1 Difference-in-Differences

This exercise uses data from Feb-Mar and Nov-Dec 1992 on employment at fast food restaurants in the US States of New Jersey and Pennsylvania taken from Card and Krueger in The American Economic Review, Vol. 84(4). The data are described in the output below. We are interested in how the minimum wage affects employment decisions in these restaurants. In April 1992, New Jersey increased the minimum wage from \$4.25 to \$5.05. In Pennsylvania, the minimum wage was unchanged, and we assume it to be \$3.80 for this exercise.

Problem 1.1. Describe the data.

Solution. Using the -codebook- command, we learn the following about each variable:

1. "sheet" is an integer denoting store ID.
   - ▷ There are 408 unique stores identified.
2. "post" is a binary variable which is equal to 1 if the observation is after the law and 0 if the observation is before the law.
   - ▷ There are 408 observations with value equal to 1 and 408 with value equal to 0. This tells us that the panel is balanced.
3. "chain" is a number between 1 and 4 that denotes the specific store name.
   - ▷ Value of 1 denotes Burger King, and there are 342 observations.
   - ▷ Value of 2 denotes KFC, and there are 158 observations.
   - ▷ Value of 3 denotes Roy Rogers, and there are 198 observations.
   - ▷ Value of 4 denotes Wendy's, and there are 118 observations.
4. "state" is a binary variable that is equal to 1 if the observation is from NJ and 0 if the observation is from PA.
   - ▷ 156 out of 816 observations are from NJ, whereas 660 are from PA.
5. "empft" is a numeric variable $\in [0, 60]$ that represents the number of full-time employees.
   - ▷ The mean number of full-time employees is 8.26, and the 90th percentile is 20. This suggests that the distribution of full-time employees is highly skewed.
6. "hrsopen" is a numeric variable $\in [7, 24]$ that is the number of hours open per day.
   - ▷ The mean number of hours open per day is 14.4717, and the 90th percentile is 17 hours.
7. "nregs" is a numeric variable $\in [2, 8]$ that is the number of cash registers in the store.
   - ▷ Majority of the observations fall between 2 and 5. There are also 28 missing observations.
8. "minwage" is numeric variable $\in [3.8, 5.05]$ that is the minimum wage rate in US dollars.
   - ▷ It only takes on three values: 3.8, 4.25, and 5.05. Since the panel is balanced, we have the same number of observations for 4.25 and 5.05.

9. “temp” is a binary variable that is equal to 1 if more than 75% of the employees are part-time.

   ▷ 406 out of 816 observations are related to stores with more than 75% employees part-time.

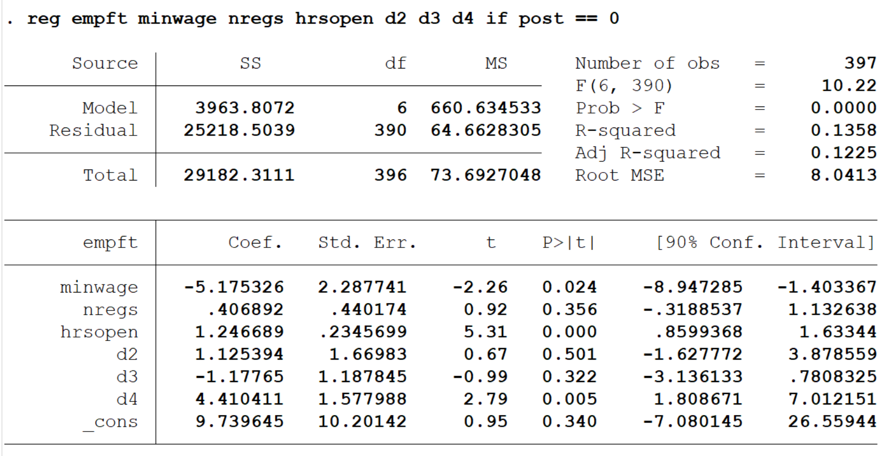
10. “d1” through “d4” are binary variables denoting the specific chain of the restaurant.

Problem 1.2. Estimate the following regression on the sample of fast food restaurants in Feb-Mar 1992:

$$\text{empft}_{ikt} = \alpha + \gamma\text{minwage}_{kt} + \beta_1\text{nregs}_{ikt} + \beta_2\text{hrsopen}_{ikt} + \sum_{j=2}^{4}\eta_j d_j + \epsilon_{ikt}$$

where i denotes a restaurant, k denotes a state, and $t = 0$ if the observation is from Feb-Mar and $t = 1$ if the observation is from Nov-Dec.

Solution. We run the following regression:



```
. reg empft minwage nregs hrsopen d2 d3 d4 if post == 0

      Source |       SS           df       MS      Number of obs   =       397
-------------+----------------------------------   F(6, 390)       =     10.22
       Model |   3963.8072         6  660.634533   Prob > F        =    0.0000
    Residual |  25218.5039       390  64.6628305   R-squared       =    0.1358
-------------+----------------------------------   Adj R-squared   =    0.1225
       Total |  29182.3111       396  73.6927048   Root MSE        =    8.0413

------------------------------------------------------------------------------
       empft |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
     minwage |  -5.175326   2.287741    -2.26   0.024    -8.947285   -1.403367
       nregs |    .406892    .440174     0.92   0.356    -.3188537    1.132638
     hrsopen |   1.246689   .2345699     5.31   0.000     .8599368     1.63344
          d2 |   1.125394    1.66983     0.67   0.501    -1.627772    3.878559
          d3 |   -1.17765   1.187845    -0.99   0.322    -3.136133    .7808325
          d4 |   4.410411   1.577988     2.79   0.005     1.808671    7.012151
       _cons |   9.739645   10.20142     0.95   0.340    -7.080145    26.55944
------------------------------------------------------------------------------
```

▷ The estimate of γ is -5.175.

Problem 1.3. Interpret the coefficient γ and calculate at 90% confidence interval.

Solution. γ reflects the impact on employment per dollar change in the minimum wage rate. The coefficient γ of -5.175 has a t-statistic of -2.26, implying that we can reject the null hypothesis and thereby argue that the parameter of interest is statistically distinct from zero. The 90% confidence interval is given as

$$[-8.947285, -1.403367]$$

which does not include zero. We can interpret the magnitude of the coefficientas saying that a \$1 increase in minimum wage is associated with a decrease in -5.175 employeers after controlling for restaurant size (proxied by the number of cash registers in the store), hours of operation, and restaurant type. Note that we cannot interpret this estimate as causal since it is likely to suffer from an omitted variable bias.

Problem 1.4. Use the Sum of squares table from the regression output to calculate the R^2 and the standard error of the regression (Root MSE).

Solution. The R^2 can be computed as

$$R^2 = \frac{MSS}{TSS} = \frac{3963.8072}{29182.3111} = 13.58\%$$

The standard error of the regressor is given as

$$SE = \sqrt{MSE} = \sqrt{64.6628305} = 8.041$$

Problem 1.5. Give an economic interpretation of the coefficients $\eta_2 \sim \eta_4$. What might explain the relatively large coefficient on d_4?

Solution. Note that $d_2 = 1$ only if the observation corresponds to KFC; $d_3 = 1$ only if the observation corresponds to Roy Rogers; and $d_4 = 1$ only if the observation is from Wendy's. Each coefficient can be interpreted as additional increase in employees from a \$1 increase in minimum wage when compared to Burger King. For example, since $\eta_2 = 1.125$, a \$1 increase in the minimum wage for KFC is associated with 1.125 more employees than the increase in employees for Burger King.

Note that η_4 is larger than η_2 and η_3. This could be (1) Wendy's sells types of food that is more labor intensive (ex. Gourmet burger vs. ready-made burger) or (2) has less automation in operations of the branch.

Problem 1.6. Test $H_0 : \eta_2 = \eta_3 = 0$.

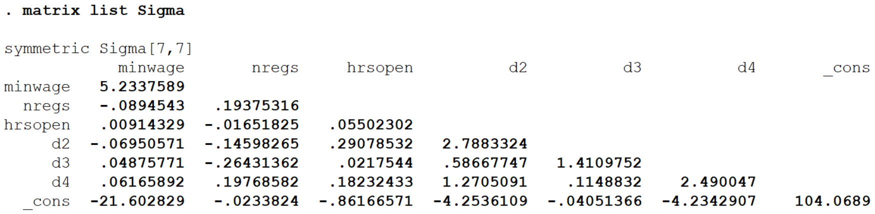
Solution. We test H_0 using Stata. We obtain an F-statistic value of 1.00, which corresponds to a p-value of 0.3673. Since the p-value is greater than 0.05, we *cannot* reject the hypothesis H_0 at the $\alpha = 0.05$ level.

Problem 1.7. Test $H_0 : \eta_2 = \eta_3$ using the estimated covariance matrix of the coefficients. Verify your answer by running the test in Stat using -test- and/or p by performing an F-test.

Solution. Alternatively, we can use the covariance matrix of the coefficients to calcualte the standard error of $\hat{\eta}_2 - \hat{\eta}_3$ to test whether this is significantly different form zero. Since:

$$\text{Var}\left[\hat{\eta}_2 - \hat{\eta}_3\right] = \text{Var}\left[\hat{\eta}_2\right] + \text{Var}\left[\hat{\eta}_3\right] + 2\,\text{Cov}\left[\hat{\eta}_2, -\hat{\eta}_3\right]$$

and we have the following covariance matrix:



```
. matrix list Sigma

symmetric Sigma[7,7]
            minwage       nregs     hrsopen          d2          d3          d4       _cons
minwage   5.2337589
  nregs   -.0894543   .19375316
hrsopen   .00914329  -.01651825   .05502302
     d2  -.06950571  -.14598265   .29078532   2.7883324
     d3   .04875771  -.26431362    .0217544   .58667747   1.4109752
     d4   .06165892   .19768582   .18232433   1.2705091    .1148832    2.490047
  _cons  -21.602829   -.0233824  -.86166571  -4.2536109  -.04051366  -4.2342907    104.0689
```

▷ Since $\text{Var}\left[\hat{\eta}_2\right] = 2.7883324$, $\text{Var}\left[\hat{\eta}_3\right] = 1.4109752$, and $\text{Cov}\left[\hat{\eta}_2, -\hat{\eta}_3\right] = -0.58667747$, we have

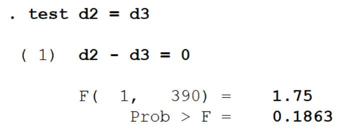
$$\text{Var}\left[\hat{\eta}_2 - \hat{\eta}_3\right] = 3.0260 \Rightarrow SE\left[\hat{\eta}_2 - \hat{\eta}_3\right] = 1.73954$$

▷ The t-statistic is then given as

$$\frac{1.125 + 1.178}{1.73954} = 1.324$$

which has a p-value of 0.1863 under the t distribution with 369 degrees of freedom.

Running the F-test in Stata, we obtain:



```
. test d2 = d3

 ( 1)  d2 - d3 = 0

       F(  1,   390) =    1.75
            Prob > F =    0.1863
```

and thus we verify that the calculation is indeed correct.

We now want to control for potential selection issues by using the panel structure of our data.

Problem 1.8. Explain why the previous estimate of γ is likely to suffer from omitted variable bias.

Solution. Given the level of autonomy that each state enjoys, it is very likely that there are non-wage state-level regulations that affect both employment and the minimum wage. For example, if one state announced a subsidy on the fast-food industry while the other did not, our OLS regression would not be taking this effect into account.

Problem 1.9. Assume that $\epsilon_{ikt} = \mu_k + \zeta_t + u_{ikt}$ and that $\mathbb{E}\left[u_{ikt} | X_{ikt}\right] = 0$ where X_{ikt} is the vector of RHS variables except -minwage-. Explain how you can then use the increase in the minimum wage in New Jersey and a difference-in-differences (DD) model to identify the effect of the minimum wage on employment. Given an example where the necessary assumption(s) are violated.

Solution. Now we assume that $\epsilon_{ikt} = \mu_k + \zeta_t + u_{ikt}$ and that $\mathbb{E}\left[u_{ikt} | X_{ikt}\right] = 0$ where X_{ikt} is the vector of RHS variables except -minwage-. The key idea is that DiD allows us to control for (1) unobservables that change similarly across states and (2) unobservables that are constant within the state.

▷ We plug this parametric specification into our original regression specification to obtain:

$$\text{empft}_{ikt} = \alpha + \gamma \text{minwage}_{kt} + \beta_1 \text{nregs}_{ikt} + \beta_2 \text{hrsopen}_{ikt} + \sum_{j=2}^{4} \eta_j d_j + \left(\mu_k + \zeta_t + u_{ikt}\right)$$

▷ To deal with μ_k and ζ_t which are unobserved, we can use a difference-in-differences model to estimate the effect of the minimum wage.

- * First, for each $k \in [NJ, PA]$, estimate the change $\text{empft}_{ik2} - \text{empft}_{ik1} \equiv \Delta\text{empft}_{ik}$:

$$\Delta\text{empft}_{ik} = \gamma\,[\Delta\text{minwage}_{k}] + \beta_1\,[\Delta\text{nregs}_{ik}] + \beta_2\,[\Delta\text{hrsopen}_{ik}] + \Delta\zeta + \Delta u_{ik}$$

- * Second, compute the difference between $\Delta\text{empft}_{i[NJ]}$ and $\Delta\text{empft}_{i[PA]}$:

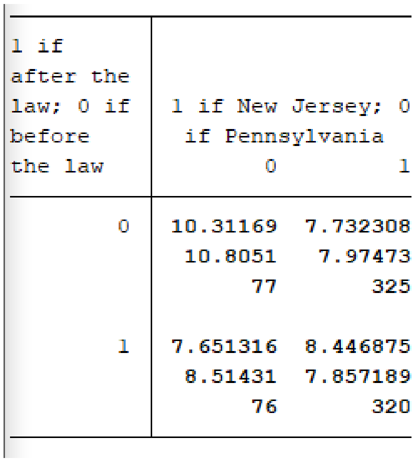
$$\begin{aligned}\Delta\text{empft}_{i[NJ]} - \Delta\text{empft}_{i[PA]} = \gamma\,[\Delta\text{minwage}_{NJ} - \Delta\text{minwage}_{PA}] + \beta_1 \left[\Delta\text{nregs}_{i[NJ]} - \Delta\text{nregs}_{i[PA]}\right] \\ + \beta_2 \left[\Delta\text{hrsopen}_{i[NJ]} - \Delta\text{hrsopen}_{i[PA]}\right] + \left[\Delta u_{i[NJ]} - \Delta u_{i[PA]}\right]\end{aligned}$$

▷ The key step in the computation above is that $\Delta\zeta \equiv \zeta_1 - \zeta_0$ is the same for both NJ and PA. This is the common trend assumption.

▷ An example in which the common trend assumption is violated when there is a non-wage state-level regulation change that affects the employment trend. One example could be a subsidy to these restaurants that only happens in NJ or PA. Another could be a health-care reform that raises the cost of employing for each firm that happens for only one state.

Problem 1.10. Generate a table of means, a table of standard errors and a table of frequencies for -empft- in each state and each time period (post = 1 and post = 0).

Solution. We obtain the following table:

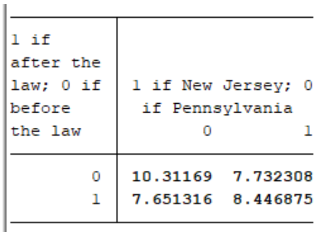


```
1 if       |
after the  |
law; 0 if  |  1 if New Jersey; 0
before     |   if Pennsylvania
the law    |          0          1
-----------+-----------------------
         0 |   10.31169   7.732308
           |    10.8051    7.97473
           |         77        325
           |
         1 |   7.651316   8.446875
           |    8.51431   7.857189
           |         76        320
```

Note that the first row in each cell is the mean; the second row is the standard error; and the last row is the number of observations (frequencies) for corresponding observations.

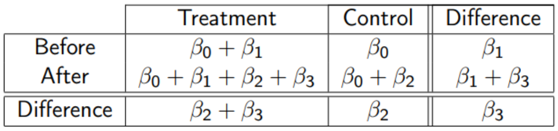
Problem 1.11. Using these statistics, calculate a DD estimate of the impact of the minimum wage law on employment.

Solution. Just focusing on the means, we obtain the following table:



```
1 if       |
after the  |
law; 0 if  |  1 if New Jersey; 0
before     |   if Pennsylvania
the law    |          0          1
-----------+-----------------------
         0 |   10.31169   7.732308
         1 |   7.651316   8.446875
```

- ▷ We can compare the result to the following table from the lecture:



	Treatment	Control	Difference
Before	$\beta_0 + \beta_1$	β_0	β_1
After	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$	$\beta_1 + \beta_3$
Difference	$\beta_2 + \beta_3$	β_2	β_3

- ▷ To get our desired estimate – which is β_3 in the table from lecture – we can compute the following difference:

$$(8.446875 - 7.732308) - (7.651316 - 10.31169) = 3.374941$$

Thus, the point DD estimate is 3.374941. A $1 increase in the minimum wage is associated with an increase in employment of around 3.375 employees.

Problem 1.12. Specify and estimate the corresponding regression.

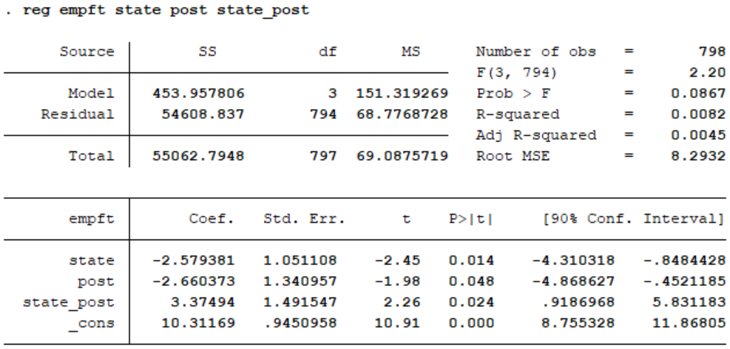
Solution. The corresponding regression is of the following form:

$$\text{empft}_{ikt} = \alpha + \beta_1 \text{state}_{ikt} + \beta_2 \text{post}_{ikt} + \beta_3 \left(\text{state}_{ikt} \times \text{post}_{ikt}\right) + \nu_{ikt}$$

Note that to be consistent with the previous steps, we do not add covariates explicitly. But the regression formulation lends itself to covariates very naturally:

$$\text{empft}_{ikt} = \alpha + \beta_1 \text{state}_{ikt} + \beta_2 \text{post}_{ikt} + \beta_3 \left(\text{state}_{ikt} \times \text{post}_{ikt}\right) + X_{kt}^{'} \pi + \nu_{ikt}$$

The result from the stata is the following:



```
. reg empft state post state_post

      Source |       SS           df       MS      Number of obs   =       798
-------------+----------------------------------   F(3, 794)       =      2.20
       Model |  453.957806         3  151.319269   Prob > F        =    0.0867
    Residual |   54608.837       794  68.7768728   R-squared       =    0.0082
-------------+----------------------------------   Adj R-squared   =    0.0045
       Total |  55062.7948       797  69.0875719   Root MSE        =    8.2932

------------------------------------------------------------------------------
       empft |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
       state |  -2.579381   1.051108    -2.45   0.014    -4.310318   -.8484428
        post |  -2.660373   1.340957    -1.98   0.048    -4.868627   -.4521185
  state_post |    3.37494   1.491547     2.26   0.024     .9186968    5.831183
       _cons |   10.31169   .9450958    10.91   0.000     8.755328    11.86805
------------------------------------------------------------------------------
```

- ▷ We find $\beta_3 = 3.37494$ which corresponds to the manually computed DD estimate from above.

Problem 1.13. How much does this suggest that the minimum wage affects full time employment in fast food restaurants?

Solution. The DiD point estimate of 3.37494 corresponds to an increase in employment. This result suggests that the increase in minimum wage from $4.25 to $5.05 increased full-time employment among the fast-food restaurants by 3.37494 employees.

Problem 1.14. Explain why the t-test from the regression above may understate the uncertainty in the effect of the minimum wage on full time employment. How could you correct the standard error? Compare the t-values with and without this correction.

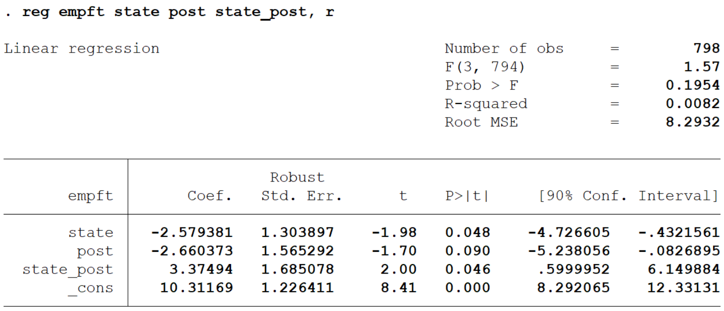
Solution. There can be two main issues: group-time correlation and serial correlation.

1. Group-time correlation: Observations in state k at time t are unlikely to be independent, since state-time random shocks can cause correlation between residuals within state-time. To deal with this issue, one can assume $\epsilon_{ikt} = \nu_{kt} + e_{ikt}$ and use WLS, or use cluster-robust standard errors.

2. Serial correlation also may be an issue since the outcome variable (employment) is typically highly positively serially correlated.

In both cases, the correlation deflates the standard errors, thereby understating the uncertainty in our estimate. The intuition is that previously you were not allowing for correlation between observations, so once you take this into account, the standard errors should increase.

To correct for this, we may consider cluster the standard errors at the state-time level. Since cluster-robust variance estimators are only valid when the number of clusters is sufficiently large, however, and since there are only 4 groups generated by clustering at the state-time level, this step is not recommended. Instead, I adjust for heteroskedasticity using the robust command in Stata:



```
. reg empft state post state_post, r

Linear regression                               Number of obs     =        798
                                                F(3, 794)         =       1.57
                                                Prob > F          =     0.1954
                                                R-squared         =     0.0082
                                                Root MSE          =     8.2932

------------------------------------------------------------------------------
             |               Robust
       empft |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
       state |  -2.579381   1.303897    -1.98   0.048    -4.726605   -.4321561
        post |  -2.660373   1.565292    -1.70   0.090    -5.238056   -.0826895
  state_post |    3.37494   1.685078     2.00   0.046     .5999952    6.149884
       _cons |   10.31169   1.226411     8.41   0.000     8.292065    12.33131
------------------------------------------------------------------------------
```

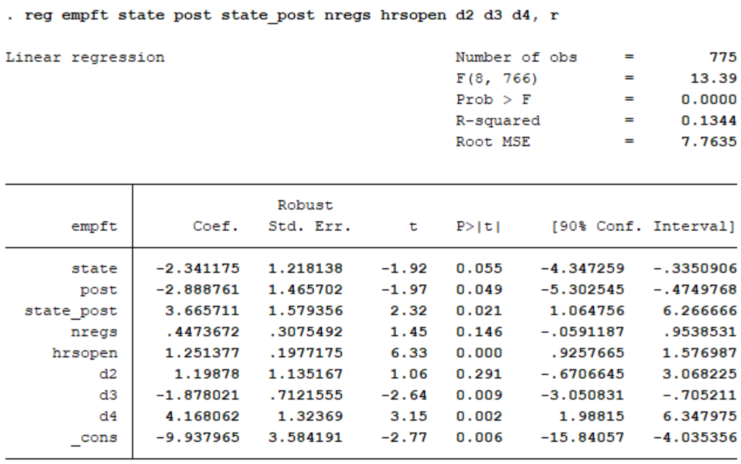
Note that the coefficient of interest is still the coefficient on *state_post*. If we had more groups we can cluster or do block bootstrap.

Problem 1.15. What regression would you run to estimate the DD model including control variables? Run the regression using robust standard errors.

Solution. We would run a regression of the following form:

$$\text{empft}_{ikt} = \alpha + \beta_1 \text{state}_{ikt} + \beta_2 \text{post}_{ikt} + \beta_3 \left(\text{state}_{ikt} \times \text{post}_{ikt}\right) + X_{kt}^{'}\pi + \nu_{ikt}$$

where X includes nregs and hrsopen and the dummies corresponding to each restaurant chain We obtain the following output from Stata:



```
. reg empft state post state_post nregs hrsopen d2 d3 d4, r

Linear regression                               Number of obs     =        775
                                                F(8, 766)         =      13.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1344
                                                Root MSE          =     7.7635

-------------------------------------------------------------------------------
             |               Robust
       empft |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+-----------------------------------------------------------------
       state |  -2.341175   1.218138    -1.92   0.055    -4.347259   -.3350906
        post |  -2.888761   1.465702    -1.97   0.049    -5.302545   -.4749768
  state_post |   3.665711   1.579356     2.32   0.021     1.064756    6.266666
       nregs |   .4473672   .3075492     1.45   0.146    -.0591187    .9538531
     hrsopen |   1.251377   .1977175     6.33   0.000     .9257665    1.576987
          d2 |    1.19878   1.135167     1.06   0.291    -.6706645    3.068225
          d3 |  -1.878021   .7121555    -2.64   0.009    -3.050831    -.705211
          d4 |   4.168062    1.32369     3.15   0.002      1.98815    6.347975
       _cons |  -9.937965   3.584191    -2.77   0.006    -15.84057   -4.035356
-------------------------------------------------------------------------------
```

▷ We obtain a point estimate of 3.665711, which is close to the DiD estimate without covariates from before (3.3749).

We could also consider interacting the covariates with the state dummy. This allows there to be differential slope on the covariates for each state. The trade-off is that now there are more things to estimate if we include them.

Problem 1.16. How might you test the key identifying assumptions underlying your DiD estimation in this application, and in general?

Solution. The key identifying assumption is that of common trends, but this is inherently untestable since we don't see what employment in NJ would have been at post $= 1$ if the minimum wage had not been raised. One way to get as close to testing is to look at the pre-treatment values of employment over time to see if both states display similar time trends. If they do not, then it would be difficult to believe that they would be displaying similar trends after the treatment.

Another test we may consider is the one analogous to the one that we saw in class where we examine the post-trend after the impact of minimum wage has effectively "died out." We saw an example like this for the Card & Krueger (2000).