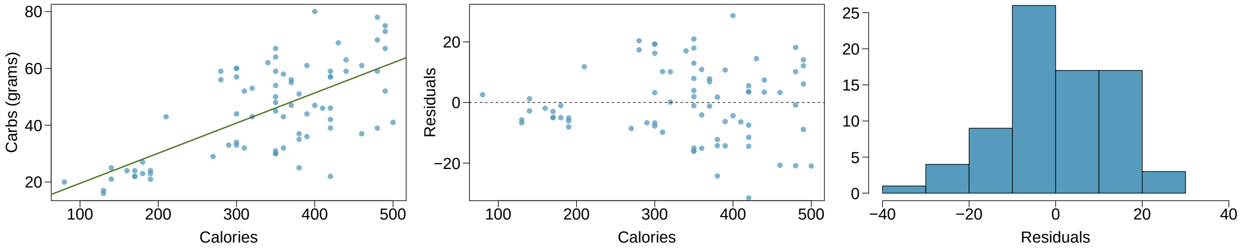
Chapter 8 - Introduction to Linear Regression

Adam Gersowitz

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.

There is a positive linear relationship between caloris and carbohydrates.

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

The explanatory variable is calories and the response variable is carbs.

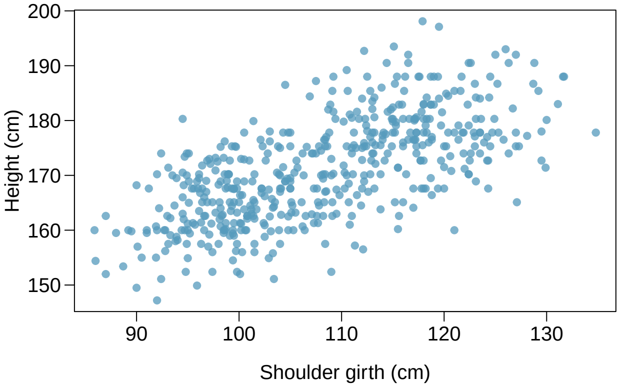
(c) Why might we want to fit a regression line to these data?

We would want to fit a regression line to better predict an estimate of carbs base on calories.

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

The data is linear with nearly normal residuals. However, the variability does not appear to be constant as the variability is much larger as the number of calories increases.

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.

There is a strong positive 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?

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.

y = a+bx

b=.67*(9.41/10.37) b=0.608

a=y-bx a=171.14-0.608*107.2 a=105.96

y=105.96+0.608x

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

The slope is positive and for each increase in shoulder girth there is an increase of 0.608 in height. 105.96 is the height when shoulder girth is 0.

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

0.67ˆ2=.4489

The Rˆ2 is .4489. This means 44.89% of variability in height can be predicted 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.

y=105.96+0.608*100

166.76 cm

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

R=160-166.76=-6.76

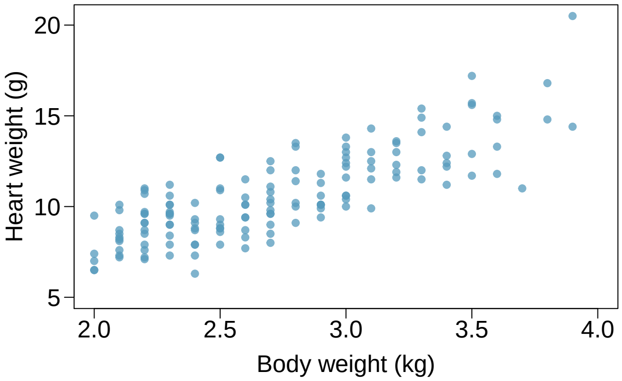
The residual is -6.76 which shows the model overestimated the students height by 6.76 cm.

(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 child would be much smaller than anyone in the original sample. Additoinally if the original sample was all adults the musculature of their shoulders will be much different than a child which will cause the prediciton of height to be inaccurate.

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.

y=-0.357 +4.034x

(b) Interpret the intercept.

-0.357 g is the predicted heart weight when the body weight = 0

(c) Interpret the slope.

For each additional kg of body weight hear weight is predicted to increase 4.034 g

(d) Interpret R^2.

64.66% of the variability in hear weight can be explained by body weight.

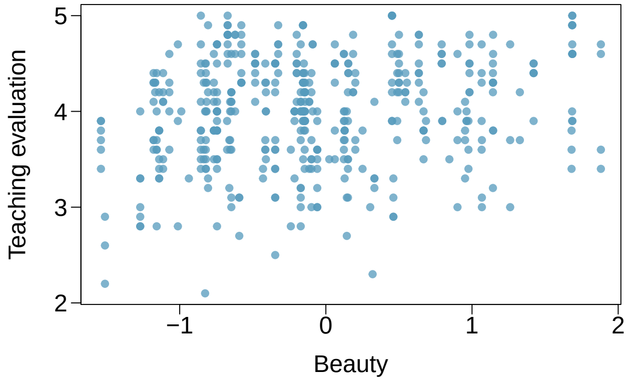
(e) Calculate the correlation coefficient.

sqrt(.6466)=.804

The correlation coefficient is 0.804

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.

y=a+bx b=(y-a)/x

b=(3.9983-4.01)/-0.0883 b=0.1325

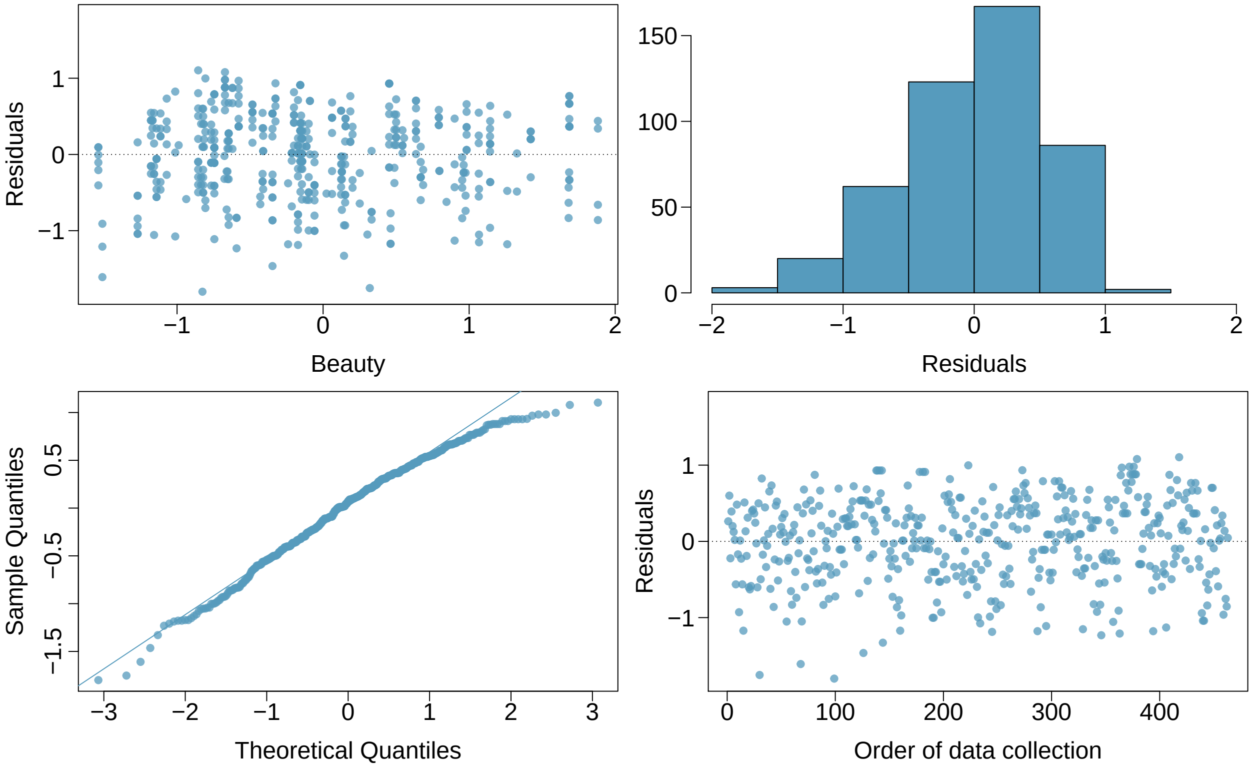
(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

Yes, the slope above is positive which indicates teh relationship is postive although somewhat weak.

(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, constant variability, and nearly normal residuals

Yes, judging by the residual graphs below you can see the relationship is linear and has contstant variabilityas well as nearly normal residuals.


Residuals
1
0
−1
−1
0
1
2
Beauty
150
100
50
0
−2
−1
0
1
2
Residuals
Sample Quantiles
0.5
−0.5
−1.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Residuals
1
0
−1
0
100
200
300
400
Order of data collection