Lab - Week 8

Kimberly Boswell

11/22/2023

Panel Data Example: Airline Services

We will consider a case study about costs of airline services. The response variable is cost, and the inputs are firm output, price of fuel, and load. The data consist of 6 firms and 15 years of annual observations from 1970 - 1984. Description of the variables are as follows:

- firm - factor indicating airline firm.
- year - factor indicating year.
- output -output revenue passenger miles index number.
- cost - total cost (in USD 1000).
- price - fuel price.
- load - average capacity utilization of the fleet.

Exercises

1) How can we describe the data in terms of width and length?
2) How do you transform the dataset into panel data
3) Through data visualization, create scatterplots of your data distinguishing each unit/individual. How can we view individual heterogeneity?

For the following relationship:

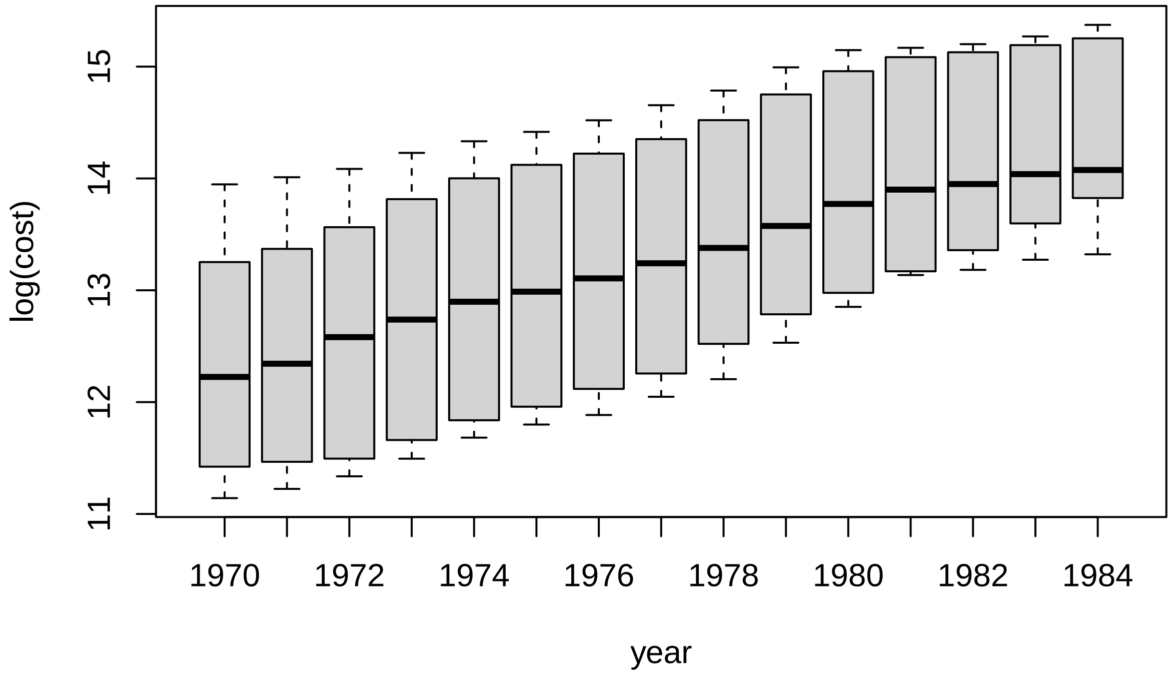
$$log(cost) = \alpha_0 + \alpha_1 log(output) + \alpha_2 log(output)^2 + \alpha_3 log(price) + \alpha_4 load + e$$

3) Create a pooled model, fixed effect model with individual effects, time effects and both 4) Create a random effects model 5) Perform the relevant tests to determine the type of model that should be used.

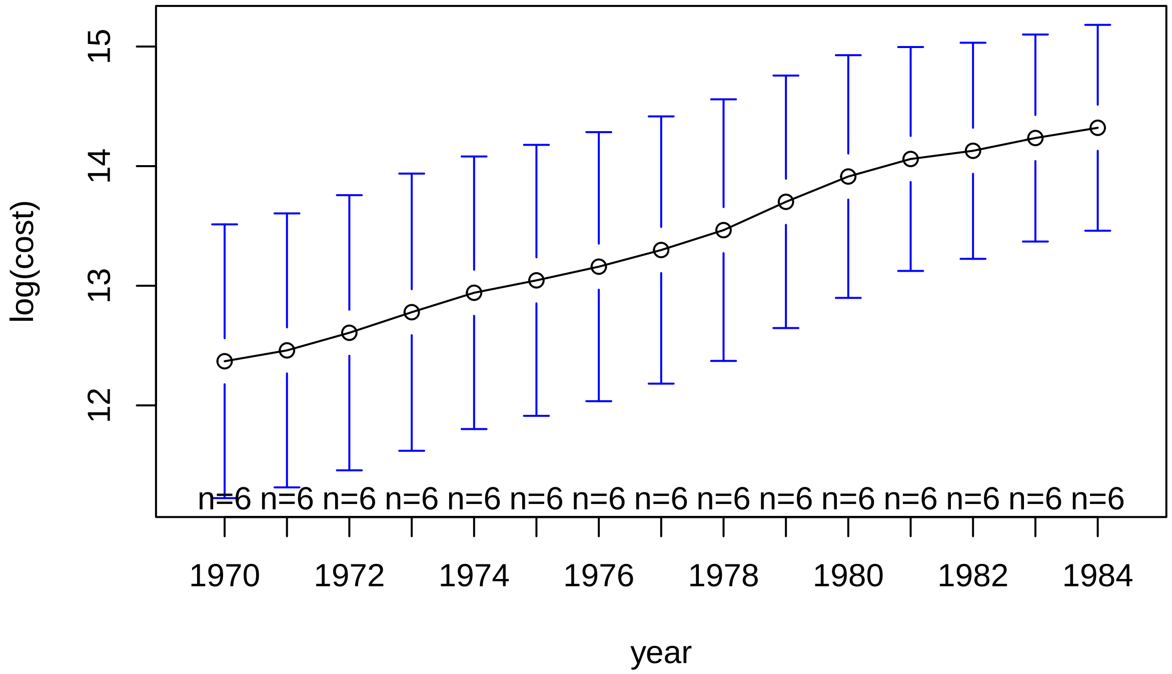
Data Visualization

We can start by exploring heterogeneity across time (year) and individuals/firms in this case (id).

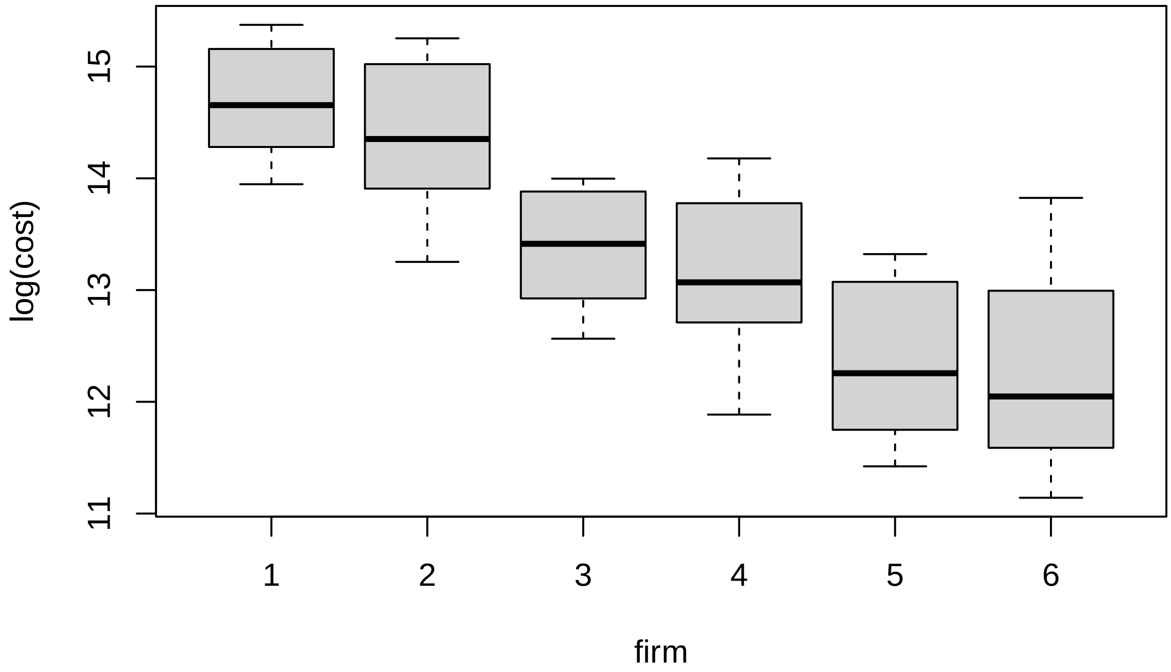
```
data(USAirlines)

# Heterogeneity across time:
scatterplot(log(cost) ~year|firm, data=USAirlines)
```



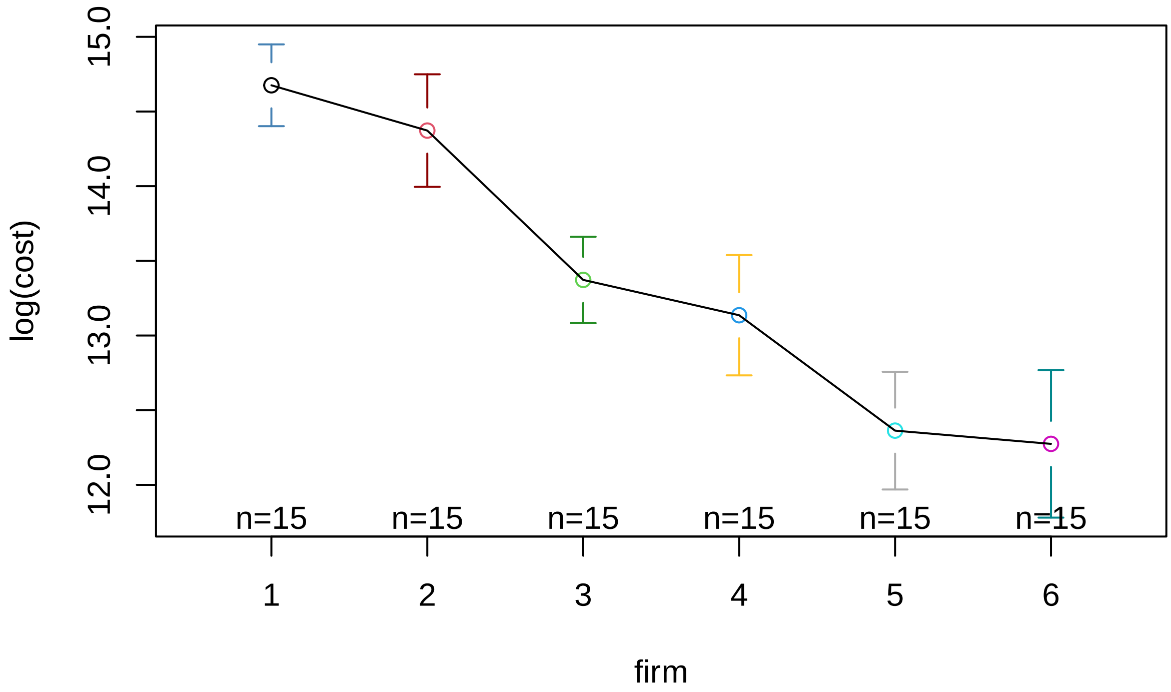
```
plotmeans(log(cost) ~year, data = USAirlines)
```



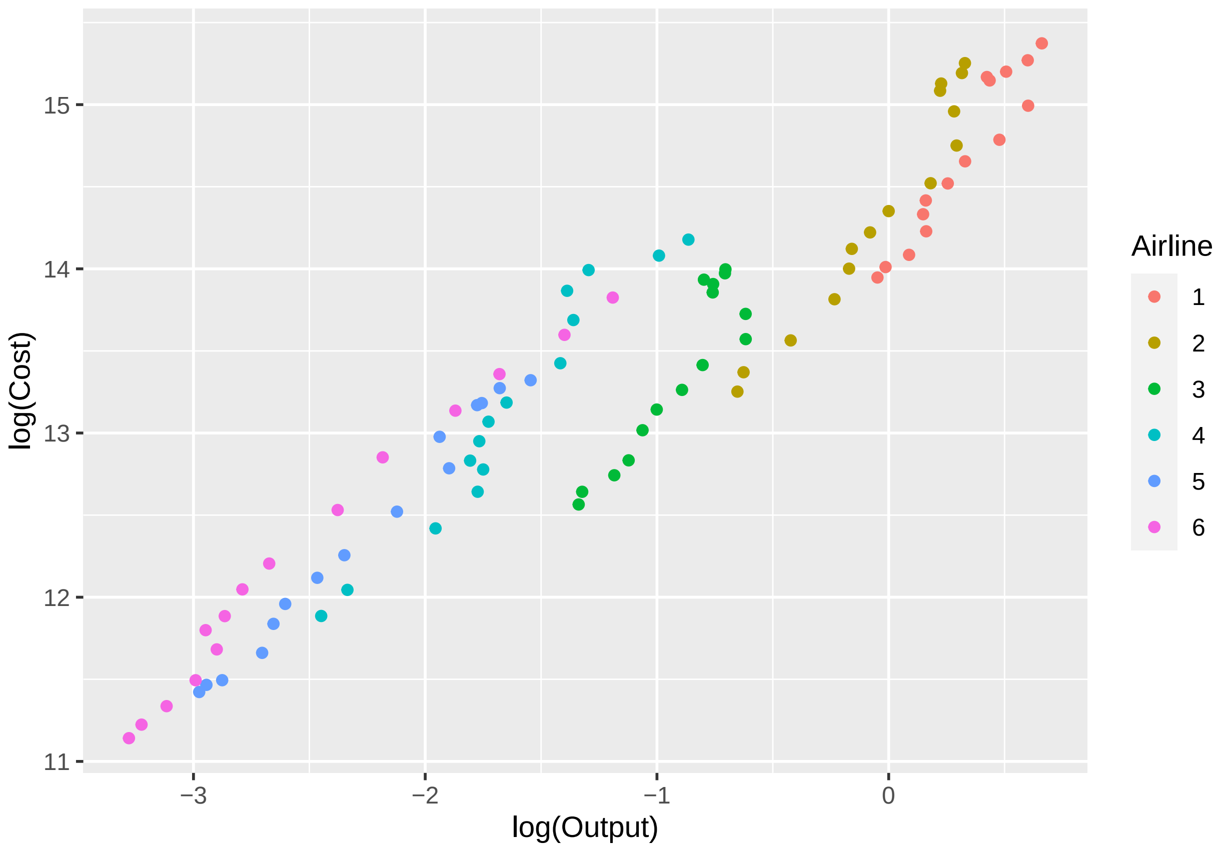
```
# Heterogeneity across firms
scatterplot(log(cost) ~firm|year, data=USAirlines)
```



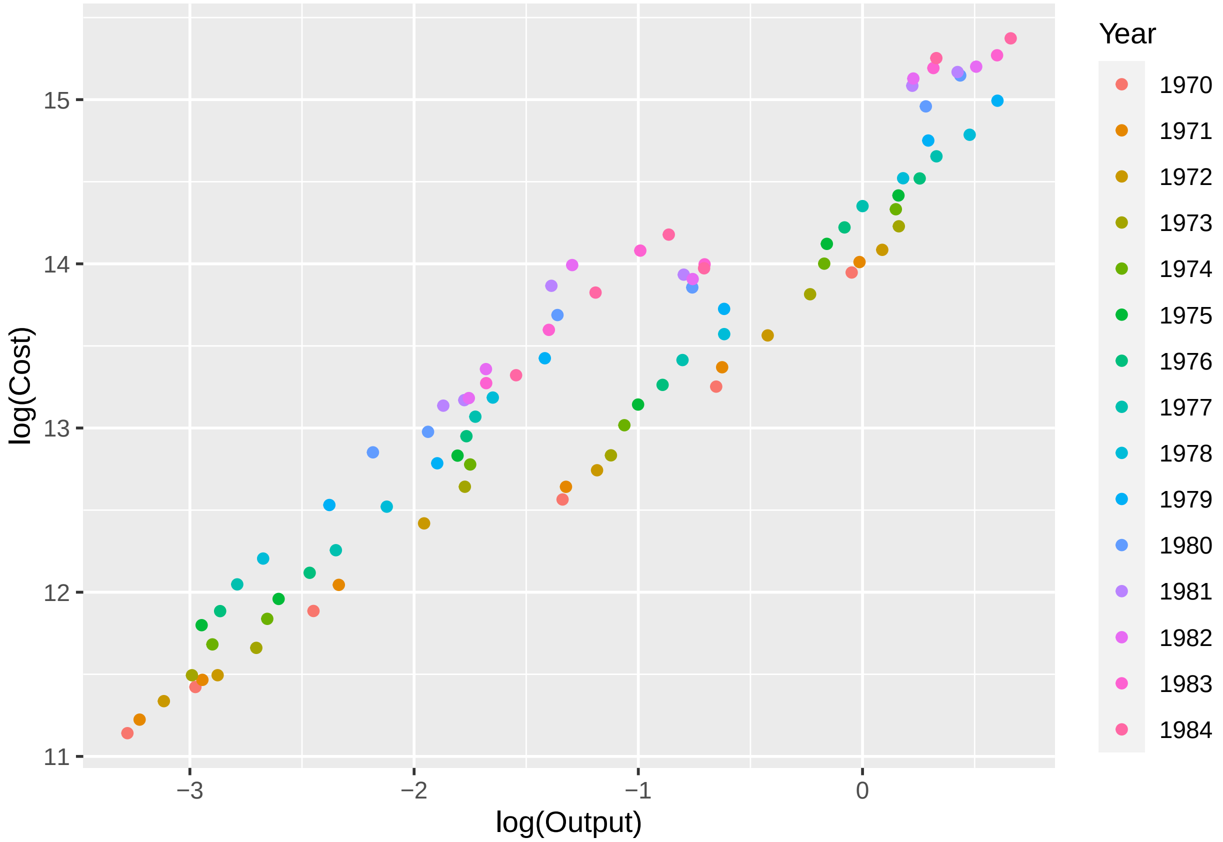
```
plotmeans(log(cost) ~firm, data = USAirlines,col =
palette( c( "steelblue", "darkred", "forestgreen", "goldenrod1", "gray67", "turquoise4" )),
barcol =
palette( c( "steelblue", "darkred", "forestgreen", "goldenrod1", "gray67", "turquoise4" )) )
```



```
# Visualize Cost vs Output by Firm
ggplot(USAirlines, aes(x=log(output), y=log(cost), colour=factor(firm))) +
  geom_point() +
  xlab("log(Output)") +
  ylab("log(Cost)") +
  scale_colour_discrete(name="Airline")
```



```
# Visualize Cost vs Output by Year
ggplot(USAirlines, aes(x=log(output), y=log(cost), colour=factor(year))) +
  geom_point() +
  xlab("log(Output)") +
  ylab("log(Cost)") +
  scale_colour_discrete(name="Year")
```



Fixed Effect Estimator

Using the plm library:

```
usair <- pdata.frame(USAirlines, c("firm", "year"))
fm_full <- plm(log(cost) ~ log(output) + I(log(output)^2) + log(price) + load,
  data = usair, model = "within", effect = "twoways")
fm_time <- plm(log(cost) ~ log(output) + I(log(output)^2) + log(price) + load,
  data = usair, model = "within", effect = "time")
fm_firm <- plm(log(cost) ~ log(output) + I(log(output)^2) + log(price) + load,
  data = usair, model = "within", effect = "individual")
fm_no <- plm(log(cost) ~ log(output) + I(log(output)^2) + log(price) + load,
  data = usair, model = "pooling")
```

Random Effects Model

The random effects model elaborates on the fixed effects model by recognizing that, since the individuals in the panel are randomly selected, their characteristics, measured by the intercept should also be random. The intercept here, unlike the fixed effects model is constant across individuals, but the error term incorporates both individual specifics and the initial regression error term.

```
#Fixed Effects:
fm_firm<-plm(log(cost) ~ log(output) + I(log(output)^2) +
                         log(price) + load, data = usair,
                         model = "within", effect = "individual")
#Random Effects:
fm_rfirm <- plm(log(cost) ~ log(output) + I(log(output)^2) +
                            log(price) + load, data = usair,
                            model = "random")
```

Random Effects Test

We then test for random effects, for which the null is that there are no differences among individuals. This implies that the individual-specific random variable has zero variance. To do this, we use `plmtest`

```
wageReTest <- plmtest(fm_no, effect="individual")
wageReTest
```

```
##
##  Lagrange Multiplier Test - (Honda)
##
## data:  log(cost) ~ log(output) + I(log(output)^2) + log(price) + load
## normal = 18, p-value <2e-16
## alternative hypothesis: significant effects
```

Our test shows that the null hypothesis of zero variance in individual-specific errors is rejected; therefore, heterogeneity among individuals may be significant. So we should not use the pooled model.

Hausman Test

To check which model is more appropriate, of the FEM and REM, we can run a Hausman Test. We assume consistency of our estimates, which implies that they both converge to the true parameter. The null implies that individual random effects are exogenous. Based on a large p-value, we would "fail to reject H0", and conclude that we should use the random effects model.

```
phtest(fm_firm, fm_rfirm)
```

```
##
##  Hausman Test
##
## data:  log(cost) ~ log(output) + I(log(output)^2) + log(price) + load
## chisq = 5.6, df = 4, p-value = 0.2
## alternative hypothesis: one model is inconsistent
```

Our p-value is well above the usual acceptable limits of significance, so we would fail to reject the null, which implies that the individual random effects are exogenous, so we should use the REM.

If this was not the case, and we were instructed to used fixed effects, it is likely that we would want to use the Hausman Taylor estimator. We need to stipulate the instruments we are using for this regression. Note that the number of time-varying variables should be at least as much as the number of time-invariant ones. We would find this by setting "model" to ht in the plm function. Be sure to include all the instruments, i.e. the time-varying and time-invariant exogenous regressors.

Table 1: Linear Probability Model for the *auto* Problem

term	estimate	std.error	statistic	p.value
(Intercept)	0.4848	0.0714	6.785	0
dtime	0.0703	0.0129	5.467	0

Linear Probability Model

1) Using the transport dataset in POE5Rdata, create the following regression:

$$auto = \beta_0 + \beta_1 dtime + e_i$$

2) Correct for heteroskedasticity
3) Use a probit and logit model to evaluate
4) What is the predicted probability with DTIME = 30

```
data(transport)
auto.ols<-lm(auto~dtime,data=transport)
kable(tidy(auto.ols), digits=4,align='c', caption=
  "Linear Probability Model for the $auto$ Problem")
```

```
summary(auto.ols)
```

```
##
## Call:
## lm(formula = auto ~ dtime, data = transport)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6564 -0.1438  0.0278  0.1529  0.8246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4848     0.0714    6.79  1.8e-06 ***
## dtime         0.0703     0.0129    5.47  2.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.327 on 19 degrees of freedom
## Multiple R-squared:  0.611,  Adjusted R-squared:  0.591
## F-statistic: 29.9 on 1 and 19 DF,  p-value: 0.0000283
```

```
cov1 <- hccm(auto.ols,type="hc1")
hcErrors <- coeftest(auto.ols,vcov.=cov1)

hcErrors
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.48480    0.07120    6.81  1.7e-06 ***
## dtime        0.07031    0.00851    8.26  1.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
auto.probit <- glm(auto~dtime,
              data=transport, family=binomial(link="probit"))

predict(auto.ols, newdata=data.frame(dtime = 30))
```

```
##     1
## 2.594
```