UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Exam: **ECON4130 – Statistics 2**

Date of exam: Wednesday, November 27, 2019 **Grades are given:** 18. December 2019

Time for exam: 2.30 p.m. – 5.30 p.m. (3 hours)

The problem set covers 7 pages

Resources allowed:
- Open book examination, where all written and printed resources – as well as two alternative calculators - are allowed

The grades given: A-F, with A as the best and E as the weakest passing grade. F is fail.

Problem 1

A. Let V be a continuous random variable (rv) with cumulative distribution function (cdf)

$$F_V(v) = P(V \le v) = \begin{cases} 0 & \text{if } v \le 0 \\ v^{10} & \text{if } 0 < v < 1 \\ 1 & \text{if } v \ge 1 \end{cases}$$

i) Let $f_V(v)$ denote the probability density function (pdf) of V. Find $f_V(v)$ and make a rough sketch of its graph.

ii) Determine the quantile function, v_p, of V for $0 < p < 1$ (i.e., find v_p such that $F_V(v_p) = p$ for $0 < p < 1$), and explain why $P(v_{0.025} \le V \le v_{0.975}) = 0.95$.

iii) Calculate the expected number of observations larger than 0.9 among 100 independent observations of V.

[**Hint:** You may utilize a suitable binomial distribution.]

B. Let Z be another continuous rv with pdf, $f_Z(z)$, given by

$$f_Z(z) = \begin{cases} \frac{2}{\alpha^2} z & \text{for } 0 < z < \alpha \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > 0$ is a parameter.

i) Show that, for $0 < z < \alpha$, the cdf of *Z* is given by $F_Z(z) = \left(\frac{z}{\alpha}\right)^2$. Find the median of *Z* expressed by α.

ii) Let μ_r denote the r-th moment, $E\left(Z^r\right)$, of *Z*. Show that $\mu_r = \frac{2}{r+2}\alpha^r$ for $r > 0$, and that $\text{var}(Z) = \frac{\alpha^2}{18}$.

C.

i) Suppose that the true value of α (in **B**.) is unknown. Find the moment method estimator (mme) for α based on an iid (independent and identically distributed) sample, $Z_1, Z_2, \ldots, Z_n$, where each Z_i is distributed as *Z*.

ii) Find the expectation and the variance of the mme expressed by α.

D.

An alternative way to estimate α is to use the largest one of $Z_1, Z_2, \ldots, Z_n$ in section **C(i)**. Let *Y* be the largest Z_i (i.e., $Y = \max\{Z_1, Z_2, \ldots, Z_n\}$).

i) Explain why

$$P(Y \le y) = P(Z_1 \le y)P(Z_2 \le y)\cdots P(Z_n \le y) = \left(\frac{y}{\alpha}\right)^{2n} \text{ for } 0 < y < \alpha .$$

ii) Show that $\hat{\alpha} = \frac{2n+1}{2n}Y$ is an unbiased estimator of α.

[Hint. Find $E(Y)$ first. **]**

iii) Find $\text{var}(\hat{\alpha})$ and show that $\hat{\alpha}$ is consistent.

[Hint. You may use that $\text{var}(Y) = \frac{n}{(n+1)(2n+1)^2}\alpha^2$, a formula you do not have to prove here. **]**

E. We have $n = 5$ independent observations of *Z* given in table 1.

Table 1 Observations of *Z*

i	1	2	3	4	5	Sum
z_i	1.42	4.80	1.91	4.75	4.54	17.42

i) Estimate α from the data in table 1 using both the mme (section **C**) and $\hat{\alpha}$ from section **D**. Which of the two estimates do you prefer? Give a reason for your answer.

ii) Show that $V = \frac{Y}{\alpha}$ is distributed as *V* in section **A** if $n = 5$, where *Y* is defined in **D**. Use this to derive the formula for an exact 95% confidence interval (CI) for α .

iii) Calculate the observed interval derived in **ii)** using the data in table 1.

F. It turns out the maximum likelihood principle does not work properly in this case.

i) Which one of the conditions for "good behavior" of the maximum likelihood estimator, is not fulfilled under the model in **C**?

ii) Write up the log likelihood function, $l(\alpha)$, under the model in **C** for $Z_1, Z_2, \ldots, Z_n$, and explain why the maximization of $l(\alpha)$ does not lead to any sensible estimator of α .

Problem 2

A political scientist developed a questionnaire to determine political tolerance scores (*Y*) for a random sample of faculty members at her university. The score was constructed in such a way that the higher the score, the more tolerant the individual. She wanted to compare mean scores, controlling for age (*X*), in each of three academic groups: 1. full professors, 2. associate professors, and 3. assistant professors.

There were $n = 30$ individuals in all who answered the questionnaire, with 10 in each of the three groups. The resulting data[1] are given in appendix **A1**.

We characterize the three groups by two dummy variables, d_1 and d_2, where d_1 indicates full professor and d_2 associate professors. The possible values of d_1, d_2 are shown in **table 2**:

Table 2

	d_1	d_2
Group 1 Full professors	1	0
Group 2 Associate professors	0	1
Group 3 Assistant professors	0	0

[1] The data are taken from Kleinbaum, Kupper, Muller & Nizam, «*Applied Regression Analysis and Other Multivariate Methods*», 3, edition, Duxbury Press 1998

Assume that the regression of *Y* with respect to X, d_1, d_2 is homoscedastic with regression function

(1) $$\mu(x, d_1, d_2) = E\left(Y \mid X = x, d_1, d_2\right) = \alpha + \beta x + \gamma_1 d_1 + \gamma_2 d_2 + \gamma_3 x \cdot d_1 + \gamma_4 x \cdot d_2$$

and condtional variance, $\text{var}(Y \mid x, d_1, d_2) = \tau^2$ (constant). Note that the model has one response, *Y*, three explanatory variables, X, d_1, d_2, and five regressor variables, $z_1 = x,\ z_2 = d_1,\ z_3 = d_2,\ z_4 = x \cdot d_1,\ z_5 = x \cdot d_2$. In addition, we assume that the conditional distribution of *Y*, given $X = x$, and d_1, d_2, is normal with expectation $\mu(x, d_1, d_2)$ and constant variance τ^2.

A.

i) The model (1), that we will refer to as *the full model* below, implies different regression functions for the three groups of faculty members. For example, for assistant professors (group 3) we have $E(Y \mid x, 0, 0) = \alpha + \beta x$, i.e., a straight line in *x*.
Write up the corresponding regression lines for group 1 and group 2 expressed by parameters from (1).

ii) What restrictions on the parameters in (1) would imply that the three regression lines in **i)** are parallel (i.e., that they have the same coefficient of increase or decrease)?
Suppose you want to test the hypothesis of parallel regression lines against the full model (1) with a likelihood ratio test (LR). How many degrees of freedom would you use in the approximate chi-square distribution that is used for the LR test statistic under the null hypothesis?

iii) The full model (1) is linear in the parameters and may be estimated from the data using the ordinary least squares (OLS) method under the following corresponding model for the data:

$$Y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \gamma_3 x_i \cdot d_{i1} + \gamma_4 x_i \cdot d_{2i} + e_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

where, given all ages and groups, the error terms, $e_1, e_2, \ldots, e_n$ are assumed to be iid and normally distributed, $e_i \sim N(0, \tau^2$.

The Stata output for this estimation is given in appendix **A2**. According to the output it appears that none of the regression coefficients (except the constant term) are significantly different from zero. Does this mean that there is evidence that the five regressor variables do not have explanatory power for *Y*, or, is there some other evidence in the output that points to the opposite conclusion that the five regressors, taken together, actually do have explanatory power? Explain.

B. Initial considerations led to, among other things, the conclusion that *X* (age) does not seem to matter much for group 3 (assistant professors), at least for the observed age range in group 3 - a conclusion you do not need to investigate here. As a result the following homoscedastic model was suggested:

(2) $\mu(x, d_1, d_2) = E\left(Y \mid X = x, d_1, d_2\right) = \alpha + \gamma_1 d_1 + \gamma_3 x \cdot d_1 + \gamma_4 x \cdot d_2$

$\text{var}(Y \mid x, d_1, d_2) = \tau^2$ (constant)

i) In what way may we consider model (2) to be a special case of the full model (1)? Describe the regression line for group 3 (assistant professors) implied by model (2), instead of by model (1), as derived in **A(i).**

ii) We want to test model (2) against model (1). Set up an F-test for this problem based on the OLS output for model (1) and (2) given in appendix **A2-3 (**under the corresponding model conditions for the data as for model (1)).
Calculate the test and verify that the p-value is larger than 10% (using, e.g., a table of quantiles of the F-distribution). Comment briefly on the result considering this test a specification test for model (2).

C. Assume model (2) to be true. The estimated model (2), according to appendix **A3** becomes (with 3 decimals precision for the parameter estimates):

(3)
$$\hat{E}(Y \mid x, d_1, d_2) = \hat{\alpha} + \hat{\gamma}_1 d_1 + \hat{\gamma}_3 \cdot x d_1 + \hat{\gamma}_4 \cdot x d_2 =$$
$$= 4.971 + (3.241) \cdot d_1 - (0.086) \cdot x d_1 - (0.002) \cdot x d_2$$

i) The mean scores in the data are 4.099, 4.855, 4.995 for group 1,2 3 respectively. It seems unfair to compare them directly since the differences may be partly due to differences in mean ages in the three groups. The function $\mu(x, d_1, d_2) = E\left(Y \mid x, d_1, d_2\right)$, however, represents the population means of *Y* in the three groups at age *x,* and allows us to control for the influence of age by comparing the estimated population means using *the same* age for the three groups. One natural choice is to choose the mean age in the whole sample as the common age, i.e., choose $x = 40$.

Calculate the estimated mean scores at age 40 (i.e., $\hat{E}(Y \mid 40, d_1, d_2)$), in all the three groups.

ii) Write μ_j for $E\left(Y \mid x, d_1, d_2\right)$ in group *j*, $j = 1, 2, 3$. According to Stata the estimated standard error for $\hat{\mu}_1 = \hat{E}(Y \mid 40, 1, 0)$ in group 1 is $SE(\hat{\mu}_1) = 0.2693$. Use this to calculate a 95% confidence interval for μ_1.

[Hint. You may use, without justification here, that $T = \frac{\hat{\mu}_1 - \mu_1}{SE(\hat{\mu}_1)}$ is approximately standard normally distributed for large and fixed *n*, considering $n = 30$ as sufficient, and provided that $SE(\hat{\mu}_1)$ is a consistent estimator.

Alternatively, if you prefer, you may use that *T* in the present model is exactly *t*-distributed with 26 =30-4 degrees of freedom. This result, that you do not need to justify here, follows from general regression theory and applies to any linear combination of parameters in the regression function. The degrees of freedom are obtained as *n* minus the number of parameters in the regression function (i.e., 4 in this case). The *t*-distribution property, however, is strongly dependent on the normality assumption of the error terms, while the first statement on the approximate standard normal distribution does not depend on normality of the error terms. **]**

iii) We want to test, at the level of significance 5%, if the expected mean scores at age 40 for full professors (group 1) is significantly different from the expected mean scores at the same age for assistant professors (group 3).

Explain briefly why this amounts to test the null hypothesis, $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, where $\delta = \gamma_1 + 40\gamma_3$.

Let the estimator of δ be $\hat{\delta} = \hat{\gamma}_1 + 40\hat{\gamma}_3$. Calculate the standard error of $\hat{\delta}$ based on the estimated covariance matrix of the estimated regression parameters (obtained from Stata):

Table 3 The estimated covariance matrix for the regression parameters in model (2)

	$\hat{\gamma}_1$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\alpha}$
$\hat{\gamma}_1$	1.29729	-0.0254	0.0009	-0.0374
$\hat{\gamma}_3$	-0.0254	0.00053	0	0
$\hat{\gamma}_4$	0.0009	0	0.00005	-0.0009
$\hat{\alpha}$	-0.0374	0	-0.0009	0.03743

Using that $\frac{\hat{\delta}}{SE(\hat{\delta})}$ is approximately standard normally distributed (or, if you prefer, *t*-distributed with 26 degrees of freedom as in the hint of **ii)**) if $\delta = 0$, perform the test of H_0 against H_1. (You do not need to justify the distribution claims here.). Comment on the result of the test.

Appendix (for problem 2)

A1 Data. Political tolerance and age for 30 individuals.

	Group 1: Full professors ($d_1 = 1 \quad d_2 = 0$)										Mean
x Age	65	61	47	52	49	45	41	41	40	39	48
Y Polit. Tol.	3.03	2.70	4.31	2.70	5.09	4.02	3.71	5.52	5.29	4.62	4.099
	Group 2: Associate professors ($d_1 = 0 \quad d_2 = 1$)										Mean
x Age	34	31	30	35	49	31	42	43	39	49	38.3
Y Polit. Tol.	4.62	5.22	4.85	4.51	5.12	4.47	4.50	4.88	5.17	5.21	4.855
	Group 3: Assistant professors ($d_1 = 0 \quad d_2 = 0$)										Mean
x Age	26	33	48	32	25	33	42	30	31	27	32.7
Y Polit. Tol.	5.20	5.86	4.61	4.55	4.47	5.71	4.77	5.82	3.67	5.29	4.995

Total mean age (30 individuals): $\bar{x} = 40$ years

A2 OLS results for the full model (1).

```
      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(5, 24)        =      5.02
       Model |  10.2093433         5  2.04186866   Prob > F        =    0.0027
    Residual |  9.76275337        24   .40678139   R-squared       =    0.5112
-------------+----------------------------------   Adj R-squared   =    0.4093
       Total |  19.9720967        29  .688692988   Root MSE        =    .63779

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.013213   .0294789    -0.45   0.658    -.0740544    .0476285
          d1 |   2.784902   1.515906     1.84   0.079    -.3437738    5.913579
          d2 |  -1.223434   1.509929    -0.81   0.426    -4.339775    1.892908
         xd1 |  -.0724738   .0377859    -1.92   0.067      -.15046    .0055123
         xd2 |     .03022   .0416451     0.73   0.475    -.0557312    .1161712
       _cons |   5.427064   .9848334     5.51   0.000     3.394468    7.459661
------------------------------------------------------------------------------
```

A3 OLS results for the reduced model (2).

```
      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(3, 26)        =      8.59
       Model |  9.93900849         3  3.31300283   Prob > F        =    0.0004
    Residual |  10.0330882        26  .385888007   R-squared       =    0.4976
-------------+----------------------------------   Adj R-squared   =    0.4397
       Total |  19.9720967        29  .688692988   Root MSE        =     .6212

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d1 |   3.240801   1.138985     2.85   0.009     .8995829    5.582018
         xd1 |  -.0856868   .0230232    -3.72   0.001    -.1330116    -.038362
         xd2 |  -.0024108   .0070316    -0.34   0.734    -.0168645     .012043
       _cons |   4.971166   .1934596    25.70   0.000     4.573504    5.368828
------------------------------------------------------------------------------
```