COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 17, 2007, 9:00 A.M.–1:00 P.M.

STATISTICS 664 QUESTIONS

Answer all parts. Closed book, calculators allowed. It is important to show all working, especially with numerical calculations. Statistical tables are provided. You may freely quote results from the course notes or text without proof, but to the extent that it is feasible to do so, state precisely the result you are quoting.

Chronic bronchitis (defined as persistent cough and phlegm) is known to be associated with exposure to high levels of sulfur dioxide (SO_2) in the air. Table 1 gives results of one epidemiological survey with average concentrations of SO_2 listed against the prevalence ratios. The exact sample size in each area is lost, and we regard it as 2000 when necessary.

Table 1. Average concentrations of SO_2 and prevalence ratios of chronic bronchitis.[a]

SO_2, mg/day/100 cm^2	Prevalence ratio
0.21	0.035
0.28	0.033
0.27	0.031
0.15	0.030
0.15	0.029
0.14	0.027
0.13	0.027
0.14	0.025
3.4	0.078
2.75	0.059
2.75	0.052
2.1	0.048
1.6	0.038
1.55	0.037
1.15	0.032
1.0	0.027
0.9	0.024

[a]Data of Mantel and Bryan (*4*).

A quick look at the scatter plot indicates that the simple linear regression model may not fit the data well, and we consider instead the following piece-wise linear regression model (PLRM):

$$y_i = \beta_{0,1} + x_i\beta_{1,1} + \epsilon_{i,1}, \text{ for } x_i \leq u, \quad (1)$$

$$y_i = \beta_{0,2} + x_i\beta_{1,2} + \epsilon_{i,2}, \text{ for } x_i > u. \quad (2)$$

where $i = 1, 2, \ldots, n$, $\epsilon_{i,1}$ and $\epsilon_{i,2}$ have i.i.d. normal distribution with mean zero and $\mathrm{Var}(\epsilon_{i,k}) = \sigma_k^2$, and we require that $\mathrm{E}[y_i]$ is a continuous function of x_i, i.e.,

$$\beta_{0,1} + u\beta_{1,1} = \beta_{0,2} + u\beta_{1,2}. \tag{3}$$

Without loss of generality we assume x_i is of increasing order, i.e., $x_1 \le x_2 \le \ldots \le x_n$. For problem 1-4, we assume we know that $x_l < u < x_{l+1}$.

1. (15 pts) Let $\boldsymbol{\beta}_k = (\beta_{0,k}, \beta_{1,k})'$, $k = 1, 2$, $\mathbf{x}^{(1)} = (x_1, \ldots, x_l)'$, $\mathbf{x}^{(2)} = (x_{l+1}, \ldots, x_n)'$, $\mathbf{y}^{(1)} = (y_1, \ldots, y_l)'$, $\mathbf{y}^{(2)} = (y_{l+1}, \ldots, y_n)'$. One way to estimate $\widehat{\boldsymbol{\beta}}_k$, $\widehat{\sigma}_k$, and $\widehat{u}$ is to fit simple linear regression model to $\mathbf{x}^{(k)}$ and $\mathbf{y}^{(k)}$, and estimate $\widehat{u}$ using (3). Write down the explicit formula for computing $\widehat{\boldsymbol{\beta}}_k$, $\widehat{\sigma}_k$, and $\widehat{u}$, which we will refer to as the local least square estimator.

2. (15 pts) Derive the maximum likelihood estimator of u, $\boldsymbol{\beta}_k$ and σ_k^2 for $k = 1, 2$. Are they the same as the local least square estimator?

3. (10 pts) Under the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, give an unbiased estimator of σ^2 and determine its distribution.

4. (10 pts) Derive an F-test to compare the simple linear regression model (SLRM) and the PLRM, and give the distribution of the F statistic.

5. (15 pts) In practice one usually do not know l for which $x_l < u < x_{l+1}$. Describe a method for determining l. If l is determined from the data, do you expect the F-test in (4) to be more or less likely to reject the null hypothesis of SLRM? Explain your reasoning.

   For problem 6-8, we use the following model, which is a special case of PLRM and often referred to as the hockey stick regression model (HSRM):

   $$y_i = \beta_{0,1} + \epsilon_{i,1}, \text{ for } x_i \le u, \tag{4}$$
   $$y_i = \beta_{0,2} + x_i\beta_{1,2} + \epsilon_{i,2}, \text{ for } x_i > u. \tag{5}$$

   For the SO_2 data, u can be interpreted as the threshold above which SO_2 is positively related to chronic bronchitis. Thus estimating the threshold u is of particular importance, which can be used for determining the air quality standard for sulfur dioxide.

6. (15 pts) Suppose from previous survey we know that the threshold u for SO_2 is between 1.15 and 1.55. Estimate u using the data in Table 1.

   Instead of using $\widehat{u}$ as the safe concentration level, it is preferable to use a lower confidence limit of u to provide some assurance that the true u is above the safe concentration level.

7. (10 pts) Describe an exact procedure for computing the confidence interval of u under the assumption $\sigma_1^2 = \sigma_2^2$.

8. (10 pts) Give a lower confidence limit of u with 95% confidence level.

Solutions, 2006 CWE 174

1. Standard results.

2. u, $\boldsymbol{\beta}_k$ the same, σ_k^2 different.

3. Linear combination of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$. Scaled χ^2 with $n-4$ d.f.

4. Refer to the general theorem on F-test. $\mathrm{F}_{2,n-4}$.

5. ML can be used. More likely to reject H_0.

6. $\widehat{u} = 1.210$

7. Similar to Feiller's method.

8. 0.904