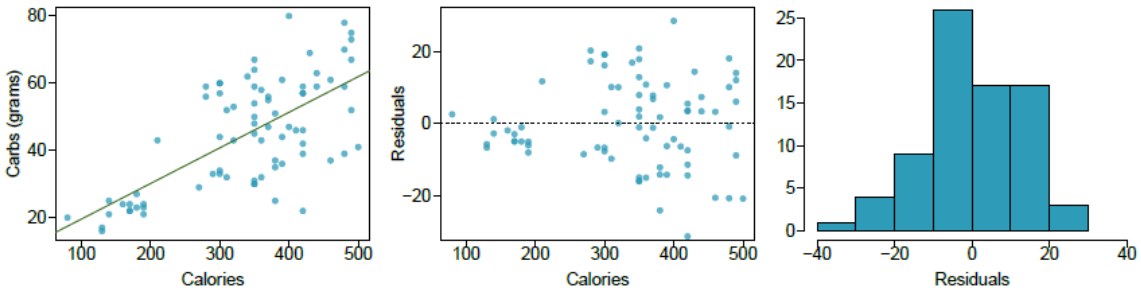
Chapter 7 HW

Kishore Prasad

7.24 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.[21] 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?

(a) From the scatter plot it is evident that the relationship between number of calories and amount of carbohydrates is positive. As amount of carbohydrates increases the calories also increases and vice versa.

(b) Calories is the Explanatory variable and Carbohydrates is the response variable.

(c) Given any food item, we may want to predict the amount of carbohydrates based on the number of calories. To do this, we can make use of the fitted regression line.

(d)

Linearity: There is a linear trend available. However, the scatter plot indicates that the data is widely dispersed.

Nearly normal residuals: There is a slight left skew in the residuals but we can consider this to be normal.

Constant variability: The residuals scatter plot does not show a constant variability. The values are nearer to the residuals line at the lower end of the scale and are quite scattered towards the higher end. So I would not consider this to be constant variability.

Independent observations: The items are presumed to be independent of each other since the ingredients and process for manufacture might be different for each item.

I would conclude that the data meets the requirements for fitting a least squares line.

7.26 Body measurements, Part III. Exercise 7.15 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?

(a) The equation for the regression line is given by the formula:

$$\hat{y} = \beta 0 + \beta 1 * x$$

Calculating $\beta 0$ and $\beta 1$ as below:

```
beta_1 <- (9.41 / 10.37) * 0.67
beta_0 <- (beta_1 * (0 - 107.2)) + 171.14
```

We get the regression equation as :

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

(b) The intercept of 105.9650878 is the height when the sholder girth is 0. The Slope of 0.6079749 indicates that height increases by 0.6079749 cms for every 1 cms increase in girth.

(c) Computing R2, we get R2 = 0.4489. This means that the linear model explains about 45% of the variation in height data.

(d) Following is the calculation for predicting the height for a student of shoulder girth of 100 cms:

```
pred_height <- beta_0 + beta_1 * 100
```

The predicted height is 166.76 cms.

(e) Calculating the residual as below:

```
actual_height <- 160

residual <- actual_height - pred_height
```

The residual of -6.7625805 indicates that we have over estimated the height

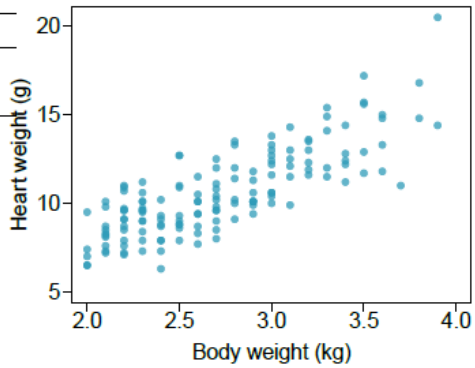
(f) From the original exercise (Ex 7.15), we see that the minimum shoulder girth was about 85 cms. Hence, the shoulder girth of 56 cms for a 1 year old does not fall within the original data that was used to generate the regression line. We would be extrapolating if we were to predict the height with this shoulder girth. Extrapolation is an unreliable estimate and should not be used.

7.30 Cats, Part I. 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.




```
beta_0 <- -0.357
beta_1 <-  4.034
```

(a) The following is the linear model:

$$\hat{heartwgt} = -0.357 + 4.034 * bodywgt$$

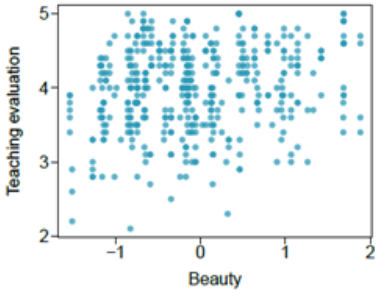
(b) The intercept of -0.357 tells us that for a cat with 0 KG body weight, the heart will weigh -0.357 gms.

(c) The slope of 4.034 tells us that the heart weight increases by 4.034 grams for every 1 kg increase in body weight.

(d) R^2 of 64.66 % tells us 64.66% of the variation in the heart weight is explained by the linear model.

(e) The correlation coefficient is the square root of R^2 and when calculated is : 0.8041144

7.40 Rate my professor. 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.[24] The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

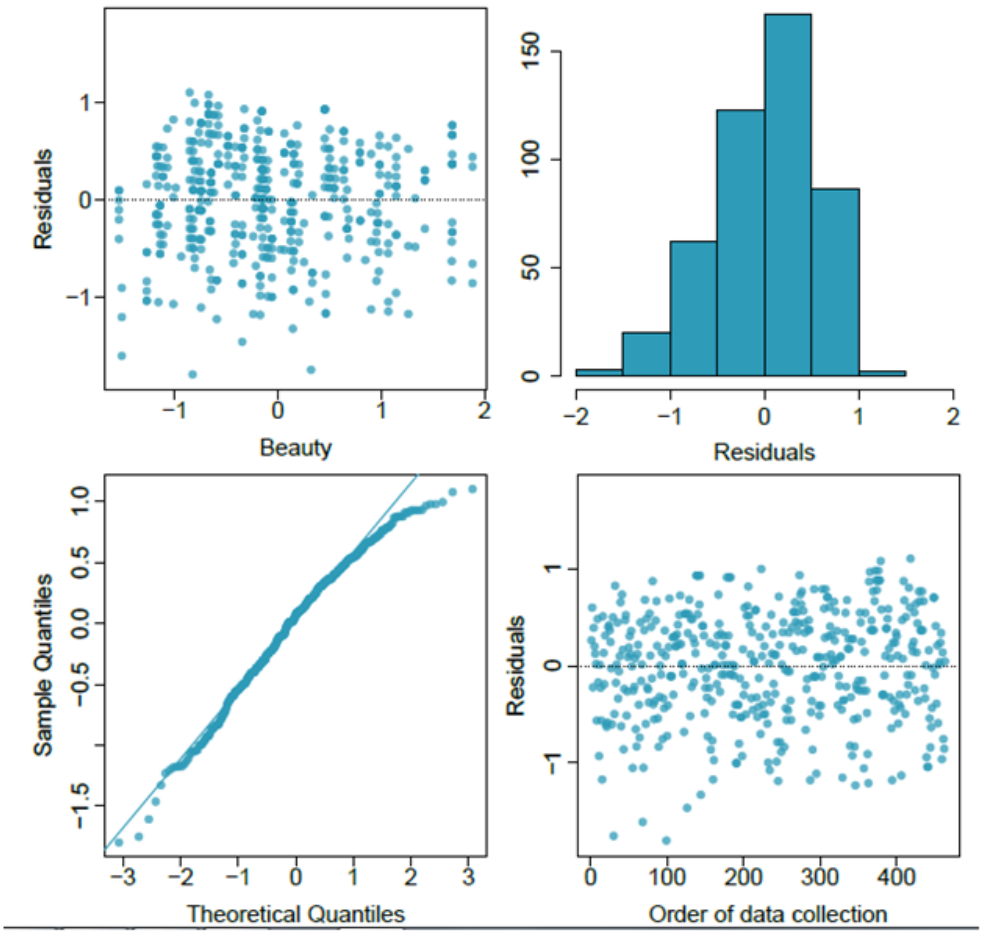


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000

(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.



(a) Below is the calculation for beta_1:

```
beta_1 <- (3.9983 - 4.010) / -0.0883
beta_1
```

```
## [1] 0.1325028
```

Slope is 0.1325028

(b) The p-value for the slope is 0.0000. This means that the slope is not 0 on a two tail test. Given that the t value is 4.13 and half the two-tail p-value is still 0.0000, it means a positive relationship between teaching evaluation and beauty.

(c)

Linearity: The data points seem to be clustered. We cannot confirm or deny the trend based on visual inspection.

Nearly normal residuals: There is a slight left skew in the histogram but we can consider this to be nearly normal.

Constant variability: Except for some outliers, the residuals scatter plot does show a constant variability.

Independent observations: Given that the number of professors (463 professors) are less than 10% of the population, I would assume these to be independent observations.