INVESTIGATING THE EFFECT OF BODY MASS INDEX ON THE RISK PRE-ECLAMPSIA.

32102717

Abstract

Pre-eclampsia complicates up to 6% of pregnancies in the UK alone[5], and can lead to severe complications for both mother and child, including eclampsia and HELLP syndrome[4]. Given a data set from a case-control study of 750 mothers, we wish to investigate the relationship between Body Mass Index (BMI) and pre-eclampsia. To do this we take known information about the variables associated with BMI and/or pre-eclampsia, and create a directed a-cyclic graph (DAG) to aid in building a logistic regression model. We also employ simulations to help find the model with the least bias. Our findings suggest that higher BMIs increase the risk of contracting pre-eclampsia, however, there is a great deal of uncertainty in our estimates of the odds-ratios. We suggest that further studies be undertaken with larger and better distributed sample populations, that also include data on more variables not included in our DAG.

1. Introduction

Pre-eclampsia is a hypertensive disease that occurs during pregnancy[1], and is estimated to complicate up to 6% of pregnancies in the UK, with 1-2% being severe cases[5]. It usually occurs after the 20th week of pregnancy or postpartum, and is categorised by high blood pressure (hypertension) and protein in the urine (proteinuria)[1;5]. Pre-eclampsia can cause a great number of complications for both mother and child. For the mother, these include (but are not limited to) eclampsia a form of seizures, and HELLP syndrome[4]. It is estimated that $1,000$ babies in the UK die each year due to pre-eclampsia complications[4]. The only known cure for pre-eclampsia is to give birth; labour will usually be induced, or a caesarian-section performed, at 37-38 weeks[6]. In severe cases, this may need to occur sooner, and thus pre-eclampsia is a leading cause of pre-mature birth[6].

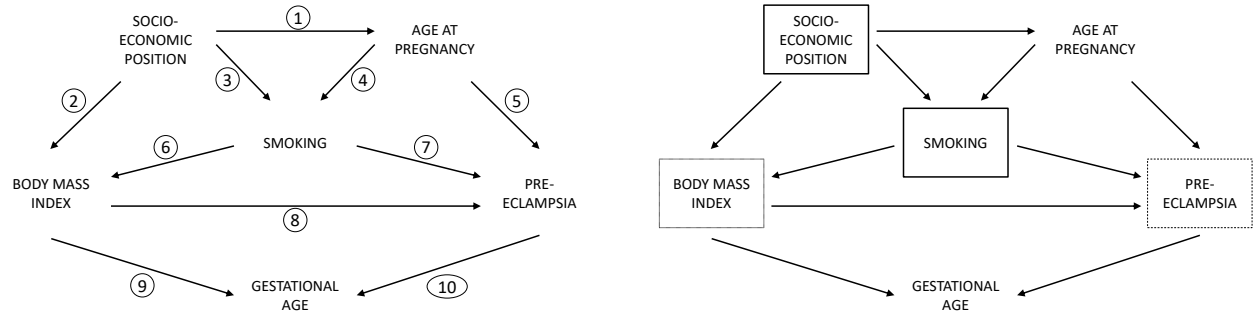
We wish to investigate whether Body Mass Index (BMI) value has an effect on the risk of contracting pre-eclampsia. BMI is a measure of whether a person's weight is "healthy" for their height[8]. It is calculated by dividing a persons weight in kilograms by their height in meters squared, kg/m^2[8]. BMI is divided into categories: ($<$18.50) "Underweight", (18.50 - 24.99) "Normal range", (25 - 29.99) "Overweight", (30 - 34.99) "Obese class I (low risk)", (35 - 39.99) "Obese class II (moderate risk)", and ($\geq$ 40.00) "Obese class III (high risk)"[8]. Previous epidemiological studies have shown that a BMI ≥ 35 is a key risk factor for pre-eclampsia[1;3]. Other risk factors include having diabetes, pre-existing high blood pressure, pre-eclampsia in a previous pregnancy, multifetal pregnancy (twins or more), a family history of pre-eclampsia, lupus, advanced maternal age ($>$ 40 years), and more[1;5].

The data pertains to a case-control study of 750 mothers which measured socio-economic position (lower to upper-upper class), age at pregnancy (in years), smoking habits (yes/no), Body Mass Index value, gestational age (in weeks), and whether or not the mother suffered pre-eclampsia during the pregnancy. The patients were aged between approximately 18 and 35 years of age, with BMIs ranging from "Underweight" to "Obese class III (high risk)". There are a only few cases than spill over into "Underweight" or "Obese class III (high risk)", and they are all on the borderline, so we have opted to include these individuals in the "Normal" or "Obese class II (moderate risk)" groups, respectively, for analysis purposes. We have approximately equal numbers of cases and controls, however some groups have very small sample sizes (as shown in Table 1).

2. Theory and Methodology

We wish to investigate the relationship between Body Mass Index (BMI) score and the occurrence of pre-eclampsia. We have data on a range of variables that have known relationships to either BMI, pre-eclampsia, or both. These relationships are demonstrated by the directed a-cyclic graph

(DAG) in Figure 1a. It has been found that higher Body Mass Index values are a risk factor for pre-mature birth (arrow (9))[2]. As previously noted, the only cure for pre-eclampsia is to give birth, so it is a major cause of pre-mature birth (arrow (10))[6].



(A) (B)

FIGURE 1. (A) The directed a-cyclic graph representing our understanding of the relationship between Body Mass Index and the occurrence of pre-eclampsia in pregnant women. (B) A directed a-cyclic graph showing the variables we intend to condition on in our model to remove as much bias as possible, and find the direct effect of BMI on pre-eclampsia.

Our aim is to condition on the variables that minimise the bias introduced into a generalised linear model that we fit based on the DAG. Bias can take the form of either confounder bias or collider bias. We wish to include confounders to remove confounder bias, but exclude colliders to remove collider bias. In an ideal DAG, we would know every associated variable that affected the model, and be able to condition a model that had no bias, such that we could isolate the direct effect of our exposure variable on our outcome variable. In reality, the DAGs created for these relationships are extremely complex, and they do not contain all the associated variables, as such there is usually no model from the DAG that completely removes bias. The goal then is to choose the model that has the least bias, informed by the "rules of the DAG", simulation, and previous knowledge.

To attempt to minimise bias we can simulate data that has the same causal structure as described by our DAG, with known coefficients, and then fit all viable models. The model that manages to estimate the known coefficients of our exposure variable (BMI) with the least error will be the one that contains the least bias. When simulating the data we have chosen to try and match the distribution of each variable, and then also stratify variables when appropriate. We have not tried to scale the variables to match their associated real world data.

Our response variable, Y, is binary, in that it measures whether or not a patient had pre-eclampsia during the observed pregnancy. Thus we let each individual have a Bernoulli distribution; $Y_i \sim \text{Bernoulli}(p_i)$, where

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \epsilon_i, \tag{1}$$

and $\epsilon_i \sim \text{Normal}(0, \sigma^2)$. This is known as logistic regression, and we interpret $\frac{p_i}{1-p_i}$ as that individuals odds of having pre-eclampsia. The coefficients in the model, β_j, can be interpreted as the log(odds ratios) for being in one catergory of a variable compared to being in the baseline catergory for the same variable, when every other variable is kept constant. For continuous variables, the odds-ratios are easiest to interpret when they are stratified, thus we have decided to stratify BMI based on the thresholds described by the World Health Organisation[8]. It is possible to interpret odds ratios on continuous scales, however it is slightly more complex and requires making assumptions about the linearity of the odds ratio as BMI increases, which does not seem appropriate in this context.

3. Results and Analysis

Table 1 provides an overview of the data, demonstrating how the cases and controls of pre-eclampsia are distributed through the different categories in each variable. In particular we notice that for the group with a BMI ≥ 35 we have very few controls, and a small number of observations over all compared to the other groups. Similarly with Socio-Economic Position, the middle 4 classes all have very few observations.

Continuous Variables	Mean	Std. Deviation	Min/Max
Age at Pregnancy	25.48 Years	4.42	(17.94, 35.51)
Gestational Age	39.91 Weeks	2.23	(34.38, 45.07)
Body Mass Index	28.50	4.99	(18.47, 40.42)
Categorical Variables	**Levels**	**Cases**	**Controls**
Smoking	No	169	233
	Yes	196	152
Body Mass Index	< 25	76	147
	25 - 29.99	113	147
	30 - 34.99	100	62
	≥ 35	76	29
Socio-Economic Position	Lower	113	55
	Lower-Middle	33	20
	Middle	39	18
	Upper-Middle	13	16
	Upper	14	22
	Upper-Upper	153	254

Table 1. A summary of the data, organised by variable. BMI is included twice as it is a variable we chose to stratify for analysis, but was initially continuous.

Our aim is to identify the model that best allows us to find the true direct effect of BMI on pre-eclampsia, given our DAG is correct. We first consider the "rules of the DAG" to try and eliminate some of the possible models. We can see from Table 2 that no model removes all possible bias, and there are no clear models that reduce the bias the most (as the amount of bias added by each biasing element is unknown). The only conclusion we can discern is that if we were to include either both Smoke and SEP or Smoke and AgePreg we would be able to close all back door paths and not include collider bias from smoking. One example is shown in Figure 1b.

As a solution to this issue we can simulate data. We simulated data based on the relationships described in Figure 1a, and fit models derived from the DAG to identify which introduced the least bias into the coefficients for the different levels. Since "gestational age" is a collider for BMI and pre-eclampsia, we assumed that every model would automatically not include it. We set all the coefficients in the logistic regression for pre-eclampsia to 0.1, and from our simulation we discovered that the model that does not include AgePreg introduces the least bias in the coefficients for BMI (on average), if the assumed DAG in Figure 1a is true. This is demonstrated by Figure 2a. Thus the simulation suggests that the best model for pre-eclampsia is given by:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) =& \beta_0 + \beta_1 x_{\text{Smok}=1,i} + \beta_2 I_{\text{BMI}=2,i} + \beta_3 I_{\text{BMI}=3,i} + \beta_4 I_{\text{BMI}=4,i} \\ &+ \beta_5 I_{\text{SEP}=2,i} + \beta_6 I_{\text{SEP}=3,i} + \beta_7 I_{\text{SEP}=4,i} + \beta_8 I_{\text{SEP}=5,i} + \beta_9 I_{\text{SEP}=6,i} + \epsilon_i,\end{aligned}$$

where $\epsilon_i \sim \text{Normal}(0, \sigma^2)$. The causal diagram representing this model is given in Figure 1b, and Table 3 gives the details of the coefficient estimates for this model, as well as their 95% confidence intervals.

4. Discussion

From Table 4 we can see that the confidence intervals cross 1 for all BMI groups. For the two lower groups (25 - 29.99) and (30 - 34.99), Normal and Overweight respectively (as defined in §1),

Included	Confounder Bias?								Collider Bias?
	$SEP_{2,3}$	$SEP_{2,1}$	$SEP_{2,15}$	$SEP_{2,37}$	$SEP_{2,147}$	$Age_{4,5}$	$Age_{46,5}$	$Smoke_{6,7}$	$Smoke_{3,4}$
None	X	X	✓	✓	✓	X	✓	✓	X
Smoke	✓	X	✓	X	X	✓	X	X	X
SEP	X	X	X	X	X	X	✓	✓	X
Age	X	✓	X	✓	X	X	X	✓	X
SEP, Age	X	X	X	X	X	X	X	✓	X
SEP, Smoke	X	X	X	X	X	✓	X	X	X
Smoke, Age	✓	✓	X	X	X	X	X	X	X
SEP, Smoke, Age	X	X	X	X	X	X	X	X	✓

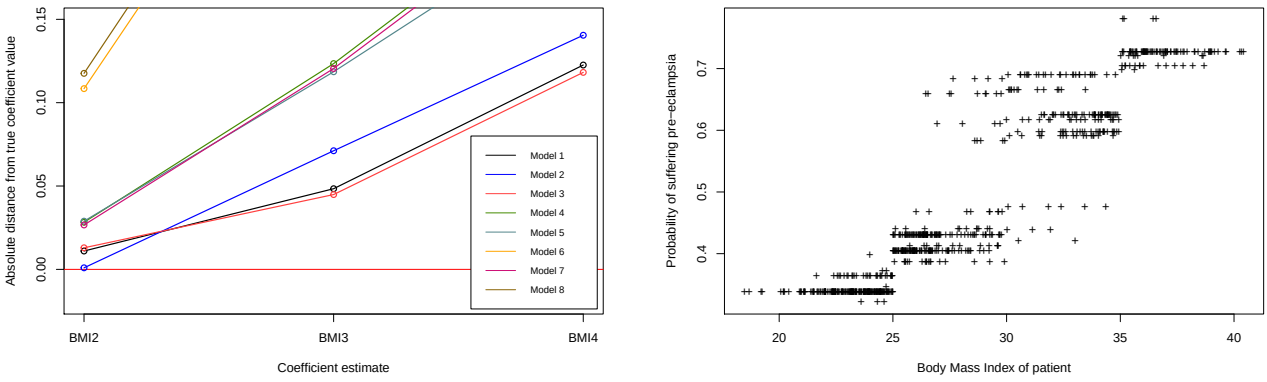
TABLE 2. A table that describes which biasing elements are present in each model. A (✓) indicates the bias is present, and an (X) indicates it is not. An example of notation: $SEP_{2,15}$ is the confounding bias introduced by not including SEP while it confounds on BMI (follow arrow (2) in Figure 1a) and pre-eclampsia (follow arrow (1) then arrow (5)).

	(Intercept)	Smoke	Body Mass Index		
Coef. Estimate	-0.68	0.11	0.28	0.31	0.78
95% CI	(-0.96, -0.39)	(-0.23, 0.45)	(-0.13, 0.69)	(-0.58, 1.20)	(-0.27, 1.84)
	Socio-Economic Position				
Coef. Estimate	-0.07	0.15	1.05	0.73	0.76
95% CI	(-0.79, 0.64)	(-0.64, 0.94)	(0.19, 1.90)	(-0.17 1.62)	(-0.15, 1.67)

TABLE 3. Details of the model coefficient estimates and their 95% confidence intervals.

BMI Catagory	Cases	Controls	Odds-ratio (95% CI)
< 25	76	147	1
25 - 29.99	113	147	1.32 (0.88, 1.99)
30 - 34.99	100	62	1.37 (0.56, 3.32)
≥ 35	76	29	2.19 (0.76, 6.27)

TABLE 4. Details of the odds-ratio for each group within BMI, and their 95% confidence intervals.



(A) (B)

FIGURE 2. (A) A graph that shows the absolute distance from the true coefficients of the mean coefficient estimates for BMI from 100 simulations. Details of each model can be found in appendix Table 5. (B) A plot of the probability of suffering pre-eclampsia for each patient in the study against their respective BMI values.

we would expect this as they are not listed as risk factors in the literature, thus we would not expect them to show evidence of increasing risk. We would, however, expect the Obese (≥ 35) group to have a lower confidence interval limit above 1, as it has been shown to be a risk factor for pre-eclampsia[1;3]. We can see an obvious trend of increasing uncertainty with decreasing sample size in each of the groups. The confidence interval for obese may contain 1 because of this uncertainty. We would advise that if this relationship is to be investigated again that the sample size be large enough to give the statistical analyses the necessary power to detect whether a BMI ≥ 35 is a risk factor for pre-eclampsia. We can, however, note the increasing trend in both the point estimates and the confidence intervals for the odds-ratio of BMI, suggesting that even with our uncertainty, a higher BMI is indicative of increased risk of pre-eclampsia. This is supported by Figure 2b. Similarly for the coefficient estimates, almost all of the confidence intervals include 0, however there is an obvious increasing pattern within the BMI category, suggesting that larger BMIs have a greater effect on pre-eclampsia compared to smaller one. We can also see that having a large BMI has one of the largest effects on risk of pre-eclampsia, though there is more uncertainty in this. For instance, its point estimate is slightly higher than that of SEP group 6 (lower class), but its confidence interval completely envelops that of lower class.

As discussed in §2 we have introduced bias into our model, and so the odds-ratios we return don't represent the true odds-ratios, however, they are the least biased of any possible according to our simulation. From Figure 2a we can see that model 3 (our chosen model) has the least bias for BMI3 and BMI4, however model 2 has the least bias for BMI2. In certain situations it may be appropriate to estimate each odds-ratio from the model that has the least bias for it, though we would need to confirm that this makes sense methodologically. There may also be biases introduced by latent variables, for instance, it is known that Pregnancy-Associated Plasma Protein A has an effect on both gestational age and pre-eclampsia[7]. It is worth noting that using classical variable selection methods, such as step-wise selection, does not account for confounders or colliders. The model given by step-wise selection, starting from the full model, suggests including the variables: socio-economic position, smoking, Body Mass Index, and gestational age, which would contain a large amount of bias. Including gestational age in particular causes the model to suggest that increasing BMI reduces the risk of pre-eclampsia.

We suggest that if this relationship is to be investigated further, that it be done with a larger and more evenly distributed sample size. Future investigations should also include additional known associations and risk factors that were not considered in this investigation.

References

[1] American College of Obstetricians and Gynecologists. Hypertension in Pregnancy. `https://www.acog.org/~/media/Task%20Force%20and%20Work%20Group%20Reports/public/HypertensioninPregnancy.pdf`, 2013. [Online; accessed 2018-02-08].

[2] Maki Murakami, Masahide Ohmichi, Tatsuya Takahashi, Akiko Shibata, Akira Fukao, Nobuyuki Morisaki, and Hirohisa Kurachi. Prepregnancy body mass index as an important predictor of perinatal outcomes in japanese. *Archives of Gynecology and Obstetrics*, 271(4):311–315, Apr 2005.

[3] NHS. Pre-eclampsia: Causes. `https://www.nhs.uk/conditions/pre-eclampsia/causes/`, 2015. [Online; accessed 2018-02-08].

[4] NHS. Pre-eclampsia: Complications. `https://www.nhs.uk/conditions/pre-eclampsia/complications/`, 2015. [Online; accessed 2018-02-08].

[5] NHS. Pre-eclampsia: Overview. `https://www.nhs.uk/conditions/pre-eclampsia/`, 2015. [Online; accessed 2018-02-08].

[6] NHS. Pre-eclampsia: Treatment. `https://www.nhs.uk/conditions/pre-eclampsia/treatment/`, 2015. [Online; accessed 2018-02-08].

[7] Gordon C. S. Smith, Emily J. Stenhouse, Jennifer A. Crossley, David A. Aitken, Alan D. Cameron, and J. Michael Connor. Early pregnancy levels of pregnancy-associated plasma protein a and the risk of intrauterine growth restriction, premature birth, preeclampsia, and stillbirth. *The Journal of Clinical Endocrinology & Metabolism*, 87(4):1762–1767, 2002.

[8] World Health Organisation. BMI classification. `http://apps.who.int/bmi/index.jsp?introPage=intro_3.html`, 2018. [Online; accessed 2018-02-08].

5. Appendix

Model	Included Variables
Model 1	BMI, SEP, AgePreg, Smoke
Model 2	BMI, AgePreg, Smoke
Model 3	BMI, SEP, Smoke
Model 4	BMI, SEP, AgePreg
Model 5	BMI, AgePreg
Model 6	BMI, Smoke
Model 7	BMI, SEP
Model 8	BMI

Table 5. Details of the models that are demonstrated in Figure 2a.

```
1
2  model1coef <- data.frame("case" = seq(1, 100, by = 1),"int" = NA,
3                           "bmi2" = NA, "bmi3" = NA, "bmi4" = NA,
4                           "smok1" = NA, "age" = NA, "se2" = NA, "se3" = NA,
5                           "se4" = NA, "se5" = NA, "se6" = NA)
6
7  model2coef <- model1coef
8  model3coef <- model1coef
9  model4coef <- model1coef
10 model5coef <- model1coef
11 model6coef <- model1coef
12 model7coef <- model1coef
13 model8coef <- model1coef
14
15
16 for(q in 1:100){
17   set.seed((19*q))
18
19   n = 1000
20
21   sep = rmultinom(n=n, size = 1, prob = rep(1/6, 6))
22
23   se = 0
24   for(i in 1:ncol(sep)){
25     for(j in 1:nrow(sep)){
26       if(sep[j,i] == 1){se[i] = j}
27     }
28   }
29
30   summary(as.factor(se))
31
32   #summary(as.factor(sample(x=c(1,2,3,4,5), size=n, replace=TRUE, prob=rep(1/5, 5))))
33
34
35   agepreg = se + rnorm(n)
36
37   eta.smok = 0.01*se + 0.01*agepreg + 0.01*rnorm(n)
38   prob.smok = exp(eta.smok) / (1 + exp(eta.smok))
39   smok = rbinom(n=n, size=1, prob=prob.smok)
40   summary(as.factor(smok))
41
42   BMI = se + smok + rnorm(n)
43   hist(BMI)
```

```
for(i in 1:n){
  if(BMI[i] < 2){BMI[i] <- 1}
  else if(BMI[i] < 4 & BMI[i] > 2){BMI[i] <- 2}
  else if(BMI[i] < 6 & BMI[i] > 4){BMI[i] <- 3}
  else if(BMI[i] > 6){BMI[i] <- 4}
}
summary(as.factor(BMI))

eta.pe = 0.1*BMI + 0.1*smok + 0.1*agepreg + rnorm(n)
prob.pe = exp(eta.pe) / (1 + exp(eta.pe))
PE = rbinom(n=n, size=1, prob=prob.pe)
summary(as.factor(PE))

sim.data <- data.frame(pe = PE, bmi = as.factor(BMI), smok = as.factor(smok), agepreg = agepreg, se = as.
  factor(se))

model1 <- glm(pe ~ ., data= sim.data, family = binomial)
model1coef$int[q] <- coef(model1)[1]
model1coef$bmi2[q] <- coef(model1)[2]
model1coef$bmi3[q] <- coef(model1)[3]
model1coef$bmi4[q] <- coef(model1)[4]
model1coef$smok1[q] <- coef(model1)[5]
model1coef$age[q] <- coef(model1)[6]
model1coef$se2[q] <- coef(model1)[7]
model1coef$se3[q] <- coef(model1)[8]
model1coef$se4[q] <- coef(model1)[9]
model1coef$se5[q] <- coef(model1)[10]
model1coef$se6[q] <- coef(model1)[11]


model2 <- glm(pe ~ .-se, data= sim.data, family = binomial)
model2coef$int[q] <- coef(model2)[1]
model2coef$bmi2[q] <- coef(model2)[2]
model2coef$bmi3[q] <- coef(model2)[3]
model2coef$bmi4[q] <- coef(model2)[4]
model2coef$smok1[q] <- coef(model2)[5]
model2coef$age[q] <- coef(model2)[6]


model3 <- glm(pe ~ .-agepreg, data= sim.data, family = binomial)
model3coef$int[q] <- coef(model3)[1]
model3coef$bmi2[q] <- coef(model3)[2]
model3coef$bmi3[q] <- coef(model3)[3]
model3coef$bmi4[q] <- coef(model3)[4]
model3coef$smok1[q] <- coef(model3)[5]
model3coef$se2[q] <- coef(model3)[6]
model3coef$se3[q] <- coef(model3)[7]
model3coef$se4[q] <- coef(model3)[8]
model3coef$se5[q] <- coef(model3)[9]
model3coef$se6[q] <- coef(model3)[10]


model4 <- glm(pe ~ .-smok, data= sim.data, family = binomial)
model4coef$int[q] <- coef(model4)[1]
model4coef$bmi2[q] <- coef(model4)[2]
model4coef$bmi3[q] <- coef(model4)[3]
model4coef$bmi4[q] <- coef(model4)[4]
model4coef$age[q] <- coef(model4)[5]
model4coef$se2[q] <- coef(model4)[6]
```

```
  model4coef$se3[q] <- coef(model4)[7]
  model4coef$se4[q] <- coef(model4)[8]
  model4coef$se5[q] <- coef(model4)[9]
  model4coef$se6[q] <- coef(model4)[10]


  model5 <- glm(pe ~ .-se-smok, data= sim.data, family = binomial)
  model5coef$int[q] <- coef(model5)[1]
  model5coef$bmi2[q] <- coef(model5)[2]
  model5coef$bmi3[q] <- coef(model5)[3]
  model5coef$bmi4[q] <- coef(model5)[4]
  model5coef$age[q] <- coef(model5)[6]


  model6 <- glm(pe ~ .-se-agepreg, data= sim.data, family = binomial)
  model6coef$int[q] <- coef(model6)[1]
  model6coef$bmi2[q] <- coef(model6)[2]
  model6coef$bmi3[q] <- coef(model6)[3]
  model6coef$bmi4[q] <- coef(model6)[4]
  model6coef$smok1[q] <- coef(model6)[5]


  model7 <- glm(pe ~ .-smok-agepreg, data= sim.data, family = binomial)
  model7coef$int[q] <- coef(model7)[1]
  model7coef$bmi2[q] <- coef(model7)[2]
  model7coef$bmi3[q] <- coef(model7)[3]
  model7coef$bmi4[q] <- coef(model7)[4]
  model7coef$se2[q] <- coef(model7)[5]
  model7coef$se3[q] <- coef(model7)[6]
  model7coef$se4[q] <- coef(model7)[7]
  model7coef$se5[q] <- coef(model7)[8]
  model7coef$se6[q] <- coef(model7)[9]


  model8 <- glm(pe ~ .-smok-agepreg-se, data= sim.data, family = binomial)
  model8coef$int[q] <- coef(model8)[1]
  model8coef$bmi2[q] <- coef(model8)[2]
  model8coef$bmi3[q] <- coef(model8)[3]
  model8coef$bmi4[q] <- coef(model8)[4]

}

  means <- data.frame("model" = seq(1, 8, by = 1),"int" = NA,
                      "bmi2" = NA, "bmi3" = NA, "bmi4" = NA,
                      "smok1" = NA, "age" = NA, "se2" = NA, "se3" = NA,
                      "se4" = NA, "se5" = NA, "se6" = NA)

  means[1,] <- colMeans(model1coef)
  means[2,] <- colMeans(model2coef)
  means[3,] <- colMeans(model3coef)
  means[4,] <- colMeans(model4coef)
  means[5,] <- colMeans(model5coef)
  means[6,] <- colMeans(model6coef)
  means[7,] <- colMeans(model7coef)
  means[8,] <- colMeans(model8coef)

  meandis <- abs(abs(means) - 0.1)

  plot( c(3:5), as.numeric(meandis[1, 3:5]), ylim = c(-.02, 0.15), type = "o", col = "black", lwd = 1.5,
```

```
        xaxt = "n", xlab = "Coefficient estimate", ylab = "Absolute distance from true coefficient value")
abline(h=0, col = "red")
lines(c(3:5), as.numeric(meandis[2, 3:5]), type = "o", col = "blue", lwd = 1.5)
lines(c(3:5), as.numeric(meandis[3, 3:5]), type = "o", col = "brown1", lwd = 1.5)
lines(c(3:5), as.numeric(meandis[4, 3:5]), type = "o", col = "chartreuse4", lwd = 1.5)
lines(c(3:5), as.numeric(meandis[5, 3:5]), type = "o", col = "cadetblue4", lwd = 1.5)
lines(c(3:5), as.numeric(meandis[6, 3:5]), type = "o", col = "orange", lwd = 1.5)
lines(c(3:5), as.numeric(meandis[7, 3:5]), type = "o", col = "deeppink3", lwd = 1.5)
lines(c(3:5), as.numeric(meandis[8, 3:5]), type = "o", col = "darkgoldenrod4", lwd = 1.5)
legend(4.55, 0.08, legend=c("Model 1", "Model 2", "Model 3","Model 4","Model 5","Model 6","Model 7", "Model 8"),
       col=c("black", "blue", "brown1", "chartreuse4", "cadetblue4", "orange", "deeppink3", "darkgoldenrod4"), lty=1, cex=0.75)
axis(1, at=3:5, labels= c("BMI2", "BMI3", "BMI4"))

# Model 3 is the best model
# glm(pe ~ .-agepreg, data= sim.data, family = binomial)

#####Data#####

summary(data)

datafac <- data
datafac$pe <- as.factor(data$pe)
datafac$smok <- as.factor(data$smok)
for(i in 1:nrow(datafac)){
  if(datafac$bmi[i] < 25){datafac$bmi[i] <- 1} #Normal + 1 stragler
  if(datafac$bmi[i] < 30 & datafac$bmi[i] > 25){datafac$bmi[i] <- 2} # overweight
  if(datafac$bmi[i] < 35 & datafac$bmi[i] > 30){datafac$bmi[i] <- 3} # Class I obese (low risk)
  if(datafac$bmi[i] > 35){datafac$bmi[i] <- 4}# Class II obese (moderate risk) +few straglers
}
datafac$bmi <- as.factor(datafac$bmi)

summary(datafac)

summary(datafac$pe[which(datafac$sep == 5)])

##### Models #####

modelfinal <- glm(pe ~ .-gestage-agepreg, data= datafac, family = binomial)
summary(modelfinal)

cbind(coef(modelfinal), confint.default(modelfinal))

exp(cbind(coef(modelfinal), confint.default(modelfinal)))
#bmi2        1.3245873 (0.8803780 1.9929296)
#bmi3        1.3654466 (0.5619611 3.3177465)
#bmi4        2.1858688 (0.7625313 6.2660017)


prob <- predict(modelfinal, datafac, type = "response")
plot(x = data$bmi, y= prob, pch = "+", ylab = "Probability of suffering pre-eclampsia", xlab = "Body Mass Index of patient")
```