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The dataset is outdated since data is actually available up to November 2020 on Capital Bikeshare's website (as of December 18, 2020), but this limited dataset will still work for the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 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: `temp` (0-1, normalized temperature in Celsius. Divided by 41)

x_9: `atemp` (0-1, normalized "feels like" temperature in Celsius. Divided by 50)

x_{10}: `hum` (percent humidity)

x_{11}: `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.

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`.

Although there are three different dependent variables,

It could sense to treat `hour` as a categorical variable (treat `hour` as 23 independent variables, one for each hour minus the constant term), but in this case, we will try to fit `hour` in terms of a polynomial curve.

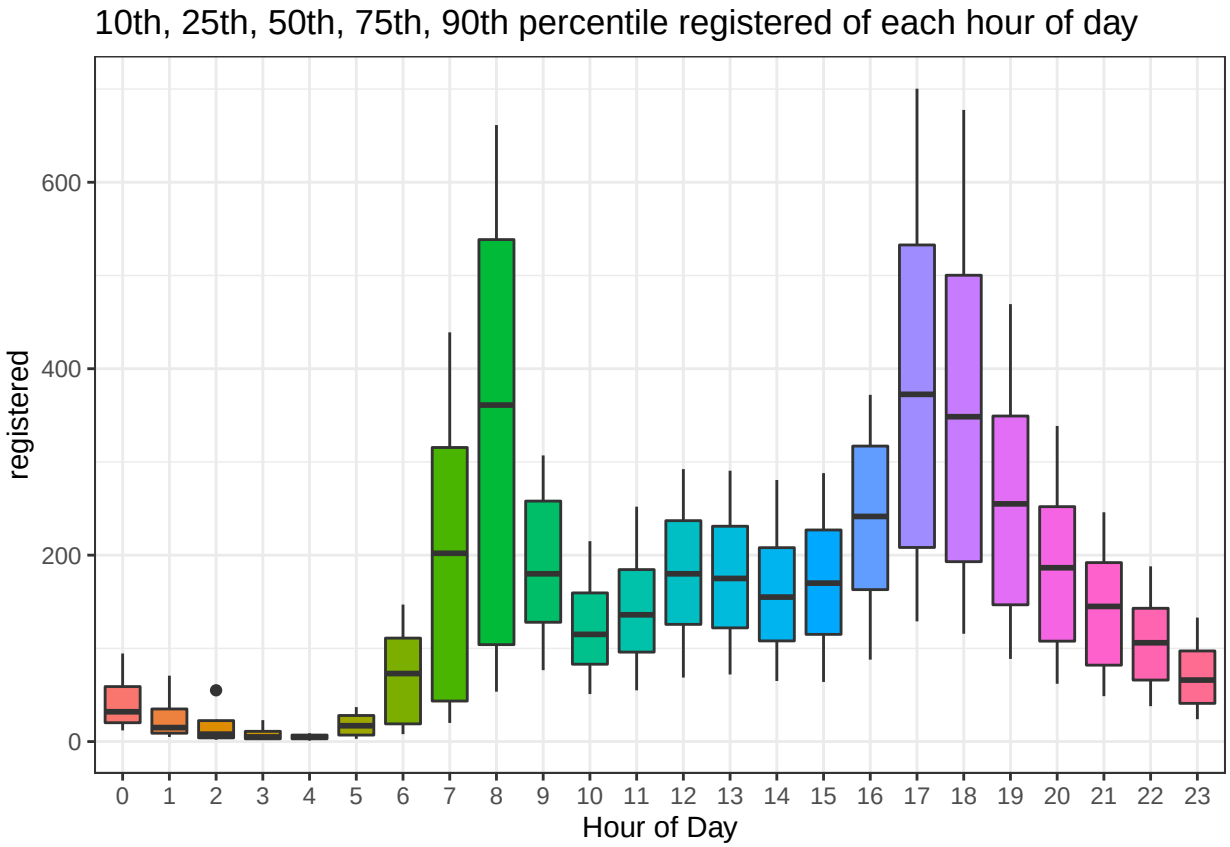


Figure 1: Boxplot of registered users for Jan 1, 2011 to December 31, 2011 for each hour of day. The low whisker represents 10th percentile, the box represents the interquartile range, the top whisker represents the 90th percentile

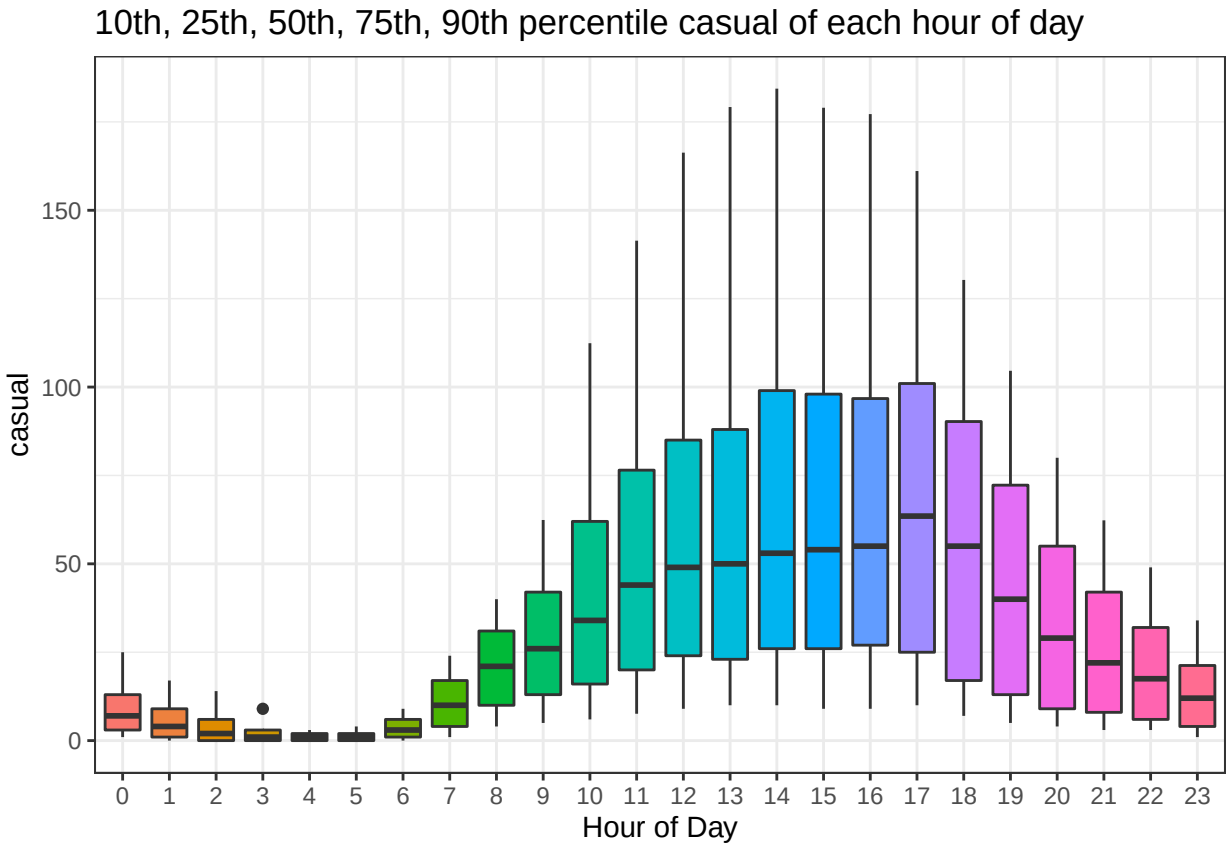


Figure 2: Boxplot of casual users for Jan 1, 2011 to December 31, 2011 for each hour of day. The low whisker represents 10th percentile, the box represents the interquartile range, the top whisker represents the 90th percentile

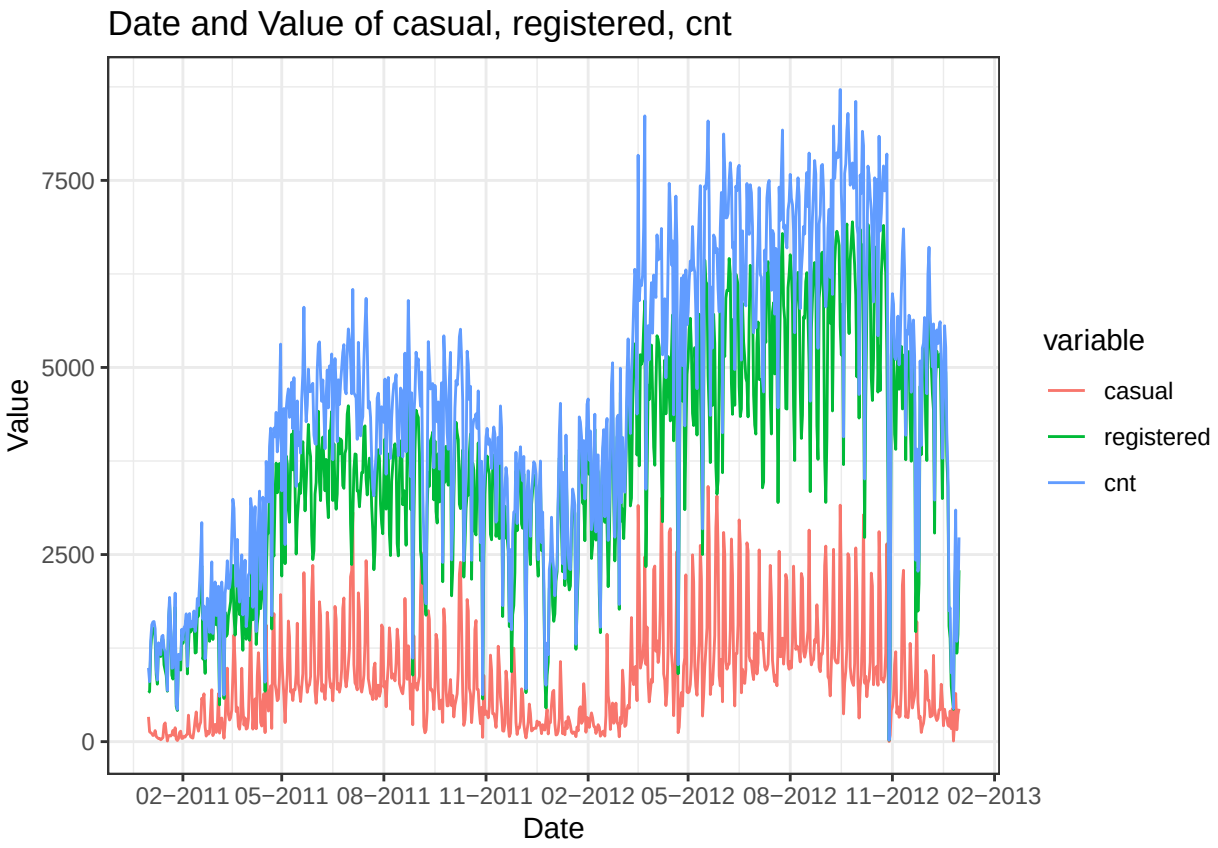


Figure 3: Number of casual, registered, and sum of casual and registered for all 731 days in the dataset

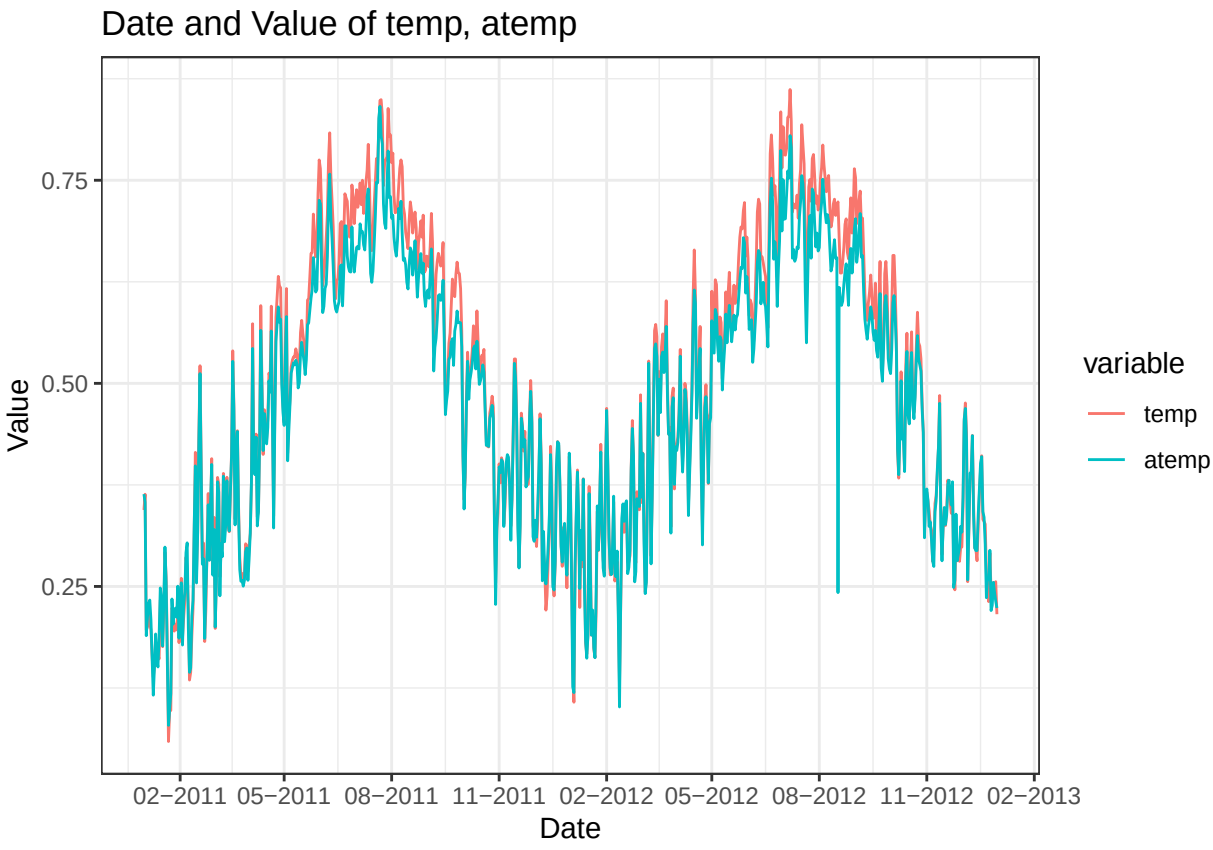


Figure 4: Normalized temperature and "feels like" temperature for all 731 days in the datset

2. Design Matrix

Table 1: R-squared values of y

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

expr	min	lq	mean	median	uq	max	neval
design_matrix()	19.831	20.682	23.796	21.782	24.580	57.546	100
design_matrix_Cpp()	8.825	9.160	10.775	9.532	10.675	22.519	100
mode.matrix.lm()	19.763	20.907	26.939	22.960	28.242	184.484	100

3. Normal Equation

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

expr	min	lq	mean	median	uq	max	neval
normal_equations(A, b)	6.045	6.601	9.521	7.551	9.587	78.223	100
normal_equations_Cpp(A, b)	4.806	5.100	6.516	5.872	7.349	13.222	100

4. QR Decomposition

expr	min	lq	mean	median	uq	max	neval
qr.solve(A, b)	5.315	5.743	8.441	6.576	8.769	23.07	100
qr_solve_Cpp(A, b)	5.277	5.629	6.554	6.176	6.970	10.95	100

5. Singular Value Decomposition

expr	min	lq	mean	median	uq	max	neval
svd_solve(A, b)	9.591	10.415	13.097	11.413	14.236	26.682	100
svd_solve_Cpp(A, b)	8.219	8.881	10.449	9.491	11.173	26.194	100

expr	min	lq	mean	median	uq	max	neval
normal_equations(A, b)	5.999	6.281	7.377	6.588	7.369	15.332	100
normal_equations_Cpp(A, b)	4.791	4.983	5.500	5.305	5.702	8.665	100
qr.solve(A, b)	5.390	5.761	7.674	6.124	7.306	17.659	100
qr_solve_Cpp(A, b)	5.317	5.615	6.218	5.936	6.386	9.740	100
svd_solve(A, b)	9.675	10.128	12.649	10.710	13.361	30.296	100
svd_solve_Cpp(A,b)	8.107	8.510	9.575	9.102	10.133	14.946	100

expr	min	lq	mean	median	uq	max	neval
qr.solve(A, b)	10.311	10.876	14.072	11.902	17.806	28.556	100
qr_solve_Cpp(A, b)	11.867	12.481	13.767	12.926	14.112	28.182	100
svd_solve(A, b)	20.678	22.220	28.869	26.057	29.520	189.340	100
svd_solve_Cpp(A,b)	18.160	19.295	21.691	20.492	23.977	37.163	100

6. Final Regression Model

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

$y_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_4 + \beta_4 x_4^2 + \beta_5 x_4^3 + \beta_6 x_4^4 + \beta_7 x_4^5 + \beta_8 x_4^6 + \beta_9 x_4^7 + \beta_{10} x_8$

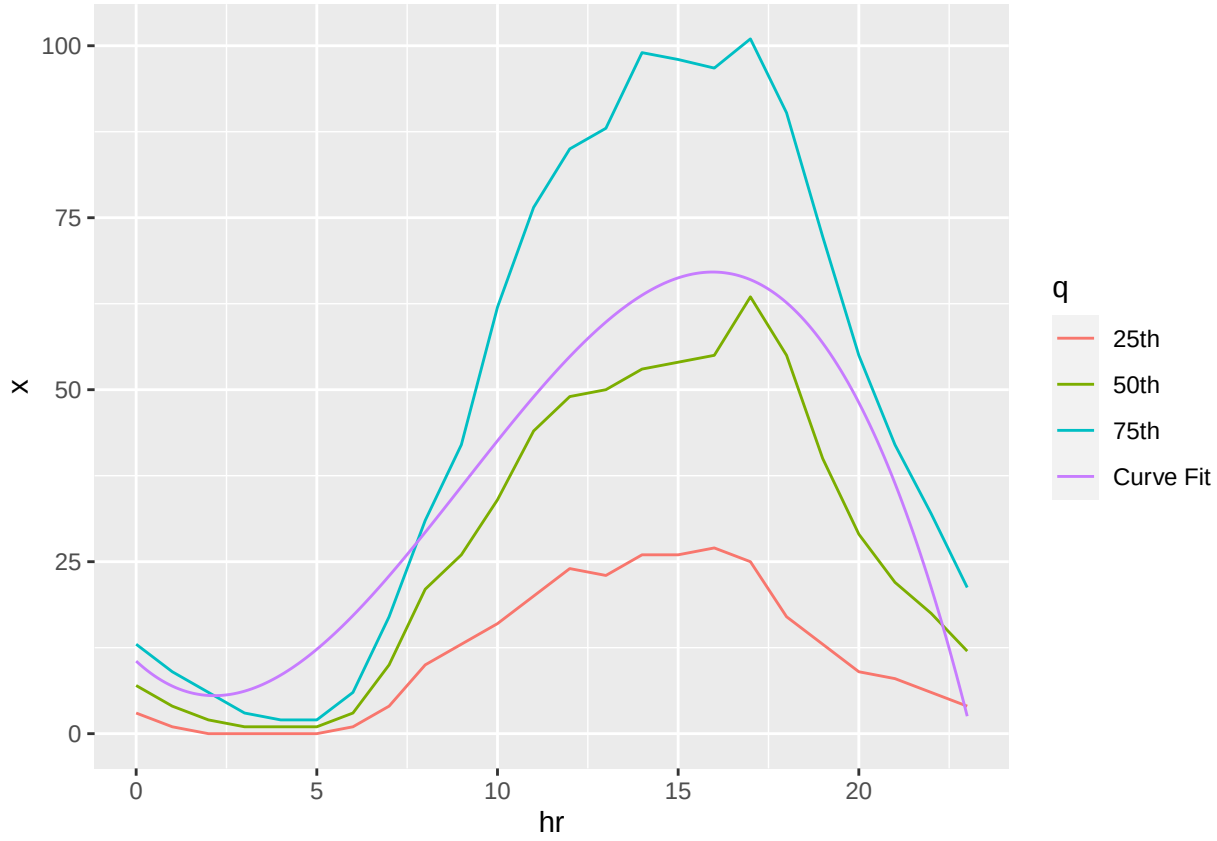
With the data inputted, we have

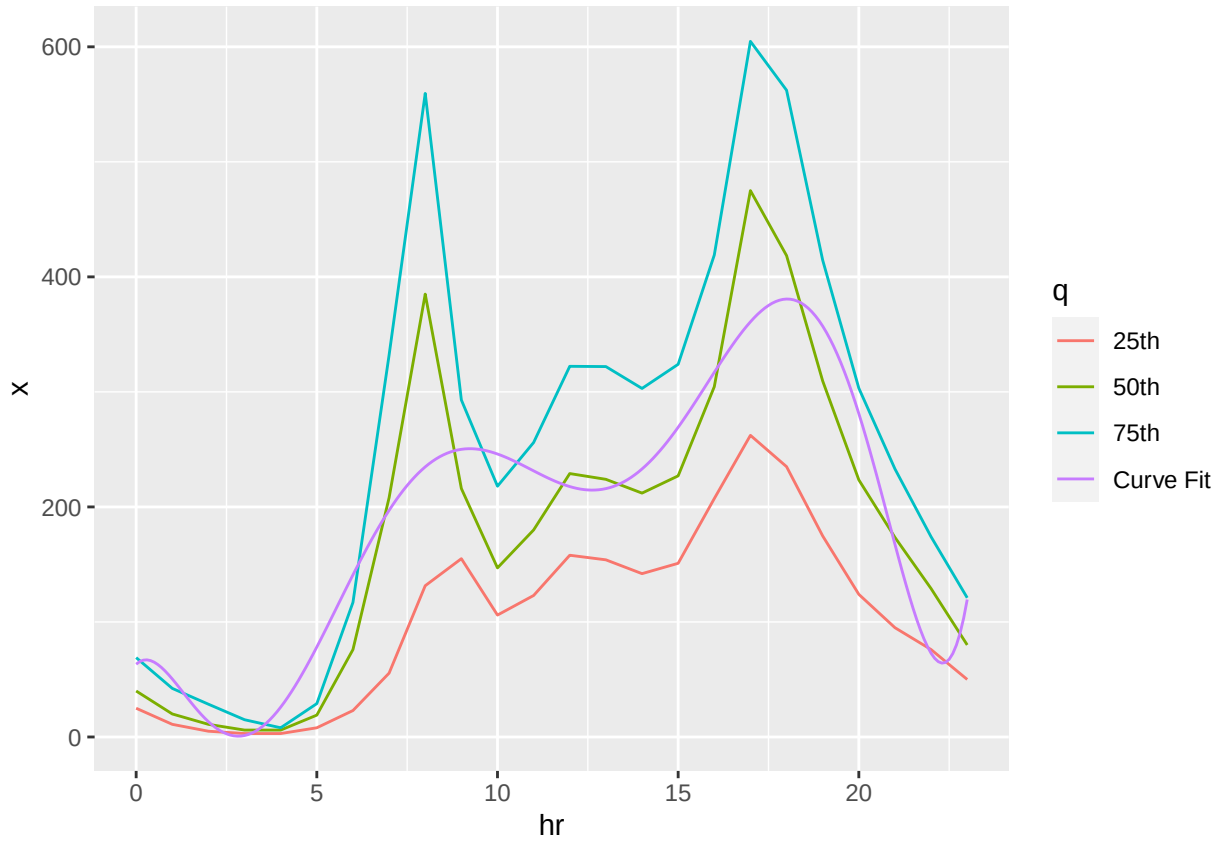
$y_1 = -42.145 + 0.6875 x_1 + 12.887 x_2 - 4.859 x_4 + 1.275 x_4^2 - 0.047 x_4^3 + 89.528 x_8$

$y_3 = -145.05 + 17.18 x_1 + 88.57 x_2 + 29.45 x_4 - 63.199 x_4^2 + 24.27 x_4^3 - 3.62 x_4^4 + 0.258 x_4^5 - 0.0088 x_4^6 + 0.000116 x_4^7 + 243.131 x_8$

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

```
## `summarise()` regrouping output by 'hr' (override with `.groups` argument)
## `summarise()` regrouping output by 'hr' (override with `.groups` argument)
```




600
400
200
0
x
0
5
10
15
20
hr
q
25th
50th
75th
Curve Fit