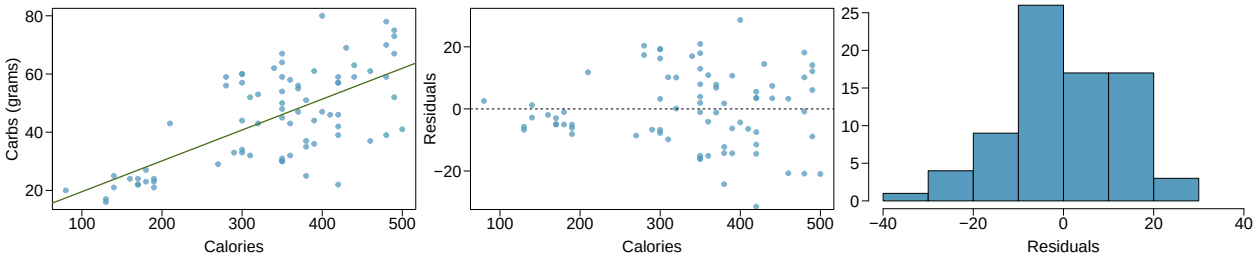
Chapter 8 - Introduction to Linear Regression HomeWork_8

Salma Elshahawy

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.

The relationship between calories and amount of carbo seems to be linear but not strong

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

The explanatory is Calories, and the response variable is Carbs

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

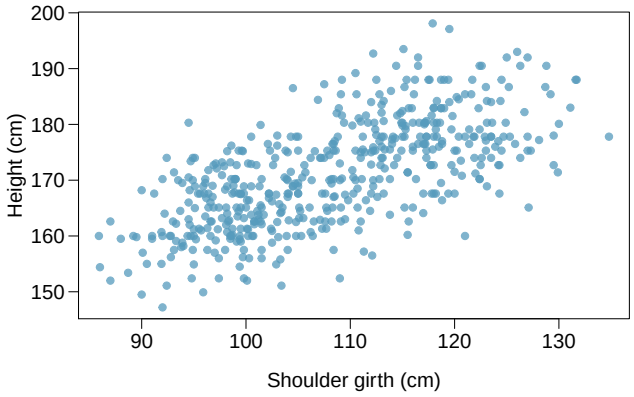
We need to do that to have a prediction of the amount of Carbs in the response variable

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

Linearity = > yes, it follows a normal distribution according to histogram Nearly normal residuals => faily yes;however the histogram is not symmetrical, this may opt out this condition Constant Variability => The variablility is not constant, where there are more residuals on the rightside of the scatter plot I don't think that this model can fit the regression model due to lack of variance and unsymmetrical residuals

---

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.

The relationship seems to be linear

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

It should remain the same; however the points will be more squished to the left

---

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.

$\widehat{x} = 107.2$
$\bar{y} = 171.14$
$s_{\bar{x}} = 10.37$
$s_{\bar{y}} = 9.41$
$R = 0.67$ $b1 = \frac{s_{\bar{y}}}{s_{\bar{x}}} * R$ $b1 = \frac{9.41}{10.37} * 0.67$ $(y - \bar{y}) = b1(x - \bar{x})$ $(y - 171.14) = 6.9479 * (x - 107.2)$

The equation of regression will be:

$y = 105.965 + 0.607 * x\ height = 105.965 + 0.607 * shoulder\ \ girth$

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

The slope is: 0.607 means that to increase height, the shoulder length should be increased by this slope which is 0.607 The intercept: for a shoulder girth of 0 cm, the average height increases by 105.965 cm

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

$R^2 = 0.67^2$
$R^2 = 0.4489$

44.89% of the variation 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.

$height = 105.965 + 0.607 * 100\ height = 166.762$

The height should be 166.76

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

$e_i = y_i - \bar{y}$
$e_i = 160 - 166.76$
$e_i = -6.76$

This means by a negative sign that the model overestimated the height by 6.76

(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?

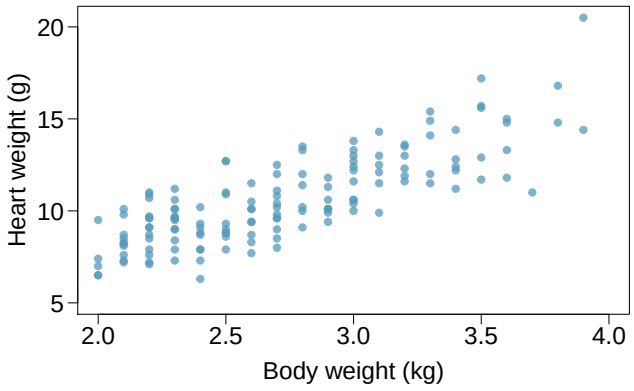
the average shoulder girth for this sample is 107.2 which means that 56 point will be far away. So I think that we cannot use this model to predict such prediction

---

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.

*heart_weight = -0.357 + 4.034 * body_weight*

(b) Interpret the intercept.

Intercept means that for body weight 0, the average heart weight is -0.357 grams

(c) Interpret the slope.

The slope means that to increase heart_weight , the body weight should be increased by 4.035

(d) Interpret R^2.

64% R-squared means that the variablility in heart weight of cats can be explained by body_weight

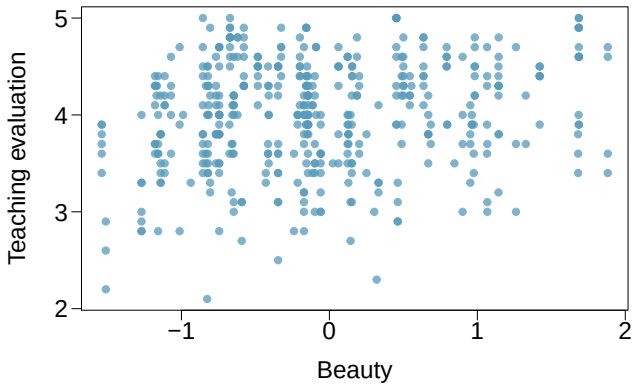
(e) Calculate the correlation coefficient.

```
R <- sqrt(0.6466)
R
```

```
## [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.

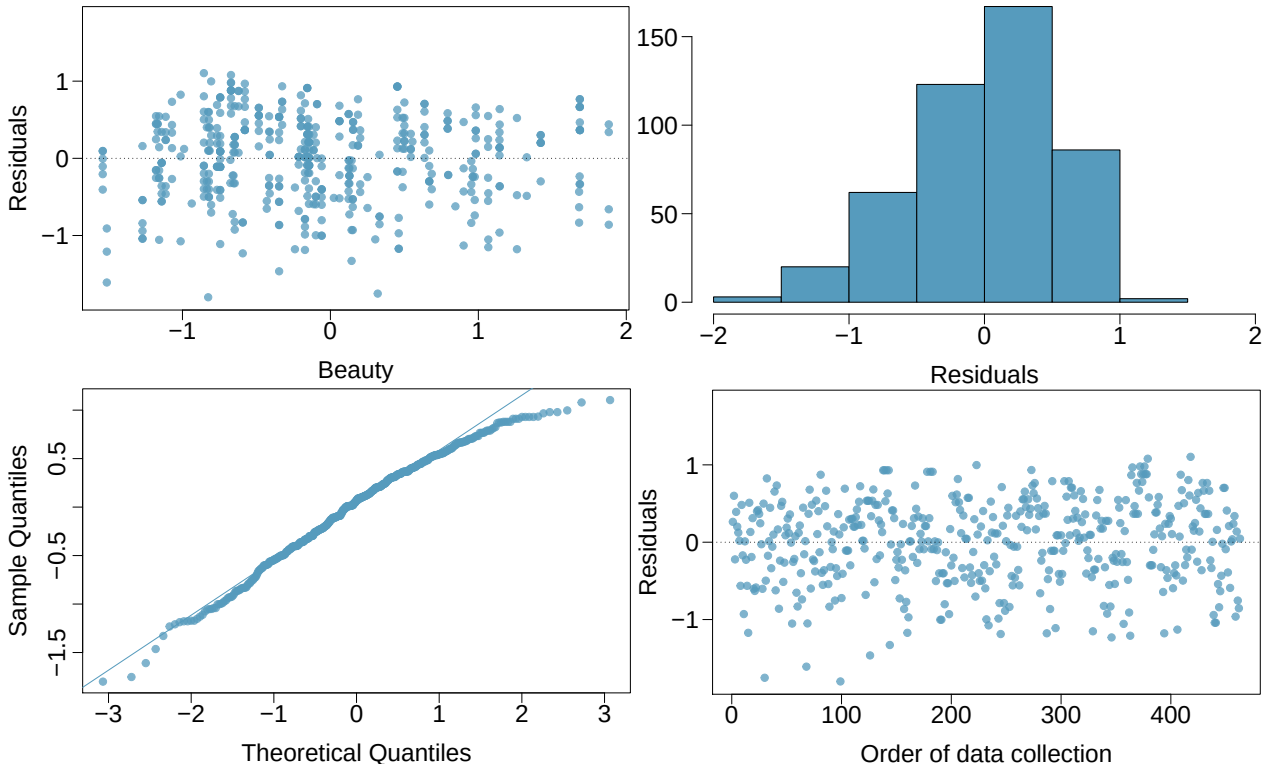
considered that both x-hat and y-hat both located on the regression line, we can have the following equation of the regression line

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

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

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

Since the slope is positive the relationship is positive. If we set up a hypothesis test with $H_0 : \beta_1 = 0$ *and* $H_A : \beta_1 > 0$*, then based on the summary table the p−value is nearly 0. And this is for a two-sided test, so it'll be even closer to 0 for a one-sided test. We reject the null hypothesis. There is convincing evidence that the relationship between teaching evluation and beauty is positive.*



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

Linearity: Based on the scatterplot, there may be a weak linear relationship. There is no evident pattern in the residual plot. Nearly normal residuals: The histogram of the residuals exhibits a left skew. Additionally, the points seem to move away from the normal probability line on each end. However, the bulk of the data is very close to the line. I would conclude that the distribution of residuals is nearly normal. Constant variability: Based on residual plot, there appears to be constant variability in the data. Independent observations: Observations are not a time series, and can be assumed to be independent (unless there is evidence that students copied each other's evaluations). I believe all conditions are satisfied for this linear model.