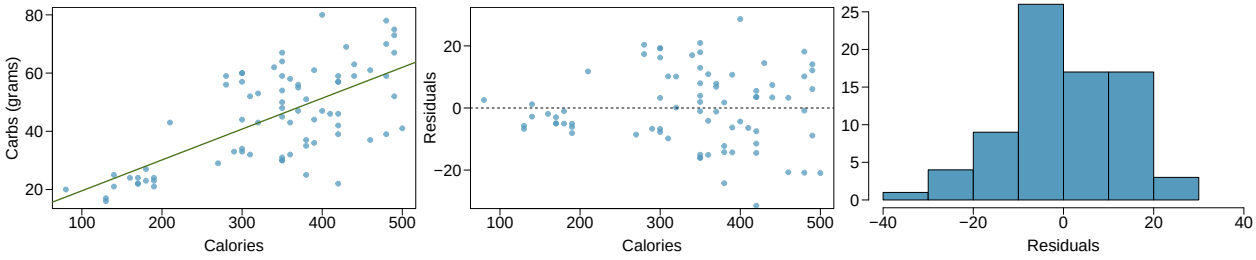
Chapter 8 - Introduction to Linear Regression

Subhalaxmi Rout

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.
(b) In this scenario, what are the explanatory and response variables?
(c) Why might we want to fit a regression line to these data?
(d) Do these data meet the conditions required for fitting a least squares line?

Answer

(*a*) The relationship between number of calories and amount of carbohydrates is positively correlated.

(*b*) In this scinario, the amount of carbohydrates is the response and calories is the explanatory variable.

(*c*) We would want to fit a regression line to these data to calculate the number of carbohydrates in an item by knowing the number of calories.

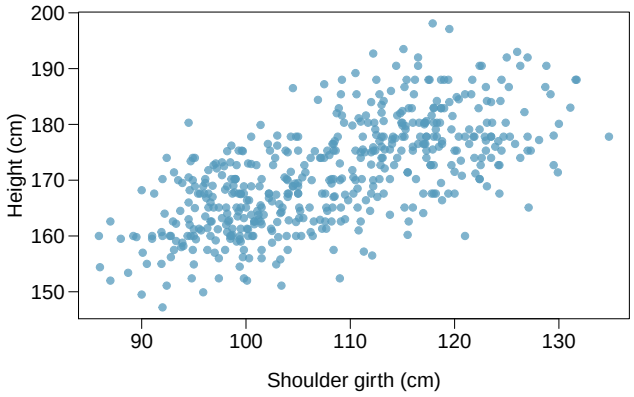
(*d*) Yes, the data meets the conditions required for fitting a least squares line.

Linearty: The data seems to linear

Nearly normal residuals: The residuals seems normal distribution but slightly left skewed.

Constant variability: This appears to be roughly the same degree of residuals above and below the horizonal line, suggesting constant variability.

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

Answer

(*a*) Shoulder girth and Height are positively correlated with each other.

(*b*) The scale of the variable should not effect either the model or the relationship. Only the coefficients of the line would change.

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.
(b) Interpret the slope and the intercept in this context.
(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.
(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.
(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.
(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

(*a*) The equation for the regression line is y = mx + c where y = predicted values m = slope c = intercept

```
# calculate slope
m <-  round(0.67 * (9.41/10.37),3)
# calculate intercept
c <- round(171.14 - m * 107.2,3)
paste0("The equation for the regression line for height = ",m," girth"," + ",c)
```

```
## [1] "The equation for the regression line for height = 0.608 girth + 105.962"
```

(*b*) For each additional cm in shoulder girth the model predicts an additional 0.608 cm in height. At a shoulder girth of 0 cm we would expect a height of 105.962 cm. 0 shoulder girth does not possible in this context so, the intercept serves only to set the height of the line.

(*c*) The value of R given : 0.67

```
R <- 0.67
R ^ 2
```

```
## [1] 0.4489
```

```
paste0("R2 = ",R ^ 2,",this means linear model accounts for about 45 percent of the variability in heig
```

```
## [1] "R2 = 0.4489,this means linear model accounts for about 45 percent of the variability in height.
```

(*d*) Given, shoulder girth of the student is 100 cm.

```
height <- c + m * 100
paste0("The students predicted height to be ", height ,"cm based on the model")
```

```
## [1] "The students predicted height to be 166.762cm based on the model"
```

(*e*) A residual means the difference between the observed and predicted value of the response variable. There are 2 types of residuals 1. Positive residuals 2. Negative residuals

```r
residual <- 160 - height
paste0("The residual is ",residual)
```

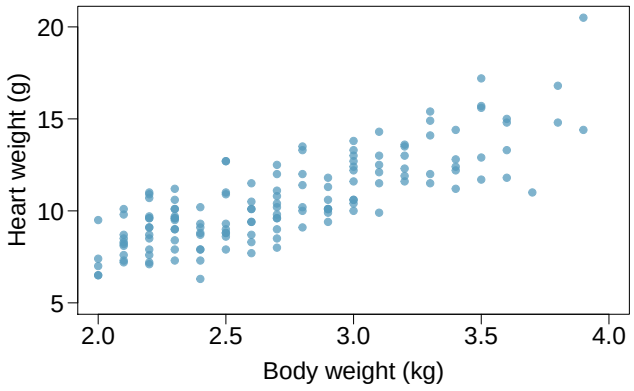
```
## [1] "The residual is -6.762"
```

(f) It will not be appropriate to use linear model to predict the height of one year child because the smallest shoulder girth in the data set this model is based on is about 80 cm but this child has shoulder girth of 56 cm.

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.
(b) Interpret the intercept.
(c) Interpret the slope.
(d) Interpret R^2.
(e) Calculate the correlation coefficient.

Answer

(*a*) Linear model y = mx + c Here m = 4.043 c = -0.357 heart weight = 4.043 * body weight - 0.357

(*b*) At a body weight of 0 kg we would expect a heart weight of -0.357 gm. 0 body weight and -0.357g heart weight is quite impossible. So the intercept serves only to set the height of the line in the model.

(*c*) The model predicts an additional 4.043 gm in heart weight for each additional kg in body.

(*d*) 64.66% of the variability in heart weight is explained by body weight in this model.

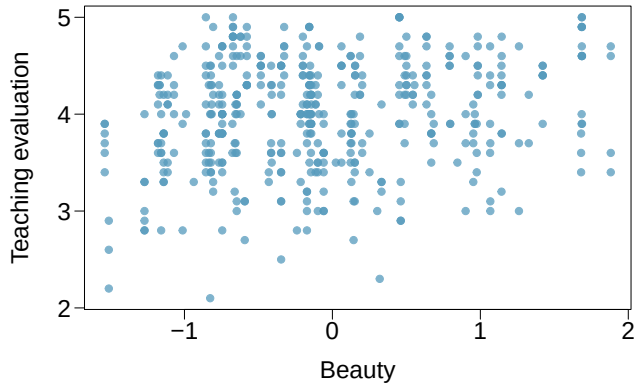
(*e*) Correlation coefficient = square root of (R^2)

```
r <- round(sqrt(0.6466),3)
paste0("The corelation coefficient is ", r)
```

```
## [1] "The corelation coefficient is 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.

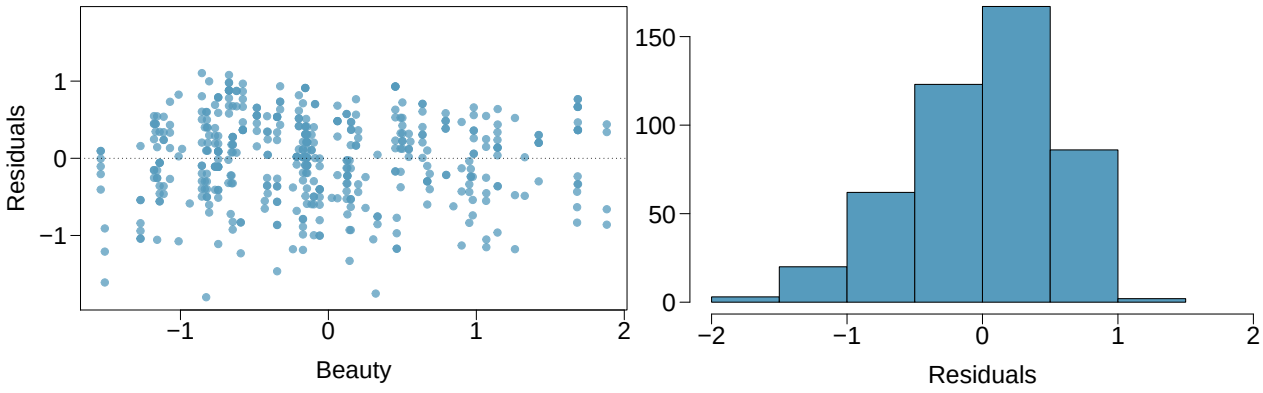
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000

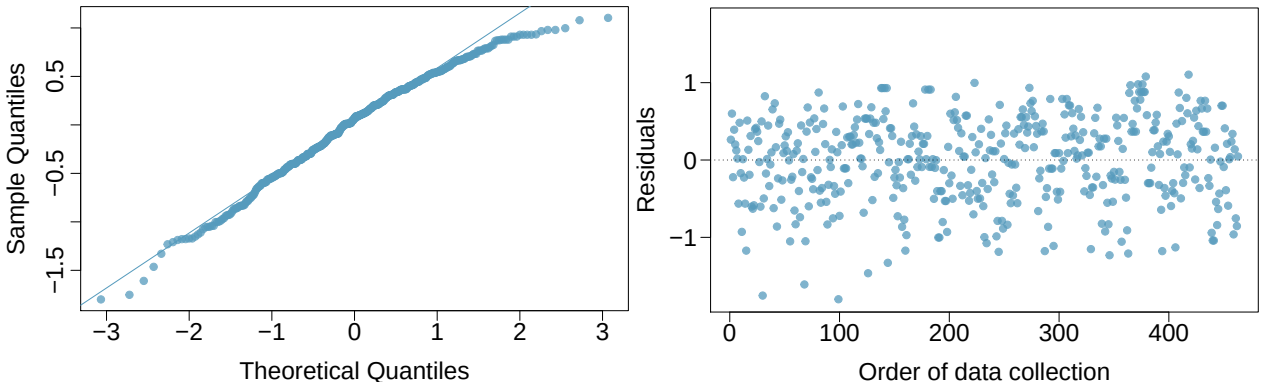


(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

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

(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

(*a*) Linear model y = m*x + c

```
y <- 3.9983
x <- -0.0883
c <- 4.010
m <- round((y -c)/x,3)
paste0("The slope is ", m)
```

```
## [1] "The slope is 0.133"
```

(*b*) The value of slope is 0.133 (positive). So we can conclude that this data shows convincing evidance of the slope being positive. This product will never be negative.

(*c*) Conditions for linear regression:

Linearity: There seems to be some trend on the scatterplot, so we can confirm linearity.

Normal Residuals: The Histogram and qqplot shows that the residuals appear to be normal with slight left skewed.

Constant Variability: Residual plot confirms this. Yes, all the conditions are satisfied.