Assignment 2:

Logistic Regression Models for Secondary Infertility Data

1 Introduction

1.1 Motivation and research questions

This enquiry is motivated by data taken from a 1970's study investigating secondary infertility in women after induced illegal abortions performed by dilation and curettage.[1] Dilation and curettage (D&C) is the surgical process of dilating the cervix and scraping off the lining (endometrium) of the uterus.[2] Despite being advised against by the World Health Organisation, D&C is still performed by some countries including Mexico and Japan. The process of scraping and may play a causal role in endometrial thinning.[3] Additionally, D&C increases the risk of pre-term births, uterine perforations, and intrauterine adhesions.[4] It is perhaps these adverse events which affect the risk of secondary infertility. However, the data from this study shows not every female who had an induced abortion by D&C was subsequently diagnosed with secondary infertility. For this reason, the aim of this enquiry is to understand how other variables play a role in increasing the probability of secondary infertility. Furthermore, the data analysis will assess which variables are good predictors that a patient is likely to be diagnosed with secondary infertility.

1.2 Data overview

The dataset used in this enquiry comes from the inbuilt *infert* dataset in R. The original data is acquired from a study investigating secondary infertility in women after induced illegal abortions performed by dilation and curettage.[5] The study uses 83 patients diagnosed with secondary infertility by the First Department of Obstetrics and Gynaecology of the University of Athens Medical School from 1973 – 1974. A diagnosis of secondary infertility was accepted if there had been a previous conception. This is reflected on the *parity* variable, which shows the minimum value for previous pregnancies is 1. Each of the 83 patients was paired two control subjects that matched in terms of age group, level of education, number of abortions and number of previous pregnancies. The matched case method can be seen in the *stratum* variable, which shows for each case of

[1] Trichopoulos, et al. (1976) 646.
[2] *Dilation and Curettage,* OxfordReference. Website.
[3] Azumaguchi , et al (2017) 523.
[4] Küng, et al. (2021) 1.
[5] Trichopoulos, et al. (1976) 646.

infertility there is data from three women. Importantly, the study indicates all control patients are fertile. Therefore, the data is binary as there are only two possibilities for the patient's medical condition: fertile and infertile. Case is the dependant variable.[6]

$$Case = \begin{cases} 0\ if\ fertile \\ 1\ if\ Infertile \end{cases}$$

For readability purposes, *case* was converted into a factor to ensure it was treated as a categorical variable.[7] The fertile group was called *control* to show they were part of the control group, while the infertile group was called *infertile* show their diagnosis as infertile.

1.3 Exploratory data analysis

It seems that *spontaneous* is a good predictor for whether someone is likely to be diagnosed with secondary infertility or not. In this dataset *spontaneous* refers to the number of spontaneous abortions. Spontaneous abortions can include miscarriages and stillbirths.[8] Box plots show that while there are equal cases of infertility and fertility for induced abortions, as the number of spontaneous abortions goes up, so does infertility. (Figures 1 & 2) The mean number of spontaneous abortions in the control group is 0 while in the infertile group the mean number is 1. Suggesting the number of spontaneous abortions reflects a female's fertility status. It must be noted that although the maximum number for *spontaneous* shown is 2; this could indicate the female has had more than two spontaneous abortions as the data only documents 2+ as the maximum number.[9] Logistic regression plots of the variables *case* and *induced*, and *case* and *spontaneous* also highlight the trend that as the number of spontaneous abortions increases, so does the likelihood of infertility. (Figures 3 & 4)

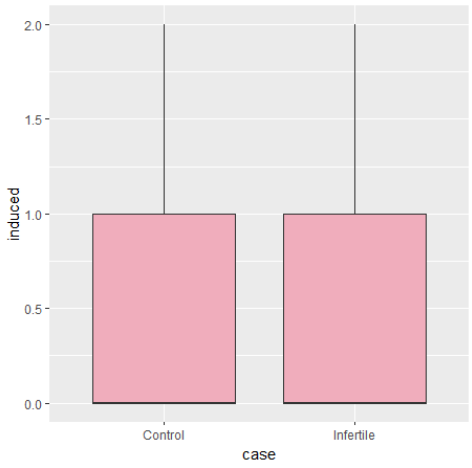


Figure 1: Boxplots showing *case* at the x axis and *induced* at the y axis.

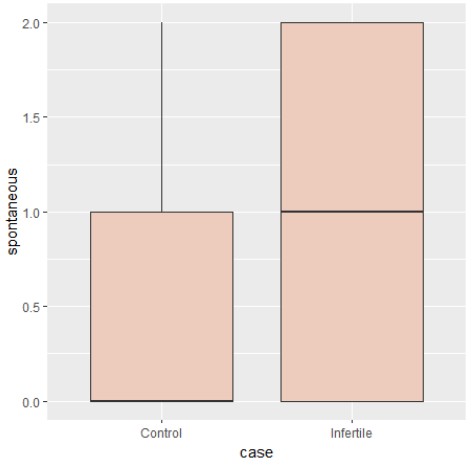


Figure 2: Boxplots showing *case* at the x axis and *spontaneous* at the y axis. In the infertile group the number of spontaneous abortions has significantly increased.

[6] James, et al. (2021) 132.
[7] *Factors in R*, StatBerkeley.edu, Website.
[8] Trichopoulos, et al. (1976) 646.
[9] Trichopoulos, et al. (1976) 646.

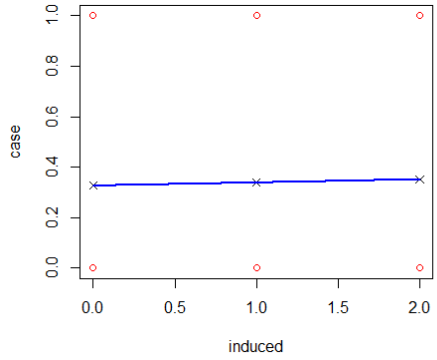


Figure 3: Logistic regression model plotted to show the relationship between *induced* and the probability that a case is infertile

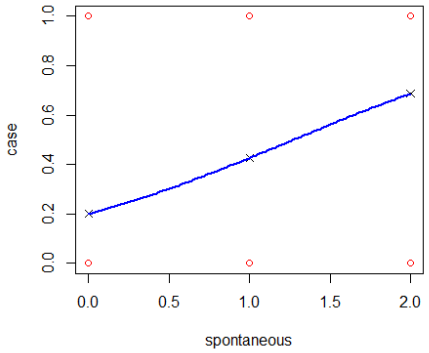


Figure 2: Logistic regression model plotted to show the relationship between spontaneous and the probability that a case is infertile

Variable	Pearson's product-moment correlation *r*	Spearman's rank correlation rho r_s	Notes
Spontaneous	0.3659506	0.3599685	Number of spontaneous abortions
Induced	0.01947074	0.0190926	Number of induced abortions
Parity	-4.895674e-20	0	Number of previous children
Age	-7.008532e-20	0	Age of patient
Education	NA	NA	Education in years

Table 1: Pearson's correlation coefficient (r), Spearman's rho correlation coefficient (r_s)

Table 1 shows bivariate correlations between *case* and other variables. As the education variable is nonnumeric, the test results returned NA. The Pearson correlation measures the strength and direction of linear relationships between pairs of continuous variables while Spearman's rank correlation measures for two variables even if they are not linearly related. [10] However, the correlation results are similar. At 0.367 (r) and 0.359 (r_s) there is a slight positive linear relationship between spontaneous abortions and infertility.[11] This is indicative of a correlation between spontaneous and case. However, the coefficient is still not very high. At 0.019 (r) and 0.019 (r_s) there is a no discernible linear relationship between induced abortions and infertility. As parity and age are both have a 0 (r_s) coefficient there is no increasing or decreasing relationship between case and these variables.

2 Data analysis

2.1 The modelling processes

Logistic regression models are most useful in this enquiry as the outcome is qualitative. Through a logistic regression model, we can find the probability of a case being fertile or infertile given any variable: $Pr\,(case\ =\ Infertile \mid X)$

The initial modelling process involved looking at a pairs plot to decide which variables to include in the model. Quantitative variables such as *induced, spontaneous* and *age* were tested first. (See Appendix 1 for pairs plot, all logistic regression models tested and their AIC value.) Eventually two

[10] *Spearman's Rank Correlation*. UniversityOfTexasAustin, Website.
[11] *Pearson Correlation and Linear Regression*. UniversityOfTexasAustin, Website.

models were chosen based on the significance of their z-test statistic and their low Akaike information criterion (A.I.C.) values.

ModelA (Figure 5) incorporates the variables *induced, spontaneous, parity* and *education*:

$$ln\left(\frac{p\,(induced)}{1-p(induced)}\right) = \beta_0 + \beta_1\ Spontaneous_{\,1} + \beta_1\ Parity_{\,1} + \beta_1\ Education_{\,1}$$

$$\text{ModelA} = \text{glm}(\text{case} \sim \text{induced} + \text{spontaneous} + \text{parity} + \text{education})$$

The summary of ModelA (figure 5) shows in the z-test all the variables are either less than 0 or greater than 0. This indicates all variables included play a role in the model. Notably, *induced,* and *spontaneous* show β>0, indicating as the number of abortions increases, the probability that case = infertile also increases. *Parity* and *education* show β<0. This indicates as the education level and number of previous pregnancies increases, the probability that case = infertile gets smaller. In ModelA, the highest z-value is *spontaneous* at 4.506, indicating *spontaneous* is important and a good indicator for whether female has secondary infertility.

The stars indicate *education* is only partially significant in ModelA. However, this variable is still included in the model as without *education* the AIC number was higher. (See appendix 1) Moreover, *education's* estimated coefficient is noteworthy. The negative coefficients far from 0 indicate the lower the *education* the higher the probability a diagnosis of secondary infertility. Therefore, education can be regarded as an important socio-economic factor effecting the probability that case = infertile.

Figure 5: Summary of logistic regression ModelA tested on training data

```
Call:
glm(formula = case ~ induced + spontaneous + parity + education,
    family = binomial(), data = inferTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7775  -0.7870  -0.5160   0.7967   2.1469

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.1481     1.1395   1.007 0.313696
induced           1.3182     0.4380   3.010 0.002614 **
spontaneous       2.0499     0.4550   4.506 6.62e-06 ***
parity           -0.9622     0.2893  -3.326 0.000881 ***
education6-11yrs -2.1349     1.0351  -2.063 0.039154 *
education12+ yrs -1.9744     1.0704  -1.845 0.065110 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 156.58  on 122  degrees of freedom
Residual deviance: 125.87  on 117  degrees of freedom
AIC: 137.87

Number of Fisher Scoring iterations: 4
```

ModeB (Figure 6) incorporates the variables *induced, spontaneous, parity* and *age:*

$$ln\left(\frac{p\,(induced)}{1-p(induced)}\right) = \beta_0 + \beta_1\ Spontaneous_{\,1} + \beta_1\ Parity_{\,1} + \beta_1\ Age_{\,1}$$

$$\text{ModelB} = \text{glm}\ (\text{case} \sim \text{induced} + \text{spontaneous} + \text{parity} + \text{age})$$

Figure 6: Summary of logistic regression ModelB tested on training data.

```
Call:
glm(formula = case ~ induced + spontaneous + parity + age, family = binomial(),
    data = inferTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6303  -0.8075  -0.5258   0.8757   2.1717

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.68408    1.45902  -1.840 0.065820 .
induced      1.32260    0.41504   3.187 0.001439 **
spontaneous  1.99444    0.43202   4.617  3.9e-06 ***
parity      -0.87894    0.26194  -3.355 0.000792 ***
age          0.05671    0.04386   1.293 0.196017
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 156.58  on 122  degrees of freedom
Residual deviance: 128.43  on 118  degrees of freedom
AIC: 138.43

Number of Fisher Scoring iterations: 4
```

In ModelB (figure 6) the highest z value is *spontaneous* at 4.617. As in ModelA this indicates *spontaneous* strongly correlates with cases of secondary infertility. Moreover, at a negative z value of -3.55 *parity* again indicates a lower number of previous pregnancies increases the probability of being diagnosed with secondary infertility. ModelB does not indicate the *age* is significant to the model. However, *age* will still be used in the logistic regression model, as including age resulted in a lower AIC number. (See appendix) Overall, it is possible ModelA is better fitted to the data than ModelB as it has the lowest AIC number.

2.2 Model assumptions and predictions

Importantly, at -0.96 and - 0.89 Models A and B indicate a negative value for *parity* increases the probability that case = infertile. This suggests the fewer previous children or conceptions the higher the probability of being diagnosed with secondary infertility. This is visualised on figure 7. This graph uses ggplot2 to illustrate how ModelB predicts the probability of secondary infertility given *parity* at a 95% confidence interval. Induced abortions are plotted along the x axis and predicted probabilities of secondary infertility are plotted along the y axis. As illustrated by the graph, a woman who has had 2 induced abortions and 5 previous conceptions has less than 0.25 probability of being infertile. However, a woman with 2 induced abortions but only 1 previous conception has just over 87.5 probability of being infertile. This indicates *parity*

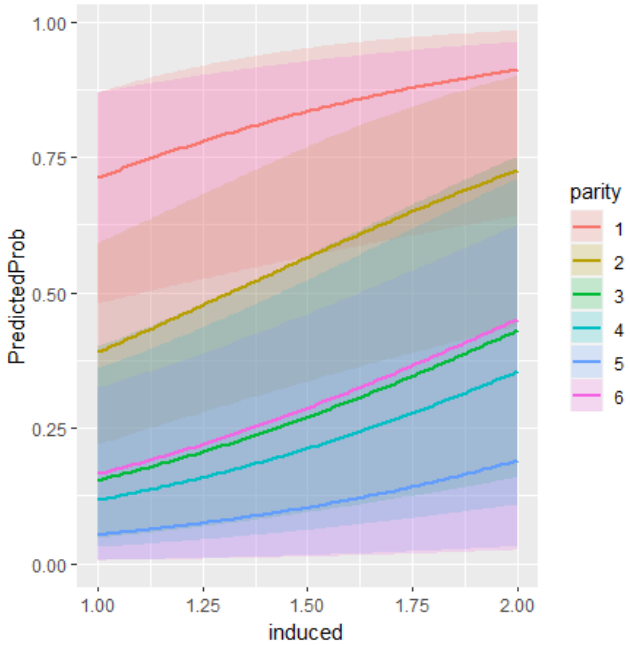


Figure 7: ggplot graph showing the predicated probability that case = infertility given number of induced abortions and number of previous pregnancies (*parity*).

effects the probability of secondary infertility. However, parity alone is not a good predictor for secondary infertility. This is because overall, the plot shows having more than one induced abortion increases the risk of secondary infertility.

By substituting the estimated regression coefficients from the model summary, we can make predictions.[12] For example, a woman with 1 induced abortion, 3 spontaneous abortions and 1 previous child with 6 years of education has an estimated 0.48 probability of secondary infertility. This is demonstrated by the binary logistic regression model for ModelA: [13]

$$\hat{p}(X) = \frac{e\ 1.15 + 1.32 * 1 + 2.05 * 3 - 0.88 * 1 - 2.13 * 6}{1 + e\ 1.15 + 1.32 * 1 + 2.05 * 3 - 0.88 * 1 - 2.13 * 6} = 0.4845017819$$

Similarly, using the binary logistic regression model for ModelB , if a woman is 33 years of age and has had 1 induced abortion, 3 spontaneous abortions and 1 previous child, her probability of secondary infertility is 0.04.

$$\hat{p}(X) = \frac{e\ -2.69 + 1.32 * 1 + 1.99 * 3 - 0.88 * 1 + 0.57 * 33}{1 + e\ -2.69 + 1.32 * 1 + 1.99 * 3 - 0.88 * 1 + 0.57 * 33} = 0.03809773175$$

These two probabilities are very different, yet they have three variables in common: *induced, spontaneous* and *parity*. Only *age* and *education* are different. It is possible ModelA predicts a higher chance of infertility than ModelB. This indicates the models may also perform differently and one may be more reliable.

2.3 Model diagnostics

The first diagnostic tool analysed is the QQ plot showing the residuals versus the expected order statistics of the standard normal distribution.[14] Both models come reasonably close to the straight line. (Figures 8 & 9) However, both models show residuals are noticeably spilt into two tails. On the bottom tail the residuals are lower than expected towards the middle, while on the upper tail the residuals are somewhat larger than expected. These plots can both be described as S-shaped, which is indicative of heavy tails or an excess of extreme values relative to the normal distribution.[15] The

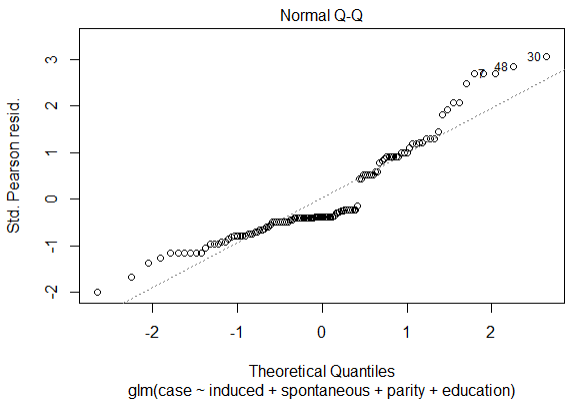


Figure 8: Normal QQ plot of ModelA

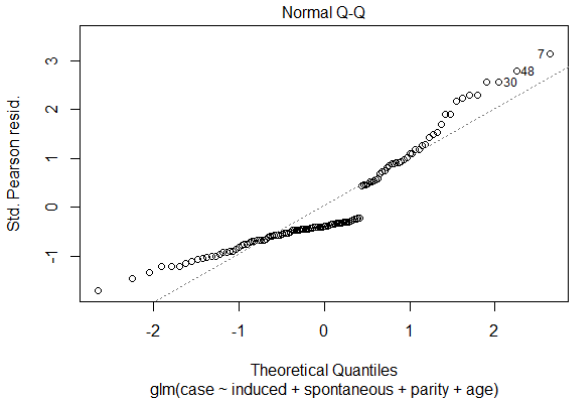


Figure 9: Normal QQ plot of ModelB

[12] James, et al. (2021) 139.
[13] Logistic Regression, PennStateUniversity.edu, Website.
[14] *Generalized Linear Models, Regression Diagnostics*, DataPrinceton.edu, Website.
[15] *Generalized Linear Models, Regression Diagnostics*, DataPrinceton.edu, Website.

curve on ModelA (figure 8) seems to be more extreme than ModelB (figure 9). Moreover, the residuals on ModelA are less predictable, some of them differing more in height and being larger at the upper tail than on ModelB. For this reason, it is possible ModelB performs better than ModelA.

The second diagnostic tool used is a confusion matrix. This provides useful diagnosis of the accuracy of the model in making predictions at a given threshold value. If the probability of infertility is greater than this threshold value, we predict infertility. However, if the probability of infertility is less than the threshold value, we predict the patient is fertile. In this case it would be better to over predict a case of infertility, so these patients could be helped sooner. The ROC curve indicates 0.3 is a good threshold value to use as it has a high sensitivity, being able to correctly classify most true infertile cases. (See appendix 2) The probability that case = infertile given any variable and a threshold value of 0.3 is written as:

$$Pr\ (case\ =\ Infertile \mid X\ =\ x)\ > 0.3$$

		True fertility status		
		Control	Infertile	Total
Predicted fertility status	Control	54	28	82
	Infertile	10	31	41
	Total	64	59	123

Table 2: ModelA glm(case ~ induced + spontaneous + parity + education) tested on training data at a threshold of 0.3

The confusion matrix shows at a threshold of 0.3 ModelA correctly predicts 54 cases of fertile patients. There are 28 mistakes indicating cases are infertile when they are fertile. There are 31 cases of true infertility predicted. 10 infertile cases are mistakenly classified as fertile. The Overall accuracy of ModelA is 0.69 ((true negative + true positive) / number of observations). A baseline prediction would predict the most frequent outcome for all observations and have an accuracy of 0.67. This means Model1 means beats the baseline accuracy by 2 percent. The overall error rate is the (false positive + false negative) / number of observations which yields 0.31.

		True fertility status		
		Control	Infertile	Total
Predicted fertility status	Control	52	30	82
	Infertile	11	30	41
	Total	63	60	123

Table 3: ModelA glm(case ~ induced + spontaneous + parity + age) tested on training data at a threshold of 0.3

ModelB does not perform as accurately as ModelA. It mistakes 30 cases of fertility for infertility. It is equally as reliable as predicting cases of true infertility, or true positive cases. The overall accuracy of ModelB is 0.66 therefore ModelB is less accurate than ModelA and the baseline. The true error rate is 0.33 which means ModelB makes more errors than ModelA. Overall, the confusion matrices indicate ModelA performs better than ModelB.

3 Resampling and validation

This enquiry uses cross validation as it is a useful approach when the outcome variable is qualitative, such as these logistic regression models with only two outcomes: fertile and, infertile.[16] The validation approach will also use k-fold cross validation, where k= 5. This is because k-fold validation empirically yields test error rate estimates which neither suffer from excessively high bias nor high variance.[17] The validation data is split into 5 folds as the data set is small. There are only 246 observations in the entire data set, therefore each fold is likely to have 48-50 observations. Splitting the validation data into 10 folds would likely distribute the data too thinly to produce useful test error rate estimates.

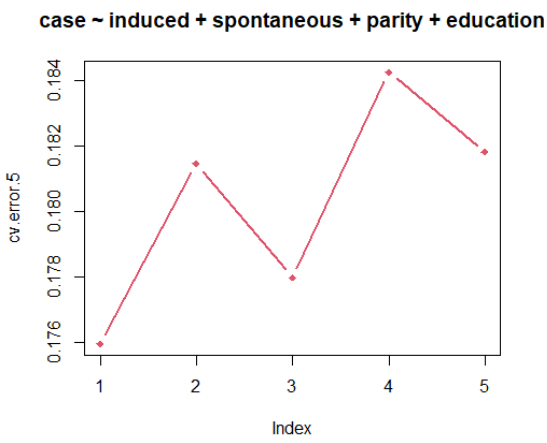


Figure 10: K-fold CV plot of CV errors for ModelA. The order of the polynomials used is displayed on the x-axis

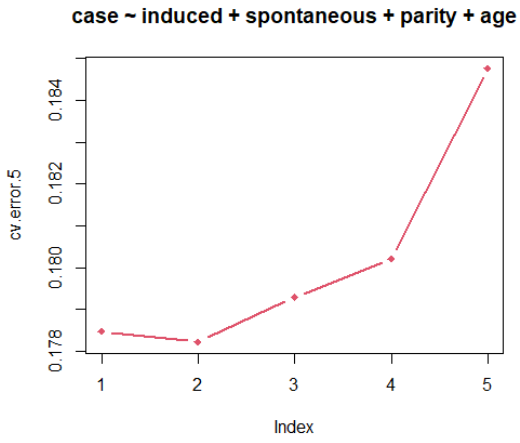


Figure 11: K-fold CV plot of CV errors for ModelB. The order of the polynomials used is displayed on the x-axis

The process of validating the logistic regression models involved setting a random seed and creating a vector to store all the CV errors corresponding to the polynomial fits of orders one to five.[18] The test error is quantified by the number of misclassified observations.[19] This is recorded as the CV error and plotted. The graphs show 5-fold CV error rates which result from fitting five logistic regression models to the data, using polynomial functions of the predictors up to the fifth order.[20] The CV errors ModelA to the four most significant digits are: 0.176 0.181 0.178 0.184 0.182. The CV errors for ModelB to the four most significant digits are: 0.178 0.178 0.179 0.180 0.185. The two models yield notably different results when validated. In ModelA the CV error rate is at its minimum when the first order of polynomial is used. While for ModelB the CV error rate reaches a minimum when the second order of polynomials is used. However, in both models, there is no evidence that higher order polynomials produce lower test errors. Using one or two order polynomials would likely lead to good test set performance.

K-fold cross validation indicates ModelB performs best. Previously the confusion matrices showed ModelB performed less accurately than ModelA when tested on the training data. However, when validated on the entire data set, this ModelB is proven to be more stable than ModelA. With more points at lower CV errors and a steady increase in error rate as the polynomial degree rises. For this

[16] James, et al. (2021) 206.
[17] James, et al. (2021) 206.
[18] James, et al. (2021) 215.
[19] James, et al. (2021) 206.
[20] James, et al. (2021) 208.

reason, it is important to use ModelB as a final logistic regression model to predict the probability of secondary infertility.

4 Discussion and limitations

4.1 Empirical findings

This enquiry found there are three crucial variables which contribute towards the likelihood a patient is diagnosed with secondary infertility. These are *induced, spontaneous* and *parity*. Importantly, without at least one induced abortion, a diagnosis of secondary is irrelevant to this study. The lower the *parity* value or the fewer previous pregnancies, the higher the probability that case = infertile. The higher the number of spontaneous abortions, the higher the probability of case = infertile. Of these three variables, only *spontaneous* is likely to be a good indicator for secondary infertility. This is indicated by *spontaneous*' z-value on both models. *Spontaneous* consistently showed the highest z-value and thus the highest relationship with infertile cases.

The significance of *age* and *education* are debatable in this enquiry. The confusion matrices indicated when *education* was accounted for, predictions were more accurate. This may suggest education is a socio-economic factor which reflects the conditions the abortion was performed in. It is possible to assume females with a lower education have abortions in worse conditions. This would reflect the study's statement that 'the circumstances of the operation would appear to be of crucial importance' whether a patient becomes infertile after an induced abortion.[21] For this reason education is an important variable in the model.

The logistic regression summary did not indicate *age* was significant when using the training data. However, *age* is likely to be a useful factor in predicting the probability that case = infertile. This is namely because when validated on unseen data, the model accounting for age was more stable and had the lowest number of CV errors.

Overall, I suggest ModelA is the better model. This is because all its variables are proven to be significant, it has the lowest AIC, and it is most reliable in correctly classifying cases of fertility and infertility.

4.2 Reliability and validity

The main issue with the two models is that the diagnostics and validation seemingly show different results for the reliability of each model. This reduces the validity of both models. I included models A and B for their low AIC numbers, and to investigate how *age* and the socio-economic variables *parity* and *education* affect the probability a case is infertile. (See Appendix 1.2) However, had I perceived these differences in diagnostics and validation I would have tested for a logistic regression model which only fitted for *induced, spontaneous* and *parity.* This may have produced a more consistently high performing model.

A limitation with the dataset is that the *spontaneous* variable only goes up to 2, while in the original data set the number of spontaneous abortions was 2 +. This could mean the models may not be valid for data which accounts for a higher parity. Had *spontaneous* accounted for higher number of spontaneous abortions than 2, the models may have been more accurate in predicting the likelihood that a case is infertile.

[21] Trichopoulos, et al. (1976) 647.

The small number of observations also causes limitations and makes the models less reliable. The training data has 123 observations and the whole data set has 246 observations. Therefore, there are arguably not enough data points to train a model that accurately predicts whether a case = infertile. The data set is possibly large enough to make useful assumptions but not large enough to make confident predictions.

As the dataset accounts for illegal abortions in the 1970's the models may not be valid or reliable in predicting cases of secondary infertility on current data. This is because the circumstances in which D&C are legally performed today may be very different to the circumstances in this data set.[22] The models were trained on data accounting for unsafe and illegal conditions, which presumably causes more cases of secondary infertility. Current data on D&C may be taken from safer settings. Therefore, it is possible if tested on current data the models may over predict cases of secondary infertility.

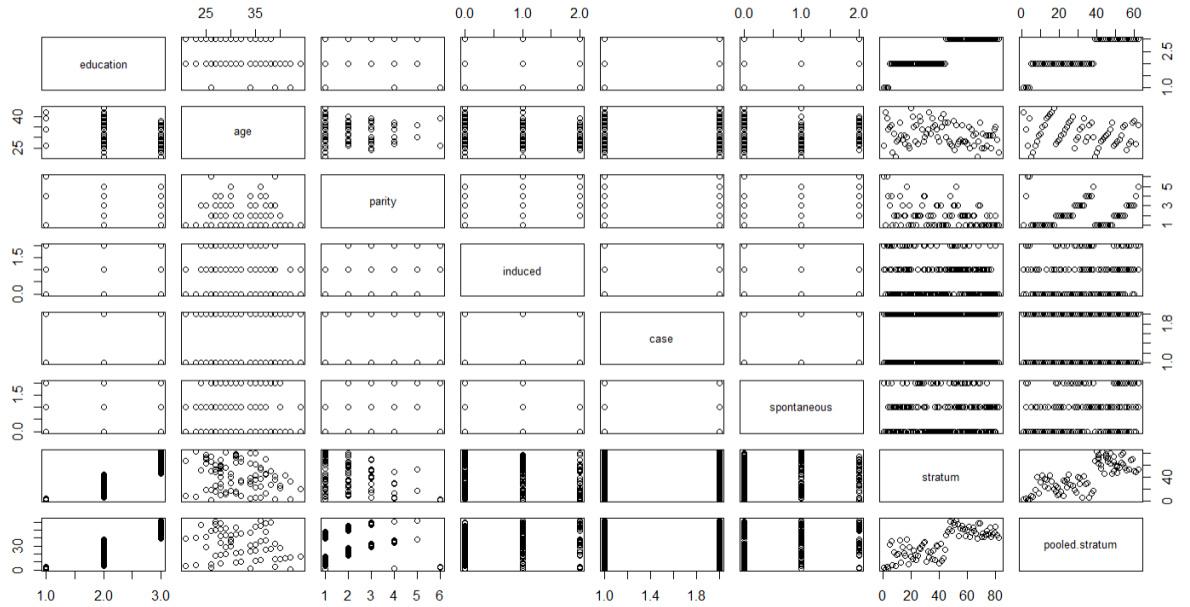
5 Conclusion

Through this enquiry, we have found how age and socio-economic factors education and parity play a role in the probability of a female being diagnosed with secondary infertility after an induced abortion. Notably, females with a low education may be more likely to be diagnosed with secondary infertility. Therefore, females with a lower education may have abortions in worse conditions. This reflects the statement that 'the circumstances of the operation would appear to be of crucial importance' whether a patient becomes infertile after an induced abortion.[23] The number of previous children and conceptions also affected the probability of becoming infertile. Females with the lowest predicted probability of having secondary infertility had more previous pregnancies. Overall, the main aspect to take away from this enquiry is that there is indeed a 'causal relationship of moderate strength' between induced abortions by dilation and curettage and secondary infertility.[24] Both models show the variable *induced* strongly correlates with cases of infertility. Moreover, the correlation coefficients for *induced* are consistently the second highest coefficient. *Spontaneous* shows the highest correlation with cases of infertility and is likely a good predictor for whether a female will be diagnosed with secondary infertility or not. This suggests when a patient suffers multiple spontaneous abortions after an induced abortion, they are likely to be diagnosed with secondary infertility. For this reason, we can assume induced abortions by D&C increase the risk of spontaneous abortions—a good indicator for secondary infertility.

Page count: 9-10

[22] Trichopoulos, et al. (1976) 647.
[23] Trichopoulos, et al. (1976) 647.
[24] Trichopoulos, et al. (1976) 647.

Appendix 1: Pairs plot and logistic models tested



1. Figure 12: Pairs plot of all variables in the *infert* data set

2. List of logistic regression models tested and their AIC number:

   glm(case ~ induced + spontaneous + parity) AIC: 138.13

   glm(case ~ induced + spontaneous) AIC: 148.04

   glm(case ~ induced + spontaneous + parity + age + education) AIC: 138.83

   glm(case ~ induced + spontaneous + parity + age) AIC: 138:43

   glm(case ~ induced + spontaneous + parity + education) AIC: 137.87

Appendix 2: ROC curves indicating good and bad threshold values

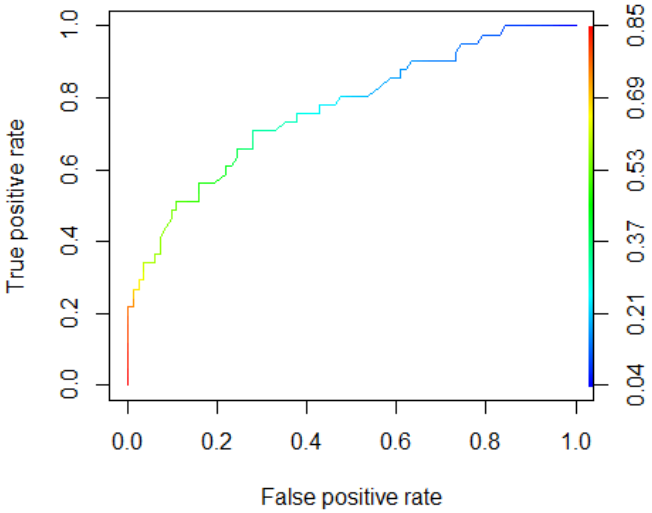


Figure 13: ROC curve for ModelA

glm(case ~ induced + spontaneous + parity + education)

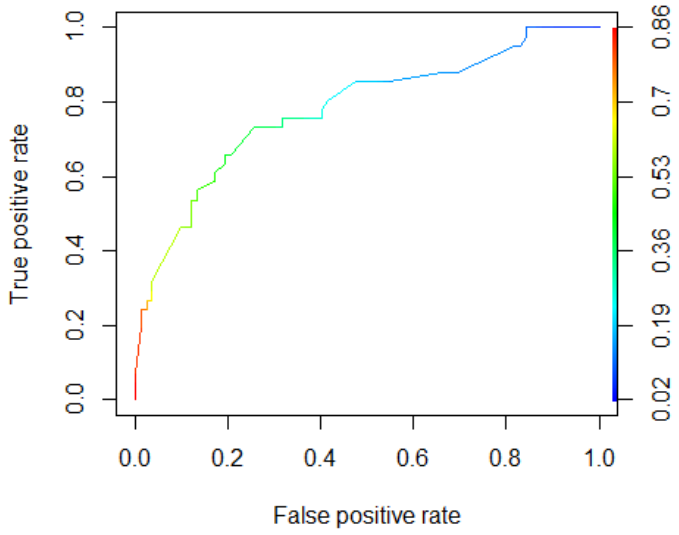


Figure 14: ROC curve for ModelB

glm(case ~ induced + spontaneous + parity + age)

Appendix 3: R script

```
data(infert)

#take away the misfitting data, the article mentioned that one patient did not have two control matches infer is our new data set

infer <- infert[(infert$stratum != 74),]

dim(infer)

#less educated implies more likely to get induced abortion, however this does not result in infertility

library(ggplot2)

ggplot(infer, aes(x=education, y=induced))+ geom_boxplot(fill='#F0ADBC')

library(ggplot2)

ggplot(infer, aes(x=education, y=spontaneous))+ geom_boxplot(fill='#509491')

library(ggplot2)

ggplot(infer, aes(x=case, y=induced))+ geom_boxplot(fill='#E6FFFB')

library(ggplot2)

ggplot(infer, aes(x=case, y=spontaneous))+ geom_boxplot(fill='#EDCCBE')

library(ggplot2)

ggplot(infer, aes(x=case, y=parity))+ geom_boxplot(fill='#EDCCBE')

str(infer)

summary(infer)

#parity = number of previous preganancies

View(infer)

pairs(infer)

#we make the dependent variable a factor

infer$case = factor(infer$case)

levels(infer$case) = c('Control', 'infertile')

summary(infer$case)

#here we see for each recorded case of infertility there are two control patients

#This is reflected in the stratum

infer$stratum = factor(infer$stratum)

infer$pooled.stratum = factor(infer$pooled.stratum)

summary(infer)
```

```
# use table to perform a baseline method for predicting the average outcome for all our data points

table(infer$case)

#separate training set and validation set

install.packages("caTools")

library(caTools)

#now randomly spilt the data into a training set and a testing set

set.seed(88)

split = sample.split(infer$case, SplitRatio = 0.50)

split

inferTrain = subset (infer, split == TRUE)

inferTest = subset (infer, split == FALSE)

nrow(inferTrain)

nrow(inferTest)

########################################

#Model 1

########################################

infertilityLog = glm(case ~ induced + spontaneous + parity + education, data = inferTrain, family =
binomial())

summary(infertilityLog)

plot(infertilityLog)

#check for colinearity

cor(inferTrain[c("induced", "spontaneous","parity")])

# look at Preliminary Correlations

cor.test(as.numeric(infer$case=="Infertile"), infer$parity, alternative = "two.sided", method =
"pearson")

cor.test(as.numeric(infer$case=="Infertile"), infer$parity, alternative = "two.sided", method =
"spearman")

glm.fit = glm(as.numeric(infer$case=="Infertile") ~ induced, data = infer, family = "binomial")

summary(glm.fit)

summary(glm.fit$fitted.values)


plot(infer$induced, as.numeric(infer$case=="Infertile"),col="red",xlab="induced",ylab="case")
```

```
points(glm.fit$data$induced,glm.fit$fitted.values, col = "black", pch = 4)

curve(predict(glm.fit,data.frame(induced = x),type="resp"),col="blue",lwd=2,add=TRUE)

#Analyse predictions--------case ~ induced + spontaneous + parity + education

residuals.glm(infertilityLog,type="pearson")

plot(residuals.glm(infertilityLog))

residuals.glmD(infertilityLog,type="deviance")

plot(residuals.glmD(infertilityLog))

predictTrain = predict(infertilityLog, type ="response")

summary(predictTrain)

tapply(predictTrain, inferTrain$case, mean)

table(inferTrain$case, predictTrain > 0.3)

install.packages("ROCR")

library(ROCR)

ROCRpred = prediction(predictTrain, inferTrain$case)

ROCRperf = performance(ROCRpred, "tpr", "fpr")

plot(ROCRperf, colorize = TRUE, print.cuttoffs.at = seq(0,1,0.1), text.adj=c(-0.2,1.7))

#####################################

#Model 2

#####################################

infertilityLog0 = glm(case ~ induced + spontaneous + parity + age, data = inferTrain, family =
binomial())

summary(infertilityLog0)

plot(infertilityLog0)

#check for colinearity

cor(inferTrain[c("induced", "spontaneous","parity", "age")])

#Analyse predictions--------case ~ induced + spontaneous + parity + age

predictTrain0 = predict(infertilityLog0, type ="response")

summary(predictTrain0)

tapply(predictTrain0, inferTrain$case, mean)

table(inferTrain$case, predictTrain0 > 0.3)

ROCRpred0 = prediction(predictTrain0, inferTrain$case)

ROCRperf0 = performance(ROCRpred0, "tpr", "fpr")
```

```
plot(ROCRperf0, colorize = TRUE, print.cuttoffs.at = seq(0,1,0.1), text.adj=c(-0.2,1.7))
########################################
#Resampling
########################################
library (boot)
set.seed (17)
cv.error.5 <- rep (0, 5)
for (i in 1:5) {
glm.fit <- glm (case ~ induced + spontaneous + parity, data = infer, family = binomial())
cv.error.5[i] <- cv.glm (infer, glm.fit , K = 5)$delta[1]
}
cv.error.5
plot(cv.error.5,type="b",col=2,pch=18,lwd=2, main="case ~ induced + spontaneous + parity")
####################################


inferTrain$parity <- factor(inferTrain$parity)
mylogit <- glm(case ~  induced + spontaneous + parity +age, data = inferTrain, family = "binomial")
summary(mylogit)


newdata1 <- with(inferTrain, data.frame(spontaneous = mean(spontaneous), age=mean(age), parity
= factor(1:6), induced = mean(induced)))
newdata1
newdata1$parityP <- predict(mylogit, newdata = newdata1, type = "response")
newdata1
newdata2 <- with(inferTrain, data.frame(induced= rep(seq(from = 1, to = 2, length.out = 100),
                                6), spontaneous = mean(spontaneous),age= mean(age), parity =
factor(rep(1:6, each = 100))))
newdata3 <- cbind(newdata2, predict(mylogit, newdata = newdata2, type = "link",
                           se = TRUE))
newdata3 <- within(newdata3, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
```

```
  UL <- plogis(fit + (1.96 * se.fit))
})
head(newdata3)
library(ggplot2)
ggplot(newdata3, aes(x = induced, y = PredictedProb)) + geom_ribbon(aes(ymin = LL,
                                                                        ymax = UL, fill = parity), alpha = 0.2) + geom_line(aes(colour =
parity),
                                                                                                                            size = 1)
```

Bibliography

Books:

James, G. et al. (2021) An Introduction to Statistical Learning with Applications in R, New York

Journals:

Azumaguchi, A. et al. (2017) 'Role of dilatation and curettage performed for spontaneous or induced abortion in the etiology of endometrial thinning', Journal of Obstetrics and Gynaecological Research (43, 3), 523–529.

Küng, S. et al. (2021) 'Factors Affecting the persistent use of sharp curettage for abortion in public hospitals in Mexico', Women's Health (17) 1-11.

Trichopoulos, D. et al. (1976) 'Induced Abortion and Secondary Infertility', British Journal of Obstetrics and Gynaecology (88) 645-650.

Websites:

Stat.Berkeley.edu,(2006) 'Factors in R' <https://www.stat.berkeley.edu/~s133/factors.html>[Last accessed 16th December 2021]

MIT Open Courseware, (2017) 'The Analytics Edge: 3: Logistic Regression' <https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/logistic-regression/> [Last accessed 16th December 2021]

Oxford Reference, (2014) 'Dilation and Curettage' <https://www.oxfordreference.com/view/10.1093/oi/authority.20110810104746617> [Last accessed 16th December 2021]

Pennsylvania State University, (2018) 'Logistic Regression' <https://online.stat.psu.edu/stat462/node/207/>[Last accessed 16th December 2021]

Rodriguez, G (2021) '*Generalized Linear Models, Regression Diagnostics*' <https://data.princeton.edu/wws509/notes/c2s9>[Last accessed 16th December 2021]

Rodriguez, G (2021) '*Generalized Linear Model,s Regression Diagnostics for Binary Data*' < https://data.princeton.edu/wws509/r/c3s8 >[Last accessed 16th December 2021]

Rodriguez, G (2021) '*Generalized Linear Models, Logit Models for Binary Data*' <https://data.princeton.edu/wws509/notes/c3s1>[Last accessed 16th December 2021]

The University of Texas Austin, (2015) 'Pearson Correlation and Linear Regression' <http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/> [Last accessed 16th December 2021]

The University of Texas Austin, (2015) 'Spearman's Correlation Rank'<http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/rankcor/>[Last accessed 16th December 2021]

UCLA,(2021) 'Logit Regression | R Data Analysis Examples '<https://stats.oarc.ucla.edu/r/dae/logit-regression/ >[Last accessed 16th December 2021]