Altamed Data Beta Regression

Phil

11/22/2019

Introduction to Beta Regression

Typical logistic regression is often interpreted in the transformed outcome variable (log odds) and usually rely on Gaussian distribution assumption for transformed outcomes. Now, instead of assuming that our transformed outcome is normally distributed, let's assume our outcome begins by being from the Beta distribution. This is the Beta regression proposed in 2004 by Ferrari and Cribari-Neto.

The Beta regression model is based on an alternate parameterization of the Beta distribution, typically expressed as

$$f(x,p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1} \quad 0 < x < 1,$$

where $p, q > 0$ and Γ is the gamma function. Ferrari and Cribari-Neto proposed a different way of expressing the density to give a more interpretative parameterization. Set $\mu = \frac{p}{p+q}$ and $\phi = p + q$. We can then rewrite the density as

$$f(x,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1} \quad 0 < x < 1,$$

and now further, we've bounded μ and ϕ such that $0 < \mu < 1$ and $\phi > 0$. Now, we can rewrite the distribution of y as $y \sim \mathcal{B}(\mu, \phi)$. We now have properties such that $E(y) = \mu$ and $Var(y) = \frac{\mu(1-\mu)}{1+\phi}$.

Now, let's take a sample $y_1, \dots y_n$ such that each $y_i \sim \mathcal{B}(\mu_i, \phi)$ for $i = 1, \dots n$. We can define our regression model as

$$g(\mu_i) = x_i^T \beta$$

where $\beta = (\beta_1, \dots \beta_k)^T$ is a $k \times 1$ vector of unknown regression parameters, $x_i = (x_{i1}, \dots, x_{ik})^T$ is the vector of k predictors and η_i is a linear predictor. Further, g is a link function, which transforms our variable to real number line. I should note that one of main advantages of the beta regression as compared to logistic regression is that we allowing the variance of the response to be larger than it would be in logistic regression. This is in part due to the fact that the variance depends on the mean. A further extension of this model by Smithson and Verkuilen allows ϕ to vary as a series of predictors based on the mean.

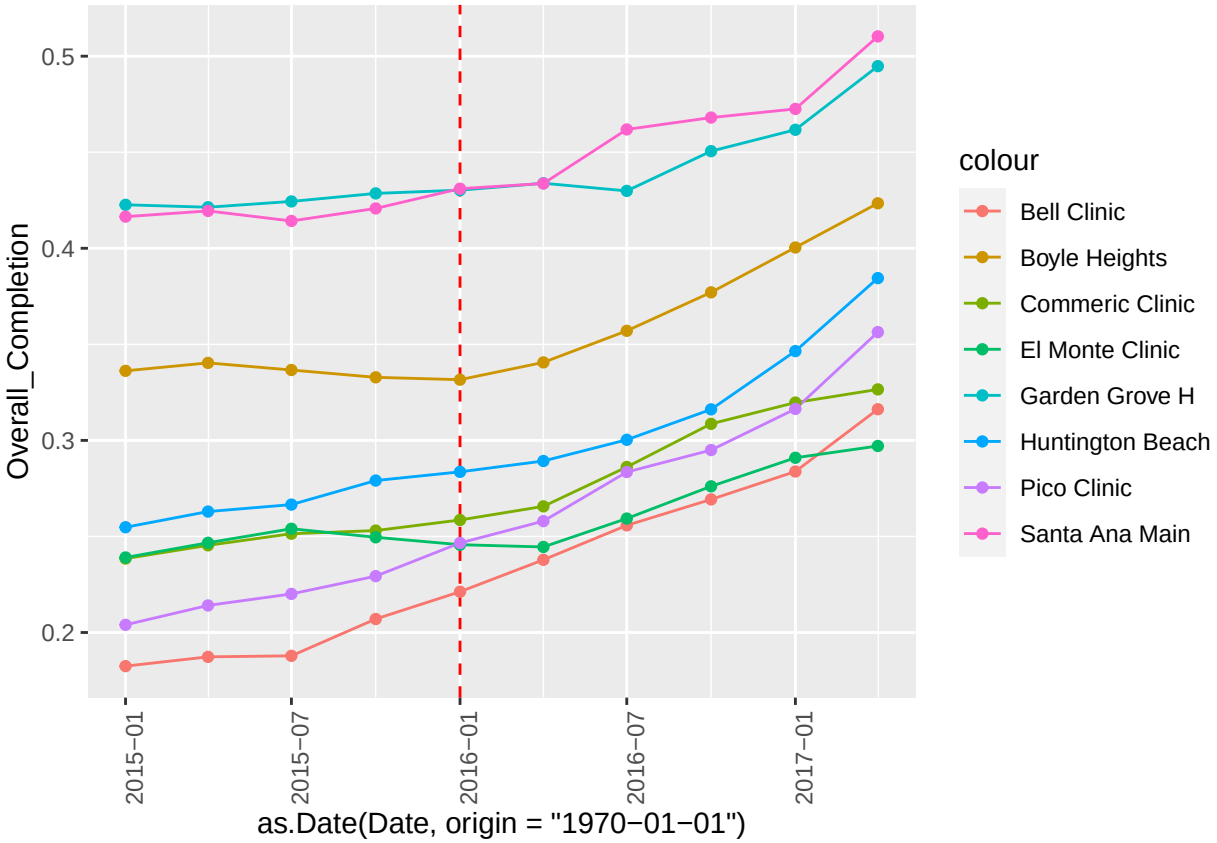
The authors states "the beta regression model shares some properties (such as linear predictor, link function, dispersion parameter) with generalized linear models (GLMs; McCullagh and Nelder 1989), but it is not a special case of this framework (not even for fixed dispersion)." It is because the beta density is not a member of the exponential family.

Altamed Dataset

Going to load in the Altamed data to see if the Beta regression is a good modeling technique for thet data.

Looking at a subset of the data, here's how it looks

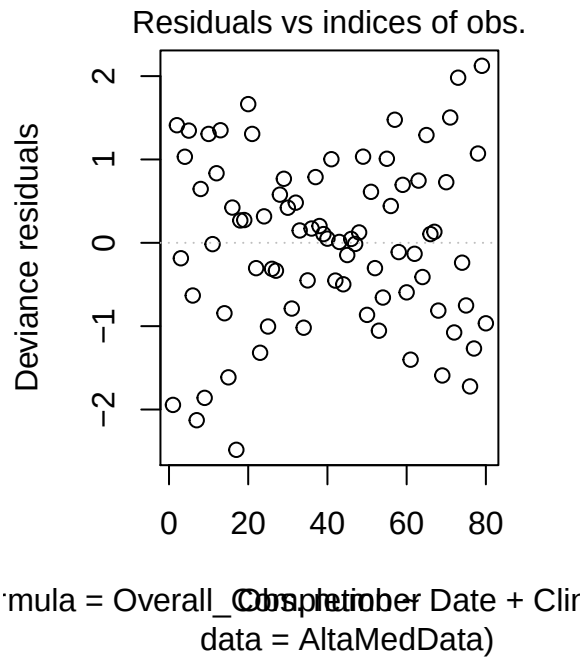
```
##             Clinic       Date Overall_Completion Overall_Initiation Time
## 1      Bell Clinic 2015-01-01          0.1824916          0.4101010    1
## 2    Boyle Heights 2015-01-01          0.3362137          0.6763551    1
## 3  Commeric Clinic 2015-01-01          0.2384805          0.6018291    1
## 4  El Monte Clinic 2015-01-01          0.2390511          0.5255474    1
## 5   Garden Grove H 2015-01-01          0.4226328          0.6734411    1
## 6 Huntington Beach 2015-01-01          0.2547893          0.5210728    1
##   Treatment
## 1         0
## 2         0
## 3         0
## 4         0
## 5         1
## 6         0
```

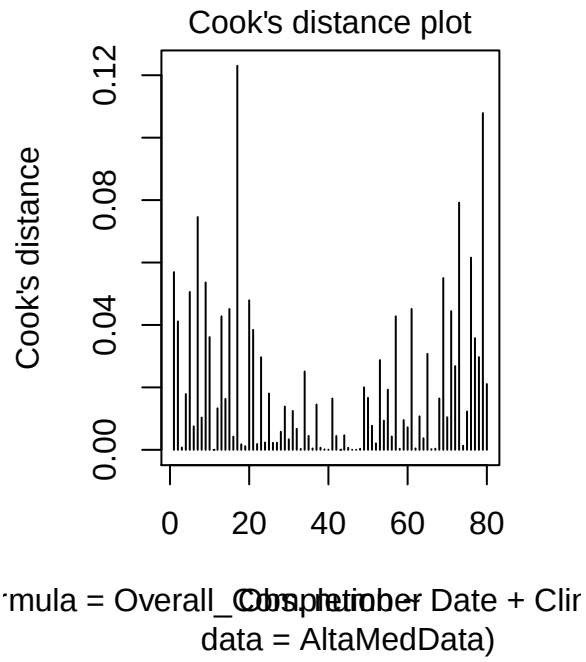


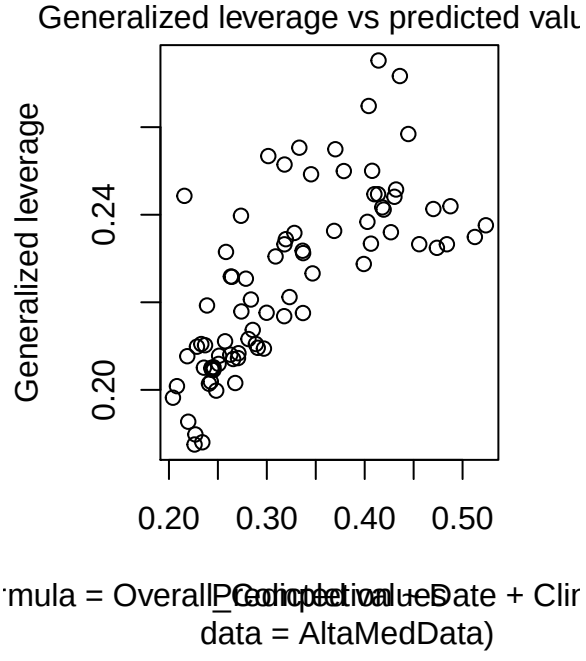
Let's start by fitting a very basic beta regression to our data, using completion percentages as the outcome.

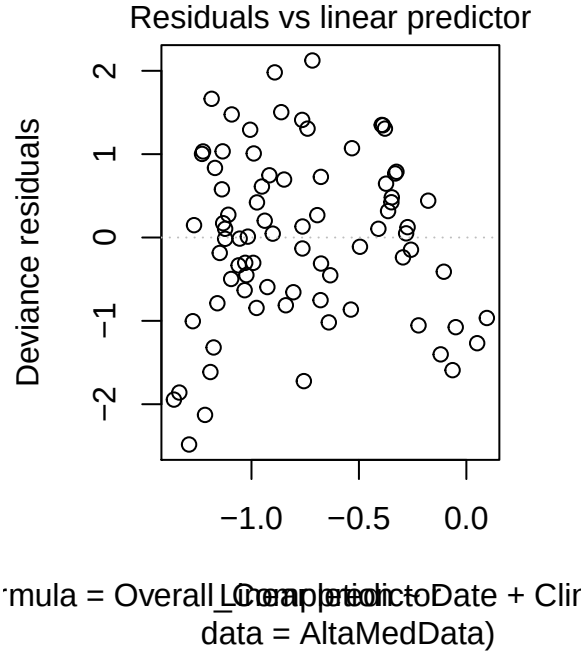
```
##
## Call:
## betareg(formula = Overall_Completion ~ Date + Clinic + Treatment,
##     data = AltaMedData)
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -2.8704 -0.7762  0.0559  0.8261  2.4231
##
## Coefficients (mean model with logit link):
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.36089    0.02995 -45.445  < 2e-16 ***
## Date2015-04-01           0.02500    0.03135   0.798  0.42510
## Date2015-07-01           0.03995    0.03130   1.276  0.20180
## Date2015-10-01           0.05737    0.03155   1.818  0.06904 .
## Date2016-01-01           0.09353    0.03116   3.002  0.00268 **
## Date2016-04-01           0.13075    0.03104   4.213 2.52e-05 ***
## Date2016-07-01           0.19624    0.03121   6.288 3.21e-10 ***
## Date2016-10-01           0.26892    0.03107   8.655  < 2e-16 ***
## Date2017-01-01           0.35476    0.03053  11.621  < 2e-16 ***
## Date2017-04-01           0.46915    0.03035  15.458  < 2e-16 ***
## ClinicBoyle Heights      0.59678    0.02801  21.306  < 2e-16 ***
## ClinicCommeric Clinic    0.21319    0.02893   7.370 1.71e-13 ***
## ClinicEl Monte Clinic    0.13551    0.02917   4.645 3.40e-06 ***
## ClinicGarden Grove H     0.94203    0.02760  34.130  < 2e-16 ***
## ClinicHuntington Beach   0.32906    0.02856  11.522  < 2e-16 ***
## ClinicPico Clinic        0.14473    0.02916   4.964 6.92e-07 ***
## ClinicSanta Ana Main     0.95663    0.02792  34.259  < 2e-16 ***
## Treatment                0.03045    0.01486   2.049  0.04047 *
##
## Phi coefficients (precision model with identity link):
##       Estimate Std. Error z value Pr(>|z|)
## (phi)   1272.6      201.2   6.327  2.5e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 235.1 on 19 Df
## Pseudo R-squared: 0.974
## Number of iterations: 40 (BFGS) + 2 (Fisher scoring)
```

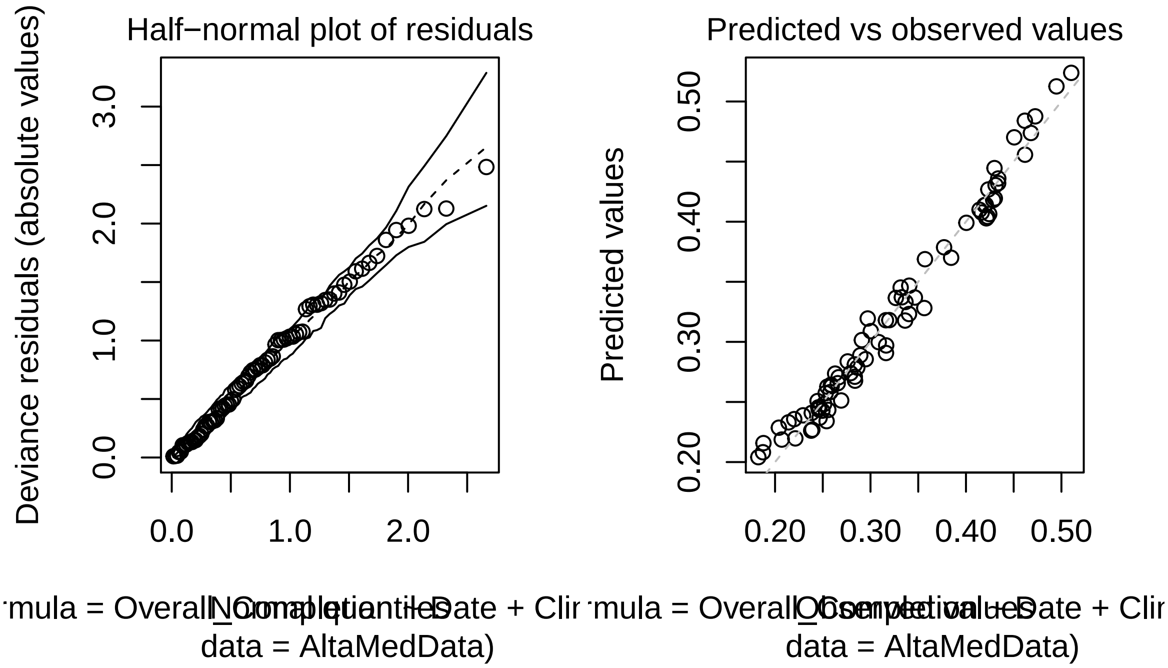
Appears to be a decent fit to the data based solely on the pseudo r-squared value. Let's look at some model diagnostics for this model:


Residuals vs indices of obs.
Deviance residuals
2
1
0
−1
−2
0
20
40
60
80
mula = Overall_Completion ~ Observation Number + Date + Clir
data = AltaMedData)


Cook's distance plot
Cook's distance
0.12
0.08
0.04
0.00
0
20
40
60
80
mula = Overall_Completion ~ Observation Number + Date + Clir
data = AltaMedData)


Generalized leverage vs predicted valu
Generalized leverage
0.24
0.20
0.20
0.30
0.40
0.50
mula = Overall_Completion ~ Predicted Values + Date + Clir
data = AltaMedData)

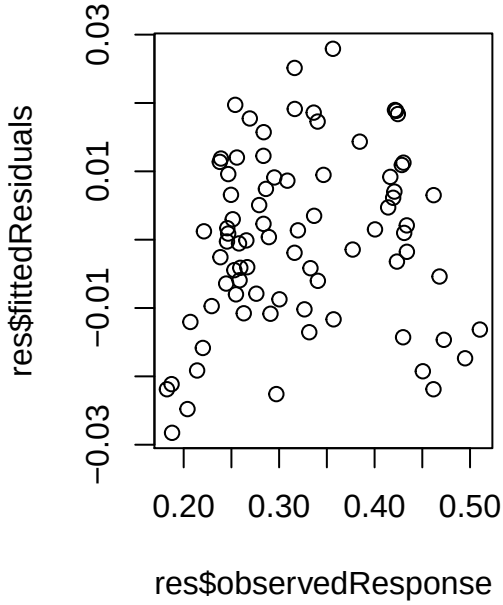

Residuals vs linear predictor
Deviance residuals
2
1
0
−1
−2
−1.0
−0.5
0.0
mula = Overall_Completion ~ Linear predictor + Date + Clir
data = AltaMedData)

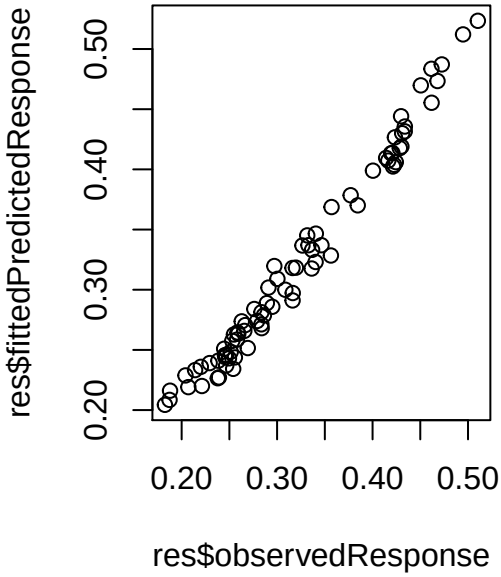

Half−normal plot of residuals
Deviance residuals (absolute values)
0.0
1.0
2.0
3.0
Normal quantiles
Predicted vs observed values
Predicted values
0.20
0.30
0.40
0.50
Observed values
data = AltaMedData)
data = AltaMedData)

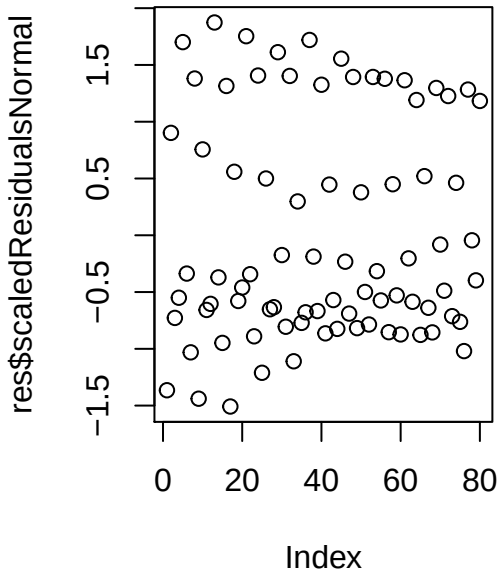
However, we are not particularly interested in inference about clinic differences. So we can make clinic a random effect. To do that, we will need to use another R package, GLMMTMB.

```
##  Family: beta  ( logit )
## Formula:          Overall_Completion ~ Date + Treatment + (1 | Clinic)
## Data: AltaMedData
##
##      AIC      BIC   logLik deviance df.resid
##   -390.2   -359.3    208.1   -416.2       67
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  Clinic (Intercept) 0.1216   0.3487
## Number of obs: 80, groups:  Clinic, 8
##
## Overdispersion parameter for beta family (): 1.15e+03
##
## Conditional model:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.94618    0.12556  -7.536 4.85e-14 ***
## Date2015-04-01  0.02497    0.03291   0.759  0.44797
## Date2015-07-01  0.03996    0.03287   1.216  0.22403
## Date2015-10-01  0.05727    0.03317   1.727  0.08423 .
## Date2016-01-01  0.09351    0.03277   2.853  0.00433 **
## Date2016-04-01  0.13077    0.03265   4.005 6.20e-05 ***
## Date2016-07-01  0.19615    0.03286   5.969 2.38e-09 ***
## Date2016-10-01  0.26882    0.03274   8.210  < 2e-16 ***
## Date2017-01-01  0.35476    0.03216  11.031  < 2e-16 ***
## Date2017-04-01  0.46910    0.03197  14.675  < 2e-16 ***
## Treatment       0.03075    0.01565   1.966  0.04934 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

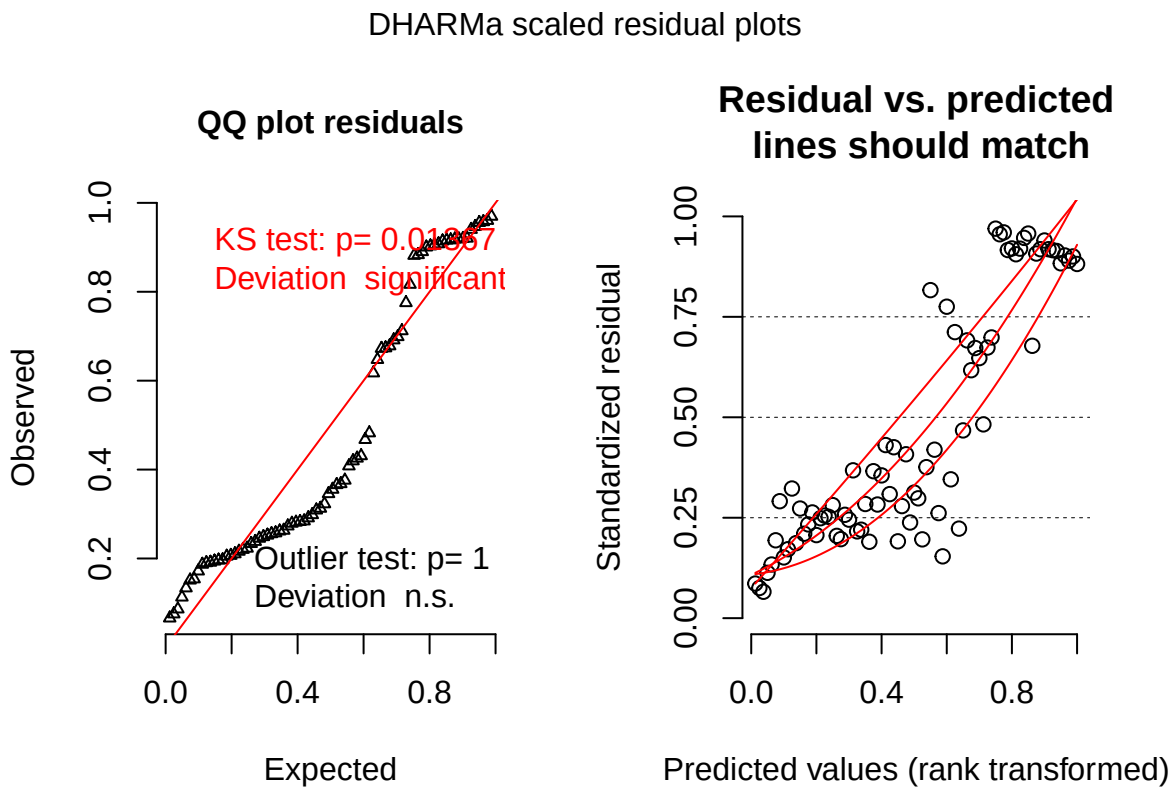
We can plot the residuals of this model, based on another package DHARMa, using simulations.


res$fittedResiduals
0.03
0.01
−0.01
−0.03
0.20
0.30
0.40
0.50
res$observedResponse


res$fittedPredictedResponse
0.50
0.40
0.30
0.20
0.20
0.30
0.40
0.50
res$observedResponse


res$scaledResidualsNormal
1.5
0.5
−0.5
−1.5
0
20
40
60
80
Index

This last plot is the residual plot taking the random effects into account, which is an issue with this package. Since our random effects are very significant, it creates a rather drastic change in our residual plots, but I figured I would show them anyway.



Regression coefficients are interpreted similarly to logistic regression. For example, a one-unit increase in a predictor X_i corresponds to a e^{β_i} increase in the odds. Looking at the output from our original model without the random intercept. We would interpret the intercept β_0 as the log(odds) of the percentage completed at time 1, at the Bell Clinic. So, $e^{\beta_0} = exp{-}1.36 = 0.25519$, which is the odds of the percentage at bell clinic at time 1. To get an estimated percentage, set $\frac{p}{(1-p)} = 0.25519$. Solving for p, we get $p = 0.2033$. If we wanted to get a predicted percentage completed for the El Monte Clinic at 2016-01-01, we would first calculate the odds, which would be $e^{-1.36574+0.16074+0.07862} = 0.32420475594$. Then, as above, we would calculate $\frac{p}{(1-p)} = 0.32420475594$, obtaining $p = 0.24482977763$. Now suppose El Monte was in the treatment group (it was not, but for interpretation's sake, suppose it is). Then, we could obtain the odds at El Monte for 2016-01-01 in the treatment group, which would $e^{-1.36574+0.16074+0.07862+0.03557} = 0.33594426861$. As above, we would calculate $\frac{p}{(1-p)} = 0.33594426861$, obtaining $p = 0.25146578079$. If we wanted to get the odds ratio for treatment group, we we take the ratio of the odds we calculated: $\frac{0.33594426861}{0.32420475594} = 1.03621018031$. Taking the log of this value, we get $log(1.03621018031) = 0.3557$, which is our estimated treatment coefficient, which is the log odds ratio!

If we want to evaluate the interpretation of coefficients, for a binary coefficient, holding other predictors constant, each β term would be the odds of $\frac{p}{1-p}$. Suppose that we were interested in the coefficient interpretation of the treatment coefficient, calculated as 0.03557. Then, we would take $\frac{p}{1-p} = e^{0.0355} = 1.036$. That is, the odds for completion percentage is 1.036 larger in the treatment group than in the groups, holding all other predictors constant. Suppose we have a continuous predictor with a coefficient of -0.02. Then, we would also calculate the estimated change in odds (or odds ratio) as $exp^{-0.02} = 0.98$. That is, for every unit increase in that predictor, we would expect the odds for completion percentage to decrease by 2%.

Variations to the model

Let's use initiation as the outcome and see how the mixed model performs:

```
##  Family: beta  ( logit )
## Formula:          Overall_Initiation ~ Date + Treatment + (1 | Clinic)
## Data: AltaMedData
##
##      AIC      BIC   logLik deviance df.resid
##   -343.2   -312.2    184.6   -369.2       67
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  Clinic (Intercept) 0.1336   0.3655
## Number of obs: 80, groups:  Clinic, 8
##
## Overdispersion parameter for beta family ():  653
##
## Conditional model:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.330398   0.132405   2.495 0.012583 *
## Date2015-04-01  0.016601   0.040376   0.411 0.680956
## Date2015-07-01  0.046572   0.040301   1.156 0.247850
## Date2015-10-01  0.111981   0.041251   2.715 0.006636 **
## Date2016-01-01  0.135312   0.040579   3.335 0.000854 ***
## Date2016-04-01  0.181043   0.040618   4.457 8.30e-06 ***
## Date2016-07-01  0.275080   0.041608   6.611 3.81e-11 ***
## Date2016-10-01  0.410293   0.041953   9.780  < 2e-16 ***
## Date2017-01-01  0.451023   0.041391  10.897  < 2e-16 ***
## Date2017-04-01  0.499648   0.041629  12.002  < 2e-16 ***
## Treatment      -0.001681   0.020256  -0.083 0.933869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And for the fixed effects model:

```
##
## Call:
## betareg(formula = Overall_Initiation ~ Treatment + Clinic + Date,
##     data = AltaMedData, link = make.link("logit"))
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -2.7838 -0.5643  0.0041  0.5847  2.9997
##
## Coefficients (mean model with logit link):
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.250175   0.034953  -7.158 8.22e-13 ***
## Treatment              -0.002271   0.019198  -0.118 0.905827
## ClinicBoyle Heights     0.870287   0.034833  24.985  < 2e-16 ***
## ClinicCommeric Clinic   0.629080   0.034076  18.461  < 2e-16 ***
## ClinicEl Monte Clinic   0.156050   0.033386   4.674 2.95e-06 ***
## ClinicGarden Grove H    0.966853   0.035200  27.468  < 2e-16 ***
## ClinicHuntington Beach  0.356561   0.033512  10.640  < 2e-16 ***
```

```
## ClinicPico Clinic       0.565937   0.033921  16.684  < 2e-16 ***
## ClinicSanta Ana Main    1.101224   0.036195  30.424  < 2e-16 ***
## Date2015-04-01          0.016695   0.038302   0.436 0.662922
## Date2015-07-01          0.046600   0.038247   1.218 0.223072
## Date2015-10-01          0.112243   0.039130   2.868 0.004125 **
## Date2016-01-01          0.135417   0.038481   3.519 0.000433 ***
## Date2016-04-01          0.181091   0.038504   4.703 2.56e-06 ***
## Date2016-07-01          0.275362   0.039483   6.974 3.07e-12 ***
## Date2016-10-01          0.410593   0.039819  10.312  < 2e-16 ***
## Date2017-01-01          0.451137   0.039293  11.481  < 2e-16 ***
## Date2017-04-01          0.499807   0.039554  12.636  < 2e-16 ***
##
## Phi coefficients (precision model with identity link):
##       Estimate Std. Error z value Pr(>|z|)
## (phi)    725.9      114.7   6.329 2.47e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood:   210 on 19 Df
## Pseudo R-squared: 0.965
## Number of iterations: 35 (BFGS) + 2 (Fisher scoring)
```

Given our data structure, we might also consider removing the logit function entirely and working with percentages. I think this could be a viable approach since our predictors are all indicators. So far, can only do this with a fixed effects approach. Would make the coefficients much more interpretable. Has a higher likelihood value and improved psuedo r-squared value compared to the logit function.

```
##
## Call:
## betareg(formula = Overall_Initiation ~ Treatment + Clinic + Date,
##     data = AltaMedData, link = make.link("identity"))
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -2.9526 -0.7470  0.0973  0.7456  2.7216
##
## Coefficients (mean model with identity link):
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.4424520  0.0083819  52.787  < 2e-16 ***
## Treatment               0.0002696  0.0042679   0.063 0.949625
## ClinicBoyle Heights     0.2051169  0.0079036  25.952  < 2e-16 ***
## ClinicCommeric Clinic   0.1512809  0.0080684  18.750  < 2e-16 ***
## ClinicEl Monte Clinic   0.0383740  0.0082332   4.661 3.15e-06 ***
## ClinicGarden Grove H    0.2241355  0.0078348  28.608  < 2e-16 ***
## ClinicHuntington Beach 0.0879600  0.0081763  10.758  < 2e-16 ***
## ClinicPico Clinic       0.1388438  0.0080993  17.143  < 2e-16 ***
## ClinicSanta Ana Main    0.2510318  0.0078210  32.097  < 2e-16 ***
## Date2015-04-01          0.0036263  0.0089689   0.404 0.685975
## Date2015-07-01          0.0108552  0.0089454   1.213 0.224939
## Date2015-10-01          0.0257466  0.0090356   2.849 0.004379 **
## Date2016-01-01          0.0311148  0.0089168   3.489 0.000484 ***
## Date2016-04-01          0.0415453  0.0088754   4.681 2.86e-06 ***
## Date2016-07-01          0.0628766  0.0089414   7.032 2.03e-12 ***
## Date2016-10-01          0.0919124  0.0088552  10.380  < 2e-16 ***
## Date2017-01-01          0.1004806  0.0086825  11.573  < 2e-16 ***
## Date2017-04-01          0.1113960  0.0086467  12.883  < 2e-16 ***
##
## Phi coefficients (precision model with identity link):
##       Estimate Std. Error z value Pr(>|z|)
## (phi)    732.7      115.8   6.329 2.48e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 210.3 on 19 Df
## Pseudo R-squared: 0.9648
## Number of iterations: 34 (BFGS) + 2 (Fisher scoring)
```