UNIVERSITY OF OSLO
DEPARTMENT OF ECONOMICS

Exam: **ECON4130 – Statistics 2**

Date of exam: Monday, November 21st 2016 **Grades are given:** December 13th 2016

Time for exam: 09.00 a.m. – 12.00 noon

The problem set covers 6 pages + 3 pages Appendix

Resources allowed:
- Open book exam. All written and printed resources – as well as calculator - is allowed

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

Problem 1

Let Y be a continuous random variable (rv) with probability density function (pdf) given by

(1) $$f(y)=\begin{cases}\frac{1}{\beta}y^{\frac{1}{\beta}-1} & \text{for } 0<y\leq 1\\ 0 & \text{otherwise}\end{cases}$$

where $\beta>0$ is parameter in the distribution of Y.

A. Assume $\beta=1/2$ (so that $1/\beta=2$).

i. Find the cumulative distribution function (cdf) of Y and make sketches of the graphs of both the pdf and of the cdf.

ii. What are the probabilities $P(Y>0.6)$ and $P(Y\leq 2)$?

iii. Find the median of Y.

B. Now let $\beta>0$ be arbitrary.

i. Show that $E\left(Y^r\right)=\frac{1}{1+\beta r}$, where r is any real number such that $r>-\frac{1}{\beta}$.

ii. Show that $X=-\ln(Y)$ is exponentially distributed with parameter $\lambda=\frac{1}{\beta}$ (written in short, $X\sim\exp\left(\frac{1}{\beta}\right)$)

iii. Explain why $E\left(-\ln(Y)\right)=\beta$ and $\mathrm{Var}\left(-\ln(Y)\right)=\beta^2$.

C. *Y* represents the percentage of the bus capacity used by the passengers on a particular bus tour, as assessed by the driver and expressed as a share between 0 and 1. For example, if the driver thinks that 25% of the full capacity of the bus was used during a particular tour, *Y* gets the value 0.25. **Table 1** gives the capacity shares assessed by the driver from $n=30$ similar tours along a given route. We assume that they are observations of rv's, $Y_1, Y_2, \ldots, Y_{30}$ that are independent and identically distributed (iid) with a common pdf as given in (1).

Table 1 **30 independent observations of *Y*:** $y_1, y_2, \ldots, y_{30}$

0.11	0.54	0.81	0.57	0.91	0.79	0.10	0.30	0.30	0.35
0.73	0.47	0.25	0.36	0.43	0.10	0.73	0.86	0.51	0.26
0.88	0.98	0.82	0.75	0.46	0.61	0.38	0.66	0.64	0.72

To help the calculations below, we give $\sum_{i=1}^{30} y_i = 16.38$, $\sum_{i=1}^{30} \ln(y_i) = -22.8598$

i. Show that the maximum likelihood estimator (mle) for β is (for *n* observations)

$$\hat{\beta} = -\frac{1}{n}\sum_{i=1}^{n}\ln(Y_i)$$

ii. Also find the moment method estimator (mme), $\tilde{\beta}$, for β based on $E(Y_i)$.

iii. Calculate both the mme- and the mle estimates for β.

D. **i.** Develop an approximate 95% confidence interval (CI) for β based on the mle, $\hat{\beta}$. If you need Slutsky's lemma in your development, explain how it is used. Also calculate the observed value of the CI.

ii. The bus company has earlier operated with a mean capacity share of $E(Y)=0.70$ for the bus route in question. Is there evidence in the data against this hypothesis (considering the alternative, $E(Y)\neq 0.70$), using a level of significance approximately 5%? If yes, does the evidence indicate that the mean capacity share has increased or gone down?

> **[Hint.** Use, for example, the information in the CI calculated in part **i.** to answer the question.]

E. Show that, in this case, the mle, $\hat{\beta}$, is unbiased and has the smallest variance among all estimators of β that are unbiased.
[Hint. Remember the results obtained in section **1Biii**.]

Problem 2

Introduction. As part of an investigation of how people project their self-image through objects they own, a data set[1] (shown in the appendix **A1**) was collected based on a random sample of $n = 40$ people owning a car.

The main question behind the data was how a person's extroversion affects the amount of time spent looking after his or her car. Here the term "extroversion" means the degree to which the person is extrovert - measured by a score from a psychological test.

However, since it is known that extroversion is related to both gender and age, it is reasonable that the latter two variables should be controlled for.

The following variables were observed:

Variable name	Description
Y	Time respondent spends looking after the car (in minutes per week)
X	Extroversion score
G	Gender (0=female, 1=male)
A	Age (in years)

We assume that, for the data, (Y_i, X_i, G_i, A_i), $i = 1,2,\ldots,n$ $(n = 40)$ are *iid* (independent and identically distributed) vectors.

In this problem we will focus mainly on the three following regression models - estimated by OLS in appendices **A4, A5**, and **A6**, respectively:

Model 1: Regression function: $\mu_1(x) = E(Y \mid X = x) = \alpha_0 + \alpha_1 x$

Variance function: $\mathrm{Var}(Y \mid X = x) = \tau_1^2$ (constant).

Model 1 for the data:

$$Y_i = \alpha_0 + \alpha_1 X_i + e_{1i}, \quad i = 1,2,\ldots,n$$

where, for given fixed values of all the X_i's, the error terms, $e_{11}, e_{12}, \ldots, e_{1n}$ are *iid* and normally distributed, $e_{1i} \sim N(0, \tau_1^2)$, $i = 1,2,\ldots,n$.

Model 2:

Regression function:

$$\mu_2(x,g,a) = E(Y \mid X = x, G = g, A = a) = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 a$$

Conditional variance function: $\mathrm{Var}(Y \mid x, g, a) = \tau_2^2$ (constant).

Model 2 for the data:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 A_i + e_{2i}, \quad i = 1,2,\ldots,n$$

[1] Described in Miles, J., & Shevlin, M. 2001. *Applying Regression and Correlation*. London: Sage Publications

where, for given fixed values of all the X_i's, G_i's, and A_i's, the error terms, $e_{21}, e_{22}, \ldots, e_{2n}$ are *iid* and normally distributed, $e_{2i} \sim N(0, \tau_2^2), \quad i = 1, 2, \ldots, n$.

Model 3 (with first order interactions as product terms):

Regression function:

$$\mu_3(x, g, a) = E(Y \mid X = x, G = g, A = a) =$$
$$= \gamma_0 + \gamma_1 x + \gamma_2 g + \gamma_3 a + \gamma_4 xg + \gamma_5 xa + \gamma_6 ga$$

Conditional variance function: $\mathrm{Var}(Y \mid x, g, a) = \tau_3^2$ (constant).

Model 3 for the data:

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 G_i + \gamma_3 A_i + \gamma_4 X_i G_i + \gamma_5 X_i A_i + \gamma_6 G_i A_i + e_{3i}, \quad i = 1, 2, \ldots, n$$

where, for given fixed values of all the X_i's, G_i's, and A_i's, the error terms, $e_{31}, e_{32}, \ldots, e_{3n}$ are *iid* and normally distributed, $e_{3i} \sim N(0, \tau_3^2), \quad i = 1, 2, \ldots, n$.

Questions.

A. Since $(Y_i, X_i, G_i, A_i), \quad i = 1, 2, \ldots, n \quad (n = 40)$ are *iid* vectors they have a common expected value, $\mu' = (\mu_Y, \mu_X, \mu_G, \mu_A)$, and a common covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} & \sigma_{YG} & \sigma_{YA} \\ \sigma_{XY} & \sigma_X^2 & \sigma_{XG} & \sigma_{XA} \\ \sigma_{GY} & \sigma_{GX} & \sigma_G^2 & \sigma_{GA} \\ \sigma_{AY} & \sigma_{AX} & \sigma_{AG} & \sigma_A^2 \end{pmatrix}$$

μ and Σ are consistently estimated (which you can take as facts that you do not need to justify) by the sample means, $\hat{\mu}' = (\bar{Y}, \bar{X}, \bar{G}, \bar{A})$ and the sample covariance matrix

$$\hat{\Sigma} = \begin{pmatrix} S_Y^2 & S_{YX} & S_{YG} & S_{YA} \\ S_{XY} & S_X^2 & S_{XG} & S_{XA} \\ S_{GY} & S_{GX} & S_G^2 & S_{GA} \\ S_{AY} & S_{AX} & S_{AG} & S_A^2 \end{pmatrix}$$

respectively, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \ldots$ etc, and

$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}), \quad \ldots$ etc.

Observed values for $\hat{\mu}, \hat{\Sigma}$ are given in appendices **A2** and **A3**.

i. Consider the regression in model 1 that has the constant conditional error variance, $\tau_1^2 = \text{var}(Y \mid X = x) = \text{var}(e_1 \mid x) = \sigma_Y^2(1-\rho^2)$, where $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$. Set up a consistent estimator for τ_1^2 based on $\hat{\Sigma}$, and explain why it is consistent.

ii. Calculate the estimate for τ_1^2 based on the estimator you obtained in **i.** and compare with the corresponding estimate from the OLS output in appendix **A4** for model 1.

It can be shown (that you do not have to do) that these two ways of estimating the error variance τ_1^2 are algebraically equivalent (the two estimators are in fact equal). In practice, however, the two estimates will be somewhat different due to rounding errors. How big is the difference between the two estimates in this case?

B. **i.** Consider regression model 3 (with interaction terms). Find the ceteris paribus (cet. par.) effect on the conditional expected *Y* for a unit change in *X* (i.e., for *X* going from $X = x$ to $X = x+1$, while keeping *G* and *A* fixed), expressed by the regression coefficients in model 3.

ii. The cet. par. effect of *X* in **i.** varies with the values of G and A as they vary in the population. Estimate the population mean value of this effect (as expressed by $E(G)$ and $E(A)$), and compare with the cet. par. effect of a unit change of *X* estimated in model 2.

C. Test whether there is strong evidence in the data that (first order) interactions should be included in the regression model, i.e., if there is evidence that some of the coefficients, $\gamma_4, \gamma_5, \gamma_6$ are different from 0. In other words, test model 2 versus model 3 using the output in appendix **A5** and **A6**. State the null- and alternative hypotheses you are using, and formulate a conclusion based on a 5% level of significance.

[Hint. If the table of quantiles of the F-distribution you are using should happen not to contain exactly the critical value you are looking for (e.g., when the degrees of freedom you need are not represented in the table), just guess roughly on a critical value based on the two nearest values in the table.]

D. We want to choose one of the three regression models in the introduction for interpretation purposes. Now, all three models may be true at the same time, but they measure different things. So the question is not of which one of the three is the true model, but rather which one is most suitable (or preferable) for studying the influence of *X* on *Y*.

i. In the special case that $E(G \mid X = x) = c_0 + c_1 x$, and $E(A \mid X = x) = d_0 + d_1 x$, show that the relationship between the regression coefficient α_1 in model 1 and the regression coefficients in model 2 is given by

$$\alpha_1 = \beta_1 + \beta_2 c_1 + \beta_3 d_1$$

[**Hint.** Use "the law of total expectation" on $(Y \mid X = x)$ in model 2, i.e., $E(Y \mid x) = E\left[E(Y \mid X, G, A) \mid X = x\right] = E\left[E(Y \mid x, G, A) \mid x\right] = \ldots$etc, where the model 2 regression function is used on the right side.]

ii. Which of the three models do you think is best for describing the influence of *X* on *Y*? Give a reason for your answer, i.e., present a short argument for your choice based on your common sense and the results so far in problem 2. In addition, appendix **A7,** that shows the results of regressing *X* on *G* and *A*, may be relevant for your discussion.

Appendix - Stata Output for Problem 3

A1 Data set

Variable name	Variable label
Y	Minutes per week car care
X	Extroversion score
A	Age
G	Gender (Male=1 Female=0)

Obs.	Y	X	A	G		Obs.	Y	X	A	G
1	46	40	55	1		21	54	60	49	0
2	79	45	43	1		22	73	47	49	1
3	33	52	57	0		23	19	18	48	0
4	63	62	26	1		24	36	16	29	0
5	20	31	22	0		25	31	36	58	0
6	18	28	32	0		26	71	24	24	1
7	11	2	26	0		27	15	12	21	0
8	97	83	29	1		28	40	32	29	1
9	63	55	40	1		29	61	46	45	1
10	46	32	30	0		30	45	26	28	1
11	21	47	34	0		31	42	40	37	0
12	71	45	44	1		32	57	46	44	1
13	59	60	49	1		33	34	44	22	0
14	44	13	22	1		34	26	3	38	0
15	30	7	34	0		35	47	25	24	0
16	80	85	47	1		36	42	43	34	1
17	26	61	48	0		37	44	41	26	1
18	33	26	22	1		38	59	42	26	1
19	7	3	24	0		39	59	42	26	1
20	50	29	50	0		40	27	36	25	1

A2 Sample means and standard deviations (n=40 observation units)

```
    Variable |        Mean    Std. Dev.         Min          Max
-------------+---------------------------------------------------
           Y |      44.475    20.91863           7           97
           X |      37.125    19.68624           2           85
           G |        .525    .5057363           0            1
           A |        35.4    11.32232          21           58
```

A3 Sample covariance matrix (n=40 observation units)

```
             |        Y        X        G        A
-------------+------------------------------------
           Y |  437.589
           X |  275.862  387.548
           G |  7.15449  4.08654  .255769
           A |  51.4205  87.4103 -.266667  128.195
```

A4 OLS output for model 1

```
      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =   30.93
       Model |   7658.1408     1   7658.1408           Prob > F      =  0.0000
    Residual |   9407.8342    38  247.574584           R-squared     =  0.4487
-------------+------------------------------           Adj R-squared =  0.4342
       Total |   17065.975    39  437.589103           Root MSE      =  15.735

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .7118141   .1279847     5.56   0.000     .4527227    .9709055
       _cons |    18.0489   5.363344     3.37   0.002      7.19138    28.90642
```

A5 OLS output for model 2

```
      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  3,    36) =   21.91
       Model |  11027.5184     3  3675.83946           Prob > F      =  0.0000
    Residual |  6038.45662    36  167.734906           R-squared     =  0.6462
-------------+------------------------------           Adj R-squared =  0.6167
       Total |   17065.975    39  437.589103           Root MSE      =  12.951

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .4652158   .1294747     3.59   0.001      .202629    .7278026
           G |   20.67178   4.641239     4.45   0.000     11.25891    30.08465
           A |   .1269033   .2055044     0.62   0.541     -.289879    .5436856
       _cons |    11.8588    7.24045     1.64   0.110    -2.825512    26.54311
```

A6 OLS output for model 3

```
      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  6,    33) =   11.33
       Model |  11488.8977     6  1914.81628           Prob > F      =  0.0000
    Residual |  5577.07731    33  169.002343           R-squared     =  0.6732
-------------+------------------------------           Adj R-squared =  0.6138
       Total |   17065.975    39  437.589103           Root MSE      =      13

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .6853658    .487486     1.41   0.169     -.306432    1.677164
           G |   6.510446   14.57921     0.45   0.658    -23.15117    36.17206
           A |   .5041919   .4985523     1.01   0.319    -.5101204    1.518504
          XG |   .2892336   .2652183     1.09   0.283     -.250357    .8288243
          XA |  -.0105246   .0122332    -0.86   0.396    -.0354132    .0143641
          GA |   .0962131   .4370057     0.22   0.827    -.7928816    .9853078
       _cons |   3.940182   16.71364     0.24   0.815    -30.06398    37.94435
```

A7 OLS output for regressing *X* on *G* and *A*

```
      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  2,    37) =    9.45
       Model |  5108.52787     2  2554.26393           Prob > F      =  0.0005
    Residual |  10005.8471    37  270.428301           R-squared     =  0.3380
-------------+------------------------------           Adj R-squared =  0.3022
       Total |   15114.375    39  387.548077           Root MSE      =  16.445

------------------------------------------------------------------------------
           X |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           G |   16.72462    5.21244     3.21   0.003     6.163215    27.28603
           A |   .7166445   .2328252     3.08   0.004      .244896    1.188393
       _cons |   2.975357   9.180457     0.32   0.748    -15.62602    21.57673
```