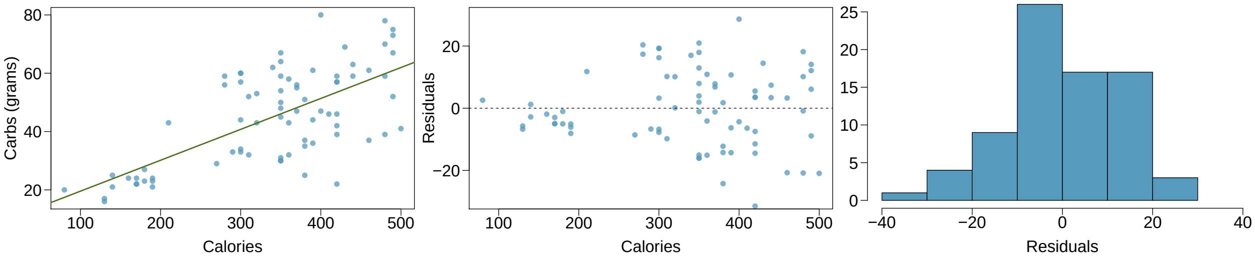
Chapter 8 - Introduction to Linear Regression

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.

It appears to have a positive, medium-strong relationship.

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

Our explanatory variable is calories and the response variable is carbohydrates

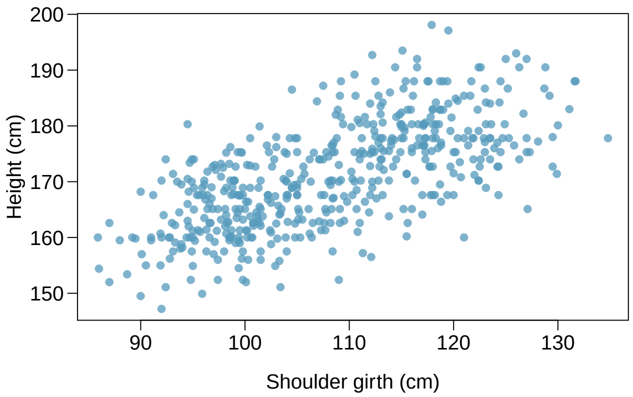
(c) Why might we want to fit a regression line to these data?

To make predictions on new data.

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

Our first condition is linearity, we see in the first graph that there is a linear relationship between both variables. Our second condition is normality. Looking at the histogram for the residuals, the distribution is slightly skewed but appears to be mostly normal. However, for the constant variability condition, it does not appear we achieved that. Looking at the second plot, it does not appear to have constant variability.

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.

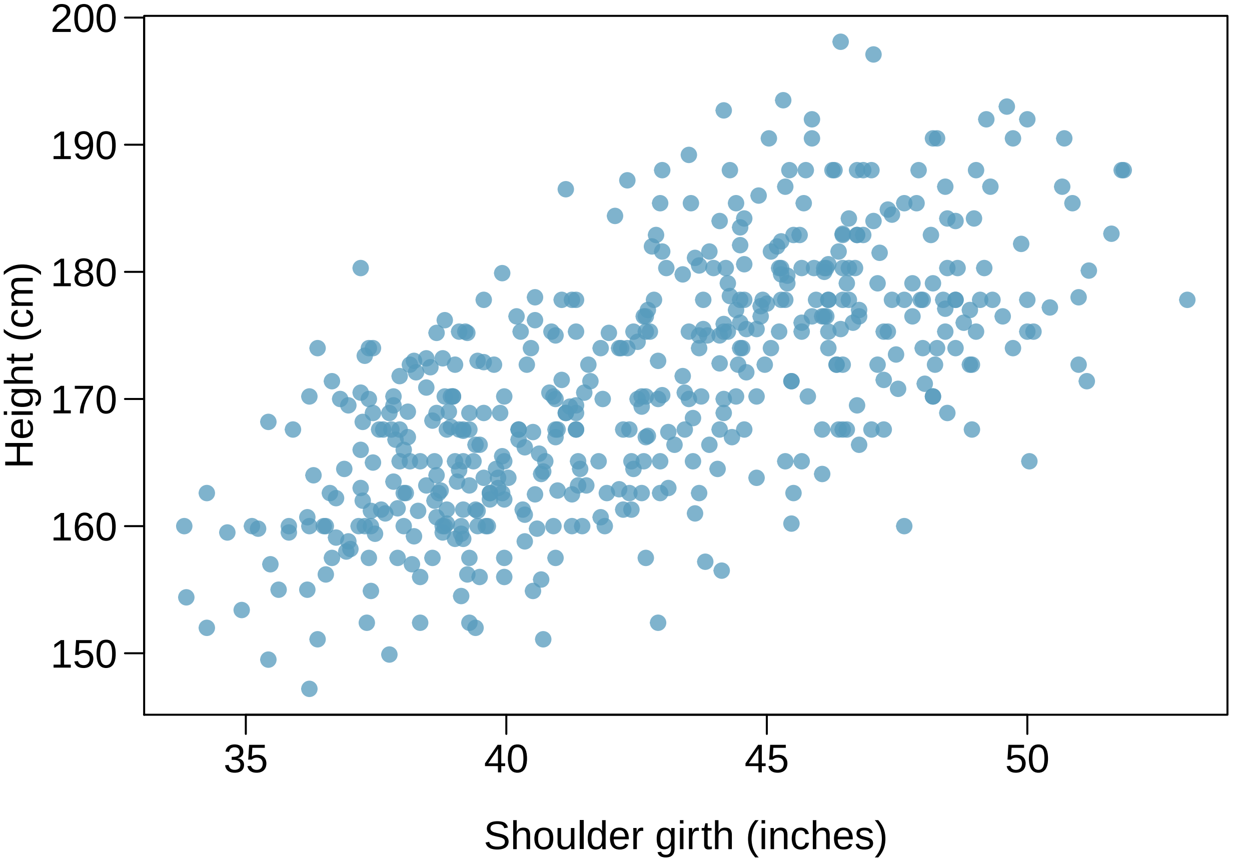
It appears that shoulder girth and height share a linear positive, medium strength relationship

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

The relationship would remain the same regardless. However, the points on the graph would shift downwards, as shown below.

```
# plot height vs. shoulder girth ------------------------------------
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)

bdims$sho_gi_in <- bdims$sho_gi * 0.393701
plot(bdims$hgt ~( bdims$sho_gi_in),
     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.

First we start by calculating the slop: $b_1 = R\,X\,\frac{s_y}{s_x} = .67$ X 9.41 / 10.37 = 0.6079749 Now we use the fact that the regression line passes through $(\bar{x}, \bar{y}) : \bar{y} = b_0 + b_1 \times \bar{x}$, therefore we have the point (107.20,171.14).

$y - y_0 = b_1(x - x_0)$
$y - 171.14 = 0.67.(x - 107.20)$

Which simplifies down to,

$\widehat{height} = 105.9651 \,+\, 0.608. \times Shoulder\,girth$

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

b_1: For each increase in should girth, the model predicts an additional .608 cm in height. b_0: Tells us what the height in cm is when shoulder girth is zero. This information is useless in this context as a person cannot have a shoulder girth of 0.

(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 is our correlation value squared, or .67 squared. Which in this case is .4489. It is a measure of how close our data are fitted to the regression line. We have a moderate value.

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

Using our equation from above $\widehat{height} = 105.9651 \,+\, 0.608. \times Shoulder\,girth$:

```
105.9651 + 0.6079749*100
```

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

Our prediction is 166.76 cm tall.

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

```
160-166.76
```

```
## [1] -6.76
```

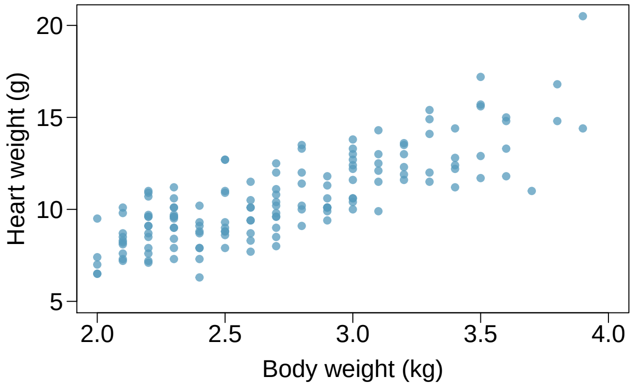
The negative value indicates that our model overestimated the value.

(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 value of 56 cm does not fit within the data used to create our model, nor do we have an individuals below the age of 18. This would not be an appropriate model to use for predictions.

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.

$\widehat{HeartWeight} = -0.357 + 4.034 \times BodyWeight$:

(b) Interpret the intercept.

It tells us that the expected heart weight in cats with 0 kg body weight is -0.357. Not meaningful in this case again.

(c) Interpret the slope.

For each additonal kg of body weight, we epect an increase in heart weight of about 4.034 grams.

(d) Interpret R^2.

Our R^2 is 64.66%. It means we can describe a moderate amount of variation with our model.

(e) Calculate the correlation coefficient.

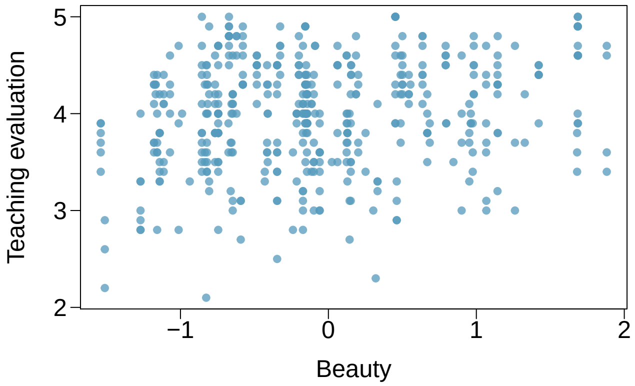
We need simply take the squure root of our R^2:

```
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.

```
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
```

As shown above, our slope is 0.13300.

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

We have a significant p-value in our model, indicating that we do have a relationship occuring between our variables. However, looking at both the graphic and our values, it's clear that if anything, it's a small relationship.

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

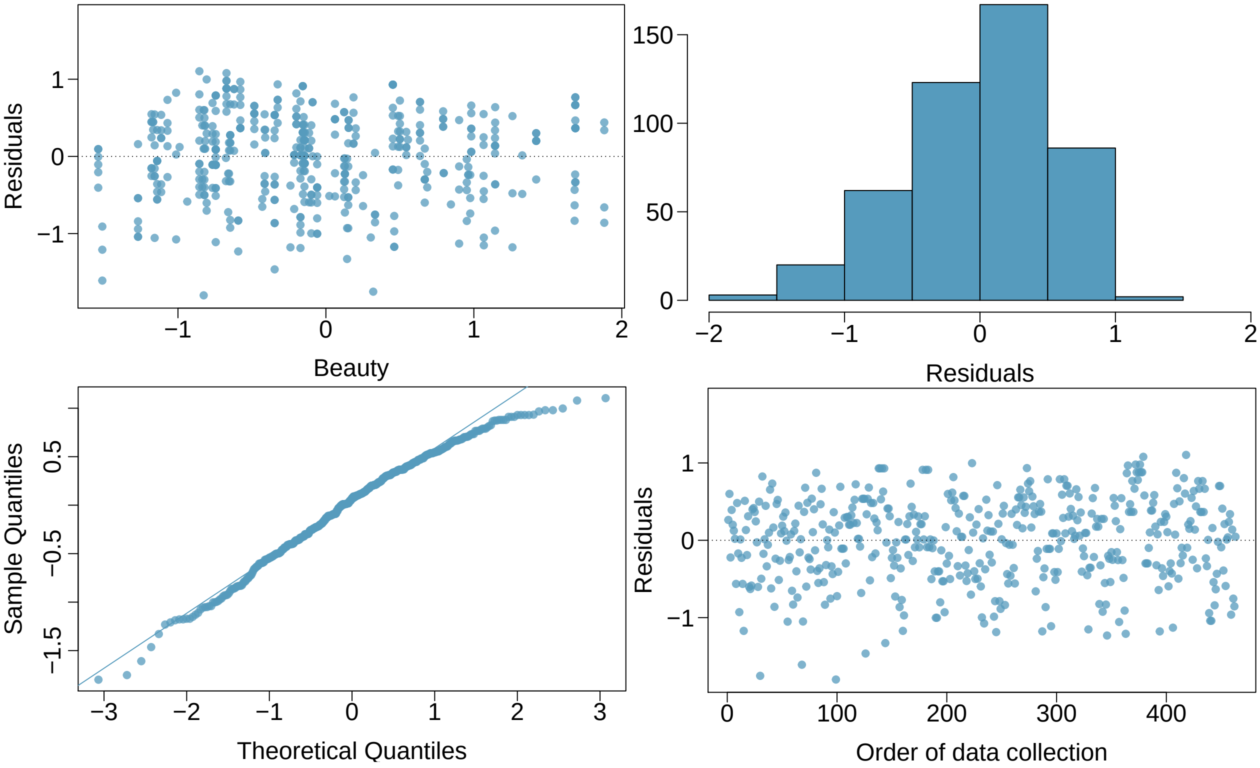
There does appear to be a somewhat weak positive relationship between our variables, so we can conclude linearity is achieved.

For normality, it appears our histogram is mostly normal, however this clearly some outliers causing some right skewing. Some extreme value handling might be needed. However, it should be okay, as shown in the normal probability plot, the majority of our values follow the line.

We can also see from both scatter plots that there is no clear pattern occurring, our residuals are random. Constant variability is achieved.

Using the fourth graphic, showing order of data collection, we can see no pattern is occurring here as well. We can conclude that our variables are independent from each other.

With this in mind, we can perform linear regression on the data and conclude that there is convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive


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