CE 687 Assignment 3

Due (March 25, 2025)

1. To log or not to log [20 points]

[Relevance:When working with linear regression, researchers often times seek to ensure that the predictions are not negative by transforming the dependent variable from $\mathbf{y}$ *to* $log(\mathbf{y})$*. Changing the scale of the dependent variable, also changes the relationship with explanatory variables and the impact of the model assumptions.]*

Collecting mode-specific traffic volumes at every transportation facility is infeasible (presently). Consequently, it is common practice to collect data of traffic counts at some representative sample locations and develop a regression model, which can be applied to other locations in the road network. A data set pertaining to pedestrian crossing volumes was curated for intersections in California, so as to be able to estimate a pedestrian exposure model that can be applied to intersections along the California highway system.

Table 1 provides a summary of the key variables used in the study. The variables pertain to demographics, built environment and network characteristics in the vicinity of the intersection. For variables that can be quantified at varying buffer areas around the intersection, such as population, street segments, number of schools, etc., different alternative were calculated at 3 different buffer distances–half-mile (H), quarter-mile(Q), and tenth-mile (T). The dependent variable is also available as *AnnualEst* or *logAnnualEst*, depending on the choice of model formulation. Supplementary files provided along with the assignment provide more details about the data sources of the explanatory variables.

Table 1: Summary statistics of some key variables

Variable	**Min**	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	**Max**
AnnualEst (*dependent*)	185.0	44439.2	156901.5	860765.9	583282.8	34787649.0
logAnnualEst (*dep, transformed*)	5.2	10.7	12.0	11.9	13.3	17.4
District	1.0	4.0	7.0	6.4	7.0	12.0
PopT	0.0	70.3	171.3	241.1	340.3	3799.9
WalkComT	0.0	0.2	1.8	10.4	7.4	653.9
HseHldT	0.0	24.0	64.3	101.6	128.7	2020.5
PrincArt	0.0	0.0	1.0	0.8	1.0	2.0
MinorArt	0.0	0.0	0.0	0.5	1.0	2.0
Collector	0.0	0.0	0.0	0.1	0.0	2.0
Int4way	0.0	0.0	1.0	0.7	1.0	1.0
SchoolsT	0.0	0.0	0.0	0.1	0.0	2.0
Signal	0.0	0.0	0.0	0.5	1.0	1.0
MaxSlope	0.0	0.0	0.0	0.0	0.0	0.2
EmpSF_T	0.0	18750.0	98750.0	363512.6	262500.0	12042500.0
EmpT	0.0	57.2	201.0	678.8	526.0	15270.0
StMetersT	1564.2	7064.4	9008.1	9133.9	11183.1	22508.7
StSegT	2.0	65.0	94.0	101.9	129.8	521.0

To also evaluate the performance of the predictions outside of the training data used to estimate the model, the total sample of 1297 observations was split into 80% training and 20% test data.

Table 2 shows the parameter estimates of a model estimated using *AnnualEst* as a dependent variable. Similarly, table 3 shows the parameter estimates of a model estimated using *logAnnualEst* as a dependent variable.

As discussed in the lectures, the goodness-of-fit of models can be evaluated using a variety of metrics: R^2, Adjusted R^2, Log-likelihood, AIC, BIC, root mean square error (RMSE), among others. However, note that

Table 2: Model 1 estimates (Dependent variable: *AnnualEst*)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	278190.36	112948.22	2.46	0.01
PopT	2593.62	316.64	8.19	0.00

Table 3: Model 2 estimates (Dependent variable: *logAnnualEst*)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.98	0.08	145.31	0.00
PopT	0.00	0.00	18.72	0.00

both models use different dependent variables, so their scales are not comparable. Under some instances, comparisons can be made by transforming the predictions to a common scale. For example, the root mean square error (RMSE) can be computed for model 2 by back-transforming the model's prediction using the exponential function ($exp(log\widehat{AnnualEst})$. Note that this prediction represents the median estimate of AnnualEst and not the mean, which also involves the estimation of $E[exp(\epsilon)|\mathbf{X}] = exp(\sigma^2/2)$, wherein $\sim N(0, \sigma^2)$ represents the error term. RMSE (at a common scale) can then be estimated for both training and test data.

Table 4 shows some of the commonly discussed model performance metrics. Please note that the training RMSE is worse than the test RMSE, which predicts that the preliminary models are currently underfitting the test data, and thus can be improved.

Table 4: Common model performance metrics

Metric	Model 1	Model 2
R^2	0.06	0.26
Adjusted R^2	0.06	0.26
Log-likelihood	-16515.63	-2028.42
AIC	33037.26	4062.84
BIC	33052.04	4077.62
RMSE(training, converted to scale of AnnualEst)	2646493	6485983220
RMSE(test, converted to scale of AnnualEst)	1993407	16762899

Given this information, and using the code provided with the assignment (*pedestrian_ exposure_ model.R*) as a starting point, please answer the following questions:

(i) Which of model 1 or 2 would you recommend for the application of estimating pedestrian volumes? Please discuss the metrics or visualizations that have been utilized by you in order to arrive at this conclusion. **[2 points]**

(ii) Summarize the pedestrian volumes by each district. Compute the sample means, standard deviations and confidence intervals of the population means. **[2 points]**

(iii) Summarize the pedestrian volumes by each district. Compute the sample means, standard deviations and confidence intervals of the population means. **[3 points]**

(iv) Conduct a two-sample test to compare the population means of pedestrian volumes in district 4 vs district 7. You may refer to the two-sample test discussion in reference by Washington et al., as well as the codes provided in `https://uc-r.github.io/t_test`. **[3 points]**.

(v) Please estimate and present correlations as part of the exploratory analysis that leads to the variable selection process for both the linear and the log-linear model. You are also encouraged to explore transformation of variables to their log or exponential counterparts, or converting them to indicator variables, as discussed in class. Also, note that several variables may also be correlated among themselves. You may consider the discussion in corrplot package as well `https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html`. **[5 points]**

(vi) Present revised models for both the linear and log-linear models and interpret the coefficients along with their statistical significance. Do they match your a-priori hypotheses? Utilize relevant model performance metrics and statistical tests (wherever applicable) to compare the updated models. Finally, identfy and explain the process undertaken to determine the final model across the linear and log-linear specification. **[5 points]**

Please present the discussions in a write-up, with model outputs in tables. The write-up needs to be self-explanatory and the relevant code needs to be shared as supplementary material for verification purposes.

Note: This dataset is being shared strictly for assignment purposes and students are striclty prohibited to share or upload the content in public domain.