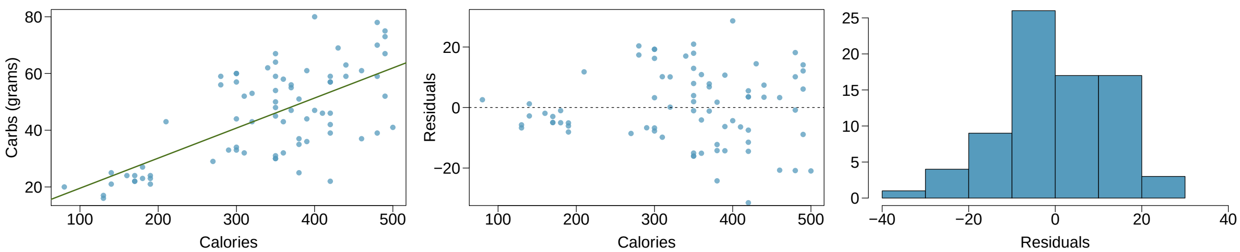
Chapter 8 - Introduction to Linear Regression

Homework completed by Chad Bailey

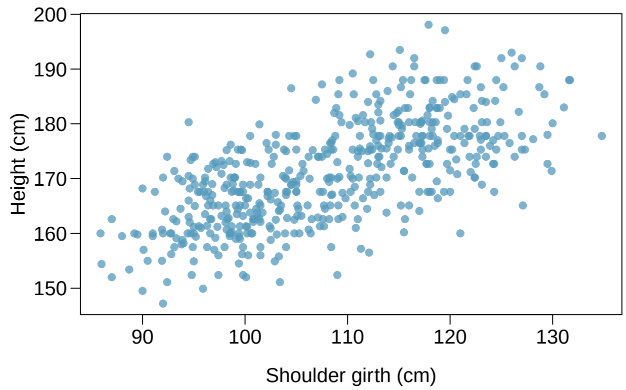
Nutrition at Starbucks, Part I. (8.22, p. 326) The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we are interested in predicting 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?

```
## Student response to problem 8.22

##  (a) calories and amount of carbohydrates have a modestly strong positive relationship
##  (b) the explanitory variable is 'calories' and the response variable is 'carbs'
##  (c) fitting a regression line to these data help in visually exploring the relationship
##      between these two variables
##  (d) No, these data do not meet all of the conditions for fitting a least squares line.
##        Linearity:  yes, the data show a visually linear trend
##        Nearly Normal Residuals: yes, residual histogram is nearly normal in shape
##        **Constant Variability: no, the residual scatter shows a weak but noticable
##                                relationship that as calories increase so does residual
##        Independent observations: yes, the observations 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. 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?

```
##  Student response to probelm 8.13

##  (a)   Shoulder girth and height have a relatively strong positive relationship

##  (b)   If shoulder girth was changed to be measured in inches while height remained
##        in centimeters, the slope of the regression line would become steeper as
##        inches are larger and therefore would compress the x-axis. However, values
##        such as R or R-squared would remain unchanged.
```

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?

```
##  Student response to problem 8.24

##  Given values
    ##  xbar  = mean shoulder girth = 107.20 cm
    ##  x_sd  = sd shoulder girth   = 10.37 cm
    ##  ybar  = mean height         = 171.14 cm
    ##  y_sd  = sd height           = 9.41 cm
    ##  R                           = 0.67

##  Derived values
    ##  b1    = regresion slope     = (y_sd / x_sd) * R
    ##                              = 0.67 * (9.41 / 10.37)
    ##                              = 0.61

    ##  b0:   = intercept           = ybar - b1*xbar
    ##                              = 171.14 - 0.67*107.20
    ##                              = 99.32


##  (a) hieght_hat = 0.61 * shoulder_grith_hat + 99.32

##  (b) slope:  for every 1 cm increase in shoulder_girth there is a 0.67 increase
##              in height
##      intercept:  an individual with shoulder_girth = 0 would be predicted to have
##                  a height of 99.32 cm. Such a height likely indicates the model
##                  was built on a sample of random adults and therefore is only
##                  appropriate to apply the model to adults.

##  (c) R-squared = R^2 = 0.67^2 = 0.4489
##      This means that 44.89% of the variance observed in height can be explained
##      by the variable shoulder_girth



##  (d) Using the model, a student with shoulder girth of 100cm would be predicted to
##      have a height of 160.32 cm
    0.61 * 100 + 99.32
```

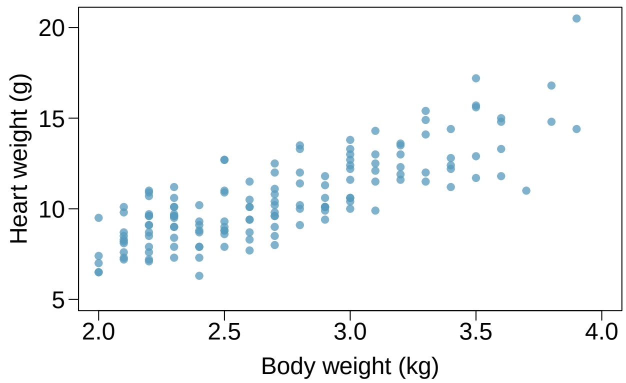
```
## [1] 160.32
```

```
##  (e) The residual for the student in part (d)
##        residual  = actual - prediction
##                  = 160 - 160.32
##                  = -0.32
##
##      The residual would be -0.32 which means that the model overpredicted the
##      individual's height by 0.32cm.

##  (f) No, this model would not be appropriate to use to predict the height of a
##      one-year old with a height of 56cm. As explained in part (b) the model's
##      intercept of 99.32 likely indicates the model was built on a sample of
##      random adults and therefore is only appropriate to apply the model to adults.
```

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

$s = 1.452$ $R^2 = 64.66\%$ $R^2_{adj} = 64.41\%$

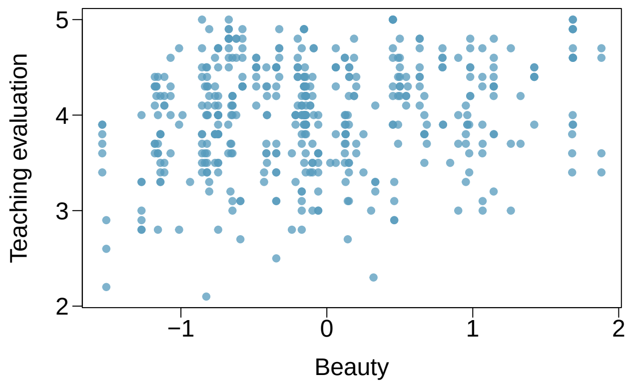


(a) Write out the linear model.
(b) Interpret the intercept.
(c) Interpret the slope.
(d) Interpret R^2.
(e) Calculate the correlation coefficient.

```
## Student response to problem 8.26

##  (a) linear model:   y = -0.357 + 4.034 * x

##  (b) The intercept is -0.357. This means if the cat's body weight was zero (0.0)
##      the model would predict a heart weight of -0.357. This indicates that model
##      does not have a great fit on the lower end cat body weight since the
##      intercept is not possible to be observed.

##  (c) The slope is 4.034. This means the model predicts an increase of 4.034g for
##      heart weight for every increase of 1kg in the cat's body weight.

##  (d) The R^2 value is 0.6466. This means 64.66% of the variation in heart weight
##      is explained by body weight.

##  (f) Calculating the correlation coeffient (R).
##        R = sqrt(R^2); and the sign is the same as B1 (slope) coefficient
##          = sqrt(0.6466)
##          = 0.8041
##      The correlation coefficient (R) is 0.8041.
```

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.

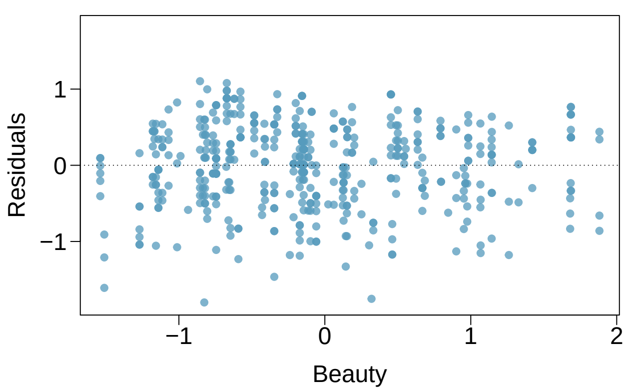
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000

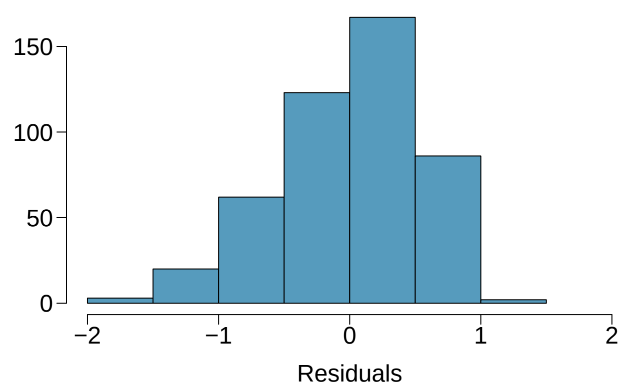


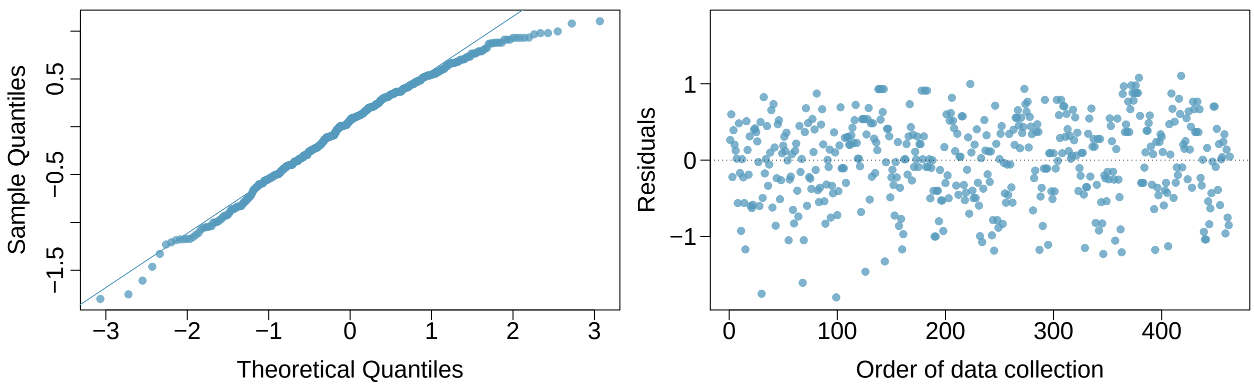
(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(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.








```
## Student response to problem 8.44

##   (a) Calculating the slope (b1) of the model
##         b1   = (ybar - b0) / xbar
##              = (3.9983 - 4.010) / (-0.0883)
##              = 0.1325

##   (b) Yes, these data provide convincing evidence that the slope of the
##       relationship is positive. The slope coefficient is positive and
##       visually the scatterplot shows a slight positive relationship.

##   (c) Evaluating the conditions of linear regression:
##
##         Linearity:  the data appear to have a mostly linear relationship in the
##                     Beauty vs Teaching evaluation scatterplot.

##         Nearly normal residuals:  the residuals appear nearly normally distributed
##                                   via the histogram of residuals

##         Constant variability: residuals appear to have no correlation with Beauty
##                               score in scatterplot of Beauty vs Residuals

##         Independent observations: residuals appear to have no correlation with
##                                   order of data collection meaning the data are
##                                   all likely independent observations.
```