# AEDE 6330 HW2

Blain Morin

11/6/2020

In this document, I present the R code that does the equivalent STATA calculations. A log file with the STATA commands is attached in a separate file.

## Part 1: Getting to know your data

1. Download and load the data:

```
### Read in data
df <- read_excel("Dare_geocoded.xls")
```

2. (a) Open the dataset in the Stata Data Editor (that is, type edit in the Stata command window). What do the different colors for some variables mean?

The different colors in the STATA editor distinguish differnt variable types. For example, strings are color coded in red.

2. (b) Close the Data Editor and examine the data for missing values (hint, you may want to use the codebook command in the Stata command window). Are there any? How many observations do you have? How many cities do they come from? How many years does the dataset span?

There are 632 missing values in the "ownership" column. The data contain 2,987 observations. There are three cities in the data. The years range from 2006 to 2015.

3. Label the variables without labels in the Variables window.

Done in the STATA log file.

4. Generate a treatment variable (Nourish) to capture the impact of beach nourishment. Given that during the study period, beach nourishment took place only in the city of Nags Head, generate a treatment variable (Nourish) that is equal to 1 (that is, the observation is considered treated), if the observation is in Nags Head and 0 if located Kitty Hawk and Duck. How many treatment and control observations do you have in the data?

```
## [1] "Treatment observations = 1050"
```

```
## [1] "Control observations = 1937"
```

There are 1,050 observations in the treatment group and 1,937 observations in the control group.

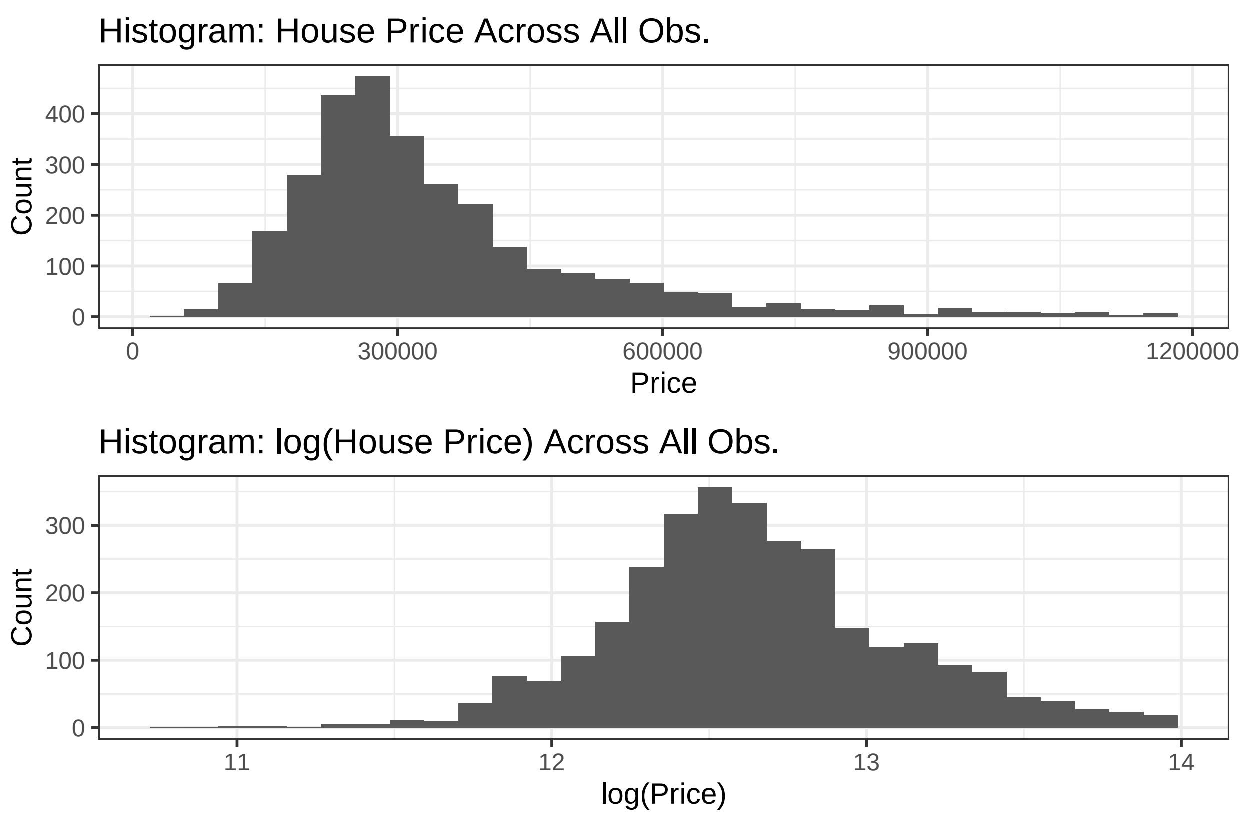
5. Generate summary statistics (means, standard deviations, min, max) for the structural house variables (that is, #bathrooms, bedrooms etc), the house price, and location variables (ocean front property, distance to beach and access point), by treatment group.

Table 1: Summary Stats, by Treatment Group

| Variable | Mean | sd | Min | Max |
|---|---|---|---|---|
| **Nourished Treatment Group** | | | | |
| Bathrooms | 3.19 | 1.53 | 1.00 | 10.00 |
| Bedrooms | 4.06 | 1.26 | 2.00 | 10.00 |
| Stories | 1.76 | 0.54 | 1.00 | 3.00 |
| Sqft | 20.33 | 8.31 | 6.08 | 53.68 |
| Price | 433415.92 | 208454.08 | 80000.00 | 1175000.00 |
| OceanFront = 1 | 0.09 | 0.29 | 0.00 | 1.00 |
| Shore Distance | 429.90 | 347.92 | 0.00 | 2728.72 |
| Access Distance | 334.66 | 194.62 | 23.29 | 1169.49 |
| **Not Nourished Control Group** | | | | |
| Bathrooms | 2.40 | 0.99 | 1.00 | 7.50 |
| Bedrooms | 3.32 | 0.84 | 1.00 | 8.00 |
| Stories | 1.34 | 0.51 | 1.00 | 3.00 |
| Sqft | 16.29 | 6.49 | 5.25 | 52.22 |
| Price | 296771.06 | 130174.80 | 50000.00 | 1150000.00 |
| OceanFront = 1 | 0.02 | 0.15 | 0.00 | 1.00 |
| Shore Distance | 896.22 | 739.83 | 25.40 | 4670.84 |
| Access Distance | 839.37 | 720.73 | 9.57 | 4606.86 |

6. Generate a histogram of the house values. Generate a ln(house price) variable and plot that. How does the histogram change due to the log transformation?

Figure 1: Compare log Transformation



We see that the log(Price) look more normally distributed than the nominal value.

7. Generate an indicator variable PostTreat that is equal to 1 if the house sale took place after 2010 and 0 otherwise. How many observations were sold after 2010? Before 2010?

```
### Generate PostTreat

df = df %>%
  mutate(PostTreat = ifelse(sale_year > 2010, 1, 0))

paste("Houses sold after 2010 =" , sum(df$PostTreat == 1))
```

```
## [1] "Houses sold after 2010 = 1472"
```

```
paste("Houses sold 2010 or prior =", sum(df$PostTreat == 0))
```

```
## [1] "Houses sold 2010 or prior = 1515"
```

There were 1,472 houses sold after 2010. There were 1,515 houses in 2010 or prior.

8. Generate a variable PostSandy that indicates if the sale took place before or after Hurricane Sandy in 2012. PostSandy=1 if the sale took place after 2012 and 0 otherwise. How many observations are these in the before and after groups?

```
## [1] "Houses sold after Sandy = 862"
```

```
## [1] "Houses sold before Sandy = 2125"
```

There are 862 sales after Sandy, and 2,125 before sandy.

9. How many observations are there in the following categories:

- Nourished and after the treatment took place

```
sum(df$Nourish == 1 & df$PostTreat == 1)
```

```
## [1] 539
```

- Nourished and before the treatment took place

```
sum(df$Nourish == 1 & df$PostTreat == 0)
```

```
## [1] 511
```

- Not nourished before the treatment took place

```
sum(df$Nourish == 0 & df$PostTreat == 0)
```

```
## [1] 1004
```

- Not nourished and after the treatment took place.

```
sum(df$Nourish == 0 & df$PostTreat == 1)
```

```
## [1] 933
```

## Part 2: Estimation

1. Running a basic difference-in-difference model (4 pts total). In this part, you will estimate an equation like this (Eq. 1 in Qiu & Gopalakrishnan 2018):

$$ln(Price_{ijt}) = \alpha_0 + \alpha_1 X_i + \beta_1 Nourish + \beta_2 Nourish * PostNourish + \eta_j + \zeta_t + \epsilon_{ijt}$$

a. Write out and execute the command in Stata. Be very careful the syntax for the fixed effects variables.

Here is the code that runs he regression (these match the STATA output in my log file):

```r
### Convert factors
df = df %>%
  mutate(Nourish = as.factor(Nourish)) %>%
  mutate(PostTreat = as.factor(PostTreat)) %>%
  mutate(blkgrp = as.factor(blkgrp)) %>%
  mutate(sale_year = as.factor(sale_year)) %>%
  mutate(OceanFront = as.factor(OceanFront)) %>%
  mutate(PostSandy = as.factor(PostSandy))

### Regression Specification, here block group and sale year are specified as fixed effects
model1 = lm(log(price) ~ bathroom + bedroom + stry +
              sqft + OceanFront +
              I(dist_Shoreline_m / 10) +
              I(dist_AccessPoint_m / 10) +
              Nourish*PostTreat + blkgrp + sale_year,
            data = df)
```

Table 2 displays the regression results.

Table 2: Diff-in-Diff Regression Results

| | *Dependent variable:* |
|---|---|
| | ln(Sale Price) |
| # of Bathrooms | 0.064*** (0.008) |
| # of Bedrooms | 0.032*** (0.008) |
| # of Stories | 0.137*** (0.011) |
| Living Area (100 Sqft) | 0.021*** (0.001) |
| Ocean Front = Yes | 0.219*** (0.024) |
| Distance to Shoreline (10m) | −0.001*** (0.0002) |
| Distance to Access Point (10m) | 0.001*** (0.0003) |
| Nourish = 1 | −0.103 (0.142) |
| Post Treat = 1 | −0.254*** (0.032) |
| Nourish*PostNourish | −0.008 (0.019) |
| Constant | 12.029*** (0.041) |
| Location (Block Group) Fixed Effects? | Yes |
| Year Fixed Effects? | Yes |
| Observations | 2,987 |
| $R^2$ | 0.709 |
| Adjusted $R^2$ | 0.707 |
| Residual Std. Error | 0.244 (df = 2958) |
| F Statistic | 257.716*** (df = 28; 2958) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

b. What is the interpretation of each of the coefficients in the Stata output? Which one is the impact of the treatment?

- **# of Bathrooms**: Every additional bathroom is associated with a 6% increase in the sale price of a house (on average, all else equal).
- **# of Bedrooms**: Each additional bedroom is associated with a 3% increase in the sale price of a house (on average, all else equal).
- **# of Stories**: Each additional floor of a house is associated with a 13.7% increase in the sale price of

a house (on average, all else equal).

- **Living Area**: Each additional 100 sqft is associated with a 2% increase in the sale price of a house (on average, all else equal).
- **OceanFront**: Being directly on the ocean is associated with a 21.9% increase in the sale price of a house (on average, all else equal).
- **Distance to Shoreline**: Every 10 meters further from the beach is associated with a .1% decrease in the sale price of a house (on average, all else equal).
- **Distance to Access Point**: Every 10 meters further from a beach access point is associated with a .1% increase in the sale price of a house (on average, all else equal).
- **Nourish = 1**: Sales in Nags Head before nourishment are associated 10% decrease in the sale price of a house (on average, all else equal), compared to the other towns before nourishment.
- **PostTreat = 1**: Sale prices after 2010 are associated with a 25% decrease in the sale price of a house (on average, all else equal), compared to prices before and including 2010.
- **Nourish x PostTreat (This is the impact of the treatment)**: This is the marginal effect of the nourishment on housing prices. In this case, having beach nourishment is associated with a .8% decrease in sale prices in the sale price of a house (on average, all else equal).
- **Fixed Effect Variables**: These are "nuisance variables" that absorb variation between years and block groups (they are the average differences between years and between locations).

c. Why is this a difference-in-difference model? What is the model testing (intuitively)?

Here we have two groups: a treatment and control. There is a difference in housing prices before the treatment time and after the treatment time for both groups. In short, we are looking at the difference in these differences. This attempts to isolate the difference caused by the treatment.

d. Is the impact of the treatment a MWTP?

Yes, assuming we have captured all relevant covariates and the parallel trend assumption holds, then we can say that this treatment effect is the marginal willingness to pay for beach nourishment.

2. Running a triple difference model (4 points total). In this part, you will estimate an equation like this (Eq. 3 in Qiu & Gopalakrishnan 2018):

$$ln(Price_{ijt}) = \alpha_0 + \alpha_1 X_i + \gamma_1 Nourish + \gamma_2 OF + \gamma_3 OF * Nourish +$$
$$\gamma_4 Nourish * PostSandy + \gamma 5 OF * PostSandy + \gamma 6 OF * Nourish * PostSandy + \eta_j + \zeta_t + \eta_{ijt}$$

a. Write out and execute the command in Stata. Be very careful the syntax for the fixed effects variables.

Here is the R code that specifies the model:

```
model2 = lm(log(price) ~
              bathroom +
              bedroom +
              stry +
              sqft +
              I(dist_Shoreline_m / 10) +
              I(dist_AccessPoint_m / 10) +
              OceanFront*Nourish*PostSandy +
              blkgrp +
              sale_year,
            data = df)
```

Table 3 displays the regression results.

b. What is the interpretation of each of the coefficients in the Stata output? Which one is the impact of the treatment?

The non-interaction terms have a similar interpretation to those in part 1. Here I will only interpret the interaction terms:

Table 3: Triple Diff Regression Results

| | *Dependent variable:* |
|---|---|
| | ln(Sale Price) |
| # of Bathrooms | $0.065^{***}$ (0.008) |
| # of Bedrooms | $0.032^{***}$ (0.008) |
| # of Stories | $0.137^{***}$ (0.011) |
| Living Area (100 Sqft) | $0.020^{***}$ (0.001) |
| Distance to Shoreline (10m) | $-0.001^{***}$ (0.0002) |
| Distance to Access Point (10m) | $0.001^{***}$ (0.0003) |
| Ocean Front = Yes | $0.127^{***}$ (0.047) |
| Nourish = 1 | $-0.106$ (0.142) |
| Post Sandy = 1 | $-0.255^{***}$ (0.032) |
| OceanFront * Nourish | $-0.005$ (0.059) |
| OceanFront * PostSandy | $0.247^{***}$ (0.079) |
| Nourish * PostSandy | $-0.028$ (0.021) |
| OF * Nourish * PostSandy | 0.022 (0.096) |
| Constant | $12.028^{***}$ (0.041) |
| Location (Block Group) Fixed Effects? | Yes |
| Year Fixed Effects? | Yes |
| Observations | 2,987 |
| $R^2$ | 0.713 |
| Adjusted $R^2$ | 0.710 |
| Residual Std. Error | 0.242 (df = 2955) |
| F Statistic | $236.360^{***}$ (df = 31; 2955) |
| *Note:* | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ |

- **OceanFront x Nourish**: Houses that were on the ocean and in the treatment group are associated with a marginal .5% decrease in sales prices (on average, all else equal).
- **OceanFront x PostSandy**: Houses that were on the ocean post Sandy are associated with a marginal 24.7% increase in sales prices (on average, all else equal).
- **Nourish x PostSandy**: Houses that were in the treatment group post Sandy are associated with a marginal 2% decrease in sales prices (on average, all else equal).
- **OceanFront x Nourish x PostSandy (This is the impact of the treatment on ocean front properties)**: Houses that were in the treatment group and on the ocean post Sandy are associated with a marginal increase in sales price of 2.2% (on average, all else equal).

c. Why is it a triple difference model? That is, what is the model testing? How does it compare to the simple difference-in-difference model above?

Similar to part 1, we are comparing the difference in trends between the treatment and control group. However, here we add another comparison, which is the marginal effect of the treatment for homes on the ocean. Essentially, we are estimating the nourishment effect on house sale prices specifically for the properties on the water.

d. Is the impact of the treatment a MWTP?

Yes, this again is a marginal effect. This can be interpreted as the marginal willingness to pay for nourishment for people whose homes are on the waterfront. Like in part 1, this assumes that we have captured all relevant covariates and that the parrelel trends assumption holds.