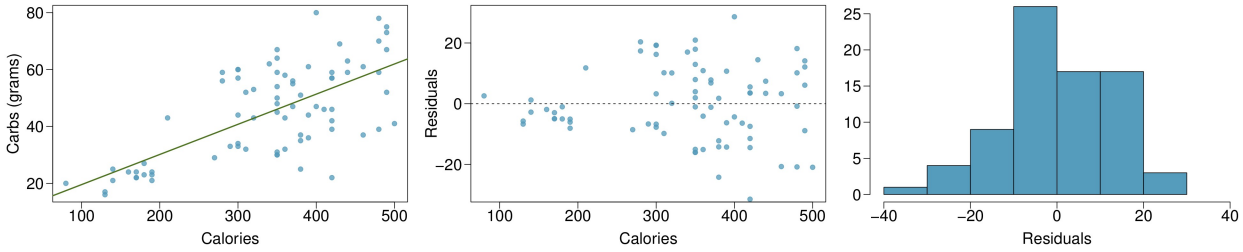
Homework 8 : Introduction to Linear Regression

Ramnivas Singh

05/02/2021

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.

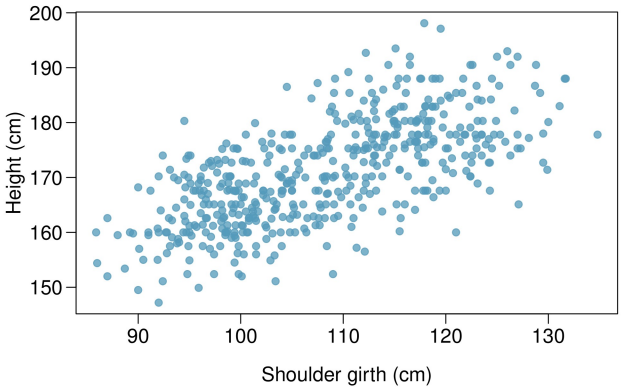


(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.
(b) In this scenario, what are the explanatory and response variables?
(c) Why might we want to fit a regression line to these data?
(d) Do these data meet the conditions required for fitting a least squares line?

Answer: (a) positive linear relationship, the more calories, the more carbs. (b) The explanatory variable is the calories, and the responce variable is the carbs. (c) So we can predict the amount of carbs one would expect for an item by it's amount of calories. (d) Pretty much. This may not be so much a sample as it is the entire population. The only concern is that residuals increase with calories count

---

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

Answer: (a) It appears linear, and positively correlated. (b) Changing the units doesn't change the relationship. It might change the slope, but not relationship between variables.

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?

Answer: (a)

```
R <- 0.67
sd_y <- 9.41
sd_x <- 10.37
b1 <- R * sd_y / sd_x
x <- 107.20
y <- 171.14
b0 <- y - b1 * x
```

$$\hat{height} = 105.9650878 + 0.6079749 * girth$$

(b)
Slope: for each 1 cm increase in girth we expect 0.608 cm increase in height.
Intercept: Theoretically if girth was 0, we'd have a base height of 106cm. This is meaningless - our data points would stop well short of a shoulder girth of 0.

(c) $R^2 = 0.4489$ - means that this percent of variablility in height is explained by girth.

(d)

```
height <- b0 + 100 * b1
```

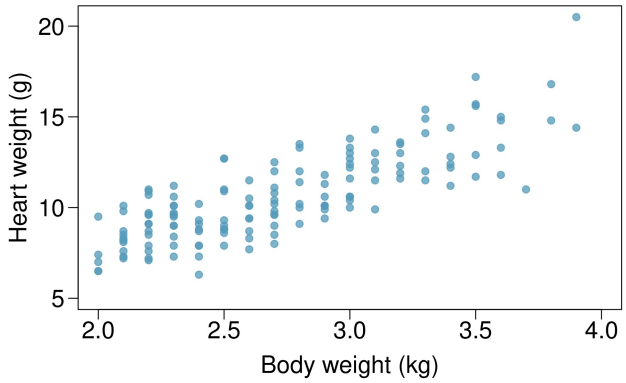
166.7625805 cm

(e) Residual: -6.7625805; this means that we overestimated the height with the model and the student is actuall a little shorter than expected.
(f) No - that would be extrapolating beyond the bounds of our dataset.

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.

Answer:

(a)

```
summary(m_cats_hwt_bwt)
```

```
##
## Call:
## lm(formula = cats$Hwt ~ cats$Bwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## cats$Bwt      4.0341     0.2503  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

$$\hat{heart\_weight} = -0.3567 + 4.0341 * body\_weight$$

(b) You cannot since *body_weight = 0* has no meaning. We'd be extrapolating beyond the available data points. The intercept merely gives us a basis point to help draw a linear line thru the available data points.

(c) Interpret the slope. for each 1 kg increase in body weight, we'd estimate an increase of 4.0341 g in heart weight.

(d) Interpret R^2.

Body weight explains 64.41% of the variation in heart weight. The remaining ~35% of variation might be due to other causes (unknown) or just random variation. We don't know.

(e)

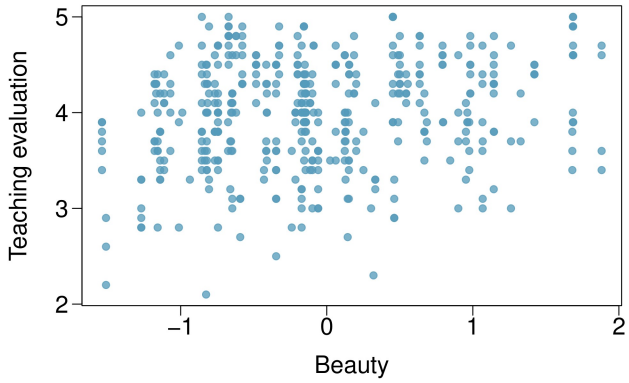
```
c <- cor(cats$Hwt,cats$Bwt)
sqrt(0.6466)
```

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

THe correlation coefficient is 0.8041274

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.

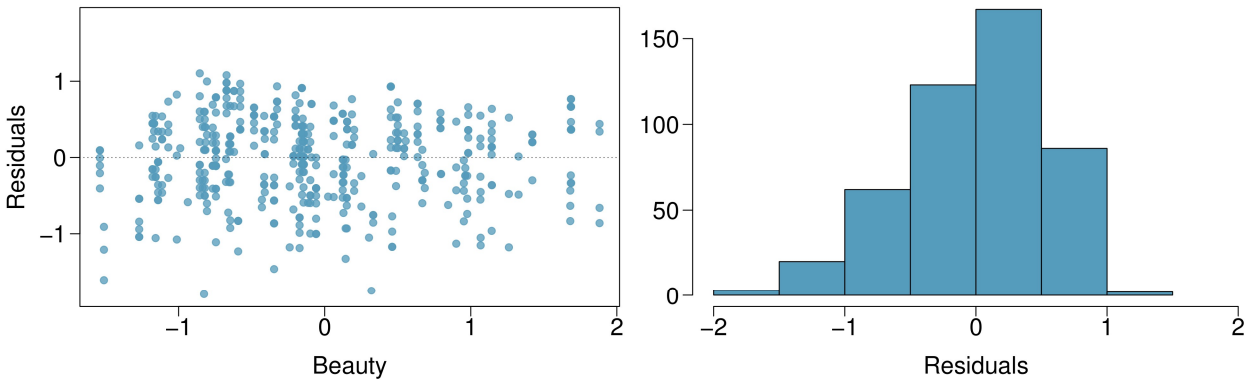
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000

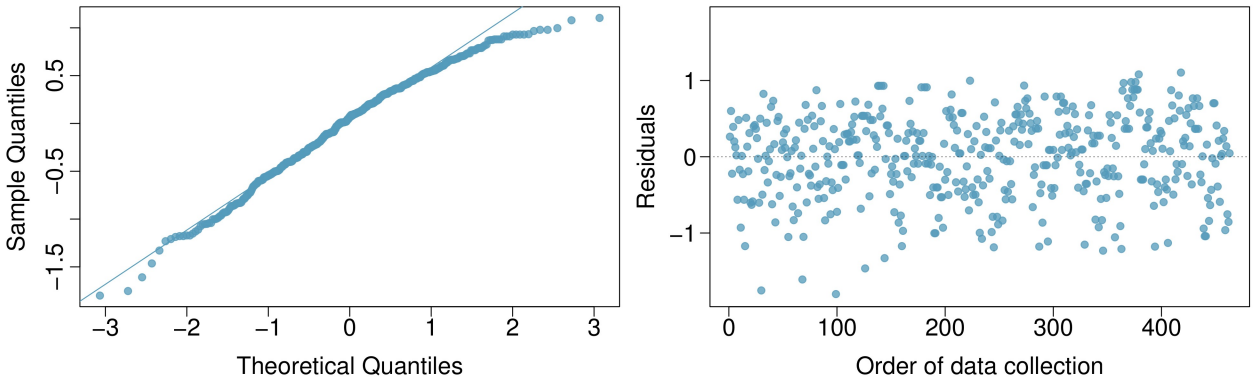


(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(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.





Answer:

(a)

```
xmean = -0.0883
ymean = 3.9983
intercept = 4.010

#intercept = ymean - slope*xmean
slope = (ymean - intercept) / xmean
print(paste("slope = ",round(slope,4)))
```

```
## [1] "slope =  0.1325"
```

```
summary(m_eval_beauty)
```

```
##
## Call:
## lm(formula = eval ~ beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
## beauty       0.13300    0.03218   4.133 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 461 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03364
## F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

(b) While there does appear to be a significant trend with beauty as a predictor of evaluation (p=0.00004247 is way blow 0.05), this trend explains very little of the variation we see in the data. Beauty only accounts for ~3.364% of the observed variation. This would suggest that while beauty is a factor, it's a minor one at best.

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