Homework 9

Denis Ostroushko

2022-11-15

18.2

In this problem we will make use a loop to make 100 logistic regression models that all use response variable derived from a PSA measurement variable. I describe the process of finding an optimal cutoff in the bullet points below

Before starting the loop:

1. Sort available values of PSA from lower to largest
2. index i denotes the i^{th} largest values of PSA from a list of available values

Loop outline:

1. Select i^{th} value of PSA from a sorted list
2. classify patients with the PSA values above selected cutoff as screen "positive", likely to be diagnosed, and "negative" otherwise
3. Fit a logistic regression :

$$ln\left[\frac{P(ScreenPositive = 1)}{1 - P(ScreenPositive = 1)}\right] = \hat{\beta}_0 + \hat{\beta}_1 * Cancer\ Positive\ Flag + \hat{\beta}_2 * Age + \hat{\beta}_3 * Age * Cancer\ Positive\ Flag$$

4. Record Odds Ratio for a person who is 65 years old, using this formula:

$$OR_i = e^{\hat{\beta}_1 + \hat{\beta}_3 * 65}$$

5. Save the value and return to step 1.

Code below implements such loop:

```
sorted_psa <- sort(psa$psa)
X <- 65

results <-
  data.frame(
    index = seq(from = 1, to = length(sorted_psa), by = 1),
    psa_used = sorted_psa,
    OR = NA,
    disease_b2 = NA,
    age_b3 = NA,
    interaction_b4 = NA,
    accuracy = NA,
    Sens = NA,
    PPV = NA
  )
```

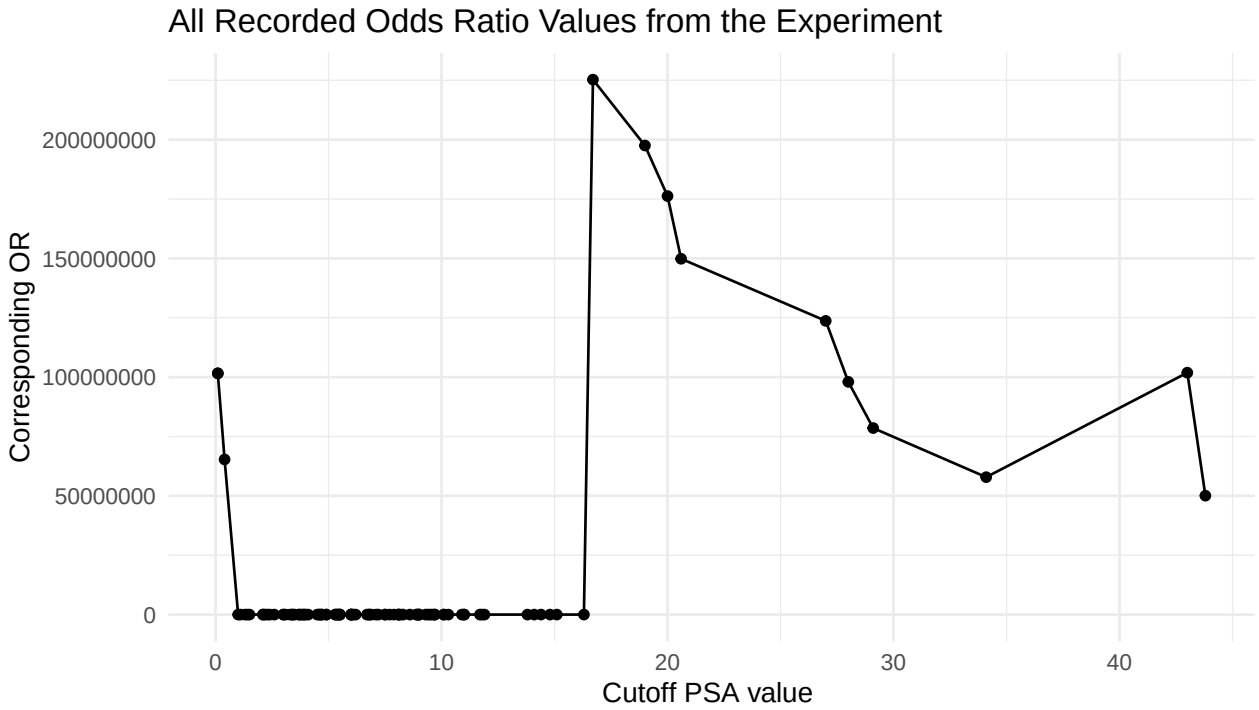
```
for(i in 1:(length(sorted_psa) - 1)){

  cutoff <- sorted_psa[i]

  psa_loop <- psa
  psa_loop$cutoff_binary <- as.factor(ifelse(psa_loop$psa > cutoff, 1, 0))

  # i = 5 is a bad fit
  # i = 25 is a good fit
  loop_glm <- glm(cutoff_binary ~ disease + age + disease:age,
                  data = psa_loop,
                  family = binomial())


  results$OR[i] <- exp(coefficients(loop_glm)[2] +
                         coefficients(loop_glm)[4] * X)

  results$disease_b2[i] <- coefficients(loop_glm)[2]
  results$age_b3[i] <- coefficients(loop_glm)[3]
  results$interaction_b4[i] <- coefficients(loop_glm)[4]

  cm <- confusionMatrix(
            data = as.factor(psa_loop$disease),
            reference = as.factor(ifelse(psa_loop$psa > cutoff, "yes", "no"))
          )

  results$PPV[i] <- cm$byClass["Pos Pred Value"]
  results$Sens[i] <- cm$byClass["Sensitivity"]
  results$accuracy[i] <- cm$byClass["Balanced Accuracy"]

}

results <- na.omit(results)
```

To find the value of PSA that maximizes OR for a patient of interest we simply select a value of PSA where OR is maximum from a data set that stores our results.

The value of PSA that maximizes OR is 16.7. So, for a person who is 65 years old, we would use values of PSA greater than 16.7 as an indicator that such patient is likely to have cancer.

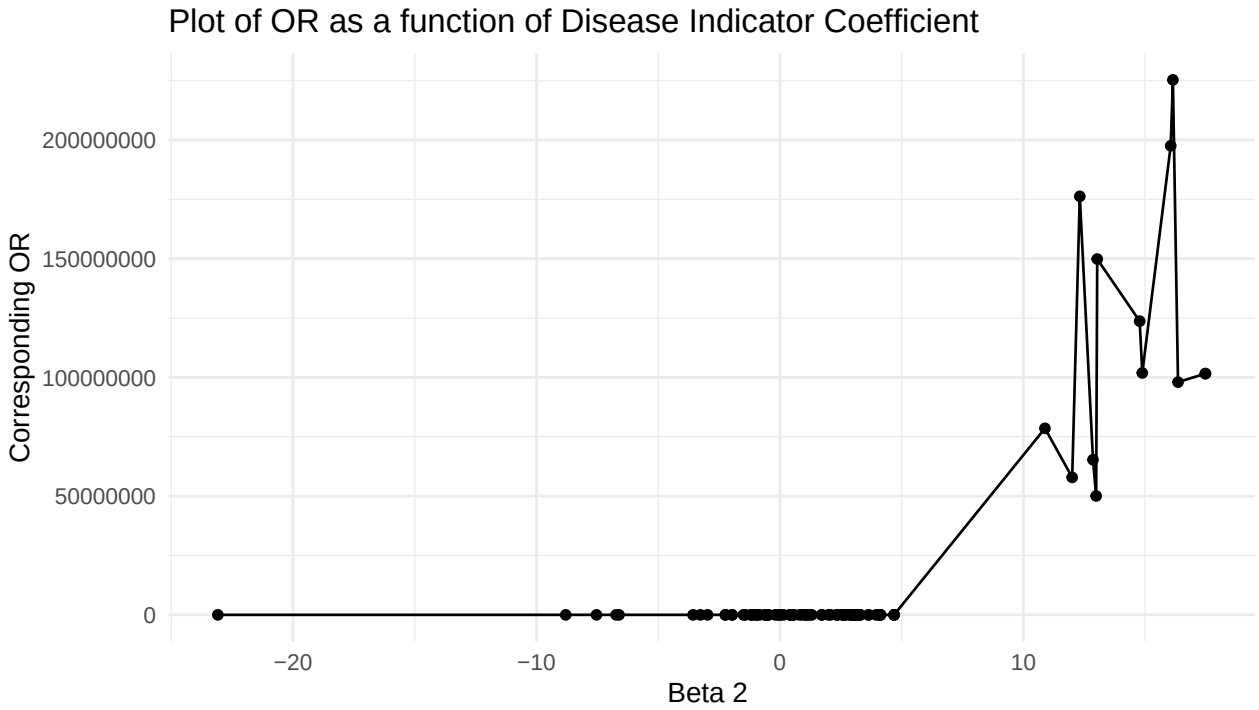
However, we also obtained some quite curious results. The value of OR we are looking for is 225283626.556382, which is just a huge value.

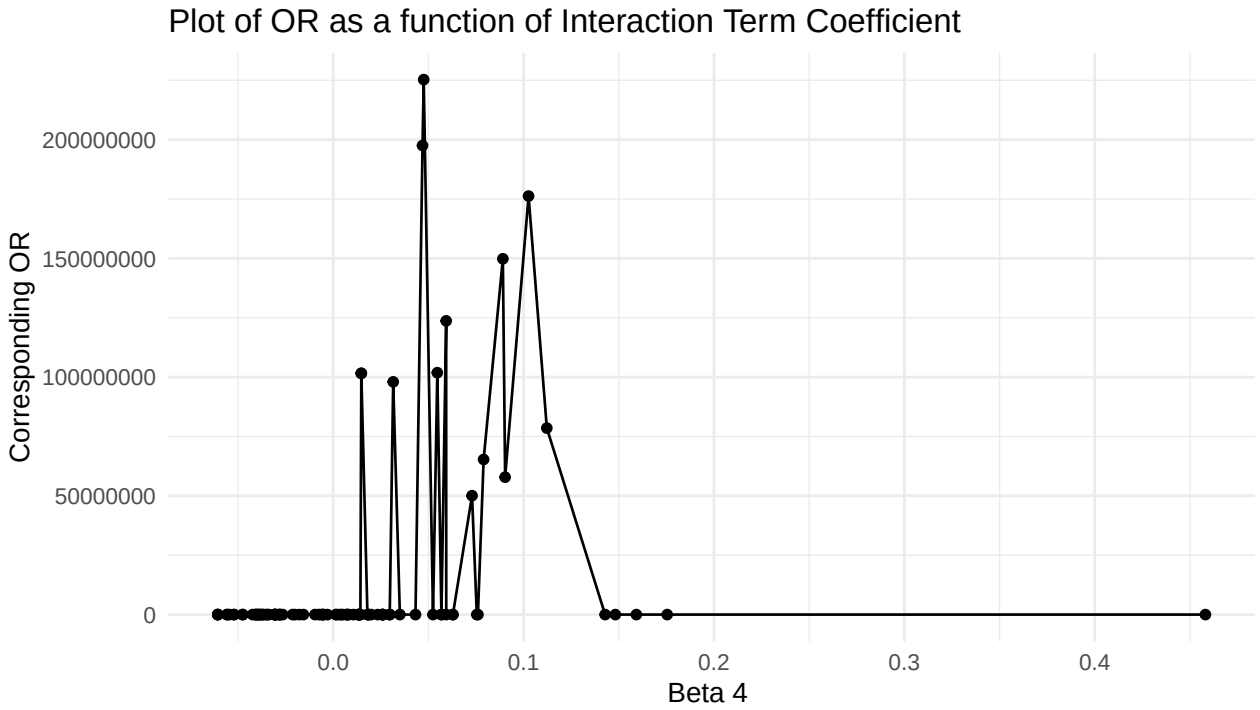
For comparison, here is a plot of all OR values plotted against the values of PSA that we used to create a binary outcome for
each iteration. As you can see, some cutoffs produce reasonable values of OR, within 0-10 range, and some are in the millions.



But how did that happen? In order to investigate these reults, we look into the coefficients of each logistic regresison model used.

We plot coefficients $\hat{\beta}_1$ and $\hat{\beta}_3$ against psa cutoff used.





So, it appears that some iterations produce models that have a huge coefficient for a disease indicator, which when exponentiated procudes a huge OR value.

19.1

19.1 - A

In this section I fit and evaluate two models. One is a strict poisson model with the assumption that mean is equal to variance, and one with a more relaxed assumption that allows for overdispersion of the response variable - the number of complaints.

Table 1: Model without Overdispersion Parameter

Predictor	Estiamte	Standard Error	Z Value	P value
(Intercept)	-7.924812	0.875680	-9.049899	0.000000
residency_flag	-0.209006	0.201152	-1.039044	0.298784
gender_flag	-0.195434	0.218153	-0.895858	0.370328
revenue	0.001576	0.002829	0.557050	0.577493
hours	0.000702	0.000351	2.002412	0.045240

Table 2: Model with Overdispersion Parameter

Predictor	Estiamte	Standard Error	T Value	P value
(Intercept)	-7.924812	1.024833	-7.732784	0.000000
residency_flag	-0.209006	0.235414	-0.887823	0.380079
gender_flag	-0.195434	0.255310	-0.765476	0.448596
revenue	0.001576	0.003311	0.475977	0.636746
hours	0.000702	0.000410	1.710982	0.095029

Overdispersion parameter for the quasi-poisson model is We can see that the model with an overdispersion parameter has the same estimates, but the standard errors are higher for the estimates. That means that without this parameter we would underestimate the variance of estimates, and potentially accept some predictors as statistically significantly related to the number of complaints, while that might not be true. For example, revenue's P-value increases from 0.0452 to 0.095 , which makes it look less important as a predictor of the number of complaints.

19.1 - B

Table 3: Model with Overdispersion Parameter

Predictor	Estiamte	Standard Error	T Value	P value
(Intercept)	-0.696508	0.928866	-0.749848	0.457846
residency_flag	-0.046898	0.238625	-0.196534	0.845213
gender_flag	-0.210065	0.228205	-0.920511	0.362965
revenue	0.003346	0.003341	1.001490	0.322764
hours	0.001118	0.000334	3.346055	0.001823

Standard errors and estimates are similar to poisson and quasipoisson models for the most part. The only difference is the residency flag variable, however, due to to high standard errors for these estimates we should not address this variability.

Interpreting Effects:

1. Hours has the only statistically significant effect between the three models. I will use quasipoisson model to interpret the results because this model should estimate the standard error of the estimate with the most accuracy.

Each additional hour worked increases the the relative risk of getting a complaint by 0.1119%, bounded by 0.0029%, 0.1405%.

19.1 - C

Observations on Estimates

- Overall, the only coefficient that changes notably is the residency flag, indicating the effect of enrollment in the residency program on the number of complaints.

  However, that coefficient is not statistically significant in any of the models, so it does not affect our model so much.

- What is more impressive is that the standard errors for each estimate is very similar between the three models, which gives mroe credibility to our results.

Residuals and Predictions

To compare the two methods we also compare the fitted values and residuals.

- Residual Plot for Quasipoisson model

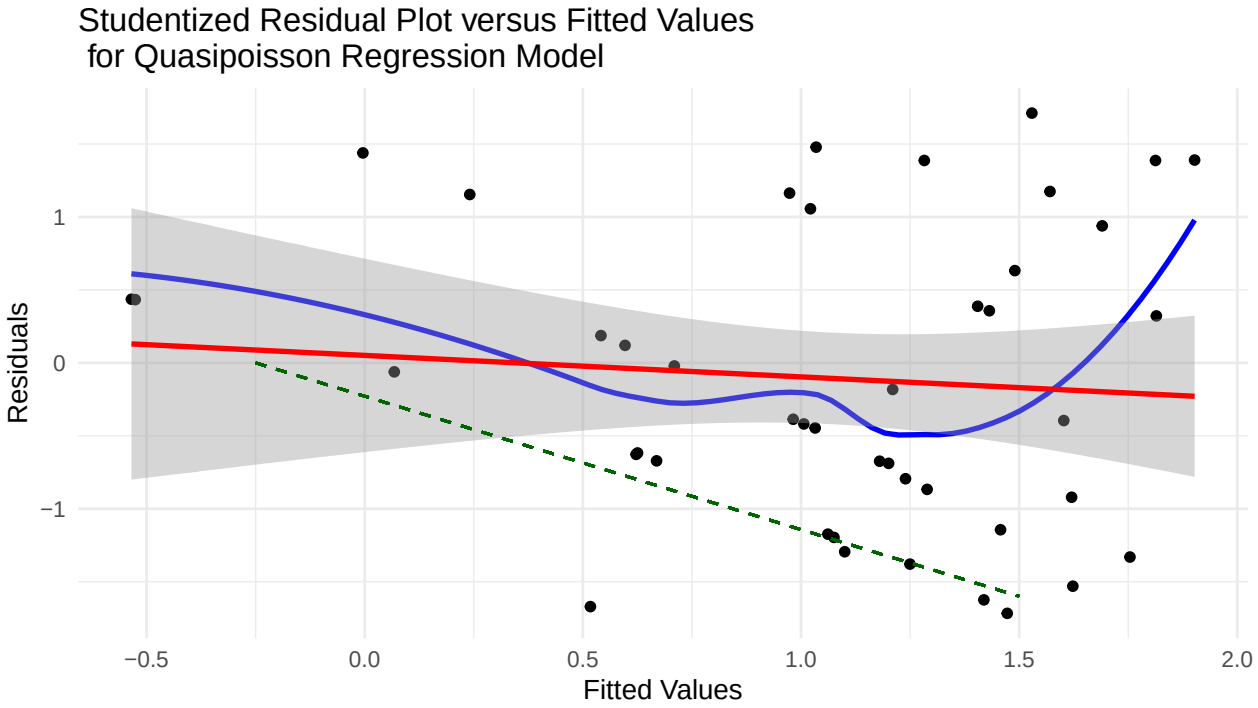
We specified the number of visits as the "offset" parameter, so out response variable is the compalint rate, number of complaints over the number of visits.

As we can see, studentized residuals fluctuate within two standard deviations from the mean zero line. As with any count data, we have the case where residuals on the lower end usually do not cross a bound, identified by the dashed green line. However, there is one outlier. We have 3 cases where fitted values are lower than 0. Fitted values are -0.5347,-0.5259,-0.0042 . Observed complaint rates for these cases are 0.001138,0.000972,0.002068.

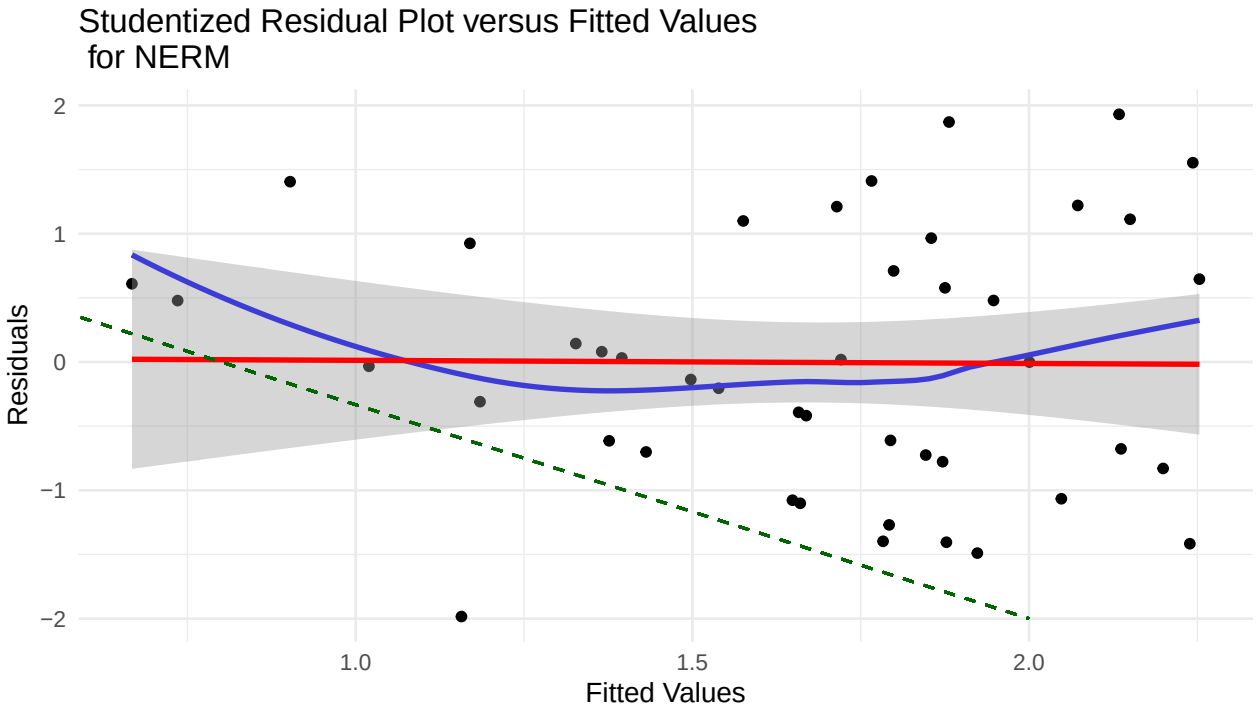
For reference, this is the distribution of complaint rates in the data:

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0006617 0.0009754 0.0013289 0.0020898 0.0030170
```

I am not sure what the implications are for the practical purposes.



- Residual Plot for NERM



Same comments apply to the cluster of residual and the one outlier below the bound.

It appears that the two model produce more of less similar results.