P8130: Biostatistical Methods I

Final Project

Group 18: Angel Garcia de la Garza, Jiaqi Li, Siling Li, Shichen Zhou

Abstract

The primary objective of this analysis is to identify variables associated with length of stay (LoS) and construct a predictive model. Length of stay is associated with increased patient care costs and depletion of hospital resources. We created a predictive model using multiple linear regression and p-value and criterion selection to decrease parameter dimensionality. Our predictive model has an R^2 = 0.1376. We found that several variables are linearly associated with length of stay, including number of visits to the emergency room before admission, age, type of insurance, and severity of comorbidity of a patient. Finally, we validated the predictive ability of this model using cross-validation.

Introduction

Length of stay (LoS), measured as the duration of time that a patient remains in an inpatient care facility, is associated with decreased costs and better patient outcomes. U.S. hospital stays cost the healthcare system more than 377.5 billion dollars per year.[1] Reducing the LoS could yield large cost savings associated with patients' care as the average cost of a hospital stay is $10,400 per day.[2,3] In this study, the Data Analytics group from Good Health Corporation is interested in improving hospital management and minimizing the cost and resource expenditure of hospitalization. The goal of this analysis is to identify variables associated with increased LoS and to build a predictive model for future use.

Data Description

The dataset provides information on 3682 records from 3612 patients in 2016 collected by Data Analytics group from Good Health Corporation. Only visits within 24 hours of hospital admission and for patients older than 17 were considered relevant for the analysis. The main outcome is length

of stay (LOSDays). The predictors in this data are: whether the patient have been admitted to the hospital within the past 30 days (is30dayreadmit); the Modified Early Warning Score (MEWS); Charlson comorbidity index (cindex); number of times the patient visited an emergency department in the six months prior to admission (evisit); whether the patient had a visit in the ICU (ICU_Flag); patients' age, gender, marital status, race, and religion; facility name; insurance type; and vital signs like BPS, temperature, heart rate, O2 saturation, respiration rate, BPD, and BMI. Table 1 shows descriptive statistics of each of the continuous measures and its association to the outcome as measured through simple linear regression (SLR). Table 2 shows the frequency of each of the groups in our categorical variables, along with the estimated effects for each category through an SLR along with the one-way ANOVA.

Data Cleaning

From the summary statistics and our histogram and scatter plots (Figure 1), we observe some problematic issues with the data. For example, some values of vital statistics are extreme: there are temperatures as high as 52°C and as low as 11°C. This doesn't make clinical sense[4], so we have removed temperature observations greater than 46°C and lower than 21°C. Furthermore, we removed observations with an oxygen saturation percentage greater than 100%. There are around 600 observations missing for Body Mass Index, and there are additionally observations that seem to be errors (BMI of less than 10). BMI is marginally correlated with the outcome, (ρ = -0.036, p-value = 0.0333). As such, we have decided to exclude this variable from the analysis since we believe missing so many observations could introduce bias, especially if they are not missing at random.

We combined some categories together since they have a small number of observations. We incorporated the groups "Anglican", "Non-Denominational", and "Mormon" into "Christian," and "Hebrew" into "Jewish". Furthermore, we grouped race into three groups, "White", "African American" and "Other" since only the first two had enough observations. MEWS determines the degree of illness of a patient based on vital signs like respiratory rate, oxygen saturation, and blood pressure, etc., indicating that fitting in both MEWS and vital signs will induce multicollinearity to the model. As a result, we chose to exclude MEWS from the analysis.

We further assumed that people in the ICU tend to stay longer in the hospital; we believe these 60 subjects differ from the rest of the population so we have excluded them and focused our analysis on only those subjects that were not in the ICU. We also explored observations by their month. We see that the observations from December were from the 29th and 30th and seem to be censored. This could introduce bias in our inference so we have further excluded these observations.
From the distribution of LoS, we observe that the distribution is skewed. We believe that this can cause our model assumptions to be violated if the residuals are not normally distributed. We chose the natural logarithm of LoS as the best transformation, which we will use as our outcome. Our final sample includes 3168 observations.

Statistical Methods

We created a predictive model using multiple linear regression. We trained the model using backwards p-value based selection procedure (at $\alpha = 0.15$) in SAS and the criterion-based procedure to find the best predictive subset of variables. We proceeded to evaluate this model by identifying potential influential observations, and refitted the model without them. We also checked for collinearity among our selected predictive variables using the variance inflation factor method. Moreover, we verified that the model assumptions hold using the variance inflation factors method and using diagnostic plots. Finally, we estimated a cross-validated mean square error and a bootstrapped mean square error to validate the predictive ability of our model.

Results

We first performed p-value based backward selection in SAS using all our covariates. This first training step excluded race and religion as predictors. We proceeded to use criterion based selection in R. We checked this across several criterions. We calculated AIC, BIC, RSS, RSQ, Adj-RSQ, and Cp. From Table 3, we observe that all the criterions converge and select the model with all the predictors with the exception of BIC, which indicates that a model with 8 variables is the best. Given the high concordance of all the criterions, we proceed with the full model. The multiple regression model with

all 14 predictors produced an $R^2 = 0.1376$, and Adj-$R^2 = 0.1312$. Unfortunately, this indicates a low predictive ability.

From Table 4, the number of times the patient visited an emergency department in the six months prior to admission, age, temperature, heart rate, and respiration rate of a patient are positively and significantly linearly associated with the length of stay, indicating that those with higher scores on these variables tend to have longer length of stays. Vital signs including blood pressure systolic, blood pressure diastolic, and O2 saturation are significantly and negatively linearly associated with the outcome.

We also obtain useful information from our categorical variables. For example, patients with severe comorbidity were expected to stay longer in the hospital than those with mild comorbidity. Additionally, patients with private insurance tend to have shorter length of stays than patients with Medicaid. Furthermore, patients who were admitted into the hospital within the past 30 days were expected to stay in the hospital longer than those who were not.

While verifying model assumptions (Figure 2) we see from the residual versus fitted values that there is constant variance around 0 and residuals seem to be independent, despite some outliers. There is some deviation from normality in the residuals as the QQ-plot has a heavy tail. There are no clear outliers as flagged by Cook's distance. We proceeded to detect outliers using influence measures. However, more than 400 observations were marked as problematic, so we ran a sensitivity analysis without those problematic observations, and found that excluding them would help normality of residuals (Figure 3).

We assessed the variability of the coefficient estimates by calculating a cross-validated MSE along with the bootstrap. Both of these methods converge and validate our model. The crude MSE is estimated to be 0.6705 while the cross-validated one is 0.6619. Figure 4 shows the bootstrapped distribution of MSE, which indicates that our crude falls well within this distribution. These results indicate that our model is valid, does not overfit the data, and has a generalizable predictive ability.

Discussion

We have been able to generate a valid model with minor predictive ability. The variables in our sample do not seem to be good single predictors of LoS and using them together in a multiple linear regression does not seem to highly increase prediction ability. From the scatter plots, we observe that there is non-linear relationship between some of the vital statistics and length of stay. Therefore, we believe that using a model that accounts for non-linearity would increase predictive ability. Examples of those include polynomial regression model and generalized additive models. We also believe that we could model this data as a survival analysis question, since we could think of the length of stay as the duration of time until a subject leaves the hospital. Another path to increasing prediction ability is to include interactions in the model. We decided not to do that since it is hard to find interpretable interactions and overusing them might lead to overfitting in the model especially if you have a small amount of observations in your stratified groups. We could have used imputation in order to account for the missing variables but that can introduce bias in itself since you are modeling the missingness without knowing the true nature of the missing data. Finally, we expect that using a generalized linear model with a different distribution might increase the prediction ability of the model. FWe could think of length of stay as a type of "count" so a Poisson regression might be appropriate in that case.

References

1. Reducing length of stay in hospital improves outcomes. (n.d.). Retrieved from https://www.healthcatalyst.com/success_stories/reducing-length-of-stay-in-hospital
2. Jaffee, E. G., Arora, V. M., Matthiesen, M. I., Meltzer, D. O., & Press, V. G. (2017, Sep. 20). *Health Literacy and Hospital Length of Stay: An Inpatient Cohort Study*. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29236095
3. New York State Department of Health. *New York State Inpatient Hospital Cost Trends, 2009-2012*. Retrieved from https://www.health.ny.gov/statistics/sparcs/sb/docs/sb10.pdf
4. Marx, John. *Rosen's emergency medicine: concepts and clinical practice*. (2006) Mosby/Elsevier. p. 2239.

Appendix

Table 1. Descriptive Summary of Continuous Variables

Variable	n	mean	sd	median	min	max	beta1	p.value	CI lower	CI upper
loshours	3612	131.1442	142.1069	92.0000	1.0000	2111.0000	0.0417	0.0000	0.0417	0.0417
ageyear	3612	65.6869	18.6904	68.0000	18.0000	105.0000	0.0379	0.0000	0.0277	0.0482
bmi	2927	28.3493	7.9933	27.1000	3.1000	122.6500	-0.0283	0.0333	-0.0543	-0.0022
bpsystolic	3607	130.5527	16.7216	129.2222	88.7826	193.9630	-0.0264	0.0000	-0.0380	-0.0149
o2sat	3609	97.8609	4.9083	97.5854	80.0000	236.5263	-0.0397	0.0482	-0.0791	-0.0003
temperature	3610	36.7300	0.8995	36.7286	11.8500	52.2750	0.1762	0.1078	-0.0386	0.3911
heartrate	3607	80.0711	13.0041	79.2000	37.5833	242.5833	0.0479	0.0000	0.0331	0.0627
respirationrate	3609	18.1960	2.6335	17.7600	12.0000	67.7179	0.2197	0.0000	0.1467	0.2928
bpdiastolic	3611	72.5175	9.8018	71.8421	29.5635	154.4000	-0.0647	0.0000	-0.0844	-0.0451

Table 2. Descriptive Summary of Categorical Variables

Variable		**frequency**	**beta1**	**p.value**	**CI lower**	**CI upper**	**anova.pval**
is30dayreadmit	0 = no admission into the hospital within past 30 days	518	-	-	-	-	7.28E-10
	1=otherwise	3094	1.7273	0.0000	1.1790	2.2756	
mews	0-1=normal	648	-	-	-	-	1.06E-10
	2-3=increase caution	1503	0.7601	0.0052	0.2271	1.2930	
	4-5=further deterioration	1013	1.5510	0.0000	0.9806	2.1214	
	>5=immediate action required	285	2.4790	0.0000	1.6730	3.2850	
cindex	0=normal	1197	-	-	-	-	8.03E-18
	1-2=mild	1325	0.6443	0.0058	0.1864	1.1021	

	3-4=moderate	424	1.7555	0.0000	1.1067	2.4044	
	>5=severe	666	2.4023	0.0000	1.8473	2.9573	
evisit	0	1133	-	-	-	-	5.86E-12
	1	744	0.5174	0.0621	-0.0262	1.0611	
	2	512	1.5575	0.0000	0.9439	2.1710	
	3	324	1.9871	0.0000	1.2612	2.7130	
	4	899	1.5728	0.0000	1.0581	2.0874	
icu_flag	1=if during hospitalization, the patient had a visit in ICU	63	-	-	-	-	0.3755
	0=otherwise	3549	-0.6671	0.3755	-2.1427	0.8085	
gender	female	1952	-	-	-	-	5.18E-04
	male	1660	0.6858	0.0005	0.2988	1.0728	
race	White	2057	-	-	-	-	
	African Amer/Black	722	-0.8237	0.0564	-1.6697	0.0223	0.3566
	Asian	249	-1.1857	0.3544	-3.6956	1.3242	
	Native Amer/Alaskan	22	-2.2226	0.4540	-8.0418	3.5965	
	Natv Hawaii/Pacf Isl	4	-0.2173	0.5206	-0.8805	0.4459	
	Other/Multiracial	508	-0.0454	0.8559	-0.5353	0.4446	
religion	Angelican	1	-	-	-	-	
	Catholic	1648	1.8227	0.7583	-9.7890	13.4344	0.3911
	Christian	892	1.8859	0.7502	-9.7287	13.5006	
	Hebrew	1	1.1667	0.8892	-15.2497	17.5831	
	Hindu	122	1.7374	0.7701	-9.9183	13.3930	
	Islam	110	0.3989	0.9465	-11.2619	12.0597	
	Jewish	515	1.8682	0.7526	-9.7512	13.4876	
	Mormon	2	-0.0833	0.9908	-14.3003	14.1337	
	No Affiliation	174	0.9282	0.8758	-10.7133	12.5696	
	Non-Denominational	1	-0.9167	0.9128	-17.3331	15.4997	
	Other	146	1.8978	0.7494	-9.7500	13.5457	
maritalstatus	Civil Union	1	-	-	-	-	
	Divorced	235	-1.8495	0.7552	-13.4797	9.7808	0.1473
	Married	1607	-1.7951	0.7618	-13.4043	9.8141	
	Separated	51	-1.9265	0.7472	-13.6453	9.7924	

	Single	951	-1.7252	0.7708	-13.3369	9.8864	
	Widowed	690	-1.0708	0.8566	-12.6848	10.5432	
facilityname	Lenox Hill Hospital	5	-	-	-	-	
	LIJ Forest Hills	559	1.1649	0.6605	-4.0349	6.3646	2.23E-04
	LIJ Valley Stream	323	0.9772	0.7134	-4.2393	6.1938	
	Long Island Jewish Hospital	813	1.2801	0.6289	-3.9124	6.4727	
	NSUH	983	2.0889	0.4301	-3.1009	7.2787	
	Plainview Hospital	323	1.6608	0.5325	-3.5558	6.8774	
	Southside Hospital	523	2.2374	0.3991	-2.9640	7.4387	
	Syosset Hospital	83	0.0395	0.9884	-5.2909	5.3698	
insurancetype	Medicaid	166	-	-	-	-	
	Medicare	1425	-0.4721	0.3307	-1.4235	0.4794	1.92E-07
	Private	1987	-1.5154	0.0015	-2.4527	-0.5781	

Figure 1: Descriptive plots of continuous variables. Lower panels show bivariate scatterplots, diagonal shows histogram and top panels show absolute value of correlation

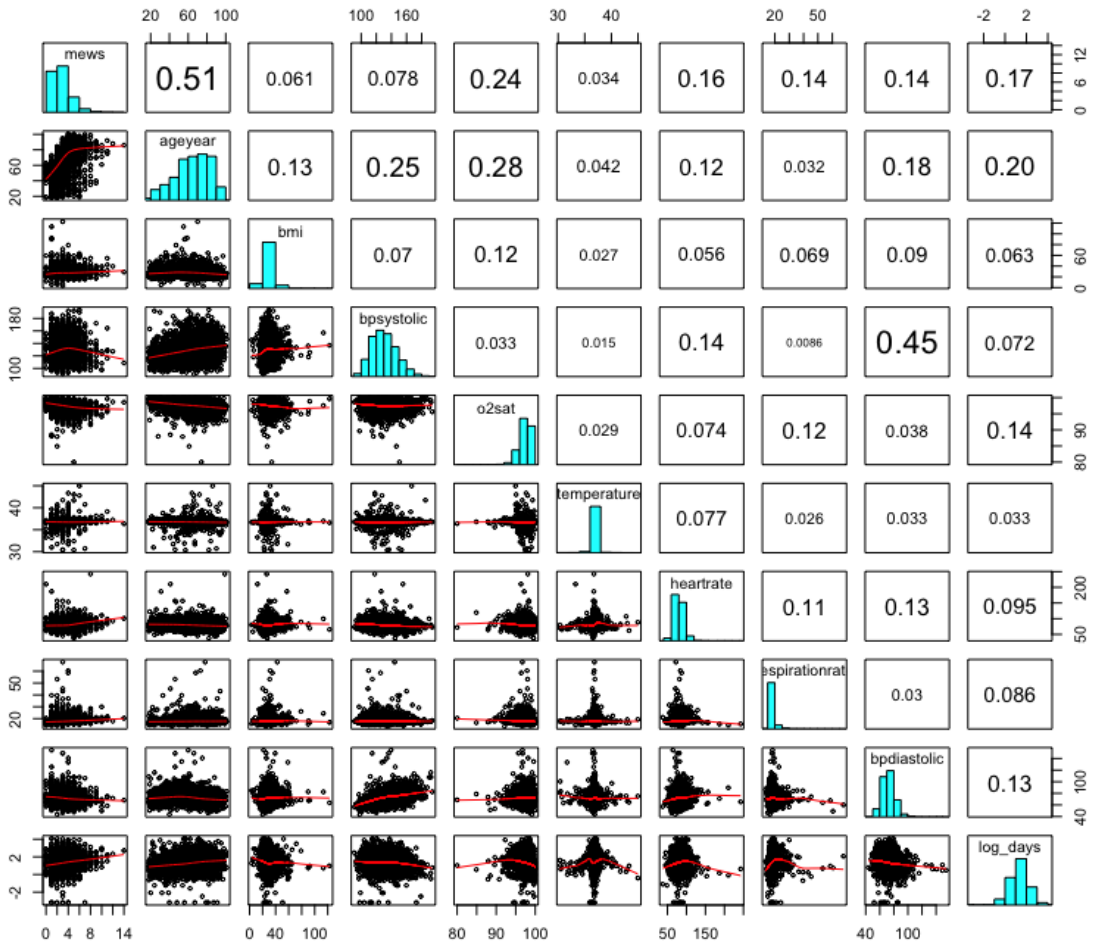

Table 3: Criterion Selection values for each model size

Number of variables	rss	rsq	adjr2	cp	bic
1	2330.060322	0.043259519	0.042957327	323.5433258	-123.9770823
2	2277.411034	0.064877717	0.064286802	246.7399265	-188.3204555
3	2242.590215	0.07917541	0.078302315	196.6214834	-229.0713288
4	2211.623993	0.091890376	0.090741961	152.2724514	-265.0598011
5	2195.230795	0.098621548	0.097196218	129.7357527	-280.5685401
6	2179.797277	0.104958668	0.10325976	108.635465	-294.8588958
7	2166.875261	0.110264546	0.108293613	91.29429613	-305.6340883
8	2155.324571	0.115007347	0.11276615	76.00567521	-314.5056617
9	2146.953224	0.118444686	0.115932337	65.47576893	-318.7733741
10	2139.639283	0.121447845	0.118664975	56.52854811	-321.5232434
11	2132.769147	0.124268776	0.12121648	48.24559635	-323.650838
12	2127.170771	0.126567511	0.123245422	41.8661662	-323.9167026
13	2122.061995	0.128665213	0.125073789	36.21954953	-323.4735094
14	2117.536921	0.130523243	0.126662579	31.44659618	-322.1752925

Table 4: Model Coefficients

Term	estimate	std.error	statistic	p.value
(Intercept)	2.1879	1.3101	1.6700	0.0950
is30dayreadmit1	0.1840	0.0437	4.2090	0.0000
cindex_catModerate	0.1228	0.0490	2.5062	0.0123
cindex_catNormal	-0.0117	0.0364	-0.3210	0.7482
cindex_catsevere	0.1579	0.0419	3.7717	0.0002
evisit1	0.0338	0.0420	0.8032	0.4219
evisit2	0.1242	0.0477	2.6054	0.0092
evisit3	0.2480	0.0560	4.4316	0.0000
evisit4	0.2272	0.0417	5.4482	0.0000
ageyear	0.0095	0.0011	8.4384	0.0000
genderMale	0.0726	0.0310	2.3416	0.0193
maritalstatusMarried	-0.0417	0.0613	-0.6809	0.4960
maritalstatusSeparated	0.1461	0.1364	1.0711	0.2842
maritalstatusSingle	0.0915	0.0649	1.4100	0.1586
maritalstatusWidowed	0.0268	0.0674	0.3974	0.6911
facilitynameLIJ Valley Stream	0.0306	0.0609	0.5024	0.6154
facilitynameLong Island Jewish Hospital	0.0364	0.0494	0.7369	0.4612
facilitynameNSUH	0.0794	0.0481	1.6524	0.0986
facilitynameOther	-0.1650	0.1023	-1.6134	0.1068
facilitynamePlainview Hospital	0.0481	0.0626	0.7684	0.4423
facilitynameSouthside Hospital	-0.0234	0.0542	-0.4322	0.6657
insurancetypeMedicare	-0.1851	0.0757	-2.4458	0.0145
insurancetypePrivate	-0.2512	0.0718	-3.4989	0.0005
bpsystolic	-0.0048	0.0011	-4.3957	0.0000
temperature	0.0677	0.0209	3.2479	0.0012
heartrate	0.0069	0.0012	5.8694	0.0000
o2sat	-0.0389	0.0102	-3.8219	0.0001
respirationrate	0.0135	0.0055	2.4682	0.0136
bpdiastolic	-0.0055	0.0019	-2.9578	0.0031

Figure 2: Diagnostic Plots

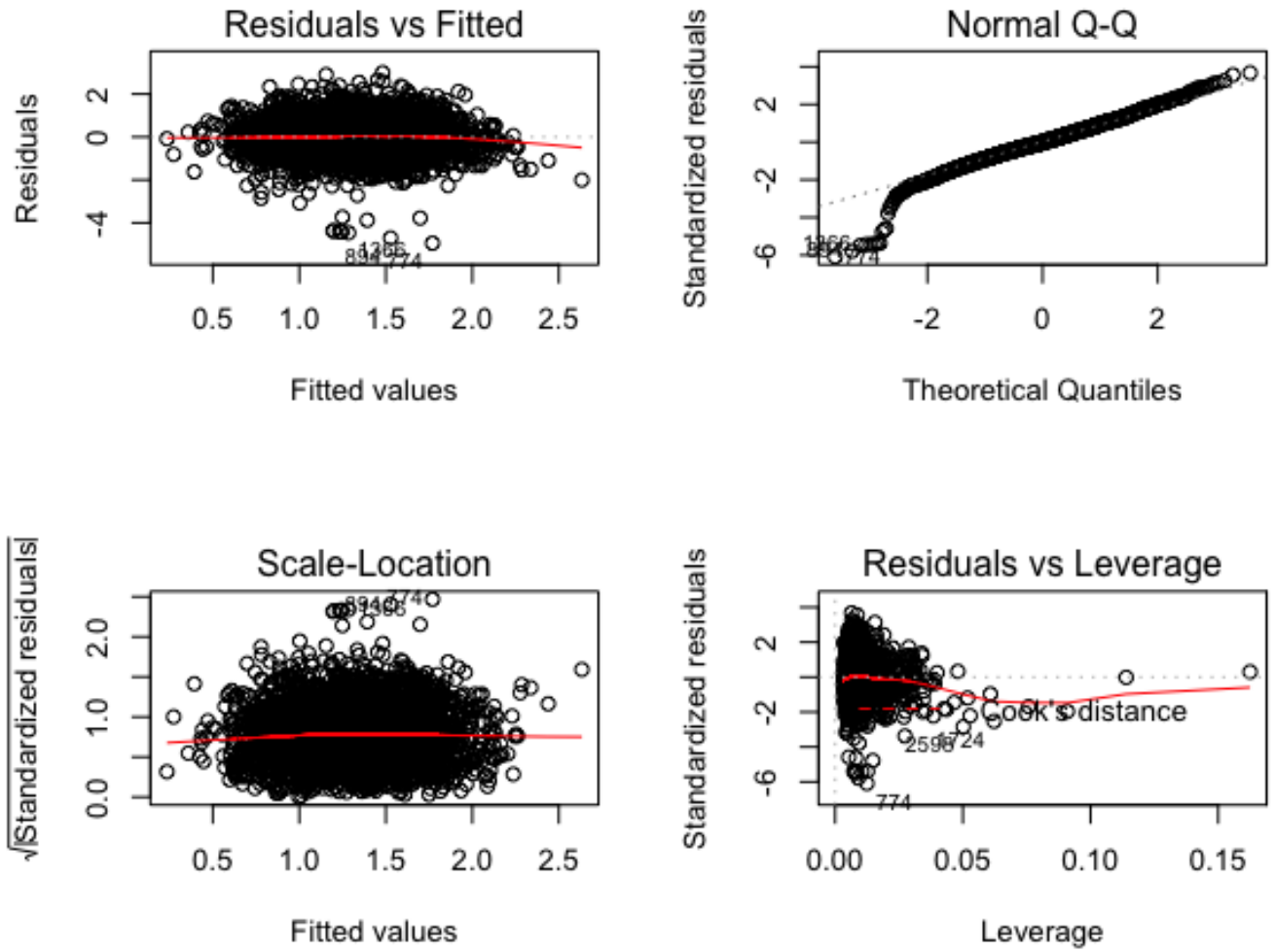

Figure 3: Bootstrap MSE

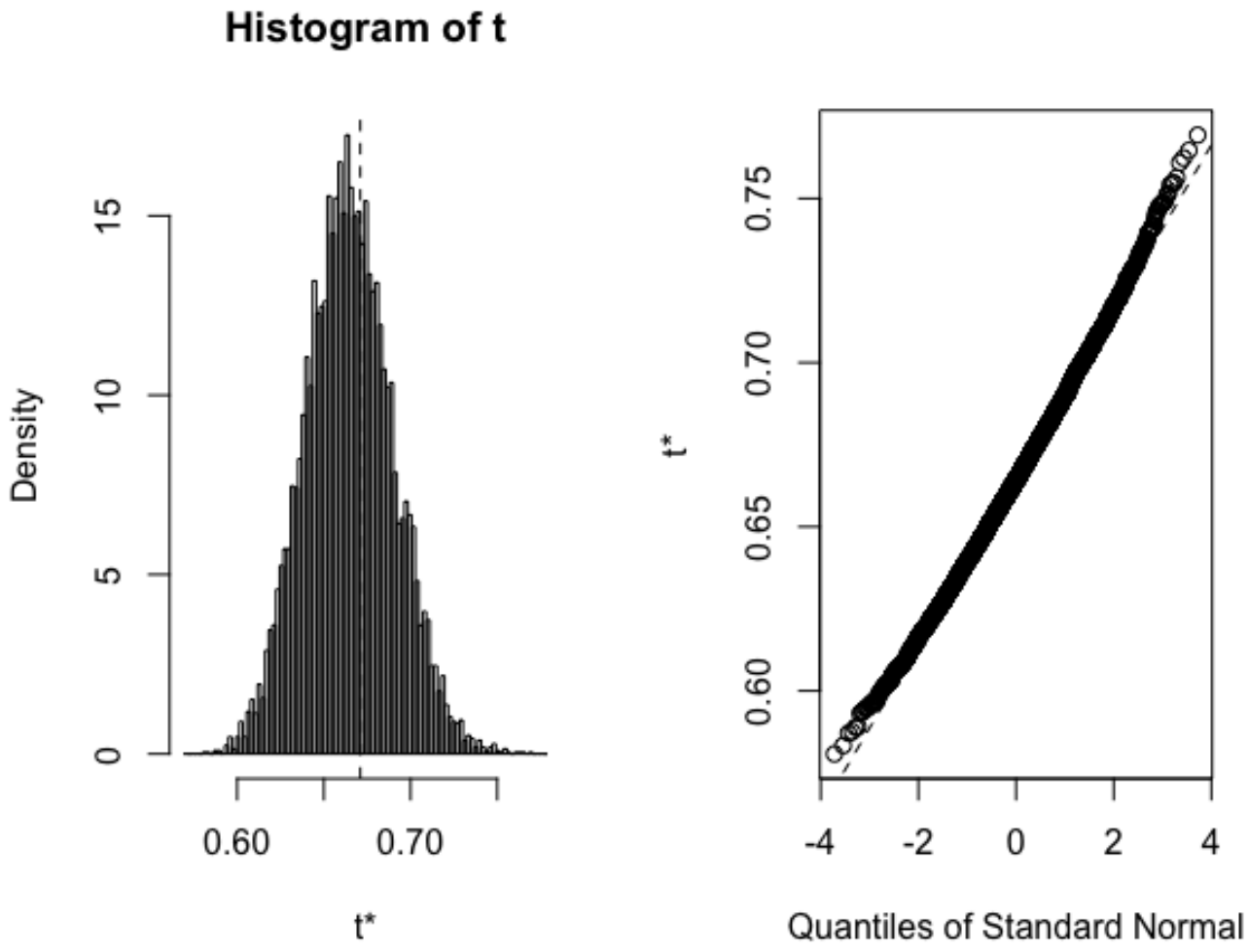

Figure 4: Refitted model without outliers