Predicting Severe COVID-19 Cases in Unvaccinated Patients

Contents

List of Figures

Abstract

COVID-19 has created a global health crisis. This study focuses on predicting severe COVID-19 symptoms in unvaccinated patients using data from Auckland General Hospital (460 patients, August-October 2022). It also examines the impact of general practitioner availability on severe cases in West Auckland (December 2021-April 2022) using a GLM. Using Bayesian, resampling, and regression methods, the study assesses weight differences by sex and whether weight and age predict illness duration. Quasi-Poisson and Negative Binomial regression models address overdispersion and estimate severe case numbers, aiding hospital staff in resource allocation and patient care.

Keywords: COVID-19, Bayesian Analysis, Quasi-Poisson Regression, Negative Binomial Regression, Severe Symptoms Prediction, New Zealand

Chapter 1

Introduction

During the peak months of August to October 2022, Auckland General Hospital operated at 95% capacity due to an influx of unvaccinated COVID-19 patients. The Hospital Board is concerned about the severity of symptoms that extend patients' stays .

This research aims to analyze the risk factors associated with severe COVID-19 symptoms and propose a predictive model for severe cases. The factors considered include patient weight, age, duration since testing positive, presence of diabetes, and sex.
The research is structured into:

1. **Statistical Analysis:** This section provides a detailed descriptive analysis, focusing on objectives such as comparing weights between females and males, and evaluating 'weight' and 'age' as predictors of 'duration' using Frequentist and Probabilistic approaches

2. **Main Analysis:** This section addresses the Hospital Board's primary concern of predicting severe COVID-19 symptoms in unvaccinated patients using Logistic models, Quantile regression, and Generalized Additive Models

3. **Evaluation of ROC Curves:** This section examines the limitations of ROC curves in evaluating model performance. .

4. **Expected Severe COVID-19 Cases:** This section forecasts severe COVID-19 cases using a separate dataset, discussing a predictive model and its assumptions

The results of this research aim to provide insights for the Hospital Board to allocate resources more effectively in managing severe COVID-19 cases.

Chapter 2

Objectives

1. **General Objective**
   This analysis aims to address the concerns of the Auckland Hospital's Hospital Board by proposing a model that predicts whether a patient with COVID-19 will develop severe symptoms and thus require hospital admission.

2. **Specific Objectives**

   - Using Resampling and Bayesian methods to analyze whether females weigh more than males on average.
   - Using Ordinary Linear Regression and Bayesian methods to determine whether 'weight' and 'age' are good predictors of 'duration'.
   - To fit various models (logistic, quantile, GAMs) and determine the "winner" model.
   - Discussing the downsides of ROC curves.
   - Determining the expected number of severe COVID-19 cases in West Auckland and understanding the effect of GPs in the area over the severe cases.
   - Discussing model assumptions.

Chapter 3

Statistical Analysis

3.1 Descriptive Statistics

- **Weight:** The average weight of the individuals in the dataset is 62.72 kgs with a standard deviation of 8.89 kgs. The minimum recorded weight is 47.04 kgs while the maximum is 77.98 kgs.
- **Age:** With a standard deviation of 10.06 years, the persons' average age is 47.62 years. This indicates that the dataset contains a significant amount of age variability. The eldest person is 64.94 years old while the youngest is 30.10.
- **Duration:** On average, 11.30 days have elapsed since the patients tested positive for COVID-19 with a range from 5 to 18 days and a standard deviation of 3.69 days.

Descriptive statistics

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
weight	460	62.7244565	8.8862452	47.0400000	55.2250000	62.6800000	70.2900000	77.9800000
age	460	47.6179130	10.0564095	30.1000000	39.2250000	47.9950000	56.7800000	64.9400000
duration	460	11.2978261	3.6930712	5.0000000	8.0000000	11.0000000	14.0000000	18.0000000

Figure 3.1: Descriptive Statistics

Frequency Distributions

- **Sex:** The dataset consists of two sex categories.Males consists of 223 individuals accounting for 48.48% , while 237 individual females making up 51.52% of the sample.
- **Diabetes:** Out of the total sample, 324 individuals (70.43%) do not have diabetes while 136 individuals (29.57%) have diabetes. This indicates that most individuals in the dataset do not suffer from diabetes.
- **COVID Severity:** Regarding the severity of COVID-19, 191 individuals (41.52%) are categorized as not having severe COVID-19, whereas 269 individuals (58.48%) are classified as having severe COVID-19 indicating that more than half of the individuals in the sample have experienced severe COVID-19.

Descriptive statistics

The FREQ Procedure

covidsever	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	191	41.52	191	41.52
1	269	58.48	460	100.00

diabetes	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	324	70.43	324	70.43
1	136	29.57	460	100.00

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	223	48.48	223	48.48
1	237	51.52	460	100.00

Figure 3.2: Frequency Distributions

3.2 Estimated Average Difference in Male and Female Weights

Resampling Methods

- **Randomisation test**

Hypotheses considered: H_0 as the mean weight of both the sexes are same and H_1 it is different

The observed difference in means (Females - Males) is approximately -0.46 kg. This indicates that, on average, females weigh slightly less than males in this dataset.
The p-value is 0.5532, so we do not have enough statistical evidence to assert that the mean weight is different between females and males i.e. we fail to reject H_0 due to the high p-value.

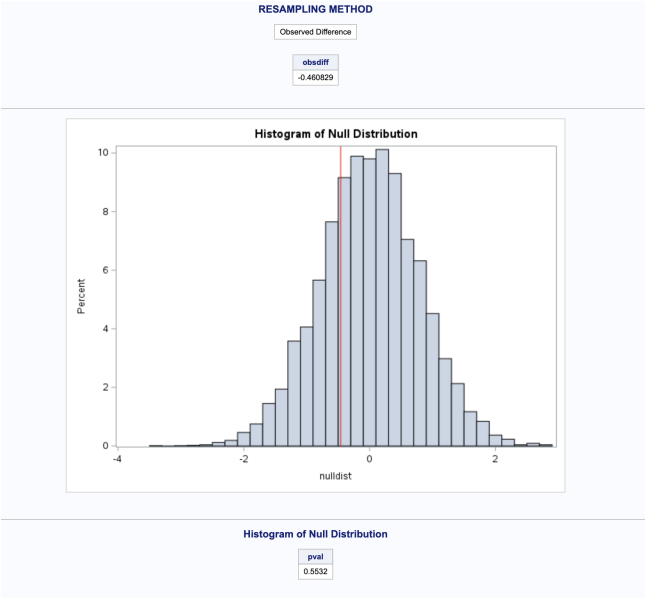


Figure 3.3: Randomisation test Output

- **Bootstrap method**

  From this method we would expect that most of the mean weight differences fluctuate between -2.092624 and 1.1684276 kg approximately (95 percent confidence interval)
  Since this interval includes 0, it supports the conclusion that there is no significant difference in mean weight between females and males.

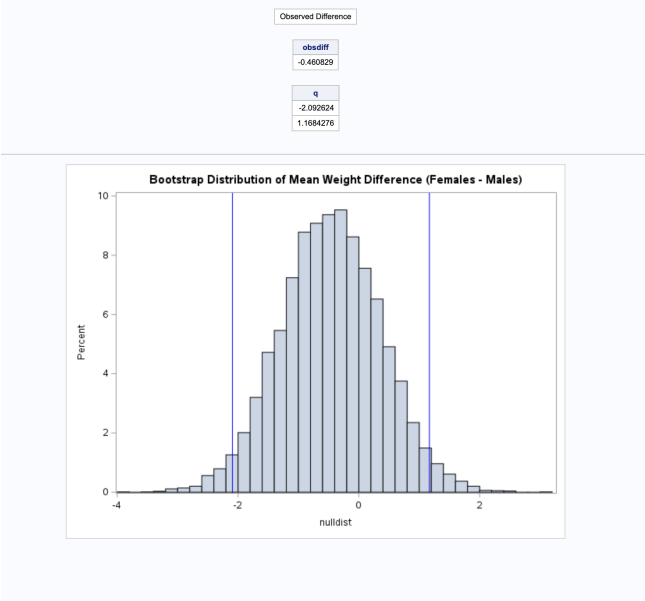


Figure 3.4: Bootstrap Output

Bayesian Method

Likelihood:

- The response variable weight is assumed to follow a normal distribution.
- For an individual i, the weight is modeled as:

$$\text{weight}_i \sim \text{Normal}(\mu_i, \sigma^2)$$

- The mean μ_i is modeled as:

$$\mu_i = \beta_0 + \beta_1 \cdot \text{gender}_i$$

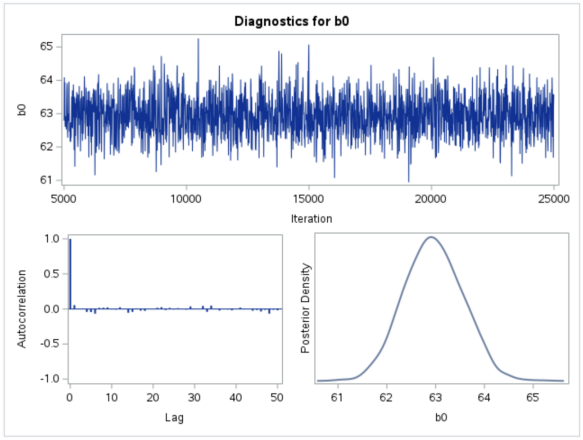
Priors:

- Priors for the regression coefficients (β_0, β_1) are set to normal distributions with mean 0 and a large variance (diffuse priors):

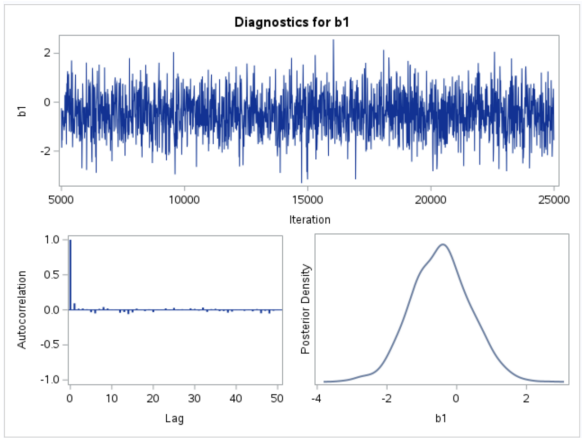
$$\beta_0, \beta_1 \sim \text{Normal}(0, 1000000)$$

- Prior for the variance parameter (σ^2) is set to an inverse gamma distribution:

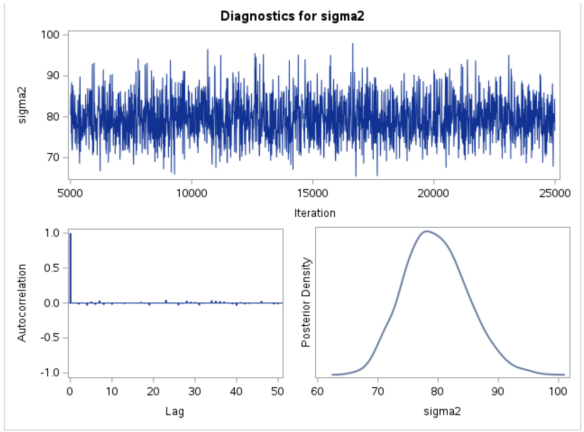
$$\sigma^2 \sim \text{Inverse-Gamma}(0.01, 0.01)$$



(a) β_0



(b) β_1



(c) σ^2

Figure 3.5: Traceplots of β_0, β_1, σ^2

Bayesian analysis

Prob	CIs weight_pred1	weight_pred2
0.025	61.322346	61.811711
0.975	63.656813	64.093852

Figure 3.6: Credible Interval

The effective sample sizes of the variables are more than 200, confirming that we have enough samples to characterize the posterior distributions.

From the credible intereval, the average weight of females is between 61.322346 and 63.656813 and for males it is between 61.811711 and 64.093852 for 95% probability.
Overlapping credible intervals indicate that there is no significant difference in the mean weights between females and males.

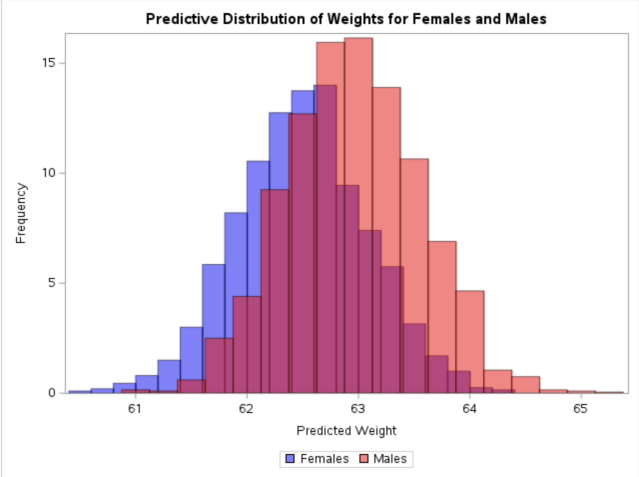


Figure 3.7: Predictive Distribution of weights

Thus from both the methodologies we can confirm that the females do not weigh more than the males on average.

3.3 Patient's 'weight' and 'age' as predictors of 'duration'

Ordinary Linear Regression

R-square and Adjusted R-Squared: The R-Square value is 0.0017 indicating that only 0.17% of the variability in the duration can be explained by the weight and age of the patients. The Adjusted R-Square is negative suggesting that the model does not fit the data well.
The F-value is 0.38 with a p-value of 0.6856 indicating that the model is not statistically significant indicating the predictors (weight and age) do not have a significant relationship with the duration.
Checking Assumptions:

- Residuals: The residuals are randomly scattered around zero suggesting that there is no clear pattern, which is a good sign for the assumptions of linear regression.
- QQ Plot and Histogram: The Q-Q plot follows a straight line indicating that the residuals are approximately normally distributed, and histogram appears symmetric and bell-shaped which aligns with the assumption of normally distributed errors.

Bayesian Linear Regression

β_0**:** The expected value of the duration when both predictors, weight and age, are equal to zero. The posterior mean estimate of β_0 is 10.6289 with a standard deviation of 1.4379.

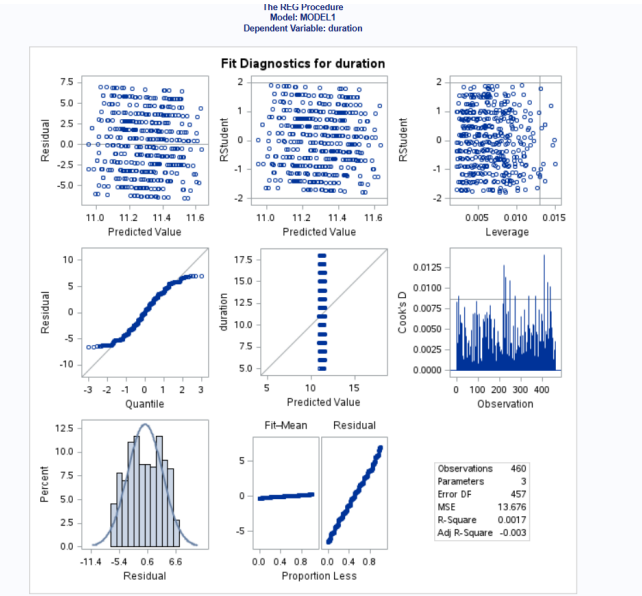


Figure 3.8: Diagnostic plot for OLS model

β_1**:** The expected change in the duration for a one-unit increase in weight, holding age constant. The posterior mean is 0.0155 with a standard deviation of 0.0198. The 95% interval [-0.0239, 0.0519] includes zero. This means there is no strong evidence to suggest that changes in weight have a substantial impact on the duration according to this model.

β_2**:** The expected change in the duration for one-unit increase in age, holding weight constant. The posterior mean is -0.00649 with a standard deviation of 0.0169. The 95% interval [-0.0389, 0.0255] includes zero, indicating that age is not a significant predictor of duration.

Markov Chains: The Markov chains for all the parameters look stationary, i.e., they fluctuate around a constant with a stable variance, which implies a good mix.
Effective Sample Size: Since the ESS of all are >200 we have enough samples to characterize the posterior distributions of these parameters.

Difference between both the approaches

The OLS model provides clear point estimates and p-values, revealing a poor model fit with an R-Square of only 0.17% and non-significant predictors. In contrast, the Bayesian MCMC model offers a richer interpretation by providing posterior distributions and 95% Highest Posterior Density (HPD) intervals, which confirm the non-significance of weight and age with greater detail on parameter uncertainty.

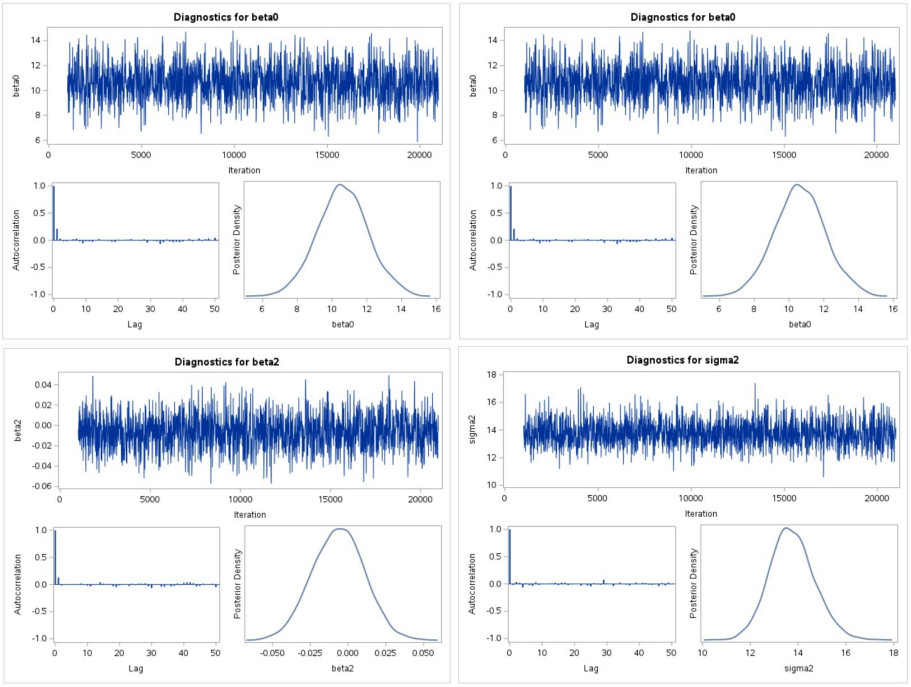


Figure 3.9: Trace plot

Bayesian MCMC for Predicting Duration

The MCMC Procedure

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
beta0	2000	10.6289	1.4379	8.0138	13.7579
beta1	2000	0.0155	0.0198	-0.0239	0.0519
beta2	2000	-0.00649	0.0169	-0.0389	0.0255
sigma2	2000	13.7469	0.9212	11.9598	15.5702

Bayesian MCMC for Predicting Duration

The MCMC Procedure

Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
beta0	1404.0	1.4245	0.7020
beta1	1330.5	1.5032	0.6652
beta2	1629.5	1.2273	0.8148
sigma2	2266.1	0.8826	1.1331

Figure 3.10: Result of Bayesian

Chapter 4

Main Analysis: Predicting Covid Severe Cases

4.1 Logistic Regression Model with Interactions

Estimated model from the results:

$$\log\left(\frac{p}{1-p}\right) = -42.6631 + 0.3443 \cdot \text{weight} + 0.3130 \cdot \text{duration} + 0.2577 \cdot \text{age} + 14.4745 \cdot \text{sex} + 11.0552 \cdot \text{diabetes} - 0.18$$

where sex $= 0$ and diabetes $= 0$ are the reference levels.
Keeping other factors constant,
For a diabetic female (*diabetes* = 1, *sex* = 1):

$$\log\left(\frac{p}{1-p}\right) = -17.1334 + 0.3443 \cdot \textit{weight} + 0.3130 \cdot \textit{duration} + 0.0761 \cdot \textit{age} \tag{4.1}$$

For a diabetic male (*diabetes* = 1, *sex* = 0):

$$\log\left(\frac{p}{1-p}\right) = -31.0079 + 0.3443 \cdot \textit{weight} + 0.3130 \cdot \textit{duration} + 0.0761 \cdot \textit{age} \tag{4.2}$$

For a non-diabetic female (*diabetes* = 0, *sex* = 1):

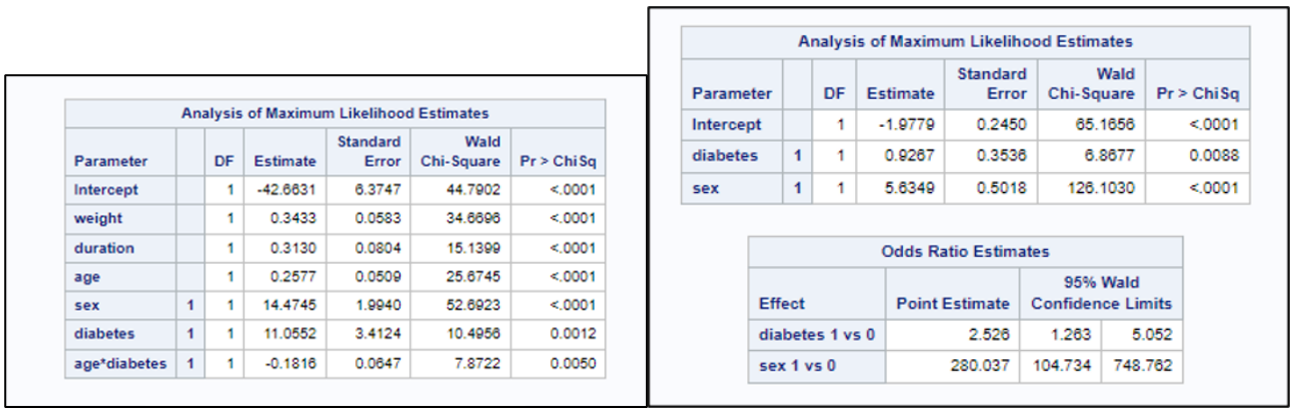
$$\log\left(\frac{p}{1-p}\right) = -27.5886 + 0.3443 \cdot \textit{weight} + 0.3130 \cdot \textit{duration} + 0.2577 \cdot \textit{age} \tag{4.3}$$

For a non-diabetic male (*diabetes* = 0, *sex* = 0, reference level):

$$\log\left(\frac{p}{1-p}\right) = -42.0631 + 0.3443 \cdot \textit{weight} + 0.3130 \cdot \textit{duration} + 0.2577 \cdot \textit{age} \tag{4.4}$$

Highlights

- A 1 kg increase in the weight of a patient increases the odds of developing severe symptoms vs not developing by 41%.
- When the number of days since the patient tested positive increases by 1, the odds of developing severe COVID-19 symptoms increase by 36.752%.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-42.6631	6.3747	44.7902	<.0001
weight		1	0.3433	0.0583	34.6696	<.0001
duration		1	0.3130	0.0804	15.1399	<.0001
age		1	0.2577	0.0509	25.6745	<.0001
sex	1	1	14.4745	1.9940	52.6923	<.0001
diabetes	1	1	11.0552	3.4124	10.4956	0.0012
age*diabetes	1	1	-0.1816	0.0647	7.8722	0.0050

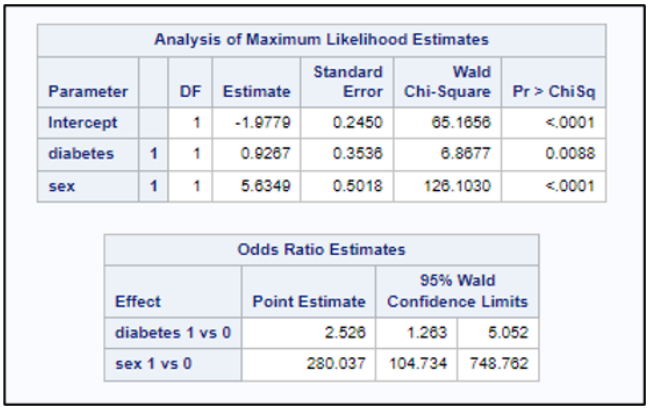
Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.9779	0.2450	65.1656	<.0001
diabetes	1	1	0.9267	0.3536	6.8677	0.0088
sex	1	1	5.6349	0.5018	126.1030	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
diabetes 1 vs 0	2.526	1.263	5.052
sex 1 vs 0	280.037	104.734	748.762

Figure 4.1: Result of logistics Model



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.9779	0.2450	65.1656	<.0001
diabetes	1	1	0.9267	0.3536	6.8677	0.0088
sex	1	1	5.6349	0.5018	126.1030	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
diabetes 1 vs 0	2.526	1.263	5.052
sex 1 vs 0	280.037	104.734	748.762

Figure 4.2: Model for Binary data

- A year increase in the age of a patient increases the odds of developing severe COVID-19 symptoms by 29.395%.
- For a diabetic patient, a year increase in age increases the odds of developing severe COVID-19 symptoms by 7.907% more than a non-diabetic patient.
- For a female, the odds of developing severe COVID-19 symptoms are 19328371% higher than for a male.
- For a diabetic patient, the odds of developing severe COVID-19 symptoms are 6306869% higher than for a non-diabetic patient.

Removing sex and diabetes from the above model would result in a lower accuracy due to their statistical significance. Due to the unusually high effect of sex and diabetes on the severity of COVID-19 symptoms, we used a separate model for them.

Estimated model for just diabetes and sex:

$$\log\left(\frac{p}{1-P}\right) = -1.9779 + 0.9267 \cdot \text{diabetes} + 5.6349 \cdot \text{sex}$$

where sex $= 0$ and diabetes $= 0$ are the reference levels.

Keeping other factors constant, the odds of developing severe COVID-19 symptoms vs not developing them, for a diabetic patient, are 2.526 times more than they are for a non-diabetic patient, and 280.037

times more for a female than for a male.
Significance: The P-values for all the parameters in both the models are < 0.05. This means that we have enough evidence to suggest that their effect on the odds of developing severe COVID-19 symptoms vs not developing is statistically significant.

4.2 Quantile regression

Quantile regression focuses on estimating the conditional quantiles of a response variable distribution through the effect of the explanatory variables. This approach does not fit our objectives since our explanatory variable only has two outcomes.

4.3 Generalized additive models (GAMs)

GAMs capture non-linear relationships between explanatory and response variables, revealing complex patterns beyond linear models. We will use this model to predict severe COVID-19 odds with continuous predictors as independent variables.

The regression analysis shows insufficient evidence for a linear effect of duration and age, but with a P-value < 0.05, weight is not relevant linearly.

Deviance analysis indicates non-significant non-parametric effects for all three variables, lacking evidence for a non-parametric relationship with severe COVID-19 odds.

Neither smoothing component is statistically significant due to large P-values, indicating insufficient evidence for their contribution to model fit.

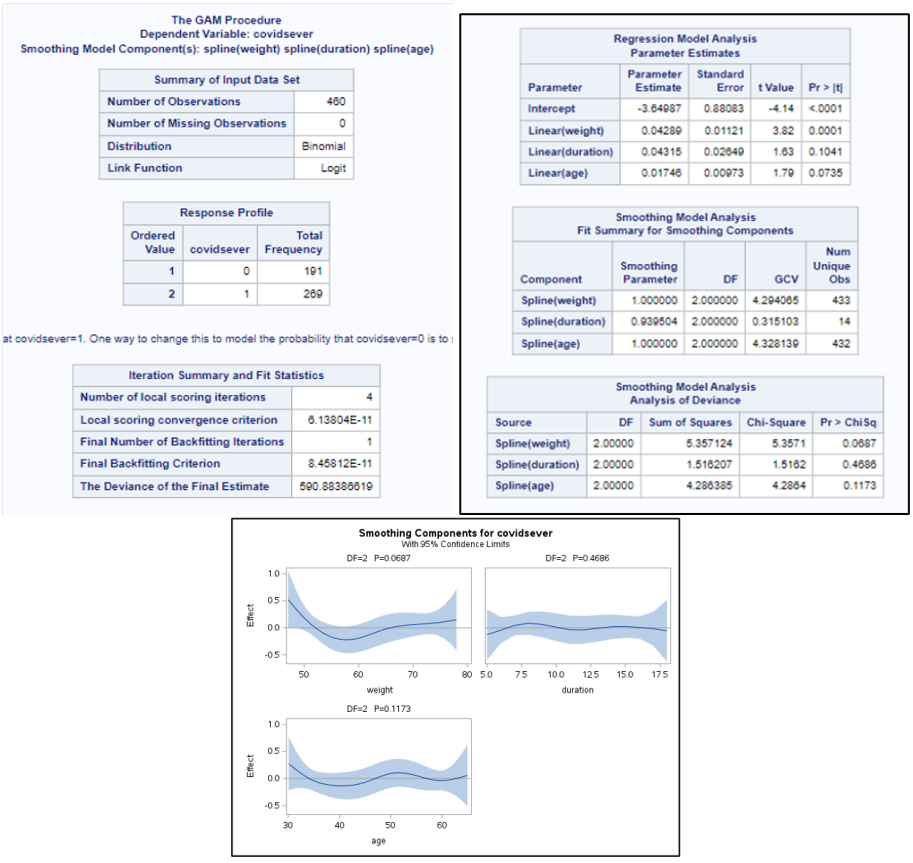


Figure 4.3: GAMs Model Output

Chapter 5

ROC CURVE

A Receiver Operating Characteristic (ROC) curve is used to evaluate the performance of binary classification models but it has some downside also, especially in our case the data is highly imbalanced which results in giving inaccurate ROC results.

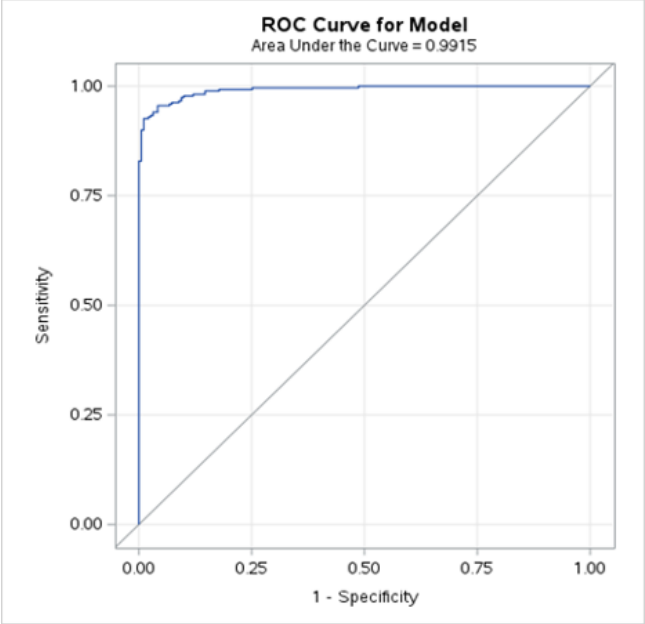


Figure 5.1: ROC Curve

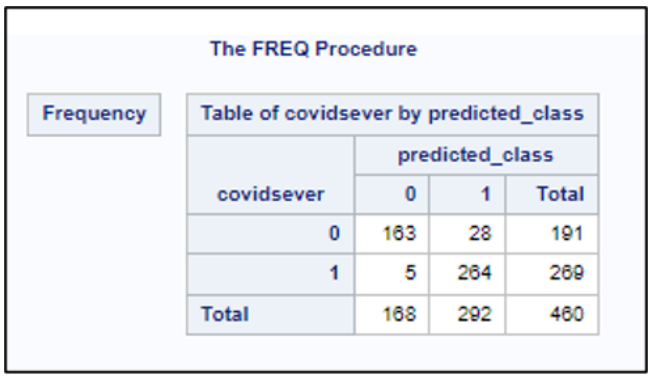
5.1 Limitations of ROC Curves

Cost and Benefits:
ROC curves give the same weight to the false negatives and the false positives. It does not consider the cost or benefit of the different types of errors. In the context of the problem, we might want to assign different weights to the false positives and false negatives, depending on the seriousness. Considering that the cost involves the potential loss of life due to severe COVID-19 cases, or unnecessary hospitalization, this might lead to huge problems.
Not Useful for Multi-Class Problems:
It may not reflect the true performance of our classifier, and this can be costly. ROC curves also do not work great with multi-class problems with more than two outcomes, which is generally a more realistic scenario.

Imbalanced Dataset:
If the dataset is imbalanced (more non-severe cases than severe cases), ROC curves may give a misleading performance. In this case, Precision-Recall (PR) curves could be more useful because they focus on the performance of the minority class.
Threshold Selection:
Selecting the optimal threshold is a crucial task in identifying severe cases, which may have implications for resource allocation in a hospital setting. Methods such as F1 score or using cost-sensitive analysis could help determine the optimal threshold.
It is clear that ROC curves should be used along with other evaluation methods and should not be the only tool used to measure the performance of a model.

5.2 Confusion Matrices

The optimal threshold was calculated using Youden's Index1.
True Positive (TP): The model correctly predicts 264 severe COVID-19 cases.
True Negatives (TN): The model correctly predicts 163 non-severe COVID-19 cases.
False Positives (FP): The model incorrectly predicts 28 cases as severe when they were not severe.
False Negative (FN): The model incorrectly predicted 5 cases are non-severe when they were severe.
Table to compare accuracy, sensitivity, specificity, and precision of the model:



The FREQ Procedure

Frequency

Table of covidsever by predicted_class

covidsever	predicted_class 0	1	Total
0	163	28	191
1	5	264	269
Total	168	292	460

Figure 5.2: Confusion Matrix

Metric	Value
Accuracy (TP+TN/Total)	92.82%
Sensitivity (TP/TP+FN)	98.14%
Specificity (TN/TN+FP)	85.34%
Precision (TP/TP+FP)	90.41%

Table 5.1: Performance Metrics

Chapter 6

Expected Covid Severe Cases with GPs' count

We are using the number of GPs as a predictor variable to estimate its effect on the number of severe cases.
The model can be formulated as follows:

$$SevereCases \sim GPS$$

$$SevereCases = \beta_0 + \beta_1 \times GPS + \epsilon_i, \quad \text{where GPS = 2, 3, 4, 5}$$

We will fit both Poisson regression and Negative binomial regression models to compare their performance.

Poisson Regression

Analysis Of Maximum Likelihood Parameter Estimates

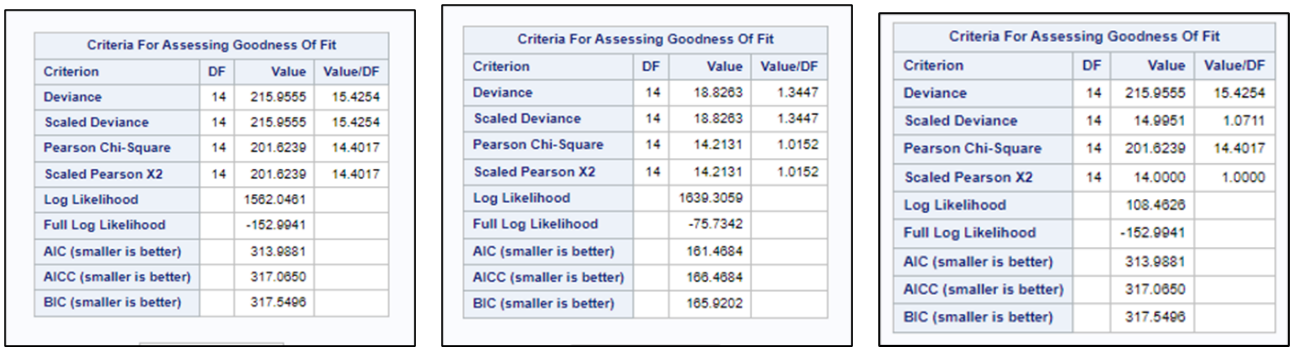
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.1744	0.0877	4.0025	4.3463	2265.32	<.0001
gps	2	1	-1.7177	0.1904	-2.0909	-1.3444	81.36	<.0001
gps	3	1	-1.0245	0.1218	-1.2632	-0.7858	70.75	<.0001
gps	4	1	-0.4435	0.1054	-0.6502	-0.2369	17.70	<.0001
gps	5	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Figure 6.1: Poisson regression results

Results suggest that 2 GPs reduce the expected severe cases by 82.053% more than when there are 5 GPs. 3 GPs reduce the expected severe cases by 64.102% more than when there are 5 GPs, and 4 GPs reduce the expected severe cases by 35.821%. All the P-values are < 0.05, implying towards the fact that the number of GPs has a statistically significant relationship with severe COVID-19 cases.

Overdispersion
Since the variance of the response factor (severe COVID-19 cases) is significantly larger than the mean, it can be said that there is presence of overdispersion. The scaled Pearson Coefficient in our estimated model is 14.4017, which means the data is over-dispersed. Hence to solve this, we should consider a Negative-Binomial regression model, which has more flexibility towards modelling the relationship between the conditional variance and the conditional mean as opposed to the Poisson regression model. The parameters for both the models are the same, however, a small Scaled Pearson X2 suggests that the overdispersion is handled



Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	14	215.9555	15.4254
Scaled Deviance	14	215.9555	15.4254
Pearson Chi-Square	14	201.6239	14.4017
Scaled Pearson X2	14	201.6239	14.4017
Log Likelihood		1562.0461	
Full Log Likelihood		-152.9941	
AIC (smaller is better)		313.9881	
AICC (smaller is better)		317.0650	
BIC (smaller is better)		317.5496	

(a) Poisson

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	14	18.8263	1.3447
Scaled Deviance	14	18.8263	1.3447
Pearson Chi-Square	14	14.2131	1.0152
Scaled Pearson X2	14	14.2131	1.0152
Log Likelihood		1639.3059	
Full Log Likelihood		-75.7342	
AIC (smaller is better)		161.4684	
AICC (smaller is better)		166.4684	
BIC (smaller is better)		165.9202	

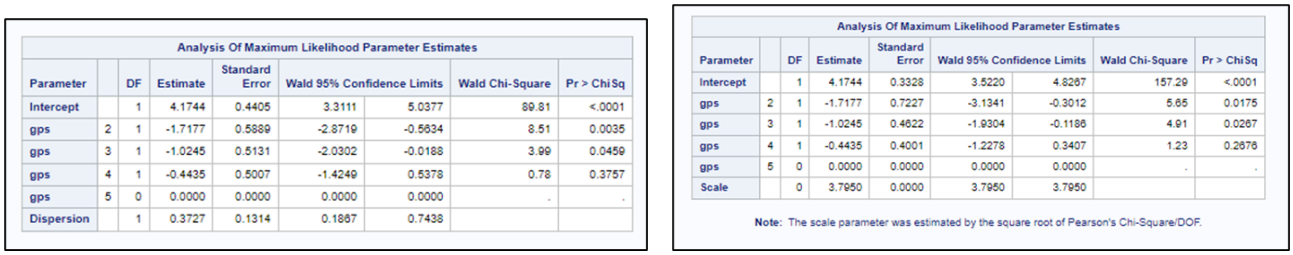
(b) Negative binomial

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	14	215.9555	15.4254
Scaled Deviance	14	14.9951	1.0711
Pearson Chi-Square	14	201.6239	14.4017
Scaled Pearson X2	14	14.0000	1.0000
Log Likelihood		108.4626	
Full Log Likelihood		-152.9941	
AIC (smaller is better)		313.9881	
AICC (smaller is better)		317.0650	
BIC (smaller is better)		317.5496	

(c) Quasi Poisson

Figure 6.2: Assessing Goodness of fit



Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.1744	0.4405	3.3111	5.0377	89.81	<.0001
gps	2	1	-1.7177	0.5889	-2.8719	-0.5634	8.51	0.0035
gps	3	1	-1.0245	0.5131	-2.0302	-0.0188	3.99	0.0459
gps	4	1	-0.4435	0.5007	-1.4249	0.5378	0.78	0.3757
gps	5	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.3727	0.1314	0.1867	0.7438		

(a) Negative binomial

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.1744	0.3328	3.5220	4.8267	157.29	<.0001
gps	2	1	-1.7177	0.7227	-3.1341	-0.3012	5.65	0.0175
gps	3	1	-1.0245	0.4622	-1.9304	-0.1186	4.91	0.0267
gps	4	1	-0.4435	0.4001	-1.2278	0.3407	1.23	0.2676
gps	5	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	3.7950	0.0000	3.7950	3.7950		

Note: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

(b) Quasi Poisson

Figure 6.3: Maximum likelihood parameter estimate

Negative Binomial or Quasi-Poisson?

While a quasi-poisson and negative binomial models address the overdispersion, the P-values in the NB model are quite high, making the parameters statistically insignificant. Relatively, the P-values for quasi-poisson model are smaller and the parameters are almost significant. when comparing the AIC and BIC values NB appears to be better model. Hence, to formulate an answer for the HB regarding the expected number of patients developing severe COVID-19 symptoms, we recommend using the NB Model due to its ability to handle overdispersion effectively. More GPs are likely to lead to better detection and reporting of severe cases. While the number of severe cases detected increases, it does not necessarily imply that the actual incidence of severe cases are rising. Instead, it suggests that more cases are being identified and reported due to increased medical surveillance and capacity.

Chapter 7

Discussion

- The analysis revealed no statistically significant difference in average weight between males and females, indicated by overlapping credible intervals. Both ordinary and Bayesian linear regressions showed that weight and age are not significant predictors of duration. Thus, gender does not influence average weight, and neither weight nor age reliably predict duration .
- Diabetic patients are 2.5 times more likely to develop severe COVID-19 symptoms compared to non-diabetics. Females are 280 times more likely to experience severe symptoms than males. These findings align with Sugiyama et al. and Kumar et al., which link comorbidities like diabetes to severe COVID-19 symptoms.
- Both Poisson and Negative Binomial regression models indicate that the number of GPs significantly impacts identifying severe COVID-19 cases. Due to overdispersion, the NB model is more appropriate for this dataset.
- The analysis suggests that increasing the number of GPs significantly increase detection of severe COVID-19 cases. Specifically, having 2 GPs can reduce severe cases by approximately 82.1% compared to having 5 GPs, with similar reductions observed with 3 and 4 GPs.

Bibliography

[1] Manoj P. & Abhinav J. (2016). ROC Curve: Making way for correct diagnosis. *SP11 - PharmaSUG 2016*, 7-9. https://www.pharmasug.org/proceedings/2016/SP/PharmaSUG-2016-SP11.pdf

[2] Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?

[3] David C. & Giuseppe J. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*. https://biodatamining.biomedcentral.com/articles/10.1186/s13040-023-00322-4

[4] Eur R. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *National Library of Medicine*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4356897/

[5] Masaya S. (2023). Tools and factors predictive of the severity of COVID-19. *National Library of Medicine*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10130545/#:~:text=Their%20meta%2Danalysis%20indicated%20the,ALT)%2C%20creatinine%2C%20and%20lactate

[6] Gangopadhyay KK. (2020). Does having diabetes increase chances of contracting COVID-19 infection?. *PubMed Central*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7263205/

[7] SAS Support. (2021). FAQ: How do I compare ROC curves? https://support.sas.com/kb/41/364.html

[8] Doe, John. (2015). Analyzing Receiver Operating Characteristic Curves. *NCBI*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4356897/

[9] Jane Smith. (2016). Advanced Statistical Methods in SAS. *PharmaSUG 2016*. https://www.pharmasug.org/proceedings/2016/SP/PharmaSUG-2016-SP11.pdf

[10] Rick Wicklin. (2018). Compare ROC Curves in SAS. https://blogs.sas.com/content/iml/2018/11/14/compare-roc-curves-sas.html

[11] Rick Wicklin. (2010). Statistical Programming with SAS/IML Software. SAS Institute Inc.

[12] Mithat Gönen. (2006). Analyzing Receiver Operating Characteristic Curves with SAS. SAS Institute Inc.

[13] SAS Institute. (1997). Advanced Techniques in SAS Programming. *SUGI 22*. https://support.sas.com/resources/papers/proceedings/proceedings/sugi22/POSTERS/PAPER219.PDF

[14] Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data? National Marine Mammal Laboratory, mailto:jay.verhoef@noaa.gov, mailto:peter.boveng@noaa.gov

Appendix

7.1 SAS CODES AND OUTPUTS

7.1.1 STATISTICAL ANALYSIS

```
/*************************************PART1************************
   *************/
/* Load the dataset */
proc import datafile="/home/u63817297/STAT802_project/ResearchTask-
   Team7_614652376.csv"
    out=covid_data
    dbms=csv
    replace;
    getnames=yes;
run;

/* Print the dataset */
proc print data=covid_data(obs=10);
run;

TITLE 'Descriptive statistics  ';
/* Descriptive statistics */
proc means data=covid_data n mean std min p25 median p75 max;
    var weight age duration;
run;

/* Frequency table for binary variables */
proc freq data=covid_data;
    tables covidsever diabetes sex;
run;
/*************************************PART1- A*********************
   ***************/
/* Separate the data into males and females */
proc sql;
    create table males as
    select weight
    from covid_data
    where sex = 0;

    create table females as
```

```
    select weight
    from covid_data
    where sex = 1;
quit;
TITLE 'RESAMPLING METHOD ';
/* Randomization test for independent samples */
/*       H_0 : The mean weight of females is the same as the mean
   weight of males.
         H_1 : The mean weight of females is different from the
            mean weight of males.*/

PROC IML;
/* Load the data */
use males;
read all var {weight} into M;
use females;
read all var {weight} into F;

/* Calculate the observed difference in means */
obsdiff = mean(F) - mean(M);
print "Observed Difference", obsdiff;

/* Combine the data */
alldata = F // M;
N1 = nrow(F);  /* Number of females */
N2 = nrow(M);  /* Number of males */
N = N1 + N2;   /* Total number of observations */
NRepl = 9999;  /* Number of permutations */
nulldist = j(NRepl, 1);  /* Allocate vector to hold permuted
   differences */

/* Perform permutation resampling */
do k = 1 to NRepl;
   x = sample(alldata, N, "WOR");  /* Permuting the data without
      replacement */
   nulldist[k] = mean(x[1:N1]) - mean(x[(N2+1):N]);  /* Mean
      difference */
end;

/* Generate a histogram of the null distribution */
title "Histogram of Null Distribution";
refline = "refline " + char(obsdiff) + " / axis=x lineattrs=(color=
   red);";
call Histogram(nulldist) other=refline;

/* Calculate the p-value */
pval = (1 + sum(abs(nulldist) >= abs(obsdiff))) / (NRepl+1);
print pval;
quit;
```

```
76
77  /*The p-value associated with the observed difference is
       approximately 0.5851
78  A high p-value indicates that there is no significant evidence to
       reject the null hypothesis.*/
79
80  /* Bootstrap to estimate the difference in mean weights */
81  proc iml;
82  /* Load the data */
83  use males;
84  read all var {weight} into M;
85  use females;
86  read all var {weight} into F;
87
88  /* Calculate the observed difference in means */
89  obsdiff = mean(F) - mean(M);
90  print "Observed Difference", obsdiff;
91
92  /* Set the parameters for bootstrap resampling */
93  N1 = nrow(F);
94  N2 = nrow(M);
95  NRepl = 10000;  /* Number of bootstrap samples */
96  call randseed(12345);  /* Set random seed */
97
98  /* Allocate vector to hold bootstrap mean differences */
99  nulldist = j(NRepl, 1);
100
101 /* Perform bootstrap resampling */
102 do k = 1 to NRepl;
103     x1 = sample(F, N1);  /* Bootstrap sample for females */
104     x2 = sample(M, N2);  /* Bootstrap sample for males */
105     nulldist[k] = mean(x1[1:N1]) - mean(x2[1:N2]);  /* Mean
           difference */
106 end;
107
108 /* Calculate the 2.5th and 97.5th percentiles for the confidence
       interval */
109 p = {0.025, 0.975};
110 call qntl(q, nulldist, p);  /* Compute quantiles */
111 print q;
112
113 /* Generate a histogram of the bootstrap distribution */
114 title "Bootstrap Distribution of Mean Weight Difference (Females -
       Males)";
115 refline = "refline " + char(q[1:2]) + " / axis=x lineattrs=(color=
       blue);";
116 call Histogram(nulldist) other=refline;
117
118 quit;
```

```
119
120 /*The observed difference in means (Females - Males) is
       approximately -0.461 kg.
121 This indicates that, on average, females weigh slightly less than
       males in this dataset.
122 The p-value is 0.5851, so we fail to reject  H_0 .
123 We do not have enough statistical evidence to assert that the mean
       weight is different between females and males.
124 i.e., when  H_0  is true it would be unlikely to observe a
       difference of -0.461 kg
125 and we would expect that most of the mean weight differences
       fluctuate between -2.092624 and 1.1684276 kg approximately.
126 The 95% confidence interval for the mean difference, given by the
       2.5th and 97.5th percentiles of the bootstrap distribution, is
       approximately (-2.092624, 1.1684276).
127 Since this interval includes 0, it supports the conclusion that
       there is no significant difference in mean weight between
       females and males.*/
128
129
130 TITLE 'Bayesian analysis ';
131 /* Bayesian analysis using PROC MCMC */
132 ods graphics on;
133 proc mcmc data=covid_data nmc=20000 thin=10 nbi=5000 seed=1 monitor
       =(b0 b1 sigma2)
134     outpost=weight_mcmc;
135     /*Parameters: (b0, b1) Regression coefficients (sigma2)
          variance.*/
136     parms b0 b1 sigma2;
137     /*Priors: Diffuse normal priors for b0 and b1*/
138     prior b0 ~ normal(0, var=1000000);
139     prior b1 ~ normal(0, var=1000000);
140     /*inverse gamma prior for sigma2*/
141     prior sigma2 ~ igamma(shape=0.01, scale=0.01);
142     /*likelihood*/
143     mu = b0 + b1*sex;
144     model weight ~ normal(mu, var=sigma2);
145 run;
146
147 /* Generate predictive distributions */
148 data topred;
149     input sex;
150     datalines;
151     1
152     0
153 ;
154 run;
155
156 data weight_mcmc_pred;
```

```
    set weight_mcmc;
    if _N_ = 1 then do;
        array sex_vals[2] _temporary_ (1 0);
    end;
    array weight_pred[2];
    do i = 1 to 2;
        weight_pred[i] = b0 + b1*sex_vals[i];
    end;
    output;
    keep weight_pred1 weight_pred2 b0 b1 sigma2;
run;

/* 95% Credible intervals for predictions */
proc iml;
    varNames = {"weight_pred1" "weight_pred2"};
    use weight_mcmc_pred;
    read all var varNames into X;
    close weight_mcmc_pred;

    Prob = {2.5, 97.5} / 100;   /* prob in (0,1) */
    call qntl(CIs, X, Prob);
    print Prob CIs[c=varNames];
run;

/* Plot the predictive distribution */
proc sgplot data=weight_mcmc_pred;
    histogram weight_pred1 / transparency=0.5 fillattrs=(color=blue
       ) legendlabel="Females";
    histogram weight_pred2 / transparency=0.5 fillattrs=(color=red)
        legendlabel="Males";
    title "Predictive Distribution of Weights for Females and Males
       ";
    xaxis label="Predicted Weight";
    yaxis label="Frequency";
run;
/*The overlapping credible intervals indicate that
 there is no significant difference in the mean weights between
    females and males.
 The means of these distributions suggest that
 the average weight for males is slightly higher than that for
    females,
 but the credible intervals overlap considerably,
 indicating that the difference is not statistically significant.*/


/*************************************PART1 B*********************
   ***************/
TITLE 'Linear Regression Analysis';
/* Fit ordinary least squares (OLS) regression model */
```

```
proc reg data=research_task;
    model duration = weight age;
    title "ordinary least squares (OLS) regression model";

run;
TITLE 'Bayesian analysis ';
proc mcmc data=research_task outpost=posterior nmc=20000 thin=10
   seed=42;
    parms beta0 0 beta1 0 beta2 0 sigma2 1;
    prior beta0 beta1 beta2 ~ normal(0, var=1000);
    prior sigma2 ~ igamma(0.01, scale=0.01);

    model duration ~ normal(beta0 + beta1*weight + beta2*age, var=
       sigma2);
    title "Bayesian MCMC for Predicting Duration";
run;
```

Bayesian analysis

The MCMC Procedure

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
b0	2000	62.9540	0.5973	61.8603	64.1283
b1	2000	-0.4618	0.8467	-2.0237	1.2543
sigma2	2000	79.5827	5.1983	70.3457	90.1259

Bayesian analysis

The MCMC Procedure

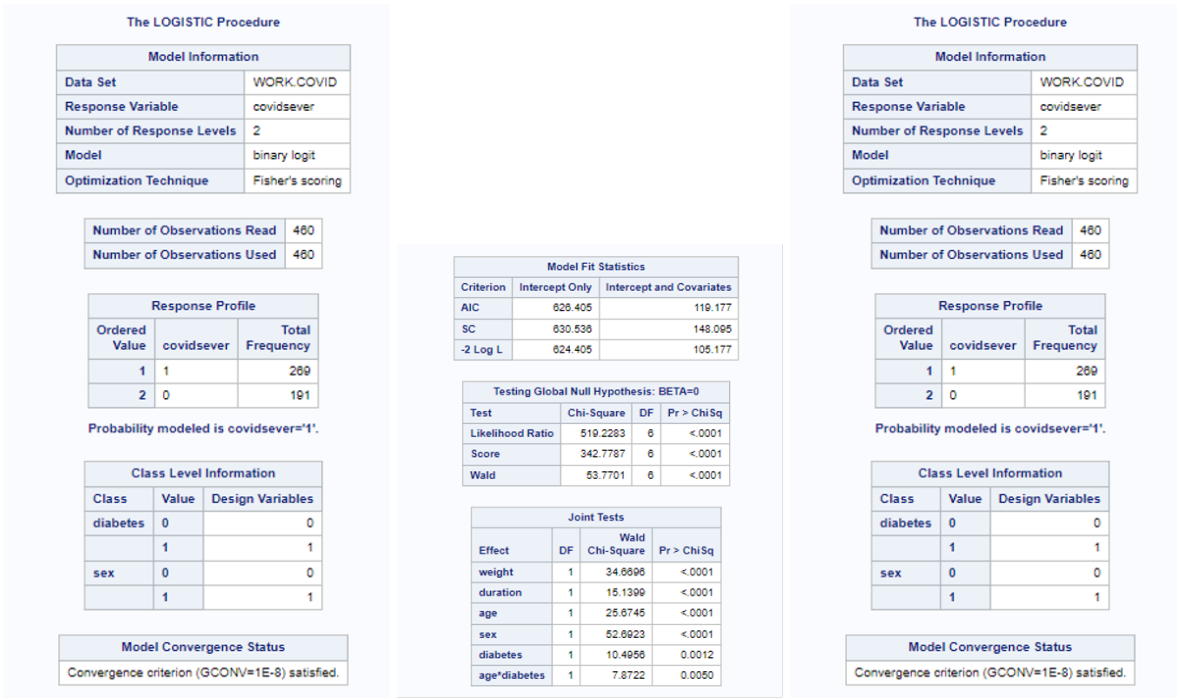
Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
b0	1802.8	1.1094	0.9014
b1	1592.1	1.2562	0.7960
sigma2	2000.0	1.0000	1.0000

Figure 7.1: Bayesian Analysis

7.1.2 ANALYSIS 1

```
/*main analysis*/
/* Logistic Model */
 PROC LOGISTIC DATA=stand_covid descending ;
 CLASS diabetes sex / param=ref ref=first ;
MODEL covidsever = weight duration age sex diabetes age|diabetes /
   outroc=rocdata;
ROC; ROCCONTRAST;
OUTPUT OUT=predicted P=pred; /* Save the predicted probabilities */
RUN;
```



The LOGISTIC Procedure

Model Information	
Data Set	WORK.COVID
Response Variable	covidsever
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	460
Number of Observations Used	460

Response Profile		
Ordered Value	covidsever	Total Frequency
1	1	269
2	0	191

Probability modeled is covidsever='1'.

Class Level Information		
Class	Value	Design Variables
diabetes	0	0
	1	1
sex	0	0
	1	1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	626.405	119.177
SC	630.536	148.095
-2 Log L	624.405	105.177

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	519.2283	6	<.0001
Score	342.7787	6	<.0001
Wald	53.7701	6	<.0001

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
weight	1	34.6696	<.0001
duration	1	15.1399	<.0001
age	1	25.6745	<.0001
sex	1	52.6923	<.0001
diabetes	1	10.4956	0.0012
age*diabetes	1	7.8722	0.0050

The LOGISTIC Procedure

Model Information	
Data Set	WORK.COVID
Response Variable	covidsever
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	460
Number of Observations Used	460

Response Profile		
Ordered Value	covidsever	Total Frequency
1	1	269
2	0	191

Probability modeled is covidsever='1'.

Class Level Information		
Class	Value	Design Variables
diabetes	0	0
	1	1
sex	0	0
	1	1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Figure 7.2: Logistic procedure

Determining optimal cut-off (Youden's Index) and generating confusion matric /* Determining optimal (where J is maximum) cut-off point */

```
DATA CUTOFF;
 SET ROCDATA;
_SPECIF_ = (1 - _1MSPEC_);
 LOGIT=LOG(_PROB_/(1-_PROB_));
 CUT_POINT=(LOGIT + 1.914)/14.744;
 J = _SENSIT_ + _SPECIF_ - 1;
 D = SQRT ((1-_SENSIT_)**2 + (1-_SPECIF_)**2);
 DIFF= ABS (_SENSIT_ - _SPECIF_);
RUN;

/* Max value of J is 0.9151794 */
PROC MEANS data=cutoff MAX;
VAR J;
RUN;
```

```
15
16 /* Confusion matrix */
17 DATA predicted;
18     SET predicted;
19     predicted_class = (pred >= 0.20357); /* You can adjust the
          threshold as needed */
20 RUN;
21
22 /* Generating Confusion Matrix */
23 PROC FREQ DATA=predicted;
24     TABLES covidsever*predicted_class / NOCOL NOROW NOPERCENT CHISQ
          ;
25 RUN;
```

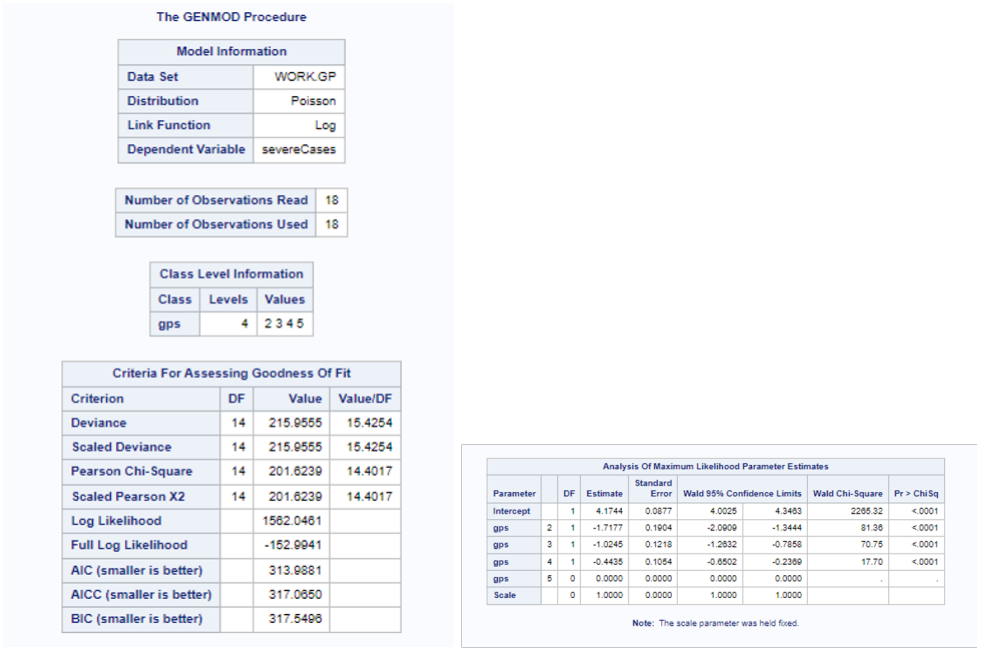
GAMs Model for Continuous Predictors:

```
1 PROC GAM DATA = covid plots = components(clm commonaxes);
2 MODEL covidsever (EVENT = '1') = spline(weight, df = 3) spline(
     duration, df = 3) spline(age, df = 3)    / dist = binary;
3 RUN;
```

7.1.3 ANALYSIS 2

Poisson Regression

```
/* Poisson Model */
PROC GENMOD DATA=gp PLOTS=all;
CLASS gps;
MODEL severeCases = gps / dist = poisson link = log;
RUN;
```



The GENMOD Procedure

Model Information	
Data Set	WORK.GP
Distribution	Poisson
Link Function	Log
Dependent Variable	severeCases

Number of Observations Read	18
Number of Observations Used	18

Class Level Information		
Class	Levels	Values
gps	4	2 3 4 5

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	14	215.9555	15.4254
Scaled Deviance	14	215.9555	15.4254
Pearson Chi-Square	14	201.6239	14.4017
Scaled Pearson X2	14	201.6239	14.4017
Log Likelihood		1562.0461	
Full Log Likelihood		-152.9941	
AIC (smaller is better)		313.9881	
AICC (smaller is better)		317.0650	
BIC (smaller is better)		317.5496	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.1744	0.0877	4.0025	4.3463	2265.32	<.0001
gps	2	1	-1.7177	0.1904	-2.0909	-1.3444	81.36	<.0001
gps	3	1	-1.0245	0.1218	-1.2632	-0.7858	70.75	<.0001
gps	4	1	-0.4435	0.1054	-0.6502	-0.2369	17.70	<.0001
gps	5	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Figure 7.3: Poisson Regression

Negative Binomial Regression

```
/* Negative Binomial Model */
PROC GENMOD DATA=gp PLOTS=all;
CLASS gps;
MODEL severeCases = gps / dist = negbin link = log;
RUN;
```

The GENMOD Procedure

Model Information	
Data Set	WORK.GP
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	severeCases

Number of Observations Read	18
Number of Observations Used	18

Class Level Information		
Class	Levels	Values
gps	4	2 3 4 5

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	14	18.8263	1.3447
Scaled Deviance	14	18.8263	1.3447
Pearson Chi-Square	14	14.2131	1.0152
Scaled Pearson X2	14	14.2131	1.0152
Log Likelihood		1639.3059	
Full Log Likelihood		-75.7342	
AIC (smaller is better)		161.4684	
AICC (smaller is better)		166.4684	
BIC (smaller is better)		165.9202	

Analysis Of Maximum Likelihood Parameter Estimates

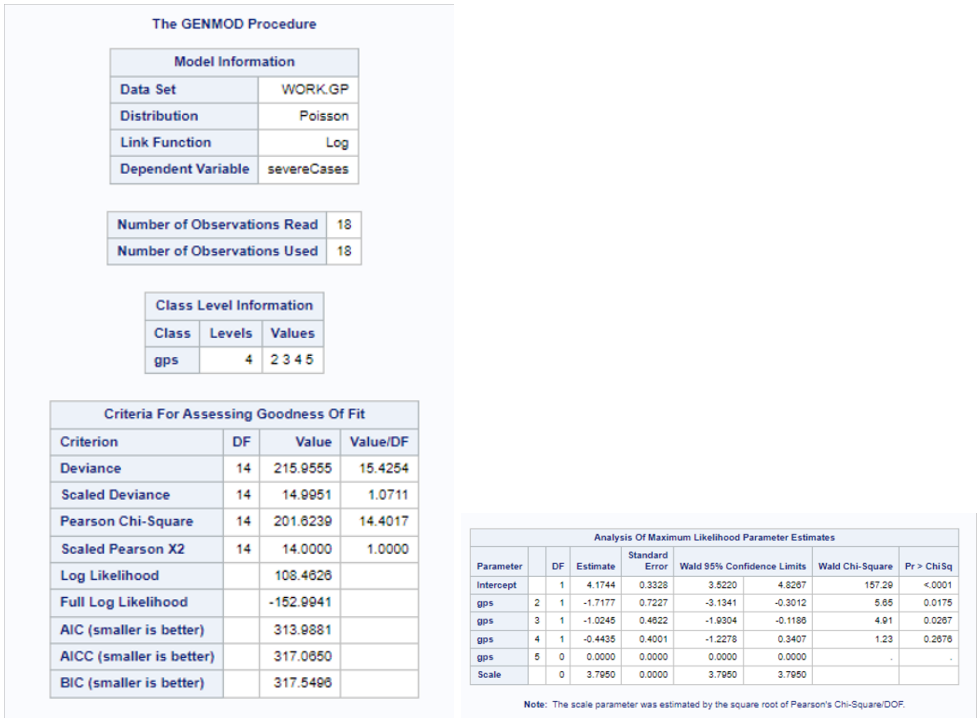
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.1744	0.4405	3.3111	5.0377	89.81	<.0001
gps	2	1	-1.7177	0.5889	-2.8719	-0.5634	8.51	0.0035
gps	3	1	-1.0245	0.5131	-2.0302	-0.0188	3.99	0.0459
gps	4	1	-0.4435	0.5007	-1.4249	0.5378	0.78	0.3757
gps	5	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.3727	0.1314	0.1867	0.7438		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

Figure 7.4: Negative Binomial Regression

Quasi-Poisson Model

```
/* Quasi-Poisson Model */
PROC GENMOD DATA=gp PLOTS=all;
CLASS gps;
MODEL severeCases = gps / dist = poisson link = log;
SCALE=PEARSON;
RUN;
```



The GENMOD Procedure

Model Information	
Data Set	WORK.GP
Distribution	Poisson
Link Function	Log
Dependent Variable	severeCases

Number of Observations Read	18
Number of Observations Used	18

Class Level Information		
Class	Levels	Values
gps	4	2 3 4 5

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	14	215.9555	15.4254
Scaled Deviance	14	14.9951	1.0711
Pearson Chi-Square	14	201.6239	14.4017
Scaled Pearson X2	14	14.0000	1.0000
Log Likelihood		108.4626	
Full Log Likelihood		-152.9941	
AIC (smaller is better)		313.9881	
AICC (smaller is better)		317.0650	
BIC (smaller is better)		317.5496	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.1744	0.3328	3.5220	4.8267	157.29	<.0001
gps	2	1	-1.7177	0.7227	-3.1341	-0.3012	5.65	0.0175
gps	3	1	-1.0245	0.4622	-1.9304	-0.1186	4.91	0.0267
gps	4	1	-0.4435	0.4001	-1.2278	0.3407	1.23	0.2676
gps	5	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	3.7950	0.0000	3.7950	3.7950		

Note: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

Figure 7.5: Quasi-Poisson Model