Econometrics I - Applications of FWL

Ryan T. Godwin

University of Manitoba

Applications of FWL

FWL helps understand common transformations of variables in econometrics:

- deviations from the mean (de-meaning or centring the variables)
- de-seasonalizing data
- de-trending data

These are non-singular linear transformations. From the invariance property we know that $\bar{\boldsymbol{y}}$ and $\boldsymbol{e}$ will be unchanged. If the transformed data is orthogonal to certain regressors, then the FWL theorem tells us those certain regressors may be dropped from the model.

These transformations can aid in the visualization of the data, interpretation of the estimated parameters, and in some cases ease computational burden.

Centring data (deviations from the mean, or de-meaning)

Simple model:

$$\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x} + \boldsymbol{\epsilon} \tag{1}$$

Can we drop the constant β_1 from model 1? The LS estimators for β_2 would be different under:

$$\boldsymbol{y} = \beta_2 \boldsymbol{x} + \boldsymbol{\epsilon} \tag{2}$$

```r
un <- read.csv("http://rtgodwin.com/data/centrethis.csv")
lm(y ~ x, data=un)
lm(y ~ x -1, data=un)
```

Figure: A least-squares line fitted through some uncentred data. The estimated intercept of $b_1 = 63.7$ is outside the range of the data, and has little economic meaning in most models.

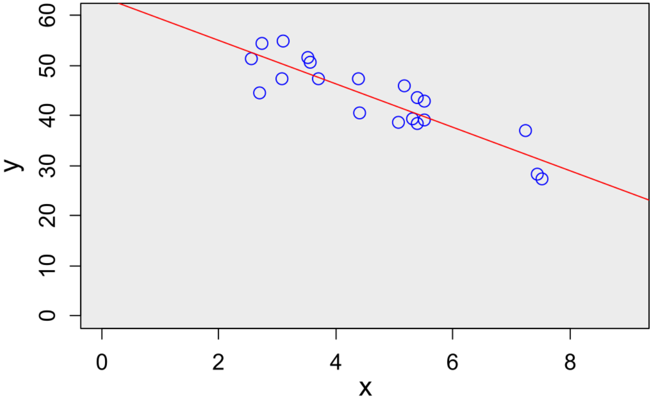
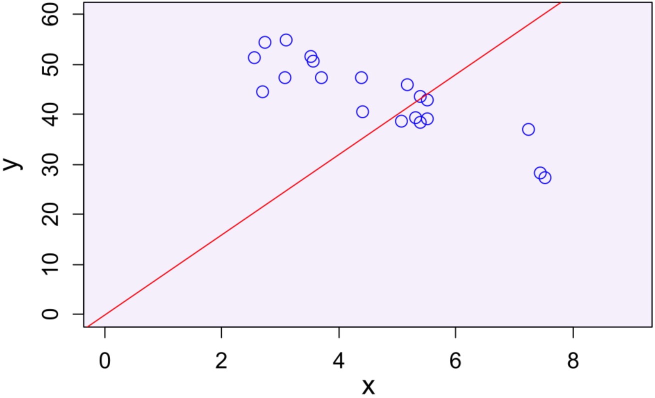


Figure: The least-squares line is forced through the origin if the model does not include an intercept.



Recall that the regression line must pass through the sample means of the data . If the data all has mean zero, then the LS line must pass through the origin anyway, and dropping the intercept has no effect.

To centre data, we *transform* it by subtracting its sample mean:

$$\boldsymbol{y}^\star = \boldsymbol{y} - \boldsymbol{i}\bar{y} \quad ; \quad \boldsymbol{x}^\star = \boldsymbol{x} - \boldsymbol{i}\bar{x}$$

where $\boldsymbol{y}^\star$ and $\boldsymbol{x}^\star$ are the centred variables, and $\boldsymbol{i}$ is a column vector of 1s:

$$\boldsymbol{i} \quad = \quad \underset{(n \times 1)}{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}$$

Estimating the model:

$$\boldsymbol{y}^{\star} = \beta_2 \boldsymbol{x}^{\star} + \boldsymbol{\epsilon} \tag{3}$$

yields identical results to model 1.

The trick is that the variables $\boldsymbol{y}^{\star}$ and $\boldsymbol{x}^{\star}$ are orthogonal to the column vector $\boldsymbol{i}$. The FWL theorem says that exclusion of this regressor ($\boldsymbol{i}$) does not affect the LS estimates.

To prove this, consider the residuals from regressions of $\boldsymbol{x}$ on a constant, and $\boldsymbol{y}$ on a constant. That is, consider the vectors $M_i\boldsymbol{y}$ and $M_i\boldsymbol{x}$.

Let:

$$M_i = I - \boldsymbol{i}\left(\boldsymbol{i}'\boldsymbol{i}\right)^{-1}\boldsymbol{i}' = I - \frac{1}{n}\boldsymbol{i}\boldsymbol{i}' \tag{4}$$

Then,

$$M_i\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{i}\bar{y} = \boldsymbol{y}^{\star}$$

The M_i matrix, when pre-multiplying a vector, creates the deviations-from-means. That is, it centres a variable.

To see how this works, multiply out $M_i\boldsymbol{y}$:

$$M_i\boldsymbol{y} = \left\{ \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} - \begin{bmatrix} 1/n & \dots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \dots & 1/n \end{bmatrix} \right\} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \begin{bmatrix} y_1 & - & y_1/n & - & y_2/n & - & \dots & - & y_n/n \\ & & & & \vdots & & & & \\ y_n & - & y_1/n & - & y_2/n & - & \dots & - & y_n/n \end{bmatrix} = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$$

The transformed (centred) variables are now orthogonal to the regressor $\boldsymbol{i}$, that is $(M_i\boldsymbol{y})'\boldsymbol{i} = 0$ and $(M_i\boldsymbol{x})'\boldsymbol{i} = 0$, and so the intercept may be dropped from the model without any substantive effect.

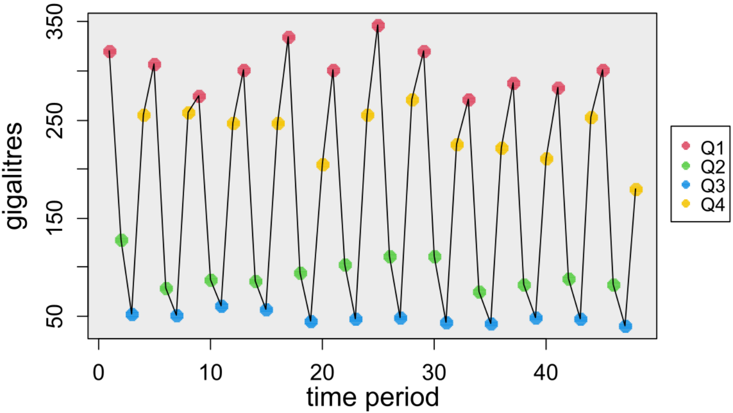
De-seasonalizing

- Time series data that is reported quarterly may have seasonal effects.
- Certain activities only take place in the summer, more presents are purchased in the 4th quarter, etc.
- Seasonality at other frequencies as well (e.g. monthly)
- It is often desirable to visualize, and work with, data that has been de-seasonalized. That is, we sometimes want to *purge* the seasonal patterns from data, and explain variation in the data that is independent from seasonal variation.

For example, see Figure 3 for the quarterly residential demand for natural gas in **Manitoba**, from 1990 Q1 to the end of 2001 Q4[1]. There is a strong seasonal component.

[1] Data from: Statistics Canada. Table 25-10-0005-01 Supply and demand of primary and secondary energy in natural units, quarterly, with data for years 1990 - 2001

Figure: Quarterly seasonality in the residential demand for natural gas in Manitoba.



To produce a graph similar to that in Figure 3, download the data in R:

```r
gas <- read.csv("http://rtgodwin.com/data/MBgas.csv")
```

Create a time-trend variable, and plot the gas consumption over time:

```r
gas$time <- 1:nrow(gas)
plot(gas$time, gas$gigalitres, type = "l",
  xlab = "time period", ylab = "gigalitres")
```

We may want to *de-seasonalize* the data. Can use quarterly dummy variables. Let $\boldsymbol{q}_1$ be a dummy variable equal to 1 if the reference period is in the first quarter, and 0 otherwise. Similar definitions follow for the dummy variables $\boldsymbol{q}_2$, $\boldsymbol{q}_3$, and $\boldsymbol{q}_4$. That is, we need to create dummy variables that look like:

$$\boldsymbol{q}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \quad ; \quad \boldsymbol{q}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix} \quad ; \quad \boldsymbol{q}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \quad ; \quad \boldsymbol{q}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

Beware “dummy variable trap”:

$$\boldsymbol{q}_1 + \boldsymbol{q}_2 + \boldsymbol{q}_3 + \boldsymbol{q}_4 = \boldsymbol{i} \tag{5}$$

Including all 4 dummies, and the intercept, would be a violation of A.2 (no perfect multicollinearity): the $(X'X)$ matrix is not invertible and the LS estimator is undefined. We must *exclude* one of the dummy variables, or exclude the intercept.

Question: Suppose we estimate a model where we exclude $\boldsymbol{q}_1$. What would change if we instead decided to exclude $\boldsymbol{q}_2$, or the intercept?

To de-seasonalize the data, we can regress the seasonal variable on the system of dummy variables, and extract the residuals. That is, the de-seasonalized variable is $M_Q\boldsymbol{y}$, where Q consists of the regressors representing the seasonal dummy variables. For example:

$$Q = \left[\begin{array}{cccc} \boldsymbol{i} & \boldsymbol{q}_2 & \boldsymbol{q}_3 & \boldsymbol{q}_4 \end{array}\right] = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{array}\right]$$

$M_Q\boldsymbol{y}$ consists of variation that is independent of the seasonal component. The de-seasonalized variable is now *orthogonal* to the seasonal dummies.

To accomplish this de-seasonalization in R, we first create the system of dummies:

```
gas$q4 <- gas$q3 <- gas$q2 <- gas$q1 <- 0
gas$q1[seq(1, n, 4)] <- 1
gas$q2[seq(2, n, 4)] <- 1
gas$q3[seq(3, n, 4)] <- 1
gas$q4[seq(4, n, 4)] <- 1
```

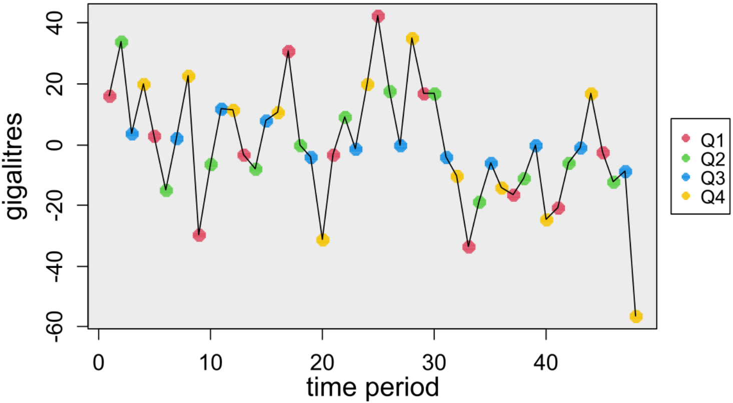
and then extract the residuals from the regression of the time series on the seasonal dummies:

```
Mgigalitres <- lm(gigalitres ~ q2 + q3 + q4, data = gas)$residuals
```

Now plot the de-seasonalized variable over time (see Figure 4):

```
plot(gas$time, Mgigalitres, type = "l")
```

Figure: De-seasonalized time series of the residential demand for gas in MB.



Questions:

1. What is the mean value of the de-seasonalized data?
2. What is the mean value of the de-seasonalized data, for the 1st quarter?
3. When is it acceptable to "drop" seasonal dummy variables from a model?

To answer question 3, consider what is *wrong* with estimating the model:

$$gas^{\star} = \beta_1 + \beta_2 temp + \epsilon \tag{6}$$

where *temp* is the mean quarterly temperature at Richardson International airport obtained from Environment Canada, and $gas^{\star} = M_q gas$ is the de-seasonalized demand for MB gas.

Estimate this model in R:

```r
summary(lm(Mgigalitres ~ gas$temp))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6650     2.8550   0.233    0.817
gas$temp     -0.2350     0.2397  -0.980    0.332

Residual standard error: 19.21 on 46 degrees of freedom
Multiple R-squared:  0.02046, Adjusted R-squared:  -0.0008304
F-statistic: 0.961 on 1 and 46 DF,  p-value: 0.3321
```

Notice that the variable `gas$temp` is *insignificant*.

Consider instead the regression model:

$$gas^{\star} = \beta_1 + \beta_2 temp^{\star} + \epsilon \tag{7}$$

where $temp^{\star} = M_Q temp$ and has been de-seasonalized. In R:

```r
Mtemp <- lm(temp ~ q2 + q3 + q4, data = gas)$residuals
summary(lm(Mgigalitres ~ Mtemp - 1))
```

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
Mtemp  -8.0933     0.7641  -10.59 4.83e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.44 on 47 degrees of freedom
Multiple R-squared:  0.7048,  Adjusted R-squared:  0.6985
F-statistic: 112.2 on 1 and 47 DF,  p-value: 4.835e-14
```

Notice that `Mtemp` is *significant*.

Questions:

1. Which model is the “correct” one to estimate, model 6 or model 7? Why?
2. Why has the intercept been dropped in model 7?
3. In general, what might be the problem with using de-seasonalized data?