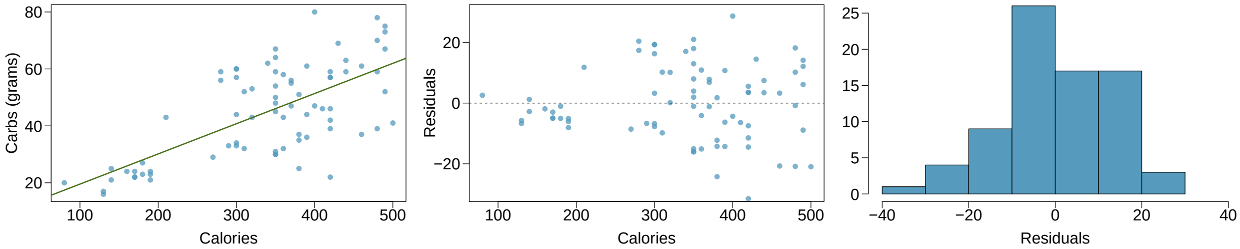
Chapter 8 - Introduction to Linear Regression

Samuel I Kigamba

11/10/2019

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.

The relationship between the number of carolies and the amount of carbohydrtes (in grams) that starbucks food menu items contain is linear as evidenced by graph above.

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

Carolies are the explanatory variable and Carbs the response variable.

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

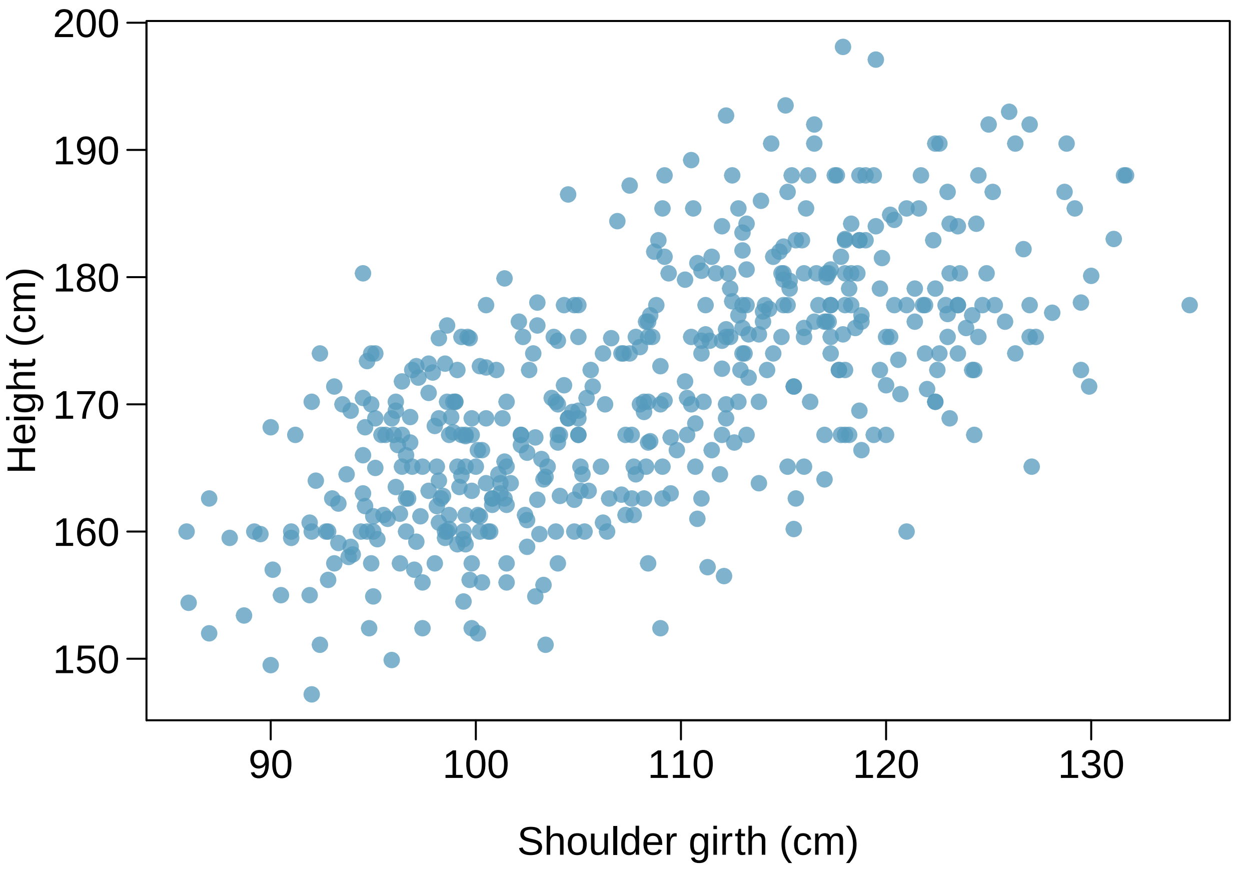
We could fit a regression line to these data to see if we could find a relationship between the two variables.

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

a. Nearly Normal Residuals: The histogram of residual seems normal.
b. Independent observations: We asssume the data to be independent.
c. Linearlity: The scatter plot shows a linear relationship between carbs and carolies, and
d. Constant Variability: There seem to be constant variability amont the variables.

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.

```r
# plot height vs. shoulder girth ------------------------------------
par(mar = c(3.8, 3.8, 0.5, 0.5), las = 1, mgp = c(2.7, 0.7, 0),
    cex.lab = 1.25, cex.axis = 1.25)
plot(bdims$hgt ~ bdims$sho.gi,
     xlab = "Shoulder girth (cm)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2])
```



(a) Describe the relationship between shoulder girth and height. 190

```
There is a positive relationship between between the height and shoulder
girth as evidenced in the graph above.
```

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

```
The positive relationship would be remain but the slope would steepen to
account for the larger change in height for each inch of shoulder.
```

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.

$$y\hat{} = „_0 + „_1 * x$$

$$„_1 = S_y/S_x * R$$

```
s_y <- 9.41
s_x <- 10.37
r <- 0.67
y_mn <- 171.14
x_mn <- 107.20
b1 <- (s_y / s_x ) *r
b0 <- y_mn - (b1*x_mn)
cat(b0, b1)
```

```
## 105.9651 0.6079749
```

Equation of regression line is:

$$y\hat{} = 105.9651 + 0.6079749 * x$$

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

```
The slope represents an increase of 0.6079 in height for every cm of shoulder.
```

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

```
r_sqr <- r^2
r_sqr
```

```
## [1] 0.4489
```

```
The linear model explains approximately 44.89% variation in height.
```

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

$$y\hat{} = 105.9651 + 0.6079749 * x$$

```
rand_std_ht <- 105.9651 + 0.6079749*100
rand_std_ht
```

```
## [1] 166.7626
```

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

```
res <- 160 - rand_std_ht
res
```

```
## [1] -6.76259
```

```
The negativity is an indication that the model overestimates the height data.
```

(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 shoulder girth of 56cm is outside of the model paramenters and might not be
appropriate for estimatiom of the childs height. This would represent an outlier in the data set.
```

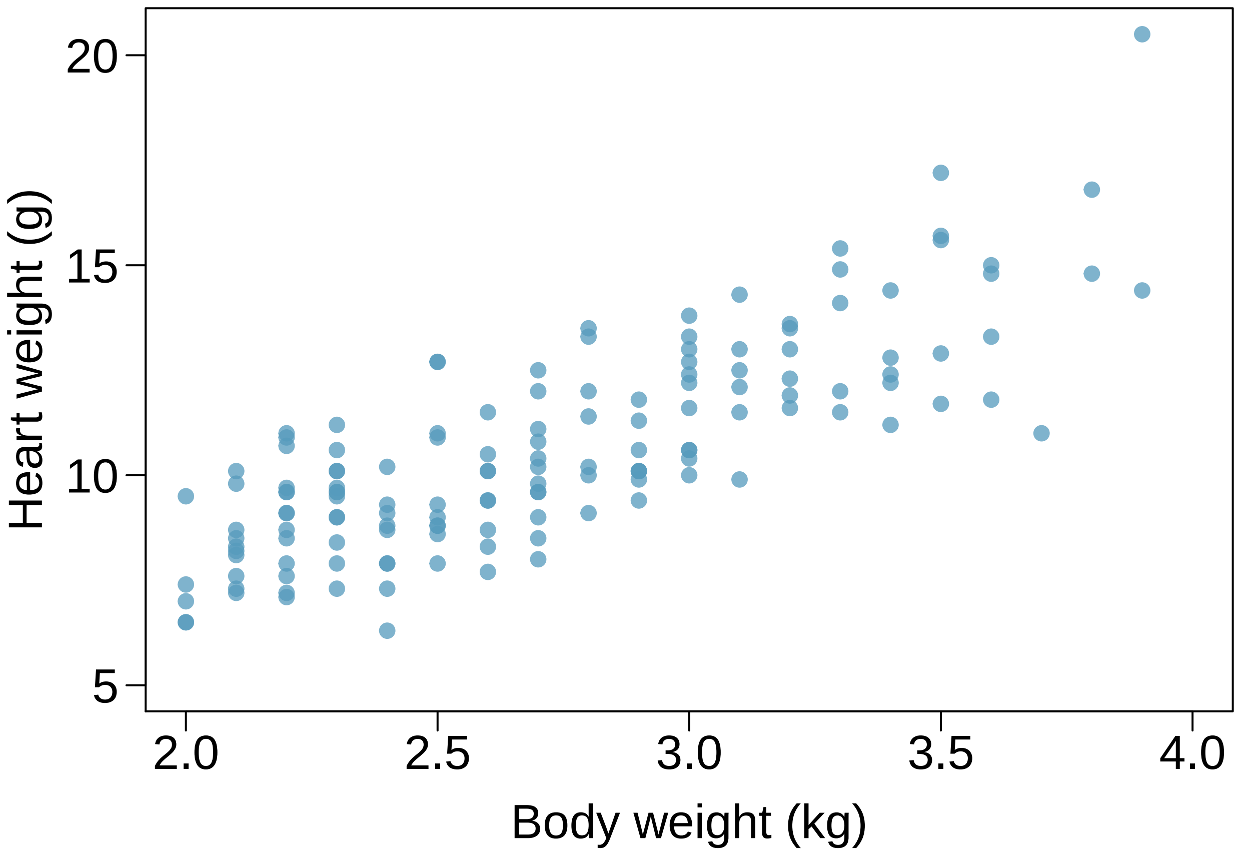
Cats, Part I. (8.26, p. 327) The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

	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
$s = 1.452$	$R^2 = 64.66\%$		$R^2_{adj} = 64.41\%$	

```r
# model heart weight vs. weight --------------------------------------
m_cats_hwt_bwt <- lm(cats$Hwt ~ cats$Bwt)
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
```

```r
# plot heart weight vs. weight ---------------------------------------
par(mar = c(3.7, 3.7, 0.5, 0.5), las = 1, mgp = c(2.5, 0.7, 0),
    cex.lab = 1.5, cex.axis = 1.5)
plot(cats$Hwt ~ cats$Bwt,
     xlab = "Body weight (kg)", ylab = "Heart weight (g)",
     pch = 19, col = COL[1,2],
     xlim = c(2,4), ylim = c(5, 20.5), axes = FALSE)
axis(1, at = seq(2, 4, 0.5))
axis(2, at = seq(5, 20, 5))
box()
```



(a) Write out the linear model.

$$y\hat{} = -0.3567 + 4.0341 * x$$

(b) Interpret the intercept.

`The intercept of -0.3567 indicates the heart weight if body was zero, however the negative weight interpretetion would be misleading, thus this is just the anchor to the regressionline.`

(c) Interpret the slope.

`The slope of 4.0341 indicates an increase of 4.0341 units of heart weight for every one unit increa`

(d) Interpret R^2.

`The R-squared of 64.41% of height is the variability that can be explained by body weight.`

(e) Calculate the correlation coefficient.

```
rsqr1 <- 0.6466
corr1 <- sqrt(rsqr1 )

corr1
```

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

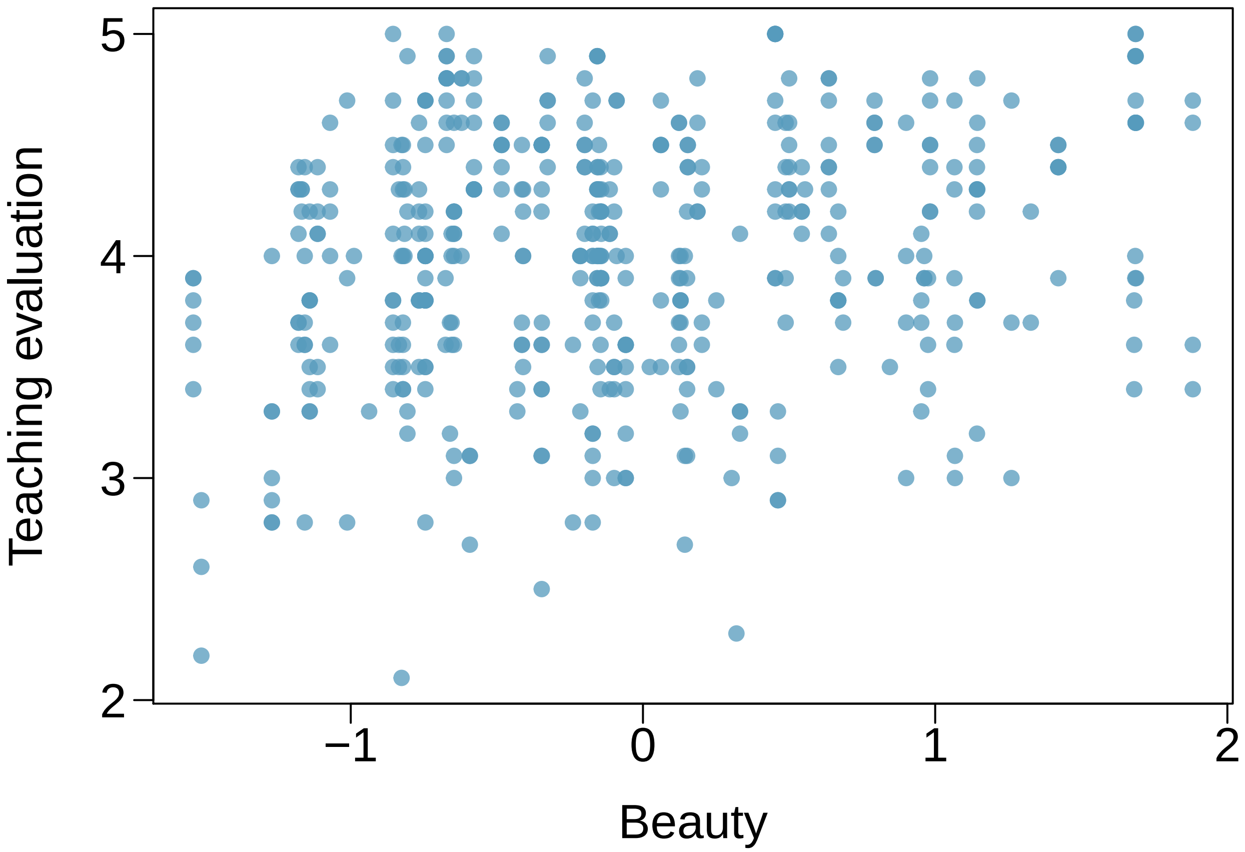
Rate my professor. (8.44, p. 340) Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures 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

```r
# model evaluation scores vs. beauty --------------------------------
m_eval_beauty = lm(eval ~ beauty)
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
```

```r
# scatterplot of evaluation scores vs. beauty -----------------------
par(mar = c(3.9, 3.9, 0.5, 0.5), las = 0, mgp = c(2.7, 0.7, 0),
    cex.lab = 1.5, cex.axis = 1.5, las = 1)
plot(eval ~ beauty,
     xlab = "Beauty", ylab = "Teaching evaluation",
     pch = 19, col = COL[1,2],
     axes = FALSE)
axis(1, at = seq(-1, 2, 1))
axis(2, at = seq(2, 5, 1))
box()
```



(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.

$$y\hat{} = „0 + „1 * x$$

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

```
## [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.

$$„_1 = S_y/S_x * R$$

```
Since both Sy and Sx are positive the data slope is positive as well.
```

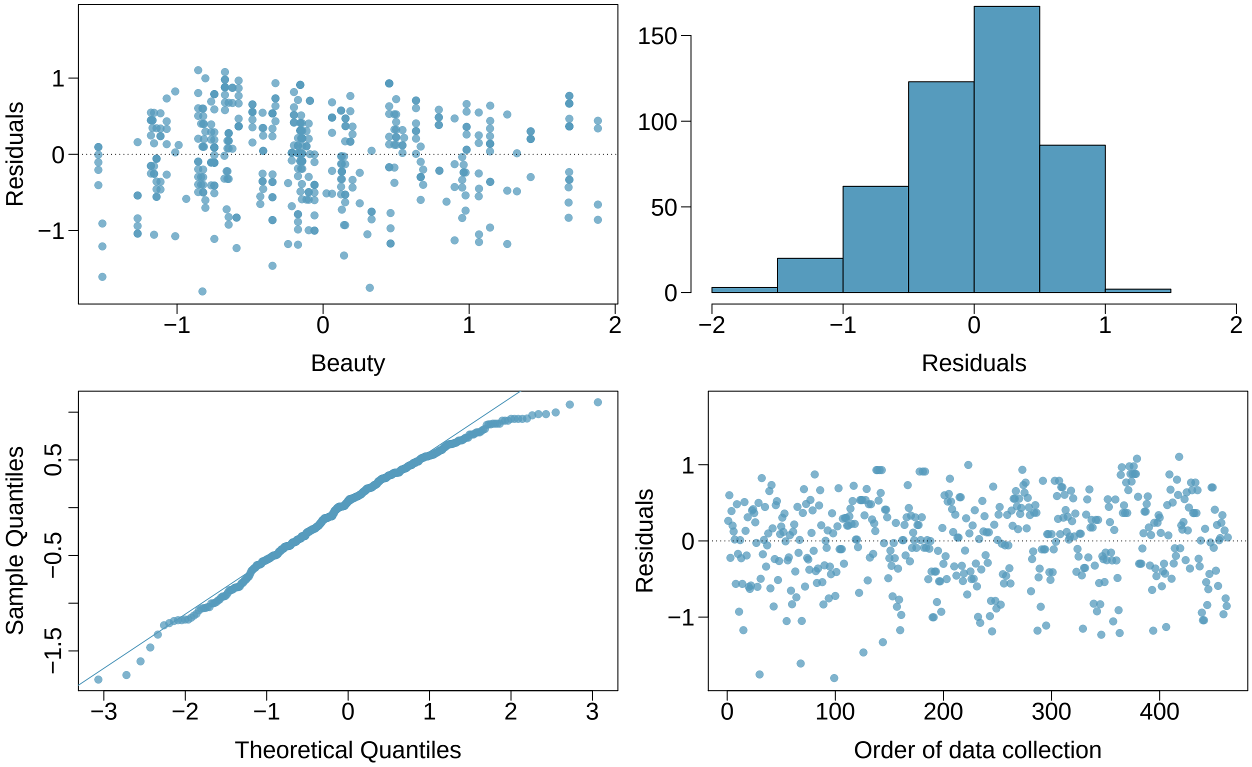
(c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.

Constant variability: The residual scatterplot indicates constant variability.

Independent observations: We assume data is independent in the sample.

Nearly normal residuals: The residuals histogram shows the data as nearly normally distributed.

Linearity: The scatter plots show the relationship between beauty and teaching evaluation as linear


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