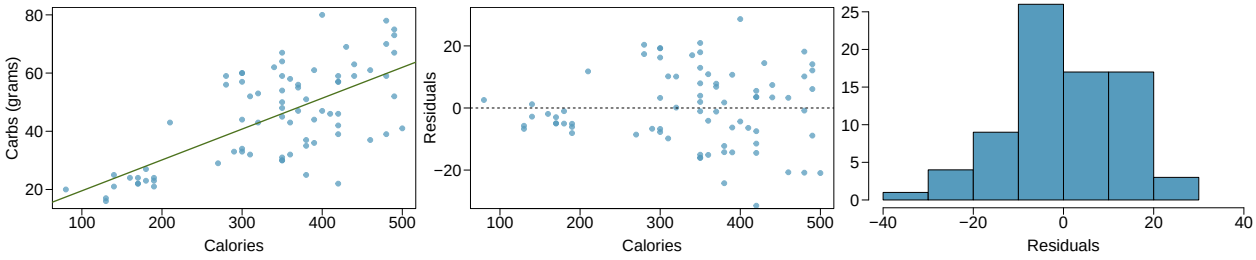
Chapter 8 - Introduction to Linear Regression

Sam Reeves

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.

It looks like there is a positive trend regarding calories ~ carbohydrates. There is stronger variance on the high end of calories, and the residuals are almost normally distributed.

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

explanatory: calories response: carbohydrates

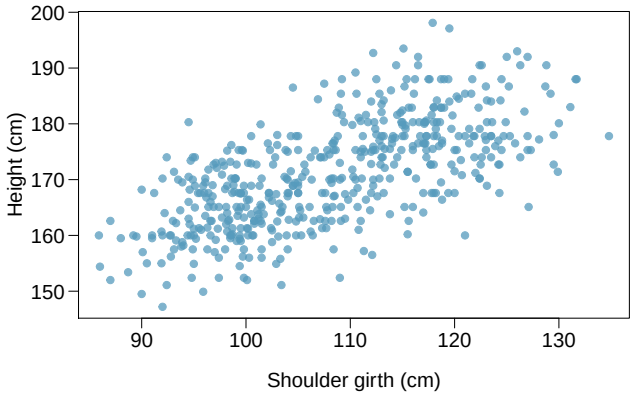
(c) Why might we want to fit a regression line to these data?

There may be other foods you can buy at Starbucks which do not have listed amounts of carbohydrates. Perhaps you know one value but not the other, and you'd like to make an assumption.

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

Unfortunately, they do not because the data do not exhibit constant variability around the trend line.

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.

```
summary(lm(hgt ~ sho_gi, bdims))
```

```
##
## Call:
## lm(formula = hgt ~ sho_gi, data = bdims)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2297  -4.7976  -0.1142   4.7885  21.0979
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 105.83246    3.27245   32.34   <2e-16 ***
## sho_gi        0.60364    0.03011   20.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.026 on 505 degrees of freedom
## Multiple R-squared:  0.4432, Adjusted R-squared:  0.4421
## F-statistic:   402 on 1 and 505 DF,  p-value: < 2.2e-16
```

There is an evident but weak correlation showing a positive linear trend.

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

The slop of the line would reflect the ratio of distance from inches to centimeters, but the relationship would not change. Perhaps you would see a larger error as a result of residuals being squared.

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.

$$y = \beta_0 + \beta_1 x$$
$$b_1 = \frac{s_y}{s_x} \times R = 0.608$$
$$b_0 = \bar{y} - b_1 \bar{x} = 105.96$$
$$y = 105.96 + 0.608x$$

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

I suppose it suggests that a person with 0 cm shoulder girth would be 107 cm tall.... This is completely meaningless.

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

$R^2 = 0.4489$ This suggests that the correlation is weak, as R^2 is closer to zero than to 1.

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

166.76cm

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

Residual in this case is -6.76cm.... It means they don't fall exactly on the line, but they are within a reasonable margin of error.

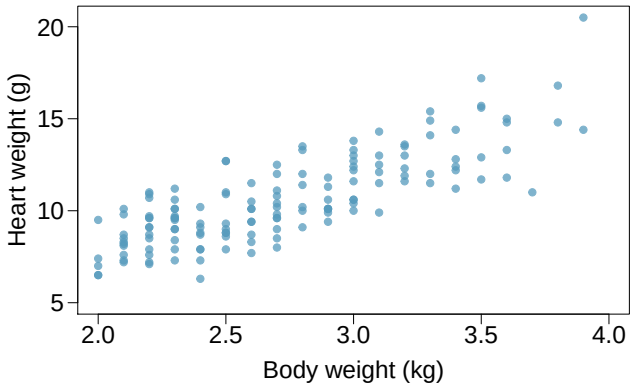
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

Absolutely not.

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.

$$y = -0.357 + 4.034x$$

(b) Interpret the intercept.

The intercept is not useful in this context!

(c) Interpret the slope.

There is a positive relationship between body weight and heart weight.

(d) Interpret R^2.

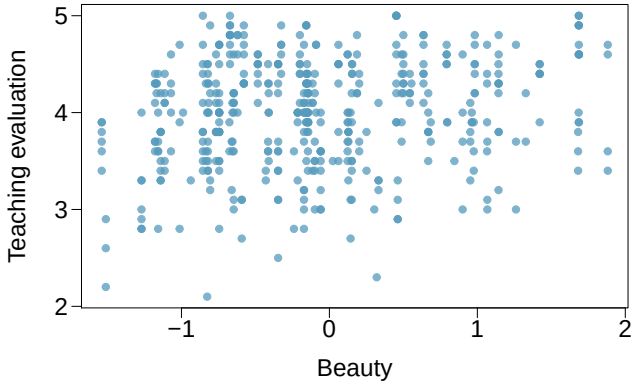
The correlation is strong, because it falls closer to 1 than to 0.

(e) Calculate the correlation coefficient.

$$R = \sqrt{R^2} \approx 0.804$$

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.

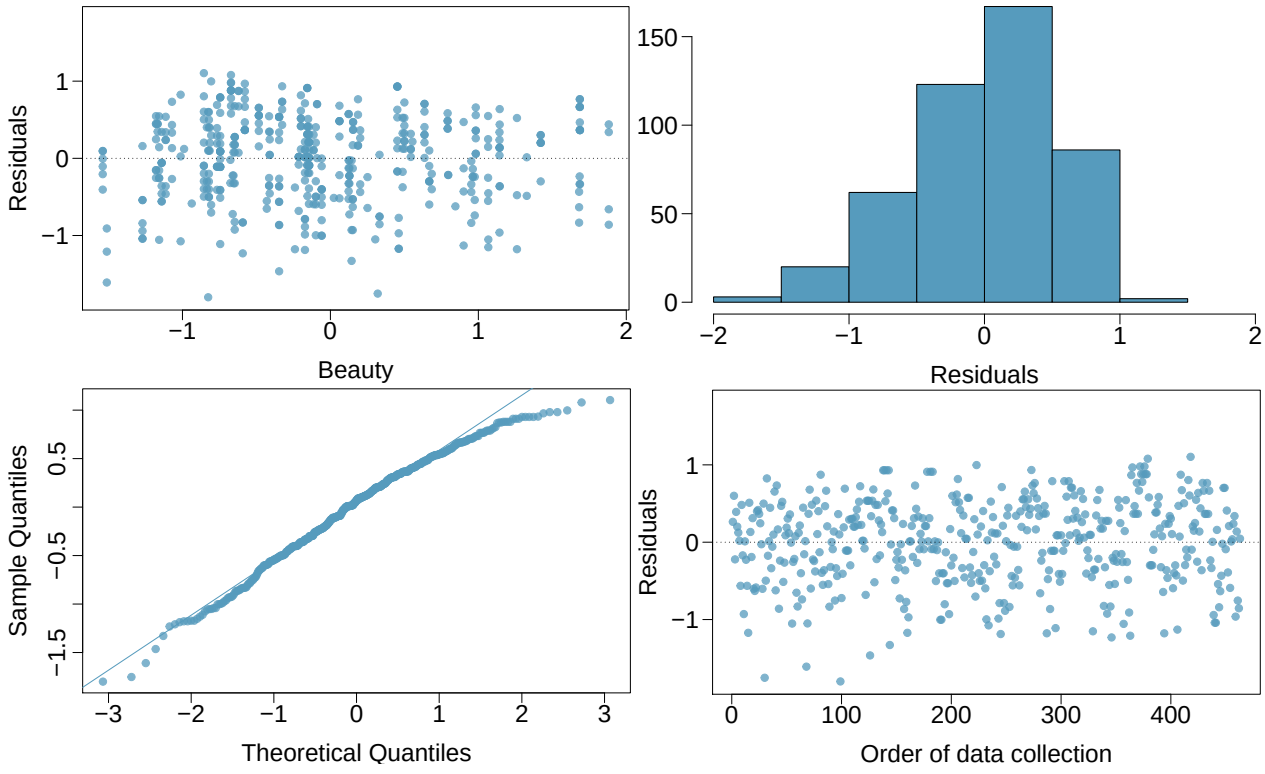
$$b_0 = \bar{y} - b_1\bar{x}\ b_1 = 0.1325$$

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

I'm not convinced. The slope is nearly zero, and though the residuals exhibit normal variance, they are not really normally distributed.

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

The data does not exhibit distinct linearity, the residuals are not quite normally distributed, the variability of residuals is basically constant, and the observations are independent.


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