1. The observations (x_i, Y_i) satisfy the equation

$$Y_i = \alpha + \beta(x_i - \overline{x}) + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

where α and β are unknown constants, $\overline{x} = n^{-1} \sum x_i$ and the ε_is are independent normal random variables with mean zero and variance σ^2. The x_is are not all equal.

Show that the covariance matrix of the least squares estimators of α and β is

$$\sigma^2 n^{-1} \left(\sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^{n} (x_i - \overline{x})^2 & 0 \\ 0 & n \end{pmatrix}.$$

How would you estimate σ^2? How would you obtain $100\,(1-\gamma)\,\%$ confidence intervals for α and β?

What estimate would you use for the value of y when $x = 0$, and what is the variance of your estimate?
Now suppose that y depends upon two explanatory variables x_i and z_i according to the model

$$Y_i = \alpha + \beta(x_i - \overline{x}) + \gamma z_i + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

where the vectors $(x_1, x_2, \dots, x_n)$, $(z_1, z_2, \dots, z_n)$ and $(1, 1, \dots 1)$ are linearly independent. Show that the variance of the least squares estimate of β is

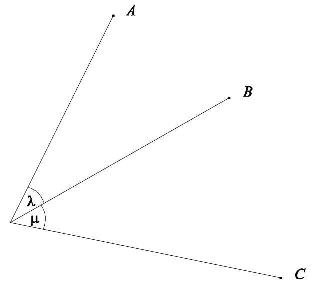
$$\frac{\sigma^2 \sum_{i=1}^{n} (z_i - \overline{z})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (z_i - \overline{z})^2 - \left(\sum_{i=1}^{n} (x_i - \overline{x})(z_i - \overline{z}) \right)^2}.$$

2. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is a random n-vector, $\mathbf{X}$ is an $n \times p$ design matrix of rank $p < n$, $\boldsymbol{\theta}$ is a p-vector of parameters and $\boldsymbol{\epsilon}$ is an n-vector of independent random variables with mean zero and variance σ^2. Derive the least squares estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and show that it is unbiased. Find its covariance matrix. Under what circumstances is the least squares estimator also the maximum likelihood estimator?

A sailor uses a sextant to measure the two angles in the horizontal plane subtended by three distant landmarks A, B, C: their true values are λ and μ.



He takes four measurements, the first two, y_1 and y_2, being of λ and μ respectively and the third and fourth, y_3 and y_4, each being of the combined angle $\lambda + \mu$. The measurements are subject to independent, normally distributed random errors with known variance σ^2. Obtain maximum likelihood estimates $\widehat{\lambda}$, $\widehat{\mu}$ of λ and μ. Show that the variance of $\widehat{\lambda}$ is $\frac{3}{5}\sigma^2$ and that the correlation coefficient between $\widehat{\lambda}$ and $\widehat{\mu}$ is $-\frac{2}{3}$.

The sailor suspects that each of the measurements has a constant bias β. Show that this bias is estimated as $y_1 + y_2 - \frac{1}{2}(y_3 + y_4)$ and show how to test the hypothesis that the bias is zero.

3. A team of researchers believe that the outcome Y of an experiment obeys the linear regression model

$$Y = \alpha + \beta x + \varepsilon$$

where α and β are unknown constants, the error $\varepsilon \sim N\left(0, \sigma^2\right)$ with σ^2 unknown, and where they can control the value of x. They conduct the experiment r times at $x = a$, r times at $x = -a$, and $n - 2r$ times at $x = 0$. Assume the associated errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent.

Derive explicit expressions for the maximum likelihood estimators of α and β. If the researchers' only interest is in estimating β, which value of r should they have chosen?

How should they estimate the value of σ^2?

Consider a new experiment about to be conducted, resulting in a new observation $\widetilde{Y}$ taken at $x = x_0$. Show how to construct

i) a $100\,(1-\delta)\,\%$ confidence interval for the mean value of $\widetilde{Y}$

ii) a $100\,(1-\delta)\,\%$ prediction interval for the new value of $\widetilde{Y}$.

4. Consider the model in Question 1 and suppose that two further independent observations, $\widetilde{Y}_1, \widetilde{Y}_2$ are to be made at the point x', that is

$$\widetilde{Y}_j = \alpha + \beta\left(x' - \overline{x}\right) + \widetilde{\epsilon}_j, \quad \text{for } j = 1, 2.$$

Find an expression for the variance τ^2 of the predictor of $\widetilde{Y}_1$ and explain how it can be used to construct a prediction interval for $\widetilde{Y}_1$. Show that the correlation between the prediction error of $\widetilde{Y}_1$ (defined as the difference between the true value of $\widetilde{Y}_1$ and the predictor of $\widetilde{Y}_1$) and the prediction error of $\widetilde{Y}_2$ is $\frac{\tau^2}{\tau^2+\sigma^2}$.

Optional Computer Exercises

1. The data file *wind.txt*, which you will find on the website, contains data from an investigation into the relationship between electrical current produced by a wind generator and wind speed (in miles per hour). Observations were recorded in the matched vectors *output* and *speed*.

(i) Load the data and use the `plot()` function to look at a scatterplot of *output* against *speed*. Also produce scatterplots of *output* against the square root of *speed*, *output* against the logarithm of *speed* and *output* against the reciprocal of *speed*. Does any of these plots show a plausible straight line relationship?

(ii) In the light of your investigations from part (i) of this question, fit the most plausible straight line. Produce a vector of fitted values and a vector of residuals.

(iii) Use the residuals to produce a normal probability plot; also produce a plot of residuals against fitted values. Comment briefly upon the suitability of your model in the light of the assumptions of the regression model.

[You may find it helpful to fit your linear model using an R command of the form `output <- lm(reponse ~variable1 + variable2 + ....` You can then obtain a model summary with `summary(output)` and fitted values and residuals are obtained respectively from `fitted(output)` and `resid(output)`.]

2. The proportion of successful putts (y) in a series of golf tournaments was recorded for different putt lengths (x), measured in feet. The data are taken from Iman, R.L. (1994) *A Data-based Approach to Statistics*, and are in the table below.

x	y	x	y
2	0.93	12	0.26
3	0.83	13	0.24
4	0.74	14	0.31
5	0.59	15	0.17
6	0.55	16	0.13
7	0.53	17	0.16
8	0.46	18	0.17
9	0.32	19	0.14
10	0.34	20	0.16
11	0.32		

(i) Enter these data into the computer and plot the variable y against x. Is it useful to model these data by fitting a regression line?

(ii) Fit the straight line $y = \alpha + \beta x$. According to your model, at what putting distance (in feet) is the chance of a successful putt equal to $\frac{1}{2}$?

(iii) Try the effect of the transformation $y_2 = \sqrt{y}$ by plotting against x. Does this transformation seem to be effective?

The usual approach to regression when the response variable y is a proportion is to transform to a new variable

$$y_3 = \log\left(\frac{y}{1-y}\right).$$

Try the effect of this transformation by plotting y_3 against x. Fit the models $y_2 = \alpha + \beta x$ and $y_3 = \alpha + \beta x$ obtaining vectors of fitted values and residuals. Which transformation do you prefer and why?

(iv) According to your preferred fitted model in part (iii), at what putting distance (in feet) is the chance of a successful putt equal to $\frac{1}{2}$?

(v) Use your fitted line to estimate the success rate at a putting distance of 40 feet. Does your answer make sense? What can you conclude?