novis-i-hwk2-2

February 18, 2025

ECON 470 Hwk2-2

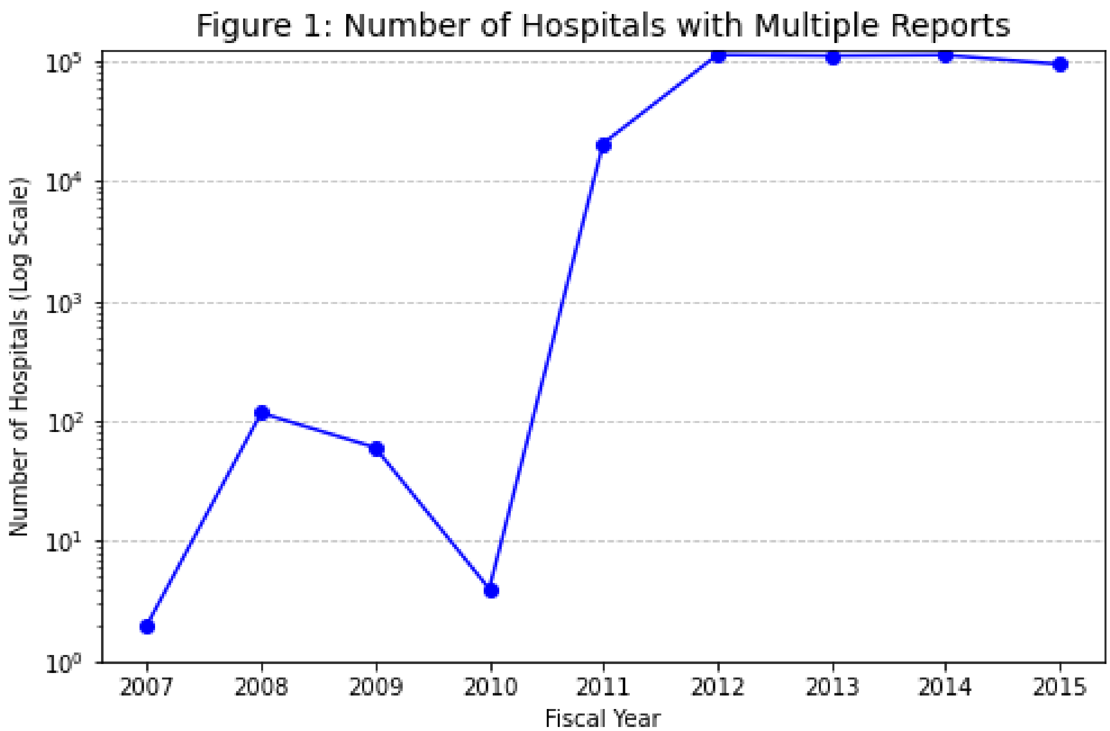
Author: Ilse Novis

Date: 2/19/2025

GitHub Repository

Question 1:
How many hospitals filed more than one report in the same year? Show your answer as a line graph of the number of hospitals over time.

```
Number of distinct providers: 6731
```

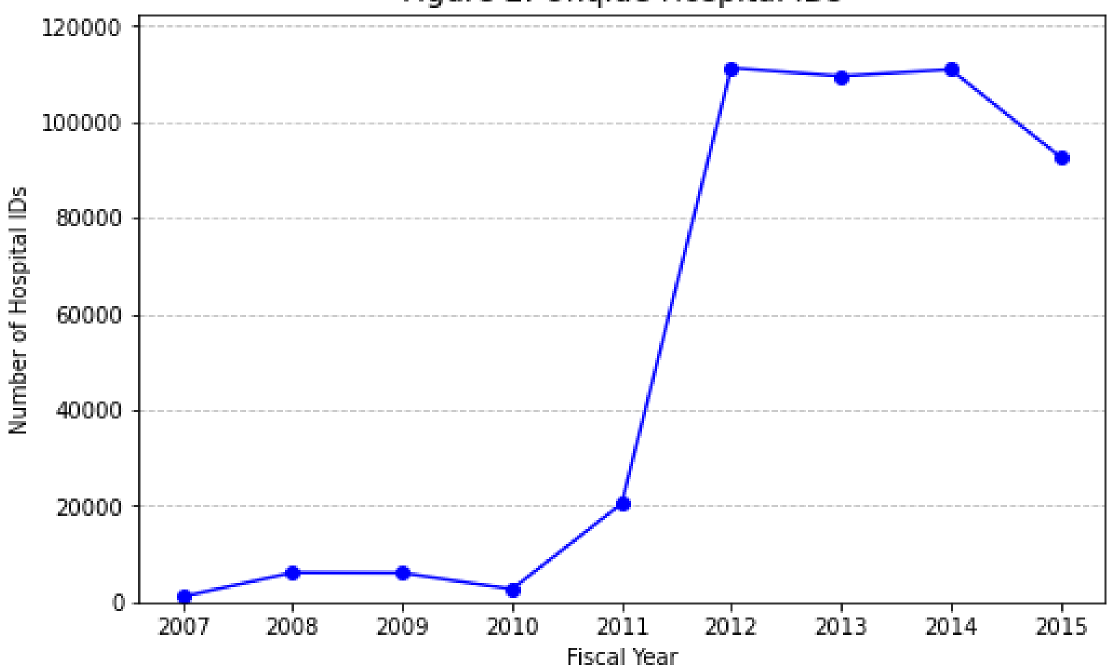


Question 2:

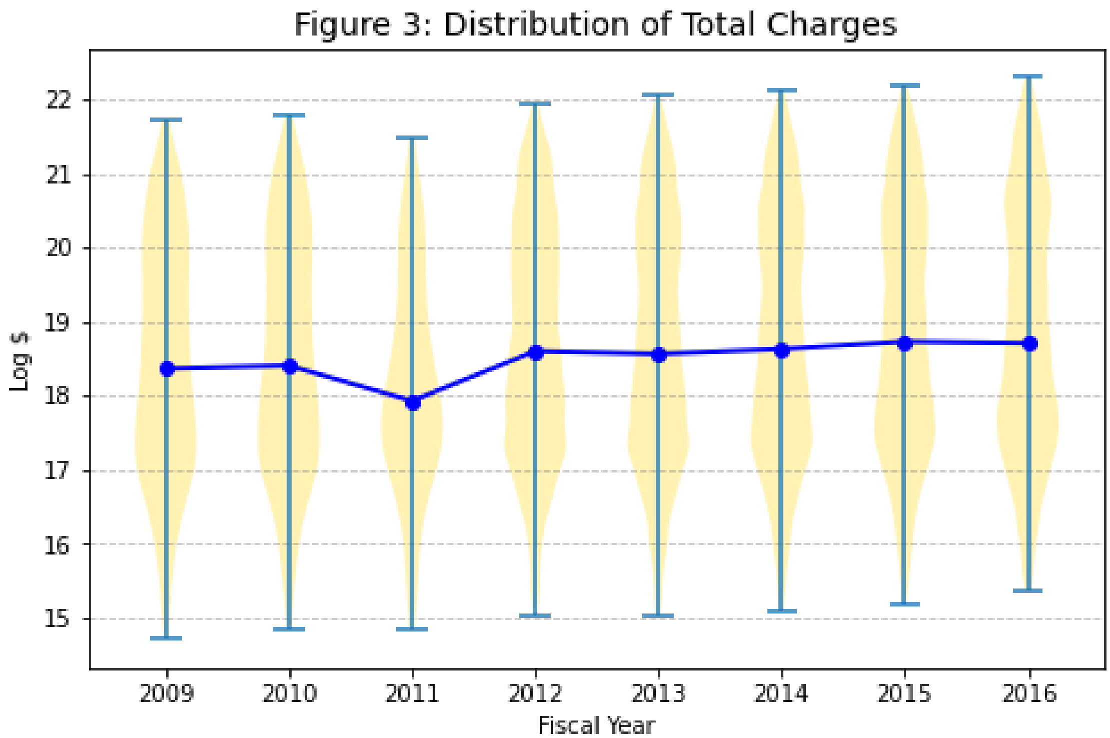
After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

	fy_year	hosp_count
0	2007	1116
1	2008	6116
2	2009	6041
3	2010	2656
4	2011	20459
5	2012	111296
6	2013	109552
7	2014	110966
8	2015	92751

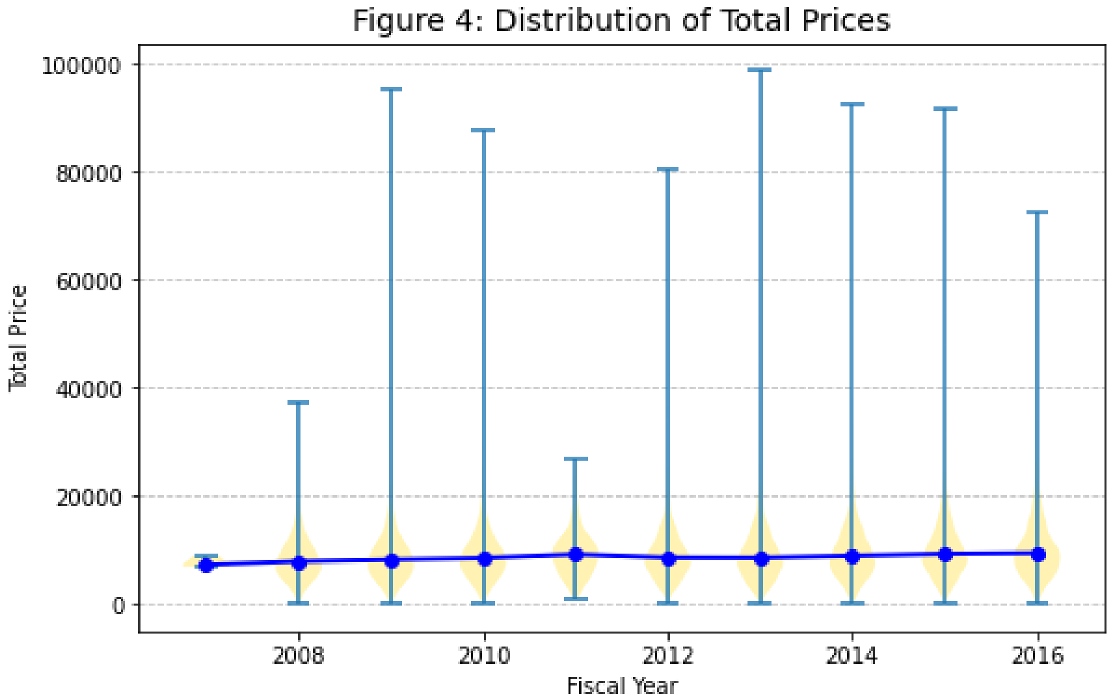
Figure 2: Unqiue Hospital IDs



Question 3:
What is the distribution of total charges (tot_charges in the data) in each year? Show your results with a "violin" plot, with charges on the y-axis and years on the x-axis.



Question 4:
What is the distribution of estimated prices in each year?



Question 5:
Calculate the average price among penalized versus non-penalized hospitals.

```
penalty
False    10200.321640
True      9748.015414
Name: price, dtype: float64
```

Question 6:
Split hospitals into quartiles based on bed size. Provide a table of the average price among treated/control groups for each quartile.

```
Table: Average Price by Treatment Status for Each Bed Size Quartile


             Control (No Penalty)  Treated (Penalty)
Bed Quartile
1                        15254.11           26868.59
2                         9263.94            9223.96
3                         9019.22            8423.95
4                        11414.75           10904.63
```

Question 7:
Use different estimators to calculate ATE.

Q7.A: Nearest Neighbor Matching (Inverse Variance)

```
Nearest Neighbor Matching ATE: 28791.02
```

Q7.B: Nearest Neighbor Matching using Mahalanobis Distance

```
Nearest Neighbor Matching (Mahalanobis) ATE: 24298.26
```

Q7.C: Propesntiy Score Matching and Weighting

```
Mean Price (Treated): 23033.46
Mean Price (Control): 9989.76
ATE (IPW): 13043.69
```

Q7.D: Simple Regression for ATE Estimation

```
provider_number                    int64
fy_start                  datetime64[ns]
fy_end                    datetime64[ns]
date_processed                    object
date_created                      object
beds                             float64
tot_charges                      float64
tot_discounts                    float64
tot_operating_exp                float64
ip_charges                       float64
icu_charges                      float64
ancillary_charges                float64
tot_discharges                   float64
mcare_discharges                 float64
mcaid_discharges                 float64
tot_mcare_payment                float64
secondary_mcare_payment          float64
street                            object
city                              object
state                             object
zip                               object
county                            object
hvbp_payment                     float64
hrrp_payment                     float64
year                               int64
source                            object
fy_year                            int64
discount_factor                  float64
price_num                        float64
price_denom                      float64
price                            float64
penalty                             bool
bed_quart                          int64
dtype: object
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
/Users/ilsenovis/Documents/GitHub/ECON470HW2/submission2/results/novis-i-hwk2-2
↪ipynb Cell 36 line <cell line: 54>()
    <a href='vscode-notebook-cell:/Users/ilsenovis/Documents/GitHub/ECON470HW2/
↪submission2/results/novis-i-hwk2-2.ipynb#X56sZmlsZQ%3D%3D?line=50'>51</a> X = 
↪sm.add_constant(X)
    <a href='vscode-notebook-cell:/Users/ilsenovis/Documents/GitHub/ECON470HW2/
↪submission2/results/novis-i-hwk2-2.ipynb#X56sZmlsZQ%3D%3D?line=52'>53</a> # 
↪Run regression
```

```
---> <a href='vscode-notebook-cell:/Users/ilsenovis/Documents/GitHub/ECON470HW2/
  ↪submission2/results/novis-i-hwk2-2.ipynb#X56sZmlsZQ%3D%3D?line=53'>54</a>␣
  ↪reg_model = sm.OLS(y, X).fit()
     <a href='vscode-notebook-cell:/Users/ilsenovis/Documents/GitHub/ECON470HW2/
  ↪submission2/results/novis-i-hwk2-2.ipynb#X56sZmlsZQ%3D%3D?line=55'>56</a> #␣
  ↪Display results
     <a href='vscode-notebook-cell:/Users/ilsenovis/Documents/GitHub/ECON470HW2/
  ↪submission2/results/novis-i-hwk2-2.ipynb#X56sZmlsZQ%3D%3D?line=56'>57</a>␣
  ↪print(reg_model.summary())

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/regression/
  ↪linear_model.py:890, in OLS.__init__(self, endog, exog, missing, hasconst,␣
  ↪**kwargs)
    887     msg = ("Weights are not supported in OLS and will be ignored"
    888            "An exception will be raised in the next version.")
    889     warnings.warn(msg, ValueWarning)
--> 890 super(OLS, self).__init__(endog, exog, missing=missing,
    891                           hasconst=hasconst, **kwargs)
    892 if "weights" in self._init_keys:
    893     self._init_keys.remove("weights")

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/regression/
  ↪linear_model.py:717, in WLS.__init__(self, endog, exog, weights, missing,␣
  ↪hasconst, **kwargs)
    715 else:
    716     weights = weights.squeeze()
--> 717 super(WLS, self).__init__(endog, exog, missing=missing,
    718                           weights=weights, hasconst=hasconst, **kwargs)
    719 nobs = self.exog.shape[0]
    720 weights = self.weights

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/regression/
  ↪linear_model.py:191, in RegressionModel.__init__(self, endog, exog, **kwargs)
    190 def __init__(self, endog, exog, **kwargs):
--> 191     super(RegressionModel, self).__init__(endog, exog, **kwargs)
    192     self._data_attr.extend(['pinv_wexog', 'wendog', 'wexog', 'weights'])

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/base/model.py:267,␣
  ↪in LikelihoodModel.__init__(self, endog, exog, **kwargs)
    266 def __init__(self, endog, exog=None, **kwargs):
--> 267     super().__init__(endog, exog, **kwargs)
    268     self.initialize()

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/base/model.py:92,␣
  ↪in Model.__init__(self, endog, exog, **kwargs)
     90 missing = kwargs.pop('missing', 'none')
     91 hasconst = kwargs.pop('hasconst', None)
---> 92 self.data = self._handle_data(endog, exog, missing, hasconst,
     93                               **kwargs)
```

```
     94 self.k_constant = self.data.k_constant
     95 self.exog = self.data.exog

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/base/model.py:132,␣
 ↪in Model._handle_data(self, endog, exog, missing, hasconst, **kwargs)
    131 def _handle_data(self, endog, exog, missing, hasconst, **kwargs):
--> 132     data = handle_data(endog, exog, missing, hasconst, **kwargs)
    133     # kwargs arrays could have changed, easier to just attach here
    134     for key in kwargs:

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/base/data.py:673,␣
 ↪in handle_data(endog, exog, missing, hasconst, **kwargs)
    670     exog = np.asarray(exog)
    672 klass = handle_data_class_factory(endog, exog)
--> 673 return klass(endog, exog=exog, missing=missing, hasconst=hasconst,
    674              **kwargs)

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/base/data.py:82, i␣
 ↪ModelData.__init__(self, endog, exog, missing, hasconst, **kwargs)
     80     self.orig_endog = endog
     81     self.orig_exog = exog
---> 82     self.endog, self.exog = self._convert_endog_exog(endog, exog)
     84 self.const_idx = None
     85 self.k_constant = 0

File ~/opt/anaconda3/lib/python3.9/site-packages/statsmodels/base/data.py:507,␣
 ↪in PandasData._convert_endog_exog(self, endog, exog)
    505 exog = exog if exog is None else np.asarray(exog)
    506 if endog.dtype == object or exog is not None and exog.dtype == object:
--> 507     raise ValueError("Pandas data cast to numpy dtype of object. "
    508                      "Check input data with np.asarray(data).")
    509 return super(PandasData, self)._convert_endog_exog(endog, exog)

ValueError: Pandas data cast to numpy dtype of object. Check input data with np
 ↪asarray(data).
```

Question 7: Final Summary Table

	Estimator	ATE Estimator
0	Nearest Neighbor Matching	28791.021695
1	Mahalanobis Distance Matching	24298.260886
2	Inverse Propensity Weighting	13043.694639

Question 8:
With these different treatment effect estimators, are the results similar, identical, very different?

The results from the different treatment effect estimators vary significantly. Nearest Neighbor Matching produced the highest estimate at ***28,791***, while Mahalanobis Distance Matching yielded a slightly lower estimate of ***24,298***. Inverse Propensity Weighting, however, resulted in a much lower estimate of ***13,043***.

These differences arise because each estimator makes different assumptions and applies different methodologies:

- ***Nearest Neighbor Matching***: pairs treated and control units based on ***similarity*** in key covariates, which can reduce bias but is sensitive to how matches are selected.
- ***Mahalanobis Distance Matching***: accounts for ***correlations*** between covariates, potentially making it more robust, which could explain why its estimate is lower than simple nearest neighbor matching.
- ***Inverse Propensity Weighting (IPW)***: adjusts for differences in the ***probability*** of treatment, effectively reweighting observations to approximate a randomized experiment. The lower estimate suggests that after adjusting for observed covariates, the estimated treatment effect is smaller.

The fact that IPW produced a much lower estimate indicates that selection into treatment might be influenced by observable factors, and adjusting for those reduces the estimated effect.

Question 9:
Do you think you've estimated a causal effect of the penalty? Why or why not? (just a couple of sentences)

While techniques like propensity score weighting and matching help reduce selection bias, they only account for observed confounders. If there are unobserved factors influencing both the penalty and hospital prices (e.g., hospital quality, patient mix), the estimates may still be biased. Without a truly randomized experiment or strong instrumental variable, we cannot definitively claim a causal relationship.

Question 10:
Briefly describe your experience working with these data (just a few sentences). Tell me one thing you learned and one thing that really aggravated or surprised you.

One thing that really aggravated me was that the datasets didn't download correctly so it took a while to actually clean/fix the data before I could merge it into the final dataset. One thing that suprised me was the large difference in the charges from the hospitals versus the actual prices