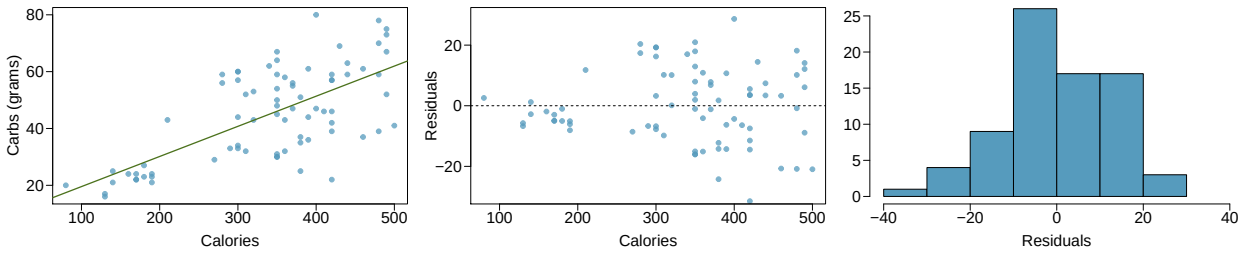
Chapter 8 - Introduction to Linear Regression

Bryan Persaud

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?

(a)

The relationship between number of calories and amount of carbohydrates that Starbucks food menu items contain is a linear relationship where most of the time if calories increases, carbohydrates increases as well.

(b)

The explanatory and response variables are calories is the explanatory variable and carbohydrates is the response variable.

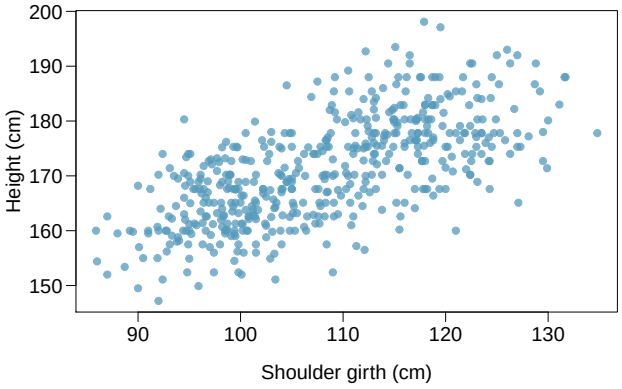
(c)

We want to fit a regression line to these data because if we are looking to track the amount of carbohydrates based on the number of calories, then a regression line would best be used to predict this.

(d)

These data do not meet the conditions required for fitting a least squares line. The conditions to be met are linearity, this is met by the scatterplot graph shows it is fairly linear. Nearly normal residuals is sort of met by we see from the histogram that the residuals are fairly normal with a slight left skew. Constant variability is not met by from the plots we see the data fits better for lower number of calories than higher numbers as there are a lot more higher values for the residuals. Therefore, the conditions are not met since constant variability is not met.

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.19 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?

(a)

The relationship between shoulder girth and height is almost always as shoulder girth increases, height increases. This shows on most cases that the longer the shoulder girth, the taller the person.

(b)

If the shoulder girth was measured in inches while the units of height remained in centimeters the relationship would remain 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.
(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?

(a)

```
shoulder_girth_mean <- 107.20
shoulder_girth_sd <- 10.37
height_mean <- 171.14
height_sd <- 9.41
correlation <- 0.67
slope <- correlation * (height_sd / shoulder_girth_sd)
slope
```

```
## [1] 0.6079749
```

```
intercept <- height_mean - slope * shoulder_girth_mean
intercept
```

```
## [1] 105.9651
```

The equation of the regression line for predicting height is height = 105.97 + 0.61 * shoulder girth.

(b)

The slope tells us the predicted increase in height, in cm, for every one cm increase in shoulder girth. For every 1 cm increase in shoulder girth, there will be an additional 0.61 cm to the height. The intercept tell us the height, in cm, when shoulder girth = 0. This is not useful information since a person cannot have a shoudler girth = 0.

(c)

```
r_squared <- correlation^2
r_squared
```

```
## [1] 0.4489
```

R^2 of the regression line for predicting height is 0.45. This means that 45% of the variability in height can be explained by the shoudler girth.

(d)

```
shoulder_girth_100 <- intercept + slope * 100
shoulder_girth_100
```

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

The predicted height for the random student with a shoulder girth of 100 cm is 166.76.

(e)

```
160 - shoulder_girth_100
```

```
## [1] -6.762581
```

The residual for the student from part (d) is -6.76. This means that the model overestimated the height of the student.

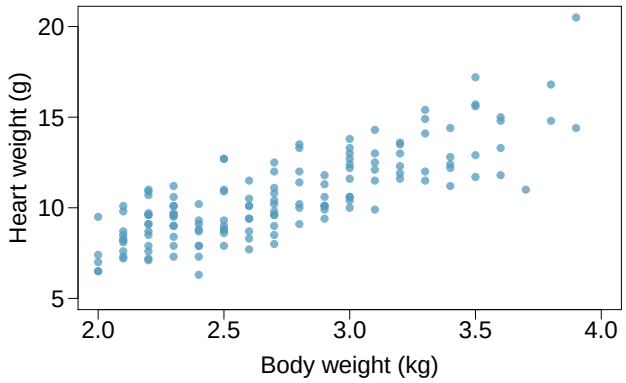
(f)

It would not be appropriate to use this linear model to predict the height of this child because the shoulder girth of 56 cm is too far off from the model's range. This model was most likely taken from adults, so it would not be appropriate to use it on children.

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.

(a)

The linear model is heart weight = (4.034 * body weight) - 0.357. This was calculated using the table above.

(b)

The intercept is -0.357 g. This is not useful since it is telling us that the height weight is -0.357 g when the body weight is 0 kg, but a cat can't have a body weight = 0 kg.

(c)

The slope is 4.034 g. This tells us that if body weight increases by 1 kg, height weight increases by 4.034 g.

(d)

R^2 = 64.66%. This means that 64.66% of the variability in heart weight can be explained by body weight.

(e)

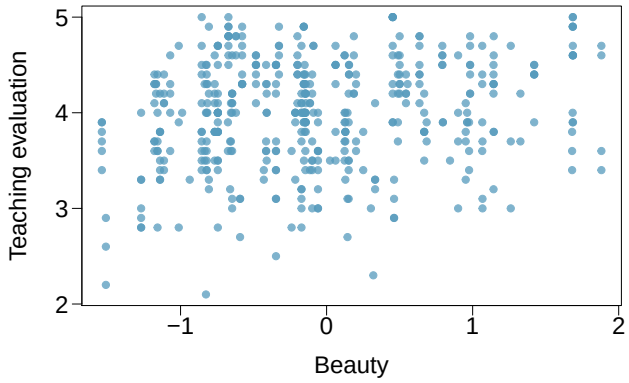
```
sqrt(0.6466)
```

```
## [1] 0.8041144
```

The correlation coefficient is 0.80.

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.

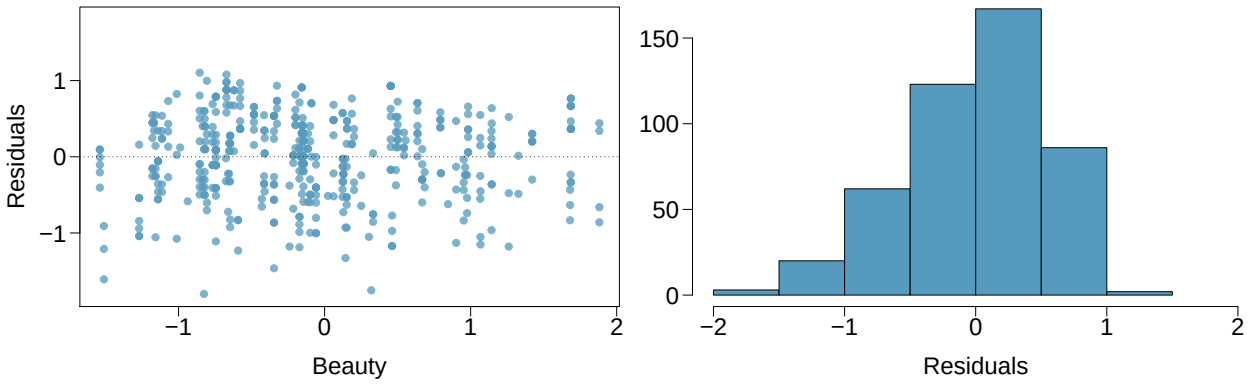
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000

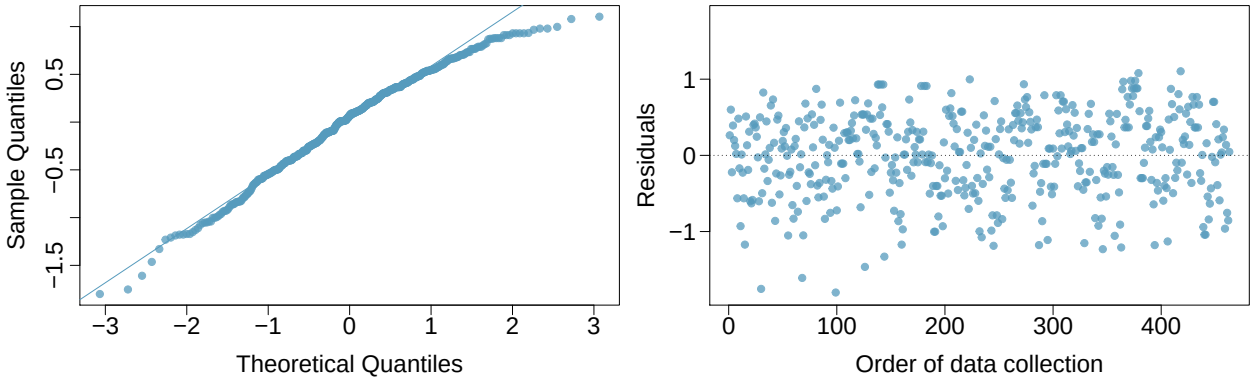


(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(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.





(a)

```
(3.9983 - 4.010) / -0.0883
```

```
## [1] 0.1325028
```

The slope is 0.133.

(b)

These data do provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive because the slope calcuated above is postive.

(c)

The conditions required for linear regression are linearity, nearly normal residuals, and constant variability. Linearity can be assumed to be met as there is no clear pattern on the scatterplot graphs though a weak linear relationship may be shown. Nearly normal residuals is met as the histogram shows a fairly normal distribution. Constant variability is met since the scatterplot graphs show we can assume that the residual variability is constant.