a5 Statistics TT 2003 **Exercises 4** Dr. Gesine Reinert

1. If gene frequencies are in equilibrium, the genotypes AA, Aa and aa occur in a population with frequencies $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2, according to the so-called *Hardy-Weinberg equilibrium*. In a sample from the Chinese population of Hong Kong in 1937, blood types occurred with the following frequencies, where M and N are erythrocyte antigens.

	M	MN	N	Total
Frequency	342	500	187	1029

Test the hypothesis that the multinomial distribution is as specified by the Hardy-Weinberg equilibrium, with unknown parameter θ, against the alternative that the multinomial distribution does not have probabilities of that specified form. Use a 5% significance level. (Hint: If X_2, X_2 and X_3 are the counts in the three cells, show that the maximum likelihood estimator is $(2X_3 + X_2)/(2n)$.)

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. Percentages of total calories obtained from complex carbohydrates were recorded for 20 male insulin-dependent diabetics who had been on a high carbohydrate diet for six months. Age (x_1), body weight (x_2) and percentage protein in diet (x_3) were tried as explanatory variables and a regression model was fitted to the data. The results were as follows:

	Coefficient	Standard error
Intercept	36.960	13.071
x_1	−0.114	0.109
x_2	−0.228	0.083
x_3	1.958	0.635

Correlations

	Intercept	x_1	x_2
x_1	−0.297		
x_2	−0.606	0.000	
x_3	−0.581	−0.216	−0.132

The residual sum of squares is 567.663 and the total sum of squares is 1092.800.

On the assumption that the normal linear model is appropriate:

(i) Calculate F-tests for (a) the whole regression, (b) the effect of age. Comment on your result for (b).

(ii) Compute 95% confidence intervals for the coefficients of x_2 (body weight), x_3 (protein) and the difference between these coefficients.

(iii) Explain how to compute a 95% confidence interval for the mean carbohydrate response when the values of age, weight and protein are 43, 120 and 16 respectively.

4. The table below gives the number of births per thousand in the states of the USA in three regions for the year 1985.

	North East	South Atlantic	South Central
	14.5	15.5	14.2
	15.5	15.5	14.0
	15.0	15.8	14.9
	14.1	15.1	16.6
	13.5	12.5	14.9
	13.9	14.3	18.2
	14.6	15.6	16.1
	14.0	16.1	18.8
	13.5	14.4	
Totals	128.6	134.8	127.7

A linear model, which allows for different birth rates in the three regions, is fitted. The residual sum of squares is 36.143 and the total sum of squares is 48.063. Test the null hypothesis that the birth rates in the regions are the same. What can you say about the differences between birth rates?