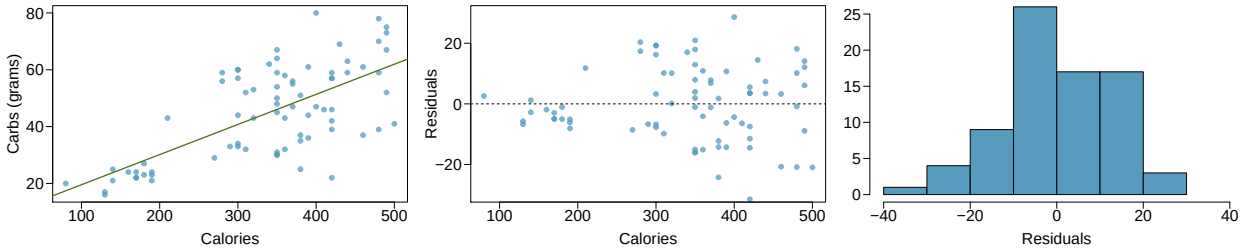
Chapter 8 - Introduction to Linear Regression

Jai Jeffryes

10/31/2019

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.

A weak, positive linear relationship, stronger in the domain of calories less than 300.

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

Explantory: Calories.

Response: Carbohydrates.

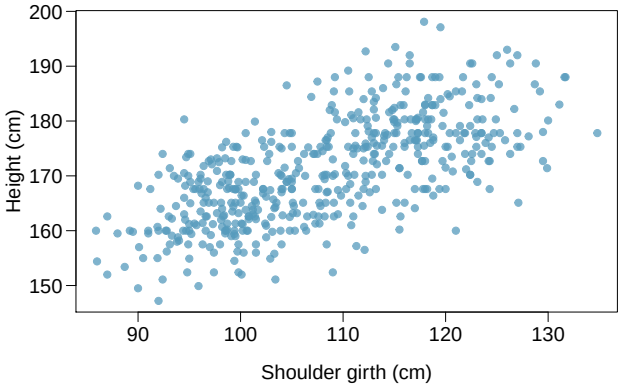
(c) Why might we want to fit a regression line to these data?

If one were restricting consumption of carbohydrates, it might be useful to predict carbohydrates on the Starbucks menu even though Starbucks does not post that nutrion data.

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

There are problems in this fit, with higher variability in the higher domain of x, calories.

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.

It is strongly, and postively linear.

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

None. Axis scale has nothing to do with it unless the linearity of the scale itself changed, like going to a logarithmic scale.

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.

$\widehat{height} = 171.14 + 0.67 * shoulder\_girth$

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

The intercept is the mean of the dependent variable, *height*, predicted if *shoulder_girth* measured 0 and that measurement were valid, which it is not however. The slope is the predicted increase in *height* given 1 unit of increase in *shoulder_girth*.

(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: 0.4489. This figure is the proportion of the height explained by variability in *shouder_girth*. (I confess I'm parroting the book in this. I don't really understand what that means.)

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

```
h <- 171.14 + 0.67 * 100
```

$height_{pred}$: 238.14 cm.

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

Residual: 78.14. We overestimated the height by this amount.

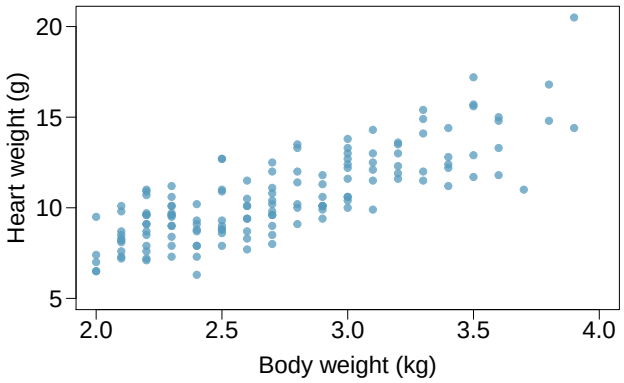
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

No. The population is made up of adults. Children are undeveloped.

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.

$\widehat{heart\_weight}$ = -0.357 + 4.034 * *body_weight*

(b) Interpret the intercept.

The intercept is -0.357. It is meaningless in this context, since an x value of 0 would represent an overall weightless cat, and its heart would be a negative weight.

(c) Interpret the slope.

Out of each kilogram of body weight, 4.034 grams of that is heart weight, on average.

(d) Interpret R^2.

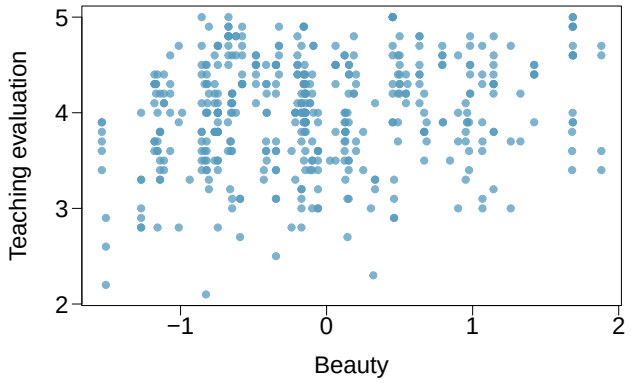
Body weight accounts for 64.66% of the variability in heart weight.

(e) Calculate the correlation coefficient.

Root of R^2: 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.

We can figure out the slope based on the calculation of the known T value.

$T = \frac{estimate - null\ value}{SE}$

$4.13 = \frac{estimate - 0}{0.0322}$

$estimate = 4.13 * 0.0322$

$estimate = 0.132986$

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

Yes, because the p value is reported as 0.0000, within our precision. This is an observational experiment, though, rather than a random controlled trial. It is suggestive, but requires more experimentation to establish causality.

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

(1). Linearity. The first plot appears linear.

(2). Nearly normal residuals. The first plot displays no prominant outliers. The histogram is bell-shaped but is pretty skewed to the left. The residuals could stand to be more normal. However, there are 463 observations, so I judge it to be okay.

(3). Constant variability. Plot 1 appears to show constant variability.

(4). Independent observations. In the fourth plot, for order of data collection, I think my eyes see a series of waves. I think these data lack independence.