ECON 124: Midterm #2

Due on Jul 9, 2025

Dr. Deniz Baglan

Alejandro Ouslan

Problem 1

Use the data in **FERTIL2.XLSX** to answer this question.

(a) Estimate the model

$$children = \beta_0 + \beta_1 age + \beta_2 age^2 \beta_3 educ + \beta_4 electricity + \beta_5 urban + \epsilon$$

And report the usual and heteroskedasticity-robust standard errors. Are the robust standard errors always bigger than the non robust ones?

Answer: It seems that the robust standard errors are generally larger than the non robust ones, but not neccesarily always the case.

Table 1: OLS Regression Results: Children ~Age, Age², Education, Electricity, Urban

Variable	**Coef.**	**Std. Err.**	**t**	**P**$>|t|$	**[0.025**	**0.975]**
const	-4.2225	0.240	-17.580	0.000	-4.693	-3.752
age	0.3409	0.017	20.652	0.000	0.309	0.373
age²	-0.0027	0.000	-10.086	0.000	-0.003	-0.002
education	-0.0752	0.006	-11.948	0.000	-0.088	-0.063
electricity	-0.3100	0.069	-4.493	0.000	-0.445	-0.175
urban	-0.2000	0.047	-4.301	0.000	-0.291	-0.109
Model statistics:						
R-squared		0.573				
Adj. R-squared		0.573				
F-statistic		1170	(Prob F-statistic = 0.000)			
No. Observations		4358				
Df Residuals		4352				
Df Model		5				
Log-Likelihood		-7806.3				
AIC		1.562e+04				
BIC		1.566e+04				
Durbin-Watson: 1.883						
Omnibus: 203.155, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 715.135, Prob(JB): 5.13×10^{-156}						
Skew: 0.014, Kurtosis: 4.984						
Cond. No.: 1.07e+04						

Problem 1 continued on next page. . .

Table 2: OLS Regression Results: Children ~Age, Age², Education, Electricity, Urban

Variable	**Coef.**	**Std. Err.**	**z**	**P>** $\lvert z \rvert$	**[0.025**	**0.975]**
const	-4.2225	0.244	-17.316	0.000	-4.700	-3.745
age	0.3409	0.019	17.780	0.000	0.303	0.379
age²	-0.0027	0.000	-7.821	0.000	-0.003	-0.002
education	-0.0752	0.006	-11.927	0.000	-0.088	-0.063
electricity	-0.3100	0.064	-4.848	0.000	-0.435	-0.185
urban	-0.2000	0.045	-4.399	0.000	-0.289	-0.111
Model statistics:						
R-squared		0.573				
Adj. R-squared		0.573				
F-statistic		1161	(Prob F-statistic = 0.000)			
No. Observations		4358				
Df Residuals		4352				
Df Model		5				
Log-Likelihood		-7806.3				
AIC		1.562e+04				
BIC		1.566e+04				
Durbin-Watson: 1.883						
Omnibus: 203.155, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 715.135, Prob(JB): 5.13×10^{-156}						
Skew: 0.014, Kurtosis: 4.984						
Cond. No.: 1.07e+04						

Python Code

```python
import polars as pl
import statsmodels.formula.api as smf

# Question
# 1a
df = pl.read_excel("data/fertil2.xlsx")
df = df.select(
    pl.col(
        [
            "children",
            "age",
            "educ",
            "electric",
            "urban",
            "spirit",
            "protest",
            "catholic",
        ]
    )
)
df = df.with_columns(age2=pl.col("age") ** 2)
df = df.to_pandas()
```

Problem 1 continued on next page. . .

```
model = smf.ols("children ~ age + age2 + educ + electric + urban", data=df).fit()
print(model.summary())

model = smf.ols("children ~ age + age2 + educ + electric + urban", data=df).fit(
    cov_type="HC1"
)
print(model.summary())
```

(b) Add the three religious dummy variables and test whether they are jointly significant. What are the p-values for the nonrobust and robust tests?
Answer: The p-values for the non-robust test is 0.0864, while the p-value for the robust test is 0.0911. It seems that robust tests are less likely to report something is significant, especially assuming standard errors are greater than non-robust ones.

Table 3: OLS Regression Results: Children ~ Age, Age², Education, Electricity, Urban, Spirit, Protest, Catholic

Variable	**Coef.**	**Std. Err.**	**t**	**P>** $\|t\|$	**[0.025**	**0.975]**
Intercept	-4.3147	0.243	-17.731	0.000	-4.792	-3.838
age	0.3419	0.017	20.696	0.000	0.309	0.374
age²	-0.0028	0.000	-10.139	0.000	-0.003	-0.002
education	-0.0762	0.006	-11.796	0.000	-0.089	-0.064
electricity	-0.3057	0.069	-4.429	0.000	-0.441	-0.170
urban	-0.2034	0.047	-4.366	0.000	-0.295	-0.112
spirit	0.1405	0.056	2.517	0.012	0.031	0.250
protest	0.0754	0.065	1.156	0.248	-0.052	0.203
catholic	0.1174	0.083	1.407	0.160	-0.046	0.281
Model statistics:						
R-squared		0.574				
Adj. R-squared		0.573				
F-statistic		732.6	(Prob F-statistic = 0.000)			
No. Observations		4358				
Df Residuals		4349				
Df Model		8				
Log-Likelihood		-7803.0				
AIC		1.562e+04				
BIC		1.568e+04				
Durbin-Watson: 1.887						
Omnibus: 202.228, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 709.030, Prob(JB): 1.09×10^{-154}						
Skew: 0.016, Kurtosis: 4.976						
Cond. No.: 1.09e+04						

Problem 1 continued on next page. . .

Table 4: F-test Results

Statistic	**Value**	**Notes**
F-statistic	2.196	-
p-value	0.0864	-
Degrees of Freedom (denominator)	4350	-
Degrees of Freedom (numerator)	3	-

Table 5: OLS Regression Results: Children $\sim$ Age, Age², Education, Electricity, Urban, Spirit, Protest, Catholic (Robust Standard Errors)

Variable	**Coef.**	**Std. Err.**	**z**	**P$>|z|$**	**[0.025**	**0.975]**
Intercept	-4.3147	0.248	-17.389	0.000	-4.801	-3.828
age	0.3419	0.019	17.807	0.000	0.304	0.379
age²	-0.0028	0.000	-7.861	0.000	-0.003	-0.002
education	-0.0762	0.006	-11.860	0.000	-0.089	-0.064
electricity	-0.3057	0.064	-4.772	0.000	-0.431	-0.180
urban	-0.2034	0.046	-4.456	0.000	-0.293	-0.114
spirit	0.1405	0.056	2.487	0.013	0.030	0.251
protest	0.0754	0.066	1.140	0.254	-0.054	0.205
catholic	0.1174	0.079	1.483	0.138	-0.038	0.272
Model statistics:						
R-squared		0.574				
Adj. R-squared		0.573				
F-statistic		727.9	(Prob F-statistic = 0.000)			
No. Observations		4358				
Df Residuals		4349				
Df Model		8				
Log-Likelihood		-7803.0				
AIC		1.562e+04				
BIC		1.568e+04				
Durbin-Watson: 1.887						
Omnibus: 202.228, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 709.030, Prob(JB): 1.09×10^{-154}						
Skew: 0.016, Kurtosis: 4.976						
Cond. No.: 1.09e+04						

Table 6: F-test Results

Statistic	**Value**	**Notes**
F-statistic	2.156	-
p-value	0.0911	-
Degrees of Freedom (denominator)	4350	-
Degrees of Freedom (numerator)	3	-

Problem 1 continued on next page. . .

Python Code

```
# assumes that the code in 1a ran
model = smf.ols(
    "children ~ age + age2 + educ + electric + urban + spirit + protest + catholic",
    data=df,
).fit()
print(model.summary())
print(model.f_test("spirit = protest = catholic = 0"))

model = smf.ols(
    "children ~ age + age2 + educ + electric + urban + spirit + protest + catholic",
    data=df,
).fit(cov_type="HC1")
print(model.summary())
print(model.f_test("spirit = protest = catholic = 0"))
```

(c) From the regression in part (b), obtain the fitted values $\hat{y}$ and the residuals, $\hat{\epsilon}$. Regress $\hat{\epsilon}^2 \sim \hat{y}$, and $\hat{\epsilon}^2 \sim \hat{y}^2$ and test the joint significance of the two regressors.

Table 7: OLS Regression Results: $\hat{u}^2 \sim \hat{y} + \hat{y}^2$

Variable	**Coef.**	**Std. Err.**	**t**	**P>** $\lvert t\rvert$	**[0.025**	**0.975]**
Intercept	0.3126	0.111	2.807	0.005	0.094	0.531
$\hat{y}$	-0.1489	0.102	-1.462	0.144	-0.348	0.051
$\hat{y}^2$	0.2668	0.020	13.607	0.000	0.228	0.305
Model statistics:						
R-squared		0.250				
Adj. R-squared		0.250				
F-statistic		726.1	(Prob F-statistic = 7.19e-273)			
No. Observations		4358				
Df Residuals		4355				
Df Model		2				
Log-Likelihood		-11803				
AIC		2.361e+04				
BIC		2.363e+04				
Durbin-Watson: 1.947						
Omnibus: 3446.975, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 119444.435, Prob(JB): 0.000						
Skew: 3.503, Kurtosis: 27.672						
Cond. No.: 31.4						

Table 8: F-test Results

Statistic	**Value**	**Notes**
F-statistic	726.11	-
p-value	7.19×10^{-273}	-
Degrees of Freedom (denominator)	4358	-
Degrees of Freedom (numerator)	2	-

Problem 1 continued on next page...

Python Code

```python
# assums that the code in problem 1 a ran
df["yhat"] = model.fittedvalues
df["u_hat"] = model.resid
df["u_hat2"] = df["u_hat"] ** 2
df["yhat2"] = df["yhat"] ** 2

model = smf.ols("u_hat2 ~ yhat + yhat2", data=df).fit()
print(model.summary())
print(model.f_test("yhat = yhat2 = 0"))
```

Problem 2

Use the data set **Movies**

Does viewing a violent movie lead to violent behavior? If so, the incidence of violent crimes, such as assaults, should rise following the release of a violent movie that attracts many viewers. Alternatively, movie viewing may substitute for other activities (such as alcohol consumption) that lead to violent behavior, so that assaults should fall when more viewers are attracted to the cinema. The dataset includes weekend U.S. attendance for strongly violent movies (such as Hannibal), mildly violent movies (such as Spider-Man), and nonviolent movies (such as Finding Nemo). The dataset also includes a count of the number of assaults for the same weekend in a subset of counties in the United States. Finally, the dataset includes indicators for year, month, whether the weekend is a holiday, and various measures of the weather.

(a) (i) Regress the logarithm of the number of assaults [ln_(assaults) = ln(assaults)] on the year and month indicators. Is there evidence of seasonality in assaults? That is, do there tend to be more assaults in some months than others? Explain.

(ii) Regress total movie attendance (attend = attend_v + attend_m + attend_n) on the year and month indicators. Is there evidence of seasonality in movie attendance? Explain.

(b) Regress ln_assaults on attend_v, attend_m, attend_n, the year and month indicators, and the weather and holiday control variables available in the data set

(i) Based on the regression, does viewing a strongly violent movie increase or decrease assaults? By how much? Is the estimated effect statistically significant?

(ii) Does attendance at strongly violent movies affect assaults differently than attendance at moderately violent movies? Differently than attendance at nonviolent movies?

(iii) A strongly violent blockbuster movie is released, and the weekend's attendance at strongly violent movies increases by 6 million; meanwhile, attendance falls by 2 million for moderately violent movies and by 1 million for nonviolent movies. What is the predicted effect on assaults? Construct a 95% confidence interval for the change in assaults.

(iv) It is difficult to control for all the variables that affect assaults and that might be correlated with movie attendance. For example, the effect of the weather on assaults and movie attendance is only crudely approximated by the weather variables in the data set. However, the data set does include a set of instruments, pr_attend_v, pr_attend_m, and pr_attend_n, that are correlated with attendance but are (arguably) uncorrelated with weekend-specific factors (such as the weather) that affect both assaults and movie attendance. These instruments use historical attendance patterns, not information on a particular weekend, to predict a film's attendance in a given weekend. For example, if a film's attendance is high in the second week of its release, then this can be used to predict that its attendance was also high in the first week of its release. (The details of

Problem 2 continued on next page. . .

the construction of these instruments are available in the Dahl and DellaVigna's paper on Canvas) Run the regression from part (b) (including year, month, holiday, and weather controls) but now using pr_attend_v, pr_attend_m, and pr_attend_n as instruments for attend_v, attend_m, and attend_n. Use this regression to answer (b)(i)–(b)(iii).

(v) It is difficult to control for all the variables that affect assaults and that might be correlated with movie attendance. For example, the effect of the weather on assaults and movie attendance is only crudely approximated by the weather variables in the data set. However, the data set does include a set of instruments, pr_attend_v, pr_attend_m, and pr_attend_n, that are correlated with attendance but are (arguably) uncorrelated with weekend-specific factors (such as the weather) that affect both assaults and movie attendance. These instruments use historical attendance patterns, not information on a particular weekend, to predict a film's attendance in a given weekend. For example, if a film's attendance is high in the second week of its release, then this can be used to predict that its attendance was also high in the first week of its release. (The details of the construction of these instruments are available in the Dahl and DellaVigna's paper on Canvas) Run the regression from part (b) (including year, month, holiday, and weather controls) but now using pr_attend_v, pr_attend_m, and pr_attend_n as instruments for attend_v, attend_m, and attend_n. Use this regression to answer (b)(i)–(b)(iii).

(vi) Based on your analysis, what do you conclude about the effect of violent movies on (short-run) violent behavior?

Problem 3

We examined **Koop and Tobias's data** on wages, education, ability, and so on. We considered the model.

$$\begin{aligned}\ln wage &= \beta_0 + \beta_1 educ + \beta_2 ability + \beta_3 experience \\ &\quad + \beta_4 motherEduc + \beta_5 FatherEduc + \beta_6 broken \\ &\quad + \beta_7 sinlings + \epsilon\end{aligned}$$

(a) We are interested in possible non-linearities in the effect of education on ln Wage. (Koop and Tobias focused on experience. As before, we are not attempting to replicate their results.) A histogram of the education variable shows values from 9 to 20, a spike at 12 years (high school graduation), and a second at 15. Consider aggregating the education variable into a set of dummy variables:

$$\begin{aligned}\text{HS} = 1 &\quad \text{if Educ} \leq 12, \quad 0 \text{ otherwise} \\ \text{Col} = 1 &\quad \text{if } 12 < \text{Educ} \leq 16, \quad 0 \text{ otherwise} \\ \text{Grad} = 1 &\quad \text{if Educ} > 16, \quad 0 \text{ otherwise}\end{aligned}$$

replace Educ in the model with (Col, Grad), making high school (HS) the base category, and recompute the model. Report all results. How do the results change? Based on your results, what is the marginal value of a college degree? What is the marginal impact on ln Wage of a graduate degree?

(b) The aggregation in part (a) actually loses quite a bit of information. Another way to introduce non-linearity in education is through the function itself. Add $educ^2$ to the equation in part (a) and recompute the model. Again, report all results. What changes are suggested? Test the hypothesis that the quadratic term in the equation is not needed—that is, that its coefficient is zero. Based on your results, sketch a profile of log wages as a function of education.

(c) One might suspect that the value of education is enhanced by greater ability. We could examine this effect by introducing an interaction of the two variables in the equation. Add the variable

$$EducAb = Educ \times ability$$

Problem 3 continued on next page...

to the base model in part a. Now, what is the marginal value of an additional year of education? The sample mean value of ability is 0.052374. Compute a confidence interval for the marginal impact on ln Wage of an additional year of education for a person of average ability.

(d) Combine the models in (b) and (c). Add both $educ^2$ and $EducAb$ to the base model in the beginning of the question and re-estimate. As before, report all results and describe your findings. If we define low ability as less than the mean and high ability as greater than the mean, the sample averages are −0.798563 for the 7,864 low-ability individuals in the sample and +0.717891 for the 10,055 high-ability individuals in the sample. Using the formulation in part (b), with this new functional form, sketch, describe, and compare the log wage profiles for low- and high-ability individuals.

(e) Suppose that you are now given the following regression model:

$$\begin{aligned}\ln(\text{wage}) = \beta_0 &+ \beta_1 \text{educ} \times \mathbf{1}(\text{educ} < \tau) + \beta_2 \text{educ} \times \mathbf{1}(\text{educ} \geq \tau) + \beta_3 \text{exp} \\ &+ \beta_4 \text{MotherEduc} + \beta_5 \text{FatherEduc} + \beta_6 \text{broken} \\ &+ \beta_7 \text{siblings} + \epsilon\end{aligned}$$

where τ is the threshold parameter, and

$$\mathbf{1}(\text{Educ} < \tau) = \begin{cases} 1 & \text{if Educ} < \tau, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{1}(\text{Educ} \geq \tau) = \begin{cases} 1 & \text{if Educ} \geq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 4

Using the **California test score date**, estimate the regression below using Nonlinear Least Squares. Report you coefficient estimates and standard errors.

$$TestScore = \beta_0(1 - e^{\beta_1(income - \beta_2)}) + \epsilon$$

Problem 5

Use the **Consumption.xlsx**. We have previously estimated the nonlinear consuption function below using nonlinear least squares in class:

$$C = \alpha + \beta Y^\gamma + \epsilon$$

Where C is the real consumption and Y is the real disposable income. Alternatively, we can assume that the error term has a normal distribution and estimate the nonlinear regression above using the maximum likeligood estimation (MLE) approach, In particular, the MLE approach maximizes the log-likelihood function given by:

$$L(\alpha, \beta, \gamma, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(C_i - \alpha - \beta Y^\gamma)^2$$

Where σ^2 is the variance of the error term. Using a statistical programming language of you choice, estimate the regression model using the maximum likelihood estimation approach. Your estimate are expected to be similar to those in Table 7.1 of Green's textbook. Please submit the following:

(a) Your code used to preform the estimation.

(b) The output of your estimation, including the estimated parameters and the error variance.