Solutions to exam questions

STK4900/9900 June 8th, 2009

Problem 1

a) For model 1 the multiple correlation coefficient is $R^2 = 0.948$, while for model 2 it is $R^2 = 0.978$. The multiple correlation coefficient measures the proportion of the variation in the observations that is explained by the model. Since model 2 explains a larger proportion of the variation than model 1, model 2 is the one to be preferred for predicting the volume of a tree.

b) We will predict the volume y of a tree with diameter $x_1 = 15$ inches and height $x_2 = 70$ feet. According to model 2 we get the prediction:

$$\begin{aligned} \lg(\hat{y}) &= -2.880 + 1.983 \cdot \lg(x_1) + 1.117 \lg(x_2) \\ &= -2.880 + 1.983 \cdot \lg(15) + 1.117 \cdot \lg(70) \\ &= 1.513 \end{aligned}$$

Thus the predicted volume of a tree with diameter 15 inches and height 70 feet is

$$\hat{y} = 10^{1.513} = 32.6 \text{ cubic feet}$$

c) We have observations (y_i, x_{i1}, x_{i2}); $i = 1, 2, \ldots, n$; and consider a linear regression model of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{1}$$

(For model 2 we have a model of this form with y_i replaced by $\lg(y_i)$ and x_{ij} replaced by $\lg(x_{ij})$ for $j = 1, 2$.)

The assumptions for a linear regression model of the form (1) are the following:

(i) Linearity: $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

(ii) Constant variance: $\text{Var}(\epsilon_i) = \sigma^2$ for all i

(iii) Uncorrelated errors: $\text{Cov}(\epsilon_i, \epsilon_l)$ for all $i \neq l$

(iv) Normally distributed errors: $\epsilon_i \sim N(0, \sigma^2)$

Assumption (i) may be checked by plotting the residuals versus each of the covariates; cf. the two bottom plots for each of the models in the problem

text. If such a plot looks like a "random scatter", it indicates that assumption (i) is reasonably well fulfilled for a covariate, while a systematic pattern of the residual (e.g. a curvature) indicates deviation from linearity. For model 2 the plots looks fine for both covariates. For model 1 the plot for height looks fine, while there is an indication of some curvature in the plot for diameter (indicating a non-linear effect of this covariate).

Assumption (ii) may be checked by plotting the residuals versus the fitted values; cf. the upper right-hand plots in the problem text. If such a plot looks like a "random scatter", it indicates that assumption (ii) is reasonably well fulfilled, while a "fan like" plot indicates that the residual variance is not constant. For model 2 the plot looks fine, while there is an indication of some curvature in the plot for model 1 (which is caused by a non-linear effect of diameter; cf. above).

Assumption (iv) may be checked by a QQ-plot of the residuals; cf. the upper left-hand plots in the problem text. The QQ-plot is approximately a straight line for both models. This indicates that model assumption (iv) is reasonably well fulfilled.

[Problems with assumption (iii) may occur when the observations are taken consecutively in time. Then a plot of the residuals versus the observation number may be used to check (iii). This is not relevant in the present example, however.]

Problem 2

a) There are two ways of reasoning that may lead to assume that the number of accidents in a group is Poisson distributed:

- If the group is reasonably large and we disregard the possibility of two or more accidents for a driver in the three months period, we may argue that the number of accidents is binomially distributed with large n and small p, and hence approximately Poisson distributed.
- If we assume that each driver has same risk of being involved in an accident every day, and accidents on one day are independent of accidents on other days, the process counting the occurrence of accidents is a Poisson process, and hence the number of accidents in a three moths period is Poisson distributed. (Note that this argument is valid also for small groups.)

(Each of these explanations is fine as an answer to the question.)

b) The rate ratios for levels 2, 3 and 4 of motor volume, compared to level 1, are:

Level 2: $RR(2) = e^{0.149} = 1.16$

Level 3: $RR(3) = e^{0.389} = 1.48$

Level 4: $RR(4) = e^{0.553} = 1.74$

According to the estimated rate ratios, level 2 has a risk of accidents that is 16% larger than level 1, while levels 3 and 4 have accidents risks that are 48% and 74%, respectively, larger than level 1.

c) The rate ratio for cars with motor volume 1–1.5 litres compared to cars with motor volume less than 1 litre is $RR(2) = e^{\beta_2}$, where β_2 is the regression coefficient for level 2 of motor volume when level 1 is the reference.

We first derive a 95% confidence interval for β_2. This takes the form:

$$\hat{\beta}_2 \pm 1.96\,\hat{\text{se}}_2$$

Inserting the estimates given in the output in the problem text, this becomes:

$$0.1493 \pm 1.96 \cdot 0.0505 \quad \text{i.e. the interval} \quad [0.0503\,,\,0.2483]$$

Exponentiating the lower an upper limits of the confidence interval for β_2, we obtain the following 95% confidence interval for rate ratio $RR(2) = e^{\beta_2}$:

$$[e^{0.0503}, e^{0.2483}] = [1.05\,,\,1.28]$$

Thus a 95% confidence interval for the rate ratio for cars with motor volume 1–1.5 litres compared to cars with motor volume less than 1 litre is from 1.05 to 1.28.

d) To see how the risk of accidents depends on the age of the driver, we compute the rate ratios for the age groups 25–29 years, 30–35 years, and more than 35 years, compared to the age group less than 25 years:

25–29 years: $RR(2) = e^{-0.191} = 0.83$

30–35 years: $RR(3) = e^{-0.345} = 0.71$

More than 35 years: $RR(4) = e^{-0.537} = 0.58$

According to the estimated rate ratios, drivers that are 25–29 years old have a risk of accidents that is 83% that of the risk of the youngest drivers, while the drivers who are 30–35 years and 35 years or more have accidents risks

that are 71% and 58%, respectively, that of the risk of the youngest drivers.

Let γ_2 be the regression coefficient for drivers 25–29 years old when the group of youngest drivers is the reference. In order to test the null hypothesis $\gamma_2 = 0$, we may use the Wald statistic $\hat{\gamma}_2/\hat{\text{se}}_2$, which is approximately standard normally distributed if the null hypothesis is true. Inserting the estimates given in the output, the Wald statistic takes the value $-0.191/0.083 = -2.30$. This gives a P-value of about 2%, so there is a significant difference in accident risks between the two youngest age groups.

e) For the model in question b with motor volume as the only covariate, the intercept is the logarithm of the risk for cars with motor volume less than 1 litre (which is the reference for the model in b).

For the model in question d with motor volume, age and area as covariates, the intercept is the logarithm of the risk for cars with motor volume less than 1 litre, with drivers less than 25 years and not living in London or other big cities (which is the reference for the model in d).

As the intercept for the model in d is for the youngest drivers, and the youngest drivers have the highest risk of accidents, this explains why the intercept for the model in question d is larger than the one for the model in question b.

f) Interaction between age and area means that the effect of age is not the same in all areas. To test the hypothesis of no interaction between age and area, we may use the G-statistic, i.e. the difference between the deviances of the fitted model without and with interaction. Here the G-statistic takes the value $G = 51.42 - 40.91 = 10.51$. This should be compared with the chi-squared distribution with $3 \cdot 3 = 9$ degrees of freedom (= the number of parameters used to model interaction between age and area). This gives a P-value larger than 10%, so the null hypothesis of no interaction between age and area is not rejected.