Problem Sheet 7 (Vacation)

1. This question gives a way of deriving least squares estimators which is different from that in lectures.

Consider the standard linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is a random n-vector with observed values $\mathbf{y}$, $\mathbf{X}$ is a $n \times p$ design matrix of rank $p < n$, $\boldsymbol{\theta}$ is a p-vector of parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with zero means and variances σ^2. Show that

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbf{B}^{-1}\mathbf{X}^T\mathbf{y})^T\mathbf{B}(\boldsymbol{\theta} - \mathbf{B}^{-1}\mathbf{X}^T\mathbf{y}) + \mathbf{y}^T(\mathbf{I} - \mathbf{X}\mathbf{B}^{-1}\mathbf{X}^T)\mathbf{y}$$

where $\mathbf{B} = \mathbf{X}^T\mathbf{X}$.

Hence find the least squares estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, i.e. the value of $\boldsymbol{\theta}$ that minimizes $(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$.

Show that $\widehat{\boldsymbol{\theta}}$ is unbiased for $\boldsymbol{\theta}$ and find its covariance matrix. Deduce that the sum of squares of residuals of the fitted model $(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}})$ is given by

$$RSS = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\widehat{\boldsymbol{\theta}}.$$

How would you estimate σ^2?

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

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

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

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

$$\begin{pmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{\sigma^2}{\sum(x_i - \overline{x})^2} \end{pmatrix}.$$

Consider estimating the value of Y when $x = 0$. What estimate would you use, and what is the appropriate variance?

Now suppose that Y_i depends on two explanatory variables x_i and z_i according to the model

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

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

$$\frac{\sigma^2 \sum (z_i - \overline{z})^2}{\sum (x_i - \overline{x})^2 \sum (z_i - \overline{z})^2 - \left(\sum (x_i - \overline{x})(z_i - \overline{z})\right)^2}.$$

3. A spark chamber consists of a series of parallel metal plates fixed in the planes $x = x_i$, $i = 1, \ldots, n$. A particle emerges at the origin travelling on a parabolic trajectory whose axis is parallel to the y-axis, and causes sparks where the trajectory crosses the plates. Photographs showing the (x, y)-projection of the trajectory are then analysed to provide coordinates (x_i, y_i), $i = 1, \ldots, n$, for the point of emission of each spark, but the measurement of the y-coordinates are subject to independent errors drawn from a common normal distribution $N(0, \sigma^2)$. Show that the method of least squares gives the equation of the (x, y)-projection of the trajectory on the photographs as

$$y = x \left[\frac{(c_1 s_4 - c_2 s_3) + (c_2 s_2 - c_1 s_3) x}{s_2 s_4 - s_3^2} \right]$$

where $s_r = \sum x_i^r$ and $c_r = \sum y_i x_i^r$.

4. Let the independent random variables X_1, X_2 have respective means μ_1, μ_2 and variances σ_1^2, σ_2^2. Use a Taylor expansion of the function $h(x_1, x_2) = x_1/x_2$ about (μ_1, μ_2) to show that, for σ_1^2 and σ_2^2 both small, the variance of X_1/X_2 is approximately given by

$$\operatorname{var}\left(\frac{X_1}{X_2}\right) \approx \frac{\sigma_1^2}{\mu_2^2} + \frac{\mu_1^2 \sigma_2^2}{\mu_2^4}.$$

The yield Y of an industrial process is known to be a quadratic function of the temperature x. Measurements of yield at different temperatures are made comprising m observations at temperature $x = x_0$ and n observations at each of $x = x_0 + 1$ and $x = x_0 - 1$. Measurement errors are independent and normally distributed with zero mean and variance σ^2. Writing the quadratic model in the form

$$E(Y \mid x) = \alpha + \beta(x - x_0) + \gamma(x - x_0)^2,$$

show that the least squares estimates of β and γ are given by

$$\widehat{\beta} = \tfrac{1}{2}(\overline{Y}_1 - \overline{Y}_{-1}), \qquad \widehat{\gamma} = \tfrac{1}{2}(\overline{Y}_1 + \overline{Y}_{-1} - 2\overline{Y}_0),$$

where $\overline{Y}_0, \overline{Y}_1, \overline{Y}_{-1}$ are the means of the observations at x_0, $x_0 + 1$, $x_0 - 1$, respectively. Show that, provided $\gamma < 0$, the maximum expected yield is obtained when $x = x_0 - \beta/2\gamma = x^*$ say, and that when σ^2 is small, the variance of the estimate $x_0 - \widehat{\beta}/2\widehat{\gamma}$ of x^* is approximately

$$\frac{\sigma^2}{8\gamma^4}\left(\frac{\beta^2 + \gamma^2}{n} + \frac{2\beta^2}{m}\right).$$

Please turn over

5. Consider the linear model

$$Y_i = a + bx_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the independent random variables ε_i are normally distributed with mean zero and unknown variance σ^2. Derive the least squares estimates $\widehat{a}$ and $\widehat{b}$ of a and b and show that they are unbiased. How would you estimate the variances of $\widehat{a}$ and $\widehat{b}$?

Show how you would use your estimates of a, b, σ^2 to construct a $100(1-\alpha)\%$ confidence interval for the predicted value of y at any given point x.

In calibration problems it is common for estimates of a and b to be used to predict a value of x for a given y. By considering the distribution of the random variable

$$W = \widehat{a} + \gamma\widehat{b} - y,$$

where $\gamma = (y-a)/b$, show how to construct a $100(1-\alpha)\%$ confidence interval for the predicted value of x.