

# **Biostatistics Methods Final Project Report**

Manqi Cai, Yixi Xu, Peng-Lin Lin, Kaitlin Maciejewski

## **Abstract**

In order to improve the hospital management and lower the expenditure on patients' care, a data analyst group in Good Health Corporation wants to find out if there is some way to predict patients' length of stay (LOS) in the hospital. Based on this, simple linear regression and multiple linear regression models were fit on a dataset collected patients' information from the year 2016. The predictors included in the model are Is\_30\_Day\_Readmit, Age, Cindex, Insurance\_Type, BP\_Systolic, Heart\_Rate, Respiration\_Rate, and the model has a coefficient of determination (R-squared) of 0.13.

Keywords: Hospital, length of stay, multiple linear regression, model diagnostic, model validation

## **Introduction**

A precise hospital resources allocation is crucial for the effectiveness of hospital management. Several methods have been developed to find better ways to allocate health-care resources.<sup>1-3</sup> One important component in these methods is an accurate prediction of patients' length of stay (LOS) in hospital. Previous research has found that hospital LOS is an indicator of quality of care in hospital, patient outcomes, and cost.<sup>9</sup> Longer LOS could be due to more severe medical impairment on admission, and is also associated with an increased risk of a hospital-acquired condition and additional illness.<sup>11</sup> Longer LOS also increases cost and financial burden.

Previous research has identified many variables associated with the LOS, including standardized or modified early warning scoring system (SEWS/ MEWS),<sup>4</sup> age, gender, emergency surgery, intensive care unit stay,<sup>5</sup> Charlson comorbidity index (Cindex),<sup>6,7</sup> and religion.<sup>8</sup> It has been found that LOS

increases when patient age increases, LOS is longer for females, unmarried patients, black patients, and those with Medicare insurance.<sup>9,10</sup>

This project aims to build a linear regression model to predict the patients' length of stay for Good Health Corporation. The dataset contains 3682 medical records from 3612 patients in 2016.

## Methods

The dataset for this study is composed of 3682 cases of medical records from 3612 patients from various health facilities in 2016. Statistical descriptions of the continuous variables are shown in Table 1.

In order to preserve independence of observations, only the first visit for each patient was included; this excluded 70 duplicated visits. Postal code and facility zip were excluded since they provide no helpful information for the analysis. Previous research shows that there was little or no association between race and LOS, so race was removed as a possible predictor.

Table 1 shows that there is some weird data in the dataset. BMI ranges from 3.10 to 122.65, temperature ranges from 11.85 C to 52.27 C, which are outside the normal ranges for these measures. There are also 697, or 19%, missing records for BMI. Since the amount of missingness can bias the results, the BMI variable was removed.

Categorical data were recode for better model building process. Religion was recoded into two groups: religious (1) or not (0). Gender was recode as male (1) and female (0). Marital status was also recode as: married (1) and unmarried (0). Charlson comorbidity index (Cindex) was reeveled into three level: normal (0), moderate (1), severe (2). Insurance was recoded into four groups: none (0), private (1), medicare (2), and medicaid (3). Complete date of admission (Admitdtm) was regrouped to month of admit from January to December. Modified Early Warning Score (MEWS) were redefined into 4 classes: normal (1), increase caution (2), moderate (3) and severe (4). Statistical descriptions of the categorical variables are shown in Table 2.

The primary outcome of the analysis was the natural logarithm of LOS. For the continuous variables, the maximum, minimum, mean, median and interquartile ranges (IQR) were used to measure the central tendency and spread of the data.

The model selection was performed in SAS. Forward, backward, and stepwise procedures all used a threshold p-value of 0.15 to select the significant predictors for the model. All procedures gave the same model, so this was the model used going forward.

For model diagnostics, residuals v.s. fitted values plot was used to detect heteroscedasticity and outliers, quantile-quantile plot (qqplot) was used to examine the normality, scale-location was used to check assumption of equal variance, residual v.s. leverage plot was used to identify the influential cases. The first model (Figure 1) produced a qq-plot with a heavy tail, showing that some residuals were not normal (Figure 2).

The studentized residuals were then used to identify potential outliers, using a threshold of 2.5. After removing these outliers, the re-fitted data held the normality assumption. However, o2saturation was now very significantly insignificant, so it was also removed in order to produce the final model (Figures 3, 4)

In terms of model validation, bootstrap method with 1000 repeats, and cross-validation with 10 folds were performed to assess the variability of coefficient estimates and model prediction.

## Results

The final dataset contains 2666 observations, with 1218 male patients and 1448 female patients. 1226 are married, 1440 single, 51 patients have been to ICU v.s. 2615 have not.

The final model is :

$$\log(\text{LoS}(days)) = 0.27X_1 + 0.01X_2 + 0.13X_3 + 0.10X_4 - 0.006X_5 + 0.005X_6 + 0.02X_7$$

where  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$ ,  $X_6$  and  $X_7$  denote Is\_30\_Day\_Readmit, Age, Cindex, Insurance\_Type, BP\_Systolic, Heart\_Rate, Respiration\_Rate, respectively. The R-squared is 0.1336 and the adjusted

R-squared is 0.1313, which indicates that 13% of the model fit is explained by the data . All the coefficients are significant (Figure 4.). Table 3 compares the simple linear regression for the predictors and the multiple linear regression.

In model validation, The mean square error (MSE) produced from Bootstrap method and the cross-validation are 0.7431896 and 0.7399242, respectively. The MSE for our final model is 0.7459, the difference between validation and final model are 0.4% and 0.8%, which much less than the threshold 10%. Therefore, after model validation, we think our model is reasonable to predict LOS.

### **Discussion and Conclusion:**

The linear model is significant and has a good performance on the validation. The R-squared show that the model can explain about 13% of the association between predictors and hospital length of stay. Compared to the previous studies, the R-squared are typically between 0.10 to 0.27,<sup>12</sup> our linear regression model has an acceptable prediction performance.

Our final model does not include gender or marriage status, though previous research showed there was an association between these variables and length of stay. A sample interpretation our model is that for each 1-year increase in age, we expect LOS to increase by 1 day. This does agree with previous research that claims increased age is associated with increased LOS.

To further improve our model, we should consider methods other than the linear model. The correlation matrix (Figure 5) does not show a significant linear trend, so this may not be the best model for the data. Plenty of models have been developed to fit the hospital length of stay. Under the general condition, one of the best models is using Bayesian model<sup>1</sup>, but the variables should be independent of each other. Under some other specific conditions, generalized linear model with normal, Poisson, negative binomial, gamma, lognormal and phase-type distributions are better.<sup>12-14</sup> The model using phase-type probability distribution can manipulate a high-skewed dataset,<sup>14</sup> which is the case of our data. We could consider it as a potential model to fit the length of stay of our dataset.

In conclusion, our model is significant and valid, with a moderate goodness-of-fit. It should be a good reference for the Good Health Cooperation to improve the hospital management. However, further work is needed to improve the performance of the model.

**Reference:**

1. Gustafson, D. H. Length of stay: prediction and explanation. *Health Serv Res* **3**, 12-34 (1968).
2. Harper, P. R. A framework for operational modelling of hospital resources. *Health Care Manag Sci* **5**, 165-173 (2002).
3. Vissers, J. M. Patient flow based allocation of hospital resources. *IMA J Math Appl Med Biol* **12**, 259-274 (1995).
4. Paterson, R. *et al.* Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. *Clin Med (Lond)* **6**, 281-284 (2006).
5. Higgins, T. L. *et al.* Early indicators of prolonged intensive care unit stay: impact of illness severity, physician staffing, and pre-intensive care unit length of stay. *Crit Care Med* **31**, 45-51, doi:10.1097/01.CCM.0000038743.29876.3C (2003).
6. Librero, J., Peiro, S. & Ordinana, R. Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days. *J Clin Epidemiol* **52**, 171-179 (1999).
7. de Groot, V., Beckerman, H., Lankhorst, G. J. & Bouter, L. M. How to measure comorbidity. a critical review of available methods. *J Clin Epidemiol* **56**, 221-229 (2003).
8. Koenig, H. G. & Larson, D. B. Use of hospital services, religious attendance, and religious affiliation. *South Med J* **91**, 925-932 (1998).
9. Shulan, Mollie and Gao, Kelly. Revisiting Hospital Length of Stay: What Matters?. *The American Journal of Managed Care* Vol 21, No 1, 71-77 (2015).
10. Weiss, Audrey J and Elixhauser, Anne. Overview of Hospital Stays in the United States, 2012. Agency for Healthcare Research and Quality Statistical Brief #180 (2014).
11. Patient-Centered LOS Reduction Initiative Improves Outcomes, Saves Costs. *Health Catalyst* (2016).
12. Verburg, I. W., de Keizer, N. F., de Jonge, E. & Peek, N. Comparison of regression methods for modeling intensive care length of stay. *PLoS One* **9**, e109684, doi:10.1371/journal.pone.0109684 (2014).
13. Peter. C. Austin, Deanna M. Rothwel, Jack V. Tu. A Comparison of Statistical Modeling Strategies for Analyzing Length of Stay after CABG Surgery. *Health Services and Outcomes Research Methodology* 3: 107. Doi: <https://doi.org/10.1023/A:1024260023851> (2002).
14. Faddy, M., Graves, N. & Pettitt, A. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value Health* **12**, 309-314, doi:10.1111/j.1524-4733.2008.00421.x (2009).

**Appendix:**

**Table 1:** Summary of Numerical Patient Data

	Min	Max	Mean	Median	Sd	Missing
Length of Stay, hours	1	2111	131.8	92.0	142.35	
Length of Stay, days	0.04	87.96	5.49	3.83	5.93	
Age	18.00	105	65.74	68.00	18.66	
BMI	3.10	122.65	28.33	27.10	7.96	697
BP Systolic mmHg	88.78	193.96	130.52	129.17	9.83	5
O2 Saturation %	80	236.53	97.85	97.58	4.86	3
Temperature °C	11.85	52.27	36.73	36.73	0.91	3
Heart Rate bpm	37.58	242.58	80.09	79.20	12.97	5
Respiration Rate bpm	12	67.72	18.20	17.76	2.65	3
BP Diastolic mmHg	29.56	154.4	72.53	71.85	16.78	1

**Table 2:** Summary of Categorical Data

ICU flag	count	Religion	count	MEWS	count	Marital status	count
0	3618	Angelican	1	0	39	Civil Union	1
1	64	Catholic	1678	1	618	Married	241
		Christian	909	2	730	Separated	1637
		Hebrew	1	3	801	Single	967
		Hindu	127	4	711	Widowed	704
		Islam	112	5	328	NA	80
		Jewish	527	>5	289		
		Mormon	2	NA	166	Evisit	count
		No Affiliation	177	0	1134	0	
		Non Denominational	1	1	752	1	
		Other	147	2	521	2	
				3	329	3	
				4	946	4	

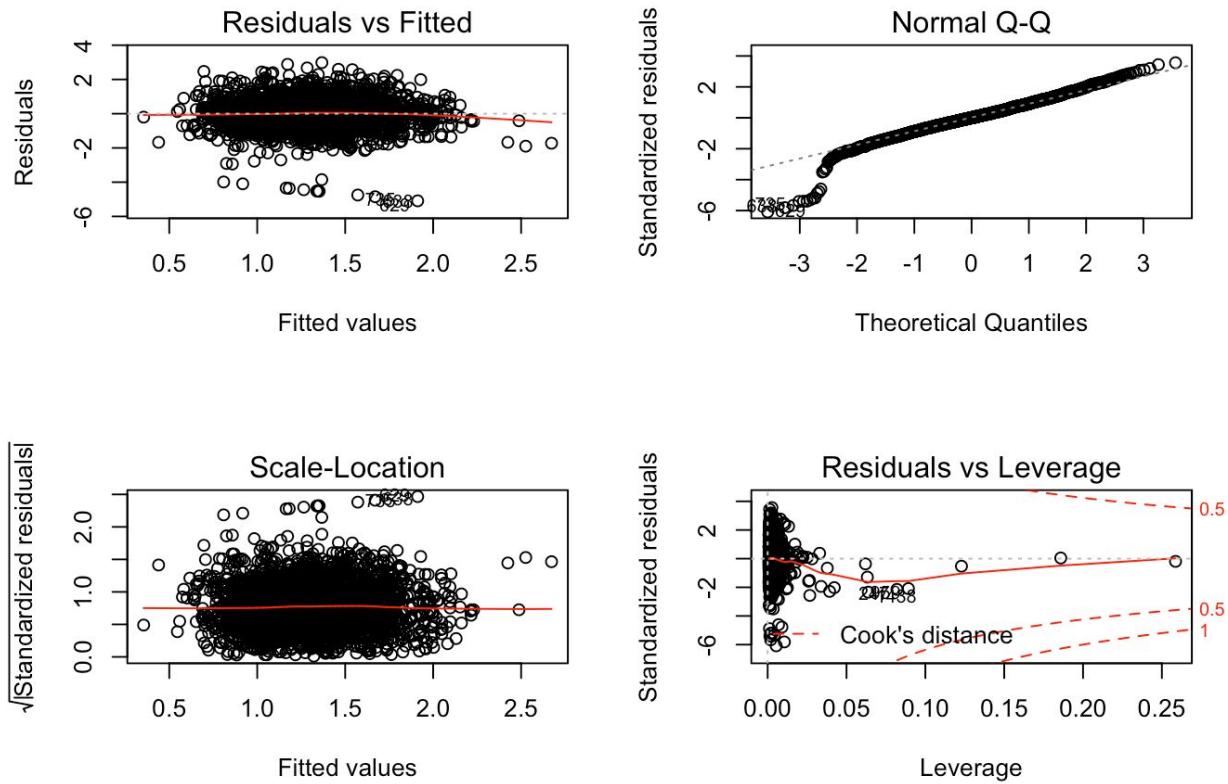
**Table 3:** Summary of continuous variable between SLR vs MLR

Variables	SLR					MLR				
	Estimate	Se	Pr(> t )	2.5%	97.5%	Estimate	Se	Pr(> t )	2.5%	97.5%
Is_30_Day_Readmit	1.832	0.265	0	1.312	2.352	0.973	0.265	0	0.453	1.493
Cindex	1.156	0.12	0	0.921	1.392	0.656	0.122	0	0.417	0.895
Evisit	0.473	0.058	0	0.359	0.587	0.291	0.059	0	0.175	0.407
Age	0.044	0.005	0	0.035	0.054	0.037	0.006	0	0.026	0.048
Gender	0.185	0.186	0.321	-0.181	0.551	0.522	0.182	0.004	0.164	0.879
Marital_Status	-0.397	0.186	0.033	-0.762	-0.032	-0.282	0.181	0.12	-0.638	0.073
Insurance_Type	0.974	0.157	0	0.666	1.281	0.363	0.157	0.021	0.055	0.672
BP_Systolic_mmHg	-0.022	0.006	0	-0.033	-0.011	-0.018	0.006	0.004	-0.031	-0.006
Heart_Rate_bpm	0.039	0.007	0	0.025	0.052	0.038	0.007	0	0.025	0.052
Respiration_Rate_bpm	0.223	0.034	0	0.157	0.289	0.161	0.032	0	0.098	0.225
BP_Diastolic_mmHg	-0.064	0.009	0	-0.083	-0.046	-0.038	0.011	0.001	-0.059	-0.016

**Figure 1: Regression 1**

```
##  
## Call:  
## lm(formula = losdays2_log ~ is30dayreadmit + ageyear + cindex +  
##     insurancetype + bpsystolic + o2sat + heartrate + respirationrate,  
##     data = ghproject_tidy)  
##  
## Residuals:  
##      Min    1Q   Median    3Q   Max  
## -5.0899 -0.4820  0.0048  0.5188  2.9834  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.7674030  0.3663017  2.095 0.036264 *  
## is30dayreadmit 0.2597481  0.0468562  5.544 3.25e-08 ***  
## ageyear       0.0092775  0.0009619  9.645 < 2e-16 ***  
## cindex        0.1364620  0.0221881  6.150 8.88e-10 ***  
## insurancetype 0.1130282  0.0284840  3.968 7.43e-05 ***  
## bpsystolic    -0.0057149  0.0010042 -5.691 1.40e-08 ***  
## o2sat         -0.0044032  0.0030484 -1.444 0.148733  
## heartrate      0.0058341  0.0012381  4.712 2.58e-06 ***  
## respirationrate 0.0195876  0.0059103  3.314 0.000931 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.8388 on 2700 degrees of freedom  
## Multiple R-squared:  0.1109, Adjusted R-squared:  0.1083  
## F-statistic: 42.09 on 8 and 2700 DF, p-value: < 2.2e-16
```

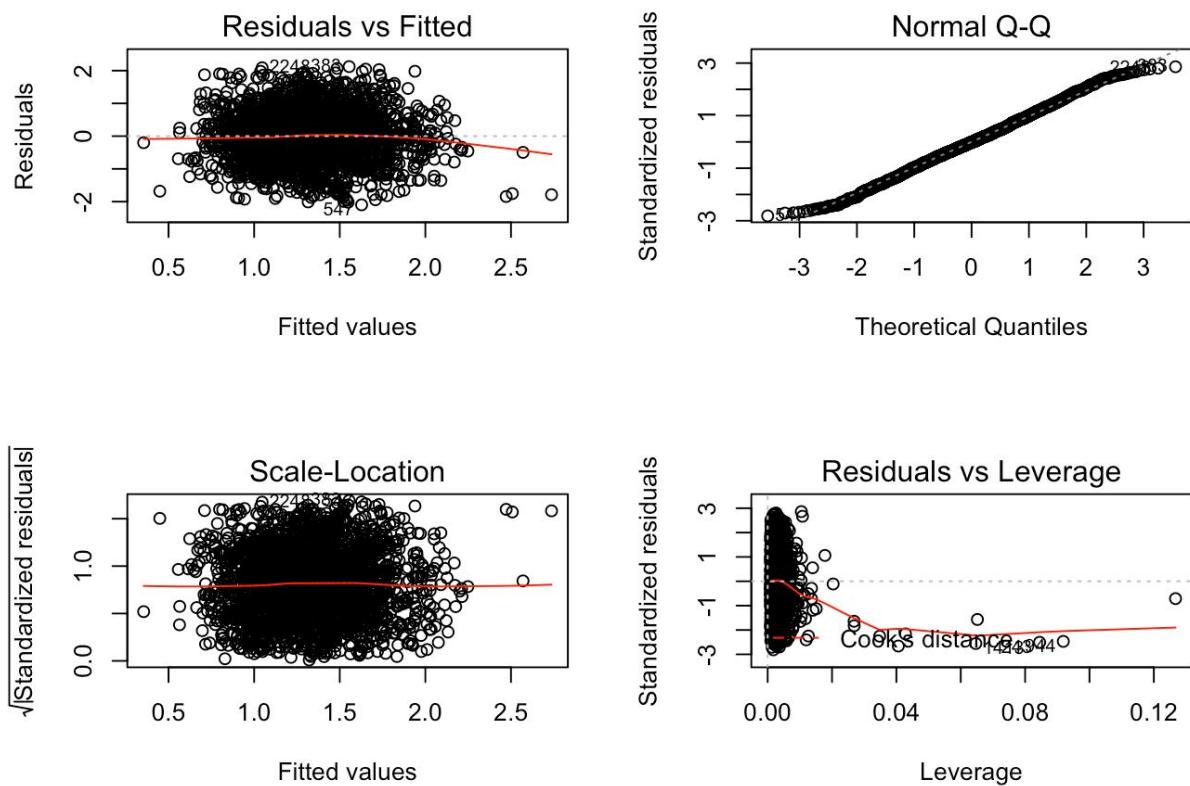
**Figure 2: Diagnostic plot for model 1**



**Figure 3: Final model**

```
##  
## Call:  
## lm(formula = losdays2_log ~ is30dayreadmit + ageyear + cindex +  
##       insurancetype + bpsystolic + heartrate + respirationrate,  
##       data = data_no_outliers)  
##  
## Residuals:  
##      Min        1Q     Median        3Q       Max  
## -2.09763 -0.48678 -0.01201  0.49003  2.12255  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.3862749  0.1803208  2.142 0.032272 *  
## is30dayreadmit 0.2665833  0.0419770  6.351 2.51e-10 ***  
## ageyear      0.0097364  0.0008654 11.250 < 2e-16 ***  
## cindex       0.1276104  0.0198784  6.420 1.61e-10 ***  
## insurancetype 0.0964752  0.0256678  3.759 0.000175 ***  
## bpsystolic    -0.0059440  0.0009012 -6.596 5.08e-11 ***  
## heartrate      0.0053471  0.0011077  4.827 1.46e-06 ***  
## respirationrate 0.0213752  0.0053345  4.007 6.32e-05 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.7459 on 2658 degrees of freedom  
## Multiple R-squared:  0.1336, Adjusted R-squared:  0.1313  
## F-statistic: 58.55 on 7 and 2658 DF,  p-value: < 2.2e-16
```

**Figure 4: Diagnostic plot for final model**



**Figure 5: Correlation Matrix for Continuous Data**

