In [ ]:

```
#Load packages
using DataFrames, CSV, Statistics, Distributions, GLM, Plots, Random, StatsModels, StatFiles, FixedEffectModels
```

In [ ]:

```
# Load DataFrames
df = DataFrame(load("fulton.dta"));
```

Problem 1

On the Fulton fish market about 35 different wholesale dealers sell fish to the customers that visit the market from Monday to Friday. There are no posted prices and dealers can (and do) charge different prices to different consumers. In this exercise you will analyze data which were recovered from a single dealer on the Fulton fish market. The data run from December 2nd, 1991 through May 8th, 1992 and record daily transactions in whiting. Whiting is a cheap fish which is oily and distinctive tasting. Whiting vary relatively little in size and quality, and there is probably little scope for substitution with other types of fish. Use fulton.dta. The dataset contains the quantity of whiting sold in pounds (qty) and the average (quantity weighted) prices in dollars per pound (price), information about the day of the week (Monday-Friday) and the weather conditions (stormy, mixed)

1.Write a two-equation system in “supply and demand form”, that is, with the same variable (typically, quantity) appearing on the left-hand side:

$$qty = \alpha_1 price + \beta_1 z_1 + u_1 \tag{1}$$

$$qty = \alpha_2 price + \beta_2 z_2 + u_2 \tag{2}$$

If $\alpha_1 = 0$ or $\alpha_2 = 0$, explain why a reduced form exists for qty. (Remember, a reduced form expresses qty as a linear function of the exogenous variables and the structural errors.) If $\alpha_1 \neq 0$ and $\alpha_2 = 0$, find the reduced form for price.

Without loss of generality, we can assume that $\alpha_2 = 0$ and $\alpha_1 \neq 0$ and then the functions can be reduced to

$$qty = \beta_2 z_2 + u_2$$

Which is a reduced form for qty.

Then we can equate the two equations to have:

$$\alpha_1 price + \beta_1 z_1 + u_1 = \beta_2 z_2 + u_2$$

And thus

$$price = \frac{\beta_2 z_2 + u_2 - \beta_1 z_1 - u_1}{\alpha_1}$$

2.If $\alpha_1 \neq 0$, $\alpha_2 \neq 0$ and $\alpha_1 \neq \alpha_2$, find the reduced for for qty. Does price have a reduced form in this case?

From equation (1), we can rewrite it as a function of $price$:

$$price = \frac{qty - \beta_1 z_1 - u_1}{\alpha_1} \tag{3}$$

Then plug it into equation (2):

$$\begin{aligned} qty &= \alpha_2 \frac{qty - \beta_1 z_1 - u_1}{\alpha_1} + \beta_2 z_2 + u_2 \\ &= \frac{\alpha_2}{\alpha_1} qty - \alpha_2 \frac{\beta_1 z_1 + u_1}{\alpha_1} + \beta_2 z_2 + u_2 \\ &= \frac{-\alpha_2 \frac{\beta_1 z_1 + u_1}{\alpha_1} + \beta_2 z_2 + u_2}{1 - \frac{\alpha_2}{\alpha_1}} \end{aligned} \tag{4}$$

Which is the reduced form for qty.

We then plug equation (2) into equation (3):

$$price = \frac{\alpha_2 price + \beta_2 z_2 + u_2 - \beta_1 z_1 - u_1}{\alpha_1}$$

Isolate $price$ to have:

$$\begin{aligned} price &= \frac{\frac{\beta_2 z_2 + u_2 - \beta_1 z_1 - u_1}{\alpha_1}}{1 - \frac{\alpha_2}{\alpha_1}} \\ &= \frac{\beta_2 z_2 + u_2 - \beta_1 z_1 - u_1}{\alpha_1 - \alpha_2} \end{aligned} \tag{5}$$

Which is the reduced form for $price$.

3.Is the condition $\alpha_1 \neq \alpha_2$ likely to be met in supply and demand examples? Explain.

Yes, in reality it is very likely that the two coefficients are not the same. The reason is that the coefficient for price on quantity represents the elasticity, and that for demand and supply comes from very different sources. In the case of fish, the demand elasticity can be determined by factors like the preference for fish and other substitutes like meat and the supply elasticity can be determined by the operation and technology of the fishing industry.

4.The following simultaneous equations model imposes the equilibrium condition that supply equals demand:

$$qty = \alpha_1 price + \beta_1 stormy + u_1 \tag{6}$$

$$qty = \alpha_2 price + \beta_2 friday + u_2 \tag{7}$$

Which would you argue is the supply equation and which is the demand equation? Explain. In what follows suppose $\beta_2 = 0$.

Equation (6) should be the supply equation because when weather is stormy, it should be become hard to catch fish, reducing the supply. Equation (7) is the demand function because people might want more fish on Friday for some reason.

$\beta_2 = 0$ means that people's consumption of fish is the same on Friday compared to other days in a week.

5.Do the supply and demand equations satisfy the order condition for identification?

Since we have two exogenous variables: stormy and friday, and also two endogenous variables, so $H = K$ which satisifies the condition for exact identification.

6.Estimate reduced form equations for qty and price. Comment on the reduced form coefficients on stormy

Replace z_1 and z_2 with $stormy$ and $friday$ from the reduced form equations (4) and (5) and isolate the exogenous variables on the right hand sides to have:

$$qty = -\frac{\alpha_2\beta_1}{\alpha_1 - \alpha_2} stormy + \frac{\alpha_1\beta_2}{\alpha_1 - \alpha_2} friday + u_{qty} \tag{8}$$

$$price = \frac{\beta_2}{\alpha_1 - \alpha_2} friday - \frac{\beta_1}{\alpha_1 - \alpha_2} stormy + u_{price} \tag{9}$$

In [ ]:

```
# Modify the data
df.stormy = Float64.(df.stormy)
df.friday = Float64.(df.friday)
df.qty = Float64.(df.qty)

# Estimate the supply function
model_qty = lm(@formula(qty ~ stormy + friday),df)
```

```
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}, GLM.DensePredChol{Float64, LinearAlgebra.Chol
eskyPivoted{Float64, Matrix{Float64}, Vector{Int64}}}}, Matrix{Float64}}

qty ~ 1 + stormy + friday

Coefficients:
──────────────────────────────────────────────────────────────────────────
                  Coef.  Std. Error      t  Pr(>|t|)  Lower 95%   Upper 95%
──────────────────────────────────────────────────────────────────────────
(Intercept)   8.58036    0.0892642  96.12    <1e-99   8.40342    8.7573
stormy       -0.356755   0.151787   -2.35    0.0206  -0.657622  -0.0558871
friday        0.221614   0.169636    1.31    0.1942  -0.114633   0.557861
──────────────────────────────────────────────────────────────────────────
```

For the stormy dummy variable, the coefficient is negative which matches the idea that during a stormy day the fish supply is decreased and it is also statistically significant at 5% level.

For the friday dummy variable, the coefficient is positive which means more fish are supplied on Friday for some reason but it is not statistically significant so we cannot reject the null hypothesis that fish supply on Friday is not different from other days.

In [ ]:

```
# Modify the data
df.price = Float64.(df.price)
# Estimate the demand function
model_price = lm(@formula(price ~ stormy + friday),df)
```

```
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}, GLM.DensePredChol{Float64, LinearAlgebra.Chol
eskyPivoted{Float64, Matrix{Float64}, Vector{Int64}}}}, Matrix{Float64}}

price ~ 1 + stormy + friday

Coefficients:
──────────────────────────────────────────────────────────────────────────
                  Coef.  Std. Error      t  Pr(>|t|)  Lower 95%  Upper 95%
──────────────────────────────────────────────────────────────────────────
(Intercept)  -0.296184    0.0435453  -6.80    <1e-09  -0.382499  -0.20987
stormy        0.336015    0.0740453   4.54    <1e-04   0.189245   0.482786
friday        0.0271893   0.0827525   0.33    0.7431  -0.13684    0.191219
──────────────────────────────────────────────────────────────────────────
```

For the stormy dummy variable, the coefficient in positive and statistically significant, meaning that the price goes up on stormy day, which match the fact that the suply on stormy day will decrease. And the Friday dummy variable is positive but not statistically significant, meaning we cannot reject the null hypothesis that price on friday is different from the others.

7.Divide the reduced form coefficient on stormy in the reduced form quantity equation by the reduced form coefficient on stormy in the reduced form price equation. What parameter are you estimating now? Interpret on the sign and size of this estimate.

From equation (8) and (9) we have:

$$\frac{-\frac{\alpha_2\beta_1}{\alpha_1-\alpha_2}}{-\frac{\beta_1}{\alpha_1-\alpha_2}} = \alpha_2$$

which is exactly the coefficient on price in the demand curve. The reason behind is that stormy shift the supply curve which help us identify the demand curve.

Numerically: $-0.356755/0.336015 = -1.0617$

The sign is negative, meaning that when price increases, quantity demand decreases, matching the normal demand curve. And 1 dollar per pound increase in price decreases the quantity demand by -1.0617 pounds.

8.Estimate the same parameter in (7) using 2SLS (-ivreg- in Stata)

In [ ]:

```
# So we essentially treat stormy as a instrument for price.
# Estimate the 2SLS model
demand_2sls= reg(df, @formula(qty ~ (price ~ stormy)))
```

```
                             FixedEffectModel
===========================================================================
Number of obs:                   111   Converged:                      true
dof (model):                       1   dof (residuals):                 108
R²:                           -0.000   R² adjusted:                  -0.009
F-statistic:                 5.40174   P-value:                       0.022
F-statistic (first stage):   21.0696   P-value (first stage):         0.000
===========================================================================
             Estimate  Std. Error    t-stat  Pr(>|t|)  Lower 95%  Upper 95%
───────────────────────────────────────────────────────────────────────────
price        -1.08241    0.46572   -2.32416    0.0220   -2.00555  -0.159272
(Intercept)   8.31379    0.114622  72.5319     <1e-92    8.08659   8.54099
===========================================================================
```

9.Can you find an explanation why we assumed that β2 = 0 (Hint: include it in the reduced form estimations)

In the reduced form estimation we have the coefficients on Friday are given by:

$$\frac{\alpha_1\beta_2}{\alpha_1-\alpha_2}$$

$$\frac{\beta_2}{\alpha_1-\alpha_2}$$

Since in the reduced form estimation we found that both coefficients are not significant, so it is safe to assume them to be 0. To achieve this, we can assume $\beta_2=0$.

Problem2

For each of the following simple regression models, find reasons why the residual may not be orthogonal to the explicative variables, and determine the direction of the bias of the OLS estimator of β_1:

1.$score=\beta_0+\beta_1 classsize+e$

Because class size may correlate with other factors that can affect score, including parents' education level and income. Factors like income are negatively correlated with class size, thus the bias of OLS estimator is downward.

2.$exam=\beta_0+\beta_1 attendance+e$

Attendance rate may correlate with the health condition of the students, which also affects their exam performance. As attendance rate is positively correlated with health and also the exam performance, the bias of OLS estimator is upward.

3.$crime = \beta_0 + \beta_1 education + e$

Education can be correlated with parents' income which also affects the probability of comitting crimes. With positive correlation between parents' income and education it is expected that the bias of OLS estimator is upward.

Problem 3

Consider the following wage equation :

$$y_i = a + bx_i + u_i$$

where x_i represent the level of education of individual i, and y_i the logarithm of its labor income. We are concerned by the potential endogeneity problems which may arise from this formulation.

1.In order to highlight these problems, we assume in this question that there exist unob- served characteristics z_i, called “ability”, which has an impact on both u_i et x_i. That is,

$$\begin{cases} u_i = bz_i + \eta_i, \\ x_i = \alpha + \beta z_i + e_i. \end{cases}$$

(a)Interpret this structural model.

This structural model accounts for the endogenity of x_i which is correlated with u_i through the equation using z_i, the unobservable ability. This ability affects both the education level and income.

(b)Assume that η_i and e_i are not correlated and have zero mean, determine the asymptotic bias of the OLS estimator $\hat{\beta}_{OLS}$. Show that it is likely to be positive.

We have

$$plim(\beta) = b_0 + \frac{Cov(x_i, u_i)}{Var(x_i)}$$

so the bias is just

$$\frac{Cov(x_i, u_i)}{Var(x_i)}$$

Then we plug in the formula for u_i and x_i,

$$bias = \frac{Cov(bz_i + \eta_i, \alpha + \beta z_i + e_i)}{Var(x_i)}$$

Since η_i and e_i is not correlated and so does e_i and z_i, we can simplify the bias formula into:

$$bias = \frac{b\beta Var(z_i)}{Var(x_i)}$$

Since $Var \geq 0$ by definition, $b > 0$ because education has positive effects on income, and $\beta > 0$ because the unobservable ability also positively affects education, the bias is positive.

(c)We are interested in computing empirically the bias of $\hat{\beta}$. Which method could we use ?

We can use 2SLS. First we need to find an instrumental variable that is correlated with education but not the error terms (in other words not correlated with income). Then we estimate the $\hat{\beta}_{2SLS}$ and also $\hat{\beta}_{OLS}$. Their difference is the bias.

2.We find a significant negative bias at the end. We interpret the previous paradox by assuming that there are measurement errors. The true model can be written :

$$y_i^* = a + bx_i^* + u_i,$$

where y_i^* is the labor income measured by y_i with error:

$$y_i = y_i^* + \nu_i$$

and x_i^* is the true level of education measure with error by x_i:

$$x_i = x_i^* + \epsilon_i$$

We assume that the (measurement) errors ϵ_i and ν_i are not correlated, iid and not correlated with x_i^*.

(a)Let $\hat{\beta}$ be the OLS estimator of the coefficient associated to x_i in the regression of y_i on x_i with a constant term. Express the asymptotic bias over b and show that the measurement error biases the estimator downward (toward 0).

By substituting the measurement error equations into the true model, we have:

$$\begin{aligned} y &= a + b(x_i - \epsilon_i) + \nu_i \\ &= a + bx_i + \nu_i - b\epsilon_i \end{aligned}$$

Then we use the general formula for bias

$$\begin{aligned} bias &= \frac{Cov(x_i, \nu_i - b\epsilon_i)}{Var(x_i)} \\ &= \frac{Cov(x_i^* + \epsilon_i, \nu_i - b\epsilon_i)}{Var(x_i)} \end{aligned}$$

Since x_i^* is not correlated with ν_i and ϵ_i and the two measurement errors are not correlated with each other,

$$bias = \frac{-bVar(\epsilon_i)}{Var(x_i)}$$

Since $Var > 0$ by definition, and $b > 0$ for positive correlation between education and income, this measurement error causes downward bias.

(b)Let $\hat{c}$ be the OLS estimator of the coefficient associated to y_i in the regression of x_i on y_i with a constant term. Express the asymptotic bias of $1/\hat{c}$ over b. Show that the bias is positive.

The OLS estimator is:

$$\hat{c} = \frac{\mathrm{Cov}(y_i, x_i)}{\mathrm{Var}(y_i)}$$

Use:

- $y_i = y_i^* + \nu_i = a + bx_i^* + u_i + \nu_i$
- $x_i = x_i^* + \varepsilon_i$

Because the measurement erros does not correlated with the other variable

$$\mathrm{Cov}(y_i, x_i) = \mathrm{Cov}(y_i^* + \nu_i,\ x_i^* + \varepsilon_i) = \mathrm{Cov}(y_i^*, x_i^*)$$

Since $y_i^* = a + bx_i^* + u_i$, we get:

$$\mathrm{Cov}(y_i^*, x_i^*) = b\mathrm{Var}(x_i^*)$$

$$\mathrm{Var}(y_i) = \mathrm{Var}(y_i^* + \nu_i) = \mathrm{Var}(y_i^*) + \mathrm{Var}(\nu_i)$$

And:

$$\mathrm{Var}(y_i^*) = b^2\mathrm{Var}(x_i^*) + \mathrm{Var}(u_i)$$

Thus:

$$\mathrm{Var}(y_i) = b^2\mathrm{Var}(x_i^*) + \mathrm{Var}(u_i) + \mathrm{Var}(\nu_i)$$

So asympotically,

$$\hat{c} \xrightarrow{p} \frac{b\mathrm{Var}(x_i^*)}{b^2\mathrm{Var}(x_i^*) + \mathrm{Var}(u_i) + \mathrm{Var}(\nu_i)}$$

Now take the reciprocal:

$$\frac{1}{\hat{c}} \xrightarrow{p} \frac{b^2\mathrm{Var}(x_i^*) + \mathrm{Var}(u_i) + \mathrm{Var}(\nu_i)}{b\mathrm{Var}(x_i^*)} = b + \frac{\mathrm{Var}(u_i) + \mathrm{Var}(\nu_i)}{b\mathrm{Var}(x_i^*)}$$

As $b > 0$ for positive correlation between education and $Var \geq 0$ by definition, the bias part is positive.

Problem 4

A simple model to determine the effectiveness of nicotine patch usage on reducing lung cancers among smokers is:

$$cancerrate = \beta_0 + \beta_1 patchuse + \beta_2 percmale + \beta_3 avgine + \beta_4 city + u$$

where *cancerrate* is the percent of smokers who have contracted lung cancer, *patchuse* is the percentage of individuals who claim to regularly use nicotine patches, *avginc* is average income, and *city* is a dummy variable indicating whether the individual lives in a city. The model is at the state level.

1.Interpreting the preceding equation in a causal, ceteris paribus way, what should be the sign of β_1? Explain the mechanism.

We are predicting the probability of getting lung cancer among somkers by using factors including the use of nicotine patches, average income, gender and living area.

It is expected that the sign β_1 is negative because the use of nicotine patches can help smokers to reduce the frequency of smoking, thus decreasing the risk of getting lung cancer.

2.Why would you suspect $patchuse$ to be correlated with the error term u?

Because there is problem with reverse causality that people with higher risk of getting lung cancer will use more nicotine patch to reduce the risk, thus making the use of nicotine patch correlates with the error term. Also there may be the issue of omitted variable, for example the mental strength to continue use nicotine patch, which is also correlated with the risk of getting cancer.

3.Assume that patch usage increases with the rate of lung cancer, so that $\gamma_1 > 0$ in the equation

$$patchuse = \gamma_0 + \gamma_1 cancerrate + \gamma_2 otherfactors + v.$$

4.What is the likely bias in estimating β_1 by OLS?

Since the patch usage is positively correlated with the rate of lung cancer and thus the error term, the bias upward. As a result, we may overestimate β_1, that is the reality the patch usage should less positive effects of rate of lung cancer.

5.What sort of variable would you look for in order to instrument $patchuse$?

So a good instrument should be correlated with the independent varaibles $patchuse$ but not the error term or $cancerrate$. One potential instrument can be the distance between population center and clinics that offer nicotine patches because it is correlated with the useage of the patch but not necessarily with the cancer rate itself.

6.Let $patchdis$ be a binary variable equal to unity if the state has a program to distribute nicotine patch. Explain how this can be used to estimate β_1 (and the other β_s) by IV. What do we have to assume about $patchdis$ in the above 2 equations?

First we estimate the equation

$$patchuse = \alpha_0 + \alpha_1 patchdis + e$$

and it is expected that α_1 is positive and statistically significant to satisfy the first condition for good instrument. and get the predicted $patc\hat{h}use$ and run the second stage regression:

$$cancerrate = \beta_0 + \beta_1 \hat{patchuse} + \beta_2 percmale + \beta_3 avgine + \beta_4 city + u$$

We need to assume that $patchdis$ is not correlated with the cancer rate otherwise it does not satisfy the exogeneity assumption.

When estimating the second stage equation, we will get a better estimation of other β_s.