Developing a Prediction Model For Composite Outcome of Tracheotomy and Death for Infants with Severe Bronchopulmonary Dysplasia

Project 2: Regression Analysis

Yiwen Liang

12/15/2023

Abstract

Background: Bronchopulmonary dysplasia (BPD) is a chronic pulmonary condition characterized by inflammatory processes and lung scarring. Given the merits and drawbacks of tracheostomy placement, it becomes crucial to determine the criteria for tracheotomy and the optimal timing for referring a patient for this procedure.

Methods: This project construct two prediction models for composite outcome of tracheotomy and death for infants with severe bronchopulmonary dysplasia at 36 weeks and 44 weeks. For both week-36 and week-44 models, we first use *lasso* and *best subset* for variable selection of fixed effects, then fit a mixed effects model using the fixed effects obtained through variable selection, incorporating random intercept for each center. The performances of these models will be evaluated using metrics including the Brier score, AUC, sensitivity, and specificity.

Results: The week-36 models have Brier scores around 0.07, and the week-44 models have Brier scores of approximately 0.06. All four models have AUC greater than 0.9 and threshold lower than 0.15. However, the calibration of all models are not ideal.

Conclusions: The models constructed with all available variables up to 44 weeks were deemed more effective. These models provide clinicians with valuable tools for early prognosis, facilitating informed discussions with the families of infants at risk.

---

1. Introduction

Bronchopulmonary dysplasia (BPD) is a chronic pulmonary condition characterized by inflammatory processes and lung scarring (Association 2023). It stands as the most prevalent complication arising from prematurity, especially in its severe manifestation. Over 10,000 infants are affected by BPD annually. Severe bronchopulmonary dysplasia (sBPD) often requires respiratory support through mechanical ventilation or oxygen therapy, sometimes leading to tracheostomy before discharge for sustained ventilator dependence. Given the merits and drawbacks of tracheostomy placement, it becomes crucial to determine the criteria for tracheotomy and the optimal timing for referring a patient for this procedure.

This project endeavors to develop a regression model to predict the composite outcome of tracheotomy and death, providing valuable insights into the indication criteria and optimal timing for tracheotomies in neonates with sBPD. The dataset provided is a national database encompassing demographic, diagnostic, and respiratory parameters of infants with sBPD admitted to collaborative Neonatal Intensive Care Units (NICUs), with known respiratory support parameters at 36 and 44 weeks post-menstrual ages (PMA).

The data used in this project is not available to the public, hence only the code used for analysis and codebook for the dataset are available at Github, https://github.com/yiwen-liang/PHP_2550_Final_Portfolio.

2. Data

The goal of this project is to develop a prediction model. The choice of the regression model and the variables to include depend significantly on the characteristics of the dataset, encompassing both predictors and outcomes. Therefore, we will initiate the process by conducting a concise exploratory data analysis (EDA) on the data.

Subsequently, we will pre-process the dataset, laying the groundwork for a detailed explanation of the model development process and the selection of techniques. All analyses are performed in R v4.2.2.

The initial phase of EDA involves identification and elimination of potential duplicate records within the dataset. We notice that the patient with `record_id` is 2000824 has 4 identical records, prompting the removal of these duplicates. Subsequently, a thorough examination of missing data in the dataset is conducted. Among the 30 variables, 26 exhibited missing values. The quantitative details, including the count and proportion of missingness for each variable, are outlined in Table 1. Variables pertaining to information at the 44-week mark displayed a considerable extent of missingness. Additionally, the variable `any_surf` has more than 40% of missing values, hence we'll not include it while constructing the model.

Table 1: Summary of Missing Values in Each Variable

Variables	N	Proportion (%)	Variables	N	Proportion (%)
center	10	1.00	inspired_oxygen.36	92	9.24
mat_ethn	57	5.72	p_delta.36	128	12.85
blength	78	7.83	peep_cm_h2o_modified.36	117	11.75
birth_hc	77	7.73	med_ph.36	30	3.01
del_method	3	0.30	weight_today.44	446	44.78
prenat_ster	35	3.51	ventilation_support_level_modified.44	424	42.57
com_prenat_ster	193	19.38	inspired_oxygen.44	448	44.98
mat_chorio	62	6.22	p_delta.44	448	44.98
gender	4	0.40	peep_cm_h2o_modified.44	446	44.78
sga	15	1.51	med_ph.44	424	42.57
any_surf	433	43.47	hosp_dc_ga	124	12.45
weight_today.36	92	9.24	Death	2	0.20
ventilation_support_level.36	30	3.01	NA	NA	NA

The identification of 10 missing values in the variable `center` prompts further investigation. Fortunately, no missing value is observed in the corresponding `record_id` variable. Leveraging the encoding pattern inherent in the `record_id`, it's deduced that all 10 records with missing values all come from center 1. Hence, we manually impute the missing values in `center` originated from center 1. Consequently, manual imputation is employed to fill in the missing values in `center` with the value 1. As shown in Table 2, we can see that the data is collected from 10 different centers, with considerable variations in the number of observations attributed to each center. Given that the data is collected from multiple sites, it's natural for us to consider it a multilevel data, and fit mixed effect models. Notably, approximately two thirds of the records are from center 2, while a mere 4 records are from center 20 and 1 record from center 21. In order to have reasonable sample size for each cluster, we'll merge center 21 to center 20.

Table 2: Summary of the Number of Observations from Each Center

center	1	2	3	4	5	7	12	16	20
N	65	630	57	60	40	32	69	38	5

Upon observing the dataset, a discrepancy is noted in the coding of the variable `mat_race` compared to the specifications outlined in the codebook. Consequently, the decision is made to disregard the variable `mat_race` and exclude it from the subsequent model fitting processes.

Table 3 provides a summary of baseline information and measurements at weeks 36 and 44, categorized by two outcomes: tracheostomy placement and death. Among all subjects in the database, it is evident that the majority (850) did not undergo tracheostomy, and only a small portion of infants died (54 out of 937). However, since the objective is to create a predictive model for the composite outcome of tracheostomy placement and mortality, developing an appropriate model is challenging when dealing with two distinct outcome variables.

Given the objectives of this project and for our convenience, we will first create a new outcome variable, Y, which is a binary indicator for negative outcomes that combines tracheostomy and death. $Y = 0$ indicates the absence of

tracheostomy and survival, while $Y = 1$ signifies the occurrence of at least one negative outcome, encompassing either undergoing a tracheostomy, death, or both.

Table 3: Summary Statistics by Tracheostomy and Death

Death	N	Tracheostomy Overall, N = 996	No, N = 850	Yes, N = 146	N	Death Overall, N = 994	No, N = 940	Yes, N = 54
mat_ethn	939				937			
1		74.00 (7.88%)	66.00 (8.22%)	8.00 (5.88%)		74.00 (7.90%)	71.00 (8.01%)	3.00 (5.88%)
2		865.00 (92.12%)	737.00 (91.78%)	128.00 (94.12%)		863.00 (92.10%)	815.00 (91.99%)	48.00 (94.12%)
bw	996	806.07 (296.62)	813.78 (295.08)	761.21 (302.58)	994	806.18 (296.83)	811.09 (288.84)	720.69 (406.03)
ga	996	25.77 (2.14)	25.76 (2.14)	25.84 (2.13)	994	25.78 (2.14)	25.77 (2.14)	25.83 (2.13)
blength	918	32.49 (3.82)	32.56 (3.79)	31.97 (3.94)	916	32.49 (3.82)	32.61 (3.75)	30.32 (4.37)
birth_hc	919	23.19 (2.76)	23.22 (2.71)	22.99 (3.07)	917	23.19 (2.76)	23.24 (2.71)	22.34 (3.54)
del_method	993				991			
1		285.00 (28.70%)	254.00 (29.95%)	31.00 (21.38%)		284.00 (28.66%)	274.00 (29.24%)	10.00 (18.52%)
2		708.00 (71.30%)	594.00 (70.05%)	114.00 (78.62%)		707.00 (71.34%)	663.00 (70.76%)	44.00 (81.48%)
prenat_ster	961	835.00 (86.89%)	712.00 (85.78%)	123.00 (93.89%)	959	833.00 (86.86%)	789.00 (86.70%)	44.00 (89.80%)
com_prenat_ster	803	610.00 (75.97%)	524.00 (75.94%)	86.00 (76.11%)	801	608.00 (75.91%)	574.00 (75.73%)	34.00 (79.07%)
mat_chorio	934	160.00 (17.13%)	138.00 (17.25%)	22.00 (16.42%)	932	160.00 (17.17%)	153.00 (17.35%)	7.00 (14.00%)
gender	992				990			
Female		408.00 (41.13%)	348.00 (41.13%)	60.00 (41.10%)		407.00 (41.11%)	390.00 (41.67%)	17.00 (31.48%)
Male		584.00 (58.87%)	498.00 (58.87%)	86.00 (58.90%)		583.00 (58.89%)	546.00 (58.33%)	37.00 (68.52%)
sga	981				979			
Not SGA		778.00 (79.31%)	678.00 (80.81%)	100.00 (70.42%)		776.00 (79.26%)	748.00 (80.86%)	28.00 (51.85%)
SGA		203.00 (20.69%)	161.00 (19.19%)	42.00 (29.58%)		203.00 (20.74%)	177.00 (19.14%)	26.00 (48.15%)
any_surf	563	461.00 (81.88%)	395.00 (80.94%)	66.00 (88.00%)	562	461.00 (82.03%)	433.00 (81.70%)	28.00 (87.50%)
weight_today.36	904	2,120.58 (413.47)	2,131.53 (410.15)	2,023.88 (432.03)	902	2,119.76 (413.56)	2,132.30 (397.21)	1,826.46 (633.79)
ventilation_support_level.36	966				964			
0		117.00 (12.11%)	111.00 (13.23%)	6.00 (4.72%)		116.00 (12.03%)	114.00 (12.45%)	2.00 (4.17%)
1		589.00 (60.97%)	560.00 (66.75%)	29.00 (22.83%)		589.00 (61.10%)	582.00 (63.54%)	7.00 (14.58%)
2		260.00 (26.92%)	168.00 (20.02%)	92.00 (72.44%)		259.00 (26.87%)	220.00 (24.02%)	39.00 (81.25%)
inspired_oxygen.36	904	0.34 (0.15)	0.32 (0.13)	0.49 (0.20)	902	0.34 (0.15)	0.33 (0.13)	0.52 (0.25)
p_delta.36	868	5.27 (9.74)	4.25 (8.87)	15.13 (12.13)	866	5.27 (9.75)	4.85 (9.36)	17.14 (12.70)
peep_cm_h2o_modified.36	879	6.33 (2.91)	6.20 (2.90)	7.50 (2.79)	877	6.34 (2.91)	6.31 (2.92)	7.13 (2.25)
med_ph.36	966				964			
0		900.00 (93.17%)	798.00 (95.11%)	102.00 (80.31%)		899.00 (93.26%)	861.00 (94.00%)	38.00 (79.17%)
1		66.00 (6.83%)	41.00 (4.89%)	25.00 (19.69%)		65.00 (6.74%)	55.00 (6.00%)	10.00 (20.83%)
weight_today.44	550	3,646.12 (682.09)	3,667.21 (665.21)	3,550.07 (750.40)	550	3,646.12 (682.09)	3,674.90 (662.68)	3,222.71 (822.51)
ventilation_support_level_modified.44	572				572			
0		269.00 (47.03%)	262.00 (56.83%)	7.00 (6.31%)		269.00 (47.03%)	268.00 (50.28%)	1.00 (2.56%)
1		146.00 (25.52%)	128.00 (27.77%)	18.00 (16.22%)		146.00 (25.52%)	141.00 (26.45%)	5.00 (12.82%)
2		157.00 (27.45%)	71.00 (15.40%)	86.00 (77.48%)		157.00 (27.45%)	124.00 (23.26%)	33.00 (84.62%)
inspired_oxygen.44	548	0.34 (0.15)	0.32 (0.13)	0.44 (0.19)	548	0.34 (0.15)	0.33 (0.13)	0.53 (0.23)
p_delta.44	548	7.62 (14.19)	4.80 (12.01)	21.10 (15.99)	548	7.62 (14.19)	6.44 (13.12)	26.75 (17.10)
peep_cm_h2o_modified.44	550	4.30 (4.46)	3.37 (4.06)	8.69 (3.59)	550	4.30 (4.46)	4.03 (4.40)	8.78 (2.71)
med_ph.44	572				572			
0		473.00 (82.69%)	413.00 (89.59%)	60.00 (54.05%)		473.00 (82.69%)	458.00 (85.93%)	15.00 (38.46%)
1		99.00 (17.31%)	48.00 (10.41%)	51.00 (45.95%)		99.00 (17.31%)	75.00 (14.07%)	24.00 (61.54%)
hosp_dc_ga	872	52.76 (26.52)	48.92 (23.62)	79.94 (29.94)	871	52.77 (26.53)	52.42 (26.65)	58.90 (23.75)
Death	994	54.00 (5.43%)	37.00 (4.36%)	17.00 (11.64%)				
Trach					994	146.00 (14.69%)	129.00 (13.72%)	17.00 (31.48%)

[1] n (%); Mean (SD)

3

Table 4: Summary of Baseline Characteristics of Infants for With and Without Negative Outcome

Variable	No, N = 808	Yes, N = 182	p-value
mat_ethn			0.3
1	62.00 (8.12%)	10.00 (5.92%)	
2	702.00 (91.88%)	159.00 (94.08%)	
bw	817.56 (285.74)	758.54 (339.64)	<0.001
ga	25.76 (2.14)	25.86 (2.14)	0.6
blength	32.66 (3.73)	31.66 (4.12)	0.002
birth_hc	23.25 (2.65)	22.94 (3.20)	0.014
del_method			0.022
1	242.00 (30.02%)	39.00 (21.55%)	
2	564.00 (69.98%)	142.00 (78.45%)	
prenat_ster	678.00 (85.71%)	154.00 (92.77%)	0.014
com_prenat_ster	498.00 (75.80%)	109.00 (76.22%)	>0.9
mat_chorio	131.00 (17.17%)	28.00 (16.87%)	>0.9
gender			0.6
Female	334.00 (41.54%)	72.00 (39.56%)	
Male	470.00 (58.46%)	110.00 (60.44%)	
sga			<0.001
Not SGA	656.00 (82.31%)	118.00 (66.29%)	
SGA	141.00 (17.69%)	60.00 (33.71%)	
any_surf	371.00 (80.65%)	86.00 (87.76%)	0.10
weight_today.36	2,143.18 (392.74)	1,983.79 (507.87)	<0.001
ventilation_support_level.36			<0.001
0	109.00 (13.64%)	7.00 (4.35%)	
1	551.00 (68.96%)	36.00 (22.36%)	
2	139.00 (17.40%)	118.00 (73.29%)	
inspired_oxygen.36	0.31 (0.12)	0.49 (0.21)	<0.001
p_delta.36	3.87 (8.43)	15.50 (12.32)	<0.001
peep_cm_h2o_modified.36	6.17 (2.91)	7.44 (2.62)	<0.001
med_ph.36			<0.001
0	767.00 (95.99%)	128.00 (79.50%)	
1	32.00 (4.01%)	33.00 (20.50%)	
weight_today.44	3,696.84 (643.51)	3,487.95 (767.37)	0.011
ventilation_support_level_modified.44			<0.001
0	261.00 (59.73%)	8.00 (6.02%)	
1	123.00 (28.15%)	22.00 (16.54%)	
2	53.00 (12.13%)	103.00 (77.44%)	
inspired_oxygen.44	0.31 (0.11)	0.45 (0.20)	<0.001
p_delta.44	3.84 (10.90)	21.65 (16.17)	<0.001
peep_cm_h2o_modified.44	3.09 (3.91)	8.68 (3.45)	<0.001
med_ph.44			<0.001
0	404.00 (92.45%)	68.00 (51.13%)	
1	33.00 (7.55%)	65.00 (48.87%)	
hosp_dc_ga	47.86 (10.18)	73.68 (29.67)	<0.001

[1] n (%); Mean (SD)

[2] Pearson's Chi-squared test; Wilcoxon rank sum test

Table 4 presents the summary of variables by the newly created outcome variable Y. Among all subjects in the database, it's evident that the majority (808) didn't undergo tracheostomy and survived, while only 182 infants

belong to the *negative outcome* group ($Y = 1$). Infants in the negative outcome group clearly exhibit smaller size, lower birth weights (`bw`), shorter birth lengths (`blength`), smaller birth head circumference (`birth_hc`), and a higher likelihood of being small for gestational age. Besides, notable differences are observed in measures at week 36 and week 44 between the two groups. Infants with negative outcomes continue to be of lighter weight, more prone to requiring ventilation support and medication, exhibiting a higher proportion of inspired oxygen, and being discharged at an older age compared to those who didn't undergo tracheostomy placement and survived.

To handle missing values, the Multiple imputation (MI) technique is applied. The missing values are imputed using values generated from distributions and relationships among observed variables in the dataset. `mice()` function from `mice` package is utilized to obtain test and train (validation) dataset separately.

3. Methods

Before variable selection, two sets of variables are considered concerning the time of measurements: baseline and 36-week information, and baseline and 44-week information. The first model aims to make the best use of all available measurements at 36 weeks, intending to assess the need for tracheostomy placement at 36 weeks. Subsequently, the second model takes into account all measurements up to 44 weeks, providing predictions for the combined outcome of tracheostomy and death in infants with sBPD at 44 weeks. Scientists have suggested that earlier tracheostomy placement may be more beneficial for growth. If the predictions obtained from the 36-week model are fairly close to the 44-week model's prediction, or at least the predictive accuracy of the 36-week model is acceptable, these models can offer clinicians valuable tools for early prognosis of the composite outcome of tracheostomy and death in infants with sBPD.

The approach involves conducting variable selection for two sets of variables using two methods introduced in class: *lasso* regression and the *best subset* selection procedure. Each method undergoes a 10-fold cross-validation, and the resulting coefficients are averaged over 5 imputed datasets to obtain the coefficients for each variable. For both *lasso* and *best subset*, variables are selected, and those being dropped will have coefficients equal to zero. Based on the results from variable selection, a generalized mixed model will be fitted, where only variables selected by *lasso* or *best subset* will be included, incorporating random intercepts with the center acting as a random effect. The results from variable selection (coefficients table) and the performance evaluation of the mixed-effects models on the test dataset will be presented in the **4. Result** section, respectively. The calibration and discrimination of the models will be evaluated using metrics such as the Brier score, AUC, sensitivity, and specificity. Since the models are trained and tested in 5 imputed datasets, and each dataset has been further divided into training and testing sets, the metrics will be pooled, and the results displayed in the table will represent the average values.

4. Results

4.1 36-Week Prediction Models

Table 5 displays the coefficients for the set of variables used in the 36-week model after being selected by lasso regression and the best subset procedure. Notably, lasso regression identifies 12 variables, while the best subset procedure results in the selection of only 7 variables. Subsequently, generalized mixed models are fitted, incorporating the subset of variables selected by each approach, and random intercept for each `center`. Table 6 provides a comprehensive evaluation of the performance of the two models, presenting multiple metrics. Interestingly, the mixed effects models, including variables selected by lasso regression and the best subset technique, exhibit very similar Brier scores. Table 6 also outlines the thresholds that maximize the sum of specificity and sensitivity. Specifically, the threshold for the model based on lasso selection is 0.1242, whereas the model based on the best subset selection has a lower threshold, equal to 0.1242. Both models demonstrate specificity values slightly above 0.85, and identical sensitivity of 0.925.

The discrimination of the models is evaluated using the AUC (Area Under the ROC Curve), which assesses their ability to distinguish between individuals with and without the specified outcome. In the context of binary outcomes, the c-statistic is equivalent to the AUC. A higher AUC value indicates better model discrimination, suggesting an increased ability to correctly classify individuals with and without the outcome. Both models demonstrate good discrimination, with an AUC of approximately 0.9247.

Calibration plots are employed to evaluate the calibration of the models, assessing the agreement between the observed frequency of events and the predicted probabilities across different percentiles. The visualization in Figure

1 helps illustrate the degree of concordance between predicted and actual distributions, and a close alignment with the $x = y$ line indicates good calibration. The calibration plots for the two models show few differences, and none of them demonstrates significantly superior calibration.

Table 5: Variable Selection for 36-Week Model According to Lasso and Best Subset

	Lasso	Best Subset
(Intercept)	-6.7055	-7.6477
mat_ethn2	0.0000	0.0000
bw	0.0000	0.0000
ga	0.0294	0.0000
blength	0.0000	0.0000
birth_hc	0.0253	0.0931
del_method2	0.0350	0.0000
prenat_sterYes	0.6222	1.3455
com_prenat_sterYes	0.0133	0.0000
mat_chorioYes	0.0000	0.0000
genderMale	0.0000	0.0000
sgaSGA	0.0483	0.0000
weight_today.36	-0.0006	-0.0012
ventilation_support_level.361	0.0000	0.0000
ventilation_support_level.362	1.3782	1.7074
inspired_oxygen.36	2.8587	2.8884
p_delta.36	0.0000	0.0000
peep_cm_h2o_modified.36	0.0020	0.0000
med_ph.361	0.3158	0.7276
hosp_dc_ga	0.0483	0.0581

Table 6: Specifity and Sensitivity at Thresholds, and AUC

	Brier.Scores	AUC	Threshold	Specificity	Sensitivity
Lasso	0.0720	0.9369	0.1242	0.8636	0.925
Best Subset	0.0713	0.9352	0.0969	0.8535	0.925

4.2 44-Week Prediction Models

Table 7 displays the coefficients for the set of variables used in the 44-week model after being selected by lasso regression and the best subset procedure. It is evident that there are more variables available at week 44. Lasso regression identifies 16 variables, while the best subset procedure results in the selection of only 8 variables. Both approaches select more 44-week measurements than 36-week measurements. Subsequently, generalized mixed models are fitted, incorporating the subset of variables selected by each approach and random intercept for each `center`. Table 8 provides a comprehensive evaluation of the performance of the two models, presenting multiple metrics. The mixed effects models, including variables selected by lasso regression and the best subset technique, exhibit close Brier scores. Table 8 also outlines the thresholds that maximize the sum of specificity and sensitivity. Specifically, the threshold for the model based on lasso selection is 0.099, whereas the model based on the best subset selection has a lower threshold, equal to 0.1132. Both models demonstrate specificity values around 0.83 and an identical sensitivity of 0.925, with specificity values lying between 0.92 and 0.95.

The AUC (area under the ROC curve) is utilized to evaluate the discrimination. Both models seem to have good discrimination, with an AUC of approximately 0.95. Calibration plots are used to assess the calibration of models. The calibration plots in Figure 2 for the lasso and ridge models exhibit few differences, and both models demonstrate
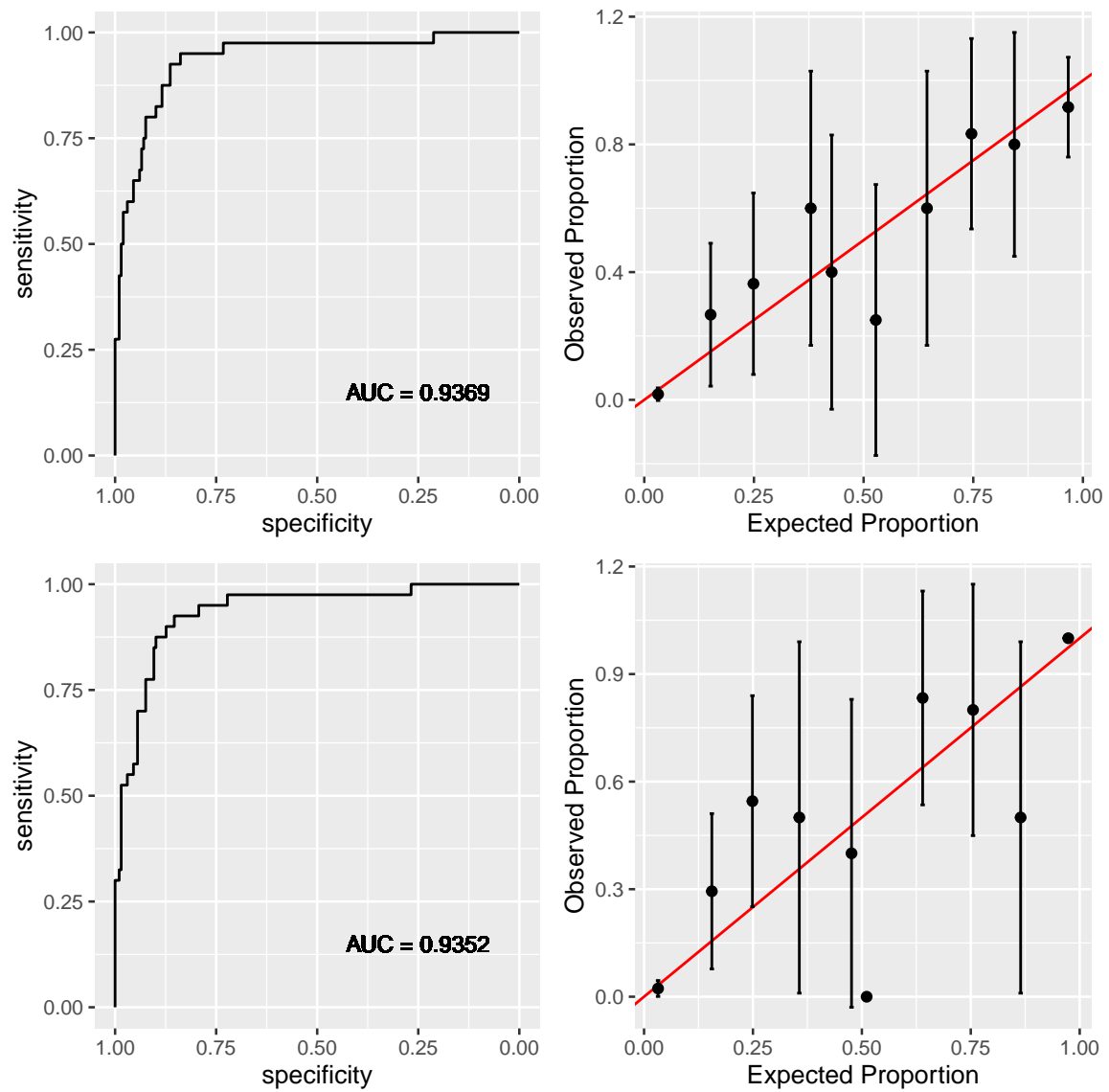
Figure 1: Calibration Plots and ROC Curves for the 36-Week Models

better calibration compared to the week-36 model. More comparisons between week-36 models and week-44 models will be discussed later.

Table 7: Variable Selection for 44-Week Model According to Lasso and Best Subset

	Lasso	Best Subset
(Intercept)	-5.2161	-6.6899
mat_ethn2	0.0000	0.0000
bw	0.0001	0.0009
ga	0.0000	0.0000
blength	0.0000	0.0000
birth_hc	0.0157	0.0000
del_method2	0.0736	0.0000
prenat_sterYes	0.8373	1.7301
com_prenat_sterYes	0.0171	0.0000
mat_chorioYes	0.0000	0.0000
genderMale	0.0122	0.0000
sgaSGA	0.0000	0.0000
weight_today.36	0.0000	0.0000
ventilation_support_level.361	0.0000	0.0000
ventilation_support_level.362	1.0486	1.3717
inspired_oxygen.36	1.6887	2.0495
p_delta.36	0.0000	0.0000
peep_cm_h2o_modified.36	0.0018	0.0000
med_ph.361	-0.0008	0.0000
weight_today.44	-0.0006	-0.0008
ventilation_support_level_modified.441	0.0000	0.0000
ventilation_support_level_modified.442	0.8394	0.7599
inspired_oxygen.44	0.3747	0.0000
p_delta.44	0.0000	0.0000
peep_cm_h2o_modified.44	0.0829	0.1007
med_ph.441	0.9004	1.2372
hosp_dc_ga	0.0419	0.0573

Table 8: Specifity and Sensitivity at Thresholds, and AUC

	Brier.Scores	AUC	Threshold	Specificity	Sensitivity
Lasso	0.0631	0.9542	0.0990	0.8283	0.950
Best Subset	0.0614	0.9524	0.1132	0.8384	0.925

5. Discussion

In this project, a total of 4 models are constructed using two different approaches, each with two different sets of initial variables. The variable selection results in Table 5 and Table 7 reveal that *best subset* is more "aggressive" in shrinking (selecting) the coefficient estimates toward zero (Hastie, Tibshirani, and Tibshirani 2020), with noticeably more zeros from *best subset* compared to *lasso*.

Most covariate effects align with the findings from EDA, including factors such as delivery method, prenatal corticosteroids, the proportion of inspired oxygen, and medication for pulmonary hypertension. Furthermore, all three models suggest that there is no discernible difference between males and females regarding the composite outcome of tracheostomy and death. However, a few discrepancies are observed. Contrary to expectations from
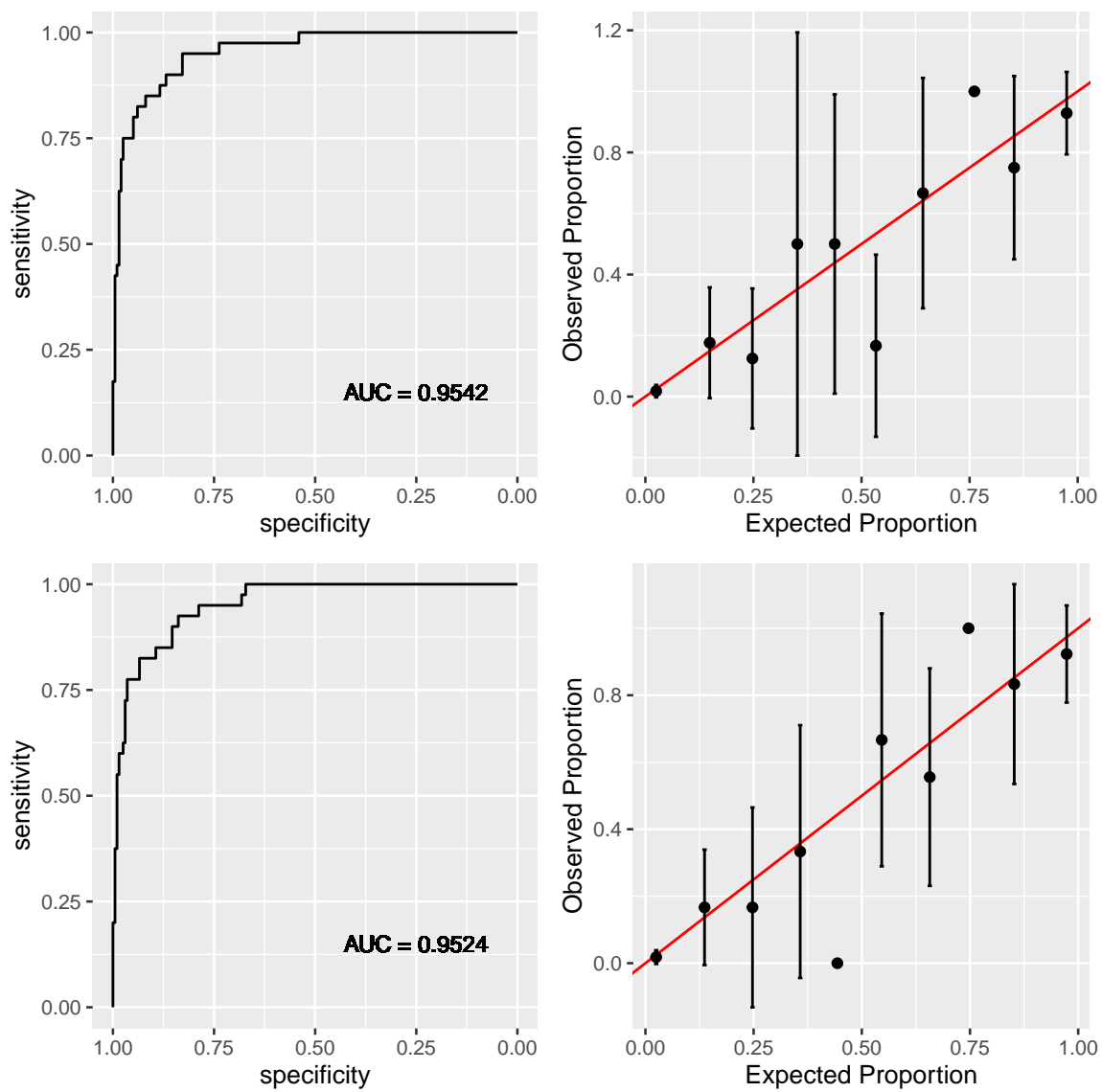
Figure 2: Calibration Plots and ROC Curves for the 44-Week Models

EDA, where birth weight and birth length are anticipated to be significant in explaining the outcome, `blength` is consistently excluded in all four models, and `bw` is only included in the week-44 model based on *lasso* regression. A potential explanation for this discrepancy may be found in the correlation matrix, which indicates a correlation between birth weight, weight at week 36, and weight at week 44. This correlation might account for the observed exclusion of `bw` and `blength` in the models.

Table 9: Correlation Matrix

	Birth Weight	Birth Length	Weight at Week 36	Weight at Week 44
bw	1.0000000	0.8533918	0.4836248	0.3827375
blength	0.8533918	1.0000000	0.4331176	0.3574033
weight_today.36	0.4836248	0.4331176	1.0000000	0.7586528
weight_today.44	0.3827375	0.3574033	0.7586528	1.0000000

All models demonstrate exceptional predictive accuracy, notably surpassing the models in the initial report that neglected to incorporate the random effects of `center`. Moreover, two week-44 models that include measurements from both 36 weeks and 44 weeks seem to exhibit superior prediction, as evidenced by Brier scores for week-44 models hovering around 0.6, while the Brier scores for week-36 models are around 0.7.

Furthermore, they exhibit commendable discrimination. The week-44 models also seem to have a larger AUC, which is more preferable. The chosen cutoff points, maximizing the sum of specificity and sensitivity, are notably low for all four models. In the context of this study, such low cutoff points are desirable as they align with the objective of minimizing the risk of overlooking any negative outcomes. However, the calibration of all models is not ideal, with wide error bars and some points located far away from the $x = y$ line. Considering multiple evaluation metrics, it can be concluded that the mixed-effects models constructed with all available variables up to 44 weeks have better performances, with few differences observed between the two approaches of variable selection.

The development of predictive models with high predictive accuracy and discrimination holds promising clinical implications. These models provide clinicians with valuable tools for early prognosis of the composite outcome of tracheostomy and death in infants with severe bronchopulmonary dysplasia. The ability to make accurate predictions allows for timely decision-making and informed discussions with the patient's family. Additionally, the coefficient estimates offer a nuanced understanding of relevant covariates, enhancing the clinician's capacity to tailor intervention strategies and improve patient care.

Our analysis has a few limitations. Firstly, we did not include any interaction terms while fitting models. As observed in the correlation matrix in Table 9, there are potentially correlated pairs of numeric variables. Additionally, due to the lack of relevant medical knowledge, some variables may have mechanistic associations that were not accounted for. Secondly, considering that the data is collected from 10 different centers, we treated this dataset as multilevel and fit mixed effect models. In this project, we used *lasso* and *best subset* for variable selection of fixed effects, then fit a mixed effects model using the fixed effects obtained through variable selection. However, some scientists argue that separating the fixed and random effects while conducting variable selection may be problematic, as the structure of the random effects will have an influence on which fixed effect variables are selected (Bondell HD 2010).

6. Conclusion

In conclusion, this study successfully developed predictive models for the composite outcome of tracheostomy and death in infants with severe bronchopulmonary dysplasia. The models constructed with all available variables up to 44 weeks were deemed more effective. These models provide clinicians with valuable tools for early prognosis, facilitating informed discussions with the families of infants at risk. However, future research endeavors should explore whether any interaction terms should be included in the model, considering our prior knowledge and the cluster-size imbalance. Additionally, it would be beneficial to investigate whether there is a better approach to perform variable selection for a mixed effects model in this study, given the large number of fixed effects and only one potential random effect.

References

Association, American Lung. 2023. "Bronchopulmonary Dysplasia." https://www.lung.org/lung-health-diseases/lung-disease-lookup/bronchopulmonary-dysplasia#:~:text=BPD%20is%20associated%20with%20infla mmation,development%20of%20BPD%20more%20likely.

Bondell HD, Ghosh SK, Krishna A. 2010. "Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models." *Biometrics* 66 (December): 1069–77. https://doi.org/10.1111/j.1541-0420.2010.01391.

Hastie, Trevor, Robert Tibshirani, and Ryan Tibshirani. 2020. "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons." *Statistical Science* 35 (November): 579–92. https://doi.org/10.1214/19-STS733.

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

# Load the packages
library(knitr)
library(tidyverse)
library(kableExtra)
library(reshape2)
library(gt)
library(gtsummary)
library(corrplot)
library(cowplot)
library(ggplot2)
library(ggpubr)
library(janitor)
library(mice)
library(nnet)
library(bestglm)
library(glmnet)
library(EnvStats)
library(L0Learn)
library(MASS)
library(lme4)
library(pROC)


#sessionInfo()

# Load the data
df <- read.csv("project2.csv")

# 1 record_id has 4 records in the dataset (3 duplicated)
# sum(duplicated(df))
df <- unique(df)

# Format the variables
# Drop mat_race during coding error
df <- df %>%
  dplyr::select(-c(mat_race))

# Re-format outcome variables to character
df <- df %>%
  mutate(Trach = case_when(Trach == "0" ~ "No",
                           Trach == "1" ~ "Yes"))
df$Trach = as.factor(df$Trach)
df$Death <- as.factor(df$Death)

# Combine center 20 and 21
df$center[df$center==21] <- 20
df$center <- as.factor(df$center)

df$mat_ethn <- as.factor(df$mat_ethn)
df$del_method <- as.factor(df$del_method)
df$prenat_ster <- as.factor(df$prenat_ster)
df$com_prenat_ster <- as.factor(df$com_prenat_ster)
```

```r
df$mat_chorio <- as.factor(df$mat_chorio)
df$gender <- as.factor(df$gender)
df$sga <- as.factor(df$sga)
df$any_surf <- as.factor(df$any_surf)

df$med_ph.36 <- as.factor(df$med_ph.36)
df$med_ph.44 <- as.factor(df$med_ph.44)
df$ventilation_support_level.36 <- as.factor(df$ventilation_support_level.36)
df$ventilation_support_level_modified.44 <- as.factor(df$ventilation_support_level_modified.44)

# 26 out of 29 variables have missing value
descript1 <- df %>%
  summarise(
    N = colSums(is.na(df)),
    prop = round(colMeans(is.na(df))*100, 2)) %>%
  mutate(Variables = colnames(df)) %>%
  filter(N != 0) %>%
  as.data.frame()

descript1 <- descript1[,c(3,1,2)]

# Display missing data summary table using kable
knitr::kable(
  list(descript1[1:13,], descript1[14:26,]),
  caption = "Summary of Missing Values in Each Variable",
  col.names = linebreak(c("Variables", "N", "Proportion (%)")),
  row.names = FALSE,
  booktabs = TRUE,
  escape = TRUE, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'),
                font_size = 8)

# Impute missing in center with 1
df$center[is.na(df$center)] <- 1

# Data was collected from 10 centers
# 2/3 are from center 2
df %>%
  group_by(center) %>%
  summarise(N = n()) %>%
  t() %>%
  kable(caption = "Summary of the Number of Observations from Each Center",
        row.names = TRUE, booktabs = TRUE, escape = TRUE, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'))

# Tableone
tbl1 <- df %>%
  dplyr::select(-c(record_id, center)) %>%
  tbl_summary(digits = list(everything() ~ c(2)),
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              by = Trach,
              missing = "no") %>%
  add_overall() %>%
  add_n() %>%
```

```r
  modify_header(label ~ "**Death**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Treatment**") %>%
  bold_labels()

tbl2 <- df %>%
  dplyr::select(-c(record_id, center)) %>%
  tbl_summary(digits = list(everything() ~ c(2)),
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              by = Death,
              missing = "no") %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Death**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Treatment**") %>%
  bold_labels()

tbl_merge(
  tbls = list(tbl1, tbl2),
  tab_spanner = c("**Tracheostomy**", "**Death**")) %>%
  as_kable_extra(caption = "Summary Statistics by Tracheostomy and Death",
                 booktabs = TRUE, escape = TRUE, align = "c") %>%
  kable_styling(latex_options = c('striped', 'HOLD_position', 'scale_down'))

# Negative outcome
df <- df %>%
  filter(!is.na(Death)) %>%
  mutate(Y = case_when(Trach=="Yes" ~ "Yes",
                       Death=="Yes" ~ "Yes",
                       (Trach=="No" & Death=="No") ~ "No"))

df$Y <- as.factor(df$Y)

# Outliers:
# hosp_dc_ga>300 weeks
# hosp_dc_ga<30 but have measures at 36 wk
df <- df %>%
  filter(!is.na(Y)) %>%
  filter((hosp_dc_ga >=30 & hosp_dc_ga < 300) | is.na(hosp_dc_ga))

# Group by outcome Y
df %>%
  dplyr::select(-c(record_id, center, Trach, Death)) %>%
  tbl_summary(digits = list(everything() ~ c(2)),
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              by = Y,
              missing = "no") %>%
  add_p() %>%
  modify_header(label = "**Variable**") %>%
  as_kable_extra(caption = "Summary of Baseline Characteristics of Infants
                 for With and Without Negative Outcome",
                 booktabs = TRUE, escape = TRUE, align = "c") %>%
  kable_styling(latex_options = c('striped', 'HOLD_position'))

df <- df %>%
  dplyr::select(-c(any_surf))
###############################
```

```r
#### Multiple Imputation ####
#############################

# MI
set.seed(2550)
samp <- sample(c(TRUE, FALSE), size = 25,
                replace = TRUE, prob = c(0.3, 0.7))
traindata <- df[!samp, ]
testdata <- df[samp, ]

imp.train <- mice(traindata, m=5, print=FALSE, seed=2550)
train_imp <- vector("list", 5)
imp.test <- mice.mids(imp.train, newdata=testdata, print=FALSE, seed=2550)
test_imp <- vector("list", 5)

for (i in 1:5){
  train_imp[[i]] <- mice::complete(imp.train, i)
  test_imp[[i]] <- mice::complete(imp.test, i)
}

# Week 36
df_36_train <- vector("list", 5)
df_36_test <- vector("list", 5)

for (i in 1:5){
  df_36_train[[i]] <- train_imp[[i]][,c(1:19,26,29)]
  df_36_test[[i]] <- test_imp[[i]][,c(1:19,26,29)]
}

# Week 44
df_44_train <- vector("list", 5)
df_44_test <- vector("list", 5)

for (i in 1:5){
  df_44_train[[i]] <- train_imp[[i]][,-c(27:28)]
  df_44_test[[i]] <- test_imp[[i]][,-c(27:28)]
}

###############
#### Lasso ####
###############

lasso <- function(df) {
  #' Runs 10-fold CV for lasso and returns corresponding coefficients
  #' @param df, data set
  #' @return coef, coefficients for minimum cv error

  # Matrix form for ordered variables
  x.ord <- model.matrix(Y ~., data = df)[,-1]
  y.ord <- as.matrix(df$Y)

  # Generate folds
  k <- 10
  set.seed(1)
  folds <- sample(1:k, nrow(df), replace=TRUE)
```

```r
  # Lasso model
  lasso_mod_cv <- cv.glmnet(x.ord, y.ord, nfolds = 10, foldid = folds,
                            alpha = 1, family = "binomial")
  lasso_mod <- glmnet(x.ord, y.ord, nfolds = 10, alpha = 1,
                      family = "binomial",
                      lambda = lasso_mod_cv$lambda.min)

  # Get coefficients
  coef <- coef(lasso_mod)
  return(coef)
}


#######################
#### Best Subset ####
#######################

L0Learn_func <- function(df) {
  #' Runs 10-fold CV for ridge and returns corresponding coefficients
  #' @param df, data set
  #' @return coef, coefficients for minimum cv error

  # Matrix form for ordered variables
  x.ord <- model.matrix(Y ~., data = df)[,-1]
  y.ord <- as.matrix(df$Y)

  # Ridge model
  L0Learn_mod <- L0Learn.cvfit(x.ord, y.ord, nFolds = 10, seed = 1,
                               loss = "Logistic", penalty = "L0")

  # Get coefficients
  min_index <- which.min(L0Learn_mod$cvMeans[[1]])
  coef <- coef(L0Learn_mod, lambda = L0Learn_mod$fit$lambda[[1]][min_index])
  return(coef)
}

df_36_test_long <- bind_rows(df_36_test)
x_vars_36 <- model.matrix(Y ~. , df_36_test_long[,-c(1,2)])

df_44_test_long <- bind_rows(df_44_test)
x_vars_44 <- model.matrix(Y ~. , df_44_test_long[,-c(1,2)])

# Lasso: variable selection
lasso_coef1 <- lasso(df_36_train[[1]][,-c(1,2)])
lasso_coef2 <- lasso(df_36_train[[2]][,-c(1,2)])
lasso_coef3 <- lasso(df_36_train[[3]][,-c(1,2)])
lasso_coef4 <- lasso(df_36_train[[4]][,-c(1,2)])
lasso_coef5 <- lasso(df_36_train[[5]][,-c(1,2)])

lasso_coef <- cbind(lasso_coef1, lasso_coef2, lasso_coef3,
                    lasso_coef4, lasso_coef5)
avg_coefs_lasso_36 <- apply(lasso_coef, 1, mean)

pred_lasso_36 <- plogis(x_vars_36 %*% avg_coefs_lasso_36)

# Best subset: variable selection
L0Learn_coef1 <- L0Learn_func(df_36_train[[1]][,-c(1,2)])
```

```r
L0Learn_coef2 <- L0Learn_func(df_36_train[[1]][,-c(1,2)])
L0Learn_coef3 <- L0Learn_func(df_36_train[[1]][,-c(1,2)])
L0Learn_coef4 <- L0Learn_func(df_36_train[[1]][,-c(1,2)])
L0Learn_coef5 <- L0Learn_func(df_36_train[[1]][,-c(1,2)])

L0Learn_coef <- cbind(L0Learn_coef1, L0Learn_coef2, L0Learn_coef3,
                      L0Learn_coef4, L0Learn_coef5)
avg_coefs_L0Learn_36 <- apply(L0Learn_coef, 1, mean)

pred_L0Learn_36 <- plogis(x_vars_36 %*% avg_coefs_L0Learn_36)

# Variable selection results
round(cbind(avg_coefs_lasso_36, avg_coefs_L0Learn_36), 4) %>%
  kbl(caption = "Variable Selection for 36-Week Model According to Lasso and Best Subset",
      col.names = linebreak(c("Lasso","Best Subset")),
      row.names = TRUE, booktabs = TRUE, escape = TRUE, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'))

# Lasso -> mixed effects model
pred_lasso_mixed_36 <- vector("list", 5)

for (i in 1:5) {
  mod_lasso <- glmer(Y ~ ga + birth_hc + del_method + prenat_ster + mat_chorio
                     + sga + weight_today.36 + ventilation_support_level.36
                     + inspired_oxygen.36 + peep_cm_h2o_modified.36 + med_ph.36
                     + hosp_dc_ga + (1 | center),
                     family = binomial, data = df_36_train[[i]],
                     control=glmerControl(optimizer="bobyqa",
                                          optCtrl=list(maxfun=2e5)))

  pred_lasso_mixed_36[[i]] <- predict(mod_lasso, df_36_test[[i]], type = "response")
}
# Lasso-36 weeks
# ROC
levels(df_36_test[[i]]$Y) = c(0, 1)
roc_lasso_36 <- roc(predictor = pred_lasso_mixed_36[[1]],
                    response = df_36_test[[i]]$Y,
                    levels = c(0, 1), direction = "<")
# coords(roc_lasso_36, "best")
p_roc_lasso_36 <- ggroc(roc_lasso_36) +
  geom_text(aes(0.25, 0.15, label = paste0("AUC = ",
                paste(unlist(round(roc_lasso_36$auc,4)))))), size=3) +
  theme_grey(base_size = 10)

# Calibration
num_cuts <- 10
calib_data <-  data.frame(prob = pred_lasso_mixed_36[[1]],
                          bin = cut(pred_lasso_mixed_36[[1]],
                                    breaks = num_cuts),
                          class = as.numeric(df_36_test[[i]]$Y)-1)
calib_data <- calib_data %>%
  group_by(bin) %>%
  summarise(observed = sum(class)/n(),
            expected = sum(prob)/n(),
            se = sqrt(observed*(1-observed)/n()))
```

```r
p_calib_lasso_36 <- ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color="red") +
  geom_errorbar(aes(x = expected,
                    ymin = observed - 1.96*se,
                    ymax = observed + 1.96*se),
                colour="black", width=.01)+
  geom_point(aes(x = expected, y = observed)) +
  labs(x = "Expected Proportion", y = "Observed Proportion") +
  theme_grey(base_size = 10)

# Best subset -> mixed effects model
pred_bs_mixed_36 <- vector("list", 5)

for (i in 1:5) {
  mod_bs <- glmer(Y ~ ventilation_support_level.36 + inspired_oxygen.36
                  + hosp_dc_ga + (1 | center),
                  family = binomial, data = df_36_train[[i]],
                  control=glmerControl(optimizer="bobyqa",
                                       optCtrl=list(maxfun=2e5)))

  pred_bs_mixed_36[[i]] <- predict(mod_lasso, df_36_test[[i]], type = "response")
}
# BS-36 weeks
# ROC
levels(df_36_test[[i]]$Y) = c(0, 1)
roc_bs_36 <- roc(predictor = pred_bs_mixed_36[[1]],
                 response = df_36_test[[i]]$Y,
                 levels = c(0, 1), direction = "<")
# coords(roc_bs_36, "best")
p_roc_bs_36 <- ggroc(roc_bs_36) +
  geom_text(aes(0.25, 0.15, label = paste0("AUC = ",
                paste(unlist(round(roc_bs_36$auc,4)))))), size=3) +
  theme_grey(base_size = 10)

# Calibration
num_cuts <- 10
calib_data <-  data.frame(prob = pred_bs_mixed_36[[1]],
                          bin = cut(pred_bs_mixed_36[[1]],
                                    breaks = num_cuts),
                          class = as.numeric(df_36_test[[i]]$Y)-1)
calib_data <- calib_data %>%
  group_by(bin) %>%
  summarise(observed = sum(class)/n(),
            expected = sum(prob)/n(),
            se = sqrt(observed*(1-observed)/n()))

p_calib_bs_36 <- ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color="red") +
  geom_errorbar(aes(x = expected,
                    ymin = observed - 1.96*se,
                    ymax = observed + 1.96*se),
                colour="black", width=.01)+
  geom_point(aes(x = expected, y = observed)) +
  labs(x = "Expected Proportion", y = "Observed Proportion") +
  theme_grey(base_size = 10)
```

```r
ggarrange(p_roc_lasso_36, p_calib_lasso_36,
          p_roc_bs_36, p_calib_bs_36, ncol = 2, nrow = 2)

brier_lasso <- c()
brier_bs <- c()
auc_lasso <- c()
auc_bs <- c()
thres_lasso <- c()
thres_bs <- c()
spec_lasso <- c()
spec_bs <- c()
sens_lasso <- c()
sens_bs <- c()
for (i in 1:5) {
  brier_lasso <- c(brier_lasso, mean((pred_lasso_mixed_36[[i]] -
                                       (as.numeric(df_36_test[[i]]$Y)-1))^2))
  brier_bs <- c(brier_bs, mean((pred_bs_mixed_36[[i]] -
                                 (as.numeric(df_36_test[[i]]$Y)-1))^2))
  auc_lasso <- c(auc_lasso, roc_lasso_36$auc)
  auc_bs <- c(auc_bs, roc_bs_36$auc)
  thres_lasso <- c(thres_lasso, coords(roc_lasso_36, "best")$threshold)
  thres_bs <- c(thres_bs, coords(roc_bs_36, "best")$threshold)
  spec_lasso <- c(spec_lasso, coords(roc_lasso_36, "best")$specificity)
  spec_bs <- c(spec_bs, coords(roc_bs_36, "best")$specificity)
  sens_lasso <- c(sens_lasso, coords(roc_lasso_36, "best")$sensitivity)
  sens_bs <- c(sens_bs, coords(roc_bs_36, "best")$sensitivity)
}

# Table: Brier scores, AUC, thresholds, specificity, and sensitivity
df_thres <- data.frame(
  "Brier Scores" = round(c(mean(brier_lasso), mean(brier_bs)), 4),
  AUC = round(c(mean(auc_lasso), mean(auc_bs)), 4),
  Threshold = round(c(mean(thres_lasso), mean(thres_bs)), 4),
  Specificity = round(c(mean(spec_lasso), mean(spec_bs)), 4),
  Sensitivity = round(c(mean(sens_lasso), mean(sens_bs)), 4))
rownames(df_thres) <- c("Lasso", "Best Subset")

kable(df_thres, row.names = TRUE, booktabs = T, escape = T,
      caption = "Specifity and Sensitivity at Thresholds, and AUC") %>%
  kable_styling(full_width = FALSE,
                latex_options = c("striped", "HOLD_position"))

# Lasso: variable selection
lasso_coef1 <- lasso(df_44_train[[1]][,-c(1,2)])
lasso_coef2 <- lasso(df_44_train[[2]][,-c(1,2)])
lasso_coef3 <- lasso(df_44_train[[3]][,-c(1,2)])
lasso_coef4 <- lasso(df_44_train[[4]][,-c(1,2)])
lasso_coef5 <- lasso(df_44_train[[5]][,-c(1,2)])

lasso_coef <- cbind(lasso_coef1, lasso_coef2, lasso_coef3,
                    lasso_coef4, lasso_coef5)
avg_coefs_lasso_44 <- apply(lasso_coef, 1, mean)

pred_lasso_44 <- plogis(x_vars_44 %*% avg_coefs_lasso_44)
```

```r
# Best subset: variable selection
L0Learn_coef1 <- L0Learn_func(df_44_train[[1]][,-c(1,2)])
L0Learn_coef2 <- L0Learn_func(df_44_train[[1]][,-c(1,2)])
L0Learn_coef3 <- L0Learn_func(df_44_train[[1]][,-c(1,2)])
L0Learn_coef4 <- L0Learn_func(df_44_train[[1]][,-c(1,2)])
L0Learn_coef5 <- L0Learn_func(df_44_train[[1]][,-c(1,2)])

L0Learn_coef <- cbind(L0Learn_coef1, L0Learn_coef2, L0Learn_coef3,
                      L0Learn_coef4, L0Learn_coef5)
avg_coefs_L0Learn_44 <- apply(L0Learn_coef, 1, mean)

pred_L0Learn_44 <- plogis(x_vars_44 %*% avg_coefs_L0Learn_44)

# Variable selection results
round(cbind(avg_coefs_lasso_44, avg_coefs_L0Learn_44), 4) %>%
  kbl(caption = "Variable Selection for 44-Week Model According to Lasso and Best Subset",
      col.names = linebreak(c("Lasso","Best Subset")),
      row.names = TRUE, booktabs = TRUE, escape = TRUE, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'))

# Lasso -> mixed effects model
pred_lasso_mixed_44 <- vector("list", 5)

for (i in 1:5) {
  mod_lasso <- glmer(Y ~ bw + birth_hc + del_method + prenat_ster + com_prenat_ster
                     + gender + ventilation_support_level.36
                     + inspired_oxygen.36 + peep_cm_h2o_modified.36 + med_ph.36
                     + weight_today.44 + ventilation_support_level_modified.44
                     + inspired_oxygen.44 + peep_cm_h2o_modified.44 + med_ph.44
                     + hosp_dc_ga + (1 | center),
                     family = binomial, data = df_44_train[[i]],
                     control=glmerControl(optimizer="bobyqa",
                                          optCtrl=list(maxfun=2e5)))

  pred_lasso_mixed_44[[i]] <- predict(mod_lasso, df_44_test[[i]], type = "response")
}
# Lasso-44 weeks
# ROC
levels(df_44_test[[i]]$Y) = c(0, 1)
roc_lasso_44 <- roc(predictor = pred_lasso_mixed_44[[1]],
                    response = df_44_test[[i]]$Y,
                    levels = c(0, 1), direction = "<")
# coords(roc_lasso_44, "best")
p_roc_lasso_44 <- ggroc(roc_lasso_44) +
  geom_text(aes(0.25, 0.15, label = paste0("AUC = ",
            paste(unlist(round(roc_lasso_44$auc,4)))))), size=3) +
  theme_grey(base_size = 10)

# Calibration
num_cuts <- 10
calib_data <-  data.frame(prob = pred_lasso_mixed_44[[1]],
                          bin = cut(pred_lasso_mixed_44[[1]],
                                    breaks = num_cuts),
                          class = as.numeric(df_44_test[[i]]$Y)-1)
calib_data <- calib_data %>%
```

```r
  group_by(bin) %>%
  summarise(observed = sum(class)/n(),
            expected = sum(prob)/n(),
            se = sqrt(observed*(1-observed)/n())))

p_calib_lasso_44 <- ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color="red") +
  geom_errorbar(aes(x = expected,
                    ymin = observed - 1.96*se,
                    ymax = observed + 1.96*se),
                colour="black", width=.01)+
  geom_point(aes(x = expected, y = observed)) +
  labs(x = "Expected Proportion", y = "Observed Proportion") +
  theme_grey(base_size = 10)

# Best subset -> mixed effects model
pred_bs_mixed_44 <- vector("list", 5)

for (i in 1:5) {
  mod_bs <- glmer(Y ~ prenat_ster + ventilation_support_level.36 + inspired_oxygen.36
                  + weight_today.44 + ventilation_support_level_modified.44
                  + peep_cm_h2o_modified.44 + med_ph.44 + hosp_dc_ga + (1 | center),
                  family = binomial, data = df_44_train[[i]],
                  control=glmerControl(optimizer="bobyqa",
                                       optCtrl=list(maxfun=2e5)))

  pred_bs_mixed_44[[i]] <- predict(mod_lasso, df_44_test[[i]], type = "response")
}

# BS-44 weeks
# ROC
levels(df_44_test[[i]]$Y) = c(0, 1)
roc_bs_44 <- roc(predictor = pred_bs_mixed_44[[1]],
                 response = df_44_test[[i]]$Y,
                 levels = c(0, 1), direction = "<")
# coords(roc_bs_44, "best")
p_roc_bs_44 <- ggroc(roc_bs_44) +
  geom_text(aes(0.25, 0.15, label = paste0("AUC = ",
               paste(unlist(round(roc_bs_44$auc,4)))), size=3) +
  theme_grey(base_size = 10)

# Calibration
num_cuts <- 10
calib_data <-  data.frame(prob = pred_bs_mixed_44[[1]],
                          bin = cut(pred_bs_mixed_44[[1]],
                                    breaks = num_cuts),
                          class = as.numeric(df_44_test[[i]]$Y)-1)
calib_data <- calib_data %>%
  group_by(bin) %>%
  summarise(observed = sum(class)/n(),
            expected = sum(prob)/n(),
            se = sqrt(observed*(1-observed)/n())))

p_calib_bs_44 <- ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color="red") +
  geom_errorbar(aes(x = expected,
```

```r
                          ymin = observed - 1.96*se,
                          ymax = observed + 1.96*se),
                  colour="black", width=.01)+
    geom_point(aes(x = expected, y = observed)) +
    labs(x = "Expected Proportion", y = "Observed Proportion") +
    theme_grey(base_size = 10)

ggarrange(p_roc_lasso_44, p_calib_lasso_44,
          p_roc_bs_44, p_calib_bs_44, ncol = 2, nrow = 2)

brier_lasso <- c()
brier_bs <- c()
auc_lasso <- c()
auc_bs <- c()
thres_lasso <- c()
thres_bs <- c()
spec_lasso <- c()
spec_bs <- c()
sens_lasso <- c()
sens_bs <- c()
for (i in 1:5) {
  brier_lasso <- c(brier_lasso, mean((pred_lasso_mixed_44[[i]] -
                                      (as.numeric(df_44_test[[i]]$Y)-1))^2))
  brier_bs <- c(brier_bs, mean((pred_bs_mixed_44[[i]] -
                                (as.numeric(df_44_test[[i]]$Y)-1))^2))
  auc_lasso <- c(auc_lasso, roc_lasso_44$auc)
  auc_bs <- c(auc_bs, roc_bs_44$auc)
  thres_lasso <- c(thres_lasso, coords(roc_lasso_44, "best")$threshold)
  thres_bs <- c(thres_bs, coords(roc_bs_44, "best")$threshold)
  spec_lasso <- c(spec_lasso, coords(roc_lasso_44, "best")$specificity)
  spec_bs <- c(spec_bs, coords(roc_bs_44, "best")$specificity)
  sens_lasso <- c(sens_lasso, coords(roc_lasso_44, "best")$sensitivity)
  sens_bs <- c(sens_bs, coords(roc_bs_44, "best")$sensitivity)
}

# Table: Brier scores, AUC, thresholds, specificity, and sensitivity
df_thres <- data.frame(
  "Brier Scores" = round(c(mean(brier_lasso), mean(brier_bs)), 4),
  AUC = round(c(mean(auc_lasso), mean(auc_bs)), 4),
  Threshold = round(c(mean(thres_lasso), mean(thres_bs)), 4),
  Specificity = round(c(mean(spec_lasso), mean(spec_bs)), 4),
  Sensitivity = round(c(mean(sens_lasso), mean(sens_bs)), 4))
rownames(df_thres) <- c("Lasso", "Best Subset")

kable(df_thres, row.names = TRUE, booktabs = TRUE, escape = TRUE,
      caption = "Specifity and Sensitivity at Thresholds, and AUC") %>%
  kable_styling(full_width = FALSE,
                latex_options = c("striped", "HOLD_position"))

cor(df[,c("bw", "blength", "weight_today.36", "weight_today.44")],
    use = "complete.obs") %>%
  kbl(caption = "Correlation Matrix",
      col.names = linebreak(c("Birth Weight","Birth Length",
                             "Weight at Week 36", "Weight at Week 44")),
      row.names = TRUE, booktabs = TRUE, escape = TRUE, align = "c") %>%
  kable_styling(full_width = FALSE,
```

```
            latex_options = c('striped', 'HOLD_position'))
```