UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics

Master's thesis

The possibility to predict gestational hypertensive disorders

Eugene Ebong Ebong
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :
dr. Ruth NYSEN

SUPERVISOR :
Dr. Dorien LANSSENS

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2019
2020

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

The possibility to predict gestational hypertensive disorders

Eugene Ebong Ebong

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

dr. Ruth NYSEN

SUPERVISOR :

Dr. Dorien LANSSENS

The possibility to predict gestational hypertensive disorders.

Biostatistics Master Thesis
2019-2020

Universiteit Hasselt

Supervisor:
Dr. Ruth Nysen

Supervisor:
Dr. Dorien Lanssens

EBONG EUGENE EBONG
Thesis presented in fulfillment of the requirements for the degree of
Master of Statistics

Submission Date: June 15, 2020

Dedication

This thesis is dedicated to my late mother Ma. Ebong Anastacia Muke. Your teachings and doctrines shall live in us all through our lives. Rest in peace mum. We love you but the Almighty loves you more.

Acknowledgement

It is a pleasure to thank many people who made this thesis possible. This work would not have been possible without the support and help of them.
First and foremost, I would like to thank the almighty GOD for having made everything possible.
Secondly, I would like to express my sincere gratitude to my supervisors Dr. Ruth Nysen and Dr. Dorien Lanssens, for their feedbacks, dedication and continuous undiluted support during this educationally and emotionally trying time in my life. They always made time out of their busy schedules to tend to my worries and were patient and motivate me when I was in a really dark place.

I also would like to extend my heartfelt gratitude to my entire family, especially my brothers, Kebulu Divine Ndode, Ebong Leslie Enang and Ebong Dieudonne Ndelle, my sister Ebote Emelda Ndelle for their prayers, support, encouragement and sponsorship. May the almighty God reward you guys abundantly.

Lastly but not the least, I would like to thank all my professors at Universiteit Hasselt for their support to expand my knowledge all through these years. And to my friends (my personal support system), I am extremely grateful for standing by me all through this period and the assistance I received in one way or another. My God bless you.

Eugene Ebong, Ebong
June 15, 2020
Hasselt, Belgium

Abstract

Background: Gestational hypertensive disorders (GHD) remain one of the most significant and intriguing unsolved problems in obstetrics and it is estimated to cause about 10% of maternal and fetal morbidity and mortality.

Objective: The aim of this project is to determine the possibility to predict the occurrence of GHD in pregnant women given some baseline measurements and mainly by studying the trend in the repeated measurements of the blood pressures (systolic (SBP) and diastolic (DBP) blood pressures) of the women in the early stage of their pregnancy (gestational age (GA) $\leq$ 27 weeks).

Methodology: The data for this analysis comes from a clinical research done in a hospital setting (Ziekenhuis Oost-Limburg (ZOL), Genk, in cooperation with the Mobile Health Unit of Universiteit Hasselt (UHasselt)) selecting pregnant women who were at risk for gestational hypertensive disorders. The outcome measure was the diagnosis of any form of gestational hypertension for the women categorised into three levels for this analysis (normal pregnancy, gestational hypertension (GH) and pre-eclampsia (PE)). The two-stage model was the best modeling approach to study the trend in the blood pressure because of their longitudinal nature while the outcome was cross sectional. In the first-stage, the SBP and DBP measurements were modeled using a quadratic regression model to obtain the patient specific estimates. These estimates were used as covariates in the second-stage model together with other variables such as age and body mass index (BMI) to model the dependent variable GHD.

Results and Conclusion: The two-stage model fitted to the data showed that the systolic blood pressure covariates in the second-stage model significantly increase the predicted probability of the GHD outcome. The overweight variable had a significant increase in their GHD predicted probability compared to obesity variable. The model classified 56.9% of the 160 women correctly. For classification per outcome diagnosis, 88% of the normal diagnosis group, 27% of the GH group and 9% of the PE were correctly classified.

Keywords: Gestational hypertensive disorders (GHD), systolic blood pressure (SBP), diastolic blood pressure (DBP), pre-eclampsia (PE), two-stage model, gestational age (GA), gestational hypertension (GH).

Contents

1 Introduction

1.1 Background

Gestational Hypertensive Disorders (GHD) remain one of the most significant and intriguing unsolved problems in obstetrics and it is estimated to cause about 10% of maternal and fetal morbidity and mortality. Gestational hypertension (GH) is defined as pregnancy-induced hypertension, measured two times with minimum six hours apart, after twenty weeks of gestation. When GH is accompanied by proteinuria (spot urine protein/creatinine ratio 30 mg/mmol or 300 mg/day or at least 1 g/L on dipstick testing), this condition is called pre-eclampsia (PE) [2]. A pregnant woman is said to be at risk or having GH when she has a systolic blood pressure (BP) $>$ 140mmHg and a diastolic BP $>$ 90mmHg

Pregnancy is the state of fertilization and development for one or more offspring within a woman's uterus. When a woman is pregnant, she needs pre-natal care and follow up especially when she is in the late stages of the second trimester going to the third trimester and preparing for birth. The traditional care they usually get is the conventional care (CC) but another method for medical management of pregnant women which has been around for a while is remote monitoring (RM), which dates back to the early 1990's and involves facilitating patients' management at home. The Pregnancy REmote MOnitoring study (PREMOM I) was designed to evaluate the benefit of the RM approach to medical management of pregnant women who are at risk or suffer from GHD.

The assessment of women with pregnancies complicated with or at risk for GHD includes a clinical follow-up, serological investigation, and fetal ultrasound evaluation. The type and frequency of follow-up depend on the kind and severity of the hypertensive disorder.
The goal of treatment is to prevent significant cerebrovascular and cardiovascular events in the mother, without compromising fetal well-being [4].

The most common management for GHD in Belgium is an admission to the prenatal observation unit for diagnostic and therapeutic follow-up before induction of labor or discharge at home. In severe cases, premature birth is indicated [5].

1.2 Research question

The PREMOM I study was a four years retrospective study done at Ziekenhuis Oost-Limburg (ZOL) Genk, a second level prenatal center in co-operation with the Mobile Health Unit of Universiteit Hasselt (UHasselt). It involved separating the study participants at risk for GHD to receiving either remote monitoring or conventional care. Some of the participants received remote monitoring on demand of the responsible obstetrician before admission or after discharge from the pre-natal observation ward. The criteria to initiate remote monitoring were GHD at gestational age $\geq$ 20 weeks where an intensive follow-up until delivery was desirable. Participants without a mobile phone, a gestational age less than 20 weeks, a fetus with congenital malformations, and those who refused informed consent were excluded and received conventional care [2].

The goal of this analysis is to be able to predict the occurrence of GHD by analysing the trend in the blood pressure measurements together with some collected patient characteristics such as age and the reason why they are considered to be at risk to develop GHD. Though the PREMOM I study was a four years retrospective study, this prediction was done by using only the data collected in the year 2015.

2 Description of the dataset

The 2015 data collected from the PREMOM I study was used for this prediction analysis with measurements up to the twenty-seventh week of gestational age. Gestational age (GA) is the time from the first day of the last menstrual cycle to the current date and it is measured in weeks. The participants in this analysis did not have the same gestational age at the time of inclusion into the study.

The data were received in three separate excel files. The first file as shown in table 1 contained the cross-sectional variables collected from the participants at the start of the study, and includes but not limited to premom identification number, the GA at inclusion, reasons why they were considered to be of high risk to develop any kind of GHD, weight, age and height. Another set of variables were collected at the time of diagnosis of the hypertension disorder or during birth. Additional participation data such as the number of systolic and diastolic blood pressures taken, the number of visits to the clinic, the number of phone call, the number of missed blood pressure measurements were also recorded though these are not shown.

The second file as shown in table 2 contained the repeated measurements of the blood pressures (SBP and DBP). It included variables like premom identification number, the day the measurement was taken (in a date-time format since several measurements were taken per day) and the measurement values.

The third data set contains information about other repeated measurements of the participants. These variables include premom identification number, weight measurements performed during the study, birth date, etc. From this data set, the weight measurements were extracted and merged with the data set for the SBP and DBP by the premom identification number and the date at which the measurement was recorded.

A complete data set used for this thesis was then obtained by merging all the sub-datasets on the patient identification number and excluding women who had no entry for the response variable (diagnosis of GHD). The weight measurements were not used in the model fitting because of very few or no observation for most of the women.

Table 1: Cross-sectional variables in the dataset

Variable	**Coding**	**Range**
Premom identification number	pat_id	Premom 03 - Premom 445
Diagnosis of Gestational hypertensive disorders	diag_gh	Normal Gestational hypertension Pre-eclampsia
Body mass index	BMI	17.57kg/m^2 - 45.71kg/m^2
Number of live births	Parity	Nulliparous Multiparous
Reason why the women were at high risk to develop GHD	RHR	Concomitant disorders GHD in (a) previous pregnancy(ies) GHD in current pregnancy Other
Age at inclusion	Age	18 years - 41years
Gestational age at the moment of inclusion	GA_inclusion	49days - 189days
Weight before pregnancy	Weight_inclusion	48kg - 140kg

Table 2: Repeated measured variables in the dataset

Variable	**Coding**	**Range**
Premom identification number	pat_id	Premom 03 - Premom 445
Systolic blood pressure	agg_systolic	62mmHg - 203mmHg
Diastolic blood pressure	agg_diastolic	43mmHg - 128.5mmHg
Weight	agg_weight	54kg - 199.3kg

3 Methods and Materials

3.1 Study Population

The PREMOM I study mainly did a preganancy outcome comparison between the participants in the remote monitoring and conventional care study groups [3, 4]. One hundred and sixty two women who were at risk of developing GHD and received remote monitoring were included in this analysis. Pregnant women participating in the prenatal remote follow-up program were given a blood pressure monitor, a weight scale and an activity tracker (Withings), which they used to perform one blood pressure measurement in the morning and one in the evening, one weight measurement a day, and track their activity throughout the day until delivery or hospital admission. These data were transmitted to an online platform which was built by the Mobile Health Unit of Universiteit Hasselt (UHasselt). [3].

It is worth mentioning that most of these participants deviated from protocol as they had more than two blood pressure measurements in some days and none on other days and same deviation was found with the weight measurements. These measurements were then aggregated using the median of the measurements on a daily and weekly basis. The end result was one blood pressure measurement for each day and for each week for each of the one hundred and sixty two women in the final data set. The activity tracker data was not of interest in this thesis and thus was not included since the main aim was to predict GHD by studying the trend in the blood pressure. Also, due to the fact that only twenty six of these participants had at least one weight measurement, this variable as mentioned earlier was not used in this thesis. The participants who did not have value for the diagnosis of the GHD were also dropped from the final data set used in the exploratory data analysis and subsequently in fitting the statistical two-stage model. Such, the final data set had a total of one hundred and sixty two participants.

3.2 Exploratory Data Analysis

This was done to get a better understanding of the trend in the SBP and DBP measurements, and also to get an insight on the distribution of key variables such as the gestational age (GA), reasons why the women were considered to be at high risk to develop GHD varies with the outcome variable (GHD). Summary statistical techniques such as frequency tables and individual profile plots were used to achieve the above analysis. Since several blood pressure measurements were recorded a day by the women, these daily measurements were aggregated using the median to have one measurement for the day. The daily median measurement was then used to study the trend in the blood pressures. Categorical variables in the dataset were coded as;

$$GHD = \begin{cases} 0, & \text{Normal outcome;} \\ 1, & \text{Gestational Hypertension (GH);} \\ 2, & \text{Pre-eclampsia (PE)} \end{cases}$$

$$Parity = \begin{cases} 0, & \text{if nulliparous;} \\ 1, & \text{if multiparous}(\geq 1)\ ; \end{cases} \qquad bmicat = \begin{cases} Healthy, & \text{if BMI} < 25; \\ Overweight, & \text{if } 25 \leq \text{BMI} < 30\ ; \\ Obese, & \text{if BMI} \geq 30; \end{cases}$$

$$RHR = \begin{cases} 1, & \text{Concomitant disorders;} \\ 2, & \text{GHD in (a) previous pregnancy(ies);} \\ 3, & \text{GHD in current pregnancy;} \\ 4, & \text{Other} \end{cases}$$

3.3 Statistical Methods

Taking into account that we have a multivariate response variable and the fact that we aimed to perform a prediction, a multinomial logistic regression model seem very plausible to be used in this context. But the data for this thesis is such that the dependent variable (diagnosis of GHD) is cross-sectional in nature while the blood pressures which we are interested to study their trend in order to be able to predict the occurrence of the dependent variable are repeatedly measured thus longitudinal in nature.

The methods for examining data with a longitudinal predictor and non-time-varying outcome do not fall in the standard concept of generalized linear mixed models as instead of having correlated outcome data, the repeated measures of the predictors are correlated [1]. There is no general consensus on how the information contained in the longitudinal exposure trajectory can be used in a multinomial regression model. Other traditional models that could be used for such analysis like ordinary linear regression and logistic regression also failed to work in this context.

Using the two-stage model, it is possible to incorporate at least two of these models to be able to study the trend in the blood pressures and be able to build a prediction model for the occurrence of GHD in the participants.

The general idea of the two-stage model is to perform a mixed-modeling where in the first-stage, the dependent variable is modeled as a function of time to obtain patient-specific estimates which are then used as the new dependent variables for the second-stage model regressed on the other predictors in the data set [7]. The modeling done in this thesis is based on the methodology as described in sections 3.3.1 and 3.3.2.

3.3.1 First-stage model

In this stage, the longitudinal time-varying blood pressure measurements (SBP and DBP) are first modeled using a quadratic regression model as a function of time in an attempt to capture the evolution of the patient-specific plots (figures 1 - 6) to obtain patient-specific extimates.

The formulation of this model is given as

$$SBP_i = s_{0i} + s_{1i}t_i + s_{2i}t_i^2 + \epsilon_{si} \tag{1}$$

$$DBP_i = d_{0i} + d_{1i}t_i + d_{2i}t_i^2 + \epsilon_{di} \tag{2}$$

where $i = 1, ..., 162$; s_{0i} and d_{0i} are the patient-specific intercept for the SBP and DBP models, s_{1i}, d_{1i} and s_{2i}, d_{2i} are the patient-specific estimates for the days and days-squared terms in the two first-stage models, $\epsilon_{si} \sim N(0, \sigma_s^2)$; $\epsilon_{di} \sim N(0, \sigma_d^2)$

3.3.2 Second-stage model

In the second-stage, the patient-specific estimates obtained from the first-stage model are used as predictors together with other predictors such as age and BMI. Since the outcome variable is ordinal in nature, a multinomial logistic regression model with a cumulative probit link function was used to fit the second-stage model. A major drawback of this model is that prediction uncertainty of the patient-specific estimates from the first-stage model is not accounted for in the second-stage analysis, which may lead to biased results. But it naturally accounts for between subject heterogeneity [1].

The formulation of this cumulative probit model is given as shown in equation 3;

$$\phi^{-1}[P(GHD_i \leq j|\boldsymbol{x})] = \alpha_j + \beta'\boldsymbol{x}, \quad j = 1, ..., J-1 \tag{3}$$

$$P(GHD_i \leq j|\boldsymbol{x}) = \phi(\alpha_j + \beta'\boldsymbol{x})$$

where GHD takes a value j if the i^{th} ordinal observations falls in the j^{th} category. $\boldsymbol{x}_j^T$ is a p-vector of regression variables for the parameters, β without a leading column for an intercept and ϕ^{-1} is the inverse of the cumulative density function of the standard normal distribution [N(0,1)]. The intercepts α_j are strictly ordered;

$$-\infty \equiv \alpha_0 \leq \alpha_1 \leq ... \leq \alpha_{J-1} \leq \alpha_J \equiv \infty$$

4 Results and Discussion

4.1 Exploratory Data Analysis

Frequency tables were produced for each level of the outcome variable (GHD) and for the whole dataset on when the first measurements were taken for the women and the total number of measurements the women had.

The women in the dataset had an age range from 18 to 41 years old with the mean of 30 years and median age of 31 years old. About half (52.5%) of the women had a normal diagnosis of GHD. The results obtained also showed that 33.3% were diagnosed gestational hypertension while 14.2% were found to have pre-eclampsia. These results are shown in tables 3 and 4 respectively.

Table 3: Distribution of Age

Minimum	**Mean**	**Median**	**Maximum**
18	30	31	41

Table 4: Distribution of gestational hypertensive disorders

Diagnosis	**Frequency**	**Percent**
Normal	85	52.5
Gestational hypertension	54	33.3
Pre-eclampsia	23	14.2
Total	162	100

As shown in table 5, the earliest blood pressure measurement was recorded on the 49^{th} day of gestation. The lates day of blood pressure measurement was recorded on the 189^{th} day of gestation. The median time for the first measurement recorded was around the 140^{th} day of gestation.

The total number of days with blood pressure measurements is shown in table 6. The minimum recorded was 1 day while the maximum was 140 days of measurement. While the mean and median number of days were 47 and 42 respectively. Two women had their total number of measurements less than 5 days, thus they were excluded from the analysis phase.

Table 5: Day of first BP measurement

Diagnosis	**Mininium**	**Mean**	**Median**	**Maximum**
Normal	56	134	140	186
Gestation hypertension	49	124	133	182
Pre-eclampsia	64	146	147	189

Table 6: Total number of days with BP measurements

Diagnosis	**Mininium**	**Mean**	**Median**	**Maximum**
Normal	4	41	36	115
Gestation hypertension	8	57	50	140
Pre-eclampsia	1	43	35	126

For each of the levels of the outcome variable (GHD), women were selected at random and their profiles for the systolic and diastolic blood pressures plotted as a function of the number of days of measurement. These plots as shown in figures 1 to 6 were used to study the trend of the blood pressure variables. They were noticeable variation of blood pressure measurement within each participant daily measurements. Overall, there was a decrease in trend but this decrease was prominent until the 125^{th} day after which it started increasing. This is clearly seen in figure 5 for the systolic blood pressure for women with a diagnosis of pre-eclampsia. The plots suggested the presence of a curvature and thus modeling the blood pressures in the first-stage model will require a polynomial term to capture this curvature. Using a quadratic time effect term in the stage-one models easily captured this trend in the blood pressure measurement.

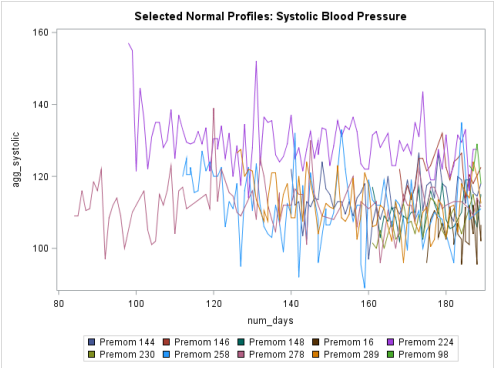

Figure 1: SBP Normal profiles

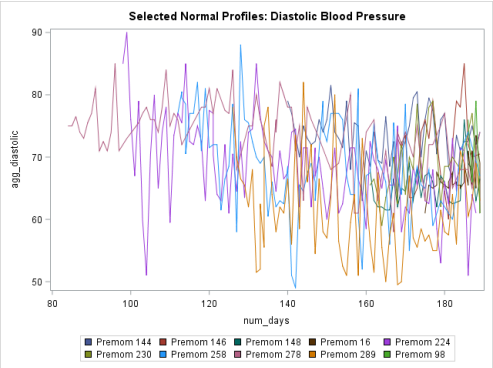

Figure 2: DBP Normal profiles

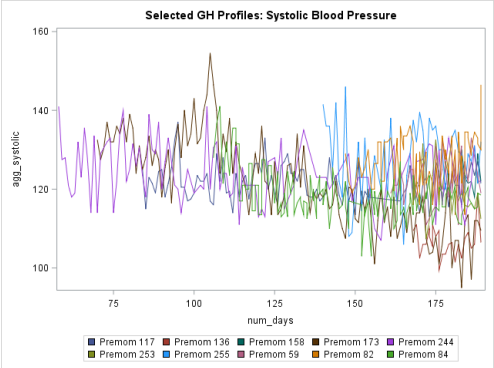

Figure 3: SBP GH profiles

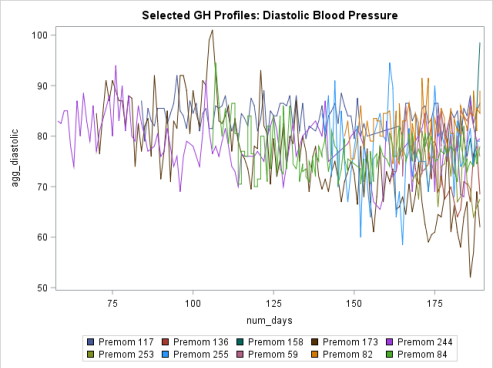

Figure 4: DBP GH profiles

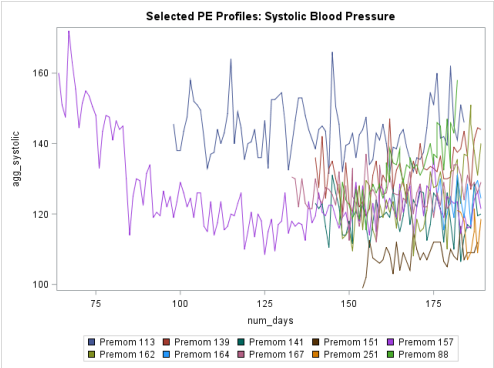

Figure 5: SBP PE profiles

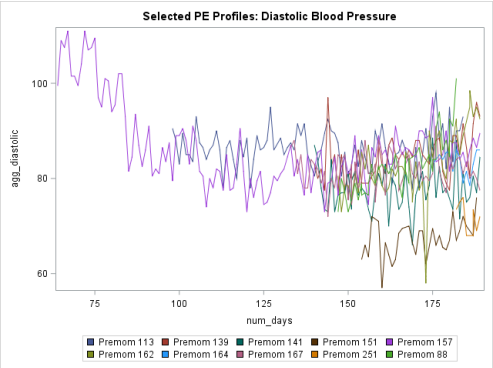

Figure 6: DBP PE profiles

Table 7 shows the distribution of the BMI across the various categories of the outcome variable. 61 of the participants had healthy weight, 47 were overweight while 54 were obese. Meanwhile table 8 describe the distribution for the reasons why the women were at high risk (RHR) across the different outcomes. The RHR describes four different categories, concomitant disorders (1), GHD in (a) previous pregnancy(ies) (2), GHD in current pregnancy (3) and other (4). Majority of the participants had GHD in current pregnancy, while Minority had other (4) except in the gestational hypertension category where the minority had concomitant disorder (1). Table 9 shows the distribution of the number of live births (parity) across the different outcomes. Majority of the participants were multiparous except for the pre-eclampsia outcome where the majority were nulliparous.

Table 7: Distribution of BMI

	Healthy	**Overweight**	**Obese**	**Total**
Normal	34	20	31	85
Gestation hypertension	20	19	15	54
Pre-eclampsia	7	8	8	23
General	61	47	54	162

Table 8: Distribution of reasons for high risk

	1	**2**	**3**	**4**	**Total**
Normal	17	24	29	15	85
Gestation hypertension	7	11	28	8	54
Pre-eclampsia	4	7	9	3	23
General	28	42	66	26	162

1=Concomitant disorders 2=GHD in (a) previous pregnancy(ies)
3=GHD in current pregnancy 4=Other

Table 9: Distribution of parity

	Nulliparous	**Multiparous**	**Total**
Normal	34	51	85
Gestation hypertension	22	32	54
Pre-eclampsia	14	9	23
General	70	92	162

4.2 Statistical Analysis

4.2.1 First-stage model

The first-stage model was fitted using the blood pressures as the dependent variable regresssed on the number of days and quadratic days effect variable as predictors.

For the DBP first-stage model, some of the participants' specific fit plots are shown in figures 7 and 8. Overall, there seemed to be a good fit when using the quadratic time effect term in the model as the fitted plots for most patients follow the overall trend seen in each participant. There were two participants with biased results due to less data points (1 and 4 total measurements), and as such they were excluded from the analysis part of this thesis. The R-square values range from 0.0004 to 0.785 while the root mean square error (RMSE) range from 1.303 to 13.034.

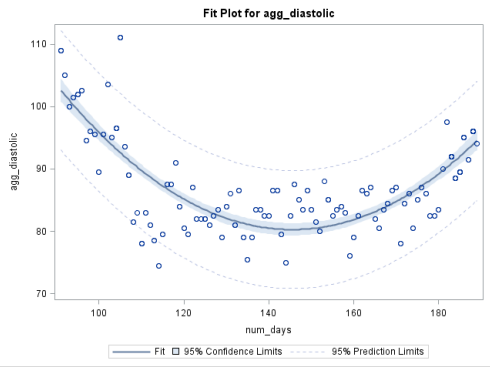

Figure 7: Premom 218 DBP fit plot

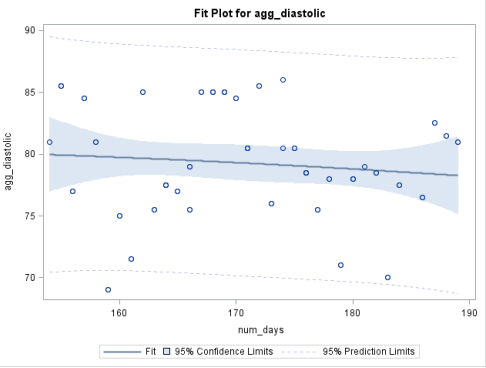

Figure 8: Premom 99 DBP fit plot

For the SBP first-stage model, some women specific fit plots are shown in figures 9 and 10. There was a good fit using the quadratic days effect in the model as shown in the fit plots of the participants. The same 2 participants with biased results in the DBP had biased results still. The range for the R-square and the root mean squared error (RMSE) values were 0.000685 to 0.798 and 1.415 to 15.187 respectively.

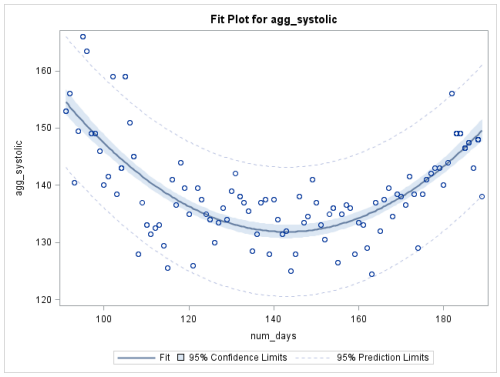

Figure 9: Premom 218 SBP fit plot

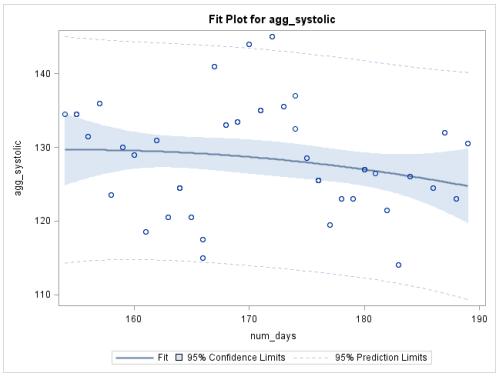

Figure 10: Premom 99 SBP fit plot

4.2.2 Second-stage model

The participant specific estimates from the two first-stage models were then merged with the other variables from the main data set and the resulting data set was used to fit the second-stage model (equation 3). This model can be re-written as;

$$\begin{aligned} P(GHD \leq j) = \phi(&\alpha_j + \beta_1 * Parity0 + \beta_2 * Age + \beta_3 * BMI\_cat1 + \beta_4 * BMI\_cat2 \\ &+ \beta_5 * RHR1 + \beta_6 * RHR2 + \beta_7 * RHR3 + \beta_8 * Intercept_{DBP} \\ &+ \beta_9 * Days_{DBP} + \beta_{10} * Days^2_{DBP} + \beta_{11} * Intercept_{SBP} \\ &+ \beta_{12} * Days_{SBP} + \beta_{13} * Days^2_{SBP}) \end{aligned}$$

where α_j is the intercept, $j = 1, 2$ and ϕ is the cumulative distribution function of the standard normal.

A cumulative probit model was fitted in order to utilize the ordinality of the dependent variable and convergence of the model was attained. A Score test for common slopes (table 10) was performed (χ^2=17.67, p-value=0.1706) and it showed that there was no evidence against the use of common slopes. Deviance goodness of fit statistics had a p-value=0.726 signaling that the model had a good fit overall and there was no evidence against the proposed model. The fit statistics are shown in tables 11 and 12.

Table 10: Score test for equal slopes assumption

Chi-Square	**DF**	**Pr >Chisq**
17.6669	13	0.1706

Table 11: Deviance and Pearson Goodness-of-Fit Statistics

Criterion	**Value**	**DF**	**Value/DF**	**Pr >Chisq**
Deviance	289.76	305	0.95	0.726
Pearson	302.91	305	0.99	0.523

Table 12: Model Fit Statistics

Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	316.862	319.757
BIC	323.013	365.884
-2Log$L$	312.862	289.757

The likelihood ratio test (table 13) was done to test if at least one of the predictors' regression coefficient present in the model is not equal to zero. This test was significant (p-value=0.0404) meaning that the fitted model is significantly better than an intercept only model since at least one of the predictors regression coefficient was significantly different from zero. This is further seen with the significant p-values of the type 3 test for the systolic blood pressure (SBP) effects in the model (table 14). This can be interpreted as the SBP variables in the model significantly improve the model fit while the others do not and can be removed. But clinical relevance now takes precedence in this analysis and no variable was removed. Table 15 shows the maximum likelihood estimates of the parameters present in the model.

Table 13: Testing Global Null Hypothesis: BETA=0

Test	**Chi-Square**	**DF**	**Pr>ChiSq**
Likelihood Ratio	23.1054	13	0.0404
Score	19.1441	13	0.1187
Wald	17.257	13	0.1878

Table 14: Type 3 Analysis of Effects

Effect	**DF**	**Wald Chi-Square**	**Pr>ChiSq**
Parity	1	1.9595	0.1616
Age	1	0.0069	0.9339
BMI cat	2	3.8791	0.1438
Reason High Risk	3	5.0448	0.1685
Intercept DBP	1	0.124	0.7248
Days DBP	1	0.1462	0.7022
Days^2 DBP	1	0.1701	0.68
Intercept SBP	1	4.5227	0.0334
Days SBP	1	4.3174	0.0377
Days^2 SBP	1	4.1365	0.0420

Interpretation of the results

Interpretation of the parameter estimates in probit regression is not as straightforward as the interpretations of parameter estimates in linear regression or logit regression. The increase in probability attributed to a one-unit increase in a given parameter is dependent both on the values of the other parameters and the starting value of the given parameter [6].

However, there are limited ways in which one can interpret the individual parameter estimates. A positive estimate means that an increase in the parameter leads to an increase in the predicted probability. A negative estimate means that an increase in the parameter leads to a decrease in the predicted probability [6].

Table 15: Analysis of Maximum Likelihood Estimates

Parameter		**DF**	**Estimate**	**SE**	**Wald Chi-Square**	Pr>ChiSq
Intercept	2	1	-3.3229	1.0081	10.865	0.0010
Intercept	1	1	-2.1924	0.9941	4.8639	0.0274
Parity	0	1	0.3093	0.2210	1.9595	0.1616
Age		1	-0.0019	0.0224	0.0069	0.9339
BMI cat	1	1	0.2593	0.2499	1.0767	0.2994
BMI cat	2	1	0.4860	0.2468	3.8786	0.0489
Reasons High Risk	1	1	0.0571	0.3475	0.0270	0.8696
Reasons High Risk	2	1	0.4950	0.3314	2.2307	0.1353
Reasons High Risk	3	1	0.5220	0.2950	3.1302	0.0769
Intercept DBP		1	0.0011	0.0031	0.1240	0.7248
Days DBP		1	0.1718	0.4493	0.1462	0.7022
Days^2 DBP		1	26.360	63.912	0.1701	0.6800
Intercept SBP		1	0.0117	0.0055	4.5227	0.0334
Days SBP		1	1.8762	0.9030	4.3174	0.0377
Days^2 SBP		1	299.5	147.3	4.1365	0.0420

The two cumulative models derived from table 15 are;

$$\begin{aligned} P(GHD = 2) = \phi(&-3.3229 + 0.3093 * Parity0 - 0.0019 * Age + 0.2593 * BMI\_cat1 \\ &+ 0.486 * BMI\_cat2 + 0.0571 * RHR1 + 0.495 * RHR2 \\ &+ 0.522 * RHR3 + 0.0011 * Intercept_{DBP} + 0.1718 * Days_{DBP} \\ &+ 26.36 * Days^2_{DBP} + 0.0117 * Intercept_{SBP} + 1.8762 * Days_{SBP} \\ &+ 299.5 * Days^2_{SBP}) \end{aligned} \tag{4}$$

$$\begin{aligned} P(GHD \geq 1) = \phi(&-2.1924 + 0.3093 * Parity0 - 0.0019 * Age + 0.2593 * BMI\_cat1 \\ &+ 0.486 * BMI\_cat2 + 0.0571 * RHR1 + 0.495 * RHR2 \\ &+ 0.522 * RHR3 + 0.0011 * Intercept_{DBP} + 0.1718 * Days_{DBP} \\ &+ 26.36 * Days^2_{DBP} + 0.0117 * Intercept_{SBP} + 1.8762 * Days_{SBP} \\ &+ 299.5 * Days^2_{SBP}) \end{aligned} \tag{5}$$

From equation 4, we are fitting the probability for a pre-eclampsia diagnosis. Evaluating all the predictors in the model at zero, the constant term -3.3229 is the predicted probability of pre-eclampsia for a multiparous obese woman who had other reasons to have GHD to be $\phi(-3.3229) = 0.000445$. But since zero is not in the range of values for the continuous variables, this probability will best be evaluated when these variables are mean-centered. Thus taking the mean of the age and the blood pressure estimates from the first-stage models, the predicted probability of a participant to have pre-eclampsia compared to having gestational hypertension or normal diagnosis of GHD is given as $\phi(-3.3229 + 1.42) = 0.0285$.
So, the predicted probability of a 31 years old multiparous obese woman with average evolution of blood pressure measurements to have pre-eclampsia is about 2.85%.

From equation 5, we are fitting the probability for the diagnosis to be pre-eclampsia or gestational hypertension. Similarly, the predicted probability to have pre-eclampsia or gestational hypertension compared to normal diagnosis when evaluating the continuous predictors at their mean and the categorical at the reference category $\phi(-2.1924 + 1.42) = 0.2201$.
So, the predicted probability of a 31 years old multiparous obese woman with average evolution of blood pressure measurements to have pre-eclampsia or gestational hypertension is about 22%.

The SBP women-specific estimates from the first-stage model significantly increase the predicted probability of the women GHD. While the other predictors in the model increase the predicted probability to predict GHD, the age variable reduces the predicted probability of GHD in the predictors.

4.3 Model performance

After fitting the model, it was of interest to check how the set of predictors in the model could have classified our one hundred and sixty participants into the various levels of the outcome variable. This was done by examining the association between the predicted probabilities and the observed response as presented in table 16. The Somers' D statistics which is a measure of agreement between pairs of ordinal variables takes values between -1 (total disagreement) to 1 (total agreement) had a value of 0.317. This shows a fair agreement between the predicted probabilities and the observed response.

A cross tabulation between the predicted response from the model and the observed response was performed as shown in table 17. Of the 160 participants used in modelling, 126 (78.75%) were predicted to have a normal outcome compared to 84 (52.5%) that was observed, 31 (19.38%) were predicted to have gestational hypertension compared to 54 (33.75%) that was observed. A large difference was seen in the pre-eclampsia category where 3 (1.88%) were predicted compared to 22 (13.75%) that was observed. A total of 91 (56.9%) out of 160 participants were correctly classified in their observed outcome.

Table 16: Association of Predicted Probabilities and Observed Responses

Percent Concordant	65.6	Somers' D	0.317
Percent Discordant	33.8	Gamma	0.319
Percent Tied	0.6	Tau-a	0.189
Pairs	7572	c	0.659

Table 17: Observe and predicted response value

	Predicted Response			
Observed Response	0	1	2	Total (Percentage)
0	74	10	0	84 (52.5)
1	38	15	1	54 (33.75)
2	14	6	2	22 (13.75)
Total (Percentage)	126 (78.75)	31 (19.38)	3 (1.88)	160 (100)

0=Normal outcome 1=Gestational hypertension 2=Pre-eclampsia

5 Possible Drawbacks

The following are limitations of the study

- Most of the women did not follow the protocol as outlined in the PREMOM I study. This resulted in a highly unbalanced dataset with sparse measurements of the key predictor variables (SBP and DBP) whose trend we were interested in studying for the prediction of the dependent variable (GHD).
- This prediction is based on observations gotten from 160 women which might not be representative of the actual pregnant women population. Therefore, conclusions drawn from this analysis might be subjective to this dataset only.
- The best linear unbiased predictors (BLUP) estimates or clustering uncertainty from the first-stage model is not accounted for in the second-stage analysis, which may lead to biased results.

6 Conclusion

The objective of this thesis was to find the possibility to predict the occurrence of gestational hypertensive disorders (GHD) in pregnant women who were at most in the second trimester of their pregnancy (GA $\leq$ 27weeks) by studying primarily the trend in their blood pressure measurements. This analysis was done in the statistical softwares R and SAS9.4 using the two-stage model methodology. Since the women were task to make repeated measurements of their blood pressure measurements two times daily, we had several repeated measurements across the days. These measurements were aggregated to have just one measurement a day by taking the median measurement for each day there was a blood pressure measurement.

Results from the second-stage model show that all the women-specific SBP estimates from the first-stage model significantly increase the predicted probability of the GHD of the women. The other continuous variables were found to increase the predicted probability of the women except the age covariate. The modeled categories of the categorical predictors (parity, BMI and reasons for high risk) all increase the predicted probabilities of the women GHD compared to the their reference categories.

The model correctly predicts the outcome for 91 (56.9%) of the women in the thesis. It performed well predicting the outcome for the women with observed normal diagnosis ($\approx$88%) and its performance decreases as the outcome level increases with $\approx$27.8% for the gestational hypertension and $\approx$9% for pre-eclampsia.

References

[1] Yin-Hsiu Chen, Kelly K Ferguson, John D Meeker, Thomas F McElrath, and Bhramar Mukherjee. Statistical methods for modeling repeated measures of maternal environmental exposure biomarkers during pregnancy in association with preterm birth. environmental health. 2015.

[2] Dorien Lanssens, Thijs Vandenberk, Christophe Smeets, Hélène De Cannière, Geert Molenberghs, Anne Moerbeke, Anne van den Hoogen, Tiziana Robijns, Sharona Vonck, Anneleen Staelens, Valerie Storms, Inge Thijs, Lars Grieten, and Wilfried Gyselaers. Remote monitoring of hypertension diseases in pregnancy: A pilot study. *JMIR mHealth and uHealth*, 5:e25, 03 2017.

[3] Dorien Lanssens, Thijs Vandenberk, Christophe Smeets, Hélène De Cannière, Sharona Vonck, Vandijck D Claessens J, Heyrman Y, V Storms, Inge Thijs, L Grieten, and Gyselaers Wilfried. Prenatal remote monitoring of women with gestational hypertensive diseases: Cost analysis. *J Med Internet Res 2018;20(3):e102*, 2018.

[4] Dorien Lanssens, Sharona Vonck, Thijs Vandenberk, Cédric Schraepen, Valerie Storms, Inge Thijs, Lars Grieten, and Wilfried Gyselaers. A prenatal remote monitoring program in pregnancies complicated with gestational hypertensive disorders: What are the contributors to the cost savings?. *Telemed J E Health.*, 25:686–692, 8 2018.

[5] Renu Singh. Hypertensive disorders in pregnancy. clinical queries:. *Clinical Queries Nephrology*, 2, 2013.

[6] Introduction to SAS. Ucla: Statistical consulting group. https://stats.idre.ucla.edu/sas/output/probit-regression/ (accessed June 06,2020).

[7] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data.* Springer-Verslag, New York., 01 2000.

7 Appendix

This section is dedicated for illustrations of the issue faced when the blood pressure measurements were aggregated on a weekly basis. Many of the participants had less than five unique data points and thus the model could not be fitted. A few randomly selected profiles are shown for each outcome diagnosis as displayed in appendix A (figures 11 to 16).

Appendix B shows maximum likelihood estimates (table 18) for the model when the BLUP estimates uncertainty from the first-stage model was accounted for in the second-stage analysis. While table 19 shows the cross tabulation between the predicted response from the corrected model and the observed response.

7.1 Appendix A - Weekly Aggregated BP Measurements

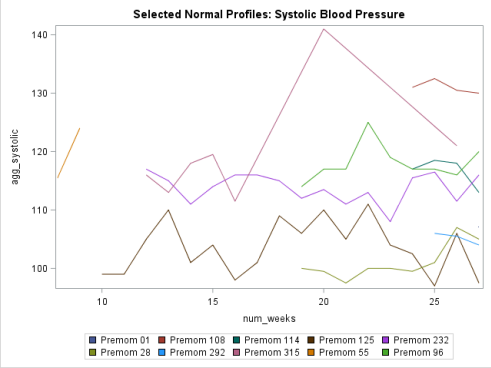

Figure 11: SBP Normal profiles

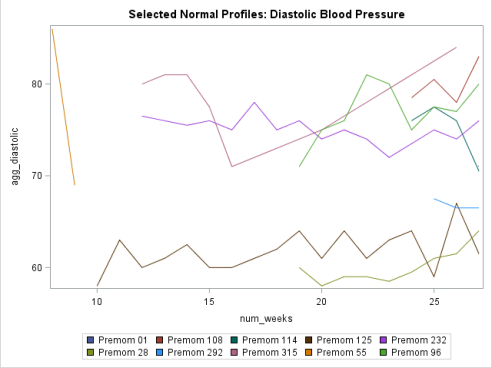

Figure 12: DBP Normal profiles

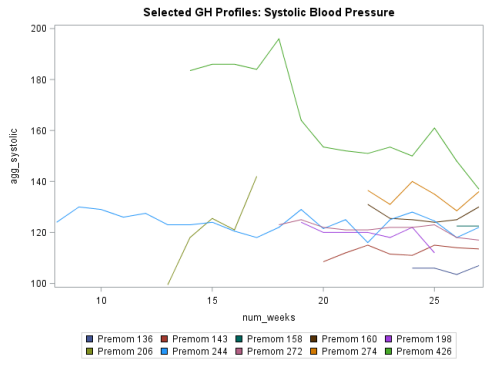

Figure 13: SBP GH profiles

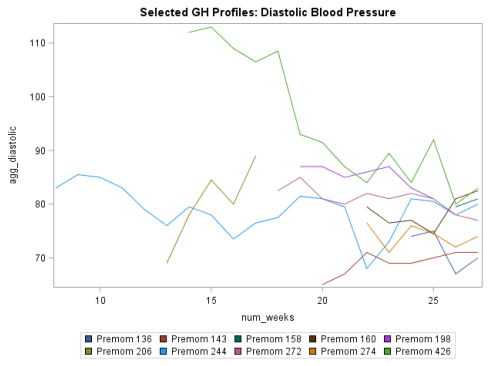

Figure 14: DBP GH profiles

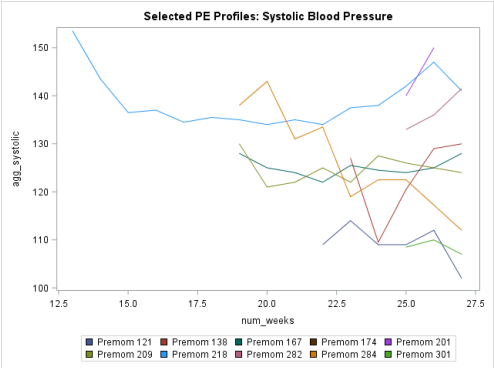

Figure 15: SBP PE profiles

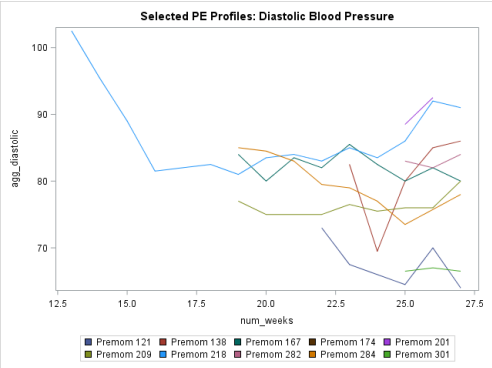

Figure 16: DBP PE profiles

7.2 Appendix B - Tables

Table 18: Analysis of Maximum Likelihood Estimates

Parameter		**DF**	**Estimate**	**SE**	**Wald Chi-Square**	Pr>ChiSq
Intercept	2	1	-2.8228	1.0685	6.9789	0.0082
Intercept	1	1	-1.6815	1.0575	2.5282	0.1118
Parity	0	1	0.2836	0.2214	1.6405	0.2003
Age		1	-0.0101	0.0231	0.1894	0.6634
BMI cat	1	1	0.2706	0.2516	1.1566	0.2822
BMI cat	2	1	0.4901	0.2515	3.7973	0.0513
Reason High Risk	1	1	0.0210	0.3480	0.0036	0.9519
Reason High Risk	2	1	0.4108	0.3358	1.4968	0.2212
Reason High Risk	3	1	0.4588	0.2982	2.3670	0.1239
Intercept DBP		1	0.000955	0.00310	0.0948	0.7581
Days DBP		1	0.1520	0.4532	0.1125	0.7374
Days^2 DBP		1	23.6910	64.7941	0.1337	0.7146
Intercept SBP		1	0.0108	0.00570	3.5596	0.0592
Days SBP		1	1.7254	0.9401	3.3682	0.0665
Days^2 SBP		1	274.5	153.5	3.1979	0.0737

Table 19: Observe and predicted response value

	Predicted Response			
Observed Response	0	1	2	Total (Percentage)
0	72	12	0	84 (52.5)
1	40	13	1	54 (33.75)
2	13	7	2	22 (13.75)
Total (Percentage)	125 (78.12)	32 (20.00)	3 (1.88)	160 (100)

0=Normal outcome 1=Gestational hypertension 2=Pre-eclampsia

7.3 Appendix C - SAS Codes

```
/*Selected individual profiles by outcome*/

proc sort data=test2;
by pat_id num_days;
run; quit;

data diag0 diag1 diag2;
set test2;
if diag_gh=0 then output diag0;
else if diag_gh=1 then output diag1;
else if diag_gh=2 then output diag2;
run; quit;

data temp1(keep=pat_id);
set diag2;
run; quit;

proc sort data=temp1 out=temp1 noduprecs;
 by pat_id;
run; quit;
```

```
proc surveyselect data=temp1 method=srs n=10 out=temp2;
run; quit;

proc sort data=temp2;
by pat_id;
run; quit;

data temp3;
  merge diag2(in=from0) temp2(in=fromtemp2);
  by pat_id;
  fromtmp = from0;
  fromtmp11 = fromtemp2;
run; quit;

 data temp4;
  set temp3;
  where (fromtmp=1) & (fromtmp11=1);
 run; quit;

 proc sort data=temp4;
by pat_id num_days;
run; quit;

proc means data=temp4 min max range;
var num_days;
run; quit;

goptions reset=all i=join noborder;
proc sgplot data=temp4;
   title 'Selected PE Profiles: Systolic Blood Pressure ';
   series x=num_days y=agg_systolic / group=pat_id  name='grouping';
   keylegend 'grouping' / type=linecolor;
   axis1 label=('Number of days Pregnant')  order=(5 to 30 by 1) minor=none;
run; quit;
title;

goptions reset=all i=join noborder;
```

```
proc sgplot data=temp4;
   title 'Selected PE Profiles: Diastolic Blood Pressure ';
   series x=num_days y=agg_diastolic / group=pat_id  name='grouping';
   keylegend 'grouping' / type=linecolor;
   axis1 label=('Number of days Pregnant')  order=(5 to 30 by 1) minor=none;
run; quit;
title;

***Fitting the 2-stage model;

***Stage 1, Diastolic BP model;

proc sort data=mt.d_dt out=mt.d_sorted noduprecs;
by pat_id;
run; quit;

proc glm data=mt.d_sorted;
model agg_diastolic=num_days num_days*num_days;
*output out=stage1_dd ParameterEstimates=dhat;
by pat_id;
run; quit;

proc reg data=mt.d_sorted outest=stage1_dd edf;
model agg_diastolic=num_days nd2;
by pat_id;
run; quit;

data reg1_dd;
set stage1_dd(keep=pat_id Intercept num_days nd2 _RMSE_ _RSQ_);
if pat_id="Premom 174" then delete;
if pat_id="Premom 98" then delete;
mse_d=_RMSE_*_RMSE_;
rename Intercept=Int_d num_days=nd1_d nd2=nd2_d _RMSE_=rmse_d _RSQ_=rsq_d;
run; quit;

***Stage 1, Systolic BP model;

ods output ParameterEstimates=parms_sd FitStatistics=fitstats_sd;
```

```
proc glm data=mt.d_sorted;
model agg_systolic=num_days num_days*num_days;
by pat_id;
run; quit;

ods output FitStatistics = fs_sd;
proc reg data=mt.d_sorted outest=stage1_sd edf noprint;
model agg_systolic=num_days nd2;
by pat_id;
run; quit;

data reg1_sd;
set stage1_sd(keep=pat_id Intercept num_days nd2 _RMSE_ _RSQ_);
if pat_id="Premom 174" then delete;
if pat_id="Premom 98" then delete;
mse_s=_RMSE_*_RMSE_;
rename Intercept=Int_s num_days=nd1_s nd2=nd2_s _RMSE_=rmse_s;
run; quit;

***Stage 2, Days' model;

data reg11_d;
set mt.d_sorted;
drop agg_systolic agg_diastolic num_days num_weeks;
if pat_id="Premom 174" then delete;
if pat_id="Premom 98" then delete;
run; quit;

proc sort data=reg11_d;
by pat_id;
run; quit;

data stage2_d;
merge reg1_dd reg1_sd reg11_d;
by pat_id;
run; quit;

proc sort data=stage2_d nodupkey;
```

```
by pat_id;
run; quit;

***Categorising BMI;

data stage2_d;
set stage2_d;
if bmi < 25 then bmicat=1;
if bmi >= 25 and bmi <30 then bmicat=2;
if bmi > 30 then bmicat=3;
run; quit;

***Probit logit model;

***Model without correction of stage 1 variability;

proc logistic data=stage2_d descending;
class rhr0 parity bmicat/ param=ref;
model diag_gh = parity age bmicat rhr0 int_d nd1_d nd2_d int_s nd1_s nd2_s
/ link=probit aggregate scale=none pprob=0.5;
output out=ppred predicted=pprobit predprobs=c;
run; quit;

proc means data=ppred;
var diag_gh pprobit;
run; quit;

*Model classification & performance;

proc sort data=ppred out=ppred_sorted nodupkey;
by pat_id;
run; quit;

title 'ppred sorted';
proc freq data=ppred_sorted;
tables  _from_*_into_;
run; quit;
title;
```

```
***Correcting for the stage 1 variability;

proc logistic data=stage2_d descending;
class rhr0 parity bmicat / param=ref;
model diag_gh = parity age bmicat rhr0 int_d nd1_d nd2_d rmse_d int_s
nd1_s nd2_s rmse_s/ link=probit aggregate scale=none ctable pprob=0.5;
output out=cppred predicted=pprobit predprobs=c;
run; quit;

proc means data=cppred;
var diag_gh pprobit;
run; quit;

*Model classification performance;

proc sort data=cppred out=cppred_sorted nodupkey;
by pat_id;
run; quit;

title 'cppred sorted';
proc freq data=cppred_sorted;
tables  _from_*_into_;
run; quit;
title;
```