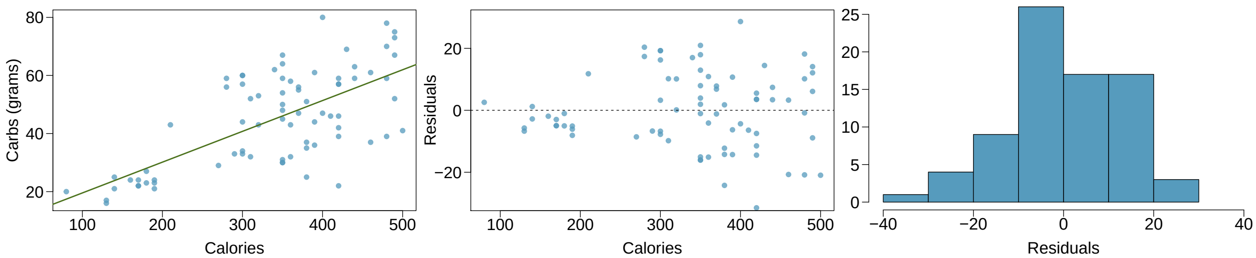
Chapter 8 - Introduction to Linear Regression

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: It is a positive linear relationship with a moderate correlation.

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

Answer: The explanatory variable is calories and the response variable is carbs

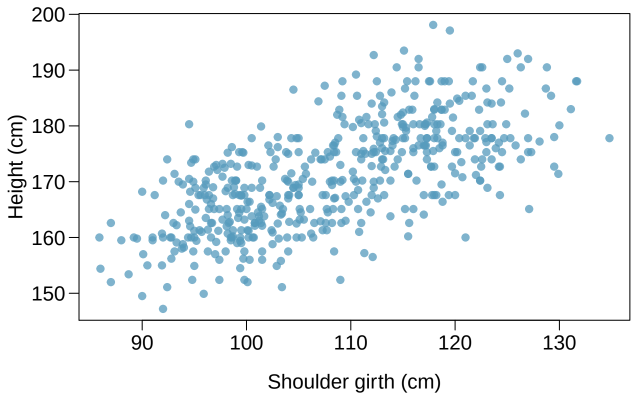
(c) Why might we want to fit a regression line to these data?

Answer: We may want to fit a regression line to these data because we want to try to predict the amount of carbs in a menu item since that information is not provided by Starbucks.

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

Answer: I would say that only 1 of the three conditions is met for linearity and that is that the data does show a positive linear trend.

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.

Answer: These is moderately strong positive relationship between shoulder girth in cm and height in cm since most of the point are tightly packed around the line.

(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 because the correlation coefficient is unitless and is not affected by changes to scale of either variable.

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.

First calculate the slope: $\beta_1 = \frac{Sy}{Sx} R$

b_1 = (9.41 / 10.37) x 0.67

b_1 = 0.91 x 0.67

b_1 = 0.61

Second calculate the point-slope equation using the (107.20,171.14):

$y - y_0 \;=\; b_1(x - x_0)$

y - 171.14 = 0.61(x - 107.20) –> Expand the right side and add 171.14 to each side

Linear model equation: $\hat{height} = 105cm + 0.61$ x girth

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

Answer: For every 100cm of girth, we would expect height to increase by 61cm over 105cm.

(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 = 0.62 therefore R^2 = .45. There is an increase of 45% in data variability by using the information about the persons girth 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.

Answer: $\hat{height}$ = 105 + 0.61 x 100 = 166cm. The prediction of the person height based on shoulder girth of 100cm is 166cm tall.

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

Answer: $e_x = y_x - \hat{y_x}$

e_x = 160 - 166 = -66cm negative residual. The negative residual indicates that the linear model overpredicted the height for student in (d).

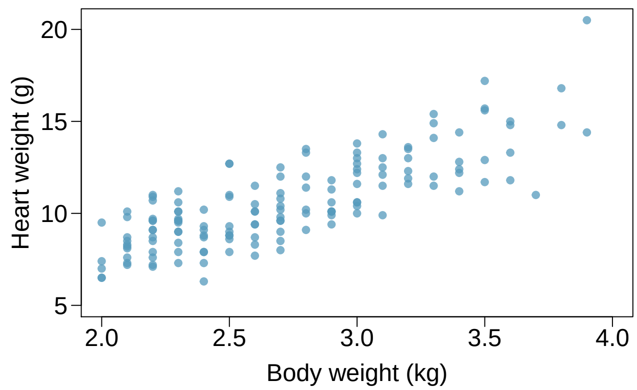
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

Answer: It is not appropriate to use this linear model on a one year child since that child at a minimum has to be 105cm tall.

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: $\widehat{heartweight}$ = -0.357 + 4.034 x catweight

(b) Interpret the intercept.

(c) Interpret the slope.

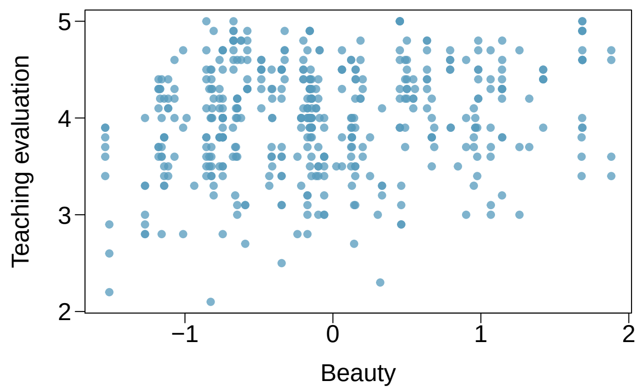
(d) Interpret R^2.

Answer: $R^2 = 64.66\%$. There is an increase of 64.66% in data variability by using the information about the person's girth for predicting height using a linear model.

(e) Calculate the correlation coefficient. Answer: The square root of 0.6466 = 0.804. The correlation coefficient (R) = .804 which indicates a very strong positive correlation between the variables.

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.

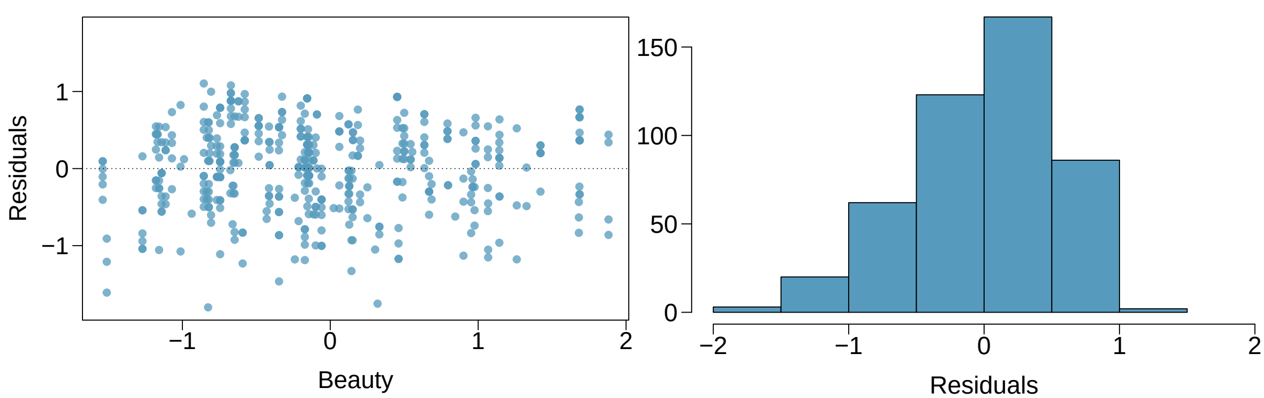
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	$\boxed{}$	0.0322	4.13	0.0000

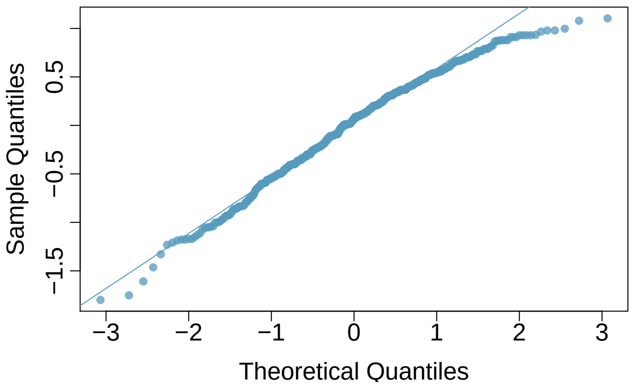


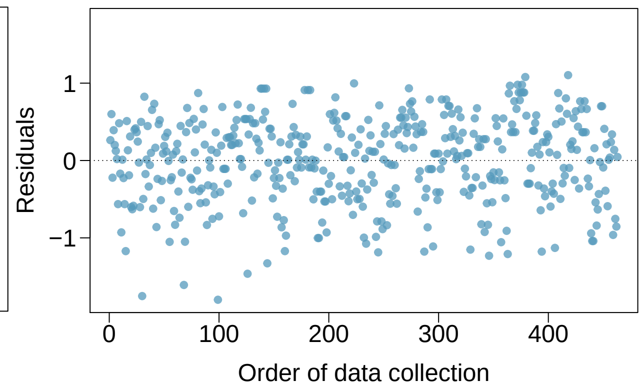
(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(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.




Sample Quantiles
Theoretical Quantiles
−1.5
−0.5
0.5
−3
−2
−1
0
1
2
3


Residuals
Order of data collection
−1
0
1
0
100
200
300
400