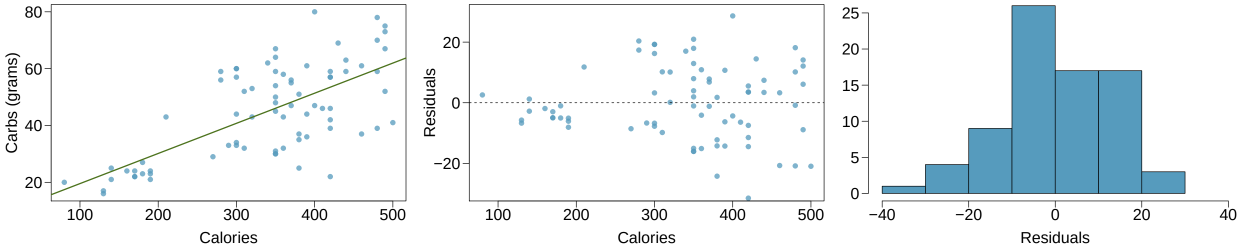
Chapter 8 - Introduction to Linear Regression

Md Forhad Akbar

Nutrition at Starbucks, Part I. (8.22, p. 326) 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.



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

Answer: The relationship between number of calories and amount of carbohydrates (in grams) is positive linear as carbohydrates (in grams) increases wiht number of calories. However, the relationship is not too strong.

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

Answer: Explanatory Variable: Calories along x-axis. Response Variable: Carbs(grams) along y-axis.

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

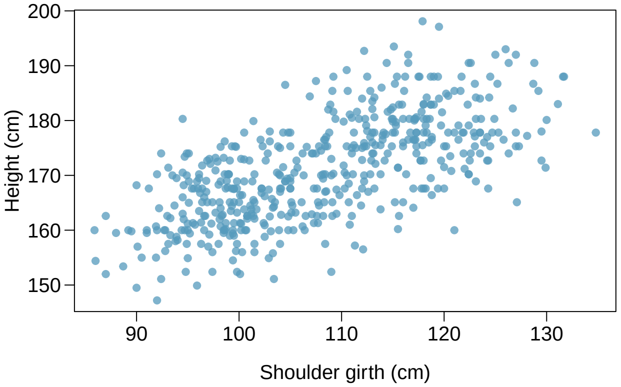
Answer:

If we can fit a regression line, then we can make predictions. Given a value of Explanatory variable, we can predict the value of Response variable. We can predict amount of carbs based on calories if we able to fit a regression line to these data.

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

Answer: The data fit a linear plot, residuals appear nearly normal. we cannot achieve constant variability. Since, these scatter plots are about Starbucks food menu items, we can't be 100% sure that they are independent.

Body measurements, Part I. (8.13, p. 316) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.19 The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.



(a) Describe the relationship between shoulder girth and height.

Answer: The relationship between shoulder girth and height is positive linear as height increases wiht shoulder girth. However, the relationship is not too strong. However, the variability increased, with higher values of Hip girth.

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

Answer: The relationship would not change but every point will move in positive direction of x-axis by a factor of 2.54 as 1 inch is equalto 2.54 cms. This would reduce the gradient of the line of linear regression.

Body measurements, Part III. (8.24, p. 326) Exercise above 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.

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

Answer:

```
# The general equation for the regression line: y = B0 + B1*x where B0 and B1 represent two model param
# X is the explanatory or predictor variable and y is the response.
# B1 is the slope, which also equals: (sample y standard deviation) / (sample x standard deviation) * R
# R is the correlation between the two variables.
# R ranges from -1 to 1, with -1 being completely negative correlation and +1 being a completely positi

# Shoulder girth is the explanatory variable
shoulder.mean <- 107.20 # in cm
shoulder.SD <- 10.37 # in cm

# Height is the response
height.mean <- 171.14 # in cm
height.SD <- 9.41 # in cm

# R for correlation
R <- 0.67

# Calculate the slope (or otherwise known as B1)
B1 <- R * (height.SD/shoulder.SD)

# Now to calculate for B0, we will use the values (x,y) = (107.20, 171.14). They are also the mean valu
# Now to rearrange the equation to solve for B0. B0 = y - B1*x
B0 <- 171.14 - B1 * 107.20

B1;B0
```

```
## [1] 0.6079749
```

```
## [1] 105.9651
```

Therefore, the final regression line is: height = 105.75 + 0.61 x shoulder girth

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

Answer:

If shoulder girth increases by 1 cm, the height increases by 0.61 cm.

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

Answer:

```r
R.squared <- R^2
paste("R squared: ", round(R.squared,3))
```

```
## [1] "R squared:  0.449"
```

This means that 44.9% of the variation found in this data is explained by the linear model i.e. explained by the shoulder girth width.

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

Answer:

```r
student.shoulder <- 100 # in cm
student.height <- B0 + B1 * student.shoulder
paste("According to the model, the estimated height of a student with a shoulder girth of 100 cm is: ",
```

```
## [1] "According to the model, the estimated height of a student with a shoulder girth of 100 cm is:  
```

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

Answer:

```r
# The residual = height(actual) - height(expected)
residual <- 160 - student.height
paste("The residual is: ", round(residual,3))
```

```
## [1] "The residual is:  -6.763"
```

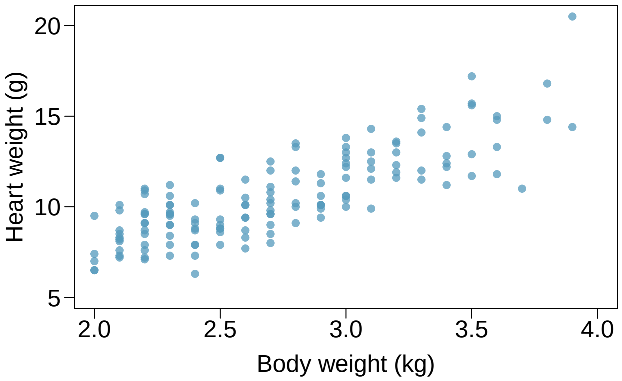
The residual is negative. So, the linear model overpredicted the height.

(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?

Answer: The original data set had a response variable values between ~80 and 140 cm. A measure of 56 is outside the sample. So, this model would not be appropriate for calculating the height of this child.

Cats, Part I. (8.26, p. 327) 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	0.607
body wt	4.034	0.250	16.119	0.000
$s = 1.452$	$R^2 = 64.66\%$		$R^2_{adj} = 64.41\%$	



(a) Write out the linear model.

Answer: for the linear model equation y = B0 + B1 * x, where B0 is the height of the intercept and B1 is the slope. Therefore, by looking at the table, the linear regression model here is:

Heart Weight (g) = -0.357 + 4.034 x Body Weight (kg)

(b) Interpret the intercept.

Answer: Expected heart weight in cats with 0 kg body weight is -0.357 g. This is not a meaningful value, it just serves to adjust the height of the regression line.

(c) Interpret the slope.

Answer: For each additional kg increase in body weight, we expect an additional 4.034 grams in the heart weight.

(d) Interpret R^2.

Answer: Body weight (in kg) explains 64.66% of the variability in the heart weight (in g) of the cat.

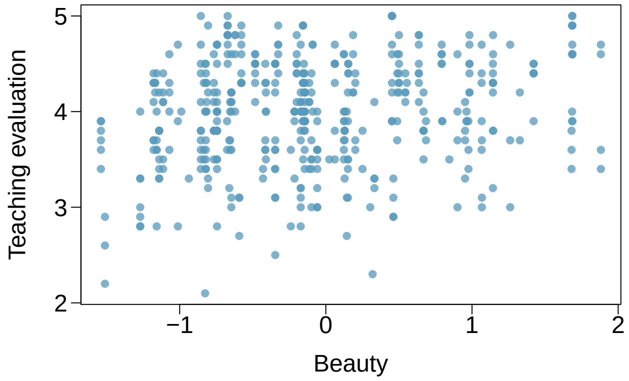
(e) Calculate the correlation coefficient. **Answer:**

```
paste("Correlation coefficient: ", round(sqrt(.6466),3))
```

```
## [1] "Correlation coefficient:  0.804"
```

Rate my professor. (8.44, p. 340) 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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	□	0.0322	4.13	0.0000



(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.

Answer: Slope can be calculated from the linear model equation y = B0 + B1 x Here, they provide enough information so that we can solve B1 (the slope) We can substitute x with the average standardized beauty score (explanatory variable) = -0.0883 and we can substitute y with the average teaching evaluation score of 3.9983 (response). We know that the averages lie on the linear regression model.We also know the Y intercept. It is on the table. Y intercept = 4.010 Therefore, the equation is now: 3.9983 = 4.010 + B1 * (-0.0883) Now solve for B1 to find the slope.

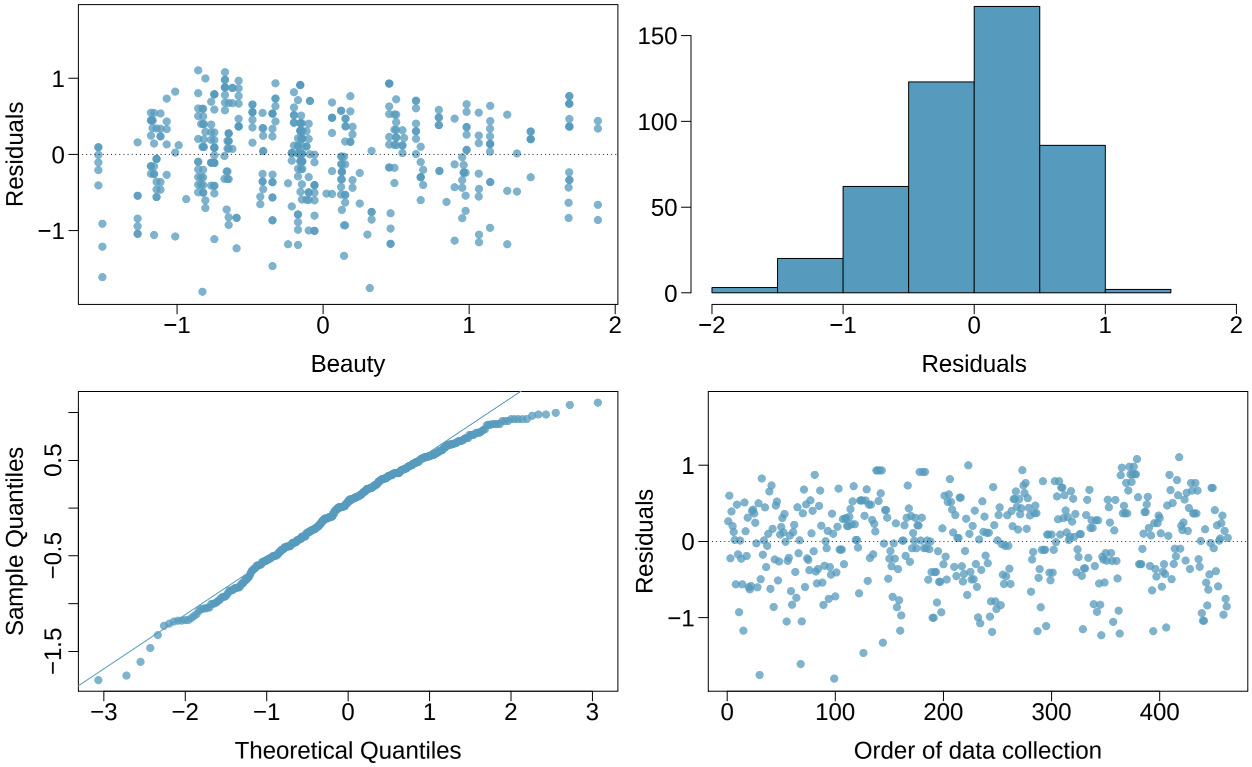
```
prof.B1 <- (3.9983 - 4.010)/(-0.0883)
paste("B1 or the slope: ", round(prof.B1, 3))
```

```
## [1] "B1 or the slope:  0.133"
```

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

Answer: While on visual inspection, if anything, it looks like there may be a near zero slope level. However, the data from the summary chart does suggest that there is some correlation between beauty and teaching evaluation. The P values in the summary chart demonstrates p's that are ~0, which suggests that we should reject the null hypothesis (null hypothesis being that there is NO correlation).

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



Answer: There are four criteria that needs to be satisfied prior to using the least squares linear regresion method.

Linearity : There does appear to be some linearity if we take a horizontal slope and place a horizontal line roughly right through the data points

Nearly normal residuals : The residuals do appear to be nearly normal, by looking at its distribution on the histogram.

Constant variability : By looking at the residual plots, there appears to be constant variability throughout

Independent observations : Looking at the order of data collection, there doesn't seem to an obvious pattern that would suggest that these observations were dependent on each other. It is likely that these were all independent observations.

Hence, we can perform the linear regression.