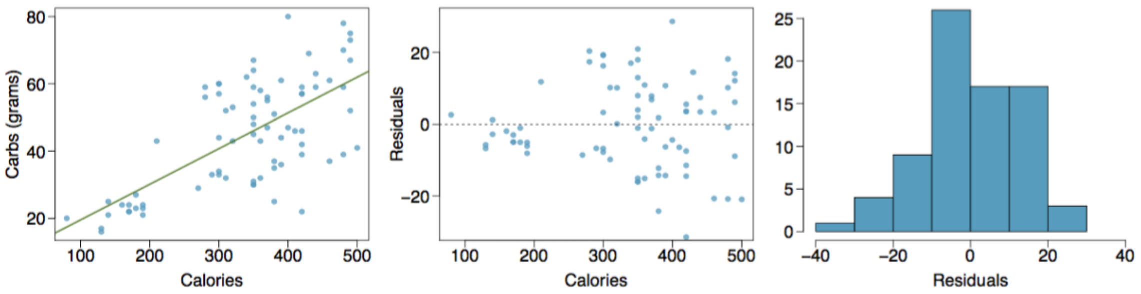
DATA-606
HW-7
Introduction to linear regression
Dhananjay Kumar

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. 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?

Solution:

a) The relationship is linear (weak positive co relation).

b) Explanatory variable: Calories
Response variable: Carbs

c) The relationship looks linear as seen from the first plot, the variance around the line is roughly constant with few outliers. Also this data is not from time-series. Therefore, Least squares regression can be applied to these data.

d) The conditions for independence are:

- Linear trend: The first plot shows the linear trend, so this condition is met.
- Residuals should be nearly normal: The histogram of the residuals is nearly normal so this condition is met.
- Constant variability: From the second graph the variance is roughly constant with few outliers.
- Observation independent: Since these observation are not from time series, we Assume they are independent.

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?

Solution:

a) Regression equation:
Height = b0 + b1 * shoulder girth

B1 = (9.41/10.37) * 0.67 = 0.61
The mean (107.20, 171.14) should lie on this line
To calculate intercept, b0
171.14 = b0 + 0.61*107.20

b0 = 105.75

b) Interpretation of slope: For one cm increase in shoulder girth there is 0.61cm increase in height.

c) R^2 = (0.67)^2 = 0.3721
This model can explain about 37 % of variation in height. This is basically describes the strength of this model.

d) Shoulder girth = 100
Height = 105.75 + 0.61*100
= 166.75

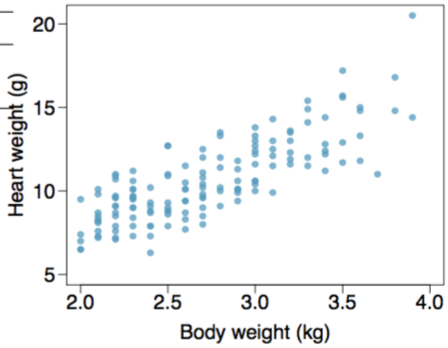
e) Actual height = 160 cm
Predicted height = 166.75
Residual = 160 – 166.75 = -6.75
The residual is negative, meaning the actual data point is below the regression line. The model overestimated the point and the error is large.

f) The data point 56cm seems out of this sample and so it will be difficult to use this model for predicting the height.

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.

Solution:

a) y = -0.357 + 4.034 * body weight
Where, y = heart weight
x = body weight

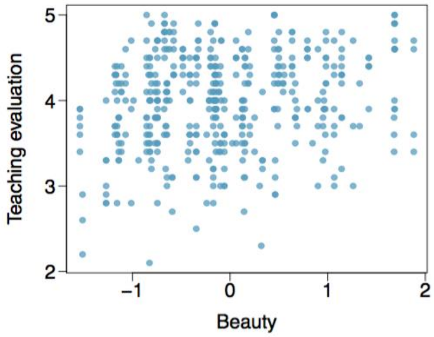
b) If the cat has 0kg body weight, it's heart would weight -0.357g

c) For every 1 kg increase in body weight of cat there is 4.034g increase in heart weight.

d) R^2 = 64.66% , this model can explain 64.66% of variation of heart weight in cats.

e) Correlation coefficient = sqrt(R^2) = sqrt(64.66) = 8.04% 0r 0.0804

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.

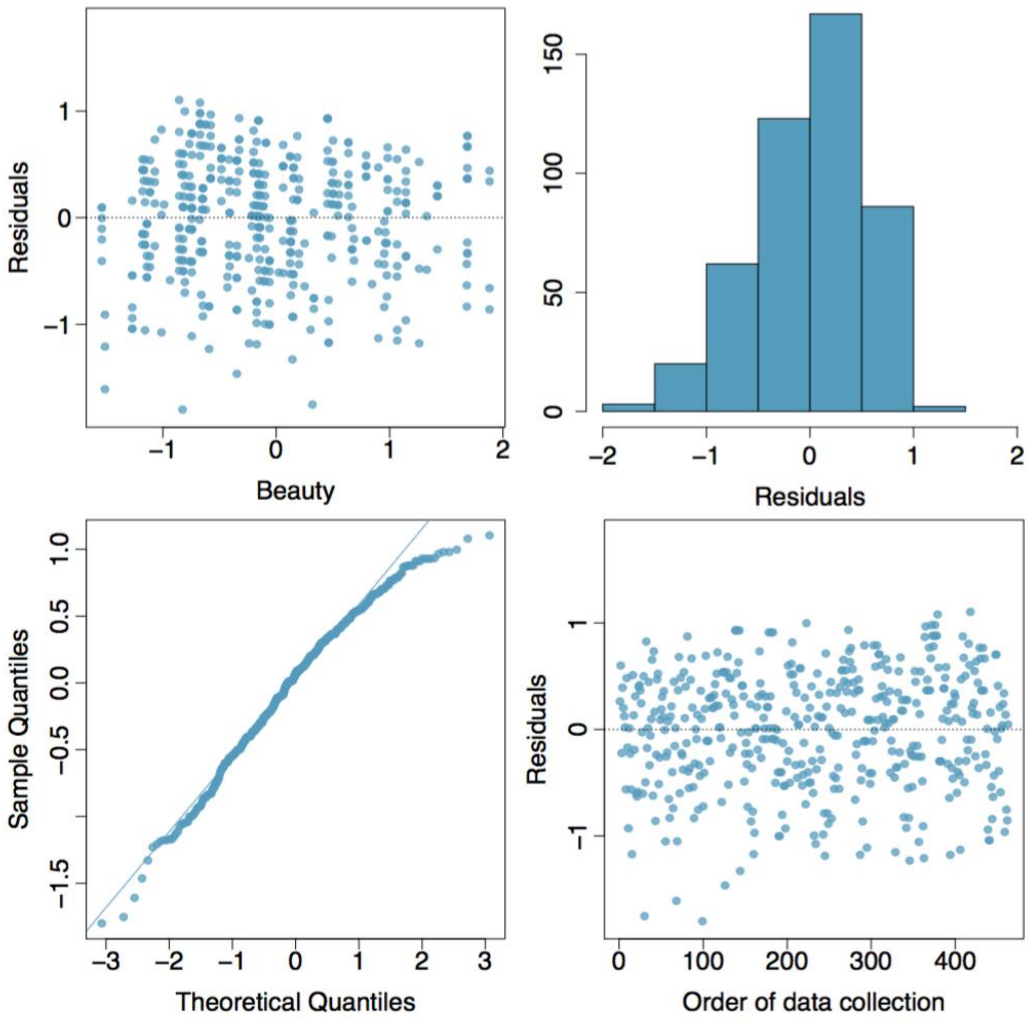


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	□	0.0322	4.13	0.0000

. (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.



Solution:

a) The mean point(-0.0883,3.9983) will lie on the regression line. The intercept 4.010 is given already
Substitute these values in the regression model to get the slope

3.9983 = 4.010 + b1 *(-0.0883)
b1 = (4.010 – 3.9983)/0.0883
0.1325

b) The slope b1 is positive(0.1325) indicating a positive correlation between beauty and teaching.

c) Conditions are:

- Linear trend: The scatter plot shows a weak linear relationship between beauty and teaching.

- Nearly normal residuals: The histogram of residuals is nearly normal. This is also shown by the quantiles plot.
- Constant variability: The scatter plot of the residuals shows roughly constant variance around the line.
- Independent observations: The data does not look a time series data. There is no other information or inferences that can be derived about how this data was collected. We assume the observations are independent.