HG
Nov. 17

ECON 4130 17H

Exercises for no-seminar week 48

(The solution set will be put on the net on Thursday 30 Nov., including the "sensorveiledninger" for regular exams 20014H and 2015H)

I)

Rice chapter 9: No. 12, 33 (**Hint:** note that there are 0 parameters under H_0 here, so the DF for the Chi-square test must be equal to the number of free parameters in the full model.)

No. 40 (Remember that $Z \sim N(0,1) \quad \Rightarrow \quad Z^2 \sim \chi_1^2$ - distributed.) (See, e.g., Rice, example C, sec 2.3, p.61)

No. 41

II) An introductory exercise on F-testing

Note. An F-test is a test for several linear restrictions, tested jointly, in a regression problem. The F-test may be looked upon as a generalization of the T-test that is a test for just a single linear restriction. Note also that the F-test may be interpreted as a likelihood ratio test (LR-test). This is justified in the appendix (optional reading) of the lecture note on F-testing. (**End of note.**)

An econometric model contains a response, Y, and 6 (exogenous) explanatory variables, $X, Z_1, Z_2, U_1, U_2, U_3$. The data are observations of $n = 22$ *iid*[1] corresponding random vectors, $(Y_i, X_i, Z_{i1}, Z_{i2}, U_{i1}, U_{i2}, U_{i3})$, and the (full) regression model is (using the observed values of the explanatory variables as fixed[2])

(1) $$Y_i = \alpha + \beta x_i + \delta_1 z_{i1} + \delta_2 z_{i2} + \gamma_1 u_{i1} + \gamma_2 u_{i2} + \gamma_3 u_{i3} + e_i \quad \text{for } i = 1, 2, \ldots, 22$$

Where, $e_1, e_2, \ldots, e_n$ are iid and normal, $e_i \sim N(0, \sigma^2)$.

[1] i.e., the joint distribution for the seven variables in one vector is the same for all *i*, and two different vectors are stochastically independent.

[2] See appendix 1 in the lecture note on prediction and the iid model for a justification of this – i.e., that we may consider the explanatory variables in a regression model as fixed numbers without loosing information. The justification is based on the maximum likelihood principle.

A. Estimating (1) by OLS gives the following table of sums of squares (using Stata terminology)

Table 1 (for full model)

Source	SS	df
Model	7817	?
Residual	3743	?
Total	11560	?

Fill in the degrees of freedom (df's) in the table. Estimate the error term variance, σ^2, using an unbiased estimator.

B. A submodel of interest is assuming both $\delta_1 = \delta_2$ and $\gamma_1 = \gamma_2 = \gamma_3$. We want to check if there is evidence in the data against this submodel using an appropriate F-test. We then need to re-estimate the model assuming the submodel (that we call the "reduced model") to be true. Using OLS for the reduced model implies that we must regress the response Y on a modified set of explanatory variables.

Write up the corresponding (to (1)) regression model in the reduced case.

[**Hint:** Introduce two new parameters, δ for the common value of δ_1, δ_2, and γ for the common value of $\gamma_1, \gamma_2, \gamma_3$, and substitute in (1). Define new regressor (i.e., explanatory) variables whenever necessary.]

C. Estimating the reduced model by OLS gives the following table of sums of squares (using Stata terminology)

Table 2 (for the reduced model)

Source	SS	df
Model	5332	?
Residual	6228	?
Total	11560	?

Use this information to perform an F-test for testing the sub-model against the more general model in (1).

Calculate the P-value, either approximately using the quantile table 5 in the back of Rice's book, or exactly using (e.g.) the "F.dist" function in Excel, or the F(df1,df2,f) – function (or Ftail(df1,df2,f)-function) in STATA.

III) Problem 2 of regular exam 2014, and Problem 2 of regular exam 2015

(Both problems are reproduced below for convenience)

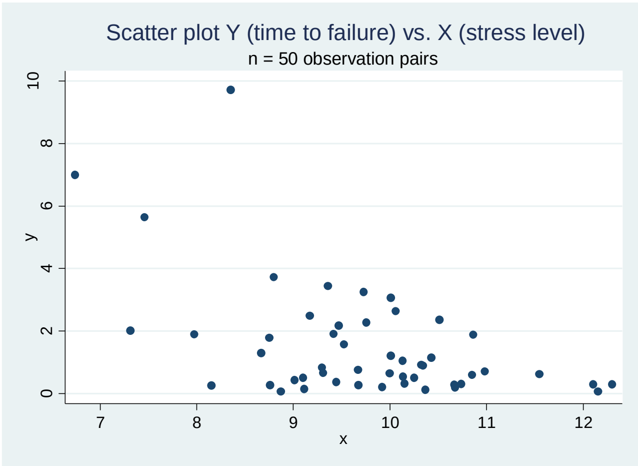
Problem 2 of regular exam 2014:

Introduction. In this problem we will look at data using a similar but more general model than the one discussed in **problem 1 (of the regular exam 2014)**.

Let Y be the time to failure of a certain component in a randomly chosen machine of a special type, and X a measure of the average intensity (stress level) of the use of the machine under regular conditions.

The data[3] consist of observations of $n = 50$ random pairs, $(X_i, Y_i),\quad i = 1, 2, \ldots, n$, which are assumed to be *iid* and representative for (X, Y). The index, *i*, refers to machine number *i* drawn from the population of machines in use. A scatter plot of the data is given in figure 1.

Figure 1



Model M1:

(a) $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are *iid* pairs, distributed as (X, Y).

(b) $X \sim N(\mu, \sigma^2),\quad$ where μ, σ^2 are unknown parameters.

(c) Given that $X = x$ is fixed, then Y is exponentially distributed with parameter $\lambda(x) = e^{-\alpha + \beta x}$, where α, β are unknown parameters.

Questions:

[3] Simulated data.

A. **i.** Suppose the true values of α and β are 5 and 0.5 respectively. Using model M1 calculate the best prediction of *Y* for a stress level $X = 12$. Describe the criterion of "best prediction" that you are using. Choose the criterion yourself (you do not have to prove that your prediction is best according to the criterion you choose).

ii. Now suppose α and β are unknown. Consider the regression function, $\mu(x) = E(Y \mid x)$, in **model M1**. Show that the relative effect of a unit change in *x* on the regression, i.e., $\theta = \dfrac{\mu(x+1) - \mu(x)}{\mu(x)}$, is a constant depending on β only.

B. **Introduction.** All machines in the population are produced at 3 different factories, called factory 1, 2, and 3. The data contains information, for each machine in the sample, which factory has produced it. We want to test the null hypothesis that there is no difference between the regression functions of the three factories, against the alternative that there may be differences.

In other words, we want to test the model (M1) against a more general model where there may be different regression functions for the three factories. We assume that possible differences may occur among the alphas only (the three betas being equal).

To formulate a more general model, dummy variables, D_1, D_2, D_3, are introduced for the factories, where $D_j = 1$ if the corresponding randomly drawn machine is produced at factory *j* and $D_j = 0$ otherwise ($j = 1,2,3$). In this way the three factories are characterized by the three vectors, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ respectively. The model is

Model M2:

(**a**) $(X_i, Y_i, D_{1i}, D_{2i}, D_{3i})$, $\;\mathrm{i} = 1, 2, \ldots, n$ are *iid* vectors, distributed as (X, Y, D_1, D_2, D_3).

(**b**) $X \sim N(\mu, \sigma^2)$, where μ, σ^2 are unknown parameters.

(**c**) Given that $X = x$, $D_1 = d_1$, $D_2 = d_2$, $D_3 = d_3$ are fixed, then Y is exponentially distributed with parameter $\lambda(x, d_1, d_2, d_3) = e^{-\alpha_1 d_1 - \alpha_2 d_2 - \alpha_3 d_3 + \beta x}$, where $\alpha_1, \alpha_2, \alpha_3, \beta$ are unknown parameters.

Questions of B:

i. The population consists of all machines in use. The relative frequencies of machines in the population from the three factories are p_1, p_2, p_3 respectively. Let U_1, U_2, U_3 denote the (absolute) frequencies in the sample of machines from the three factories respectively. Justify and write up the joint probability mass function (pmf) for U_1, U_2, U_3.

ii. Stata output for maximum likelihood estimation of both model M1 and M2 has been given at the end of the exam set. Use the output to test model M1 against M2 (i.e., test $H_0 : \alpha_1 = \alpha_2 = \alpha_3$) , and formulate a conclusion using level of significance 5%.

[**Hint.** You may assume that the conditions for "good behavior" of the mle estimators are fulfilled here.]

C. Assume the model M1 to be true.

(i) Earlier people used to believe that the true value of α was 3. Calculate approximately the p-value of testing $H_0 : \alpha \le 3$ against $H_1 : \alpha > 3$, using the Stata output.

(ii) Develop and calculate an approximate 95% confidence interval for the relative effect of a unit change in *x* on the regression, $\dfrac{\mu(x+1) - \mu(x)}{\mu(x)}$, using the Stata output. Calculate, in addition, the mle estimate of the relative effect and state the reason (based on general mle theory) why it is mle.

Stata output.

```
Model M1 **** (Reduced model)
.
< Iterations information omitted>

Maximum likelihood estimation

Log likelihood = -139.60307                        Number of obs   =         50

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /alpha |   5.318701   1.254468     4.24   0.000     2.859989    7.777413
       /beta |   .5226965    .128122     4.08   0.000     .2715819    .7738111
         /my |   9.728779   .1626116    59.83   0.000     9.410066    10.04749
      /sigma |   1.149838   .1149838    10.00   0.000     .9244735    1.375202
------------------------------------------------------------------------------
```

```
Model M2 ***** (Full model)

< Iterations information omitted>

Maximum likelihood estimation
```

```
Log likelihood = -136.97421                       Number of obs   =         50

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     /alpha1 |   4.735735   1.313504     3.61   0.000     2.161315    7.310154
     /alpha2 |   5.562711   1.276714     4.36   0.000     3.060397    8.065024
     /alpha3 |   5.548397   1.361421     4.08   0.000     2.880062    8.216733
       /beta |   .5322517   .1329083     4.00   0.000     .2717561    .7927472
         /my |   9.728779   .1626116    59.83   0.000     9.410066    10.04749
      /sigma |   1.149838   .1149838    10.00   0.000     .9244735    1.375202
------------------------------------------------------------------------------
```

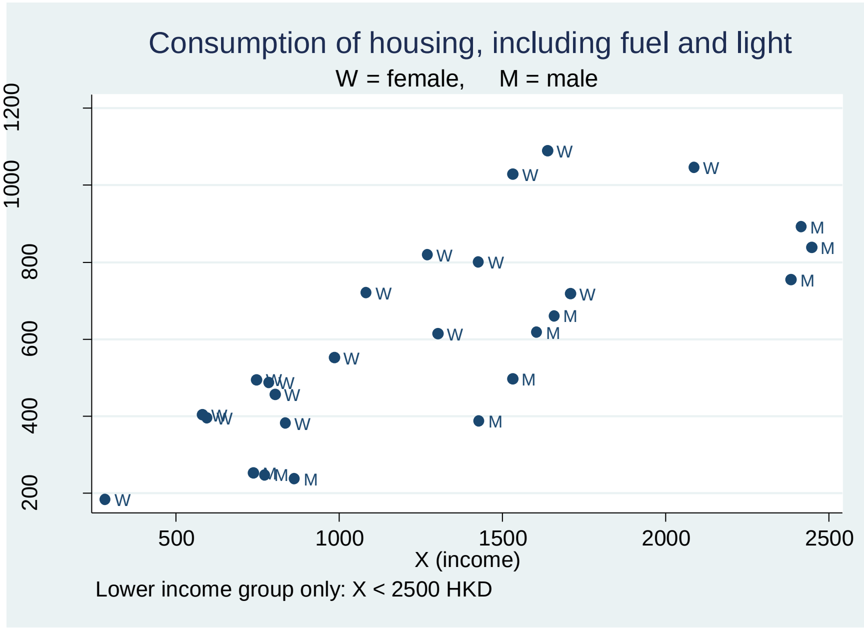
Problem 2 of regular exam 2015:

Introduction. In this problem we will look at the effect of gender on the consumption of housing, including fuel and light, for lower income (< 2500 HKD) consumers in Hong Kong. The original data (40 consumers) are given in table 1 while the lower income data from the sample (26 consumers) are plotted in figure 1.

Table 1 Consumption of housing, including fuel and light, and income for a sample of 40 Hong Kong consumers.

Women	Consumer no.	1	2	3	4	5	6	7	8	9	10
	Consumption	820	184	921	488	721	614	801	396	864	845
	Income	1271	284	3128	786	1084	1303	1428	596	2899	3258
	Consumer no.	11	12	13	14	15	16	17	18	19	20
	Consumption	404	781	457	1029	1047	552	718	495	382	1090
	Income	581	3186	804	1533	2088	986	1709	748	836	1639
Men	Consumer no.	21	22	23	24	25	26	27	28	29	30
	Consumption	497	839	798	892	1585	755	388	617	248	1641
	Income	1532	2448	3358	2416	6582	2385	1429	2972	773	10615
	Consumer no.	31	32	33	34	35	36	37	38	39	40
	Consumption	1180	619	253	661	1981	1746	1865	238	1199	1524
	Income	4004	1606	738	1659	5371	6748	9731	864	2899	5637

Figure 1 **Consumption of housing, including fuel and light, vs. income for the lower income group (26 consumers) of the sample.**



For a randomly selected consumer we define Y as the consumption (in HKD) of housing, including fuel and light, for the period in question, X the income (in HKD) for the same period, and M a dummy variable for gender (M = 0 for female and M = 1 for male).

The population of interest consists of consumers in Hong Kong with income $X < 2500$ HKD.

Model. Assume that the conditional distribution of Y given fixed values $M = m$ and $X = x$, is normal with expectation

(1) $$E(Y \mid x, m) = \beta_0 + \beta_1 x + \beta_2 m + \beta_3 m \cdot x$$

and constant variance

(2) $$\text{var}(Y \mid x, m) = \sigma^2$$

Questions.

A. **i)** The ceteris paribus (cet. par.) effect of gender is defined as the expected difference in consumption between males and females for a given income being the same for both genders. Explain why the cet. par. effect of gender is $\beta_2 + \beta_3 x$ based on the model assumption (1), where the common income for both genders is x.

ii) Find the cet. par. effect of a unit change in income *x* on the expected consumption. In what way does this effect depend on the gender?

B. **Introduction.** The corresponding model for the random mechanism behind the data is specified as

(3) $Y_i = E(Y_i \mid x_i, m_i) + e_i = \beta_0 + \beta_1 x_i + \beta_2 m_i + \beta_3 m_i \cdot x_i + e_i, \quad i = 1, 2, \ldots, n \quad (n = 26)$

where the regressors, $x_i, m_i, \quad i = 1, 2, \ldots, n$, are considered fixed numbers due to their exogeneity, and where the error terms, $e_1, e_2, \ldots, e_n$, are assumed to be iid and normally distributed random variables, $e_i \sim N(0, \sigma^2)$.
The model (3) reduces to two simple regressions, one for women (16 observation units) and one for men (10 units). Model (3) also assumes that the error variances of the two regressions are the same, i.e., $\sigma_W^2 = \sigma_M^2 = \sigma^2$, where σ_W^2, σ_M^2 are the error variances of the two regressions respectively. We should check if there is any evidence in the data against this assumption.

Questions. The two simple regressions are estimated by Stata in the appendix, see A1 and A2.

i) Use the outputs in A1 and A2 to set up unbiased estimates for σ_W^2 and σ_M^2.

ii) **Specification test**: Use the outputs in A1 and A2 to test $H_0: \sigma_W^2 = \sigma_M^2$ against $H_1: \sigma_W^2 \neq \sigma_M^2$ at the 5% level of significance.
(**Hint:** If you don't find the right critical level in the Rice table you use, e.g., if the degrees of freedom needed are not represented in the table, you can guess roughly the critical value from the nearest values in the table.)

C. We want to test if there is evidence in the data to claim that gender has an effect on the consumption in question, i.e., if the cet. par. effect of gender derived in section **A i)** is different from zero. The full (in (3)) and reduced model that can be used to test this, have been estimated in appendix **A3** and **A4**. Set up a proper null-hypothesis, perform a test at 1% level of significance, and state a conclusion.

D. Let the population mean income in the lower income group be μ_0 and μ_1 for women and men respectively, or, in other words,

$$\mu_m = E(X \mid m) = \begin{cases} \mu_0 & \text{for women} \\ \mu_1 & \text{for men} \end{cases}$$

Explain why the model assumption (1) implies that

$$E(Y\,|\,m) = \beta_0 + \beta_1\mu_m + (\beta_2 + \beta_3\mu_m)\cdot m = \begin{cases} \beta_0 + \beta_1\mu_0 & \text{for women} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\mu_1 & \text{for men} \end{cases}$$

Hint: Use the law of total expectation on the relation (1), where the outer expectation is referring to the conditional distribution of *X* given $M = m$ fixed.

Appendix: Stata Outputs for Problem 2 (regular exam 2015)

A1. Simple regression Y on X for 16 lower income WOMEN

```
      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  1,    14) =   70.86
       Model |  890213.453     1  890213.453           Prob > F      =  0.0000
    Residual |  175882.297    14  12563.0212           R-squared     =  0.8350
-------------+------------------------------           Adj R-squared =  0.8232
       Total |  1066095.75    15    71073.05           Root MSE      =  112.08

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |    .498313   .0591973     8.42   0.000     .3713473    .6252786
       _cons |   86.86374   71.14858     1.22   0.242    -65.73479    239.4623
------------------------------------------------------------------------------
```

A2. Simple regression Y on X for 10 lower income MEN

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =  124.45
       Model |  530115.571     1  530115.571           Prob > F      =  0.0000
    Residual |  34076.4291     8  4259.55364           R-squared     =  0.9396
-------------+------------------------------           Adj R-squared =  0.9321
       Total |      564192     9       62688           Root MSE      =  65.265

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .3638185   .0326123    11.16   0.000     .2886144    .4390226
       _cons |  -37.65232   55.65846    -0.68   0.518     -166.001    90.69631
------------------------------------------------------------------------------
```

A3. Full model regression for problem 2C

```
      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  3,    22) =   51.69
```

```
       Model |  1479883.74     3  493294.579           Prob > F      =  0.0000
    Residual |  209958.726    22  9543.57845           R-squared     =  0.8758
-------------+------------------------------           Adj R-squared =  0.8588
       Total |  1689842.46    25  67593.6985           Root MSE      =  97.691

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |    .498313   .0515954     9.66   0.000     .3913107    .6053152
           M |  -124.5161    103.857    -1.20   0.243    -339.9023    90.87012
          XM |  -.1344945   .0710282    -1.89   0.072     -.281798     .012809
       _cons |   86.86374   62.01187     1.40   0.175    -41.74101    215.4685
------------------------------------------------------------------------------
```

A4. Reduced model regression for problem 2C

```
      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  1,    24) =   32.17
       Model |  967783.563     1  967783.563           Prob > F      =  0.0000
    Residual |  722058.898    24  30085.7874           R-squared     =  0.5727
-------------+------------------------------           Adj R-squared =  0.5549
       Total |  1689842.46    25  67593.6985           Root MSE      =  173.45

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .3277504   .0577876     5.67   0.000     .2084826    .4470182
       _cons |   176.9169   81.91226     2.16   0.041     7.858315    345.9755
```