CE C265 & PBHLTH 285

Spring 2025

Assignment 2: Estimating a bicycle exposure model (15 points)

Submitted by: Chun-hin (Anson) Sit, Jesus Hinojosa

Feb 2025

In this assignment, you will estimate a exposure model, selecting the appropriate explanatory variables (features) and the model that performs the best, i.e. predicts the counts most accurately. The dependent variable (target) in the data is Annual Pedestrian Traffic at intersections on California state highway system and the explanatory variables (features) consist of infrastructure, demographic, climate, network connectivity, transit, employment/land use and other characteristics. Some of these characteristics are available at different radii of catchment area (buffer) around the intersection, half-mile, quarter-mile, and one-tenth of a mile. This dataset might include variables that would not be useful in the modeling process and miss some that could potentially have been useful.

Please answer the following questions:

1. Based on the columns, please list some initial hypotheses on how each variable might affect pedestrian traffic, that is whether it would have an increasing or decreasing effect, or no influence for increases in that explanatory variable. Please do not revise this after estimating your model. It is always good to note what your initial hypotheses were and how your result might conflict with some of those, since your results are dependent on the variations captured in your data set ***(2)*****.**

Answer for 1:

For our hypothesis, we are focused on the variables Annual, Int4way, SchoolsH, Signal, AllTransitH, Jobs30H, and PopT. We believe that Int4way has high pedestrian traffic because it provides pedestrians with more crossing options. SchoolsH was selected because a higher number of schools means more students, which could increase pedestrian traffic. AllTransitH was chosen because greater transit availability may incentivize more people to use it, thereby increasing pedestrian traffic. Jobs30H was included because if jobs are within 30 minutes of transit, pedestrian traffic may rise. PopT was selected because a higher population near the buffer area could lead to increased pedestrian traffic. Overall, we believe these variables will positively influence pedestrian traffic.

2. Perform initial data descriptive analysis. Include mean, standard deviations for the dependent variable (target) and the explanatory variables (features). Display correlations among them and plot a few dependent variables with the explanatory variable to observe the functional form. How does your data descriptive analysis and your knowledge of pedestrian traffic influence the variables you select for your model ***(3)***?

Answer for 2:

For our analysis, we focused on the variables Annual, Int4way, SchoolsH, Signal, AllTransitH, Jobs30H, and PopT. The Descriptive Analysis and Correlation Heat map show that Jobs30H, AllTransitH, and SchoolsH have a strong or moderate correlation. Jobs30H and Popt being 0.63 and AllTransitH and PopT being 0.62. The moderate correlations are SchoolsH and PopT being 0.47 and Annnal and Jobs30H being 0.57. The weakest are Singal, Int4way, and Snow10in have a weak correlation with PopT. This influences the variable we would select by making sure we include Jobs30H, AllTransitH, SchoolsH and exclude Singal, Intway, and Snow10in.

	Annual	Int4way	Snow10in	SchoolsH	Signal	AllTransitH	Jobs30H	PopT
count	1.301000e+03	1301.000000	1301.000000	1301.000000	1301.000000	1301.000000	1.301000e+03	1301.000000
mean	7.328024e+05	0.694081	0.212452	1.882398	0.490392	5.669776	9.730359e+05	237.709650
std	2.029549e+06	0.460973	1.013796	1.874322	0.500100	3.105202	1.733345e+06	259.902302
min	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000
25%	3.550104e+04	0.000000	0.000000	0.000000	0.000000	3.099828	6.017384e+03	66.095605
50%	1.434238e+05	1.000000	0.000000	1.000000	0.000000	6.498840	1.591286e+05	167.359809
75%	5.436381e+05	1.000000	0.000000	3.000000	1.000000	8.408637	1.064683e+06	334.644143
max	2.579122e+07	1.000000	5.600000	9.000000	1.000000	9.986393	1.050488e+07	3799.904870

Figure 1: The Descriptive Analysis

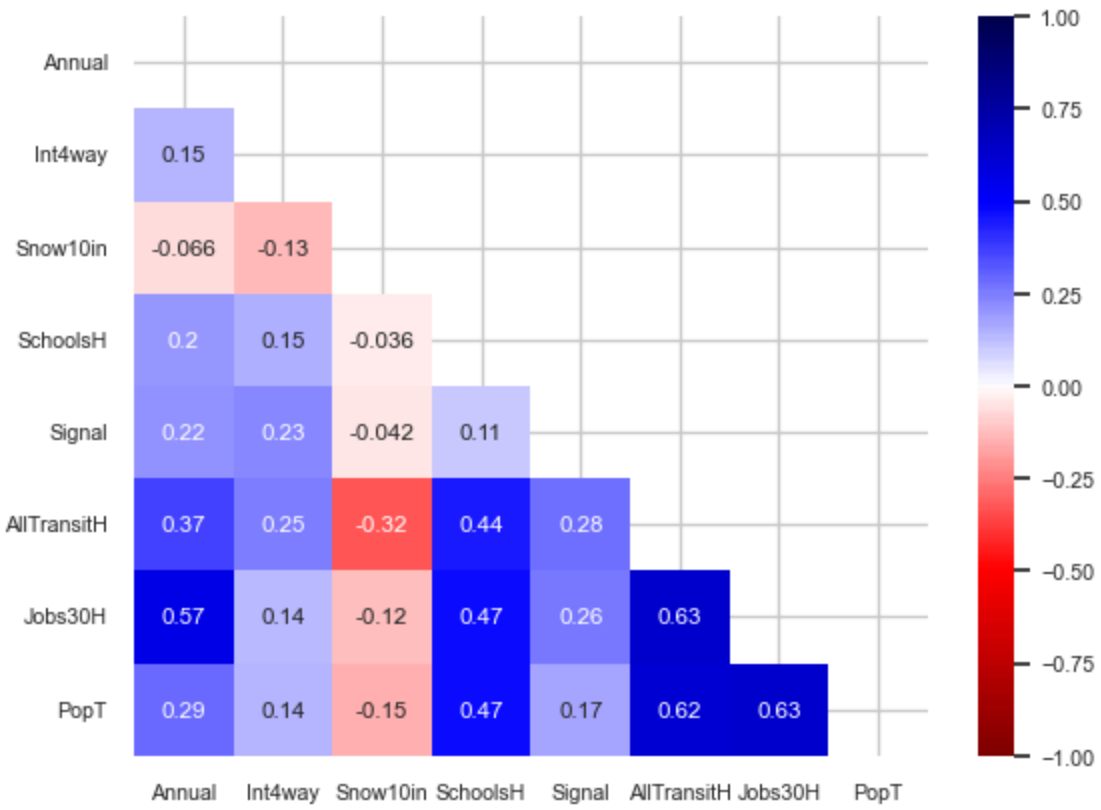


Figure 2: Correlation Heat Map

3. In the modeling stage, please begin with estimating a linear regression model, and then the count models, and any other model that you would like to estimate. For the initial model, you may want to include the explanatory variables one by one to observe the effect that each has on the significance of the other. Remember that the R-squared for LR models would mostly increase as you add more variables, whatever those may be ***(5)*.**

Answer for 3:

Before we start our analysis, we have divided the dataset into a training set and a test set. 20% of data is allocated for tests.

Linear regression model

First, we began with the simplest linear regression models. We have added the 7 selected explanatory variables one by one to observe the change of parameters.

Variables like Int4way, SchoolsH, Signal, Jobs30H, and PopT show significance, though the significance of Int4way decreases when more variables are added. In contrast, Snow10in and AllTransitH are often not significant or only borderline once additional predictors are included.

R^2 increases as more predictors are added—from 0.021 with only Int4way to 0.341 in the full model. This is expected as additional variables tend to explain more variation, though the contribution of each variable should be considered in the context of their statistical significance and potential multicollinearity.

The final model (with all 7 variables) shows a relatively high condition number (≈ 5.73e+06), which could suggest multicollinearity issues among the predictors. This is likely due to the high correlation between Jobs30H, AllTransitH and PopT.

To further interpret some of the variables:

Int4way (Four-way intersection): Positive and statistically significant. Locations with four-way intersections are associated with an increase in annual pedestrian counts by approximately 292,400, holding other variables constant.

SchoolsH (Number of schools): Negative and statistically significant. An increase in the number of school units is associated with a decrease in annual pedestrian counts by about 118,800 per school unit.

AllTransitH (All transit metric): Not statistically significant. Suggests that transit metrics may not strongly influence pedestrian counts in this model.

Poisson

Similar to the linear regression models, we have added the 7 selected explanatory variables one by one to observe the change of parameters. Every step reports z-statistics with p-values essentially 0 ($p < 0.001$), indicating that each predictor is statistically significant in its respective model. As new variables are added, some coefficients change in magnitude and even direction.

To further interpret some of the variables:

Int4way (Four-way intersection): Positive. For each additional four-way intersection, the expected annual pedestrian count increases by about 73%.

SchoolsH (Number of schools): Negative. Each additional school unitis associated with about a 13% decrease in annual pedestrian count.

AllTransitH (All transit metric): Positive. For each unit increase in the transit metric, the pedestrian count is expected to be about 40% higher.

As a special note, the pseudo R^2 in all models remains 1, which is weird and should be interpreted with caution.

Negative Binomial

Again, we have added the 7 selected explanatory variables one by one to observe the change of parameters. In every model almost all predictors have

z–statistics with $p < 0.001$. In the full model, however, "SchoolsH" ($p \approx 0.069$) is borderline and "PopT" is only marginally significant ($p \approx 0.045$), suggesting that when all predictors are included, some effects become less clear or may be partially explained by other variables (potential collinearity).

To further interpret some of the variables:

Int4way (Four-way intersection): Positive. A one-unit increase in Int4way is associated with a doubling of the expected pedestrian count (an increase of about 109%).

SchoolsH (Number of schools): Low positive. Each additional school unit is associated with about a 4% increase in the expected count; however, with a p-value of 0.069, this predictor is marginally significant. It might not be a robust predictor at a 5% significance level.

AllTransitH (All transit metric): Positive. A one-unit increase in the transit metric correlates with about a 23% increase in pedestrian count.

4. Please display some of the models you estimated including your final model. What were the reasons for selecting this model as your final model ***(2)?***

Answer for 4:

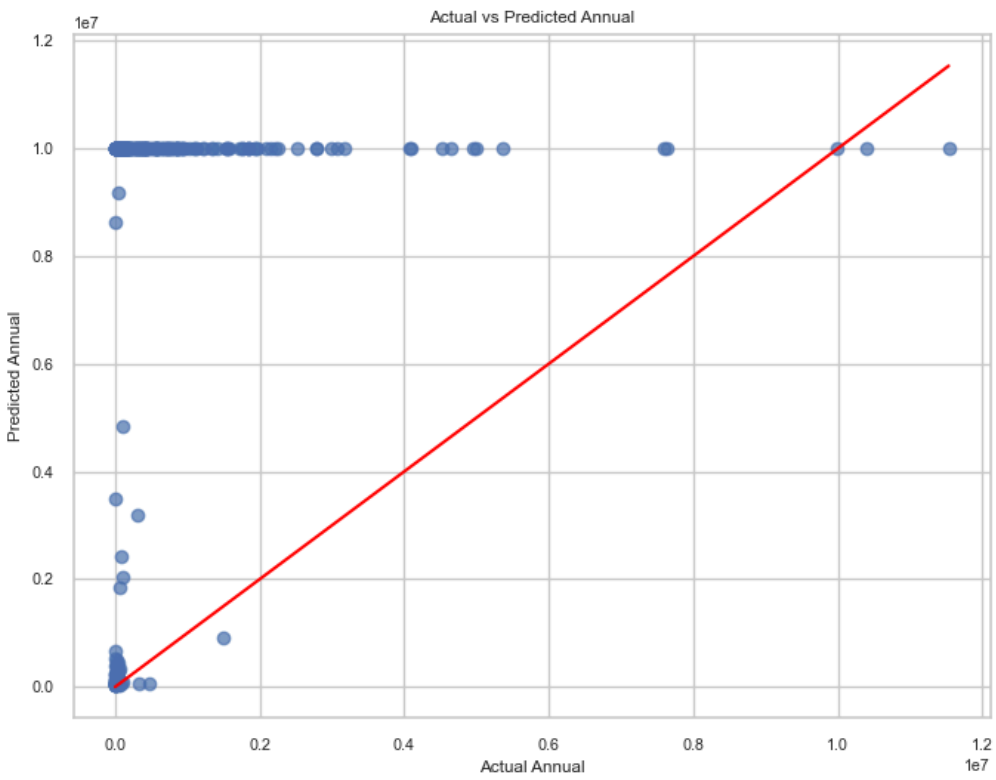
Negative Binomial would be the best model out of the 3 models tested. The annual pedestrian count is inherently a count variable, for which linear regression may perform poorly due to its assumptions of normally distributed errors and constant variance. Also, count data often exhibit variance greater than the mean. The Poisson model assumes the mean equals the variance and, as shown by its enormous deviance and Pearson chi^2 values, it fails to capture the observed variability in the data.

5. Please discuss any limitations in your model and how you could correct those. Discuss issues such as endogeneity, simultaneity etc. ***(3).***

Answer for 5:

To evaluate the performance of our Negative Binomial model using the selected predictors, we compared its estimated values with the actual test data. The results indicate that the model is performing very poorly. Key metrics are as follows:

MSE: 65909906125183.47
RMSE: 8118491.6163769895
MAE: 7104500.208496759
R^2: -26.5163803448109



Several potential issues could be contributing to this poor performance:

Endogeneity occurs when one or more predictors are correlated with the error term—often due to measurement error, omitted variables, or reverse causality. For instance, if the transit metric "AllTransitH" and pedestrian counts are determined simultaneously in certain urban settings, the model struggles to distinguish cause from effect.

Simultaneity arises when the dependent variable and one or more predictors are jointly determined. For example, areas with high pedestrian counts might attract better transit services, creating a two-way influence. This bidirectional relationship can lead to biased and inconsistent coefficient estimates.

Selecting highly correlated predictors may contribute to the model's poor performance. Overfitting or underfitting, along with improper scaling of variables, can also destabilize the model and lead to inaccurate predictions.

If given the opportunity to redo the model, it would be advisable to address these issues—by, for example, incorporating more uncorrelated variables, applying appropriate scaling or transformations, and using techniques to mitigate endogeneity (such as instrumental variables or structural modeling).

Please submit your responses in the form of a short report/write-up including your code. You could do this assignment on your own or team up with **one** other student.

Appendix:

1) Hypothesis:

Demographics:

- Pop: A higher population may lead to an increase in pedestrian traffic because more people are living near the buffer area.
- HseHold: An increase in households may increase pedestrian activity.
- White: This could have an impact due to socioeconomic factors at play.
- TransCom: More transit commuters could mean an increase in pedestrian traffic at transit hubs.
- WalkCom: The more people walking could increase pedestrian traffic.
- Degree: Higher education levels may correlate to the ability to afford a car increased cars on the road
- NoVech: Households with no car are more likely to use transit or walk, this could increase the pedestrian traffic at hubs.
- Pop25: Older population may walk less which could decrease pedestrian traffic
- WhitePct: This could impact pedestrian traffic because white people tend to be hit less when walking, possibly increasing the number of pedestrians if the white traffic is high.
- TransComPct: The higher the number of shared modes like buses or trains could increase pedestrian traffic.
- WalkComPct: The higher the walk mode share this would likely increase pedestrian traffic.
- DegreePct: High education levels could increase the car ridership lessing pedestrian traffic.

Infrastructure:

- PrincArt: The principal arteilas could have higher car traffic which would affect pedestrian traffic.
- MinAct: Minor arterial could have light traffic with minor effects on pedestrian traffic
- Collector: Collector streets could have fewer cars which would increase traffic for pedestrians.
- Int4way: This could have high pedestrian traffic because it gives pedestrians more options to cross.
- Singal: Intersections with signals could increase pedestrian traffic because pedestrians feel safer when using them.
- FC_min. FC_mac_, FC_sum: This could decrease pedestrian traffic because more cars are on the road.

Network Connectivity:

- StMetters: The increase in meters could increase traffic for pedestrians because they have more routes.
- StgSeg: More Street Segments could increase pedestrian traffic because the route is better connected

Transit (census Tract Level):

- AllTransit: With more transit, it could incentivize people to use it more increasing pedestrian traffic.
- Jobs30: If jobs are within 30 mins on transit there could be an increase in pedestrian traffic.
- Commuters: Having a higher number of transit commuters could increase pedestrian traffic at the transit hub.
- Tripswk: More transit trips per week could increase the pedestrian traffic at the transit hub.
- Routes: The more routes the transit has could increase pedestrian traffic.

Employment/Land Use:

- EmpSF: The more commercial spaces add new destinations for pedestrians increasing the traffic for them as well.
- Emp: The more employees in the area the higher chance of pedestrian traffic around the area.

Climate

- Precip05in: Maybe not enough rain to stop pedestrian traffic but enough to lessen it.
- Precip1in: Enough rain to decrease pedestrian traffic and maybe even stop it.
- Snow10in: Enough snow to decrease pedestrian traffic.
- Temp90: Enough to lower pedestrian traffic

Other:

- DistWater: Being close to a body of water could increase pedestrian traffic, think of beaches
- Schools: A higher number of schools means a higher number of students which could increase pedestrian traffic.
- MaxSlope: It could decrease traffic for pedestrians because it difficult for some

2) Code: