NCSU ST 503 HW 11

Probem 2.1,2.2,4.1 Faraway, Julian J. Extending the Linear Model with R CRC Press.

Bruce Campbell

14 November, 2017

Complete exercises # 1 (d, e, f, and h), and # 2 (a - c, h) from Chapter 2. Complete exercise # 1 (a - f) from Chapter 4.

For 2.1, you have already done parts (e, f, and h) using the full 9 variable model in the group discussion. Repeat this time for the model that is chosen in part (d). Compare your results using this model to the results from the full model.

For 4.1, the question does not match the data completely.

(1) The problem says to ignore the variable "volact". Apparently, it was already ignored well enough since it does not appear in the dataset, so that is fine.

(2) For part (b) in 4.1, it says to fit the model using the other "5 variables" as predictors. There are 6 predictors to use, 5 numeric and one binary class variable, not 5, so use the 6.

2.1 wbca analysis

The dataset wbca comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

We split the data into a training and test set for model evaluation. One third of the data is reserved for the test set.

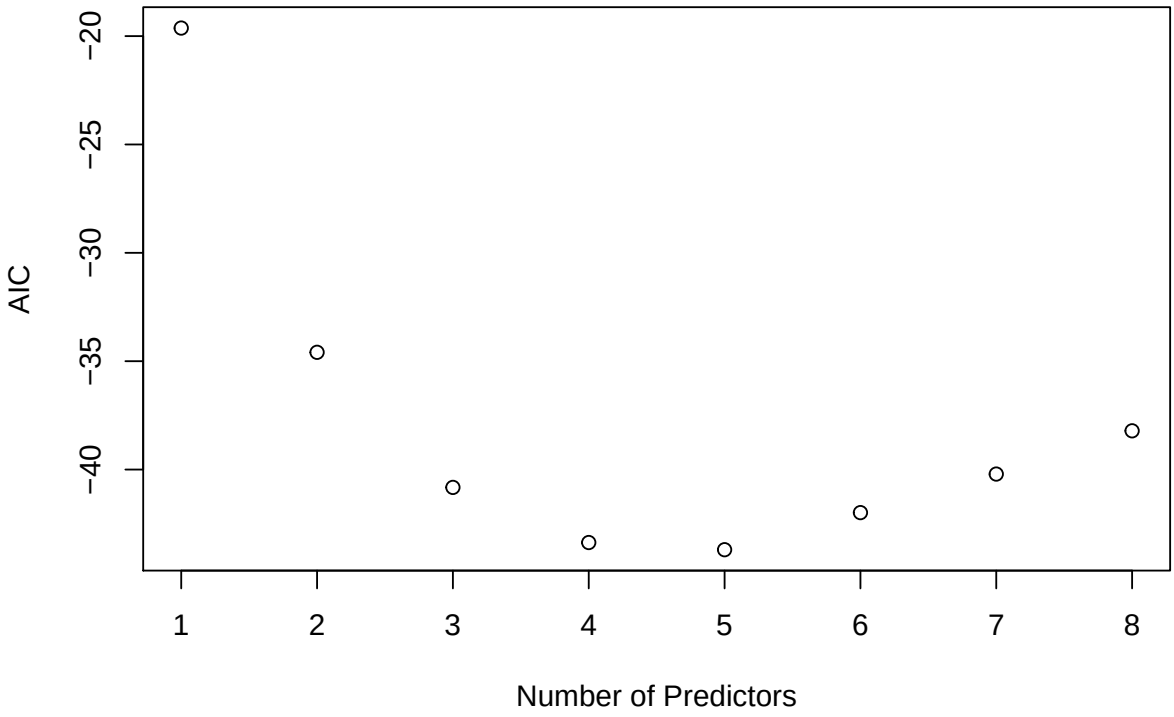
Fit a binary regression with Class as the response and the other nine variables as predictors.

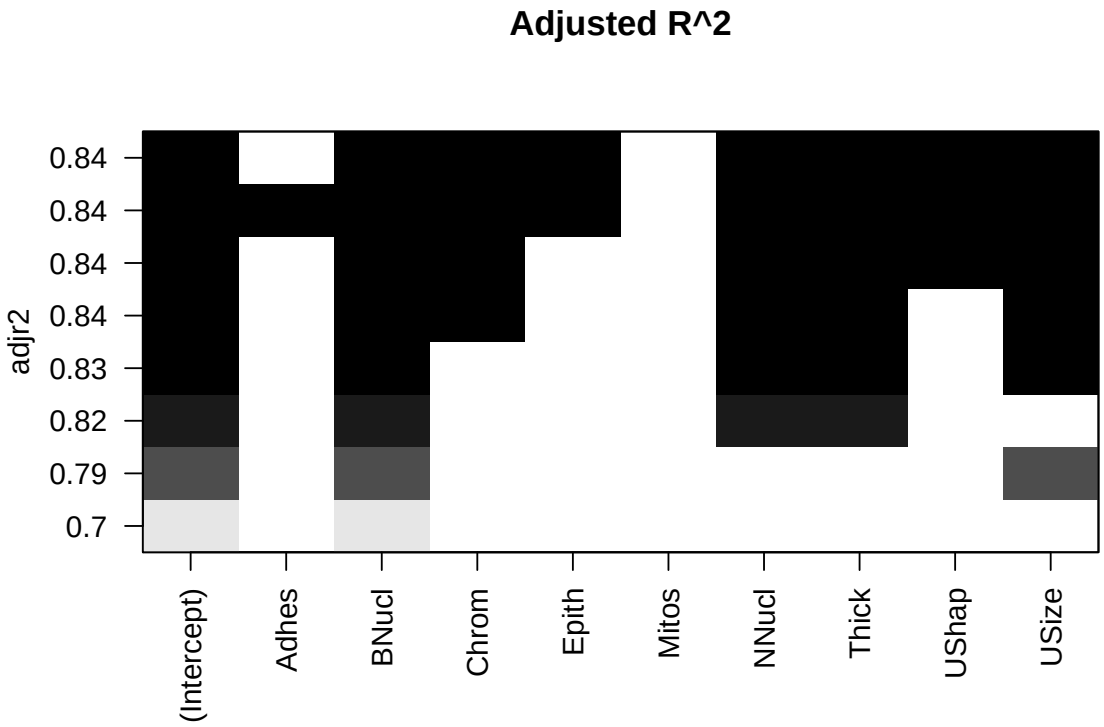
##

```
## Call:
## glm(formula = Class ~ ., family = binomial, data = DFTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3520  -0.0123   0.0406   0.0969   3.1771
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.65211    1.84168   6.327  2.5e-10 ***
## Adhes       -0.39057    0.19726  -1.980 0.047708 *
## BNucl       -0.46793    0.13566  -3.449 0.000562 ***
## Chrom       -0.65578    0.23207  -2.826 0.004716 **
## Epith       -0.02961    0.24675  -0.120 0.904469
## Mitos       -0.57934    0.51120  -1.133 0.257094
## NNucl       -0.25630    0.15143  -1.693 0.090543 .
## Thick       -0.80067    0.21174  -3.781 0.000156 ***
## UShap       -0.23802    0.26885  -0.885 0.375988
## USize        0.17773    0.24495   0.726 0.468109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 593.945  on 453  degrees of freedom
## Residual deviance:  62.044  on 444  degrees of freedom
## AIC: 82.044
##
## Number of Fisher Scoring iterations: 9
```

(d) Use AIC as the criterion to determine the best subset of variables. (Use the step function.)

```
##   (Intercept) Adhes BNucl Chrom Epith Mitos NNucl Thick UShap USize
## 1        TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 3        TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
## 4        TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 5        TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 6        TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
## 7        TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## 8        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```


AIC
−20
−25
−30
−35
−40
1
2
3
4
5
6
7
8
Number of Predictors



We see that the AIC is minimized at 4 predictors. The best 4 predictor model is $Class \sim BNucl + NNucl + Thick + USize$ We now fit that reduced model on the training data.

```
##
## Call:
## glm(formula = Class ~ BNucl + NNucl + Thick + USize, family = binomial,
##     data = DFTrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.12132  -0.04102   0.05276   0.15489   3.01438
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   8.6336     1.1110   7.771 7.80e-15 ***
## BNucl        -0.5696     0.1164  -4.892 9.97e-07 ***
## NNucl        -0.3741     0.1213  -3.084  0.00204 **
## Thick        -0.7198     0.1638  -4.394 1.11e-05 ***
## USize        -0.3936     0.1620  -2.429  0.01512 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 593.945  on 453  degrees of freedom
## Residual deviance:  82.118  on 449  degrees of freedom
## AIC: 92.118
##
## Number of Fisher Scoring iterations: 8
```

(e) Suppose that a cancer is classified as benign if p > 0.5 and malignant if p < 0.5. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.

Full model confusion matrix on training data.

```
##             class.predicted
## factor.class FALSE TRUE
##            0   157    7
##            1     7  283
```

$p = 0.5$ Full model error = `0.030837`

Reduced model confusion matrix on training data.

```
##             class.predicted
## factor.class FALSE TRUE
##            0   155    9
##            1     6  284
```

$p = 0.5$ Reduced model error = `0.0330396`

(f) Suppose we change the cutoff to 0.9 so that p < 0.9 is classified as malignant and p > 0.9 as benign. Compute the number of errors in this case.

Full model confusion matrix on training data.

```
##             class.predicted
## factor.class FALSE TRUE
##            0   163    1
##            1    12  278
```

$p = 0.9$ Full model error = `0.0286344`

Reduced model confusion matrix on training data.

```
##             class.predicted
## factor.class FALSE TRUE
##            0   164    0
```

```
##              1     14  276
```

$p = 0.9$ Reduced model error = `0.030837`

(h) It is usually misleading to use the same data to fit a model and test its predictive ability. To investigate this, split the data into two parts - assign every third observation to a test set and the remaining two thirds of the data to a training set. Use the training set to determine the model and the test set to assess its predictive performance. Compare the outcome to the previously obtained results.

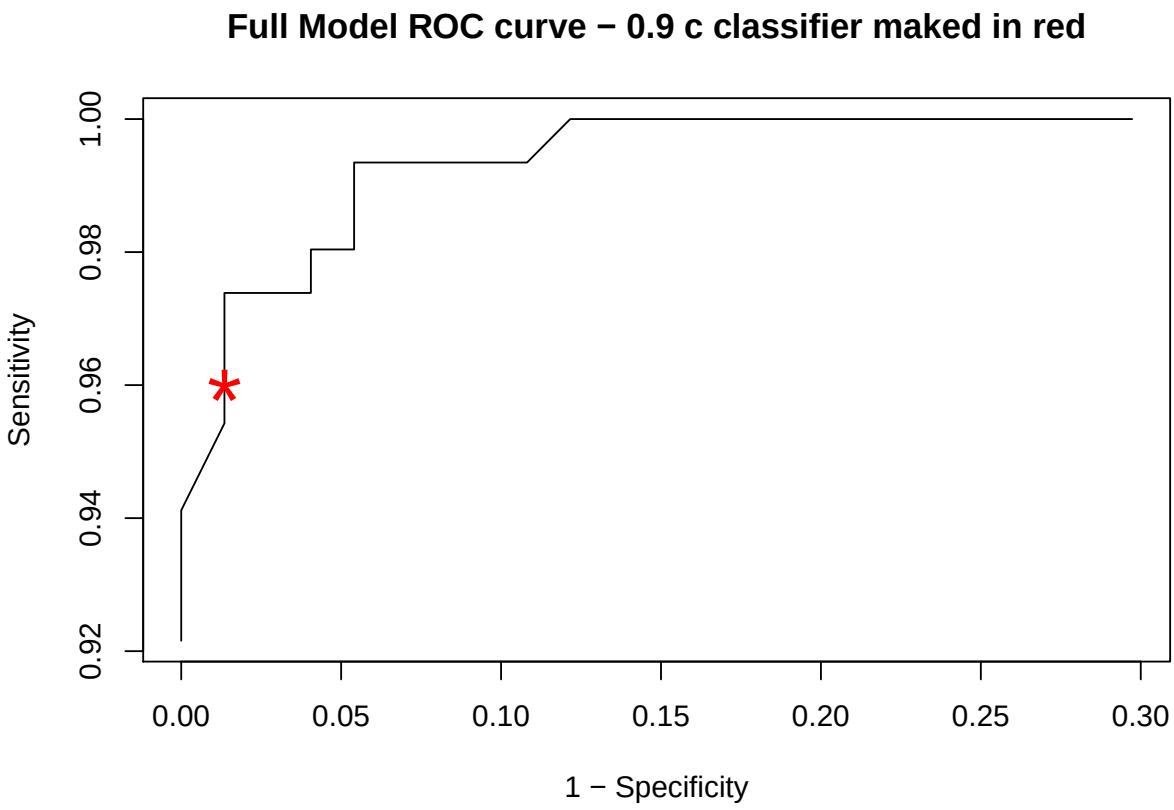
Full model test set evaluation

Table 1: Confusion matrix p=0.9

	FALSE	TRUE
0	73	1
1	6	147

Table 2: Confusion matrix p=0.5

	FALSE	TRUE
0	70	4
1	3	150



For the full model we have an accuracy of `0.030837` for the $p = 0.9$ cutoff and '`0.030837` for the $p = 0.5$ cutoff. Note that even though the accuracy is the same rate for both of these, the sensitivity and specificity will be different and there may be test implementation considerations that determine which one is preferable.

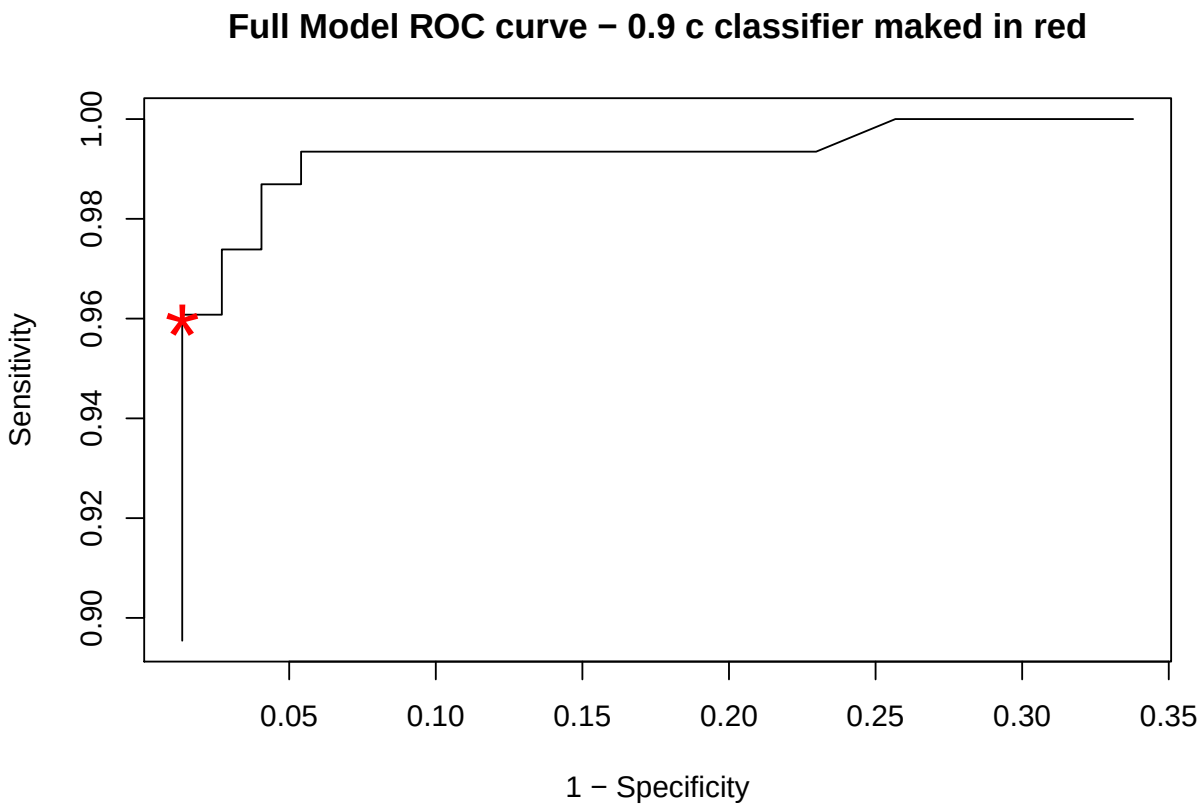
Reduced model test set evaluation

Table 3: Confusion matrix p=0.9

	FALSE	TRUE
0	73	1
1	6	147

Table 4: Confusion matrix p=0.5

	FALSE	TRUE
0	70	4
1	2	151



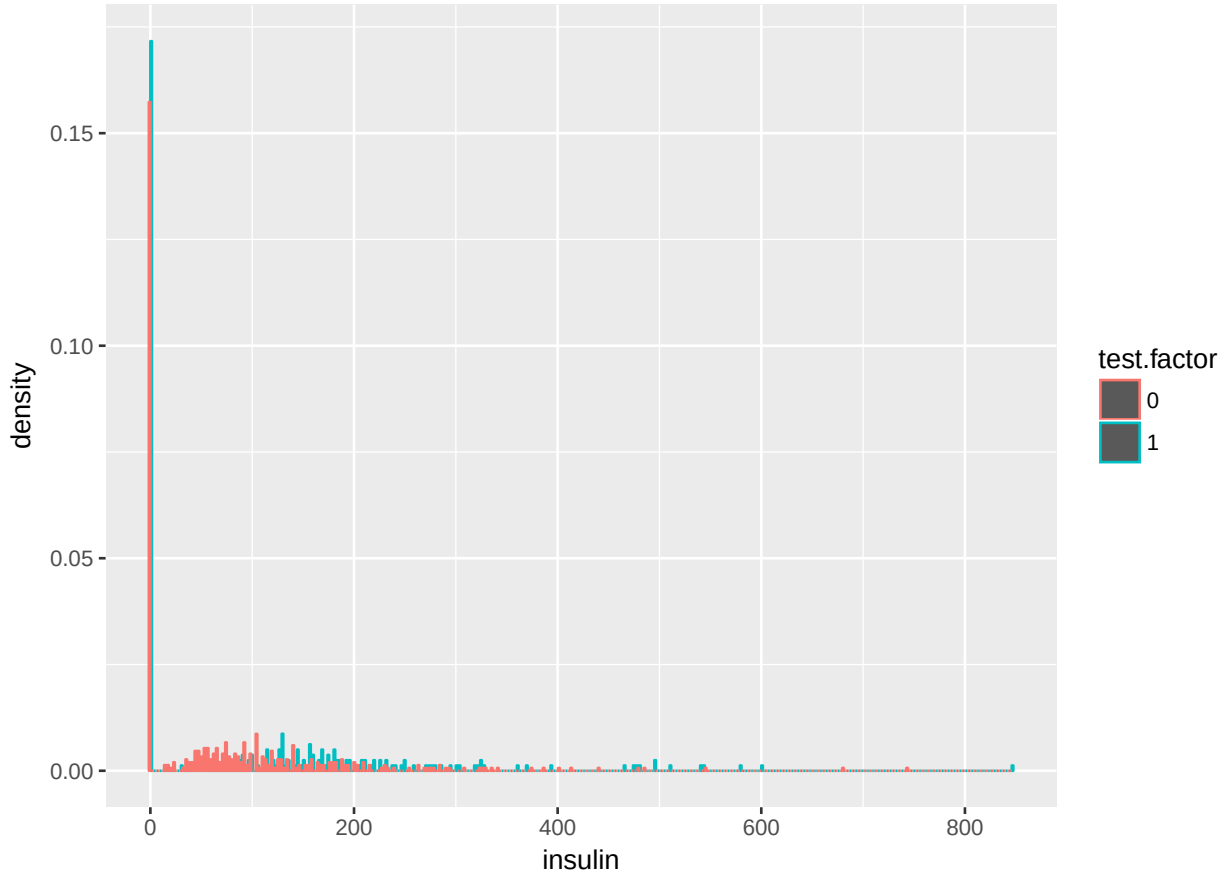
For the reduced model we have an accuracy of `0.030837` for the $p = 0.9$ cutoff and '`0.0220264` for the $p = 0.5$ cutoff.

The reduced model performs slightly better at the $p = 0.5$ cutoff. We would choose this model - all other things being equal- due to it's parsimony in the number of predictors used.

2.2 pima data analysis

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset pima.

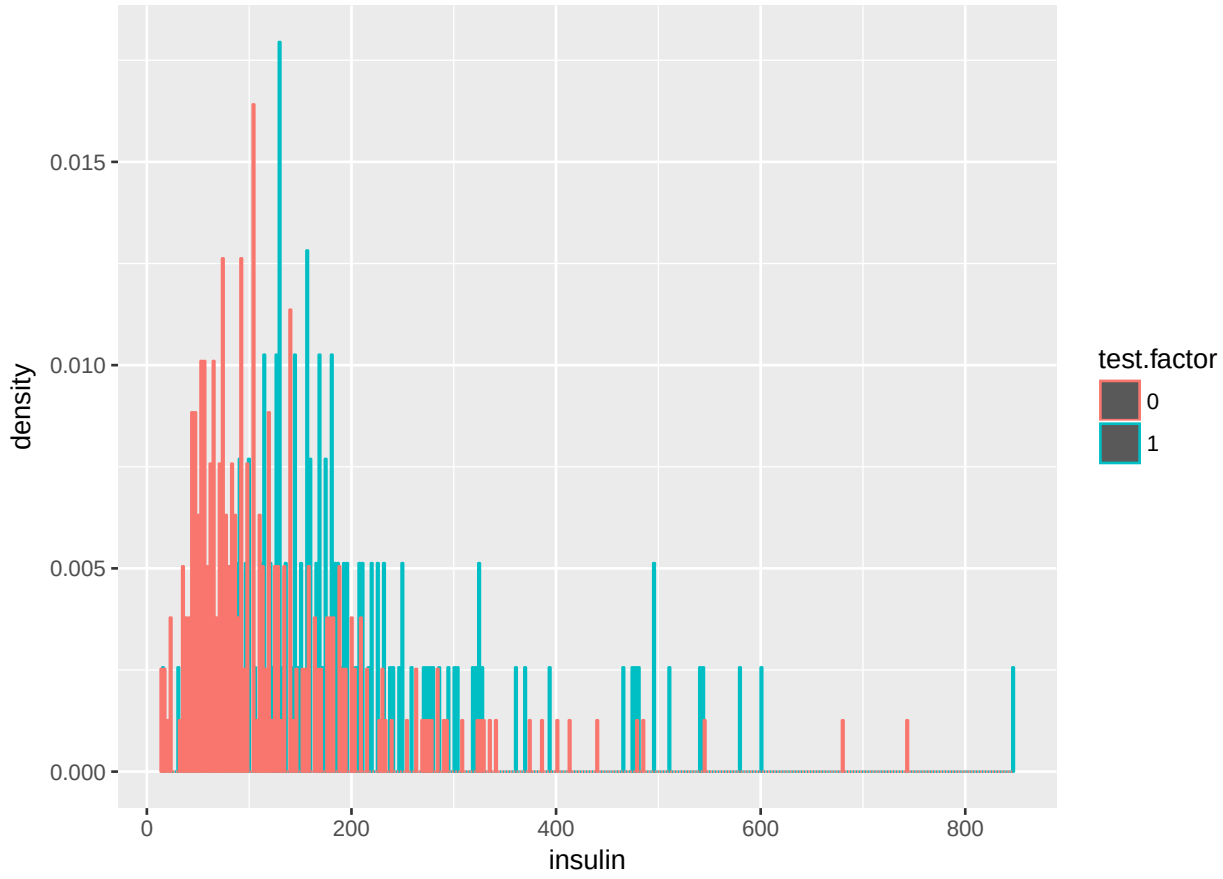
(a) Create a factor version of the test results and use this to produce an interleaved histogram to show how the distribution of insulin differs between those testing positive and negative. Do you notice anything unbelievable about the plot?



We note a number of measurements of 0 insulin. This is likely a placeholder for when no measurement was available.

(b) Replace the zero values of insulin with the missing value code NA. Recreate the interleaved histogram plot and comment on the distribution.

```
df[(df$insulin==0),]$insulin =NA
ggplot(df, aes(x=insulin, color=test.factor)) + geom_histogram(position="dodge", binwid
```



(c) Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame.

```
##
## Call:
## glm(formula = test ~ glucose + diastolic + triceps + insulin +
##     bmi + diabetes + age, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7814  -0.6675  -0.3699   0.6474   2.5697
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.016e+01  1.209e+00  -8.409  < 2e-16 ***
## glucose      3.819e-02  5.783e-03   6.605 3.97e-11 ***
## diastolic   -1.085e-03  1.174e-02  -0.092 0.926379
## triceps      1.169e-02  1.715e-02   0.681 0.495593
```

```
## insulin     -9.424e-04  1.327e-03  -0.710 0.477683
## bmi          6.660e-02  2.712e-02   2.456 0.014046 *
## diabetes     1.079e+00  4.228e-01   2.551 0.010729 *
## age          5.203e-02  1.425e-02   3.652 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 346.24  on 384  degrees of freedom
##   (376 observations deleted due to missingness)
## AIC: 362.24
##
## Number of Fisher Scoring iterations: 5
```

392 data elements were used to fit the model.

(d) Refit the model but now without the insulin and triceps predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.

```
##
## Call:
## glm(formula = test ~ glucose + diastolic + bmi + diabetes + age,
##     family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7561  -0.7236  -0.4105   0.7246   2.3652
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.872764   0.743686 -11.931  < 2e-16 ***
## glucose      0.035831   0.003551  10.092  < 2e-16 ***
## diastolic   -0.012574   0.005252  -2.394  0.01667 *
## bmi          0.093449   0.014911   6.267 3.68e-10 ***
## diabetes     0.864359   0.300074   2.880  0.00397 **
## age          0.033123   0.008124   4.077 4.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 974.75  on 751  degrees of freedom
## Residual deviance: 710.46  on 746  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 722.46
##
## Number of Fisher Scoring iterations: 5
```

Only 16 observations were removed due to missing data.

We will use $D_S - D_L \sim \chi^2_{df(L)-df(S)}$

Our test statistic is -364 with $df(L) = 384$ $df(s) = 751$

and the p-value 0.534429 is not significant so insulin and triceps are not significant in models that already have the predictors used in the smaller model.

(e) Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -10.0920180   1.0802511 -9.3423 < 2.2e-16
## glucose       0.0361890   0.0049819  7.2640 3.757e-13
## bmi           0.0744485   0.0202667  3.6734 0.0002393
## diabetes      1.0871286   0.4194084  2.5921 0.0095405
## age           0.0530121   0.0134395  3.9445 7.997e-05
##
## n = 392 p = 5
## Deviance = 347.23499 Null Deviance = 498.09781 (Difference = 150.86281)
```

The model with the minimum AIC has four predictors and the selection method found that $test \sim glucose + bmi + diabetes + age$ is the best three predictor model.

(f) Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model, but now using as much of the data as reasonable. Explain why it is appropriate to do this.

```
##
## Call:
## glm(formula = which.na ~ test, family = binomial, data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.203  -1.137  -1.137   1.218   1.218
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.09607    0.08955  -1.073    0.283
## test         0.15579    0.15152   1.028    0.304
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1064.3  on 767  degrees of freedom
## Residual deviance: 1063.3  on 766  degrees of freedom
## AIC: 1067.3
##
## Number of Fisher Scoring iterations: 3
```

We see that test is significant at a level of $\alpha = 0.3$ if this were closer to 0.2 we'd begin to wonder if there was something to investigate. This test is reasonable to execute because there may be latent variables related to the test outcome that are represented in the missing value distribution. One could imagine that insulin is hard to measure in some cases that may be related to the disease status. Likewise with the other predictors.

(g) Using the last fitted model of the previous question, what is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

```
##
## Call:
## glm(formula = test ~ glucose + bmi + diabetes + age, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8228  -0.6617  -0.3759   0.6702   2.5881
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.092018   1.080251  -9.342  < 2e-16 ***
## glucose       0.036189   0.004982   7.264 3.76e-13 ***
## bmi           0.074449   0.020267   3.673 0.000239 ***
## diabetes      1.087129   0.419408   2.592 0.009541 **
## age           0.053012   0.013439   3.945 8.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 347.23  on 387  degrees of freedom
## AIC: 357.23
##
## Number of Fisher Scoring iterations: 5
```

We know that a unit change in *bmi* yields a change in the odds ration of $exp(\beta_{bmi})$ - if all other predictors are held constant. Now we calculate the quartiles.

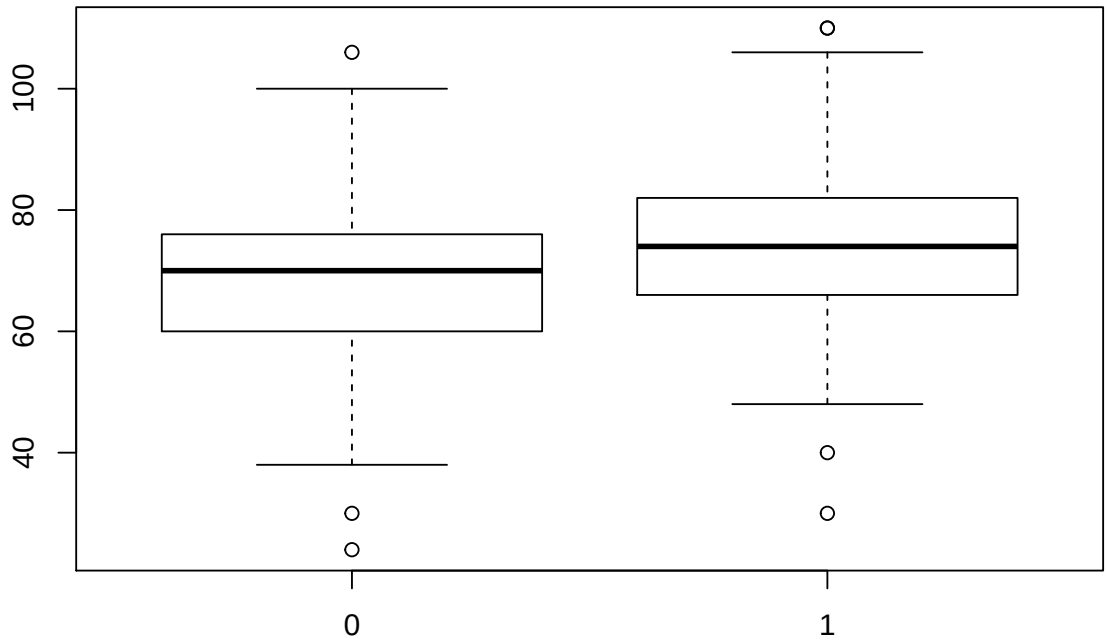
```
q1 <- quantile(df$bmi,.25)

q3 <- quantile(df$bmi,.75)

dx <- q3-q1

odds.ration.change <- exp(lm.logistic.bss$coefficients["bmi"])*dx

pander(data.frame(odds.ration.change))
```

	odds.ration.change
bmi	9.372

We see that if all other predictors are held constant - and the BMI changes from the first quratile to the third quartile, the odds ratio changes by a factor of 9.372. Making the probability of a positive test 9.372 times more likely.

(h) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

In the full model diastolic is not significant. This may be due to colliniearity. Let's plot the feature first.



We see evidence in the boxplot that elevated diastolic is associated with a positive test result. Let's create a univariate model to see.

```
##
## Call:
## glm(formula = test ~ diastolic, family = binomial, data = df.na.removed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3905  -0.9295  -0.7586   1.3246   2.1191
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.168012   0.676646  -4.682 2.84e-06 ***
## diastolic    0.034492   0.009233   3.736 0.000187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
```

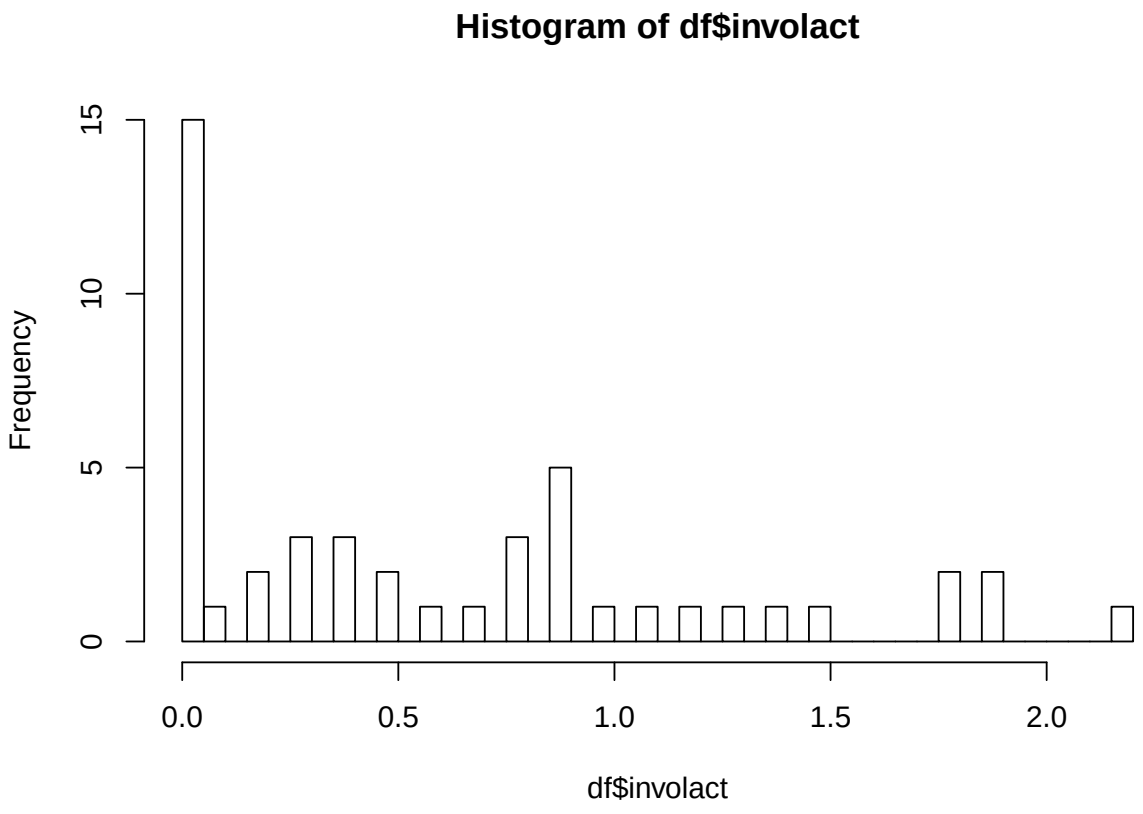
```
## Residual deviance: 483.16  on 390  degrees of freedom
## AIC: 487.16
##
## Number of Fisher Scoring iterations: 4
```

Indeed we have confirmation of a relationship between the test and diastolic.

4.1 chredlin data snalysis

The Chicago insurance dataset found in chredlin concerns the problem of redlining in insurance. Read the help page for background. Use involact as the response and ignore volact.

(a) Plot a histogram of the distribution of involact taking care to choose the bin width to illustrate the issue with zero values. What fraction of the responses is zero?



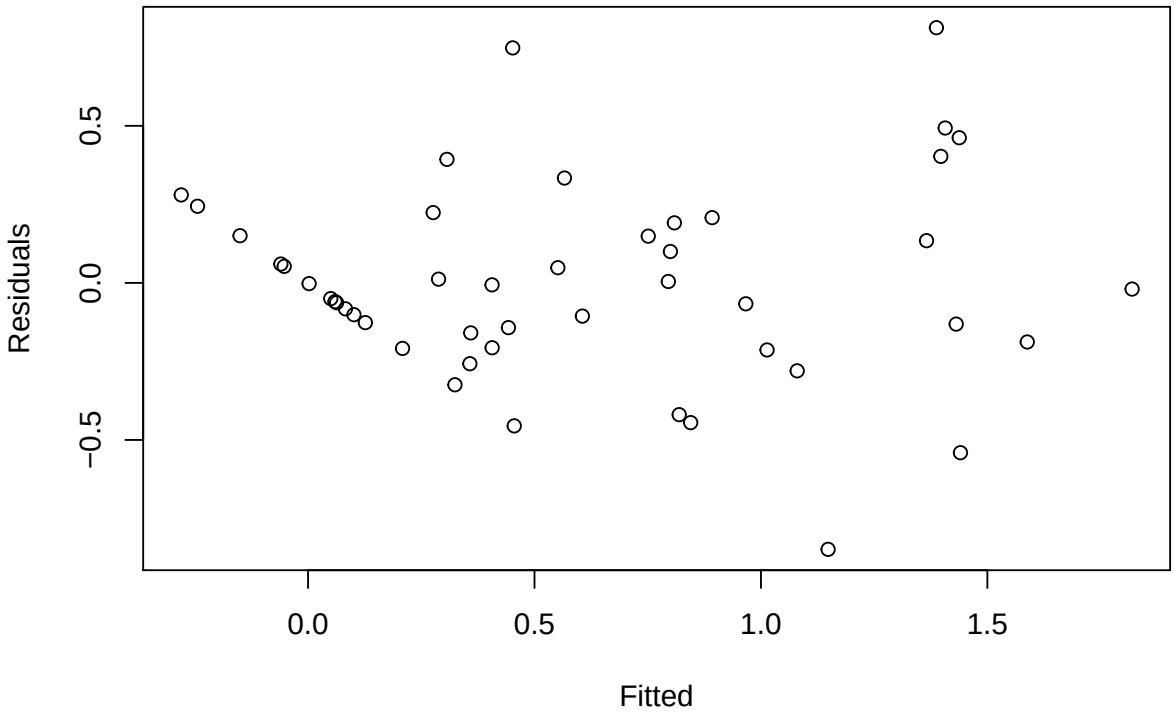
We see the proportion of zeroes is `0.3191489` Interestingly, setting freq=FALSE and using breaks did not work as expected with the `hist` function. We'll revisit this if there is time.

(b) Fit a Gaussian linear model with involact as the response with the other five variables as predictors. Use a log transformation for income. Describe the relationship between these predictors and the response.

```
##
## Call:
## lm(formula = involact ~ race + fire + theft + age + log(income) +
##     side, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84832 -0.17365 -0.01962  0.17067  0.81249
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.190806   1.114419  -1.069  0.29168
## race         0.009330   0.002847   3.277  0.00217 **
## fire         0.039994   0.008936   4.475 6.19e-05 ***
## theft       -0.010227   0.002899  -3.528  0.00107 **
## age          0.008423   0.002857   2.948  0.00532 **
## log(income)  0.343419   0.405406   0.847  0.40198
## sides        0.016287   0.124922   0.130  0.89692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3386 on 40 degrees of freedom
## Multiple R-squared:  0.7518, Adjusted R-squared:  0.7146
## F-statistic:  20.2 on 6 and 40 DF,  p-value: 1.059e-10
```

All but two of the predictors are significant in this model.

(c) Plot the residuals against the fitted values. How are the zero response values manifested on the plot? What impact do these cases have on the interpretation of the plot?



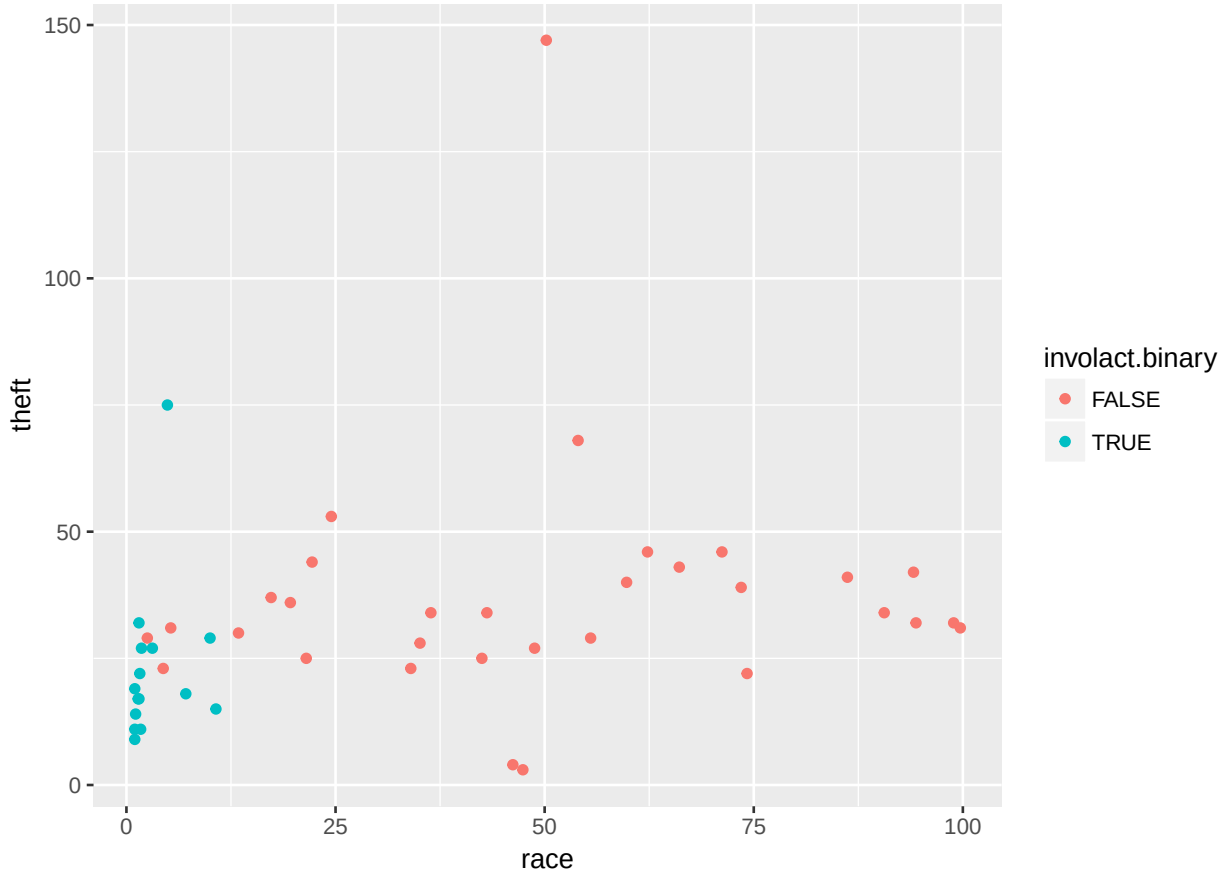
The residuals for the zero response have values have a linear association.

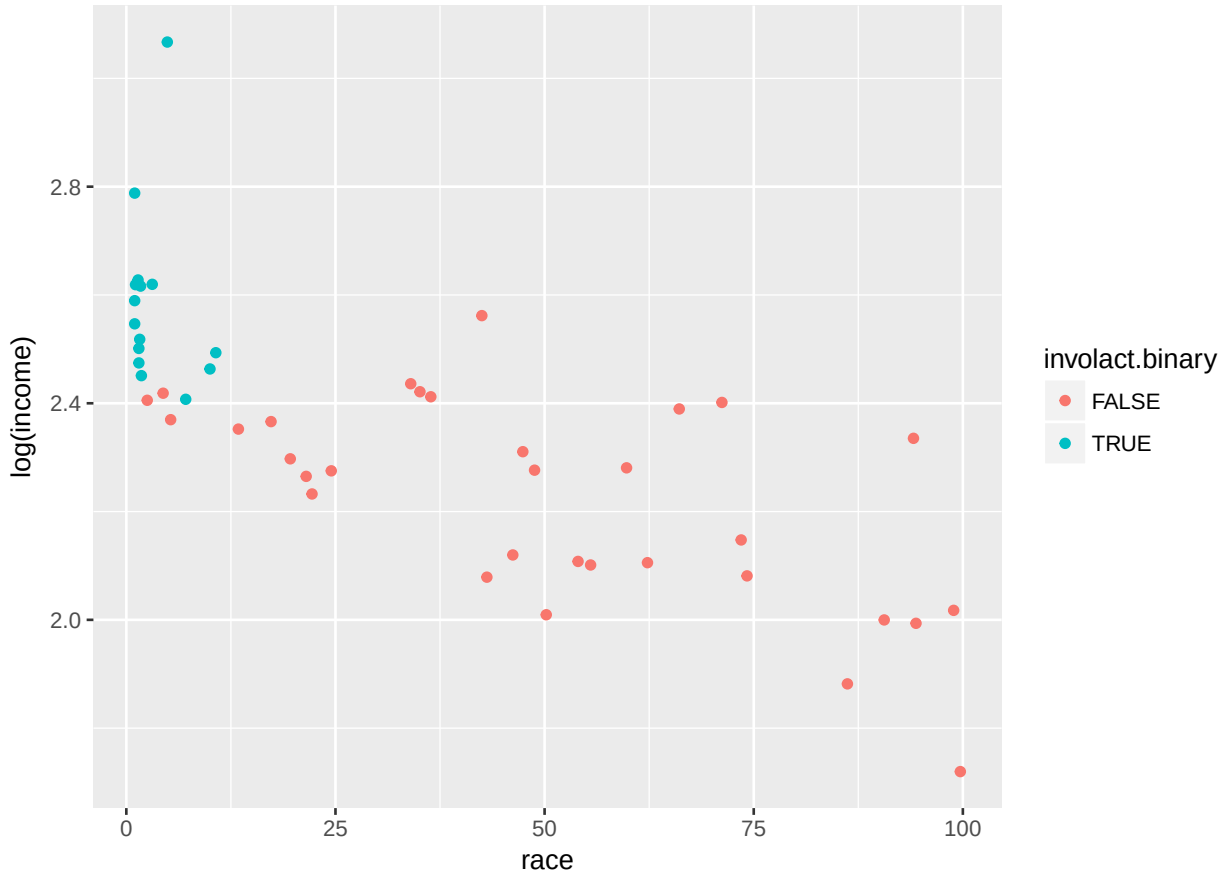
(d) Create a binary response variable which distinguishes zero values of involact. Fit a logistic regression model with this response but with the same five predictors. What problem occurred during this fit? Explain why this happened.

```
##
## Call:
## glm(formula = involact.binary ~ race + fire + theft + age + log(income) +
##     side, family = binomial, data = df)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -8.398e-05  -2.100e-08  -2.100e-08   2.100e-08   7.852e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  -1092.733 482732.211  -0.002    0.998
## race            -3.739   4174.573  -0.001    0.999
## fire            -3.861  21554.214   0.000    1.000
## theft           -3.546   2084.079  -0.002    0.999
## age             -2.635   2138.273  -0.001    0.999
## log(income)    596.601 214833.228   0.003    0.998
## sides          -62.371  28593.855  -0.002    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.8865e+01  on 46  degrees of freedom
## Residual deviance: 2.2855e-08  on 40  degrees of freedom
## AIC: 14
##
## Number of Fisher Scoring iterations: 25
```

We suspect we have perfect class separation which causes instability in the model fitting procedure. We plotted a whole bunch of 2d predictor combinations looking for this but did not encounter it. Some are below. We did not try all combinations, and some came very close. We know that if even one point crosses the linear separating hyperplane then the fitting algorithm should converge.





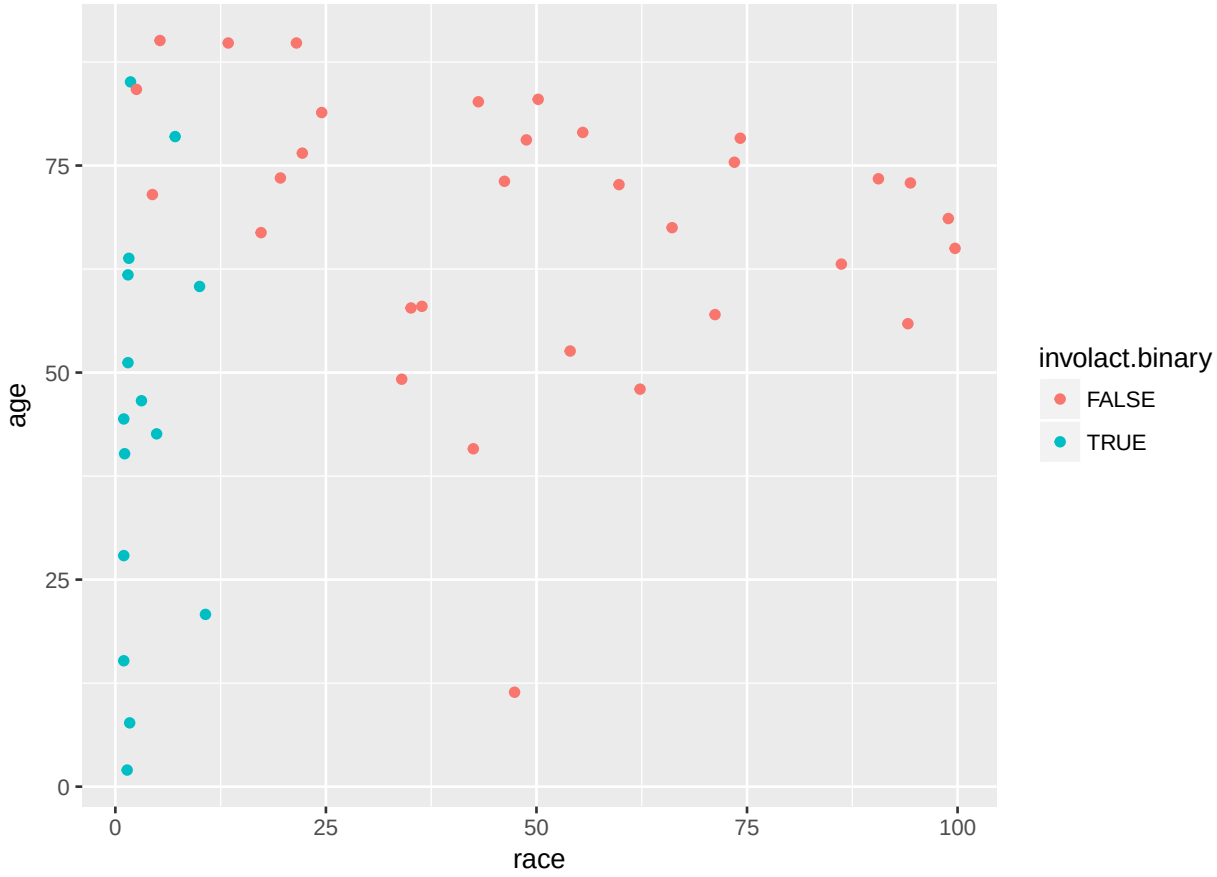
(e) Fit a smaller model using only race and age. Interpret the z-statistics. Test for the signficance of the two predictors using the difference-in-deviances test. Which test for the significance of the predictors should be preferred?

```
##
## Call:
## glm(formula = involact.binary ~ race + age, family = binomial,
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.69864  -0.04390  -0.00014   0.01286   1.50010
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.09746    7.44557   1.759   0.0786 .
## race        -0.32539    0.16602  -1.960   0.0500 *
## age         -0.14675    0.08794  -1.669   0.0952 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 58.8653  on 46  degrees of freedom
## Residual deviance:  9.2286  on 44  degrees of freedom
## AIC: 15.229
##
## Number of Fisher Scoring iterations: 10
```

The p-values indicate that all predictors are significant at a level of $\alpha = 0.1$ or below.

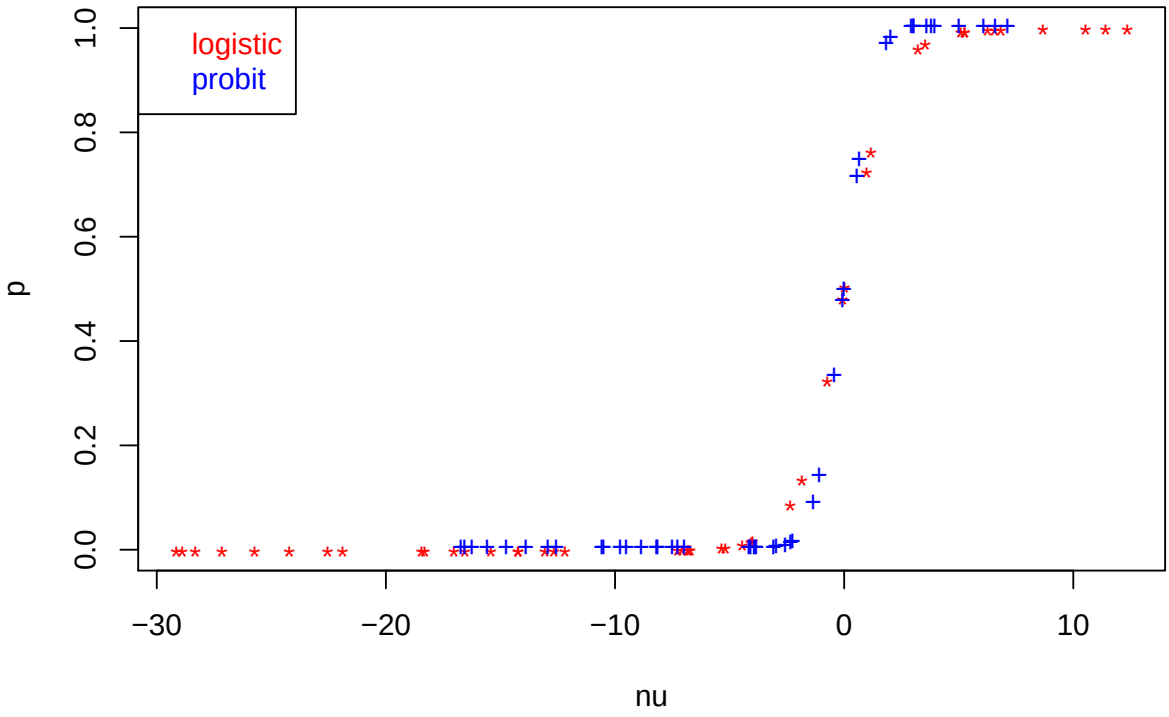
(f) Make plot of race against age which also distinguishes the two levels of the response variable. Interpret the plot and connect it to the previous model output.



(g) Refit the logit model but use a probit link. Compare the model output between the logit and probit models. Which parts are similar and which parts differ substantively? Plot the predicted values on the probability scale against each other and comment on what you see.

```
##
## Call:
## glm(formula = involact.binary ~ race + age, family = binomial(link = probit),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.65340  -0.00789   0.00000   0.00040   1.48896
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.55984    4.16874   1.813   0.0698 .
## race        -0.18655    0.08913  -2.093   0.0364 *
## age         -0.08503    0.04939  -1.722   0.0851 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 58.8653  on 46  degrees of freedom
## Residual deviance:  8.9786  on 44  degrees of freedom
## AIC: 14.979
##
## Number of Fisher Scoring iterations: 11
```

We fit $involact.binary \sim race + age$ with binomial family and probit link. We will now show the relationship between the logistic and probit models


logistic
probit
p
0.0
0.2
0.4
0.6
0.8
1.0
−30
−20
−10
0
10
nu