# IMT573 Problem Set 6: Linear Regression

Your name:

Deadline: Sat, Nov 20th 8pm

## Instructions

This problem set revolves around linear regression, in particular interpretation of linear regression results. It contains two parts:

1. Estimate linear regression models yourself and discuss the results.
2. Interpret a model results from the literature. No coding needed here!

- Please write clearly! Answer each question in a way that if the code chunks are removed from your document, the result is still readable!

## 1 Housing Values in Boston

In this problem we will use the Boston housing dataset that is available on canvas. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. It is collected in 1970s, in the good ol' days when house prices were less than $50k.

The variables in the data are:

**crim** per capita crime rate by heighborhood.

**zn** proportion of residential land zoned for lots over 25,000 sq.ft.

**indus** proportion of non-retail business acres per neighborhood.

**chas** Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

**nox** nitrogen oxides concentration (parts per 10 million).

**rm** average number of rooms per dwelling.

**age** proportion of owner-occupied units built prior to 1940.

**dis** weighted mean of distances to five Boston employment centres.

**rad** index of accessibility to radial highways.

**tax** full-value property-tax rate per $10,000.

**ptratio** pupil-teacher ratio by neighborhood.

**black** $1000(B_k - 0.63)^2$ where $B_k$ is the proportion of blacks by neighborhood.

**lstat** lower status of the population (percent).

**medv** median value of owner-occupied homes in $1000s.

Your tasks are the following:

1. Describe the data and variables that are part of the Boston dataset. Are there any missings? Any unreasonable values? Clean data as necessary.

Next, we are estimating a series of simple regression models. We are modeling the neighborhood median house price *medv*.

2. Use the following predictors: *rm*, *lstat*, *indus*, and add two additional predictors of your choice. For each predictor do the following:

   (a) Make a scatterplot that displays how *medv* is related to that predictor and add regression line to that plot. Comment the result: do you see any relationship? Anything else interesting you see?

   Hint: you can add regression line on a plot with `geom_smooth(method="lm")` if using ggplot, or with `abline(m)` where `m` is your regression model.

   (b) Fit a simple linear regression model to predict the response. Interpret the slope (the effect of your explanatory variable). Is it statistically significant?

   (c) Explain why do you think you see (or don't see) the relationship on the figure/model. For instance, why do you see that neighborhoods with more lower status people have lower house prices.

3. Comment the results: are plots where you clearly can see a relationship related to models where the effect is statistically significant?

Enough of simple regression. Now let's move to the multiple regression.

4. Fit a multiple regression model to predict the response using all the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

   Hint: R formulas have a special way to tell "all predictors"

   Hint 2: consult OIS 8.4.2 "Understanding regression output from software" (p 332)

5. Interpret the results for *rm*, *lstat* and *indus*. Are the results statistically significant?

6. How do your results from 2 compare to your results from 4? Compare the results for those predictors you used for simple regression above. Explain why do the values differ. Do they still tell the same basic story?

## 2 Interpret Regression Results

van Holm and Monaghan (2020) analyze how evictions influence social capital across neighborhoods (paper available in canvas/readings).[1] They proxy social capital with number of 311 calls. These are little bit like 911 emergency calls, just for non-urgent purposes (such as carbage or potholes on street). They estimate the model in the form

$$311\text{calls}_i = \beta_0 + \beta_1 \cdot \text{evictions}_i + \boldsymbol{\beta}_2^\top \cdot \mathbf{demographics}_i + \\ + \boldsymbol{\beta}_3^\top \cdot \mathbf{urban\ character}_i + \epsilon_i.$$

Here *evictions* is the number of evictions in neighborhood $i$, *demographics* is a vector of neighborhood demographic characteristics and *urban character* is a vector of urban environment specific variables. $\beta_1$ is the variable of interest, the effect of evictions on social capital.[2] Their results are in Figure 1. Let us focus on model 3 (the column labeled as "(3)") and *ignore the other two models*. We stress here that *ignore the logs*, assume the variables are not logged!

Answer the following questions:

1. Do neighborhoods with more evictions see more or less 311 calls? By how much?

[1] The paper suffers from a number of issues. Do not take the results literally!

[2] They actually use logs of a number of variables. We ignore logarithms in this question.

2. Is the figure statistically significant (at 5% level)?

3. How is poverty rate associated with 311 calls? How much more (or less) calls there are in neighborhoods with 10 pct point more poverty?

4. What can you tell about association of race (*white*) and calls?

5. Is older median age associated with more or less 311 calls? At which level is this statistically significant?

6. The value for housing density is *-0.13*. What does thin number mean?

7. The omitted category for city is Austin, TX. Are there more or less calls in similar neighborhoods in Philadelphia, compared to Austin? By how much?

**Table 3.** Cross-sectional regressions.

| | (1) | (2) | (3) |
|---|---|---|---|
| | Dependent Variable: Calls to 311 in 2016 (logged) | | |
| No. of evictions (logged) | − 0.057*** | − 0.021** | 0.048*** |
| | (0.011) | (0.0095) | (0.0086) |
| % White | | − 0.091*** | − 0.038 |
| | | (0.032) | (0.026) |
| % college graduates | | 0.20*** | 0.049 |
| | | (0.047) | (0.040) |
| Median age | | 0.0067*** | 0.0067*** |
| | | (0.0013) | (0.0010) |
| % in poverty | | − 0.60*** | − 0.14** |
| | | (0.066) | (0.057) |
| % never married | | 0.42*** | 0.93*** |
| | | (0.076) | (0.063) |
| % female | | − 0.31** | − 0.36*** |
| | | (0.12) | (0.100) |
| % homes vacant | | 0.78*** | 0.23*** |
| | | (0.045) | (0.039) |
| Median year home built | | − 0.011*** | − 0.011*** |
| | | (0.00066) | (0.00055) |
| Total population (logged) | | 0.70*** | 0.61*** |
| | | (0.018) | (0.016) |
| Housing density (logged) | | − 0.14*** | − 0.13*** |
| | | (0.011) | (0.012) |
| % living in same house | | 0.77*** | 0.27*** |
| | | (0.10) | (0.085) |
| Boston | | | 0.18*** |
| | | | (0.039) |
| Denver | | | 0.23*** |
| | | | (0.034) |
| Los Angeles | | | 0.20*** |
| | | | (0.028) |
| New York City | | | 0.34*** |
| | | | (0.040) |
| Philadelphia | | | − 0.56*** |
| | | | (0.040) |
| San Francisco | | | 0.61*** |
| | | | (0.046) |
| Constant | 5.64*** | 22.9*** | 22.1*** |
| | (0.011) | (1.31) | (1.09) |
| No. of observations | 7,041 | 7,041 | 7,041 |

*Note.* Robust standard errors are given in parentheses.
*$p < .1$. **$p < .05$. ***$p < .01$.

Figure 1: van Holm and Monaghan (2020) Table 3.

## 3 Extra credit (10pt = 1EC pt)

Repeat the previous question, but now take into account the fact that some of the variables are logged. Respond the questions accordingly.

Hint: consult lecture notes https://otoomet.bitbucket.io/machineLearning.pdf/ section 4.1.6 Interaction and Feature Transformations.

**Finally** tell us how many hours did you spend on this PS.

## References

van Holm, E. J. and Monaghan, J. (2020) Eviction and the dissolution of neighborhoods, *Housing Policy Debate*, pp. 1–17.