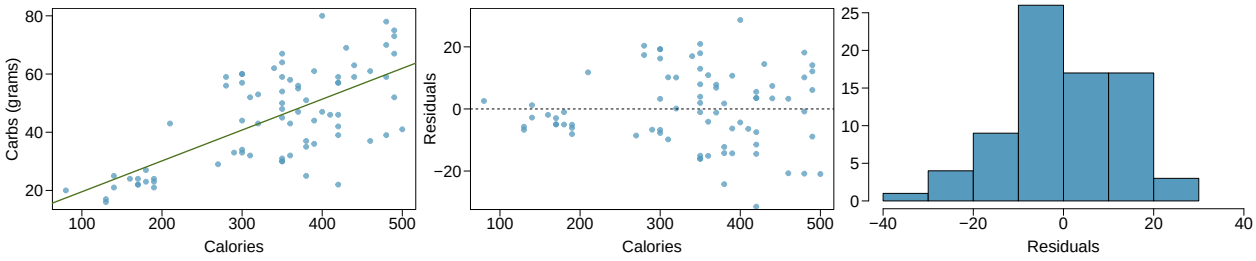
Chapter 8 - Introduction to Linear Regression

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.

There does seem to be a positive corelation between calories and carbs in starbucks menu items. The residual plot appears to be reasonably evenly distributed around zero, which is a good indicator of linear relationship.

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

Calories are the explanatory variable and carbs are the response.

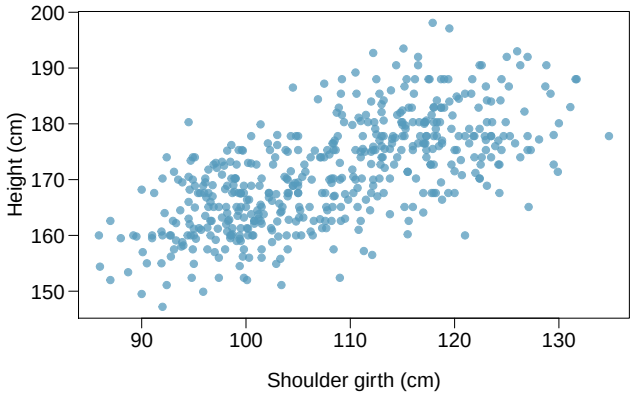
(c) Why might we want to fit a regression line to these data?

We are looking to find some way of predicting the carb content of a menu item that we do not know directly. Regression, when meaningful and well-fit can do that task well.

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

These data are independent, the residuals appear to be nearly normal, there is some linearity, but the constant variance seems dicey. It could be that the variability is increases with the explanatory variable, but it's not for sure given the residual distribution.

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.

This appear to have positive linear relationship. There do not seem to be outliers and there is positive trend.

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

The relationship would look the same, but the slope of the regression line would have an extra factor of 2.54 cm/in. By comparison, the inch measurement would be more steep.

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.

$y = b_1 \times girth + b_0$, which becomes the following given the above values.

$y = \frac{10.36}{9.41}(0.67) \times girth + (171.14 - (\frac{10.36}{9.41}(0.67)(106.20))$. We can evaluate this with R.

```
b_1<-((10.36)/9.41)*0.67
b_0<-171.14-(b_1*106.20)
paste0("y=",b_1,"*girth +",b_0)
```

```
## [1] "y=0.737640807651435*girth +92.8025462274176"
```

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

The intercept tells the experimental minimum shoulder girth is 92.80cm. The slope is the rate of change of height as a function of 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 square is simply correlation squared. $R^2 = 0.67^2 = .45$. This means that only 45% of the variance in the data can be explained by the model.

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

```
pred_height<-b_1*100 + b_0
paste0("predicted height= ",pred_height, "cm")
```

```
## [1] "predicted height= 166.566626992561cm"
```

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

```
res<- 160-pred_height
paste("residual=",res)
```

```
## [1] "residual= -6.56662699256111"
```

There is a 6.56cm difference between the predicted height and the actual height of the student. The negative value is the result of the student being shorter than predicted.

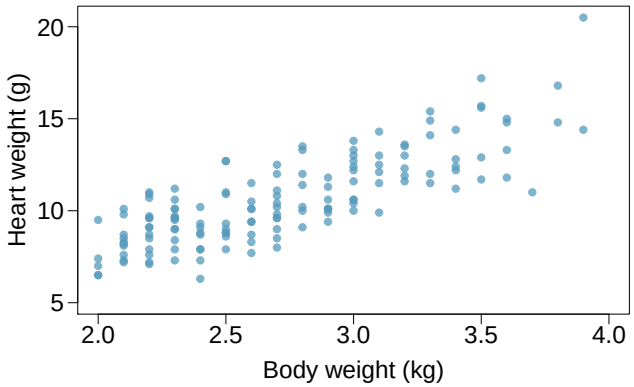
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

No. Given the R^2 value of 0.45 and the fact that a one-year old is far outside of the sample age range and that the models y-intercept indicates the lower bound for predicting heights is 92cm it would be very unwise to predict this childs height.

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.

**The linear model is $Heart = 4.034 \times body - 0.357$

(b) Interpret the intercept.

The intercept is very close to zero, as it should be. If a cat has zero body weight we should expect zero heart weight as well.

(c) Interpret the slope.

Slope is the change in heart weight as a function of body weight. The heart weight increases by 4.034g for every unit kg body weight increase.

(d) Interpret R^2.

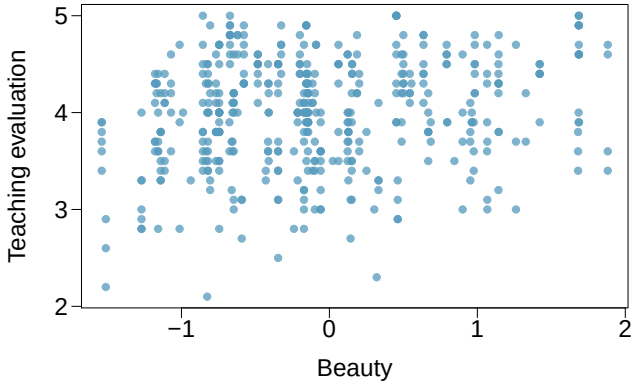
About 64% of the variance in the data can be explained by this simple model.

(e) Calculate the correlation coefficient.

The correlation is 0.804, $R = \sqrt{R^2}$

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.

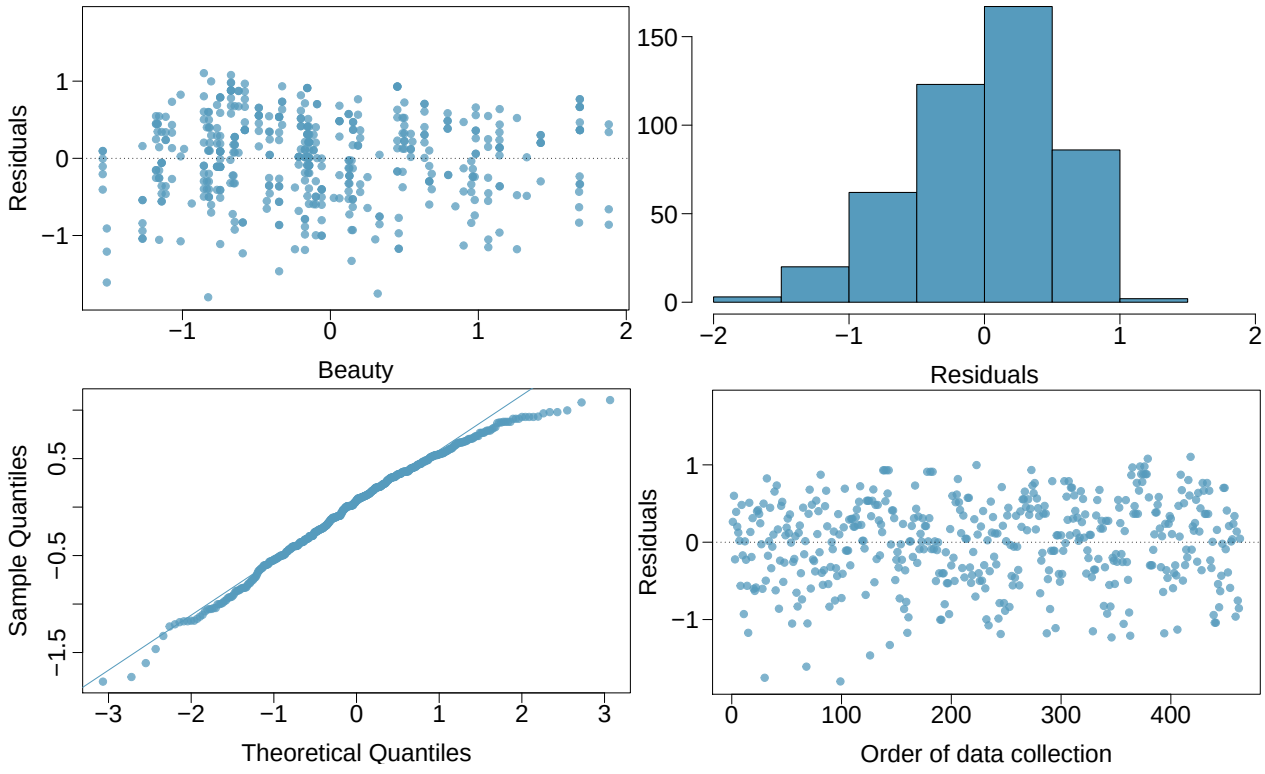
**Given $y = mx + b$ we can use the mean values to determine the slope. $3.9983 = slope \times (-0.0883) + 4.010$. Such that, $slope = 0.1325$

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

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

Ansering b & c together

Tests all seem to point to yes, there is sufficient evidence to suggest that there is a relationship between teaching evals and beauty.The p-value from the T-Score is very low and the residuals look to have constant variance and are distributed around zero. The qq plot suggest that the data is reasonably normal. They only assumption I am skeptical about is Linearity. The data do not seem to be linear even thought we can reject the null hypothesis that the slope is zero. Also the y-intercept is ~4, which means that at the lowest beauty score, the minimum teacher evaluation is around 4/5. It still doesn't seem to be a linear relationship even if the slope is non-zero.


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