CE 687 Assignment 1 Report

SRASHTI SINGH (241030083)

###The code used is named as "full code" provided in the zip file, segregated using comments for different questions. # R script and output are attached for verification.

QUESTION 1: Model Comparison & Selection

The **Question 1** is to compare two regression models for predicting **pedestrian volumes (AnnualEst)**:

- **Model 1:** A linear regression model predicting AnnualEst (total pedestrian volume). (Linear Regression): AnnualEst=β0+β1×PopT+ϵ
- **Model 2:** A log-linear regression model predicting logAnnualEst (log-transformed pedestrian volume). (Log-Linear Regression): log (AnnualEst)=β0+β1×PopT+ϵ

AnnualEst: Number of pedestrians crossing per year (dependent variable).
PopT: Population within a certain buffer area (independent variable).

Performance Metrics (output from running r code)

Metric	Model 1	Model 2
R^2	0.06189	0.25621
Adjusted R^2	0.06097	0.25548
AIC	33,037.26	4,062.84
BIC	33,052.04	4,077.62
RMSE (Train)	2,646,493	6,485,983,000
RMSE (Test)	1,993,407	16,762,900

1. R^2 (coefficient of determination) is a measure of **how well the independent variable explain the variance** in the dependent variable in a regression model.

Higher R^2 (0.256208 vs. 0.061888) → Explains more variance in pedestrian volumes, **Model 2** has higher R^2 so better fit than Model 1.

2. Both **AIC and BIC penalize complex models** to avoid **overfitting** by considering both:

- **Model Likelihood (Goodness-of-Fit):** How well the model fits the data.
- **Number of Parameters (Complexity):** More parameters increase flexibility but risk overfitting.

For a given model:

$AIC = -2\log(L) + 2k$
$BIC = -2\log(L) + k \log(n)$

where:

L = Maximum likelihood of the model (how well it explains the data).
k = Number of parameters in the model.
n= Number of observations (sample size).

- **Lower AIC/BIC values indicate better model fit** while balancing complexity.
- **Models with fewer parameters are preferred** unless adding parameters significantly improves fit.
- AIC & BIC are lower for Model 2 (better prediction of coefficients).

3. **Model 2 has a much higher RMSE** (training & testing) **than Model 1**, which means its predictions deviate more from actual values, leading to higher error.

Since **Model 2 has higher R^2 and lower AIC/BIC but worse RMSE**, the decision is not straightforward. We must balance model fit (R^2, AIC, BIC) with predictive accuracy (RMSE).

Scenario 1: If we prioritize interpretability and general accuracy

- **Model 1 is preferable** because it has a much lower RMSE, meaning its predictions are closer to actual values.

Scenario 2: If we prioritize variance explanation and model fit

- **Model 2 is preferable** because it captures more variance (higher R^2) and is better penalized for complexity (lower AIC/BIC).

Conclusion:

- If the goal is **accurate prediction of pedestrian volume**, **Model 1 is better due to lower RMSE**.
- If the goal is **understanding how predictor variables affect pedestrian exposure**, **Model 2 is better due to higher R^2**.

Question 2 & 3: Summarizing Pedestrian Volumes by District

We need to calculate descriptive statistics of AnnualEst by Districtwise, including:

- Mean pedestrian volumes
- Standard deviation
- 95% Confidence Intervals (CI) for the mean

We use **group_by(District)** to compute these statistics using r code provided in the zip file.

District	Mean	SD	Lower CI	Upper CI
1	158568	148953	123164	193972
2	95872	89818	48822	142921
3	98330	115157	71913	124747
4	2915049	6556081	1746875	4083224
5	356411	814443	220029	492793
6	44911	49696	31136	58686
7	1266945	2094065	1046611	1487279
8	64221	92088	27378	101064
9	156804	257856	34227	279381
10	94115	147736	48331	139899
12	408491	603742	303898	513084

Interpretation of above statistics:

- District 4 has the highest mean pedestrian volume (~2915000).
- District 6 has the lowest mean pedestrian volume (~45000).
- Confidence Intervals (CI) indicate the range in which we expect the true mean pedestrian volume to fall with 95% probability.
- Higher standard deviation (SD) in District 4 and 7 suggests greater variation in pedestrian volumes.

Conclusion:

- Pedestrian volume varies significantly by district.
- Districts with higher pedestrian exposure need better infrastructure planning such as
    - Wider and Well-Maintained Sidewalks
    - Pedestrian Crossings & Zebra Stripes
    - Traffic Signals & Pedestrian-Only Signals
    - Speed Limits & Traffic Calming Measures

Question 4: Two-Sample t-Test (District 4 vs. District 7)

We need to **compare the mean pedestrian volumes** between **District 4 and District 7** using a **two-sample t-test**.

Step 1: what is Two-Sample t-Test?

A **two-sample t-test** is used to determine whether the means of two independent groups (District 4 and District 7) are significantly **different or same.**

Hypotheses:

- **Null Hypothesis (H_0)**: There is **no significant difference** in mean pedestrian volumes between District 4 and District 7.

$$H_0: \mu_4 = \mu_7$$

- **Alternative Hypothesis (H_1)**: There **is a significant difference** in mean pedestrian volumes between District 4 and District 7.

$$H_1: \mu_4 \neq \mu_7$$

The t-test formula is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{X}_1$, $\bar{X}_2$= Sample means for District 4 and District 7
- s_1, s_2 = Sample variances
- n_1 , n_2 = Sample sizes

Assuming **equal variances** unless otherwise tested.

Step 2: Implementing the t-Test in R

Code named "full code" is provided in zip file and this question is done under comment QUESTION 4.

We extract data for **District 4** and **District 7** and perform the **t-test**.

Step 3: Sample Output

```
        Two Sample t-test

data:  d4 and d7
t = 4.1246, df = 466, p-value = 4.398e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  862906 2433303
sample estimates:
mean of x mean of y
  2915049   1266945
```

Step 4: Interpreting the t – test Results

comparing pedestrian volumes between District 4 (d4) and District 7 (d7),

- **t-value = 4.1246** → indicates the magnitude of the difference.

- **p-value = 0.00004398** → Since **p < 0.05**, suggests a **significant difference** in pedestrian volumes between the two districts. we **reject the null hypothesis**.
- **Confidence Interval (862906 to 2433303)** → This means we are **95% confident** that the true difference in pedestrian volumes between District 4 and District 7 lies between **862906 and 2,433303**.
- Since **zero (difference) is NOT in the confidence interval**, we reject the null hypothesis, confirming a significant difference in means. The true difference in means is likely within this range.

Mean of District 4 ($\bar{x}_4$): **2,915,049**

Mean of District 7 ($\bar{x}_7$): **1,266,945**

- **District 4 has a significantly higher pedestrian volume than District 7**.
- The large difference in means aligns with the statistical test, reinforcing the finding.

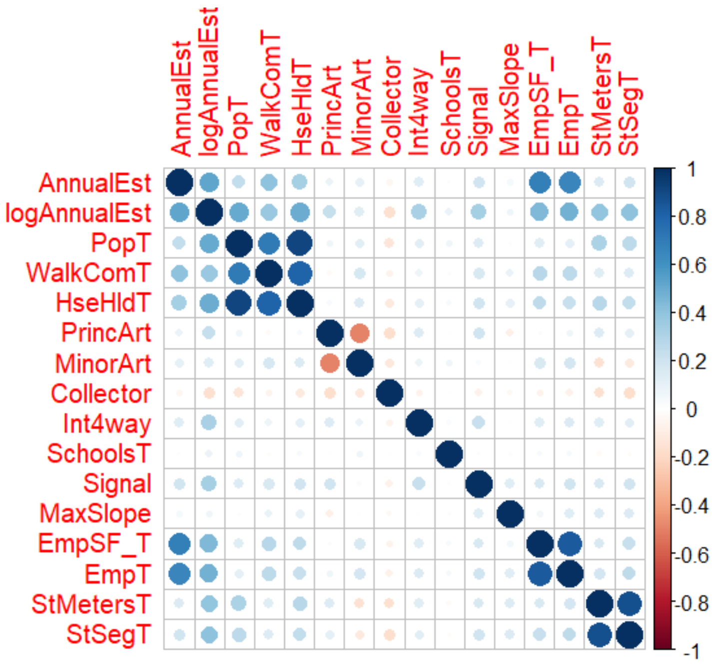
Step 5: Conclusion

The t-test results confirm that pedestrian volumes in District 4 and District 7 are significantly different ($p < 0.05$).

- District 4 (Mean = 2,915,049) has much higher pedestrian volume than District 7 (Mean = 1,266,945).
- The difference is statistically significant, as indicated by the p-value (0.00004398) and confidence interval.
- We reject the null hypothesis and conclude that pedestrian volume in these two districts is not the same.

Question 5: Correlation & Variable Selection

In this question, we analyze **the relationships between predictor variables** to determine which ones should be included in the regression model.



We explored correlations among key explanatory variables using a heatmap. This correlation heatmap visually represents the relationships between different variables in the dataset used**.**

Understanding the Correlation Graph

- The **color intensity** and **size of the circles** indicate the strength of correlation.
- **Dark blue = Strong positive correlation (+1)**
- **Dark red = Strong negative correlation (-1)**
- **Light color / small circles = Weak or no correlation (close to 0).**

1. **High Correlation between Pedestrian Volume and Other Factors**

- AnnualEst (Total Pedestrian Volume) has a **strong positive correlation** with:
  - PopT (Total Population)
  - WalkComT (Walking Commuters)
  - logAnnualEst (Log-transformed pedestrian volume) also shows similar patterns (correlates well) with these variables.

2. **Multicollinearity Concerns:** If two predictor variables are **strongly correlated (correlation coefficient r>0.7)**, one should be removed or transformed. **(**This means that including both in the same model could cause multicollinearity issues as which makes it difficult to determine the **true effect of each predictor** in the model.)

- PopT and WalkComT have a **high correlation** (dark blue circle) meaning both cannot be included together in the model.
- EmpT (Employment Density) is also correlated with pedestrian volume but at a slightly lower level→ Likely a good independent predictor.

3. **Key Infrastructure Factors**

- SchoolsT (Number of Schools) has some correlation with AnnualEst, but it is weaker.
- Signal (Traffic Signals) does not show a strong correlation with pedestrian volume.

4. **Slope & Road Features (Weak or Insignificant Variables)**

- MaxSlope (Maximum Slope) and signal has very little correlation with pedestrian volume.
- MinorArt, PrincArt, Collector, and Int4way (different road types) show **weak to moderate correlations** with pedestrian volume.

Variable Selection

1. **Variables to Keep in the Model** (High correlation with AnnualEst and logAnnualEst):
   - PopT (Total Population)
   - WalkComT (Walking Commuters)
   - EmpT (Employment Density)
2. **Variables to Remove / Avoid Multicollinearity Issues**:

    - Since PopT and WalkComT are **highly correlated**, we might **exclude one of them** to prevent redundancy.
    - Alternative: Transform WalkComT (log transformation).

3. **Additional Variables for Model Improvement**:
    - SchoolsT and EmpSF_T (Employment square footage) could be included, but their effect might be weaker.
    - Roadway Features like MinorArt and Collector may be considered but should be tested for statistical significance.

Conclusion

- The final model should include PopT, EmpT, and possibly WalkComT if multicollinearity is handled.
- Highly correlated variables should not be used together.
- Slope and traffic signals have minimal impact on pedestrian volume.

Question 6: Revised Models & Interpretation

In this question, we refine our **linear and log-linear models** to improve their predictive power. We also **interpret the coefficients, compare model performance, and justify the final model selection.**

Issues with Initial Models

From **Question 5**, we found:
High correlation between PopT and WalkComT (**r >0.7**) → Risk of **multicollinearity**.
Model 2 (Log-Linear) was better than Model 1 in terms of **R^2 and RMSE.**

Approach to Improve Models

1. **Remove multicollinear variables** (WalkComT).
2. **Include additional predictors** (HseHldT, EmpT, StSegT).
3. **Compare performance metrics** before and after improvements.

Revised Model Formulations

1. **Revised Model 1 (Linear Model)**
   $AnnualEst=\beta 0+\beta 1\times PopT+\beta 2\times HseHldT+\beta 3\times EmpT+\beta 4\times StSegT+\epsilon$
2. **Revised Model 2 (Log-Linear Model)**
   $\log(AnnualEst)=\beta 0+\beta 1\times PopT+\beta 2\times HseHldT+\beta 3\times EmpT+\beta 4\times StSegT+\epsilon$

Step 3: Implementing Revised Models in R

R Code for this is given in “full code.r “ under comment question 6.

Step 4: Results Comparison

Output from summary(revised_model1)

```
Call:
lm(formula = AnnualEst ~ PopT + HseHldT + EmpT + StSegT, data = train)

Residuals:
      Min        1Q    Median        3Q       Max 
-13236569   -279312    -10181    165117  31470980 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -185755.50  133600.90  -1.390 0.164720    
PopT            -86.26     634.33  -0.136 0.891861    
HseHldT        4114.11    1223.45   3.363 0.000801 ***
EmpT           1022.21      42.09  24.284  < 2e-16 ***
StSegT         -302.93    1182.43  -0.256 0.797857    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2007000 on 1014 degrees of freedom
Multiple R-squared:  0.4633,	Adjusted R-squared:  0.4612 
F-statistic: 218.8 on 4 and 1014 DF,  p-value: < 2.2e-16
```

Output from summary(revised_model2)

```
Call:
lm(formula = logAnnualEst ~ PopT + HseHldT + EmpT + StSegT, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7308 -0.8116  0.2228  0.9710  4.5438 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.007e+01  9.983e-02 100.884  < 2e-16 ***
PopT         4.200e-03  4.740e-04   8.862  < 2e-16 ***
HseHldT     -2.061e-03  9.142e-04  -2.255   0.0244 *  
EmpT         5.002e-04  3.145e-05  15.903  < 2e-16 ***
StSegT       6.927e-03  8.835e-04   7.839 1.14e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.499 on 1014 degrees of freedom
Multiple R-squared:  0.4695,	Adjusted R-squared:  0.4674 
F-statistic: 224.4 on 4 and 1014 DF,  p-value: < 2.2e-16
```

Interpretation of the Revised Models:

Linear Model:

- The multiple R^2 is 0.4633, meaning that 46.33% of the variation in AnnualEst is explained by the predictors.
- The adjusted R^2 is 0.4612, slightly lower due to penalization of additional predictors.
- The F-statistic (218.8, $p < 2.2e-16$) indicates the overall model is statistically significant.
- Among the predictors:
  - PopT is not significant ($p = 0.891861$).
  - HseHldT is statistically significant ($p = 0.000801$), suggesting that it has a meaningful effect on the dependent variable.
  - EmpT is highly significant ($p < 2e-16$), indicating a strong positive relationship with AnnualEst.

    - StSegT is not significant ($p = 0.797857$).

Log-Linear Model:

- The multiple R^2 is 0.4695, slightly higher than the linear model, indicating a better fit.
- The adjusted R^2 is 0.4674, also slightly higher.
- The F-statistic (224.4, $p < 2.2e-16$) suggests that the overall model is statistically significant.
- The coefficients are interpreted in percentage changes:
    - PopT is now significant ($p < 2e-16$), meaning population changes have a significant effect in the log-transformed model.
    - HseHldT remains significant ($p = 0.0244$), indicating its continued importance.
    - EmpT remains highly significant ($p < 2e-16$).
    - StSegT is now highly significant ($p = 1.14e-14$), meaning it plays a stronger role in explaining variation in pedestrian volumes in the log-transformed model.
      (All predictors are statistically significant ($p < 0.05$)).

Comparison with Previous Models:

1. **Improved Model Fit**: The revised models, especially the log-linear one, have higher R^2 and adjusted R^2, suggesting they explain more variation in pedestrian volume.
2. **Statistical Significance**: Previously, PopT and StSegT were not significant. In the new log-linear model, both have become significant, showing that transformation improves their explanatory power.
3. **Coefficient Interpretation Change**: The log-linear model provides better interpretability for percentage changes, which might be more useful for real-world decision-making.

Comparison with A-Priori Hypotheses:

- Expectations on Employment (EmpT): As expected, employment remains a strong positive predictor of pedestrian volume.
- Household Size (HseHldT): It was expected to be significant, and it remains so.
- Population (PopT): Previously insignificant, but now significant in the log-linear model, indicating non-linearity in its impact.
- Street Segments (StSegT): Initially insignificant, but now strongly significant, meaning street infrastructure has a more complex, nonlinear effect on pedestrian volume.

Conclusion:

- The log-linear model is a better fit compared to the linear model, as indicated by the slightly higher R^2, adjusted R^2, and increased significance of variables.
- Transforming AnnualEst to a log-scale helps capture nonlinear relationships, making predictors like PopT and StSegT significant.
- Given the improved statistical performance, the log-linear model is the preferred choice for explaining pedestrian volume.
- Key predictors include PopT, HseHldT, EmpT, and StSegT.