Quantile Regression for Final Hospitalization Rate Prediction

Nuoyu Li *
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
nuoyul@cs.cmu.edu

1 Introduction

Influenza (the flu) has become a serious threat to human health since its birth. In the U.S., influenza approximately causes between 9.2 million and 35.6 million illnesses, between 140,000 and 710,000 hospitalizations and between 12,000 and 56,000 deaths annually since 2010 [3]. Although protective methods - vaccination, disinfection, and antiviral drugs - can reduce the intensity and spread of viruses, they have societal and monetary costs [2, 6]. Therefore, maintaining ongoing surveillance and forecasting future trends of flu activities are critical. Accurate reports and predictions will not only detect potential flu outbreaks in certain areas, but also lead to economical applications of protective methods. For social good, researchers has been developing and improving flu prediction systems in recent years.

This project contributes to existing flu forecasting systems by analyzing hospitalization rates reported by Center for Disease Control and Prevention (CDC). Quantile regression accurately generates confidence intervals and predicts final hospitalization rates with more than 80% variance explained for each age group in every season.

2 Background

2.1 Terms and Notations

To introduce the background of the backfill prediction problem for hospitalization rates, we clarify several terms used by CDC:

1. **Epiweek**: An epidemiological week (epiweek) starts from Sunday and ends on Saturday. The first epiweek of the year is the week of the first Wednesday in that year.
2. **Flu Season**: According to CDC, a flu season lasts from the 40^{th} epiweek (early October) to the 20^{th} epiweek of the next calendar year.
3. **Hospitalization Rate**: CDC actively monitors by reporting flu-associated hospitalization cases from the Influenza Hospitalization Surveillance Network (FluSurv-NET). The following paragraph adapted from CDC FluView interface describes several basic facts of the network [5]:

   The network covers more than 70 counties across 13 different states (CA, CO, CT, GA, MD, MN, NM, NY, OR, TN, MI, OH, and UT). The network represents approximately 8.5% of US population (~27 million people). CDC defines hospitalization rate as following:

   $$\text{hosptialization rate} = \frac{\text{\# people seeing healthcare providers for influenza-like illness}}{\text{\# population}}$$

   The reports of hospitalization rate are updated each week for 5 different age groups: 0-4 years old, 5-17 years old, 18-49 years old, 50-64 years old, and more than 65 years old. The numerical

*Advisor: Roni Rosenfeld (roni@cs.cmu.edu), Co-advisor: Ryan Tibshirani (ryantibs@cmu.edu)

representation of the rate r means "r people out of 100,000 people are hospitalized". CDC reports hospitalization data for the entire network and the 13 states above from the beginning of flu season to the 17$^{\text{th}}$ epiweek of the next year.

4. **Backfill**: For a particular epiweek, CDC will report its hospitalization rate in the next epiweek. The initial report in the next week is called "the first issue" of the rate. However, since CDC works with different hospitals across the nation, the data collection usually takes significant amount of time. Therefore, CDC will continue to adjust the initial report until the flu season ends. The discrepancy between the final hospitalization rate and its first issue is defined as backfill. The definition is further explained by figure 1.

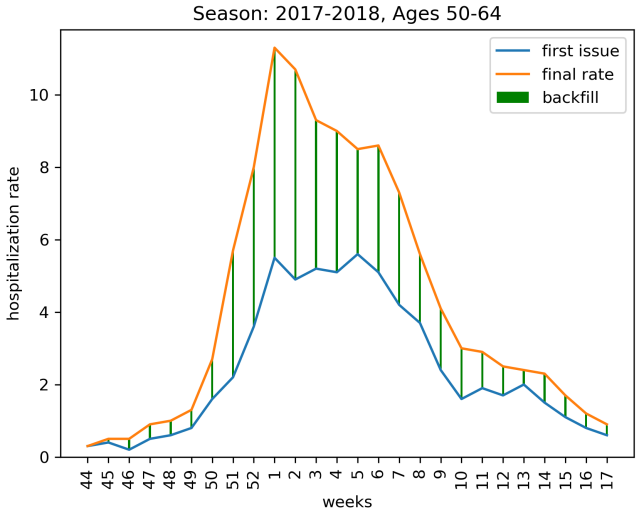


Figure 1: An Example of Backfill.

5. **Lag**: The time difference between a report of hospitalization rate and the first issue of that rate.

2.2 Problem Formulation

To formulate the problem, we firstly introduce following notations:

1. $r_{(\text{e,g});\text{l}}$: the hospitalization rate for epiweek e and age group g with lag l.
2. $r_{(\text{e,g});\text{F}}$: the final hospitalization rate for epiweek e.
3. $\hat{r}_{(\text{e,g});\text{F}}$: the estimate of final hospitalization rate.
4. $\hat{r}^{\frac{\alpha}{2}}_{(\text{e,g});\text{F}}$, $\hat{r}^{1-\frac{\alpha}{2}}_{(\text{e,g});\text{F}}$: the estimates of lower / upper bound of final hospitalization rate.
5. w: the epiweek window. i.e. maximum epiweek difference between predictive features and the predicted hospitalization rate. The definition is demonstrated by the x-axis of figure 4.

The project aims to forecast final hospitalization rates given hospitalization rates available up to their first issues. Formally, given time window w, any epiweek e, and age group g, a machine learning model $f : \mathbb{R}^{w(w+1)/2} \to \mathbb{R}$ is trained for predicting $r_{(\text{e,g});\text{F}}$:

$$\hat{r}_{(\text{e,g});\text{F}} = f(r_{(\text{e}-\text{w}+1,\text{g});0}, \cdots, r_{(\text{e}-\text{w}+1,\text{g});\text{w}-1}, r_{(\text{e}-\text{w}+2,\text{g});0}, \cdots, r_{(\text{e,g});0}) \tag{1}$$

In addition, an confidence interval estimation is usually provided in epidemic forecasting because the upper and lower bounds respectively indicate the worst and best situations. Consequently, given the significance α of confidence interval, an upper bound model $f_{1-\frac{\alpha}{2}}$ and a lower bound model $f_{\frac{\alpha}{2}}$ lead to an interval prediction:

$$\begin{aligned}
\hat{r}^{1-\frac{\alpha}{2}}_{(\text{e,g});\text{F}} &= f_{1-\frac{\alpha}{2}}(r_{(\text{e}-\text{w}+1,\text{g});0}, \cdots, r_{(\text{e}-\text{w}+1,\text{g});\text{w}-1}, r_{(\text{e}-\text{w}+2,\text{g});0}, \cdots, r_{(\text{e,g});0}) \\
\hat{r}^{\frac{\alpha}{2}}_{(\text{e,g});\text{F}} &= f_{\frac{\alpha}{2}}(r_{(\text{e}-\text{w}+1,\text{g});0}, \cdots, r_{(\text{e}-\text{w}+1,\text{g});\text{w}-1}, r_{(\text{e}-\text{w}+2,\text{g});0}, \cdots, r_{(\text{e,g});0})
\end{aligned} \tag{2}$$

Remark. The machine learning model f, $f_{1-\frac{\alpha}{2}}$, and $f_{\frac{\alpha}{2}}$ are "nowcast" sensors in epidemic forecasting because they are predicting the flu activity level of the current epiweek. We can modify the time frame of the input to create a "backcast" sensor or a "forecast" sensor.

3 Data Analysis

We obtain hospitalization rate data from FluSurv interface provided by CDC [1]. This section analyzes several aspects of the data in detail. The analysis is necessary because a solid understanding of data is critical for machine learning method selection.

3.1 Size of Data

All hospitalization rates from flu seasons 2011-12 to 2017-18 are queried successfully for 13 states and the whole network across 5 age groups. Each flu season includes 30 epiweeks. Therefore, for a single combination of location and age group, there are $30 \times 6 = 180$ hospitalization rate entries.

3.2 Missing Data

A significant amount of queried reports are empty. Empty reports are indicated by nonexistent or zero final hospitalization rate. Missing data is due to query failure and CDC report rule.

Query Failure. The FluSurv interface fails to query a significant number of existing records. For example, according to the interactive interface provided by CDC, the records for states like California and Oregon are complete for different age groups since flu season 2005-06 [4]. However, the interface is not able to return any data before season 2011-12. The situation is similar for other states.

CDC Report Rule. CDC does not report hospitalization rates until the number of cases exceeds a certain amount in a specified location. If the number of cases is lower, CDC will simply report hospitalization rate as 0. To quantify the effect of CDC report rule, the proportions of zero rates for several states and entire network are summarized in figure 2.

From the plots we observe that state-level hospitalization rates contain a significant proportion of zeros due to insufficient number of cases. Meanwhile, network-level hospitalization rates only contain a few zeros and the reports are even complete for several seasons.

3.3 Backfill Analysis

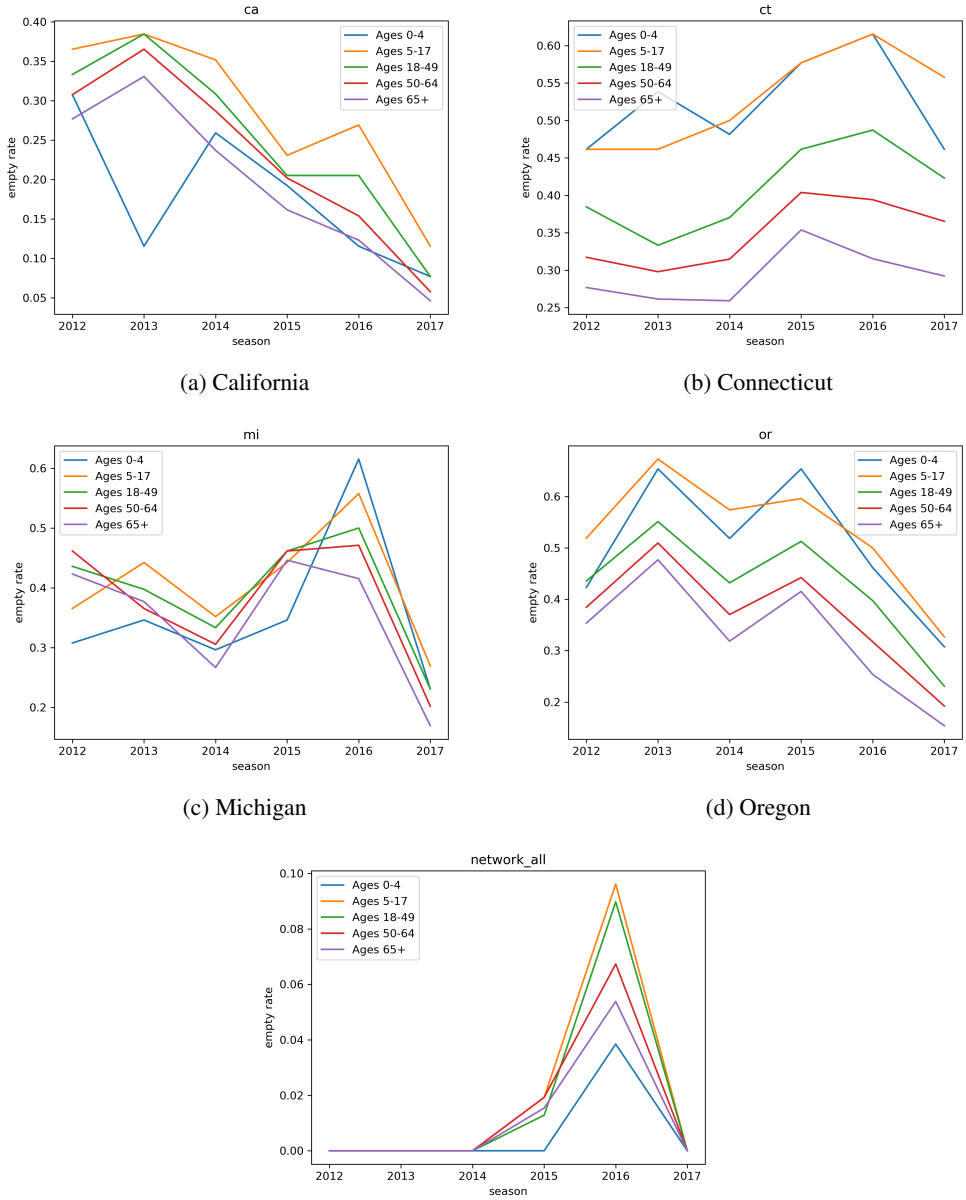
This analysis will identify the variations of backfill behavior from seasons 2011-12 to 2017-18 for different states and age groups. The results will determine the way we select training data.

To quantify backfill behavior for a flu season, we define seasonal backfill rate $br_{s,g}$ for a season s and age group g as the ratio of cumulative backfill to cumulative final hospitalization rate:

$$br_{s,g} = 1 - \sum_{e \in s} r_{(e,g),0} / \sum_{e \in s} r_{(e,g),F}$$

The seasonal backfill rates for several states and the entire network are shown in figure 3.

From the plots we firstly observe that backfill does not exist during season 2011-12. In addition, the backfill behaviors state-level hospitalization rates vary considerably across 7 seasons, while the behavior of network-level rates are relatively stable.



(a) California

(b) Connecticut

(c) Michigan

(d) Oregon

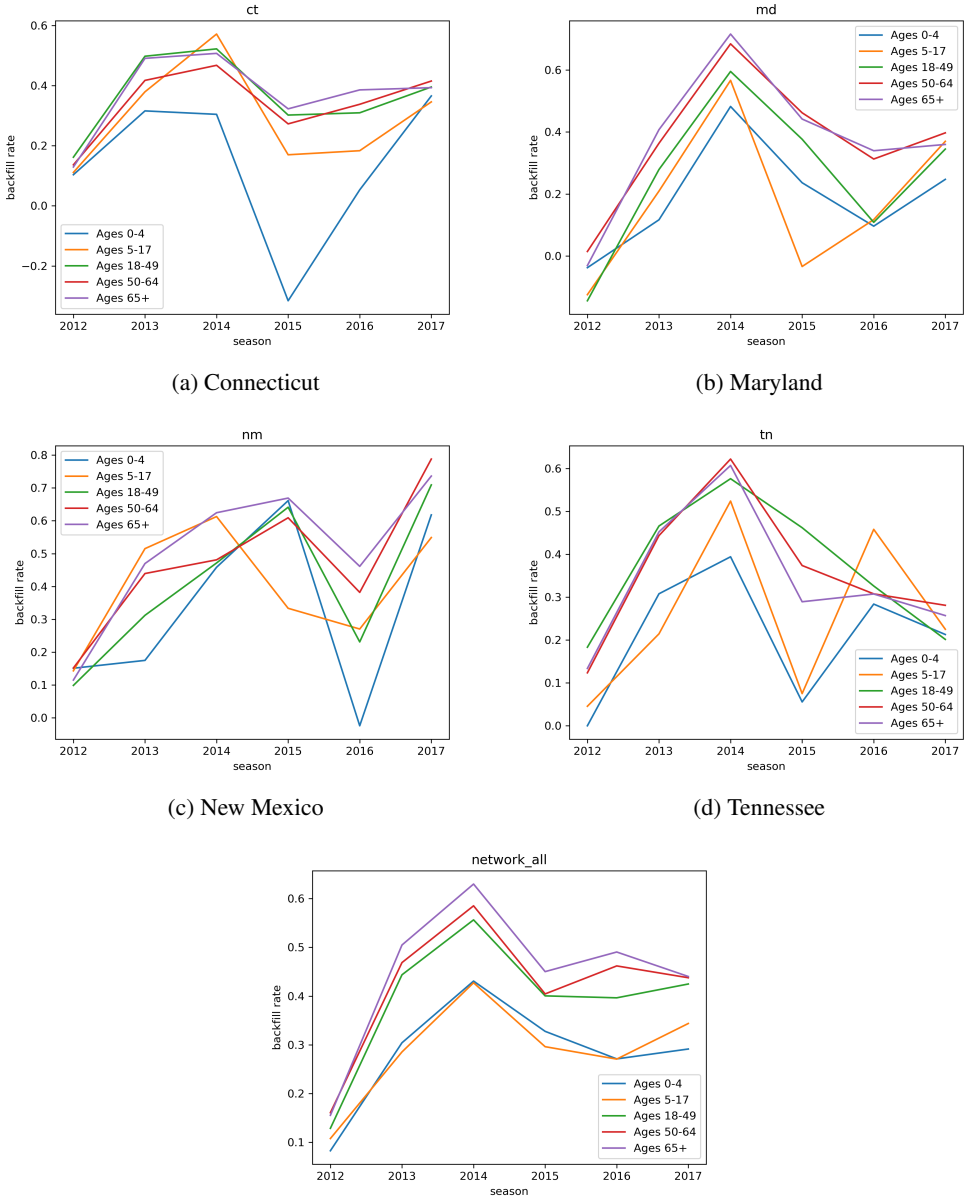
(e) Entire Network

Figure 2: Proportion of zero hospitalization rates for several representative states.

3.4 Suggestions for Machine Learning Methodology

Based on the problem formulation 1, missing data analysis, and backfill analysis, we propose several suggestions for the forecasting model:

1. The data size is small and complex models will overfit. Therefore, linear regression model and its variants are desirable choices.
2. We should use network-level data to train the machine learning model due to its high quality: most entries are accurate and the backfill pattern is relatively regular.
3. We should not use the hospitalization rates from season 2011-12 because there is no backfill behavior and the data is not representative.



(a) Connecticut

(b) Maryland

(c) New Mexico

(d) Tennessee

(e) Entire Network

Figure 3: Backfill rates for several representative states.

4 Machine Learning Methods

4.1 Point Estimation

Both mean regression and median regression are valid methods for point estimation. The only difference is the form of loss function. We denote the regression function as f, regularization function as r, and intercept as b. For f parametrized by θ trained on data $(X_1, Y_1), \cdots, (X_n, Y_n)$, mean regression minimizes quadratic loss:

$$\min_{\theta,b} \|f(\theta; X) + b\mathbf{1} - Y\|_2^2 + r(\theta) = \sum_{i=1}^{n} (f(\theta; X_i) + b - Y_i)^2 + r(\theta)$$

while median regression minimize absolute deviation:

$$\min_{\theta,b} \|f(\theta; X) + b\mathbf{1} - Y\|_1 = \sum_{i=1}^{n} |f(\theta; X_i) + b - Y_i| + r(\theta)$$

The following functional forms are considered for both mean and median regression:

1. **Linear Regression**: $f(\theta; X) = X\theta,\ r(\theta) = 0$.
2. **Ridge Regression**: for specified regularization hyperparameter λ, $f(\theta; X) = X\theta,\ r(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$.
   Remark. Ridge regression with absolute deviation loss is almost identical to support vector regression with linear kernel [7].
3. **Gradient Boosting Tree**: suppose the tree ensemble contains base tree learners $T_1, T_2, \cdots, T_M$ and the corresponding learning rates are $\gamma_1, \gamma_2, \cdots, \gamma_M$. Then we have

   $$f(\theta; X) = S_M(\theta; X) = \sum_{k=1}^{M} \gamma_k T_k(\theta_k, X),\ r(\theta) = 0$$

   The training procedure is explained as following. The first base tree learner T_1 is a constant function that minimizes loss function. For each $m > 1$, denote $L(\cdot, \cdot)$ as the loss function and compute the ensemble S_m, loss functions L_m, and gradient D_m for each training instance:

   $$\begin{aligned} S_m(\theta, X) &= \sum_{k=1}^{m-1} \gamma_k T_k(\theta_k, X) \\ L_m(\theta, X_i) &= L(S_m(\theta, X_i), Y_i),\ i = 1,\ 2, \cdots, n \\ D_m(\theta, X_i) &= -\frac{\partial L(S_m(\theta, X_i), Y_i)}{\partial S_m},\ i = 1,\ 2, \cdots, n \end{aligned}$$

   The m^{th} base tree learner fits X_i to $D_m(\theta, X_i)$ and therefore approximates the gradient descent directions. Consequently, the step from ensemble S_m to the next ensemble S_{m+1} optimizes the loss function at each training instance.

   Remark. Explicit regularizations are not applied for gradient boosting tree because the hyperparameters can directly control its complexity.

4.2 Interval Estimation

There are two potential methods for estimating a confidence interval of hospitalization rate.

1. Bootstrap training residuals from the point estimation model and obtain an interval estimate.
2. Apply quantile regression to directly estimate the upper bound and lower bound of interval.

In practice, quantile regression produces more reliable interval estimations. Although bootstrapping is able to generate true confidence intervals, it assumes that the distribution of residuals is independent of the data. However, independence assumption rarely holds.

The formulation of quantile regression method is subtly different from those of point estimation methods. Suppose the model estimates quantile q, where $0 < q < 1$. Denote regression function by f and regularization function by r, then the regression model solves the following problem:

$$\min_{\theta,b} \sum_{i=1}^{n} \max\{(1-q)(f(\theta; X_i) + b - Y_i), q(Y_i - f(\theta; X_i) - b)\} + r(\theta)$$

The loss function here is weighted absolute deviation: when $f(X_i) + b \geq Y_i$, the deviations have weight $1 - q$; when $f(X_i) + b < Y_i$, the deviations have weight q. From problem formulation 2, $f_{1-\alpha/2}$ corresponds to large q and $f_{\alpha/2}$ corresponds to small q.

Remark. If $q = 0.5$, the problem is exactly median regression. For any instance x, non-crossing condition is required to guarantee the validity of confidence interval:

$$f_{\alpha/2}(x) + b_{\alpha/2} \leq f_{0.5}(x) + b_{0.5} \leq f_{1-\alpha/2}(x) + b_{1-\alpha/2}$$

Since hospitalization rates are always positive, $x \succeq 0$ always holds. Then when f is a linear model parametrized by θ, the condition above is simplified as:

$$\begin{aligned} &{\theta_{\alpha/2}}^T x + b_{\alpha/2} \leq {\theta_{0.5}}^T x + b_{0.5} \leq {\theta_{1-\alpha/2}}^T x + b_{1-\alpha/2},\ \forall x \succeq 0 \\ \Longrightarrow\ &\theta_{\alpha/2} \preceq \theta_{0.5} \preceq \theta_{1-\alpha/2},\ b_{\alpha/2} \leq b_{0.5} \leq b_{1-\alpha/2} \end{aligned} \tag{3}$$

The three quantile regression models are jointly trained under the constraints 3.

4.3 Feature Selection

According to problem formulation 1, given a time window w for epiweek e and group g, the accessible hospitalization rates are $f(r_{\mathrm{e-w;0}}, \cdots, r_{\mathrm{(e-w,g);w}}, r_{\mathrm{(e-w+1,g);0}}, \cdots, r_{\mathrm{(e,g);0}})$.

The features can be organized in an alternative way. The earliest rate is $r_{\mathrm{(e-w+1,g);0}}$ from w epiweeks ago; the second earliest rates are $r_{\mathrm{(e-w+1,g);1}}, r_{\mathrm{(e-w+2,g);0}}$ from $w-1$ epiweeks ago ... the latest rates are $r_{\mathrm{(e-w+1,g);w-1}}, \cdots, r_{\mathrm{(e,g);0}}$ available at current epiweek. Earlier hospitalization rates should be removed in feature selection because their information is reflected in updated hospitalization rates of the same epiweeks.

The feature selection scheme is parameterized by t such that $0 \leq t \leq w$ and all selected features are available after time $e - t$. Therefore, there are w possible schemes for epiweek window w. A scheme with time window t will select following features:

$$r_{\mathrm{(e-w+1,g);w-t-1}}, r_{\mathrm{(e-w+1,g),w-t}}, \cdots, r_{\mathrm{(e-w+1,g);w-1}}, \cdots, r_{\mathrm{(e,g);0}}$$

Figure 4 is a pictorial explanation of the scheme for a particular epiweek e and age group g.

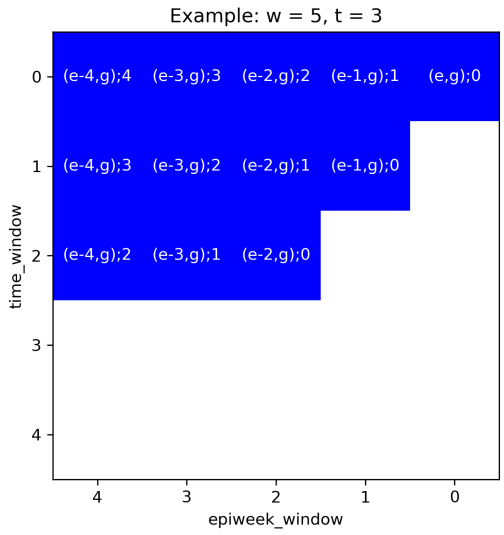


Figure 4: An example of feature selection scheme. The blue area represents selected features. The texts are the corresponding subscripts.

5 Experimental Results

5.1 Schemes of Experiments

Different point estimation and interval estimation methods are experimented under following schemes.

Data Selection. The hospitalization rates of the entire network and all 5 age groups from seasons 2013-14 to 2017-18 are used for cross-validation. For each season, training hospitalization rates are selected from previous seasons. Season 2012-13 are not included because there is no previous season with backfill behavior.

For each epiweek window w from 0 to 4, all feature selection schemes are experimented. To ensure the completeness of epiweek windows, only hospitalization rates from epiweek 44 to epiweek 17 of the next year are cross-validated. Training data from previous seasons are also selected from the same epiweek range for the same reason.

Hyperparameter Selection. For ridge regression methods, the regularization hyperparameter is $\lambda = 0.5$. For Gradient boosting tree methods, the number of estimators is set to 50 and the maximum depth for each tree is set to 3. Other parameters follow the default settings of scikit-learn gradient boosting regressor [8, 9].

Training. Two training schemes are experimented:

1. The default scheme: For each cross-validated season and an age group, use all previous hospitalization rates from season 2012-13 within the same age group to train machine learning model. If the rates across different seasons are independent and identically distributed, the scheme is optimal because it collects the most training data.
2. The one-year window scheme: For each cross-validated season and an age group, only use hospitalization rates from the previous season within the same age group to train machine learning model. The scheme can outperform the default scheme if the distribution of data depends on time.

Confidence Interval Regularization. For quantile regression models, a regularization can be performed by specifying a parameter $\epsilon > 0$ to tighten the non-crossing condition 3:

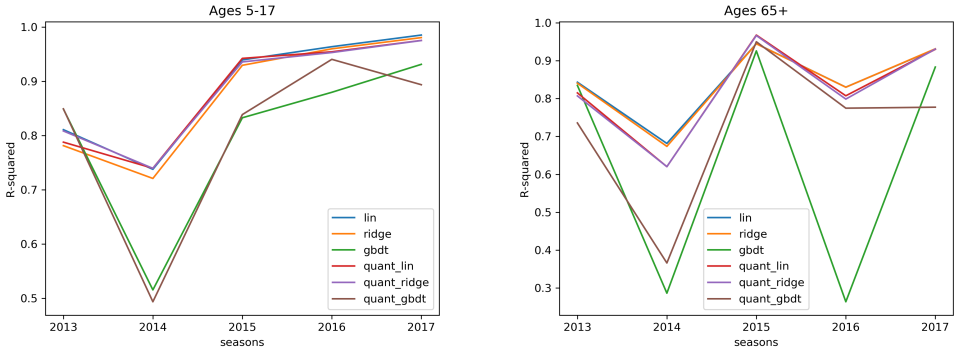
$$\theta_{\alpha/2} + \epsilon\mathbf{1} \preceq \theta_{0.5},\ \theta_{0.5} + \epsilon\mathbf{1} \preceq \theta_{1-\alpha/2}$$
$$b_{\alpha/2} + \epsilon \leq b_{0.5},\ b_{0.5} + \epsilon \leq b_{1-\alpha/2}$$

Remark. The constraints can regularize the quantile models because they indicate that the width of confidence interval is at least 2ϵ. In practice, they lead to more reliable interval estimations.

Metric. The metric used to measure predictive ability is R-squared (R^2). For a machine learning model, different combinations of epiweek window w and time window t are considered; the largest R^2 out of all combinations evaluates model performance.

5.2 Results Before Confidence Interval Regularization

Default Scheme. Different methods for point estimation are compared under the default training scheme. The results are summarized in figure 5.



(a) 5-17 (b) 65+

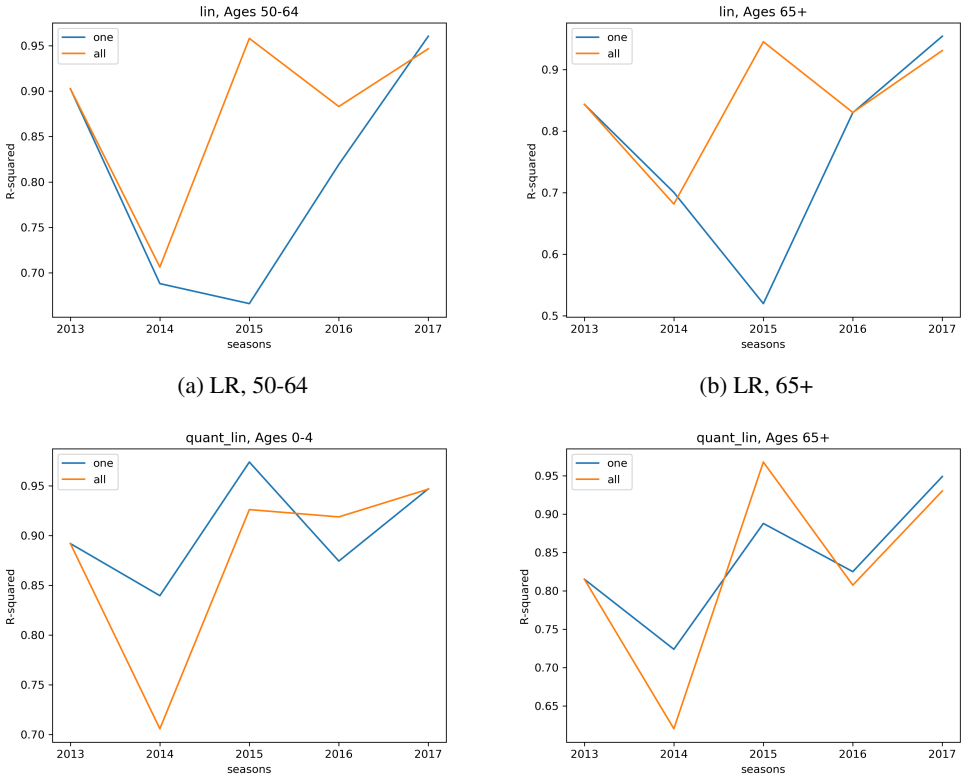
Figure 5: Representative results under default training scheme.

The results indicate that gradient boosting tree-based methods are clearly inferior: they perform significantly worse that linear or ridge regression on age groups with relatively large hospitalization rates. Therefore, only linear and ridge regression models are considered in following experiments.

One-year Window Scheme. The one-year window cross validation scheme is compared to the default scheme and representative results are summarized in figure 6.

For mean regression models, one-year window scheme is not preferable to the default scheme. When predicting age groups with high hospitalization rates, one-year window scheme performs considerably worse for both linear and ridge regression models. The non-robustness of mean regression can explain the observations.

However, for median regression models, one-year window scheme outperforms the default scheme. In particular, the out-of-sample R^2 increases by more than 5% for age group 0-4 years old, 5-17 years old, and 65+ years old. Although R^2 drops by around 5% for age group 50-64 years old, it is still above 80%. The robustness of median regression leads to the improvements above because robust models do not tend to overfit on small amount of data.



(a) LR, 50-64

(b) LR, 65+

(c) QLR, 0-4

(d) QLR, 65+

Figure 6: Representative comparisons between two training schemes. LR - mean linear regression; RR - mean ridge regression; QLR - medium linear regression; QRR - medium ridge regression.

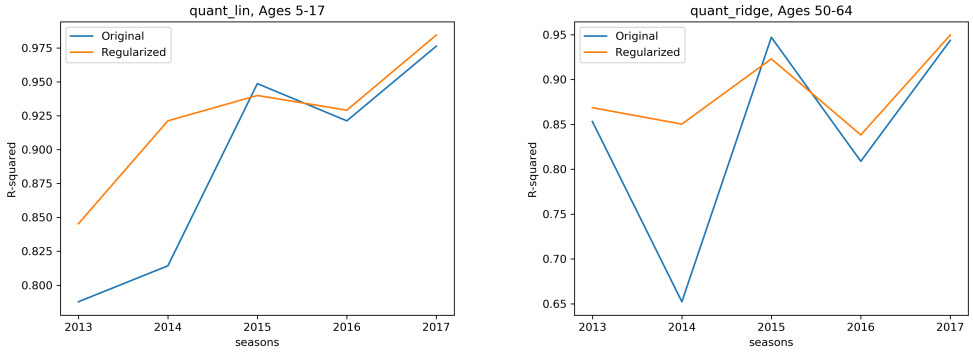
5.3 Results After Confidence Interval Regularization

Confidence interval regularization with $\epsilon = 0.05$ is applied to two median regression models with both training schemes. The effects under both training schemes are shown in figure 7 and 8.

For both linear and ridge regression models, the trick increases predictive R^2 by more than 10% for different age groups in season 2014-15. Meanwhile, there is no significant decrease in predictive R^2 for other age groups and seasons.

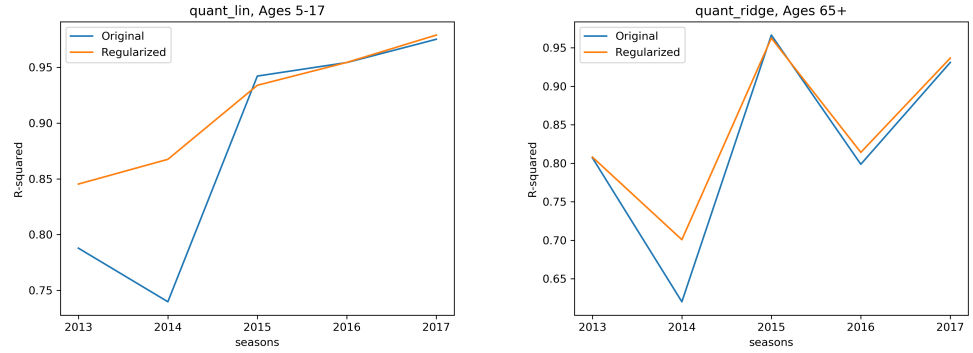
In addition, default training scheme is preferable for quantile linear model due to insufficient regularization (no ridge penalty). Meanwhile, one-year scheme is still preferable for quantile ridge model. Representative results are shown in figure 9.

As previous discussion about confidence interval regularization indicates, the trick indeed leads to more sensible lower and upper bounds, as shown in figure 10.



(a) QLR, 5-17 (b) QRR, 50-64

Figure 7: Representative effects of confidence interval regularization under one-year training scheme.

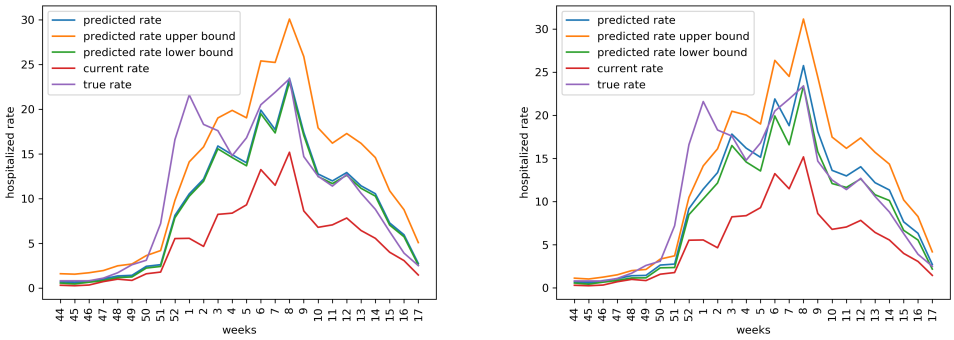


(a) QLR, 5-17 (b) QRR, 65+

Figure 8: Representative effects of confidence interval regularization under default training scheme.

(a) QLR, 5-17 (b) QRR, 50-64

Figure 9: Representative comparisons of training schemes after confidence interval regularization.



(a) Before regularization.

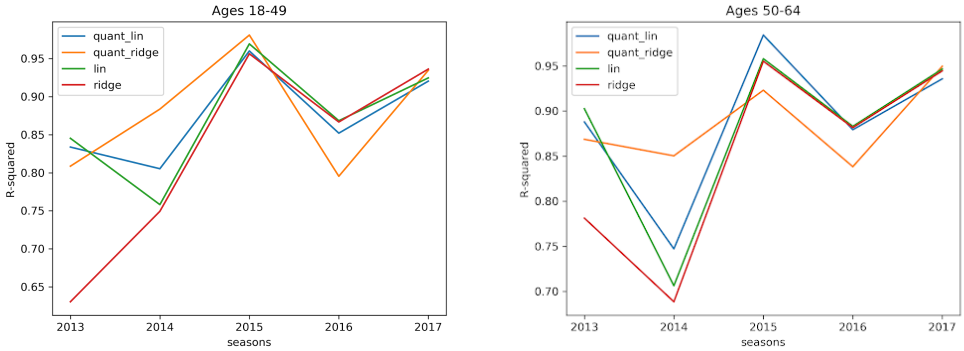
(b) After regularization.

Figure 10: Comparison before and after confidence interval regularization for hospitalization rate prediction of season 2016-17, age group 65+ with $w = 1, t = 1$.

5.4 Summary of Results

The experiments above are able to determine the optimal combinations machine learning models, cross validation scheme, and regularization techniques. The results of optimal combinations for different age groups are displayed in figure 11 with gradient boosting tree models excluded.

The results demonstrate that regularized quantile regression methods are superior to mean regression methods with stronger robustness. In particular, the predictive R^2 of quantile ridge regression exceeds 80% for every season and age group.



(a) 18-49

(b) 50-64

Figure 11: Representative comparisons of machine learning models under optimal combinations.

6 Conclusion

The project introduces sufficient background and performs preliminary analysis to develop a better understanding of hospitalization rate data from CDC. Based on the background and analysis, a hospitalization rate prediction problem is formulated. Several regression models, cross validation schemes, and regularization techniques are proposed to solve the problem. After a series of experiments, quantile regression models (with confidence interval regularization) are considered as the optimal machine learning model for the problem due to their robustness and ability to generate practical confidence intervals.

References

[1] Sandra S Chaves, Ruth Lynfield, Mary Lou Lindegren, Joseph Bresee, and Lyn Finelli. The us influenza hospitalization surveillance network. *Emerging infectious diseases*, 21(9):1543, 2015.

[2] Centers for Disease Control and Prevention. Cdc says "take 3" actions to fight the flu. https://www.cdc.gov/flu/protect/preventing.htm.

[3] Centers for Disease Control and Prevention. Disease burden of influenza. https://www.cdc.gov/flu/about/disease/burden.htm#flu-related-illness.

[4] Centers for Disease Control and Prevention. Laboratory-confirmed influenza hospitalizations. https://gis.cdc.gov/GRASP/Fluview/FluHospRates.html.

[5] Centers for Disease Control and Prevention. Fluview: influenza hospitalization surveillance network, 2013.

[6] Lisheng Gao. Ongoing influenza activity inference with real-time digital surveillance data. https://www.ml.cmu.edu/research/dap-papers/F17/dap-gao-lisheng.pdf.

[7] Changha Hwang and Jooyong Shim. A simple quantile regression via support vector machine. In *International Conference on Natural Computation*, pages 512–520. Springer, 2005.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] Gilles L. Emanuele O. Arnaud J. Jacob S. Peter P., Scott W. Gradient boosting for regression. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.