Name:

Exam H00Y0A Data and Statistical Modelling (total points:8, will be rescaled to 20)
Prof. A. Carbonez
Monday, August 29 2022, 13.00 –16.00

1. **if you specify p-values (do not say: p-value > 0.05 but say p-value = 0.076 and hence p-value > 0.05 …; do not say p-value < 0.05 but say p-value = 0.0003 and hence p-value < 0.05 … .)**
2. **Be sure to always formulate H0 and H1 in case you write and interpret a p-value.**
3. **Use =0.05 unless specified otherwise**

4. **Cell phones should be switched off and in your backpack. Backpacks on the ground.**
5. **Data is available on Toledo (> Toledo > Assignment)**
6. **If you need more white paper, raise hand. There is more white paper in the front of the room.**
7. **Write your name on this document.**
8. **This is the exam document. Write in a structured way, use nice handwriting. Be complete.**
9. **It is an open book exam, you are allowed to use your course notes, exercises (on** Toledo or printed). You can make use of R S **pt.** **Everything should be written on this document.**
10. **At the end of the exam, hand in this document in the box when leaving the room.**

Success!

Name:

1. Suppose the distribution of cholesterol values is normally distributed with mean = 220 mg/dl and standard deviation = 35 mg/dl. . **(points: 1)**
    a. What is the probability that a cholesterol level will range from 200 to 250 ? . **(points:0.5)**

$P[200 < X < 250]$

$= P[X < 250] - P[X < 200]$

pnorm(250, 220, 35)

    b. What is the cholesterol value corresponding with the 20th percentile? . **(points: 0.5)**

qnorm(0.2, 220, 35)

Name:

2. Buying a new car

A marketing research firm was engaged by an automobile manufacturer to conduct a pilot study to estimate the probability that a family will purchase a new car during the next year. A random sample of 33 suburban families was selected. Data on annual income (X1, in thousand dollars) and the current age of the oldest family car (X2, in years) were obtained.
A follow-up interview conducted 12 months later was used to determine whether the family actually purchased a new car (Y=1) or did not purchase a new car (Y=0) during the year.

. **(points: 1)**

Data: car_purchase.xlsx

Y	X1	X2
0	32	3
0	45	2
1	60	2
0	53	1
0	25	4
1	68	1
1	82	2
1	38	5
0	67	2
1	92	2
1	72	3

(this is only the first part of the data frame which you can find in car_purchase)

In order to find an appropriate model for the probability of purchasing a new car, you are going to analyze 2 models:

Model A: As explanatory variables you use annual income, age of oldest car and the interaction term.
Model B: as explanatory variables you use only annual income.

Use the appropriate partial deviance test to compare model B with model A.
State the null hypothesis, the alternative, the value of the test statistics, the p-value and the conclusion. . **(points: 1)**

(you are allowed to continue your answer at the back of this sheet).

Model: logistic regression model

Model A: $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

$p = P[Y=1]$

Model B: $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$

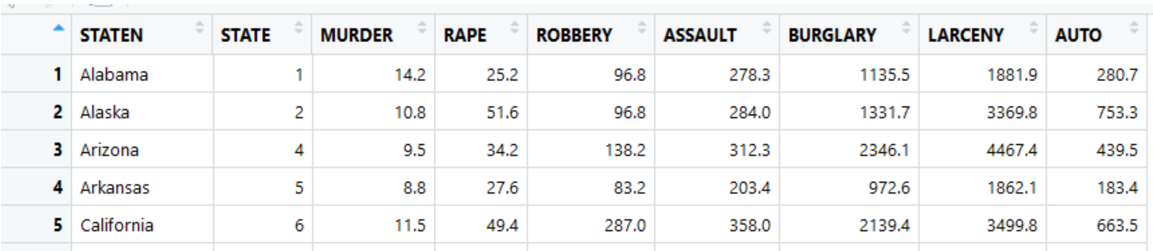
$H_0: \beta_2 = \beta_3 = 0$ vs H_1: not both are zero

Partial Dev $= 3.9$ + Pvalue $= 0.14 > \alpha$ => Not reject H_0

Name:

3. Analyze crimes data (points: 2.5 points)

Data: crimes.xlsx



	STATEN	STATE	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
1	Alabama	1	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
2	Alaska	2	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
3	Arizona	4	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
4	Arkansas	5	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
5	California	6	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5

Perform a PCA on the variables murder, rape, robbery, assault, burglary, larceny and auto. Answer the questions below.

(i) How many principal components are you going to keep? Give 2 reasons why. (points: 0.25)

(ii) Give an interpretation to these principal components (after varimax rotation). . (points: 0.5)

Name:

(iii) **What is the communality of rape in the subspace (PC1, PC2)? .** **(points: 0.75)**

(iv) **What can you say about Nevada?**
Give the principal component scores + interpret only in terms of the PC’.s **(points: 0.5)**

Name:

(v) **What can you say about Massachutes?**
Give the principal component scores + interpret only in terms of the PC’s **(points: 0.5)**

.

Name:

4. Use movies data **movies_sml** (2 point) independent, null hypothesis test (Chapt 7)

Variable name	**remark**
action	1: action movie; 0: not action movie
length	The number of minutes

We are interested to know whether the average length of an action movie is more than 17 minutes longer compared to the average length of a non-action movie.
Do a complete analysis (also assumptions), formulate H0,H1 and interpret all the conclusions.

(a) normality check

satisfied because CLT (both >25)

(b) check the equality of variance (homogeneous)

$H_0: \sigma_0^2 = \sigma_1^2$ vs $H_1: \sigma_0^2 \neq \sigma_1^2$

p value < α

Reject H_0

(c) $H_0: \mu_1 = \mu_0 + 17$ vs $H_1: \mu_1 > \mu_0 + 17$ in R movie$length ~ movie$action

$H_0: \mu_0 = \mu_1 - 17$ vs $H_1: \mu_0 < \mu_1 - 17$

$\bar{X}_0 =$ $\bar{X}_1 =$

Diff $(\bar{X}_1 - \bar{X}_0) =$

Name:

5. **Small questions. Be careful, these are multiple choice questions with GIS** correction. If your a**75. If your answer is** correct you
incorrect, you lose 0.25. **(points: 1.5)**

举例说明

1. Consider for the same data and same null hypothesis. Formulate a one –sided and a two sided hypothesis test. Then we can say

   [✓ (i)] in general, one-sided tests have more power
   (ii) In general, one-sided tests have less power
   (iii)In general, no difference in power between one-sided and two-sided test
   (iv)No decision is possible without real data

   the power is the probability to reject Ho when Ho is false

2. If you do a PC analysis for a dataset with 8 variables which are correlated with each other. Assume that we do the PC analysis on the correlation matrix. You look at the squared loadings for PC1 (the first Principal component).

What can you say about the sum of the values of these squared loadings :

(i) Will always be equal to 1
(ii) Will be larger than 1
(iii) Will be smaller than 1
(iv) We cannot know this without the data