TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
1–11

DOI: 10.1177/0361198119848701
journals.sagepub.com/home/trr
SAGE

Analysis and Application of Log-Linear and Quantile Regression Models to Predict Bus Dwell Times

Travis B. Glick[1] **and Miguel A. Figliozzi**[1]

Abstract

Understanding the key factors that contribute to transit travel times and travel-time variability is an essential part of transit planning and research. Delay that occurs when buses service bus stops, dwell time, is one of the main sources of travel-time variability and has therefore been the subject of ongoing research to identify and quantify its determinants. Previous research has focused on testing new variables using linear regressions that may be added to models to improve predictions. An important assumption of linear regression models used in past research efforts is homoscedasticity or the equal distribution of the residuals across all values of the predicted dwell times. The homoscedasticity assumption is usually violated in linear regression models of dwell time and this can lead to inconsistent and inefficient estimations of the independent variable coefficients. Log-linear models can sometimes correct for the lack of homoscedasticity, that is, for heteroscedasticity in the residual distribution. Quantile regressions, which predict the conditional quantiles, rather than the conditional mean, are non-parametric and therefore more robust estimators in the presence of heteroscedasticity. This research furthers the understanding of established dwell determinants using these novel approaches to estimate dwell and provides a relatively simple approach to improve existing models at bus stops with low average dwell times.

Understanding transit travel times and performance metrics leads to improved scheduling and on-time performance, which subsequently influences service attractiveness, cost, and efficiency (*1*). Dwell time is the amount of time a bus stops at a specific bus stop to allow passengers to board or alight and is one of the primary contributors to travel-time variability (*2*). Researchers and agencies have access to larger, more detailed, and new types of data than were unavailable in previous decades. These datasets have opened new avenues of research and allowed for improvements of past models by clarifying and reducing the assumptions of previous research, adding new variables (*3*), and improving visuals (*4*).

Until the introduction of higher resolution data, models to predict dwell times had been limited to stop level data (SLD), collected at bus stops and from onboard video data (*5*). In Portland, the Tri-County Metropolitan Transportation District of Oregon (TriMet) serves as the primary transit authority. TriMet has routinely collected automatic passenger counts (APC) and automatic vehicle location (AVL) data onboard all their buses since 1997. This collection of SLD is typical for most transit agencies. Using SLD data, previous research has confirmed, through multiple studies, that passenger movements, that is, the number of passengers entering (boarding) and exiting (alighting), are a primary contributor to dwell times (*6*, *7*). Early research excluded stop events with wheelchair activity, but newer studies have since confirmed their significance (*8*). Other research has found that door choice does not the change coefficients of passenger movements (*7*), and a non-linear relationship between dwell time and passenger movements exists (*9*).

Boardings, alightings, and lift activity are included in most models and serve as a baseline when testing new variables. Over the years, many additional variables have been confirmed as significant. For example, there are differences between the passenger boarding times of those who pay with cash versus credit cards (*10*). The time of day, onboard passenger loads, friction factors between boarding and alighting passengers (*8*), bus type (e.g., articulated buses, low-floor, or raised-floor buses) (*11*), and the number of standing passengers (*12*) have all be

[1]Transportation Technology and People (TTP) Lab, Department of Civil and Environmental Engineering, Portland State University, Portland, OR

Corresponding Author:
Address correspondence to Miguel A. Figliozzi: figliozzi@pdx.edu

shown to be significant. However, these additional variables have only marginally improved model explanatory power over the baseline models.

More recently, TriMet augmented their collection systems with high resolution data (HRD) based on timestamps and GPS locations of each of their buses in five-second intervals and made this information publically available (*13*). HRD allows for new variables to be added to models about traffic conditions and intersection behavior. Adding these variables to describe conditions about the stops themselves (e.g., proximity to a signalized intersection, existence of a designated bus bay or transit only lane, etc.) has allowed for the reduction of constraints and assumptions made in many previous studies (*3*). For example, stops near signalized intersections and transfer points were often excluded because of their tendency toward high dwells.

The linear regression models, reported in the literature, implicitly assume that regression models are a valid approach to dwell prediction; however, the distributions of dwell times are non-normal (*14*). As such, these models violate a key assumption of such tests: dependent observations should be normally distributed (*15*). While linear regression may not be ideal for dwell prediction, such models have been successfully employed within agencies to positive effect on timetable stability and on-time performance.

Alternative Modeling Methods

When the assumptions of linear regressions are violated, alternative methods may be employed that model: linear relationships between transformations of the dependent, independent, or both, variables; alternative metrics that are less reliant on previously violated assumptions; or a combination of both.

When working with non-normal data, one alternative is to transform the dependent variable into something with a normal distribution. For dwell, a Box-Cox analysis implies that the natural log transformation improves on the normality assumption of a linear regression. If none of the independent variables are transformed, a log-linear regression may be used. The coefficients of the model are related to percentage changes in dwell time rather than an additive value, as in linear regression.

Another approach to regression modeling is quantile regression. The typical regression method of least squares estimates the conditional mean of response (dependent) variable given the (independent) input variables. In contrast, quantile regression estimates the conditional quantile of the response variables based on the same inputs (*16*). Quantile regressions, specifically in their relation to dwell time, remain largely unstudied with similar research examining bus travel times rather than dwell (*17*). When coefficient values change between quantiles, it indicates that linear regression assumptions may be violated. This study will use a linear and log-linear quantile regression to examine how coefficients change and examine a practical application of the predicted coefficients.

Baseline Linear Model

This study compares results of log-linear and quantile regressions to a baseline model. The linear, log-linear, and quantile regressions were estimated using *R*, the statistical coding language. Quantile regressions were estimated using the 'quantreg' library. The contribution to the R-squared of each variable is calculated using the 'relaimpo' package.

Dependent and Independent Variables

The output (dependent) variable is Dwell for linear models. For log-linear models, *ln*(Dwell), the log transformation of Dwell, will be used. The input (independent) variables have been the subject of multiple past studies confirming their significance and will be the same for all models. Recent research used a combination of SLD and HRD variables to show that the following variables were significant and contributed to model explanatory power (*3*).

- Ons — Number of passengers boarding a bus at a specific stop (passengers board only from front door) [pax]
- Offs — Number of passengers alighting from a bus at a specific stop (passengers alight from both doors) [pax]
- Lift — Binary variable equal to 1 if the wheelchair lift is used, 0 otherwise
- Early_TP — Time ahead of schedule at a timepoint [min]
- Avg_Speed — Average speed in the segments immediately surrounding bus stop [mph]

In addition, the non-linearity of passenger movements will also be considered as Ons and Offs have been shown to add a reduced amount to dwell with each additional On or Off (*9*).

- Ons^2— The calculated square of number of boarding passengers (Ons)
- $Offs^2$— The calculated square of number of alighting passengers (Offs)

Adding to those seven variables, two additional binary variables will be used to indicate locations that have been shown to influence Dwell (*3*). The first variable is

Table 1. Independent Variable Statistics and Ranges

Continuous variables								
Variable	Mean	Min.	$q = 0.05$	$q = 0.25$	Median	$q = 0.75$	$q = 0.95$	Max.
Dwell	13.87	1	4	6	9	15	41	143
ln(Dwell)	2.31	0	1.38	1.80	2.19	2.71	3.72	4.96
Ons	1.04	0	0	0	1	1	4	9
Offs	1.08	0	0	0	1	2	4	9
Early_TP	0.06	0	0	0	0	0	0.23	5.28
Avg_Speed	15.23	0	7.32	12.41	15.46	18.29	22.22	59.01

Binary variables		
Variable	Observations	Percent of data
Lift	44,558	1.53%
Mall	159,000	5.46%
Traffic_Signal	1,545,686	53.12%
Total Observations		2,909,587

Note: Min. = minimum; Max. = maximum.

available through GTFS data that TriMet publically publishes and indicates a bus stop's proximity to a signalized intersection. Second, a variable that indicates if the bus stop is located on the downtown transit corridor (Mall) in Portland will also be included because of the overlapping of many routes and required driver behaviors. For example, drivers on the Mall are required to stop at every scheduled location regardless of passenger activity.

- Traffic_Signal — Binary variable equal to 1 if the stop is located near a signalized intersection, 0 otherwise
- Mall — Binary variable equal to 1 if the stop is located in the downtown transit corridor of Portland, 0 otherwise

Data Sources

This research includes 1,559 out of 5,818 bus stops in TriMet's service area. These stops were selected because they belong to 12 of the busiest routes in Portland. For each bus stop included in the dataset, all stop events from all possible routes servicing that stop are included. In total, 58 different routes operating for 56 weekdays between mid-September and late-November in 2017 are included. Buses stopped to serve passengers 46.8% of the time. For this study, all non-stopping buses were excluded. Each row of data is created through a combination of SLD and HRD and represents a single stop event.

Analysis of Variables

The data was cleaned for missing information (e.g., no data for GPS, or missing passenger movements), and outliers (defined as the largest 0.5% of Dwells, Ons, Offs, and Early_TP buses). The final dataset had 2,909,587 data points after removing 35,238 (1.20%) of service stops. Table 1 shows the statistics of the independent variables used in the analysis.

Baseline Model Coefficients

The results of the linear regression model to predict Dwell are shown in Table 2. These results are consistent with previous models and highly significant because of the number of data points used.

The economies of scale associated with passenger boardings and alightings can be seen through the negative values of Ons^2 and $Offs^2$. Each additional boarding or alighting passenger will contribute a smaller amount of time to dwell than the previous boarding or alighting. The total dwell time and dwell time savings for passenger movements are shown in Figure 1. According to this model, the 78th On and 51st Off would add no additional dwell time.

Alternative Model Results

Log-Linear Regressions

Using log-linear regression, the signs of the variables are the same as in the linear regression model but the adjusted R-squared increased (Table 3). Unlike linear regressions, the coefficients of the log-linear model are

Table 2. Linear Regression Results for Base Model Predicting Dwell

Variables	Coefficient	Std. error	*t*-value	Contribution to *R*-squared
Intercept	8.638***	0.0293	295	
Ons	3.338***	0.0112	298	0.0680
Offs	1.342***	0.0110	122	0.0158
Ons^2	–0.021***	0.0020	–11	0.0486
$Offs^2$	–0.013***	0.0019	–7	0.0135
Lift	37.393***	0.0576	649	0.0958
Mall	3.468***	0.0320	108	0.0064
Early_TP	17.308***	0.0222	781	0.1382
Avg_Speed	–0.126***	0.0016	–78	0.0048
Traffic_Signal	0.950***	0.0145	65	0.0069
Full Model Adjusted R-Squared				0.3982

Note: No. observations = 2,909,587.
***$p < 0.001$.

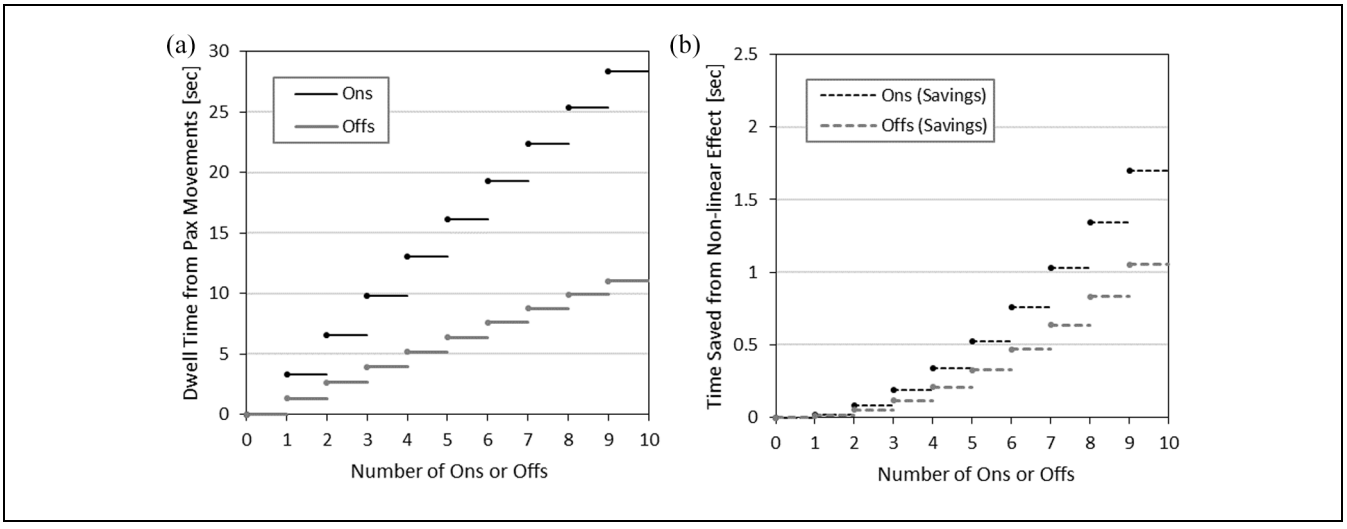

Figure 1. (*Left*) Non-linearity of Dwell; and (*right*) time savings for boarding and alighting passengers.

Table 3. Log-linear Regression Results Predicting *ln*(Dwell)

Variables	Coefficient	Std. error	*t*-value	Contribution to *R*-squared	Change to contribution
Intercept	1.9420***	0.0014	1406	na	na
Ons	0.3112***	0.0005	590	0.1346	+0.0666
Offs	0.1392***	0.0005	268	0.0351	+0.0193
Ons^2	–0.0194***	0.0001	–209	0.0714	+0.0228
$Offs^2$	–0.0066***	0.0001	–73	0.0246	+0.0111
Lift	1.2490***	0.0027	460	0.0525	–0.0433
Mall	0.1905***	0.0015	127	0.0095	+0.0031
Early_TP	0.5089***	0.0010	487	0.0590	–0.0792
Avg_Speed	–0.0087***	0.0001	–115	0.0078	+0.0031
Traffic_Signal	0.0850***	0.0007	124	0.0125	+0.0055
Full Model Adjusted R-squared				0.4071	

Note: No. Observations = 2,909,587; na = not applicable.
***$p < 0.001$.

Table 4. Linear Quantile Regression Results for Dwell

	Quantile								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Variables	Mean coefficient of significant samples ($p < 0.01$)								
Intercept	2.981	3.733	4.309	4.961	5.676	6.628	8.013	10.479	16.214
Ons	1.971	2.032	2.115	2.218	2.391	2.658	3.061	3.644	4.631
Offs	1.085	1.225	1.327	1.410	1.473	1.482	1.420	1.228	1.284
Ons^2	0.050	0.069	0.081	0.098	0.113	0.122	0.145	0.120	ã 0.077
$Offs^2$	ã 0.028	ã 0.036	ã 0.040	ã 0.045	ã 0.052	ã 0.062	ã 0.051	0.111	0.263
Lift	22.964	28.259	31.405	34.485	37.618	40.677	44.746	49.998	57.629
Mall	0.785	0.928	1.195	1.534	2.090	3.007	4.162	6.440	10.698
Early_TP	1.181	2.147	3.976	7.034	12.746	21.479	29.793	36.865	42.725
Avg_Speed	ã 0.026	ã 0.033	ã 0.036	ã 0.044	ã 0.052	ã 0.067	ã 0.089	ã 0.136	ã 0.247
Traffic_Signal	0.243	0.320	0.367	0.435	0.499	0.628	0.826	1.239	2.244
	Observed value of Dwell								
	4	6	7	8	9	11	13	17	27

not added directly to the dwell; instead, log-linear coefficients can be interpreted as percentage changes of the dwell time when the independent variable is increased by one unit. For example, the first On is likely to increase dwell by 31% based on this model, bus stops on the Mall are likely to experience 19% more dwell than stops elsewhere and stops near a Traffic_Signal will experience 8.5% longer Dwells.

One key difference between the linear and log-linear models is the contribution to the *R*-squared made by each variable. In this case, passenger movements increase by 0.1199 and now account for 65% of the model explanatory power. Based on contribution, there is a small benefit to using log-linear regressions over linear regression on these base 9 variables.

Linear Quantile Regressions

Quantile regressions do not predict the conditional mean and are used instead to estimate the conditional quantiles based on independent inputs. Quantile regression is non-parametric estimation method and not affected by the distribution of errors or the presence of outliers. Additionally, such models can provide insights into how the independent variables influence Dwells of a specific percentile, like the median (50th percentile) or the extreme lows or highs.

Quantile regressions are computationally intensive and a linear increase in data creates a non-linear increase in computational time. To overcome this limitation, bootstrapping is employed using 291 samples of 10,000 data points each. Each sample was tested using the same regression variables at each quantile in 0.1 increments. For each quantile, this method produces 291 models with coefficients and *p*-values for each independent variable (10 in total, including the intercept). For a single variable at a given quantile, the mean of all significant coefficients ($p < 0.01$) is reported as the coefficient for the quantile. The log-linear coefficients resulting from the quantile model are shown in Table 4.

The baseline linear regression (Table 2) predicted an Intercept of 8.638, but that value is more closely associated with the 70th percentile than the median value. While some variables, like Offs, remain relatively consistent, other variables like Lift, Mall, and Early_TP add substantially larger values at higher quantiles.

Another key feature of these results is the inconstant values for the non-linear passenger movement terms, Ons^2 and $Offs^2$. The reported coefficients are averages using only significant samples.

Table 5 shows the percentage of significant samples for each variable at each quantile. For Ons^2 and $Offs^2$, the majority of samples were not significant. For example, the 10th percentile reports a value of 0.050 for Ons^2, but this value comes from just 7% of the data points. The remaining 93% (272 of 10,000 samples) did not report significance.

Log-Linear Quantile Regressions

Continuing analysis for the log-linear quantile regression, the inconsistencies of the linear quantile model are mostly mitigated. Table 6 shows the coefficients for each quantile. The trend of coefficients to increase as percentile increases is still visible, but there is improved consistency of sign. Examining the coefficients for Ons and Ons^2, there are decreasing values as percentile increases. Ons is capturing some of the economies of scale discussed earlier. This is likely to be an effect of high Dwell stops tending to have more passenger movements and therefore a lower contribution from each passenger. The same trend is seen with Offs and $Offs^2$, with the exception of the 90th percentile, which experiences a sign flip.

Table 5. Significance of Bootstrapped Samples for Linear Quantile Regression

	Quantile								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Variables	Percent of significant samples ($p < 0.01$)								
Intercept	100%	100%	100%	100%	100%	100%	100%	100%	100%
Ons	100%	100%	100%	100%	100%	100%	100%	100%	100%
Offs	100%	100%	100%	100%	100%	100%	100%	99%	15%
Ons^2	7%	31%	64%	75%	76%	57%	26%	7%	5%
$Offs^2$	8%	21%	25%	24%	13%	8%	3%	4%	19%
Lift	61%	100%	100%	100%	100%	100%	100%	100%	100%
Mall	83%	94%	99%	100%	99%	100%	100%	99%	99%
Early_TP	90%	99%	100%	100%	100%	100%	100%	100%	100%
Avg_Speed	82%	96%	97%	98%	99%	99%	99%	98%	96%
Traffic_Signal	89%	99%	98%	100%	99%	99%	99%	99%	95%
	Observed value of Dwell								
	4	6	7	8	9	11	13	17	27

Table 6. Log-linear Quantile Regression Results for *ln*(Dwell)

	Quantile								
	10	20	30	40	50	60	70	80	90
Variables	Mean coefficient of significant samples ($p < 0.01$)								
Intercept	1.229	1.420	1.555	1.679	1.808	1.956	2.132	2.387	2.806
Ons	0.395	0.358	0.338	0.324	0.314	0.307	0.299	0.283	0.246
Offs	0.204	0.204	0.198	0.191	0.179	0.158	0.132	0.092	0.056
Ons^2	–0.028	–0.023	–0.021	–0.019	–0.018	–0.018	–0.018	–0.017	–0.015
$Offs^2$	–0.013	–0.013	–0.013	–0.012	–0.011	–0.009	–0.007	–0.006	0.008
Lift	0.931	1.149	1.269	1.335	1.385	1.418	1.415	1.391	1.298
Mall	0.098	0.100	0.114	0.131	0.162	0.190	0.231	0.287	0.362
Early_TP	0.148	0.210	0.288	0.406	0.533	0.675	0.805	0.903	0.945
Avg_Speed	–0.005	–0.006	–0.006	–0.006	–0.007	–0.008	–0.009	–0.011	–0.014
Traffic_Signal	0.051	0.054	0.058	0.061	0.068	0.074	0.088	0.106	0.131
	Observed value of *ln*(Dwell)								
	1.503	1.710	1.875	2.028	2.190	2.369	2.582	2.859	3.302

The number of significant samples (Table 7) is also greatly improved with the exception of the 80th and 90th percentiles for $Offs^2$, which may explain the change of sign. Yet, this overall consistency of sign and significance is evidence showing a benefit of using log-linear models over linear models.

Quantile regressions provide insights into the variability of coefficients and provide strong evidence that log-linear models are more robust than the traditionally used linear models. Additionally, they show how models based on the mean have estimated coefficients skewed toward higher percentiles.

Normality of Bootstrapped Estimates

Additional benefits of the log-linear model can be observed by comparing the distribution of the coefficient estimates from bootstrapping. The Shapiro-Wilk test tests the null hypotheses that the samples come from a normal distribution. Table 8 shows the number of variable coefficients (out of 10) that are likely to be non-normal for the linear and log-linear quantile regressions. The results indicate that the log-linear models are less likely to be non-normal.

Heteroscedasticity and Weighted Regressions

The variances of linear and log-linear models are not constant as confirmed using the Breusch-Pagan test, which tests the null hypothesis that the variances are homoscedastic (equal variances). For the models in this paper, the p-value of these tests was nearly 0, indicating heteroscedasticity, (i.e., unequal variances). Several attempts were made to correct for this. The first method added weights to each observation based on their Ons and Offs.

Table 7. Significance of Bootstrapped Samples for Log-Linear Quantile Regression

	Quantile								
	10	20	30	40	50	60	70	80	90
Variables	Percent of significant samples ($p < 0.01$)								
Intercept	100%	100%	100%	100%	100%	100%	100%	100%	100%
Ons	100%	100%	100%	100%	100%	100%	100%	100%	100%
Offs	100%	100%	100%	100%	100%	100%	100%	100%	25%
Ons^2	100%	100%	100%	100%	100%	100%	100%	100%	96%
$Offs^2$	100%	100%	100%	100%	100%	99%	74%	7%	20%
Lift	98%	100%	100%	100%	100%	100%	100%	100%	100%
Mall	55%	83%	93%	99%	100%	99%	100%	100%	100%
Early_TP	96%	100%	100%	100%	100%	100%	100%	100%	100%
Avg_Speed	82%	99%	99%	100%	100%	100%	100%	99%	99%
Traffic_Signal	89%	99%	100%	100%	100%	100%	100%	100%	99%
	Observed value of *ln*(Dwell)								
	1.503	1.710	1.875	2.028	2.190	2.369	2.582	2.859	3.302

Table 8. Number of Coefficients with Non-Normally Distributed Samples from Bootstrapping Based on *p*-value Cutoffs for Linear and Log-Linear Quantile Regression Models

		Quantile									All quantiles	
		10	20	30	40	50	60	70	80	90		
	p-value	Number of non-normally distributed sample coefficients										
Dwell	$p<0.10$	6	4	5	4	3	2	1	3	5	33	36.7%
	$p<0.05$	6	4	3	3	2	2	1	2	4	27	30.0%
	$p<0.01$	6	1	1	3	1	0	1	1	0	14	15.6%
	$p<0.001$	3	1	1	1	0	0	1	1	0	8	8.9%
ln(Dwell)	$p<0.10$	3	1	2	1	2	0	2	3	3	17	18.9%
	$p<0.05$	3	0	0	0	1	0	2	2	3	11	12.2%
	$p<0.01$	3	0	0	0	0	0	1	1	2	7	7.8%
	$p<0.001$	3	0	0	0	0	0	1	0	1	5	5.6%

$$\text{weight} = \frac{1}{(\text{Ons} + \text{Offs})^x}$$

where x could be defined as ⅓, ½, 1, 2, or 3. Approximately 15.36% of the datasets reported no passenger movements (i.e., Ons + Offs = 0) and these values were removed to provide a valid weight. By applying a weight to the regression models, some of the skew toward higher dwells could be mitigated. However, the treatment failed to correct for the heteroscedasticity. This result implies that there are factors influencing Dwell that are not included in models, which is further supported by the adjusted R-squared of each model, which suggests that only about 40% of the variance in the data is accounted for by current variables.

Applications of Quantile Coefficients

Given the skew toward higher percentiles when predicting the mean, one alternative approach is to apply a coefficient to predict the median rather than the mean Dwell or *ln*(Dwell) for each stop event. When using the median coefficients, the predicted dependent value would estimate the dwell time that half of buses are likely to exceed and half are likely to be less and this is a non-parametric estimation that does not require homoscedasticity assumptions.

In addition to applying the median, this research proposes an alternative whereby stops are assigned coefficients from the quantile regression based on the mean Dwell or *ln*(Dwell) of each given stop. The steps of this method are as follows:

1. Calculate quantile regression coefficients using all data.
2. Calculate the mean Dwell or *ln*(Dwell) for each individual stop.
3. Compare means from Step 2 to quantiles from Step 1 and, for each stop, assign coefficients from

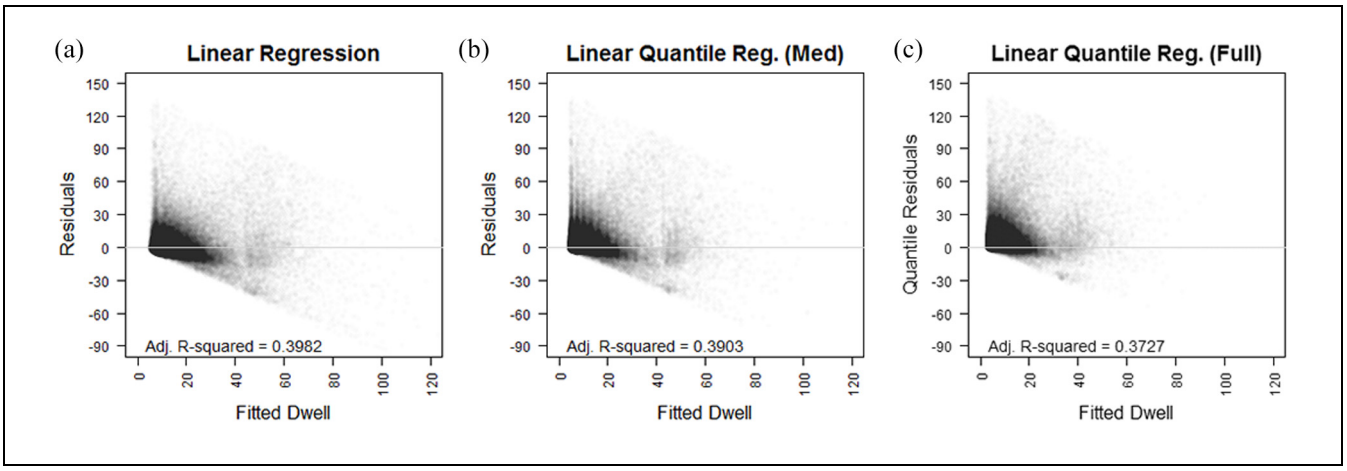

Figure 2. (*Left*) Residuals versus the fitted values of Dwell for linear regression; (*center*) median from linear quantile regression; and (*right*) full model from linear quantile regression.

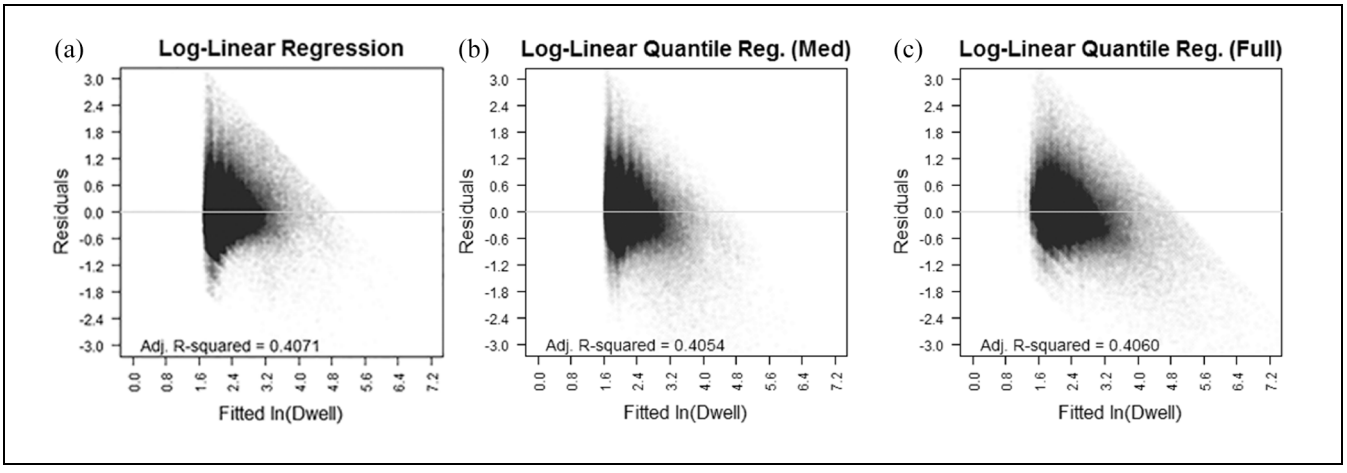

Figure 3. (*Left*) Residuals versus the fitted values of *ln*(Dwell) for log-linear regression; (*center*) median from log-linear quantile regression; and (*right*) and full model from log-linear quantile regression.

the quantile regression that most closely matches each bus stop's mean.

4. Calculate fitted values of Dwell or *ln*(Dwell) using each stop's assigned coefficients.
5. Estimate an adjusted R-squared for a quantile regression based on the fitted data, where *cor* is the correlation, n is the sample size, and p is the number of explanatory variables in the model.

$$R^2 = (cor(\text{fitted } \mathbf{Dwell},\ \text{true } \mathbf{Dwell}))^2$$

$$\text{Adj.}R^2 = 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-p-1)}$$

Figure 2 shows the residuals for predicting Dwell using linear regression, the median from linear quantile regression, and the new application using all quantiles. The first and most notable feature is the lack of symmetry around 0 in all plots, which would be expected for valid regression models. This implies that some of the assumptions of the linear regression models are not being satisfied. All three residual plots are similar; however, the full linear quantile regression shows a tendency to not over-predict values of Dwell, as evidenced by a decrease in negative residuals. Finally, the adjusted R-squared of the quantile regression is lower than the linear regression models for the median and full model.

The residuals using log-linear regression and log-linear quantile regression models (Figure 3) have a significantly improved symmetry around 0 based on visual inspection. Unfortunately, the average p-value of the Shapiro Normality test still shows heteroscedasticity problems for both the linear and log-linear models.

Table 9. Quantile Regression Model Fit for Stop Locations with Mean *ln*(Dwell) Nearest a Given Quantile

Quantile of mean *ln*(Dwell) of bus stop	Number of Obs.	Mean *ln*(Dwell)	Log-linear regression adj. R^2	Log-linear quantile regression (Median)		
				Adj. R^2	Improvement	
(0–0.15]	187	1.407	0.0889	0.1028	0.0139	15.6%
(0.15–0.25]	516	1.694	0.0425	0.0410	–0.0015	-3.6%
(0.25–0.35]	23,759	1.774	0.1168	0.1229	0.0061	5.2%
(0.35–0.45]	200,703	1.902	0.1596	0.1644	0.0048	3.0%
(0.45–0.55]	584,013	2.018	0.2116	0.2158	0.0042	2.0%
(0.55–0.65]	708,812	2.154	0.2689	0.2729	0.0040	1.5%
(0.65–0.75]	584,495	2.329	0.3174	0.3186	0.0011	0.4%
(0.75–0.85]	498,329	2.590	0.3388	0.3391	0.0002	0.1%
(0.85–1.0)	308,773	3.004	0.3437	0.3419	–0.0018	-0.5%
All Data	2,909,587	2.306	0.4071	0.4054	–0.0017	–0.4%

Note: Adj. = adjusted; obs. = observations.

Table 10. Quantile Regression Model Fit for Stop Locations with Mean *ln*(Dwell) Nearest a Given Quantile

Quantile of mean *ln*(Dwell) of bus stop	Number of obs.	Mean *ln*(Dwell)	Log-linear regression Adj. R^2	Log-linear quantile regression (Full model)		
				Adj. R^2	Improvement	
(0–0.15]	187	1.407	0.0889	0.1744	0.0855	96.1%
(0.15–0.25]	516	1.694	0.0425	0.0422	–0.0003	–0.8%
(0.25–0.35]	23,759	1.774	0.1168	0.1234	0.0067	5.7%
(0.35–0.45]	200,703	1.902	0.1596	0.1639	0.0043	2.7%
(0.45–0.55]	584,013	2.018	0.2116	0.2156	0.0040	1.9%
(0.55–0.65]	708,812	2.154	0.2689	0.2703	0.0014	0.5%
(0.65–0.75]	584,495	2.329	0.3174	0.3143	–0.0031	–1.0%
(0.75–0.85]	498,329	2.590	0.3388	0.3156	–0.0232	–6.9%
(0.85–1.0)	308,773	3.004	0.3437	0.3309	–0.0128	–3.7%
All Data	2,909,587	2.306	0.4071	0.4060	–0.0011	–0.3%

Note: Adj. = adjusted; obs. = observations.

However, the quantile regression does not require the homoscedasticity assumption. The log-linear median regression compresses the residual graphs over a slightly smaller range than the log-linear mean regression, while the using the full quantile widens the residuals.

While the log-linear model shows significant improvement over linear models, the application of the full log-linear quantile regression shows additional benefits. The first benefit is a reduction of extreme outliers at low values of fitted *ln*(Dwell). This can be confirmed by testing the correlation between the fitted values and observed values of *ln*(Dwell). Tables 9 and 10 divide the dataset into groups based on the mean value of *ln*(Dwell) of each bus stop. For each, all stop events associated with that stop are included.

Using the median or the full model resulted in improved estimates for stops with lower mean *ln*(Dwell). For stops with values around the 20th, 30th, and 40th quantiles, these improvements were 2–5% and affected 808,475 stop events (27.8% of the dataset). However, the predictions for stops with high mean *ln*(Dwell), using quantile regression reduced the model's predictive powers. However, the total change was less than a 0.5% reduction.

Conclusions

Previous research has primarily used linear regression models to test and validate potential new variables to estimate dwell times. However, an important assumption of linear regression models is homoscedasticity or the equal distribution of the residuals across all values of the predicted dwell times and this research shows that this assumption is violated in traditional Dwell linear regressions, that is, the residual distribution presents heteroscedasticity.

This research provides evidence that log-linear models (i.e., a linear regression model run on the log transformation of the dependent variables) has the potential to

improve on results obtained by traditional linear models. First, the residual plots of the log-linear models result in points distributed around 0, which is expected. While this is certainly an improvement over linear regression models, these residual plots do not have constant variances (homoscedasticity) and they are not normally distributed. But despite these issues, log-linear models proved much more robust against changes and smaller sample sizes when examined using quantile regressions.

Quantile regressions are used to model the conditional quantile of a dataset rather than the conditional mean. These coefficients can provide additional information about variable relationships that are not available with either linear or log-linear models. For example, the linear regression coefficients for Dwell tend to skew toward values near the 70th quantile. As such, the proposed approach that uses the quantile coefficients associated with the median can correct for the violation of the linear regression assumptions and therefore present alternative (less biased) coefficients. Therefore, this paper provides an alternative approach based on mean Dwell values of bus stops that reduces noise and improves correlations by 2–5% at lower quantiles. This method, which uses both log-linear and quantile methodology reduces bias for stops with a mean at lower percentiles of the dwell distribution. This alternative method is applicable at agencies with SLD collection systems, which includes most large and medium transit agencies in the US and abroad. These results require a change of analysis method, but do not require new data collection systems. As such, adoption can be inexpensive.

However, this method still suffers from the limitations of previous methods, including the inconsistent variances of the residuals, and should therefore not be viewed as a fix to Dwell modeling. Furthermore, the proposed application is just one possible way to use quantile regressions in industry and further research is required to measure the benefits of alternatives.

The results of this research have implications for past and future linear regression analysis of dwell times. The results of past dwell time studies that did not correct for heteroscedasticity are not invalid though there may be some bias in the estimation of regression coefficients and/or they may have been less efficient in the estimation of confidence levels. Regarding future studies, researchers should check for violations of the homoscedasticity assumption and make the necessary estimation adjustments, for example as proposed in this paper, to deal with violations of linear regression assumptions when dwell time is the dependent variable. Improved models for predicting Dwell can influence scheduling and transit planning and can be used to test new variables that may only apply in specific situations, like the peak-hour. Quantile regressions and log-linear modeling can be potentially useful to estimate models for network delay or bus interactions/bunching as illustrated in a recent study (*18*).

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: TBG, MAF; data collection: TBG; analysis and interpretation of results: TBG, MAF; draft manuscript preparation: TBG. Both authors reviewed the results and approved the final version of the manuscript.

References

1. Levinson, H. S. Analyzing Transit Travel Time Performance. *Transportation Research Record: Journal of the Transportation Research Board*, 1983. 915: 1–6.
2. Transportation Research Board. *TCRP Report 165: Transit Capacity and Quality of Service Manual, Third Edition*. Washington, D.C., 2013.
3. Glick, T. B., and M. A. Figliozzi. Measuring the Determinants of Bus Dwell Time. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2647: 109–117.
4. Stoll, N. B., T. B. Glick, and M. A. Figliozzi. Using High-Resolution Bus GPS Data to Visualize and Identify Congestion Hot Spots in Urban Arterials. *Transportation Research Record: Journal of the Transportation Research Board*, 2016. 2539: 20–29.
5. Fricker, J. D. Bus Dwell Time Analysis Using Onboard Video. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
6. Bertini, R. L., and A. M. El-Geneidy. Modeling Transit Trip Time Using Archived Bus Dispatch System Data. *Journal of Transportation Engineering*, Vol. 103, No. 1, 2004, pp. 56–67.
7. González, E. M., M. G. Romana, and O. Álvaro. Bus Dwell-Time Model of Main Urban Route Stops: Case Study in Madrid, Spain. *Transportation Research Record: Journal of the Transportation Research Board*, 2012. 2274: 126–134.
8. Dueker, K. J., T. J. Kimpel, J. G. Strathman, and S. Callas. Determinants of Bus Dwell Time. *Journal of Public Transportation*, Vol. 7, No. 1, 2004, pp. 21–40.
9. El-Geneidy, A. M., and N. Vijayakumar. The Effects of Articulated Buses on Dwell and Running Times. *Journal of Public Transportation*, Vol. 14, No. 3, 2011, pp. 63–86.
10. Tirachini, A. Estimation of Travel Time and the Benefits of Upgrading the Fare Payment Technology in Urban Bus Services. *Transportation Research Part C: Emerging Technologies*, Vol. 30, 2013, pp. 239–256.
11. Sun, L., A. Tirachini, K. W. Axhausen, A. Erath, and D. H. Lee. Models of Bus Boarding and Alighting Dynamics. *Transportation Research Part A: Policy and Practice*, Vol. 69, 2014, pp. 447–460.
12. Li, F., Z. Duan, and D. Yang. Dwell Time Estimation Models for Bus Rapid Transit Stations. *Journal of Modern Transportation*, Vol. 20, No. 3, 2012, pp. 168–177.

13. TriMet. *TriMet General Transit Feed Specification*. TriMet, Portland, Ore., 2016.
14. Rashidi, S., and P. Ranjitkar. Approximation and Short-Term Prediction of Bus Dwell Time using AVL Data. *Journal of the Eastern Asia Society for Transportation Studies*, Vol. 10, 2013, pp. 1281–1291.
15. Montgomery, D. *Design and Analysis of Experiments*, 6th ed. Jon Wiley & Sons Inc., New York, 2005.
16. Koenker, R. *Quantile Regression*. Cambridge University Press, Cambridge, 2009.
17. Ma, Z., S. Zhu, H. N. Koutsopoulos, and L. Ferreira. Quantile Regression Analysis of Transit Travel Time Reliability with Automatic Vehicle Location and Farecard Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2652: 19–29.
18. Glick, T., and M. Figliozzi. Analyzing the Impact of Bus Stop Queuing and Bus Interactions on Bus Dwell Times. Presented at 98th Annual Meeting of the Transportation Research Board, Washington, D.C., 2019.

The Standing Committee on Bus Transit Systems (AP050) peer-reviewed this paper (19-03527).