EXAM OF STATISTICS (DESCRIPTIVE STATISTICS AND REGRESSION)

2nd Physiotherapy Version A **May, 27 2019**

Duration: 1 hour and 15 minutes.

(5 pts.) 1. A study tries to determine the effect of smoking during the pregnancy in the weight of newborns. The table below shows the daily number of cigarretes smoked by mothers (X) and the weight of the newborn (all of them are males) (Y).

Daily num cigarettes	10.00	14.00	8.00	11.00	7.00	6.00	2.00	5.00	9.00	9.00	4.00	6.00
Weight (kg)	2.55	2.44	2.68	2.65	2.71	2.85	3.45	2.93	2.67	2.59	3.02	2.72

(a) Give the equation of the regression line of the weight of newborns on the daily number of cigarettes and interpret the slope.

(b) Which regression model is better to predict the weight of newborns, the logarithmic or the exponential?

(c) Use the best of the two previous regression models to predict the weight of a newborn whose mother smokes 12 cigarettes a day. Is this prediction reliable?

Use the following sums for the computations:
$\sum x_i = 91$ cigarettes, $\sum \log(x_i) = 23.0317$ log(cigarettes), $\sum y_j = 33.26$ kg, $\sum \log(y_j) = 12.1857$ log(kg),
$\sum x_i^2 = 809$ cigarettes^2, $\sum \log(x_i)^2 = 47.196$ log(cigarettes)^2, $\sum y_j^2 = 92.9708$ kg^2, $\sum \log(y_j)^2 = 12.4665$ log(kg)^2,
$\sum x_i y_j = 243.61$ cigarettes·kg, $\sum x_i \log(y_j) = 89.3984$ cigarettes· log(kg), $\sum \log(x_i) y_j = 62.3428$ log(cigarettes)kg, $\sum \log(x_i)\log(y_j) = 22.8753$ log(cigarettes) log(kg).

Solution

(a) $\bar{x} = 7.5833$ cigarettes, $s_x^2 = 9.9097$ cigarettes^2.
$\bar{y} = 2.7717$ kg, $s_y^2 = 0.0654$ kg^2.
$s_{xy} = -0.7176.$ cigarettes·kg
Regression line: $y = -0.0724x + 3.3208$.

(b) $\overline{\log(x)} = 1.9193$ log(cigarettes), $s^2_{\log(x)} = 0.2492$ log(cigarettes)^2.
$\overline{\log(y)} = 1.0155$ log(kg), $s^2_{\log(y)} = 0.0077$ log(kg)^2.
$s_{x\log(y)} = -0.2508$ log(cigarettes)·kg, $s_{\log(x)y} = -0.1245$ cigarettes·log(kg)
Logarithmic coef. determination: $r^2 = 0.9499$
Exponential coef. determination: $r^2 = 0.8268$
Therefore, the logarithmic models fits better the data and is better to predict the weight.

(c) Logarithmic regression model: $y = 3.7301 + -0.4994\log(x)$.
Prediction: $y(12) = 2.4892$ kg. The coefficient of determination is high but the sample size small, so the prediction is not enterely reliable.

(5 pts.) 2. The table below summarize the time that took to the runners to reach the finish in a long-distance race in Madrid:

Time (min)	Num runners
$(30, 35]$	15
$(35, 40]$	35
$(40, 45]$	40
$(45, 50]$	10

In a another race in Paris, the mean of time was 40 minutes, the standard deviation 5 minutes and the coeffitient of skewness 0.75.

(a) What percentage of runners took less than 42 minutes to reach the finish in Madrid?

(b) Compute and interpret the interquartile range of the time for Madrid race.

(c) In which race the mean of the time is more representative?

(d) In which race the time have a more simmetric distribution?

(e) In which race a time of 39 minutes to reach the finish is relatively smaller?

Use the following sums for the computations: $\sum x_i = 3975$ min, $\sum x_i^2 = 159875$ min^2, $\sum(x_i - \bar{x})^3 = -628.12$ min^3 y $\sum(x_i - \bar{x})^4 = 80701.95$ min^4.

Solution

(a)

(b) $Q_1 = 37.5$ min, $Q_3 = 42.5$ min, $IQR = 5$ min.

(c) Madrid statistics: $\bar{x} = 39.75$ min, $s^2 = 18.6875$ min^2, $s = 4.3229$ min and $cv = 0.1088$. Paris statistics: $cv = 0.125$. Thus, the mean of time in Madrid is a little bit more representative since the coef. of variation is smaller.

(d) $g_1 = -0.0778$, that is closer to 0 than the distribution of times in Paris, thus the distribution of times in Madrid is more simmetric.

(e) The standard score of the Madrid sample is $z(39) = -0.1735$ and the standard score of the Paris one $z(39) = -0.2$, thus a time of 39 min is relatively greater in the sample of Madrid.
