Interim test 3

Part 1 – logistic regression

Question 1. In a logistic regression analysis, what is the measurement level of the dependent variable?

a. Nominal
b. Ordinal
c. Interval
d. Ratio

Question 2. What is the definition of the odd ratio of a particluar event (*P* is the probability that the event occurs)?

a. P
b. $\ln(P)$
c. $\ln(1-P)$
d. $\frac{P}{1-P}$

Question 3. For a logistic regression model with one independent variable *X* we obtain the following estimation results:

		B	S.E.	Wald	df	Sig.
Step 1[a]	X	0.837	0.065	10.780	1	0.001
	Constant	-0.378	0.034	4.202	1	0.040

What is the value of the probability this model predicts for a case where $X = 1$?

a. The probability is smaller than 0.5
b. The probability is larger than 0.5
c. The probability is approximately equal to 0.5
d. The probability is approximately equal to zero

Question 4. Someone wants to test whether the probability that a student still lives in the parental home is different between male and female students. Based on the data for a sample of 500 students she estimates a binary logistic regression model. She obtains the following results (Male = 1, if the student is a male; Male = 0, if the student is a female):

		B	S.E.	Wald	df	Sig.
Step 1[a]	Male	-0.243	0.088	10.780	1	0.413
	Constant	-0.776	0.024	4.202	1	0.000

What conclusion should be drawn?

a. This test cannot be based on logistic regression analysis
b. The alternative hypothesis is accepted
c. The null hypothesis is accepted
d. The conclusion depends on Nagelkerke's R^2 of the model

Question 5. How many equations does a multinomial logistic regression model have for a dependent variable that has four categories (= values)?

a. 1 equation
b. 2 equations
c. 3 equations
d. 4 equations

Question 6. For a logistic regression model we obtain the following classification table:

Observed \ predicted	1	2	3	total
1	120	22	5	147
2	32	44	8	84
3	12	53	105	170
total	164	119	118	401

What is the value of the hit ratio?

a. 37%
b. 21%
c. 42%
d. 67%

Question 7. For a *binary* logistic regression model we find a hit ratio of 50.6%. What do you expect regarding the value of Nagelkerke's R2 of this model?

a. The value will be claose to zero
b. The value will be close to one
c. The value will be close to 0.5
d. The hit ratio does ont say anything about the R2

Question 8. In a choice experiment, home-owners can choose between 4 alternatives for making their dwelling more energy efficient: 1) insulating the house; 2) installing solar panels; 3) installing a heat pump and 4) no change. To predict the choice based on income (in Euro/month) and the current use of natural gass (in m3/year) of the household, we estimate a multinomial regression model. The 4-th alternative (no change) is used as the reference category.

The estimation results are:

Alternative		B	Std. Error	Wald	df	Sig.
insulation	Intercept	0.1730	0.0066	4.535	1	0.0332
	Income	0.000107	3.2E-10	35.778	1	0.0000
	Gass_use	0.0000694	6.6E-10	7.298	1	0.0049
solar_panels	Intercept	-1.990	0.128	30.938	1	0.0000
	Income	0.00006	5.3E-10	6.792	1	0.0092
	Gass_use	0.000877	1.21E-07	6.356	1	0.0117
heat_pump	Intercept	-0.100	0.00069	14.493	1	0.0001
	Income	0.000045	3.3E-10	6.136	1	0.0132
	Gass_use	0.000171	6.1E-09	4.794	1	0.0286

For a home-owner with an income of 2500 Euro/month and a current gass use of 1750 m3 / year – which alternatives have a higher probability than the reference category according to the model?

a. Only alternative 1
b. Only alternative 3
c. Only alternatives 1 and 3
d. Alternatives 1, 2 and 3

Part 2 - Non-parametric tests

Question 9. A manager of an office building is interested in the strength of the relationship between two variables: the amount of energy used for heating (in kilowatt) in a work space and the feeling of comfort in the work space. For a sample of 20 work spaces the energy use and feeling of comfort are determined. The feeling of comfort is determined by an expert: the expert judges the comfort by rank ordering the 20 work places from highest to lowest level of comfort.

Which test should be used to investigate the strength of the relationship between these two variables?

a. Spearman correlation
b. ANOVA
c. Pearson correlation
d. Kruskal-Wallis H test

Question 10: Which statement about non-parametric tests is NOT true?

a. A non-parametric test has more power than a parametric test for finding a difference or a relation between two variables
b. A non-parametric test does not require that the variables are normally distributed
c. A non-parametric test is also safe to use when the number of observations is small (smaller than 30, but larger than 10)
d. A non-parametric test can be used when one or both of the variables is of ordinal level

A researcher is interested in the question whether there are sex differences in car parking ability. The null hypothesis is that men and women do not differ and the alternative hypothesis is that men need less time to park the car than women do. He times how fast 14 women and 16 men can park their car (in seconds). It appears that the time variable deviates strongly from a normal distribution in both groups.

Question 11: Which parametric test would be suitable if the time variables were normally distributed?

a. Chi-square t-test
b. Independent samples t-test
c. Paired samples t-test
d. ANOVA

Since the time variables are not normally distributed and the groups in the sample are small, the researcher decides to conduct a non-parametric test. SPSS gives two tables as output. The first table is:

Ranks

	Gender	N	Mean Rank	Sum of Ranks
Seconds	1.00	16	13.64	191.00
	2.00	14	17.13	274.00
	Total	30		

SPSS determines rank scores from low to high values of the dependent variable (time in seconds). Men are coded as 1 and women as 2.

The second output table is:

Test Statistics[a]

	Seconds
Mann-Whitney U	86.000
Wilcoxon W	191.000
Z	-1.081
Asymp. Sig. (2-tailed)	0.280
Exact Sig. [2*(1-tailed Sig.)]	0.294[b]

[a]. Grouping Variable: gender
[b]. Not corrected for ties.

Question 12: Looking at the results in the above Test Statistics table, what conclusion can you draw?

a. This is a non-parametric test, therefore the null hypothesis cannot be tested
b. Reject null hypothesis; conclusion is: men have shorter average parking time than women
c. Accept null hypothesis; conclusion is: average parking time does not differ between men and women
d. Reject null hypothesis; conclusion is: average parking time differs between men and women - we cannot tell in what way

For a survey a sample of 300 persons is used. Of the persons we know the type of household they belong to (Single, Couple, Family with children). For this type-of-household variable we also know the distribution in the population. The table below shows the frequencies for the sample (observed frequencies) and the known distribution in the population (percentages).

Household type	Observed frequency	Population
Single	63	28.4 %
Couple	130	35.0 %
Family with children	107	36.6 %
Total	300	100 %

We do a Chi-square test to determine whether the sample is representative for the population in terms of household type.

Question 13: How are expected frequencies defined for this Chi-square test?

a. Expected frequencies are equal to the population percentages
b. Expected frequencies are equal frequencies for the household type categories
c. Expected frequencies are equal to the observed frequencies
d. Expected frequencies are the frequencies we expect for the sample if the distribution in the sample is the same as in the population

The expected frequencies are calculated in the proper way and the Chi-square is computed. The Chi-square is 11.808. To determine the p-value the degrees of freedom (df) for this test must be known.

Question 14: What is the value of the degrees of freedom for this test?

a. df = 5
b. df = 4
c. df = 3
d. df = 2

With the proper degrees of freedom we find that the p-value is equal to p = 0.0027.

Question 15: What do you conclude?

a. Reject the null hypothesis; conclusion is: the sample is representative
b. Reject the null hypothesis; conclusion is: the sample is NOT representative
c. The test is not reliable because the degrees of freedom is too low
d. The test is not reliable because not all conditions for a Chi-square test are met