# Homework 3

Grant Jackson

2024-09-15

## Theoretical Questions

### Question 1

i. **In words, describe how to thin about p(2).**

Think of p(2) as the expected home value for houses with 2 rooms in Massachusetts.

ii. **Say you have a sample of parcels, how would you go about estimating p(n)? How would you estimate this in a regression?**

I would sort the data into groups by the number of rooms and then calculate each groups average home value. To make the estimate in a regression, I would create dummy variables for each group of houses by room number and then run the regression equation including the dummy variable.

iii. **How does the "fully-flexible" conditional expectation function differ from a linear regression model where Home Value_i = NumRooms_i x Beta + Ui**

The "fully-flexible" conditional expectation function differs because it allows each number of rooms to have its own effect on home value, instead of assuming a constant change in home value for each additional room like in the linear model.

iv. **Why might we not believe that p(3) - p(2) be the causal effect of increasing from 4 to 5 rooms on home price?**

We might not believe this because of possible omitted variable bias like larger homes being built in better neighborhoods, or more valuable homes might have more additions built.

### Question 2

i. **Say I include a set of indicator variables for gender and for whether or not the worker has a college degree in a regression. Describe a scenario where this regression model is not the conditional expectation function (hint: think about interactions).**

Say the wage bonus for having a college degree differs between the two genders, then you will need to include an interaction term like Gender x CollegeDegree. If there is an interaction effect between gender and having a college degree or not on wages, this model wouldn't capture it.

## Coding Exercise

### Question 1

Estimate the Conditional Expectation Function:

```
# Estimate the model
model <- feols(total_value ~ i(n_rooms, ref = 4), data = data)
```

```
## NOTE: 14,022 observations removed because of NA values (RHS: 14,022).
```

```r
# Get the predicted change from 4 to 5 rooms
change_4_to_5 <- coef(model)["n_rooms::5"]

print(change_4_to_5)
```

```
## n_rooms::5
##   45071.67
```

```r
summary(model)
```

```
## OLS estimation, Dep. Var.: total_value
## Observations: 85,978
## Standard-errors: IID
##                Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)   351989.72    3435.79 102.448089  < 2.2e-16 ***
## n_rooms::1    -53466.56   24671.49  -2.167140 3.0227e-02 *
## n_rooms::2    -24753.41   14939.84  -1.656873 9.7549e-02 .
## n_rooms::3    -55198.16    6706.71  -8.230287  < 2.2e-16 ***
## n_rooms::5     45071.67    4301.57  10.477951  < 2.2e-16 ***
## n_rooms::6     99622.86    3983.15  25.011076  < 2.2e-16 ***
## n_rooms::7    173207.34    4085.42  42.396439  < 2.2e-16 ***
## n_rooms::8    258853.99    4185.39  61.846985  < 2.2e-16 ***
## n_rooms::9    378657.24    4754.73  79.638058  < 2.2e-16 ***
## n_rooms::10   368059.48    5114.08  71.969844  < 2.2e-16 ***
## n_rooms::11   383620.74    6270.91  61.174614  < 2.2e-16 ***
## n_rooms::12   304003.44    6207.83  48.970941  < 2.2e-16 ***
## n_rooms::13   372100.63    8986.89  41.404807  < 2.2e-16 ***
## n_rooms::14   314905.43    8634.27  36.471581  < 2.2e-16 ***
## n_rooms::15   238402.46    7718.22  30.888276  < 2.2e-16 ***
## n_rooms::16   306505.12   15741.01  19.471753  < 2.2e-16 ***
## n_rooms::17   288107.29   15221.92  18.927130  < 2.2e-16 ***
## n_rooms::18   241097.41   11112.36  21.696327  < 2.2e-16 ***
## n_rooms::19   389867.11   41763.69   9.335074  < 2.2e-16 ***
## n_rooms::20   374892.18   33752.63  11.107051  < 2.2e-16 ***
## n_rooms::21   350245.33   34295.85  10.212470  < 2.2e-16 ***
## n_rooms::22   546676.94  108857.14   5.021967 5.1247e-07 ***
## n_rooms::23   664310.28  188483.48   3.524501 4.2450e-04 ***
## n_rooms::24   317510.28  100790.50   3.150201 1.6321e-03 **
## n_rooms::26   250543.61  153908.90   1.627870 1.0356e-01
## n_rooms::27   746970.28  266533.75   2.802535 5.0714e-03 **
## n_rooms::30    85210.28  266533.75   0.319698 7.4920e-01
## n_rooms::33  -214756.39  153908.90  -1.395347 1.6291e-01
## n_rooms::34  -250689.72  266533.75  -0.940555 3.4694e-01
## n_rooms::37   559010.28  266533.75   2.097334 3.5967e-02 *
## n_rooms::40    49435.28   94288.70   0.524297 6.0007e-01
## n_rooms::44  -141814.72  133300.09  -1.063876 2.8739e-01
## n_rooms::45   451010.28  266533.75   1.692132 9.0624e-02 .
## n_rooms::50    52747.78   47238.15   1.116635 2.6415e-01
## n_rooms::55  -100783.06   68898.72  -1.462771 1.4353e-01
## n_rooms::60    75564.76   30371.46   2.488019 1.2848e-02 *
## n_rooms::65   192810.28  266533.75   0.723399 4.6944e-01
## n_rooms::66     4121.39   88903.62   0.046358 9.6303e-01
## n_rooms::70   152170.74   28943.34   5.257540 1.4634e-07 ***
## n_rooms::77    20510.28  266533.75   0.076952 9.3866e-01
```

```
## n_rooms::78   123510.28  266533.75   0.463395 6.4308e-01
## n_rooms::80   275646.45   27702.46   9.950252  < 2.2e-16 ***
## n_rooms::90   467325.49   39444.91  11.847548  < 2.2e-16 ***
## n_rooms::99    50610.28  266533.75   0.189883 8.4940e-01
## n_rooms::100  410225.28   48779.29   8.409825  < 2.2e-16 ***
## n_rooms::110  483564.44   77011.95   6.279083 3.4220e-10 ***
## n_rooms::120  625285.28   77011.95   8.119328 4.7492e-16 ***
## n_rooms::130  400485.28  133300.09   3.004389 2.6619e-03 **
## n_rooms::140  142210.28  266533.75   0.533554 5.9365e-01
## n_rooms::150   34760.28  188483.48   0.184421 8.5368e-01
## n_rooms::170   23010.28  266533.75   0.086332 9.3120e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 266,432.6   Adj. R2: 0.185795
```

The predicted change from the conditional expectation function is $45,071.67.

### Question 2

Estimate a Linear Regression:

```
# Estimate linear model
linear_model <- feols(total_value ~ n_rooms, data = data)
```

```
## NOTE: 14,022 observations removed because of NA values (RHS: 14,022).
```

```
# Get coefficient for n_rooms
linear_change <- coef(linear_model)["n_rooms"]
print(paste("Predicted change per room:", linear_change))
```

```
## [1] "Predicted change per room: 9333.63122318361"
```

```
# Calculate change from 4 to 5 rooms
linear_change_4_to_5 <- linear_change
print(paste("Predicted change from 4 to 5 rooms:", linear_change_4_to_5))
```

```
## [1] "Predicted change from 4 to 5 rooms: 9333.63122318361"
```

```
print(paste("Difference from CEF estimate:", change_4_to_5 - linear_change_4_to_5))
```

```
## [1] "Difference from CEF estimate: 35738.0337837171"
```

```
summary(linear_model)
```

```
## OLS estimation, Dep. Var.: total_value
## Observations: 85,978
## Standard-errors: IID
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 459901.96   1659.815 277.0802 < 2.2e-16 ***
## n_rooms       9333.63    170.982  54.5885 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 290,367.5   Adj. R2: 0.033487
```

The predicted change from the linear regression model is $9,333.63. The linear model has a much smaller value in the predicted in the change from 4 rooms to 5 rooms, with a difference of $35,738.03.

### Question 3

Use Binscatter:

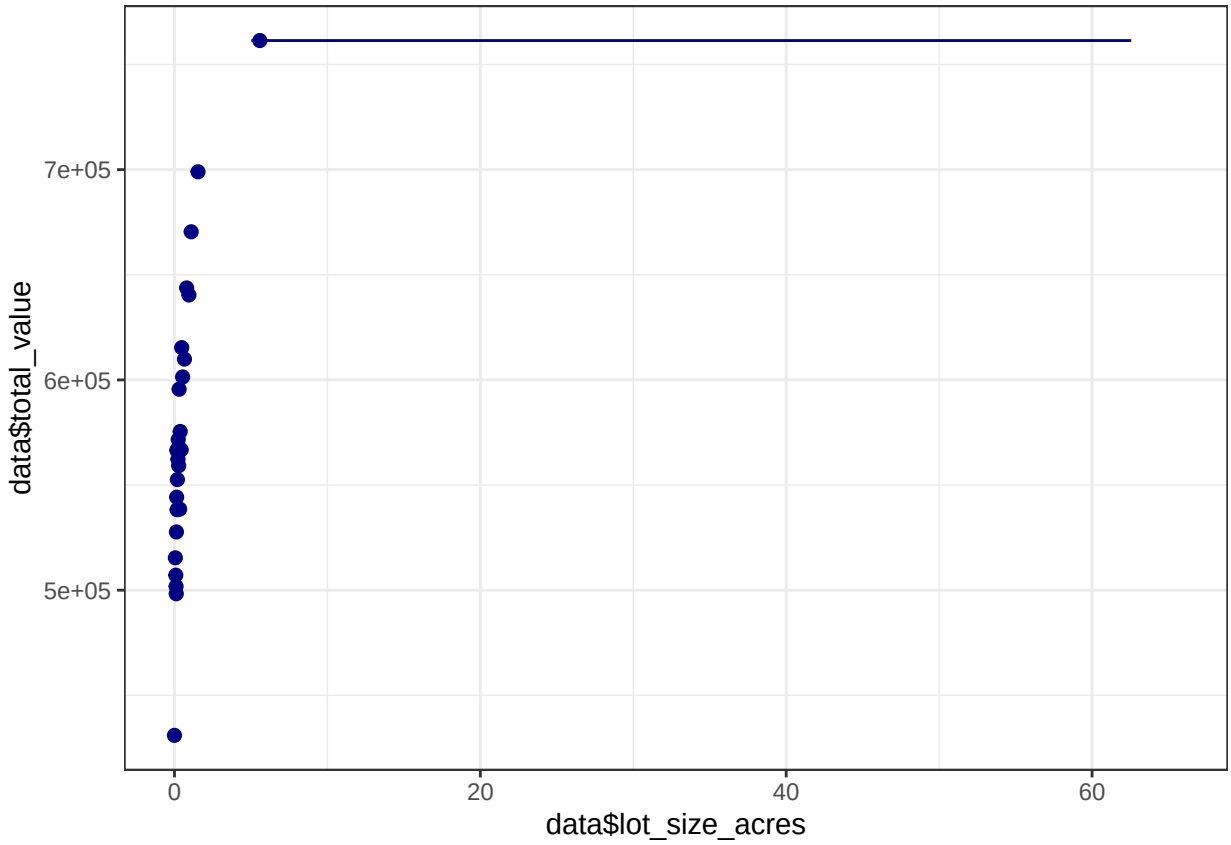
```
# Binscatter plots without covariate (simple scatter of total_value against lot_size_acres)
binsreg_plot1 <- binsreg(y = data$total_value, x = data$lot_size_acres, w = data$lot_size_acres,
                         line = c(0,0))
```

```
## Warning in binsreg(y = data$total_value, x = data$lot_size_acres, w =
## data$lot_size_acres, : To speed up computation, bin/degree selection uses a
## subsample of roughly max(5000, 0.01n) observations if the sample size n>5000.
## To use the full sample, set randcut=1.
```

```
## Warning in binsregselect(y, x, w, deriv = deriv, bins = dots, binspos =
## binspos, : Some bins have too few distinct values of x for DPI selection.
```

```
## Warning in binsreg(y = data$total_value, x = data$lot_size_acres, w =
## data$lot_size_acres, : DPI selection fails. ROT choice used.
```

```
## Warning in binsreg(y = data$total_value, x = data$lot_size_acres, w =
## data$lot_size_acres, : Repeated knots. Some bins dropped.
```




```
binsreg_plot2 <- binsreg(y = data$total_value, x = data$lot_size_acres, w = data$lot_size_acres,
                         line = c(2,2))
```

```
## Warning in binsreg(y = data$total_value, x = data$lot_size_acres, w =
## data$lot_size_acres, : To speed up computation, bin/degree selection uses a
## subsample of roughly max(5000, 0.01n) observations if the sample size n>5000.
## To use the full sample, set randcut=1.
```

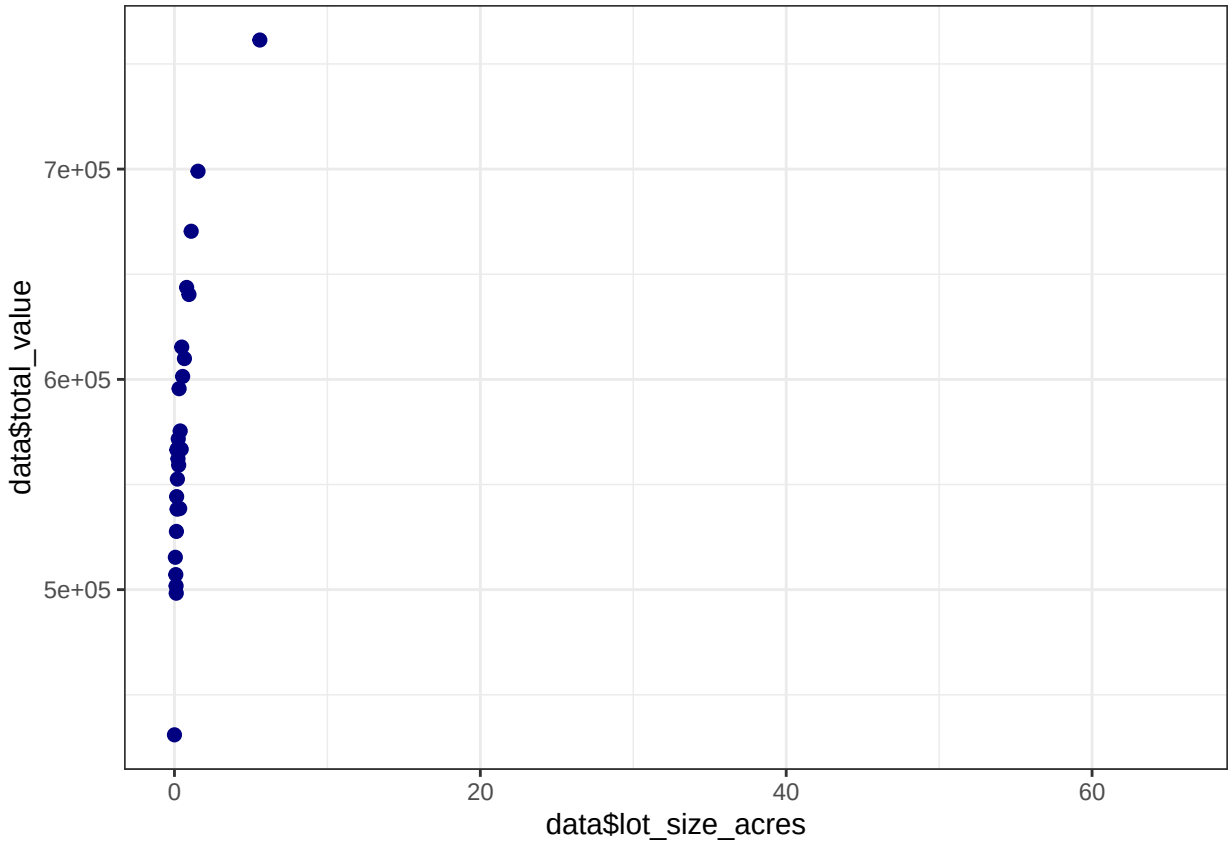
```
## Warning in binsregselect(y, x, w, deriv = deriv, bins = dots, binspos =
## binspos, : Some bins have too few distinct values of x for DPI selection.
```

```
## Warning in binsreg(y = data$total_value, x = data$lot_size_acres, w =
```

```
## data$lot_size_acres, : DPI selection fails. ROT choice used.
```

```
## Warning in binsreg(y = data$total_value, x = data$lot_size_acres, w =
## data$lot_size_acres, : Repeated knots. Some bins dropped.
```

```
## Warning in binsreg(y = data$total_value, x = data$lot_size_acres, w =
## data$lot_size_acres, : Some bins have too few distinct values of x for line.
```




```
binsreg_plot1
```

```
## Call: binsreg
##
## Binscatter Plot
## Bin/Degree selection method (binsmethod)  =  IMSE direct plug-in (select # of bins)
## Placement (binspos)                       =  Quantile-spaced
## Derivative (deriv)                        =  0
##
## Group (by)                          =  Full Sample
## Sample size (n)                     =  80839
## # of distinct values (Ndist)        =  13766
## # of clusters (Nclust)              =  NA
## dots, degree (p)                    =  0
## dots, smoothness (s)                =  0
## # of bins (nbins)                   =  26
```

```
binsreg_plot2
```

```
## Call: binsreg
##
## Binscatter Plot
```

```
## Bin/Degree selection method (binsmethod)  =  IMSE direct plug-in (select # of bins)
## Placement (binspos)                       =  Quantile-spaced
## Derivative (deriv)                        =  0
##
## Group (by)                          =  Full Sample
## Sample size (n)                     =  80839
## # of distinct values (Ndist)        =  13766
## # of clusters (Nclust)              =  NA
## dots, degree (p)                    =  0
## dots, smoothness (s)                =  0
## # of bins (nbins)                   =  26
```

## Question 4

This relationship is likely not causal due to several issues. It is likely the larger lots are in more desirable areas, more valuable properties might be more likely to have larger lots, and there could be selection bias in who is choosing to buy homes with different lot sizes.