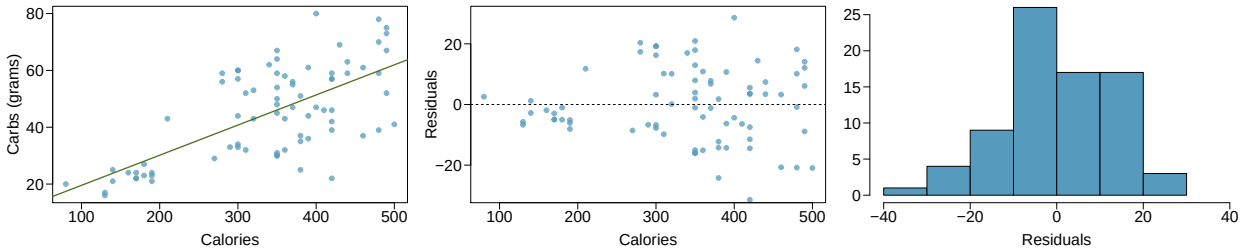
Chapter 8 - Introduction to Linear Regression

Kenan Sooklall

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.

From the first scatter plot the relationship between number of calories and amount of carbohydrates is increasing. As the number of calories goes up the more of those calories are due to carbohydrates.

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

The explanatory variable is calories and the response is carbohydrates

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

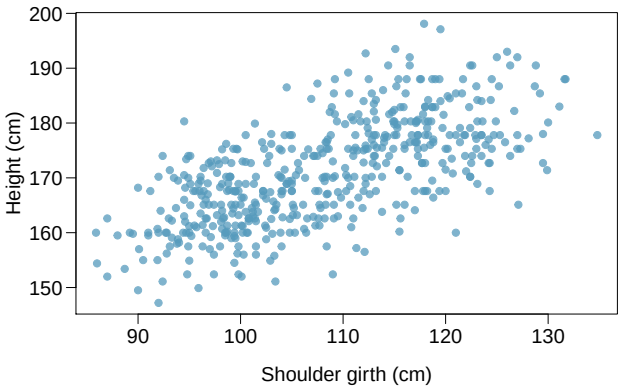
The regression line shows if there is a positive or negative relationship, in this case it's positive

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

The relationship looks linear and the residuals are normally distributed

---

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.



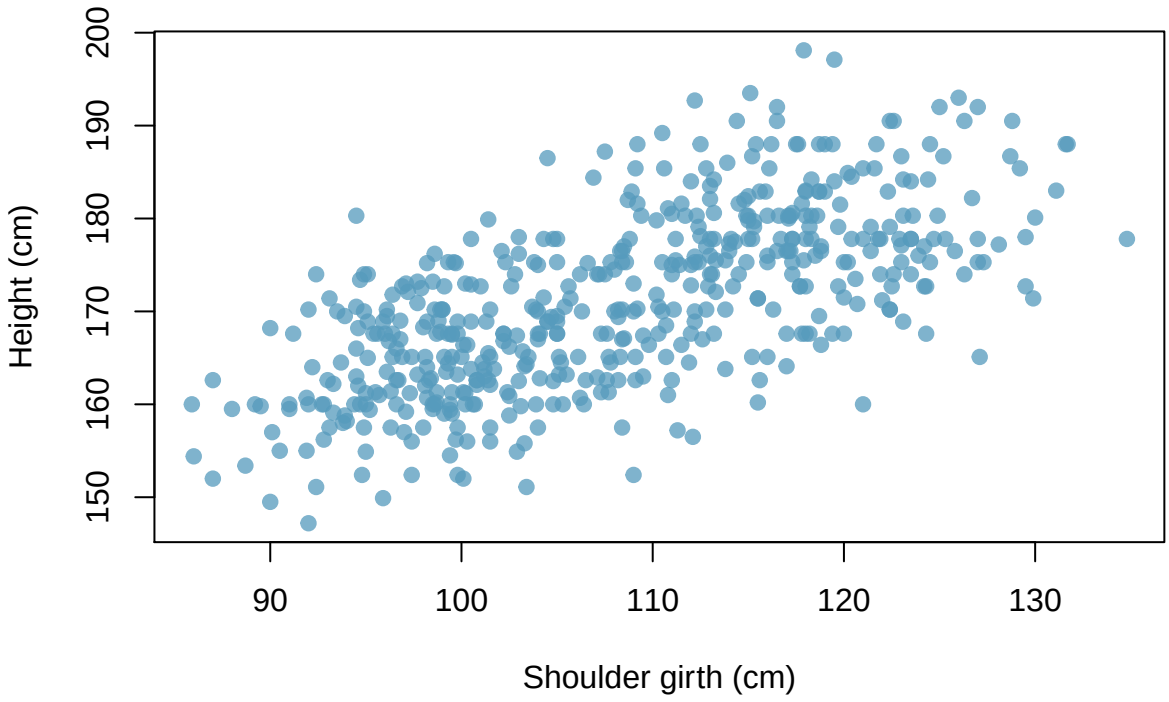
(a) Describe the relationship between shoulder girth and height.

As shoulder girth increases in cm height increases

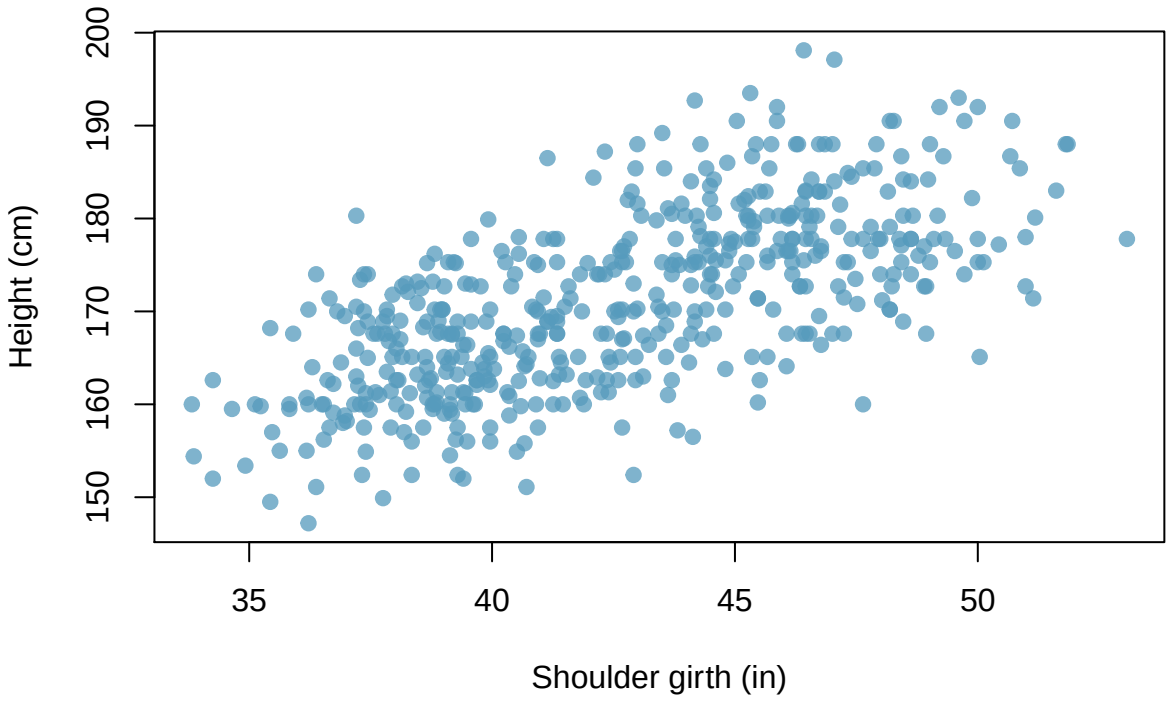
(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

```
bdims$sho_gi_in = bdims$sho_gi / 2.54

plot(bdims$hgt ~ bdims$sho_gi,
     xlab = "Shoulder girth (cm)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2])
```




```
plot(bdims$hgt ~ bdims$sho_gi_in,
     xlab = "Shoulder girth (in)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2])
```



The relationship stays the same

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.

```
smu = 107.2
ssd = 10.37

hmu = 171.14
hsd = 9.41
cor = 0.67

b1 = hsd / ssd * cor
b0 = hmu - (smu * b1)
```

$height = 105.9651 + 0.60797 * girth$

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

The average height is 106cm. An increase in shoulder girth corresponds to about 0.61cm increase in height.

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

```
ml <- lm(sho_gi_in ~ hgt, data = bdims)
summary(ml)
```

```
##
## Call:
## lm(formula = sho_gi_in ~ hgt, data = bdims)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0385 -2.1240 -0.0704  1.9163  9.1899
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.87439    2.47118  -2.782  0.00561 **
## hgt          0.28906    0.01442  20.049  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.051 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
```

About 44% of the variability in height is explained by shoulder girth.

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

```
height = 105.9651 + 0.60797 * 100
```

That student is ~167cm tall

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

167 - 160 = 7cm residual, our model overestimated the student height

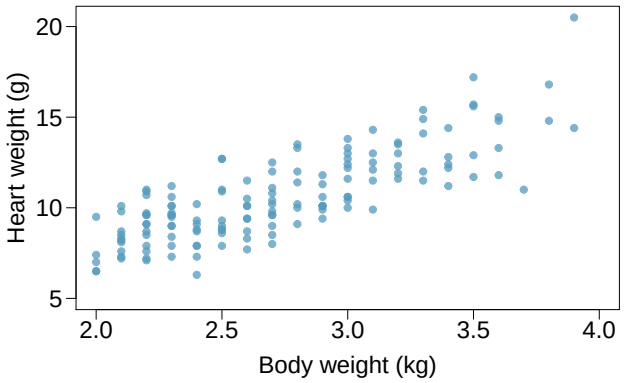
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

A shoulder girth of 56 is outside our data set and extrapolation with this model isn't appropriate.

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.

$\hat{y} = -0.357 + 4.034 * bwt$

(b) Interpret the intercept.

The intercept in this context doesn't make sense since one cann't have a negative body weight

(c) Interpret the slope.

A unit increase in body weight corresponds to 4g in heart weight

(d) Interpret R^2.

About 64.7% of the variance in heart weight is captured in body weight

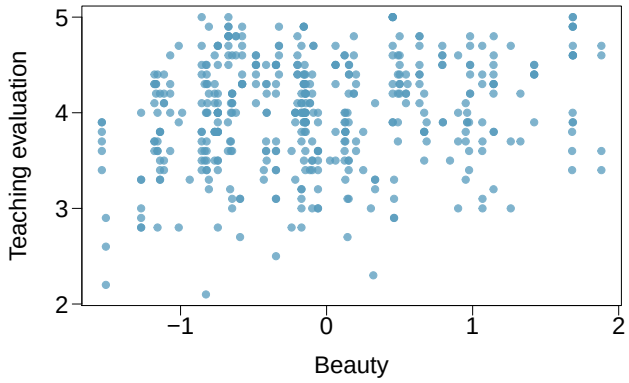
(e) Calculate the correlation coefficient.

```
sqrt(0.6466)
```

```
## [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.

```
(4.01002-3.9983)/(0-(-0.0883))
```

```
## [1] 0.1327293
```

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

```
summary(m_eval_beauty)
```

```
##
## Call:
## lm(formula = eval ~ beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
```

```
## beauty       0.13300    0.03218   4.133 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 461 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03364
## F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

The r^2 is very small (0.036) indicating there is no relationship here

(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 look like there is a linear relationship so this condition fails; however the plot below looks normal which satisfies the residual condition. The QQ plot also indicates the distribution is normal.