LeastSquares

1. Introduction

The dataset for this project is the "Bike Sharing Dataset Data Set" found in the UCI Machine Learning Repository. The dataset contains hourly count of rental bikes for all of 2011 and 2012 (January 1, 2011 to December 31, 2012) in the Capital Bikeshare System of Washington D.C. area (Washington-Arlington-Alexandria, DC-VA-MD-WV metropolitan area). The UCI Machine Learning Repository cites Hadi Fanaee-T from the "Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto" for the compilation of the data.

The dataset is outdated since data is actually available up to November 2020 on Capital Bikeshare's website (as of December 18, 2020), but this limited dataset will still work for the purposes of demonstrating linear algebra on a real world dataset.

There are two files included in the dataset: a `hour.csv` and a `day.csv`. We will use the `hour.csv` for the regression, since the `day.csv` is simply just a sumamry of the `hour.csv` file. We also made a function that easily converts the `hour.csv` to the `day.csv` called `convert_hour_to_day()`.

There are 14 different variables that are in this dataset that are potentially of interest. Two variables are not useful and immediately thrown out: `instant` (this is simply the row number of the dataset) and `dteday` (date of the year).

Denote x_n as plausible independent variables and denote y_n as plausible dependent variables.

x_1: `season` (1: spring, 2: summer, 3: fall, 4: winter)

x_2: `yr` (0: 2011, 1: 2012)

x_3: `mnth` (1 to 12)

x_4: `hour` (0 to 23)

x_5: `holiday` (whether a holiday (0 or 1) from this list of holidays)

x_6: `weekday` (0 to 6)

x_7: `workingday` (1 if weekday and not holiday, 0 otherwise)

x_8: `weathersit`: Weather conditions (1: Clear, Few clouds, Partly cloudy, Partly cloudy, 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)

x_9: `temp` (0-1, normalized temperature in Celsius. Divided by 41)

x_{10}: `atemp` (0-1, normalized "feels like" temperature in Celsius. Divided by 50)

x_{11}: `hum` (percent humidity)

x_{12}: `windspeed` (0-1, Normalized wind speed. Divided by 67)

y_1: `casual` (count of casual users)

y_2: `registered` (count of registered users)

y_3: `cnt` (count of sum of casual and registered users)

The following least squares regression exercise will try to predict the `casual`, `registered`, or `cnt` as a function of the independent variables. Also, we will try some principal components analysis (PCA) and k-nearest neighbors (kNN) with these variables.

Data Analysis

Preliminary data analysis shows that the `hour` is by far the most important independent variable for explaining the variation in the dependent variables. Thus, it is important to know how exactly the `hour` variable interacts with `registered`, `casual`, and `cnt`.

We also found that it makes sense to treat `hour` as a categorical variable (treat `hour` as 23 independent dummy variables (0 or 1 for each variable), one for each hour minus the constant term), but in this case, we will try to fit `hour` in terms of a polynomial curve to demonstrate polynomial fitting with linear algebra.

A plot of `hour` on the x-axis and `registered` or `casual` on the y-axis would us some insight of what degree polynomial for `hour` we should be looking for.

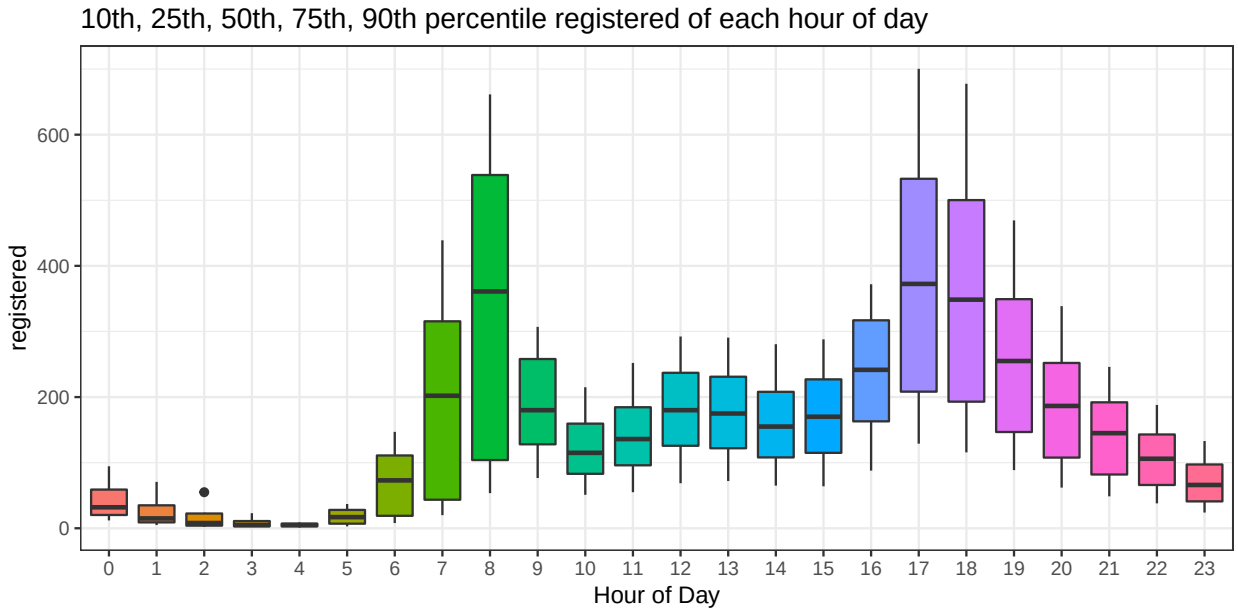

Figure 1: Boxplot of registered users for Jan 1, 2011 to December 31, 2011 for each hour of day. The low whisker represents 10th percentile, the box represents the interquartile range, the top whisker represents the 90th percentile. For example, with 731 days in the datset, the low whisker represents the 73rd lowest value.

Figure 1 for the number of registered users show a trimodal distribution with three peaks throughout the day. A possible interpretation of the three peaks is that there is an early peak for the morning commute, a central peak for lunchtime, and a late peak for the evening commute. This suggests that we need a high-degree polynomial to accurately the number of registered users throughout the day. A six-degree polynomial is the minimum degree that can represent a trimodal distribution.

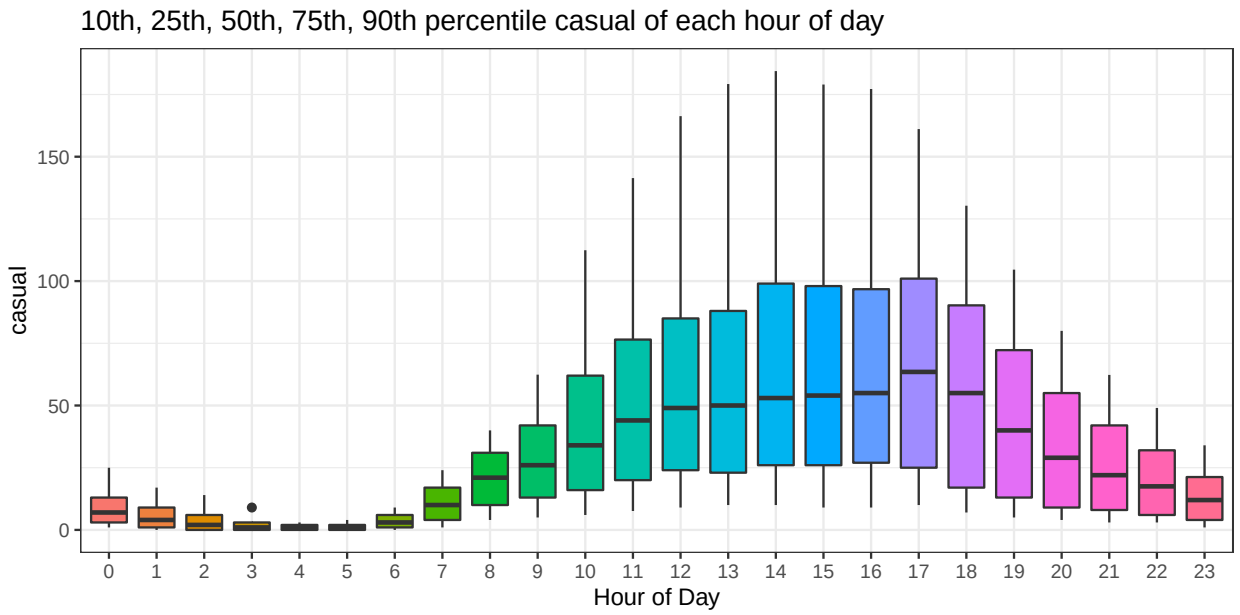

Figure 2: Boxplot of casual users for Jan 1, 2011 to December 31, 2011 for each hour of day. The low whisker represents 10th percentile, the box represents the interquartile range, the top whisker represents the 90th percentile. For example, with 731 days in the datset, the low whisker represents the 73rd lowest value.

Figure 2 for the number of casual users show a single peak distribution. A possible interpretation of this is that casual users are tourists that don't commute. Tourists are not up early in the morning and most things to do for tourists occur in the afternoon. Thus, the peak occurs at around 12 PM-6 PM. A two-degree polynomial is potentially sufficient to represent the number of casual users throughout the day.

Figure 1 and **Figure 2** show significant variation so that it is clear that the hour of the day is not the only variable influencing how many users there are for this bike sharing system. We can plot at the number of users for each day for further information.

Figure 3 shows both variation through the year and an increase in number of users from 2011 to 2012. This makes sense if this Capital Bike-Sharing was still a developing system in 2011 and not yet a mature system where the market is already saturated. Thus, we want to include a variable in our least squares regression model that includes controls for seasonal variation and the year. This would be `season` and `yr` from the list of x_n. It turns out that the `weekday` (day of the week) and `mnth` (month of the year) do not explain a large additional amount of variation, so they won't be included in the model.

But **Figure 3** still shows quite a bit of variation day to day within each season. There are quite a few weather related variables in the list of independent variables in the dataset, which would account for some of the remaining variation.

Figure 4 shows the average `temp` (actual temperature) and `atemp` ("feels like" temperature) for each day of the year. A comparison of **Figure 3** and **Figure 4** shows that temperature shows a strong negative relationship with the number of bike-sharing users, which makes sense. People do not want to bike when it is cold outside. These variables have a very high correlation with each other such that we found it to be sufficient to just add `temp`. In fact, `temp` is good enough to explain most of the weather-related variation and other variables such as `wind` (wind speed) and `hum` (humidity) are not necessary.

Thus, we choose just `season`, `yr`, `hr`, and `temp` as the independent variables and omit the rest of the variables for the least squares regression model. These are x_1, x_2, x_4, and x_9.

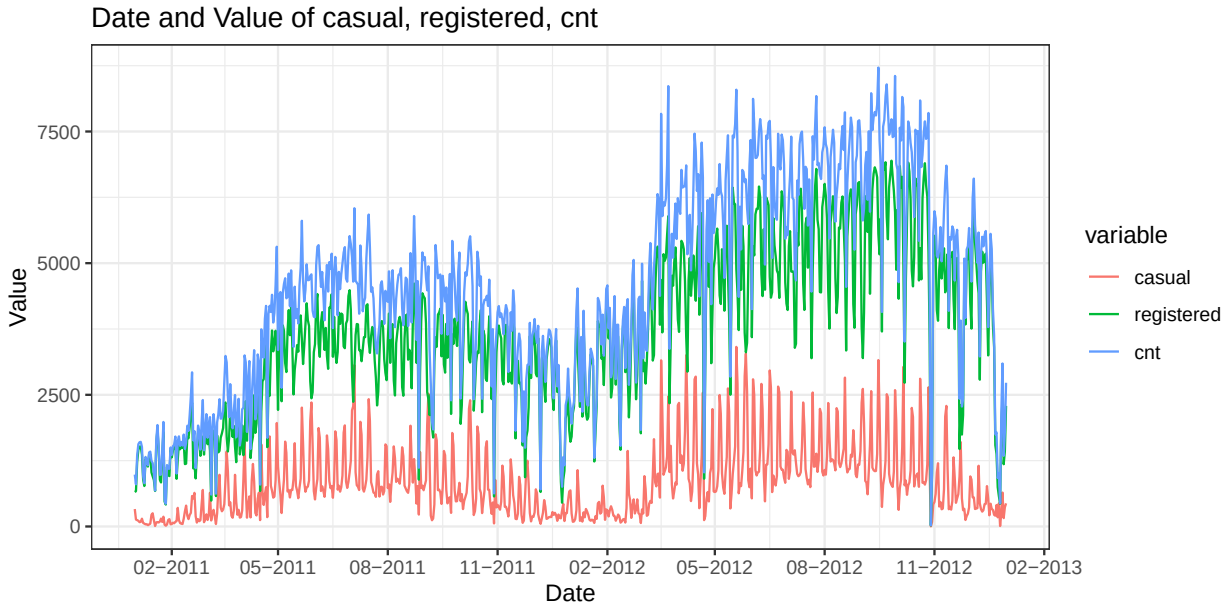

Figure 3: Number of casual, registered, and sum of casual and registered for all 731 days in the dataset

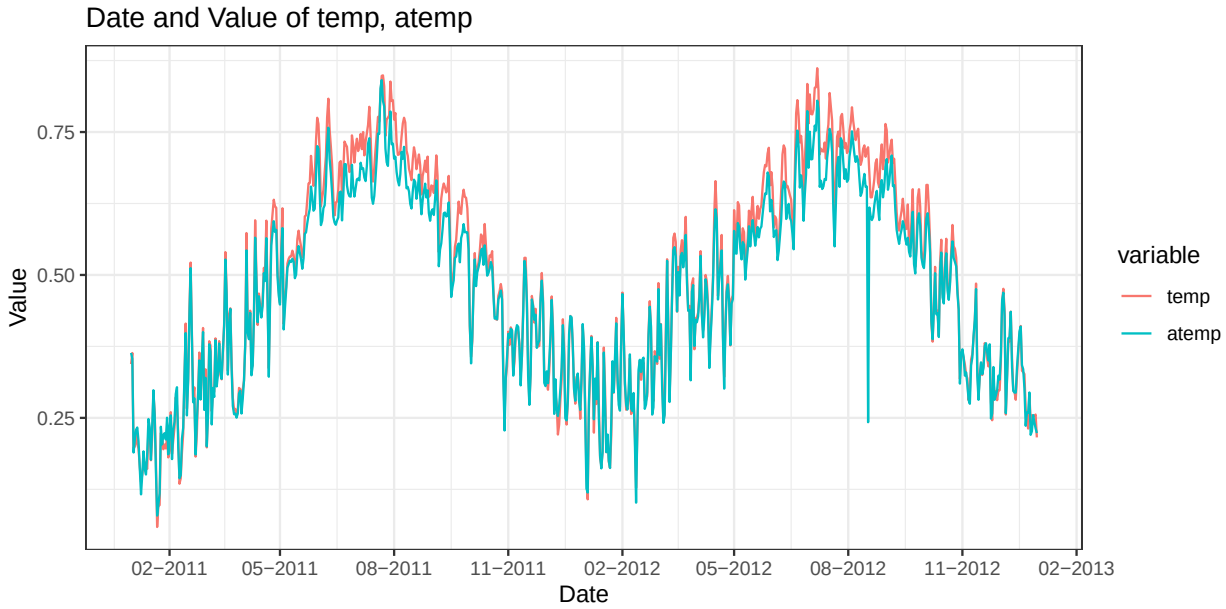

Figure 4: Normalized temperature and "feels like" temperature for all 731 days in the datset

2. Design Matrix

The design matrix A is the matrix that will be used to solve the equation $Ax = b$, where x are the coefficients, often referred to as the β's and b is the dependent variable. It is the matrix of the explanatory/independent variables.

The design matrix A will be of $\mathbb{R}^{17379\times(n+4)}$, (where n is the maximum degree of the polynomial for x_4/`hour`) since there are 17,379 rows (one for each hour in the dataset) and there are four variables other than the x_4/`hour` variable. The other four variables are the constant term, x_1 (`season`), x_2 (`yr`), and x_9 (`temp`).

The following will be the form of our design matrix $A \in \mathbb{R}^{17379\times(n+4)}$ (Note that m is left the matrix for simplification, but the of $m = 17379$):

$$A = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{14} & x_{14}^2 & \dots & x_{14}^n & x_{19} \\ 1 & x_{21} & x_{22} & x_{24} & x_{24}^2 & \dots & x_{24}^n & x_{29} \\ 1 & x_{31} & x_{32} & x_{34} & x_{34}^2 & \dots & x_{34}^n & x_{39} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{m1} & x_{m2} & x_{m4} & x_{m4}^2 & \dots & x_{m4}^n & x_{m9} \end{bmatrix}$$

Now the question remaining is the appropriate value of n. We know that the answer depends on which dependent variable we are using. If b is

Table 1: R-squared values for linear regression of each of value of n, maximum power of polynomial for `hour`

	R^2_{cnt}	R^2_{casual}	$R^2_{registered}$
x_4	0.343	0.285	0.291
$x_4 + x_4^2$	0.464	0.371	0.394
$x_4 + x_4^2 + x_4^3$	0.515	0.430	0.431
$x_4 + x_4^2 + \cdots + x_4^4$	0.516	0.440	0.436
$x_4 + x_4^2 + \cdots + x_4^5$	0.530	0.446	0.464
$x_4 + x_4^2 + \cdots + x_4^6$	0.552	0.446	0.497
$x_4 + x_4^2 + \cdots + x_4^7$	0.589	0.446	0.551
$x_4 + x_4^2 + \cdots + x_4^8$	0.599	0.446	0.565
$x_4 + x_4^2 + \cdots + x_4^9$	0.606	0.447	0.575
$x_4 + x_4^2 + \cdots + x_4^{10}$	0.606	0.447	0.575
$x_4 + x_4^2 + \cdots + x_4^{11}$	0.613	0.447	0.585
$x_4 + x_4^2 + \cdots + x_4^{12}$	0.614	0.447	0.585
$x_4 + x_4^2 + \cdots + x_4^{13}$	0.638	0.448	0.619
$x_4 + x_4^2 + \cdots + x_4^{14}$	0.638	0.448	0.619
$x_4 + x_4^2 + \cdots + x_4^{15}$	0.638	0.448	0.619

Table 2: Comparison of speed (in milliseconds) of custom design_matrix() functions with built-in R model.matrix.lm() function

expr	min	lq	mean	median	uq	max	neval
design_matrix()	19.627	20.860	29.634	22.609	30.281	172.953	100
design_matrix_Cpp()	7.689	8.132	10.826	8.583	10.822	30.184	100
model.matrix.lm()	19.797	22.946	32.592	28.127	32.765	217.489	100

3. Normal Equation

Table 3: Condition number of each value of n (maximum power of polynomial for `hour`) and relative error of normal equations versus SVD

	$\kappa(A)$	$\kappa(A)^2$	Relative error/error message
x_4	8.05×10^{1}	6.49×10^{3}	5.87×10^{-12}
$x_4 + x_4^2$	1.45×10^{3}	2.09×10^{6}	-1.1×10^{-14}
$x_4 + x_4^2 + x_4^3$	3.01×10^{4}	9.08×10^{8}	1.18×10^{-12}
$x_4 + x_4^2 + \cdots + x_4^4$	6.46×10^{5}	4.17×10^{11}	2.87×10^{-12}
$x_4 + x_4^2 + \cdots + x_4^5$	1.42×10^{7}	2.03×10^{14}	1.52×10^{-8}
$x_4 + x_4^2 + \cdots + x_4^6$	3.59×10^{8}	1.29×10^{17}	Error, Recripocal $\kappa(A)$: 5.89×10^{-18}
$x_4 + x_4^2 + \cdots + x_4^7$	1.06×10^{10}	1.12×10^{20}	Error, Recripocal $\kappa(A)$: 6.14×10^{-21}
$x_4 + x_4^2 + \cdots + x_4^8$	3.33×10^{11}	1.11×10^{23}	Error, Recripocal $\kappa(A)$: 6.35×10^{-24}
$x_4 + x_4^2 + \cdots + x_4^9$	1.10×10^{13}	1.22×10^{26}	Error, Recripocal $\kappa(A)$: 6.78×10^{-27}
$x_4 + x_4^2 + \cdots + x_4^{10}$	3.82×10^{14}	1.46×10^{29}	Error, Recripocal $\kappa(A)$: 1.29×10^{-29}
$x_4 + x_4^2 + \cdots + x_4^{11}$	1.39×10^{16}	1.94×10^{32}	Error, Recripocal $\kappa(A)$: 1.07×10^{-32}
$x_4 + x_4^2 + \cdots + x_4^{12}$	5.38×10^{17}	2.89×10^{35}	Error, Recripocal $\kappa(A)$: 7.16×10^{-35}
$x_4 + x_4^2 + \cdots + x_4^{13}$	2.21×10^{19}	4.86×10^{38}	Error, Recripocal $\kappa(A)$: 1.13×10^{-37}
$x_4 + x_4^2 + \cdots + x_4^{14}$	5.86×10^{22}	3.43×10^{45}	Error, Recripocal $\kappa(A)$: 2.67×10^{-40}
$x_4 + x_4^2 + \cdots + x_4^{15}$	4.51×10^{22}	2.04×10^{45}	Error, Recripocal $\kappa(A)$: 2.51×10^{-43}

Table 4: Comparison of speed (in milliseconds) of solving $Ax = b$ where $A \in \mathbb{R}^{17379 \times 7}$ using normal equations implemented in R vs Rcpp (C++)

expr	min	lq	mean	median	uq	max	neval
normal_equations(A, b)	5.976	6.734	8.806	7.379	9.161	31.031	100
normal_equations_Cpp(A, b)	4.931	5.243	6.160	5.742	6.741	11.054	100

4. QR Decomposition

Table 5: Comparison of speed (in milliseconds) of solving $Ax = b$ where $A \in \mathbb{R}^{17379 \times 7}$ using QR decomposition implemented in R vs Rcpp (C++)

expr	min	lq	mean	median	uq	max	neval
qr.solve(A, b)	5.391	6.532	9.216	7.226	9.042	29.765	100
qr_solve_Cpp(A, b)	5.396	5.677	6.915	6.271	7.499	14.989	100

5. Singular Value Decomposition

Table 6: Comparison of speed (in milliseconds) of solving $Ax = b$ where $A \in \mathbb{R}^{17379 \times 7}$ using SVD implemented in R vs Rcpp (C++)

expr	min	lq	mean	median	uq	max	neval
svd_solve(A, b)	9.523	11.059	13.821	12.049	14.157	38.720	100

expr	min	lq	mean	median	uq	max	neval
svd_solve_Cpp(A, b)	8.190	8.734	10.369	9.620	11.021	35.461	100

Table 7: Comparison of speed (in milliseconds) of solving $Ax = b$ where $A \in \mathbb{R}^{17379\times 7}$ using normal equations, QR decomposition, and SVD implemented in R vs Rcpp (C++)

expr	min	lq	mean	median	uq	max	neval
normal_equations(A, b)	6.058	7.092	9.496	7.808	9.591	31.948	100
normal_equations_Cpp(A, b)	4.906	5.231	6.371	5.875	6.940	13.142	100
qr.solve(A, b)	5.518	6.950	12.626	8.000	10.654	235.429	100
qr_solve_Cpp(A, b)	5.440	5.849	7.795	6.411	7.675	73.373	100
svd_solve(A, b)	9.636	12.183	16.248	14.263	18.056	66.454	100
svd_solve_Cpp(A,b)	8.298	9.201	11.884	10.529	12.813	57.141	100

Table 8: Comparison of speed (in milliseconds) of solving $Ax = b$ where $A \in \mathbb{R}^{17379\times 11}$ using QR decomposition or SVD implemented in R vs Rcpp (C++)

expr	min	lq	mean	median	uq	max	neval
qr.solve(A, b)	10.327	12.829	18.078	15.083	20.989	55.218	100
qr_solve_Cpp(A, b)	11.714	12.863	16.213	14.736	17.559	41.542	100
svd_solve(A, b)	20.475	25.185	33.355	30.504	37.977	78.544	100
svd_solve_Cpp(A,b)	17.806	19.218	24.953	20.885	23.506	140.779	100

6. Final Regression Model

$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_4 + \beta_4 x_4^2 + \beta_5 x_4^3 + \beta_6 x_9$

$y_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_4 + \beta_4 x_4^2 + \beta_5 x_4^3 + \beta_6 x_4^4 + \beta_7 x_4^5 + \beta_8 x_4^6 + \beta_9 x_4^7 + \beta_{10} x_9$

With the data inputed, we have

$y_1 = -42.145 + 0.6875 x_1 + 12.887 x_2 - 4.859 x_4 + 1.275 x_4^2 - 0.047 x_4^3 + 89.528 x_9$

$y_3 = -145.05 + 17.18 x_1 + 88.57 x_2 + 29.45 x_4 - 63.199 x_4^2 + 24.27 x_4^3 - 3.62 x_4^4 + 0.258 x_4^5 - 0.0088 x_4^6 + 0.000116 x_4^7 + 243.131 x_9$

Note that $\frac{1}{n}\sum_{i=1}^{n} x_1 = \bar{x_1} = 2.50164$, $\frac{1}{n}\sum_{i=1}^{n} x_2 = \bar{x_2} = 0.5025606$, $\frac{1}{n}\sum_{i=1}^{n} x_8 = \bar{x_8} = 0.4970$.

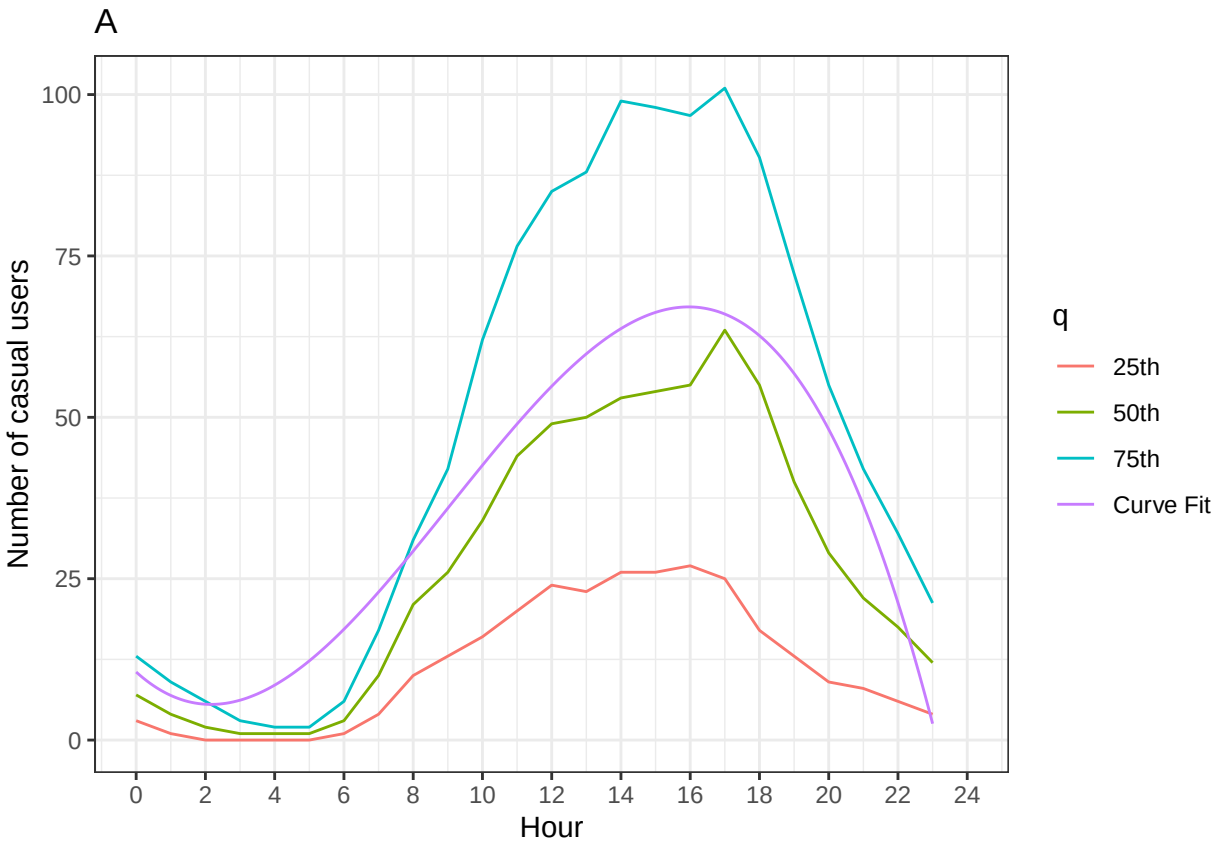
A
Number of casual users
100
75
50
25
0
0
2
4
6
8
10
12
14
16
18
20
22
24
Hour
q
25th
50th
75th
Curve Fit

Figure 5: A

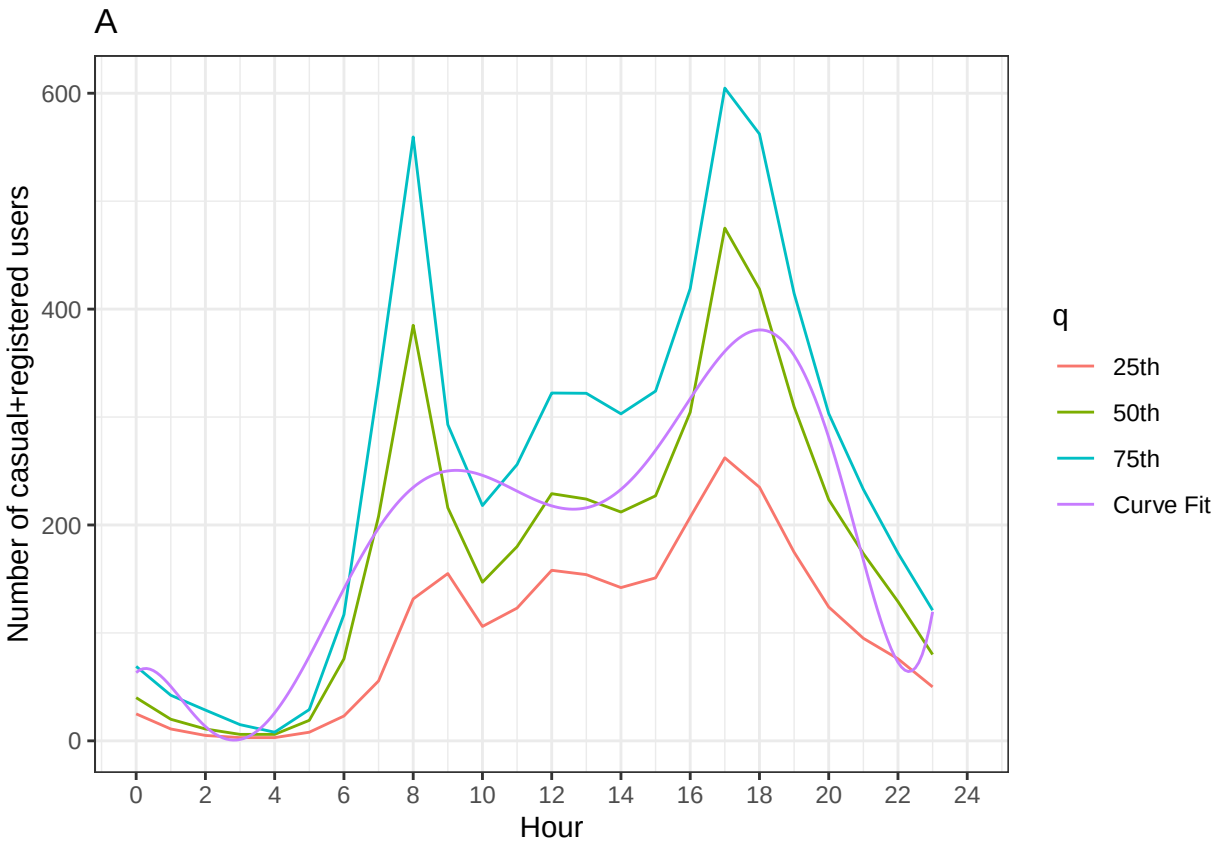
A
600
400
200
0
Number of casual+registered users
0
2
4
6
8
10
12
14
16
18
20
22
24
Hour
q
25th
50th
75th
Curve Fit

Figure 6: A