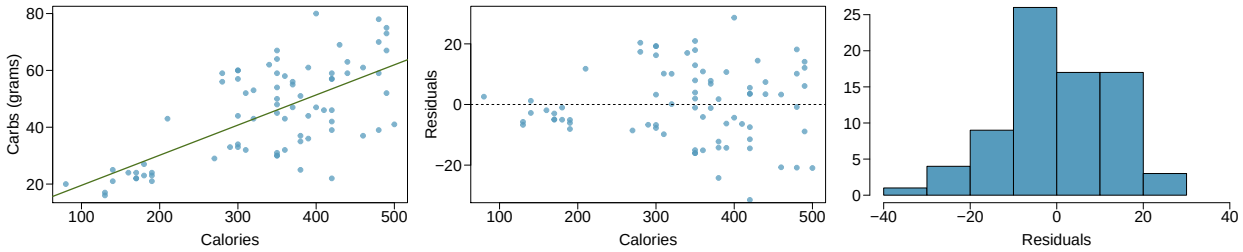
Chapter 8 - Introduction to Linear Regression

Cameron Smith

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.

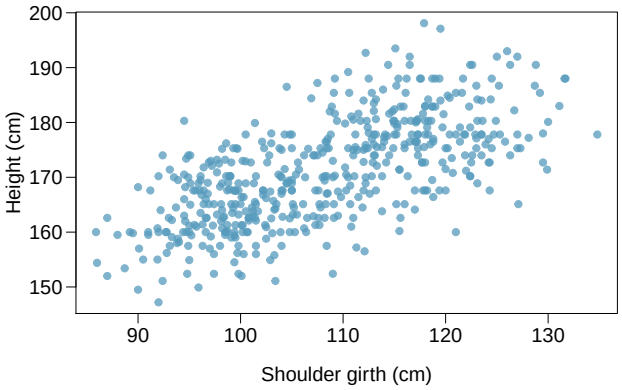


(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.
(b) In this scenario, what are the explanatory and response variables?
(c) Why might we want to fit a regression line to these data?
(d) Do these data meet the conditions required for fitting a least squares line?

Answers

a) There is a strong, positive linear correlation between the two variables.
b) The explanatory variable is calories, and the response variable is carbs.
c) Due to the strong linear correlation we can use linear regression to predict the response variable.
d) Yes, it appears so. Based on the above data and charts it is (1) linear, with (2) nearly normal residuals; (3) constant variability; and (4) we assume independent observations.

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. 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?

Answers

a) There is a strong positive linear correlation between shoulder girth and height.
b) The relationship would stay the same, but the x scale would be different. I have plotted it below for reference.

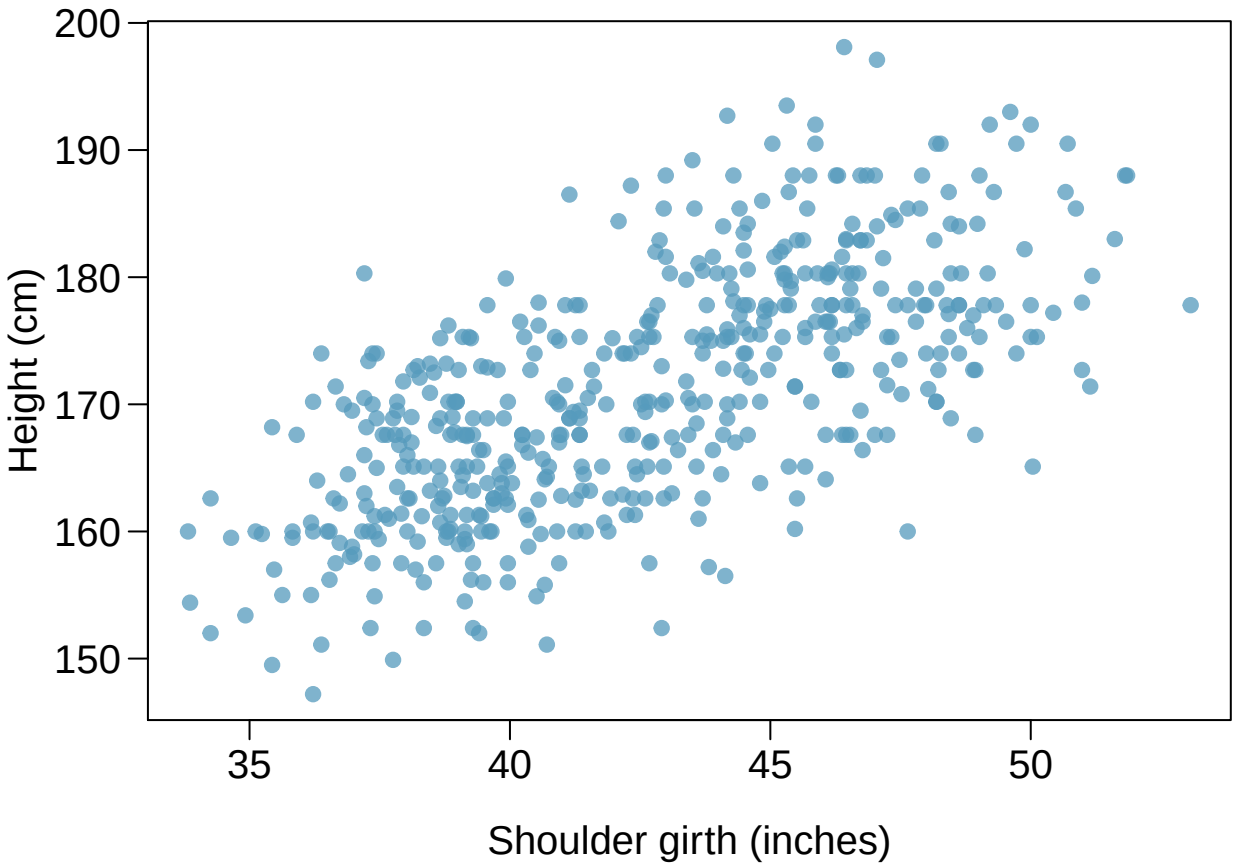
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
bdimstmp <- bdims %>% select(hgt, sho_gi)
bdimstmp <- bdimstmp %>% mutate(sho_gi_inches = sho_gi / 2.54)

# plot height vs. shoulder girth in inches-----------------------------------
par(mar = c(3.8, 3.8, 0.5, 0.5), las = 1, mgp = c(2.7, 0.7, 0),
    cex.lab = 1.25, cex.axis = 1.25)
plot(bdimstmp$hgt ~ bdimstmp$sho_gi_inches,
     xlab = "Shoulder girth (inches)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2])
```


200
190
180
170
160
150
Height (cm)
35
40
45
50
Shoulder girth (inches)

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?

ANSWERS

a)

$$\hat{y} = 105.9651 + 0.60797 \times shoulder\_girth$$

b) The slope is 0.60797, and the intercept is 105.9651. This indicates a positive correlation.
c) 44.89%
d) 166.7626
e) The residual is -6.7626 in this case. Residual is the difference between the observed value and what was predicted (i.e. the error).
f) No because it falls outside of the scope of the observed data that we have used to build the model.

```
meany <- 171.14
sdy <- 9.41
meanx <- 107.20
sdx <- 10.37
r <- 0.67

slope <- r * (sdy / sdx)
slope
```

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

```
yint <- meany - (slope * meanx)
rsquared <- round(r^2 * 100, 2)
rsquared
```

```
## [1] 44.89
```

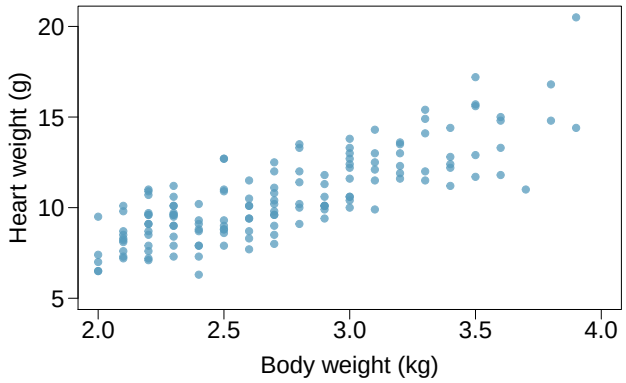
```
x <- 100

# Calculate answer for (d), Y = a + bX where a = y intercept, b = slope
ypredicted <- yint + slope * x
ypredicted
```

```
## [1] 166.7626
```

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.

ANSWERS

a)

$$\hat{y} = -0.357 + 4.034 \times body\_weight$$

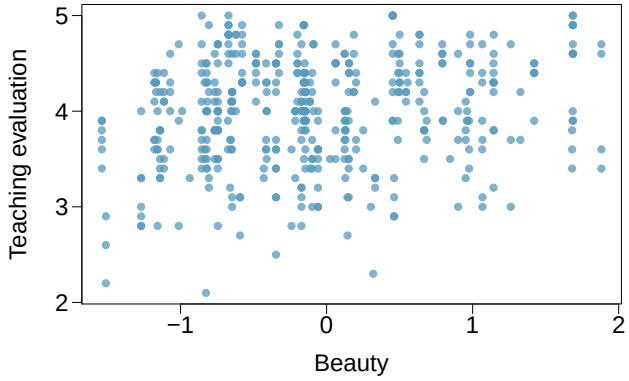
b) The intercept is -.357
c) The slope is 4.034
d) R squared is 64.66%, which means that this % of variability can be explained based on the liner model.
e) The correlation coefficient is .8041144

```
sqrt(.6466)
```

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

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.

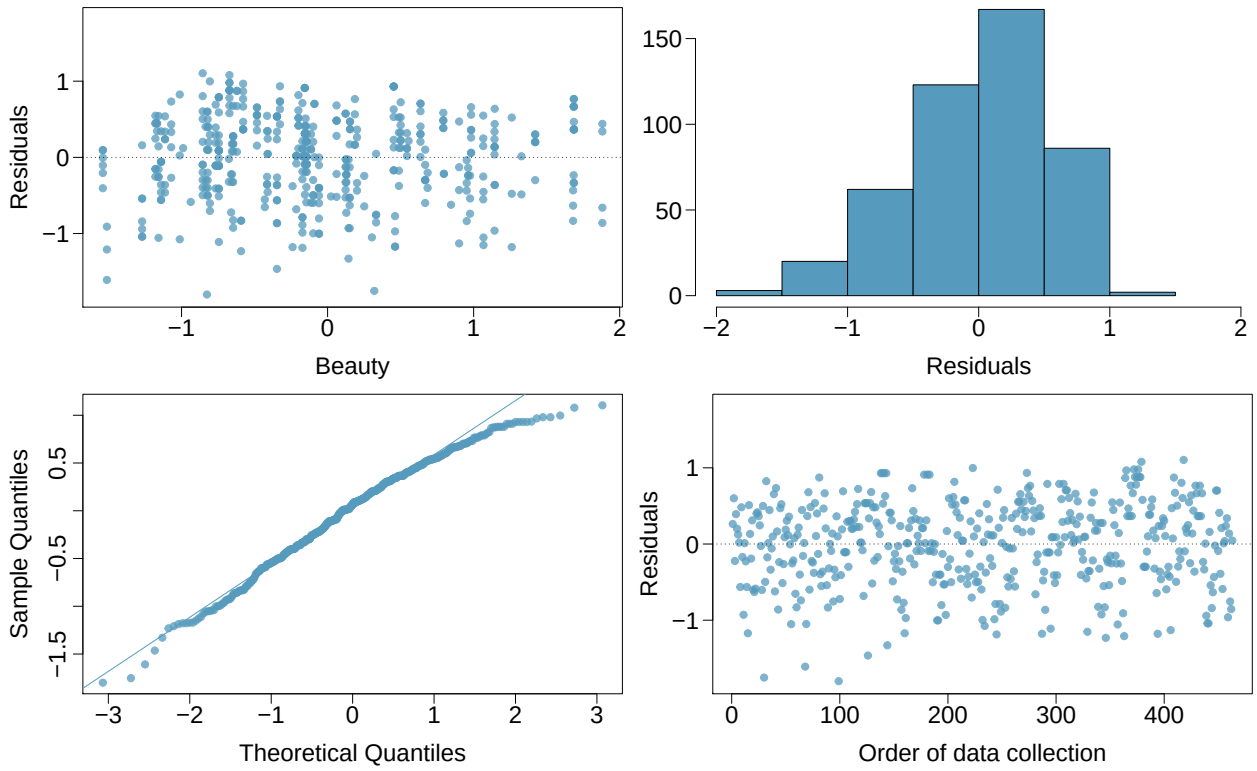
ANSWERS

a) To find the slope we just need to solve for b (since we have x and y from the means). Per the below calculation, the slope (i.e. 'b') is 0.1325028.
b) Yes, because the slope as calculated above is positive.
c) The conditions (from pg. 318) are:

- Linearity - yes, as seen in the scatter plot which shows a clear positive linear correlation
- Nearly normal residuals - yes, based on the histogram
- Constant variability - yes, based on the residuals plot
- Independent observations - assumed yes

```
# Calculate slope for (a)
meanx <- -.0883
meany <- 3.9983
yint <- 4.010
# Calculate answer for (a), Y = a + bX where a = y intercept, b = slope (solve for b)
# meany = yint + b (meanx), so...
(meany - yint) / meanx
```

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


Residuals
Beauty
Residuals
Residuals
Sample Quantiles
Theoretical Quantiles
Residuals
Order of data collection