Exam Questions For Regression Analysis **Date: 29 May 2023**

Time: 8-13. Limit to pass: 25p out of 40 points.

Permitted aids: One page (both side) Hand-written cheat sheet for the course. Pocket calculator. Please write only on one side of the paper and do not use red colour.

---

1. The following represents the average life span of smokers according to the average number of cigars they smoke in a day:

No. of cigars	5	10	15	20
Life span	65	56	48	40

   Let x represent the number of cigars and y the life span.

   (a) Find the least squares line for the data set (use two decimal places). [2]

   (b) Is the life span significantly associated with the number of smokes (do a proper hypothesis test)? [2]

   (c) Use this model to predict the average life span of a person who smokes 25 cigars daily (round-off to the nearest year). [2]

   (d) make a prediction interval for the predicted life span in the previous question [2].

   Note: Answers should be correct to four decimal places for this item.

2. Consider the model with

   $Y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, 2, 3$. We have the observations $(-2, y_1), (0, y_2), (2, y_3)$.

   (a) Find the least-square estimate of β_0 and β_1 in terms of y_1, y_2, y_3. [4]

   (b) Find the covariance between the estimated values of β_0 and β_1. [4]

3. Suppose that the error component ϵ in the multiple regression model $Y = X\beta + \epsilon$, has mean zero and covraiance matrix of $Var(\epsilon) = \sigma^2(\Omega)$, where Ω is a known n $\times$ n positive definite symmetric ma- trix and $\sigma^2 > 0$ is a constant (possibly unknown but you do not need to estimate it).

   (a) Find the generalized least square of the β. [4]

   (b) Show formally that $\hat{\beta}_{GLS}$ is an unbiased estimator of β and determine its covariance matrix. [4]

4. An insurance company wants study the "Medical Cost Personal" or the charge based on the age, gender, BMI, number of children, smoking statues, and the region (The data is not shown as it was large). The goal is to find the association of each predictor variables with the outcome variable (charge). A multivariate linear regression has been fitted. Here is the R output:

```
Call:
lm(formula = charges ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           C**      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      A**   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  B** 0.458769
regionsoutheast  -1035.0      D**  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

a. What are the value of $A**$ and $B**$, $C**$ and $D**$? [2]

b. what is the interpretions of the of the models, which variables are significantly associated with the charge? [2]

c. what will be the predicted value of charge for a 32 years old female (bmi=26) with 2 kids who is non-smoker and living in the southeast ? [2]

d. Which gender and which area cost more for the insurance company and why? [2]

5. Even before the space shuttle Challenger exploded on January 20, 1986, NASA had collected data from 20 earlier launches. One part of these data was the number of Orings that had been damaged at each launch1. In total there were six such O-rings at the Challenger. Data includes the number of damaged O-rings and the temperature (in Fahrenheit) at the time of the launch. There is concern whether damage of O-rings is related to the temperature, and hence a Poisson regression

N	x	N	x
2	53	0	70
1	57	1	70
1	58	1	70
1	63	0	72
0	66	0	73
0	66	0	55
0	67	2	75
0	67	0	76
0	68	0	76
0	69		
0	70		

Table 1: Data

model was fitted, with N, the number of damaged rings, being the response and temperature, x, the explanatory variable.

a. A model was fitted by R, see the output below, showing the fitted model. What are the values of $A**$ and $B**$? [2]

b. Is temperature a significant variable in the proposed model? why? [2]

c. Write down the fitted model. [2]

d. On the fateful day when the Challenger exploded, the temperature was 31$\circ$F. Use the fitted model to predict the expected number of damaged O-rings at that temperature. [2]

e. Based on the result in (c), would you say that the model based on Poisson regression is adequate here? [2]

```
Call:
glm(formula = NrRings ~ Temp, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9512  -0.8377  -0.6575   0.1344   2.3816

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.31821    2.94460   A**   0.0709 .
Temp        -0.09260    0.04587  B**  0.0435 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 19.918  on 19  degrees of freedom
Residual deviance: 16.119  on 18  degrees of freedom
AIC: 35.347

Number of Fisher Scoring iterations: 6
```