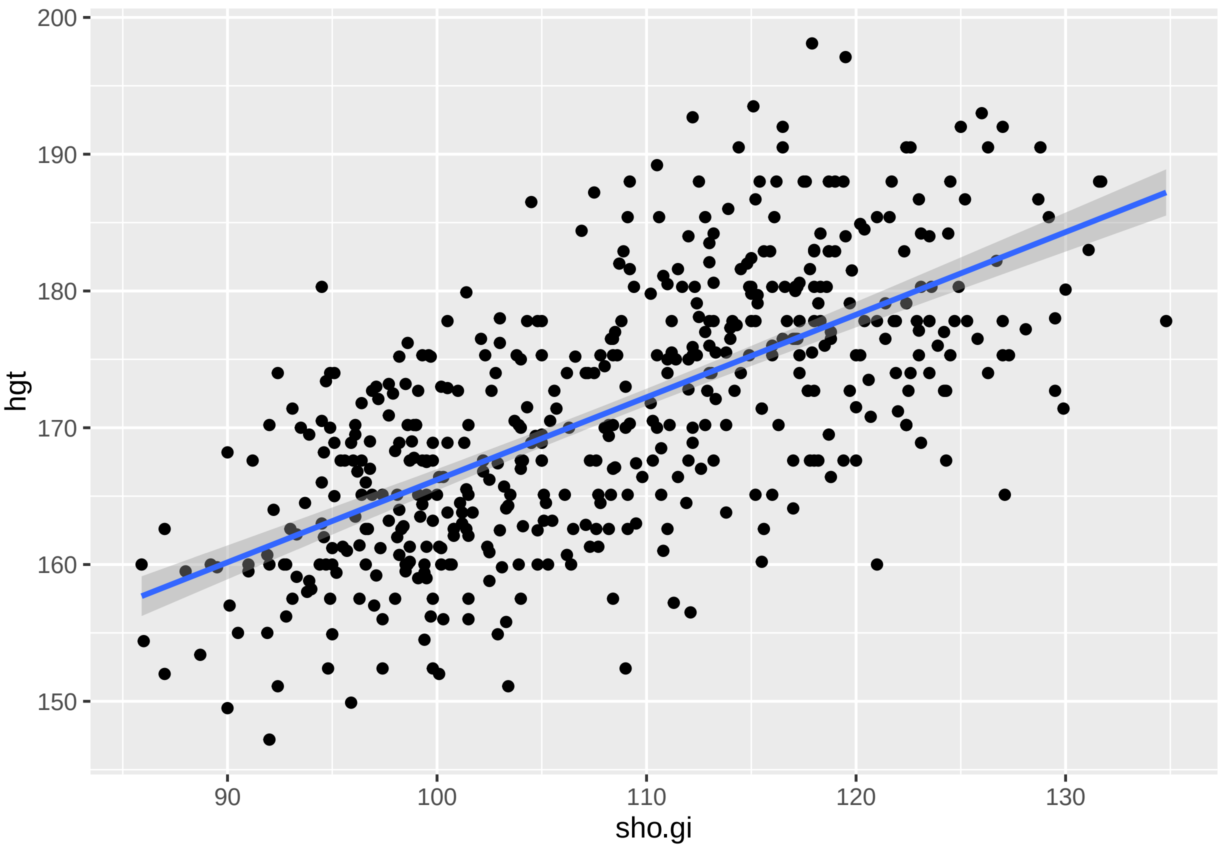
```r
sd(bdim$hgt)
```

```
## [1] 9.407205
```

```r
cor(bdim$hgt,bdim$sho.gi)
```

```
## [1] 0.6657353
```

```r
ggplot(bdim,aes(x=sho.gi,y=hgt)) + geom_point() + geom_smooth(method="lm")
```



(a) Write the equation of the regression line for predicting height.

```
p1 <- summary(lm(bdim$hgt~bdim$sho.gi))

#Equation for of regression line predicting height

paste(p1$coefficients[[1]],' + ',p1$coefficients[[2]],'*','Sholders Grim')
```

```
## [1] "105.832461820568  +  0.603644193262076 * Sholders Grim"
```

```
#Explanation

R = (sd(bdim$hgt)/sd(bdim$sho.gi))*cor(bdim$hgt,bdim$sho.gi)
```

hgt = $\beta_0 + \beta_1$ * sho.gi

(b) Interpret the slope and the intercept in this context.

```
Intercept <- p1$coefficients[[1]]
slope <- p1$coefficients[[4]]
```

Intercept means the minimum height when sholder grim is 0. The height will be atleast 105.83 cm if the sholder grim is zero. Slope means for every additional 1 cm increase in sholder grim, there will be 0.603 increase in height.

(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

```
Rsq <- p1$r.squared
```

```
R*R
```

```
## [1] 0.3643863
```

R^2 value is 44.32%. It means about 44.32% in the data variation by using information about sholder grim for predicting height using a linear model.

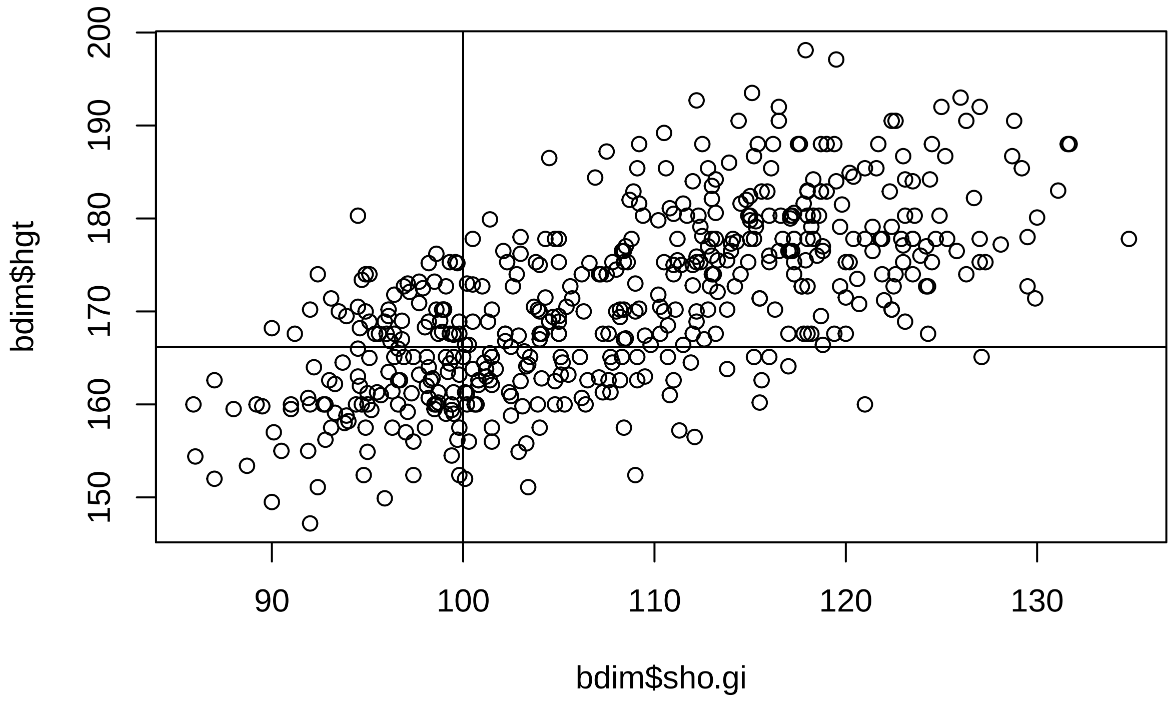
(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

```
#substitute sholder girth in the linear equation

sho.gi = 100
p1$coefficients[[1]]+p1$coefficients[[2]]*sho.gi
```

```
## [1] 166.1969
```

```
#plot
plot(bdim$sho.gi,bdim$hgt)
abline(v=sho.gi)
abline(h=p1$coefficients[[1]]+p1$coefficients[[2]]*sho.gi)
```



(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

```
 sholder.grim =(160 -p1$coefficients[[1]])/p1$coefficients[[2]]

residual = (160-100)
```

The residual for that particular student is 60. It means that the difference between the observed and the predicted value based on model of best fit. The model underestimates for that particular student.

(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

```
sho.gi = 56

p1$coefficients[[1]]+p1$coefficients[[2]]*sho.gi
```

```
## [1] 139.6365
```

Above equation shows the height of the one year old is 139.63 cm. Practically this is an extrapolation value. Because the one year old cannot be 4.6 feet. Our model does not calculate the value for 1 year old childeren.

7.30 Cats, Part I.

The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515
body wt	4.034	0.250	16.119

s = 1.452 R^2 = 64.66% R_a^2dj = 64.41%

(a) Write out the linear model.

heart_wgt = $\beta_0 + \beta_1$ * body_wgt

b0 = -0.357 b1= 4.034

heart_wgt = b0 +b1*body_wgt

(b) Interpret the intercept.

Intercept in this linear model means, if the weight of the cat is 0, then the heart weight is -0.357.

Practically both cannot happen. Cat weight cannot be 0 and the heart weight cannot be negative.

(c) Interpret the slope.

Slope in this linear model means, if there is increase in 1 kg of body weight then there is a increase of 4.034 gm of heart weight

(d) Interpret R2.

R^2 in this linear model means, 64.66% the data variation by using body weight for predicting heart weight

(e) Calculate the correlation coefficient.

```
sqrt(.6466)
```

```
## [1] 0.8041144
```

R can be calculated by taking sqrt of R-squared.

7.40 Rate my professor.

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

```
profbeauty <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20fro
```

(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

```
b1 <- (3.9983 - 4.010) / -0.0883
b1
```

```
## [1] 0.1325028
```

(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

Given the p-value for the slope is (0.0000), this provides strong evidence that the slope is not 0 (Reject null hypothesis).

(c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.

Linearity: There is a weak trend in the scatterplot. But still the linearity condition can be accepted.

Nearly normal residuals: The histogram plot shows that the residuals are nearly normal.

Constant variability: The scatterplot of the residuals does appear to have constant variability.

Independent observations: Assuming independence is there in the dataset. Each professor is independent of another.