CHPC & NITheCS Coding Summer School

Probability & Statistics

Numerical Data & Regression

paul j. van staden (PhD)
Department of Statistics
University of Pretoria

NUMERICAL DATA & REGRESSION

EXAMPLE: Arachnophobia

- In a study on arachnophobia (the fear of spiders), 24 arachnophobes (persons fearing spiders) had to interact with spiders of different sizes.

- During each interaction, the level of anxiety of the arachnophobe was measured through galvanic skin response (GSR).

EXAMPLE adapted from:

Field, A. (2009). *Discovering Statistics using SPSS (and sex and drugs and rock 'n' roll)*, SAGE Publications Ltd, London, UK.

- **Consider the following two variables for $i = 1, 2, \ldots, 24$:**

  y_i: The GSR measurement for the level of anxiety of the i th arachnophobe.

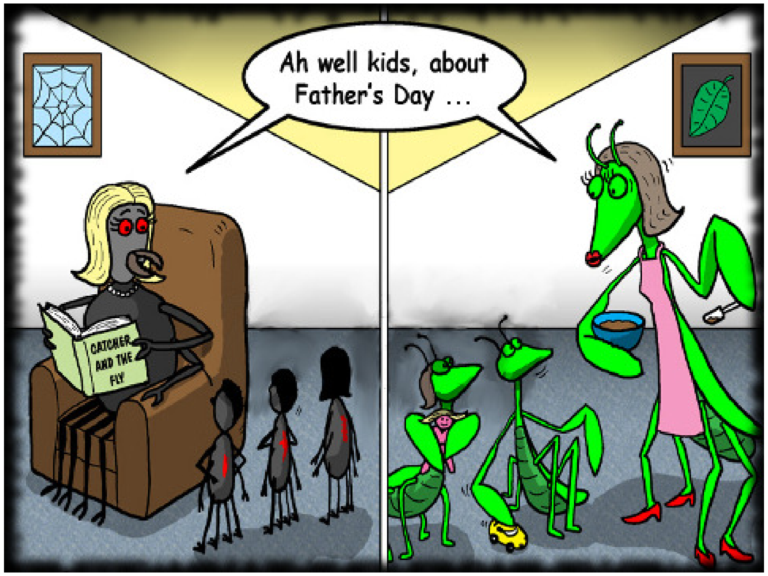
  x_i: The size of the spider in centimeters (cm) for the i th arachnophobe.

- **Calculate the correlation coefficient between the sizes of the spiders and the GSR measurements.**

$$r = 0.89$$

- **Characteristics of the correlation coefficient:**
  - r is only a measure of **LINEAR** dependence.
  - $-1 \leq r \leq 1$
  - $r_{x,y} = r_{y,x}$
  - r is independent of the units and the scale in which x and y are measured.
  - If x and y are independent, then $r = 0$.
  - But if $r = 0$, then x and y are not necessarily independent.
  - r cannot be used to describe cause-and-effect relationships.

- **Correlation analysis vs regression analysis:**
  - With <u>CORRELATION ANALYSIS</u> the strength of the linear relation between two variables is measured.
  - With <u>REGRESSION ANALYSIS</u> the population mean value of the response variable is estimated in terms of known values of the explanatory variable.



- **Simple linear regression model**

  - **Population regression line:**

$$y_i = \beta_0 + \beta_1 x_i$$

x: explanatory or predictor variable

y: response variable

β_0: intercept parameter which gives the mean value of y for $x = 0$

β_1: slope parameter which gives the change in the mean value of y for a unit increase in the value of x

- **Sample regression line:**

$$\hat{y}_i = b_0 + b_1 x_i$$

$\hat{y}$: estimator of the mean value of y for a given value of x

b_0: point estimate of the intercept parameter β_0

b_1: point estimate of the slope parameter β_1

- **Fit a linear regression model using the sizes of the spiders to explain the GSR measurements.**

$$\hat{y}_i = 3.5 + 2.8\, x_i$$

- **Interpret the parameter estimates.**

$$b_0 = 3.5$$

The mean GSR measurement for a spider of 0 cm is 3.5.

$$b_1 = 2.8$$

If the size of the spider that the arachnophobe has to interact with is increased by 1 cm, the mean GSR measurement will increase by 2.8.

Predicted values and residuals

- **Fitted regression line:**

$$\hat{y}_i = b_0 + b_1 x_i$$

- **Difference between the observed and the predicted values of the response variable:**

$$e_i = y_i - \hat{y}_i$$

y_i: observed value of the response variable

$\hat{y}_i$: predicted value of the response variable

e_i: residual

- **Predict the GSR measurement for Nosnow Cannotski who had to interact with a spider of 13 cm and calculate the corresponding residual.**

$$\begin{aligned}\hat{y}_5 &= 3.5 + 2.8\,x_5 \qquad (\text{note that } x_5 = 13)\\ &= 3.5 + 2.8 \times 13\\ &= 39.9\end{aligned}$$

$$\begin{aligned}e_5 &= y_5 - \hat{y}_5\\ &= 35 - 39.9\\ &= -4.9\end{aligned}$$

Since $e_5 < 0$, the GSR measurement of Nosnow Cannotski is <u>OVERESTIMATED</u> by the fitted regression model.

Least squares regression line

- **We want the residuals to be as small as possible.**
- **Sum of the residuals:**

  $\sum e_i = 0$: **cannot minimize this sum…**

- **Sum of the absolute residuals:**

  $\sum |e_i|$: **mathematically possible but practically not helpful to minimize this sum…**

- **Sum of the squared residuals:**

  $\sum e_i^2$: **minimizing this sum gives the LEAST SQUARES REGRESSION LINE**

- **Total variation:** $\sum(y_i - \bar{y})^2$
- **Explained variation:** $\sum(\hat{y}_i - \bar{y})^2$
- **Unexplained variation:** $\sum(y_i - \hat{y}_i)^2 = \sum e_i^2$
- **Percentage of the total variation in the response variable explained by the fitted regression line:**

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$

- **Calculate and interpret R^2 for the fitted regression line.**

$$R^2 = 0.8$$

80% of the variation in the GSR measurements is explained by the fitted regression line with the sizes of spiders as explanatory variable.