Physiotherapy Statistics Exams

Alfredo Sánchez Alberca
asalber@ceu.es
https://aprendeconalf.es

Table of contents

Preface

Statistics exam collection of the Physiotherapy degree.

1 Descriptive Statistics and Regression exam (2018/05/31)

Exercise 1.1.

The ages of a sample of patients of a physical therapy clinic are:

a. Compute the quartiles.

b. Draw the box plot and identify outliers (do not group data into intervals).

c. Split the sample into two groups, patients younger and older than 65. In which group is the mean more representative. Justify the answer.

d. Which distribution is less symmetric, the one of patients younger than 65 or the one of patients older?

e. Which age is relatively smaller with respect to its group, 50 years in the group of patients younger than 65 or 72 years in the group of patients older than 65?

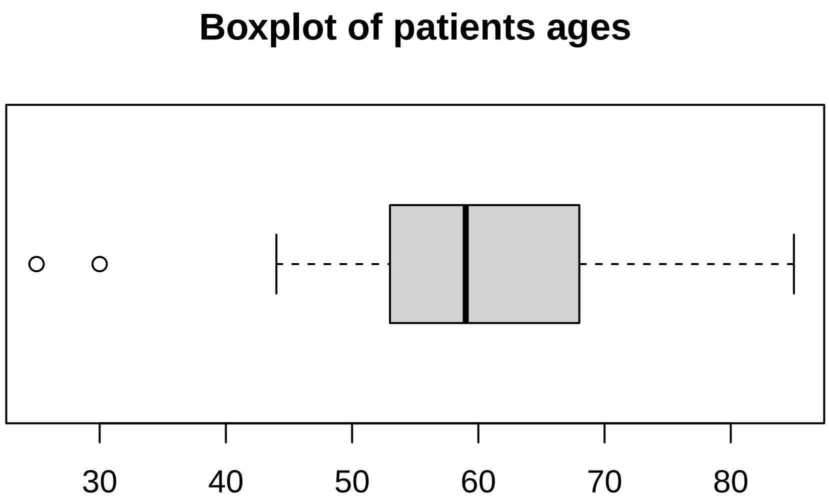
Use the following sums for the computations.
Younger than 65: $\sum x_i = 953$ years, $\sum x_i^2 = 52475$ years^2, $\sum(x_i - \bar{x})^3 = -30846.51$ years^3 and $\sum(x_i - \bar{x})^4 = 939658.83$ years^4.
Older than 65: $\sum x_i = 588$ years, $\sum x_i^2 = 43530$ years^2, $\sum(x_i - \bar{x})^3 = 1485$ years^3 and $\sum(x_i - \bar{x})^4 = 26983.5$ years^4.

Tip

a. $Q_1 = 53$ years, $Q_2 = 59$ years and $Q_3 = 68$ years.
b. There are 2 outliers: 25, 30.

25	30	44	44	51	51	53	56	57	58	58	58	59
59	61	63	63	63	66	68	70	71	72	74	82	85

Warm-up time	15	35	22	28	21	18	25	30	23	20
Injuries	42	2	16	6	17	29	10	3	12	20



a. Let x be the age in patients younger than 65 and y the age in patients older than 65.
$\bar{x} = 52.9444$ years, $s_x^2 = 112.1636$ years^2, $s_x = 10.5907$ years and $cv_x = 0.2$.
$\bar{y} = 73.5$ years, $s_y^2 = 39$ years^2, $s_y = 6.245$ years and $cv_y = 0.085$.
The mean is more representative in patients older than 65 since the coefficient of variation is smaller.

b. $g_{1x} = -1.4426$ and $g_{1y} = 0.7621$, thus the distribution of ages of people younger than 65 is less symmetric.

c. The standard scores are $z_x(50) = -0.278$ and $z_y(72) = -0.2402$, thus 50 years is relative smaller in the group of people younger than 65.

Exercise 1.2.

The table below shows the number of injuries of several teams during a league and the average warm-up time of its players.

a. Draw the scatter plot.

b. Which regression model is more suitable to predict the number of injuries as a function of the warm-up time, the logarithmic or the exponential? Use that regression model to predict the expected number of injuries for a team whose players warm-up 20 minutes a day.

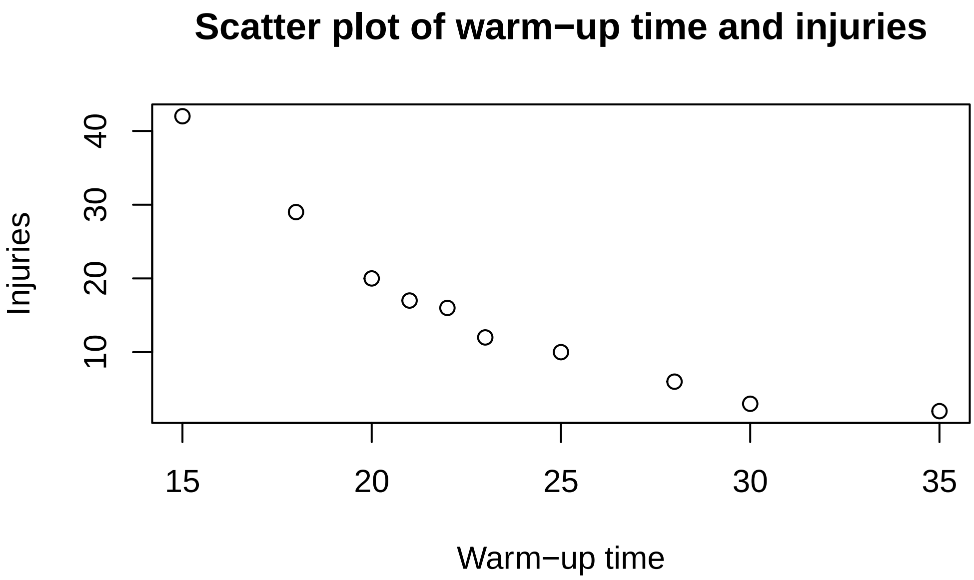
c. Which regression model is more suitable to predict the warm-up time as a function of the number of injuries, the logarithmic or the exponential? Use that regression model to predict the warm-up time required to have no more than 10 injuries in a league.

d. Are these predictions reliable? Which one is more reliable?

Use the following sums for the computations (X warm-up time and Y number of injuries):
$\sum x_i = 237$, $\sum \log(x_i) = 31.3728$, $\sum y_j = 157$, $\sum \log(y_j) = 24.0775$,
$\sum x_i^2 = 5937$, $\sum \log(x_i)^2 = 98.9906$, $\sum y_j^2 = 3843$, $\sum \log(y_j)^2 = 66.3721$,
$\sum x_i y_j = 3115$, $\sum x_i \log(y_j) = 519.1907$, $\sum \log(x_i) y_j = 465.8093$, $\sum \log(x_i)\log(y_j) = 73.3995$.

Solution

a.



a. $\bar{x} = 23.7$ min, $s_x^2 = 32.01$ min^2.
$\overline{\log(x)} = 3.1373$ log(min), $s_{\log(x)}^2 = 0.0565$ log(min)^2.
$\bar{y} = 15.7$ injuries, $s_y^2 = 137.81$ injuries^2.
$\overline{\log(y)} = 2.4078$ log(injuries), $s_{\log(y)}^2 = 0.8399$ log(injuries)^2.
$s_{x\log(y)} = -5.1446$, $s_{\log(x)y} = -2.6744$
Exponential determination coefficient: $r^2 = 0.9844$
Logarithmic determination coefficient: $r^2 = 0.9185$
So the exponential regression model es better to predict the number of injuries as a function of the warm-up time.
Exponential regression model: $y = e^{6.2168+-0.1607x}$.
Prediction: $y(20) = 20.1341$ injuries.

b. The logarithmic model is better to predict the warm-up time as a function of the number or injuries.
Logarithmic regression model: $x = 164.1851 + -47.3292\log(y)$.
Prediction: $x(10) = 55.2056112$ min.

c. Both predictions are very reliable since de deternation coefficient is very high but the last one is a little less reliable as it is for a value further from the data range.

2 Probability and Random Variables Exam (2022-05-06)

Exercise 2.1.

8% of people in a population consume cocaine. It is also known that 4%: of people who consume cocaine have a heart attack and 10%: of people who have a heart attack consume cocaine.

a. Construct the probability tree for the random experiment of drawing a random person from the population and measuring if he or she consumes cocaine and if he or she has a heart attack.

b. Compute the probability that a random person of the population does not consume cocaine and does not have a heart attack.

c. Are the events of consuming cocaine and having a heart attack dependent?

d. Compute the relative risk and the odds ratio of suffering a heart attack consuming cocaine. Which association measure is more suitable for this study? Interpret it.

Tip

Let C the event of consuming cocaine and H the event of having a heart attack.

a.

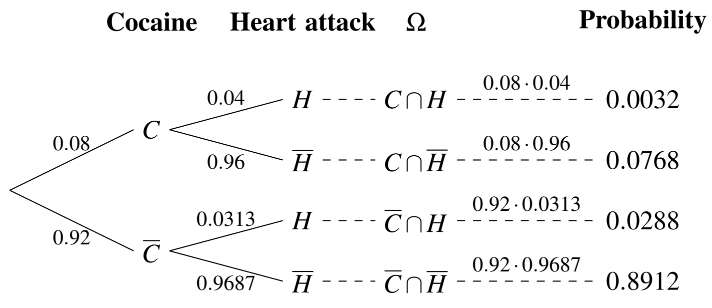


Figure 2.1: Probability tree

a. $P(\overline{C} \cap \overline{H}) = 0.8912$.

b. The events are dependent as $P(C) = 0.08 \neq P(C|H) = 0.1$.

c. $RR(H) = 1.2778$ and $OR(H) = 1.2894$. The odds ratio is more suitable as the study is retrospective. That means that the odds of having a heart attack is 1.2894 times greater if a person consumes cocaine.

Exercise 2.2.

A basketball player scores 12 points per game on average.

a. What is the probability that the player scores more than 4 points in a quarter?

b. If the player plays 10 games in a league, what is the probability of scoring less than 6 points in some game?

Solution

a. Let X be the points scored in a quarter by the player. Then $X \sim P(3)$, and $P(X > 4) = 0.1847$.

b. Let Y be the number of points scored in a game by the player. Then $Y \sim P(12)$ and $P(Y < 6) = 0.0203$.
Let Z be the number of games with less than 6 points scored by the player. Then $Z \sim B(10, 0.0203)$, and $P(Z > 0) = 0.1858$.

Exercise 2.3.

The creatine phosphokinase (CPK3) is an enzyme in the body that causes the phosphorylation of creatine. This enzyme is found in the skeletal muscle and can be measured in a blood analysis. The concentration of CPK3 in blood is normally distributed, and the interval centered at the mean with the reference values, that accumulates 99%: of the population, ranges from 40 to 308 IU/L in healthy adult males.

a. Compute the mean and the standard deviation of the concentration of CPK3 in healthy males.
Note: If you are not able to compute the standard deviation, use $\sigma = 50$ UI/L for the next parts.

b. A diagnostic test to detect muscular dystrophy gives a negative outcome when the concentration of CPK3 is below 300 UI/L. Compute the specificity of the test.

c. If the concentration of CPK3 in people with muscular dystrophy also follows a normal distribution with mean 350 IU/L and the same standard deviation, what is the sensitivity of the test?

d. Compute the predictive values of the test and interpret them assuming that the muscular dystrophy prevalence is 8%.

Solution

a. $\mu = 174$ IU/L and $\sigma = 51.938$ IU/L.

b. Specificity $= 0.9924$.

c. Sensitivity $= 0.8321$.
The test is better to confirm the disease as the specificity is greater than the sensitivity.

d. PPV $= 0.9046$. Thus, we can diagnose the disease with a positive outcome.
NPV $= 0.9855$. Thus, we can rule out the disease with a negative outcome.

3 Descriptive Statistics and Regression Exam (2022-06-06)

Exercise 3.1.

The patients of a physiotherapy clinic were asked to assess their satisfaction in a scale from 0 to 10. The assessments are summarized in the table below.

Assessment	Patients
$0-2$	3
$2-4$	12
$4-6$	9
$6-8$	18
$8-10$	22

a. Compute the interquartile range of the assessment and interpret it.

b. If it is required an assessment greater than 5 in more than 50%: of patients for the clinic to remain open, will the clinic remain open?

c. Is the assessment mean representative?

d. Compute the coefficient of kurtosis of the assessment and interpret it. Is the kurtosis normal?

e. If the assessment mean of another clinic is 6.8 and the standard deviation is 2.6, which assessment is relatively higher 6 in the first clinic or 6.2 in the second?

Use the following sums for the computations:
$\sum x_i n_i = 408$, $\sum x_i^2 n_i = 3000$, $\sum (x_i - \bar{x})^3 n_i = -548.25$ and $\sum (x_i - \bar{x})^4 n_i = 5140.45$.

Tip

Let X be the patient assessment. a. $Q_1 = 4.4444$, $Q_3 = 9.0907$ and $IQR = 4.6463$, so the central dispersion is moderate.

a. $F(5) = 0.2695$, and the percentage of patients with an assessment greater than 5 is 73.05.

Week	1	3	6	9	14	17	21	24
Grip strength	15	22	29	34	36	39	40	41

b. $\bar{x} = 6.375$, $s_x^2 = 6.2344$, $s_x = 2.4969$ and $cv = 0.3917$, thus the representativity of the mean is moderate.

c. $g_2 = -0.9335$ and the distribution is flatter than a Gauss bell, but normal, as g_2 is between -2 and 2.

d. First clinic: $z(6) = -0.1502$
Second clinic: $z(6.2) = -0.3077$.
Thus, an assessment of 6 in the first clinic is relatively higher as its standard score is greater.

Exercise 3.2.

A study tries to determine the effectiveness a training program to increase the grip strength. The table below shows the grip strength in Kg in some weeks of the training program.

a. Compute the regression coefficient of the grip strength on the weeks and interpret it.

b. According to the logarithmic regression model, what is the expected grip strength after 5 and 25 weeks. Are these predictions reliable? Would these predictions be more reliable with the linear regression model?

c. According to the exponential regression model, how many weeks are required to have a grip strength of 25 Kg?

d. What percentage of the total variability of the weeks is explained by the exponential model?

Use the following sums (X=Weeks and Y=Grip strength):
$\sum x_i = 95$, $\sum \log(x_i) = 16.7824$, $\sum y_j = 256$, $\sum \log(y_j) = 27.3423$,
$\sum x_i^2 = 1629$, $\sum \log(x_i)^2 = 43.606$, $\sum y_j^2 = 8804$, $\sum \log(y_j)^2 = 94.3237$,
$\sum x_i y_j = 3552$, $\sum x_i \log(y_j) = 342.9642$, $\sum \log(x_i) y_j = 608.4186$, $\sum \log(x_i)\log(y_j) = 60.047$.

Solution

a. $\overline{x} = 11.875$ weeks, $s_x^2 = 62.6094$ weeks^2.
$\overline{y} = 32$ Kg, $s_y^2 = 76.5$ Kg^2.

$s_{xy} = 64$ weeks·Kg.
Regression coefficient of Y on X: $b_{yx} = 1.0222$ Kg/weeek. The grip strength increases 1.0222 Kg per week.

b. $\overline{\ln(x)} = 2.0978$ ln(weeks), $s^2_{\ln(x)} = 1.05$ ln(weeks)^2 and $s_{\ln(x)y} = 8.9226$ ln(weeks)Kg.
Logarithmic regression model of Y on X: $y = 14.1729 + 8.498\ln(x)$.
Predictions: $y(5) = 27.8499$ Kg and $y(25) = 41.5268$ Kg.
Logarithmic coefficient of determination: $r^2 = 0.9912$. The predictions are not reliable because the sample size is small.
Linear coefficient of determination: $r^2 = 0.8552$.
As the linear coefficient of determination is less than the logarithmic one, the predictions with the logarithmic model are more reliable.

c. Exponential regression model of X on Y: $x = e^{-1.6345+0.1166y}$.
Prediction: $x(25) = 3.6015$ Weeks.

d. As $r^2 = 0.9912$, the exponential models explains 99.12%: of the variability of the weeks.

4 Probability and Random Variables Exam (2022-06-06)

Exercise 4.1.

A diagnostic test for a cervical injury has a 99% of sensitivity and produces 80% of right diagnosis. Assuming that the prevalence of the injury is 10%

a. Compute the specificity of the test.
b. Can we rule out the injury with a negative outcome of the test?
c. Can we diagnose the injury with a positive outcome of the test? What must the minimum prevalence of the injury be to diagnose the injury with a positive outcome of the test?

> Tip
>
> a. Specificity $= P(-|\overline{D}) = 0.7789$.
> b. Negative predictive value $= P(\overline{D}|-) = 0.9986 > 0.5$, so we can rule out the injury with a negative outcome.
> c. Positive predictive value $= P(D|+) = 0.3322 < 0.5$, so we can not diagnose the injury with a positive outcome. The minimum prevalence required to be able to diagnose the injury with a positive outcome is $P(D) = 0.1825$.

Exercise 4.2.

A pharmacy sells two vaccines A and B against a virus. The A vaccine produces 5%: of side effects, while the B vaccine produces 2%: of side effects. The pharmacy has sold 10 units of the A vaccine and 100 units of the B vaccine.

a. Compute the probability of having less than 2 side effects with the A vaccine.

b. Compute the probability of having more than 3 side effects with the B vaccine.

c. If we apply both vaccines to the same person at different moments, and assuming that the production of side effects of the vaccines are independent, what is the probability that this person will have any side effect?

Solution

a. Let X be the number of side effects in 10 applications of A vaccine. Then, $X \sim B(10, 0.05)$ and $P(X < 2) = 0.9139$.

b. Let Y be the number of side effects in 100 applications of B vaccine. Then, $Y \sim B(100, 0.02) \approx P(2)$ and $P(Y > 3) = 0.1429$.

c. Let A and B the events of having side effects with vaccines A and B respectively. $P(A \cup B) = 0.069$.

Exercise 4.3.

The length of the femur bone is normally distributed in both men and women with a standard deviation of 4 cm. It is also known that the first quartile in women is 42.3 cm, while the third quartile in men is 50.7 cm.

a. What is the difference between the means of the femur length of women and men? Remark: If you do not know how to compute the means, use a mean 44 cm for women and a mean 47 cm for men in the following parts.

b. Compute the 60th percentile of the femur length in women. What percentage of men have a femur length less than the 60th percentile of women?

c. If we pick a woman and man at random, what is the probability that neither of them has a femur length less than 45 cm?

Solution

Let X and Y be the femur length of women and men respectively. Then $X \sim N(\mu_x, 4)$ and $Y \sim N(\mu_y, 4)$.

a. $\mu_x = 44.91$ cm and $\mu_y = 48.02$ cm.

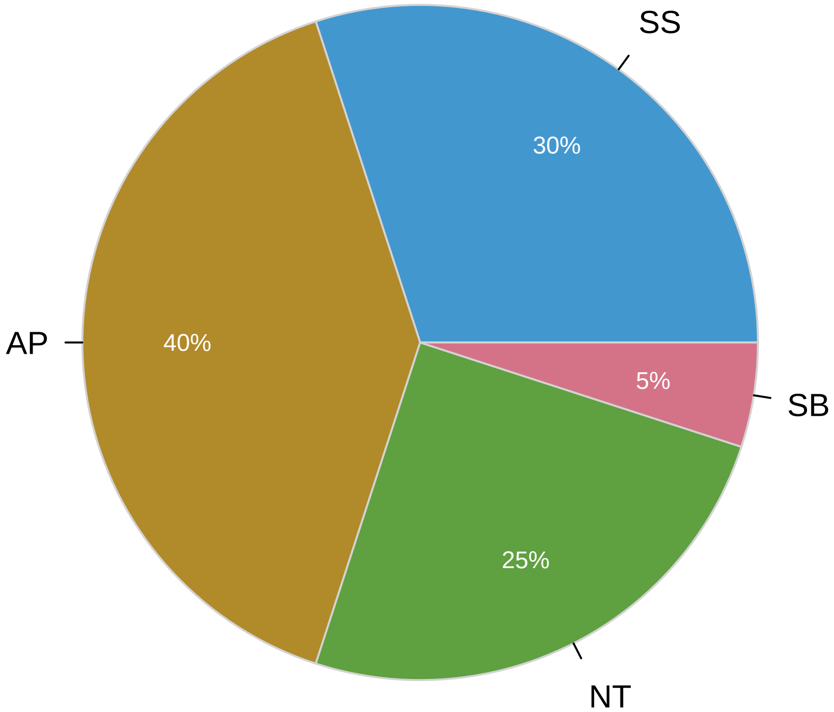
b. 60th percentile in women $P_{60} = 45.9234$ cm, and $P(Y < 45.9234) = 0.3001$, that is, a 30.01 of men have a femur lenght less than the 60th percentile of women.

c. $P(X \geq 45 \cap Y \geq 45) = 0.3805$.

5 Descriptive Statistics and Regression exam (2023/03/23)

Exercise 5.1.

The chart below shows the percentage of grades in a Statistic course with 60 students.



a. Plot the ogive of the score, assuming the following correspondence between grades and scores

Grade	Score
SS	$[0, 5)$
AP	$[5, 7)$
NT	$[7, 9)$
SB	$[9, 10]$

a. Compute the median and interpret it.

b. How many students got a score greater than 8?

c. Study the dispersion of the distribution.

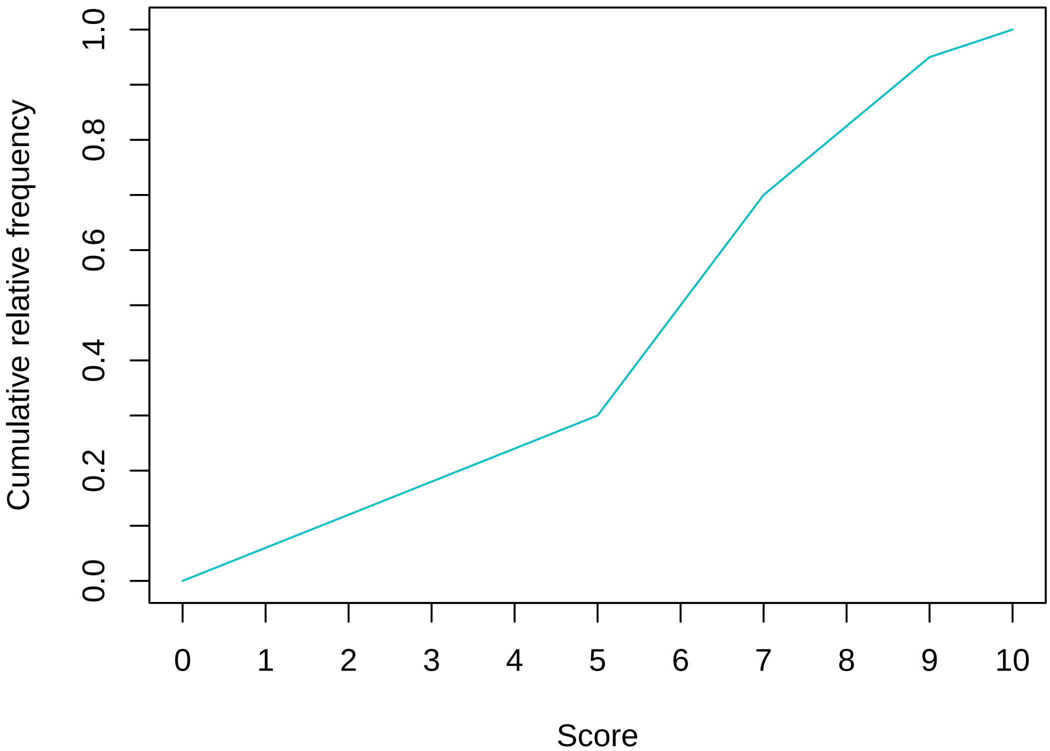
d. Study the skewness of the distribution. Is it normal?

e. If we apply the transformation $y = 10x + 5$ to the scores, how changes the representativeness of the mean. And the skewness?

Use the following sums for the computations (X = Score):
$\sum x_i n_i = 337.5$, $\sum x_i^2 n_i = 2207.25$, $\sum (x_i - \bar{x})^3 n_i = -172.55$ and $\sum (x_i - \bar{x})^4 n_i = 2870.75$.

Tip

a. Ogive



b. $Me = 6$ points.

c. $N(8) = 49.5$ students.

d. $\bar{x} = 5.625$ points, $s_x^2 = 5.1469$ points^2, $s_x = 2.2687$ points and $cv_x = 0.4033$. Thus, there is a moderate dispersion with respect to the mean.

e. $g_1 = -0.2463$ and therefore the distribution is a little bit left skewed.

f. $\bar{y} = 61.25$ points, $s_y^2 = 514.6875$ points^2, $s_y = 22.6867$ points and $cv_y = 0.3704$. As $cv_y < cv_x$ the representativeness of the mean increases. As the slope of the linear transformation is positive, the skewness does not change.

Exercise 5.2.

A study tries to determine if there is a relation between the gestation time (in weeks) and the age of the mother (in years). A sample of 40 mothers was taken and the sums below summarize the results (X=Age and Y=Gestation time):

$\sum x_i = 1262$ years, $\sum \log(x_i) = 137.0078$ log(years), $\sum y_j = 1583.6$ weeks, $\sum \log(y_j) = 147.1305$ log(weeks),
$\sum x_i^2 = 41862$ years^2, $\sum \log(x_i)^2 = 471.4222$ log(years)^2, $\sum y_j^2 = 62734.685$ weeks^2, $\sum \log(y_j)^2 = 541.2096$ log(weeks)^2,
$\sum x_i y_j = 50116.7$ years·weeks, $\sum x_i \log(y_j) = 4645.8$ years·log(weeks), $\sum \log(x_i) y_j = 5428.9192$ log(years)·weeks, $\sum \log(x_i)\log(y_j) = 504.0696$ log(years)·log(weeks).

a. Which regression models, linear, exponential or logarithmic, explains better the relation between the age and the gestation time?

b. Use the best model to predict the gestation time for a mother 45 years old. Is this prediction reliable?

c. According to the linear model, how much increases or decreases the gestation time for every year of the mother?

Solution

a. Linear model: $\overline{x} = 31.55$ years, $s_x^2 = 51.1475$ years^2.
$\bar{y} = 39.59$ weeks, $s_y^2 = 0.999$ weeks^2.
$s_{xy} = 3.853$ years·weeks.
$r^2 = 0.2905$.

Exponential model: $\overline{\ln(y)} = 3.6783$ ln(weeks), $s_{\ln(y)}^2 = 0.0006$ ln(weeks)^2
$s_{x\ln(y)} = 0.0958$ years· ln(weeks).
$r^2 = 0.2882$.

Logarithmic model: $\overline{\ln(x)} = 3.4252$ ln(years), $s_{\ln(x)}^2 = 0.0536$ ln(years)^2
$s_{\ln(x)y} = 0.1195$ ln(years)weeks.
$r^2 = 0.2668$.

As the linear coefficient of determination is greater, the linear model explains better the relation between de gestation time and the age of the mother.

b. Linear regression model of Y on X: $y = 37.2133 + 0.0753x$.
Predictions: $y(45) = 40.6032$ weeks.
The predictions are not reliable because the coefficient of determination is pretty low.

c. Regression coefficient of Y on X: $b_{yx} = 0.0753$ weeks/year. The gestation time increases 0.0753 weeks per year.