ECON 124: Problem Set #3

Due on Jun 5, 2025

Dr. Deniz Baglan

Alejandro Ouslan

Problem 1

The following sample moments for $x = [1, x_1, x_2, x_3]$ were computed from 100 observations produced using a random number generator:

$$X'X \begin{bmatrix} 100 & 123 & 96 & 109 \\ 123 & 252 & 125 & 189 \\ 96 & 125 & 167 & 146 \\ 109 & 189 & 146 & 168 \end{bmatrix} \quad X'y = \begin{bmatrix} 460 \\ 810 \\ 615 \\ 712 \end{bmatrix} \quad y'y = 3924$$

The true model underlying these data is $y = x_1 + x_2 + x_3 + \epsilon$.

1. Compute the simple correlation among the regressors.

$$\begin{bmatrix} 1.0000 & 0.6093 & 0.9186 \\ 0.6093 & 1.0000 & 0.8716 \\ 0.9186 & 0.8716 & 1.0000 \end{bmatrix}$$

2. Compute the ordinary least squares coefficients in the regression of y on a constant, x_1, x_2, and x_3.

$$\hat{\beta} = \begin{bmatrix} -0.4022 \\ 6.1234 \\ 5.9097 \\ -7.5256 \end{bmatrix}$$

3. Compute the ordinary least squares coefficients in the regression of y on a constant x_1 and x_2, on a constant, X_1 and x_3, and on a constant, x_2 and x_3.

$$\text{Regression of } y \text{ on a constant, } x_1, x_2 : \quad \hat{\beta} = \begin{bmatrix} -0.2264 \\ 2.2801 \\ 2.1061 \end{bmatrix}$$

$$\text{Regression of } y \text{ on a constant, } x_1, x_3 : \quad \hat{\beta} = \begin{bmatrix} -0.0696 \\ 0.2292 \\ 4.0254 \end{bmatrix}$$

$$\text{Regression of } y \text{ on a constant, } x_2, x_3 : \quad \hat{\beta} = \begin{bmatrix} -0.0627 \\ -0.0918 \\ 4.3585 \end{bmatrix}$$

4. Compute the variance inflation factor associated with each variable.

Variance Inflation Factors (VIFs):

$$\text{VIF}(x_1) = 258.40$$
$$\text{VIF}(x_2) = 168.07$$
$$\text{VIF}(x_3) = 676.27$$

5. The regressors are obviously badly collinear, Which is the problem variable? Explain
The most problematic variable is x_3 with a VIF of 676.27

Python Code

```
import numpy as np
from numpy.linalg import inv
```

Problem 1 continued on next page. . .

```python
XTX = np.array(
    [
        [100, 123, 96, 109],
        [123, 252, 125, 189],
        [96, 125, 167, 146],
        [109, 189, 146, 168],
    ]
)

XTy = np.array([460, 810, 615, 712]).reshape(-1, 1)

yTy = 3924


def main() -> None:
    # Problem 1a
    M = np.array([[252, 125, 189], [125, 167, 146], [189, 146, 168]])
    std_devs = np.sqrt(np.diag(M))

    correlation_matrix = M / np.outer(std_devs, std_devs)
    print(np.round(correlation_matrix, 4))

    # Problem 1b
    print(inv(XTX) @ XTy)

    # Problem 1c
    XTX_1 = np.delete(np.delete(XTX, 3, 0), 3, 1)
    XTy_1 = np.delete(XTy, 3, 0)
    print(inv(XTX_1) @ XTy_1)

    XTX_2 = np.delete(np.delete(XTX, 2, 0), 2, 1)
    XTy_2 = np.delete(XTy, 2, 0)
    print(inv(XTX_2) @ XTy_2)

    XTX_3 = np.delete(np.delete(XTX, 1, 0), 1, 1)
    XTy_3 = np.delete(XTy, 1, 0)
    print(inv(XTX_3) @ XTy_3)

    # Problem 1d
    XTX_no_const = np.delete(np.delete(XTX, 0, 0), 0, 1)

    stds = np.sqrt(np.diag(XTX_no_const))

    R = XTX_no_const / np.outer(stds, stds)

    R_inv = np.linalg.inv(R)
    VIFs = np.diag(R_inv)

    for i, vif in enumerate(VIFs, start=1):
```

Problem 1 continued on next page. . .

```
        print(f"VIF for x_{i}: {vif:.2f}")


if __name__ == "__main__":
    main()
```

Problem 2

A multiple regression of y on a constant x_1 and x_2 produces the following resutls:

$$\hat{y} = 4 + 0.4x_1 + 0.9x_2 \quad R^2 = \frac{8}{60} \quad e'e = 520, \quad n = 29,$$

$$X'X = \begin{bmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{bmatrix}$$

Test the hypothesis that the two slopes sum to 1

Python Code

```python
import numpy as np
from scipy import stats


def main() -> None:
    e_e = 520
    n = 29
    k = 3
    XTX = np.array([[29, 0, 0], [0, 50, 10], [0, 10, 80]])
    beta_hat = np.array([4, 0.4, 0.9])

    sigma_squared = e_e / (n - k)

    XTX_inv = np.linalg.inv(XTX)

    R = np.array([[0, 1, 1]])
    r = 1

    Rb_minus_r = R @ beta_hat - r
    denominator = R @ XTX_inv @ R.T * sigma_squared
    F_stat = (Rb_minus_r**2) / denominator

    df1 = 1
    df2 = n - k

    p_value = 1 - stats.f.cdf(F_stat, df1, df2)

    print("F-statistic:", F_stat[0][0])
    print("p-value:", p_value)
```

Problem 2 continued on next page. . .

```
if __name__ == "__main__":
    main()
```

Problem 3

The application in Chapter 3 used 15 of the 19,919 observations in Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. (See Appendix Table F3.2.) We will use the full data set for this exercise. The data may be downloaded from the *Journal of Applied Econometrics* data archive at **link**. The data file is in two parts. The fist file contains the panel of 19,919 obseervations on variables:
To create the data set for this exercise, it is necessary to merge these two data files. The ith observations in the first file will be replicated T_i times for the set of T_i observations in the first file. The *person id* variable indicates which rows must contain the data from the second file. (How this preparation is carried out will) vary from one computer package to another.) (*Note:* We are not attempting to replicate the data set.) Let

$$X_1 = [constant, education, experience, ability]$$

$$X_2 = [mother'seducation, father'seducation, brokenhome, numberofsiblings]$$

1. compute the full regression of $(\ln wage \sim X_1)$ and $(\ln wage \sim X_2)$

Table 1: OLS Regression Results: ln(wage) $\sim X_1$

Variable	**Coef.**	**Std. Err.**	**t**	**P**$>|t|$	**[0.025**	**0.975]**
const	1.0272	0.030	34.194	0.000	0.968	1.086
education (x_1)	0.0738	0.002	33.312	0.000	0.069	0.078
experience (x_2)	0.0395	0.001	43.958	0.000	0.038	0.041
ability (x_3)	0.0829	0.005	18.020	0.000	0.074	0.092
Model statistics:						
R-squared		0.173				
Adj. R-squared		0.173				
F-statistic		1253	(Prob F-statistic = 0.000)			
No. Observations		17919				
Df Residuals		17915				
Df Model		3				
Log-Likelihood		-12283				
AIC		24570				
BIC		24600				
Durbin-Watson: 0.801						
Omnibus: 1110.415, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 2075.096, Prob(JB): 0.000						
Skew: -0.458, Kurtosis: 4.393						
Cond. No.: 130						

2. Use the F test to test the hypothesis that all coefficients except the constant term are zero. $\beta_1 = \beta_2 = \beta_3 = 0$) is tested using the F test.

$$F = 1252.94, \quad p\text{-value} = 0.000, \quad \mathrm{df}_{\mathrm{num}} = 3, \quad \mathrm{df}_{\mathrm{denom}} = 17915$$

Since the p-value is effectively zero, we reject the null hypothesis and conclude that the regressors are jointly significant.

Table 2: OLS Regression Results: ln(wage) $\sim X_2$

Variable	**Coef.**	**Std. Err.**	**t**	**P>** $\|t\|$	**[0.025**	**0.975]**
const	2.0119	0.019	104.391	0.000	1.974	2.050
mother's education (x_1)	0.0100	0.002	5.538	0.000	0.006	0.014
father's education (x_2)	0.0151	0.001	10.727	0.000	0.012	0.018
broken home (x_3)	-0.0861	0.011	-7.964	0.000	-0.107	-0.065
number of siblings (x_4)	0.0020	0.002	1.034	0.301	-0.002	0.006
Model statistics:						
R-squared		0.027				
Adj. R-squared		0.027				
F-statistic		123.2	(Prob F-statistic = 6.81e-104)			
No. Observations		17919				
Df Residuals		17914				
Df Model		4				
Log-Likelihood		-13746				
AIC		27500				
BIC		27540				
Durbin-Watson: 0.782						
Omnibus: 383.928, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 580.233, Prob(JB): 1.01e-126						
Skew: -0.229, Kurtosis: 3.753						
Cond. No.: 85.8						

3. Use the F statistic to test the joint hypothesis that the coefficient on the four household variables in X_2 are zero

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

The F test statistic is

$$F = 123.18, \quad p\text{-value} = 6.81 \times 10^{-104}, \quad \text{df}_{\text{num}} = 4, \quad \text{df}_{\text{denom}} = 17914$$

Since the p-value is extremely small, we reject the null hypothesis and conclude that the household variables are jointly significant.

4. Use a Wald test to carry out the test in part *c*.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$F = 123.18, \quad p\text{-value} = 6.81 \times 10^{-104}, \quad \text{df}_{\text{num}} = 4, \quad \text{df}_{\text{denom}} = 17914$$

Python Code

```
import numpy as np
import polars as pl
import statsmodels.api as sm


def main():
    df1 = pl.read_excel("data/timeinvar.xlsx")
    df2 = pl.read_excel("data/timevar.xlsx")
    df = df1.join(df2, on="id", how="left", validate="m:m")
```

Problem 3 continued on next page. . .

```
    # 3b

    X_1 = df[["edu", "exper", "ability"]].to_numpy()
    X_1 = sm.add_constant(X_1)
    y = df[["lwage"]].to_numpy()
    res1 = sm.OLS(y, X_1).fit()
    print(res1.summary())

    X_2 = df[["meduc", "feduc", "brokenhome", "siblings"]].to_numpy()
    X_2 = sm.add_constant(X_2)
    y = df[["lwage"]].to_numpy()
    res2 = sm.OLS(y, X_2).fit()
    print(res2.summary())

    # 3b
    A = np.identity(len(res1.params))
    A = A[1:, :]

    f_test_result = res1.f_test(A)
    print(f_test_result)

    # 3c
    A = np.identity(len(res2.params))
    A = A[1:, :]

    f_test_result = res2.f_test(A)
    print(f_test_result)

    # 3d
    res2.wald_test(A)


if __name__ == "__main__":
    main()
```

Problem 4

In a paper in 1963, Mare Nerlove analyzed a cost function for 145 American electric companies. The attached data file, contains the data and the description file. Nerlove was interested in estimating a cost function: $TC = f(Q, PL, PF, PK)$.

1. First estimate an unrestricted Cobb-Douglas specification

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + \epsilon_i$$

   Report parameter estimates and standard errors.
2. What is the economic meaning of the restriction $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$?
   It means that if the cost cost of capital fuel and labor double cost will also double
3. Estimate the regression in (a) by constrained least squares $\beta_3 + \beta_4 + \beta_5 = 1$. Report your parameter estimates and standard errors.

Problem 4 continued on next page. . .

Table 3: OLS Regression Results: Cobb-Douglas Cost Function

Variable	**Coefficient**	**Std. Error**	**t-stat**	**P-value**	**[0.025**	**0.975]**
const	-3.5265	1.774	-1.987	0.049	-7.035	-0.018
$\log Q$	0.7204	0.017	41.244	0.000	0.686	0.755
$\log PL$	0.4363	0.291	1.499	0.136	-0.139	1.012
$\log PK$	-0.2199	0.339	-0.648	0.518	-0.891	0.451
$\log PF$	0.4265	0.100	4.249	0.000	0.228	0.625
Model Statistics:						
R-squared		0.926				
Adj. R-squared		0.924				
F-statistic		437.7	(Prob F-statistic $= 4.82 \times 10^{-78}$)			
No. Observations		145				
Df Residuals		140				
Df Model		4				
Durbin-Watson: 1.013						
Omnibus: 51.403, Prob(Omnibus): 0.000						
Jarque-Bera (JB): 175.700, Prob(JB): 7.03×10^{-39}						
Skew: 1.303, Kurtosis: 7.721						
Condition Number: 506						

Table 4: Constrained GLS Regression Results: $\beta_3 + \beta_4 + \beta_5 = 1$

Variable	**Coefficient**	**Std. Error**	**z-stat**	**P-value**	**[0.025**	**0.975]**
const	-4.6908	0.885	-5.301	0.000	-6.425	-2.956
$\log Q$	0.7207	0.017	41.334	0.000	0.687	0.755
$\log PL$	0.5929	0.205	2.898	0.004	0.192	0.994
$\log PK$	-0.0074	0.191	-0.039	0.969	-0.381	0.366
$\log PF$	0.4145	0.099	4.189	0.000	0.221	0.608
Model Statistics:						
No. Observations		145				
Df Residuals		141				
Log-Likelihood		-67.838				
Deviance		21.640				
Pearson Chi2		21.6				

Problem 4 continued on next page. . .

4. Test $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$ using a Wald statistic.

$$W = 0.0000, \quad p\text{-value} = 1.0000$$

Since the p-value is 1, we fail to reject the null hypothesis and conclude that the linear constraint is consistent with the data.

Python Code

```python
import polars as pl
import statsmodels.api as sm
import numpy as np

from scipy.stats import chi2


def main() -> None:
    df = pl.read_excel("data/Nerlove1963.xlsx").to_pandas()
    for col in df.columns:
        df[f"{col}_log"] = np.log(df[col])
    X = df[["output_log", "Plabor_log", "Pcapital_log", "Pfuel_log"]]
    X = sm.add_constant(X)
    y = df["Cost_log"]

    model = sm.OLS(y, X).fit()

    print(model.summary())

    # 4c
    glm_model = sm.GLM(y, X, family=sm.families.Gaussian())
    constraint = "Plabor_log + Pcapital_log + Pfuel_log = 1"

    model_constrained = glm_model.fit_constrained(constraint)
    print(model_constrained.summary())
    # 4d
    R = np.array([[0, 0, 1, 1, 1]])
    q = np.array([1])

    beta_hat = model_constrained.params.values
    cov_beta = model_constrained.cov_params().values

    W = (R @ beta_hat - q) @ np.linalg.inv(R @ cov_beta @ R.T) @ (R @ beta_hat - q)
    p_value = 1 - chi2.cdf(W, df=1)

    print(f"Wald statistic: {W:.4f}")
    print(f"p-value: {p_value:.4f}")


if __name__ == "__main__":
    main()
```

Problem 5

Replicate Example 7.12 income elasticity of credit card expenditures in Green's textbook.the data set can be downloaded form the link below: