Clinical Trial of Patients with Respiratory Illness

Michael Nash

September 13, 2018

I present the following as an example of how to analyze categorical data with repeated measures in R. I use a didactic dataset originally found in chapter 15 of Stokes, Davis and Koch (1995) and also included in Fitzmaurice, Laird & Ware (2011). It can be found on the website for the latter text (https://content.sph.harvard.edu/fitzmaur/ala/). Here is an excerpt from this website explaining how the data were obtained:

The data are from a clinical trial of patients with respiratory illness, where 111 patients from two different clinics were randomized to receive either placebo or an active treatment. Patients were examined at baseline and at four visits during treatment. At each examination, respiratory status (categorized as 1 = good, 0 = poor) was determined.

The treatment outcome is the patient's respiratory status. I sometimes refer to good and poor respiratory status as GRS and PRS, respectively. The purpose of the trial is to assess the effect of the active treatment on respiratory status, although one might also be interested in how other factors are related respiratory status. Patients' gender, respiratory status and age at a pre-treatment baseline visit, and at which of two study centers they were enrolled are also known.

I will begin by reading the data into R from the text file supplied on the textbook website. I will perform some exploratory data analysis, followed by statistical analysis to determine the effect of the treatment and subject characteristics on the outcome of respiratory status.

ID	Center	Treatment	Gender	Age	Base	V1	V2	V3	V4
1	1	P	M	46	Poor	Poor	Poor	Poor	Poor
2	1	P	M	28	Poor	Poor	Poor	Poor	Poor
3	1	A	M	23	Good	Good	Good	Good	Good
4	1	P	M	44	Good	Good	Good	Good	Poor
5	1	P	F	13	Good	Good	Good	Good	Good
6	1	A	M	34	Poor	Poor	Poor	Poor	Poor

The first few rows of data in a 'wide' format, with outcome measurements taken at the five timepoints for each subject shown in a single row of the table. Although only the first six rows are shown, the table contains 111 rows, one for each subject. From left to right, the dataset contains the following variables:

Center - Each subject is enrolled at one of two study centers, numbered 1 and 2.
ID - Each subject has a unique ID, numbering from 1 to 111.
Treatment - Each subject receives one of two treatments: A for active or P for placebo.
Age - Subject's age in years at baseline visit.
Base - Subject's respiratory status at baseline visit.
V1, V2, V3, V4 - Subject's respiratory status at one of four visits.

Confusingly, subject IDs number 1 to 56 at the first study center and 1 to 55 at the other, such that pairs of subjects share an ID. I created a unique ID for each subject and removed the original, non-unique IDs.

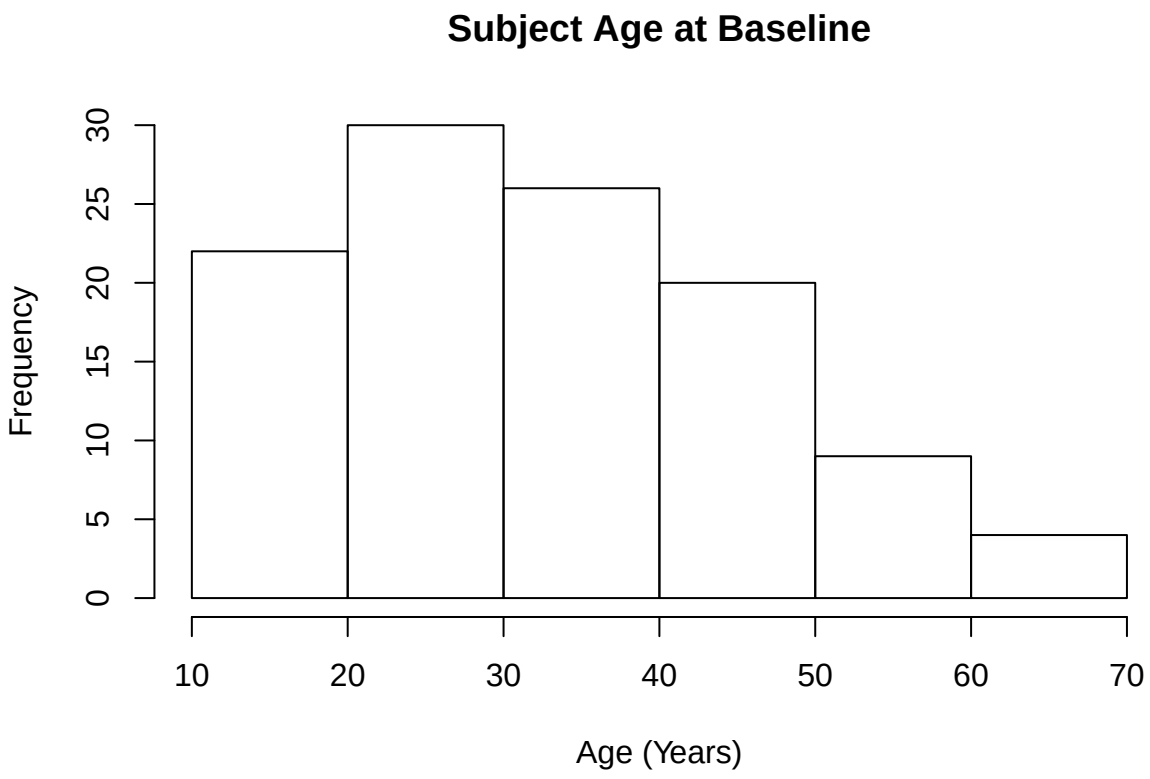
```
## [1] "Summary of 'Complete Cases'"
```

Mode	TRUE
logical	111

111 of 111 'complete cases' indicates that every subject has complete data.

Center	Treatment	Gender	Base	V1	V2	V3	V4
1:56	P:57	F:23	Poor:61	Poor:46	Poor:51	Poor:46	Poor:53
2:55	A:54	M:88	Good:50	Good:65	Good:60	Good:65	Good:58

The numbers of subjects at each of the two centers, and assigned to placebo and active treatments, are roughly equal. All subjects are male or female. Male subjects outnumber female subjects almost 4 to 1. Subjects with poor respiratory status outnumber those with good respiratory status at baseline, whereas the reverse is true at all subsequent visits. This could indicate that the treatment works, or just that the subjects tend to get better over time.




```
## [1] "Summary statistics for baseline age:"
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
11	23	31	33.27928	43	68

Subjects' ages range from 11 to 68 at baseline. I do not know how far apart post-treatment visits are from one another or from the baseline visit, so it is possible that subjects are actually older than this at later visits.

These data come from a randomized trial. The purpose of randomization in a clinical trial or any other experiment is to prevent confounding, which is when a third variable is related to both the outcome one wishes to understand and the exposure whose affect one is studying, distorting the effect of the exposure. For example, in a non-randomized trial, doctors might assign patients to therapies based on the seriousness

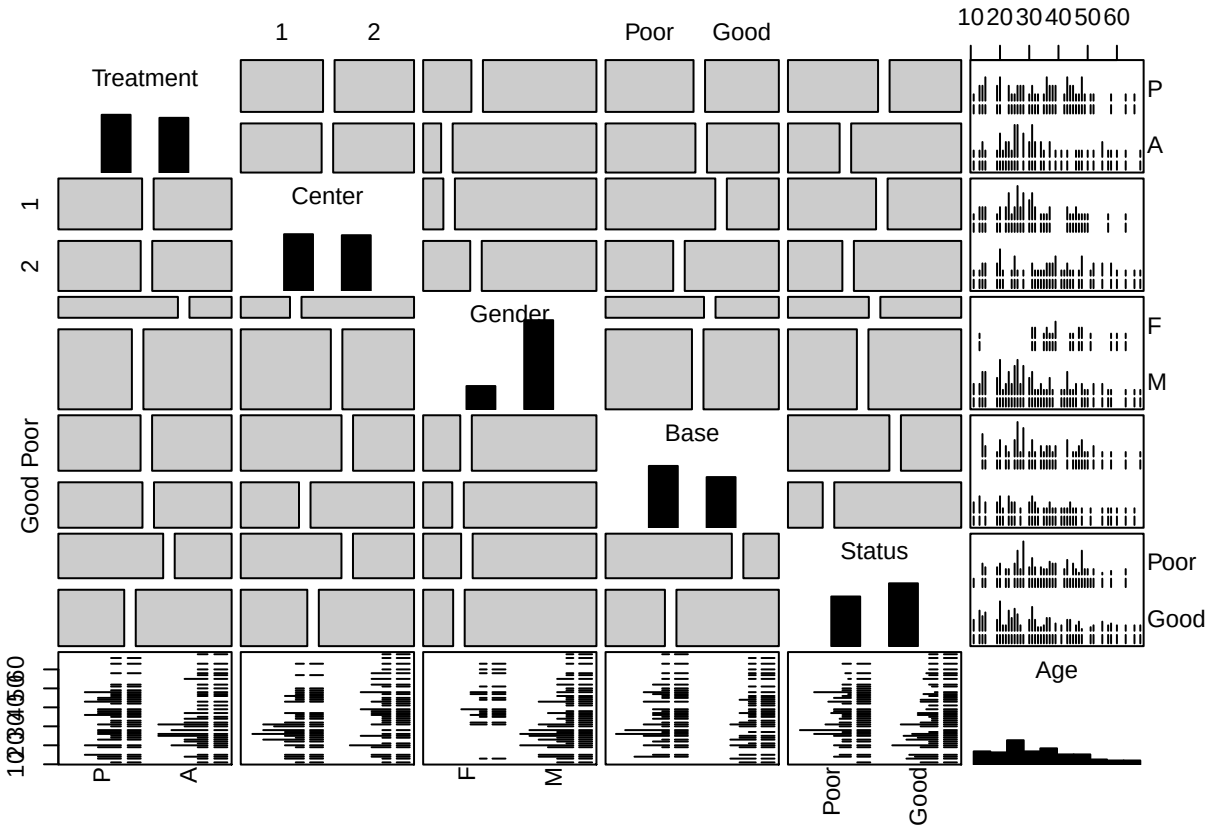
of their conditions, with more severely ill patients receiving a more aggressive treatment. One would then say that the treatment condition is confounded with the severity of illness. One might observe that patients receiving the more aggressive treatment tend to have a worse prognosis and erroneously assume that the difference in outcomes is caused by the difference in treatment received, when it is really caused by the difference in disease severity.

Randomization avoids this situation by making individuals with different characteristics equally likely to be assigned to a given exposure condition, so that there are no other systematic differences between subjects exposed to different conditions as part of the experiment. Crucially, this is true for all possible characteristics of the subjects, even those that are unknown. For this reason, unlike an observational study, one can infer a causal relationship between exposure and outcome from the results of a randomized experiment. The characteristics of subjects assigned to different treatments can still differ by chance, and it is worth checking whether this is the case for characteristics one can measure and suspect might be important in determining the outcome.

Below is a plot showing the pairwise relationships among the variables of treatment, study center, gender, baseline respiratory status ('base'), respiratory status at each post-treatment visit ('status'), and age. For each combination of categorical variables, a rectangle represents the overlap between two categories, with the size of the rectangle being proportional to the number of subjects contained in this overlap. A bar plot shows the number of individuals at each age grouped by each dichotomous variable. I will not be performing any hypothesis tests, because I already know that any differences between treatment groups occurred by chance, and they could still affect the results even if they are not statistically significant.

```
## Loading required package: grid
```

```
## Loading required package: lattice
```



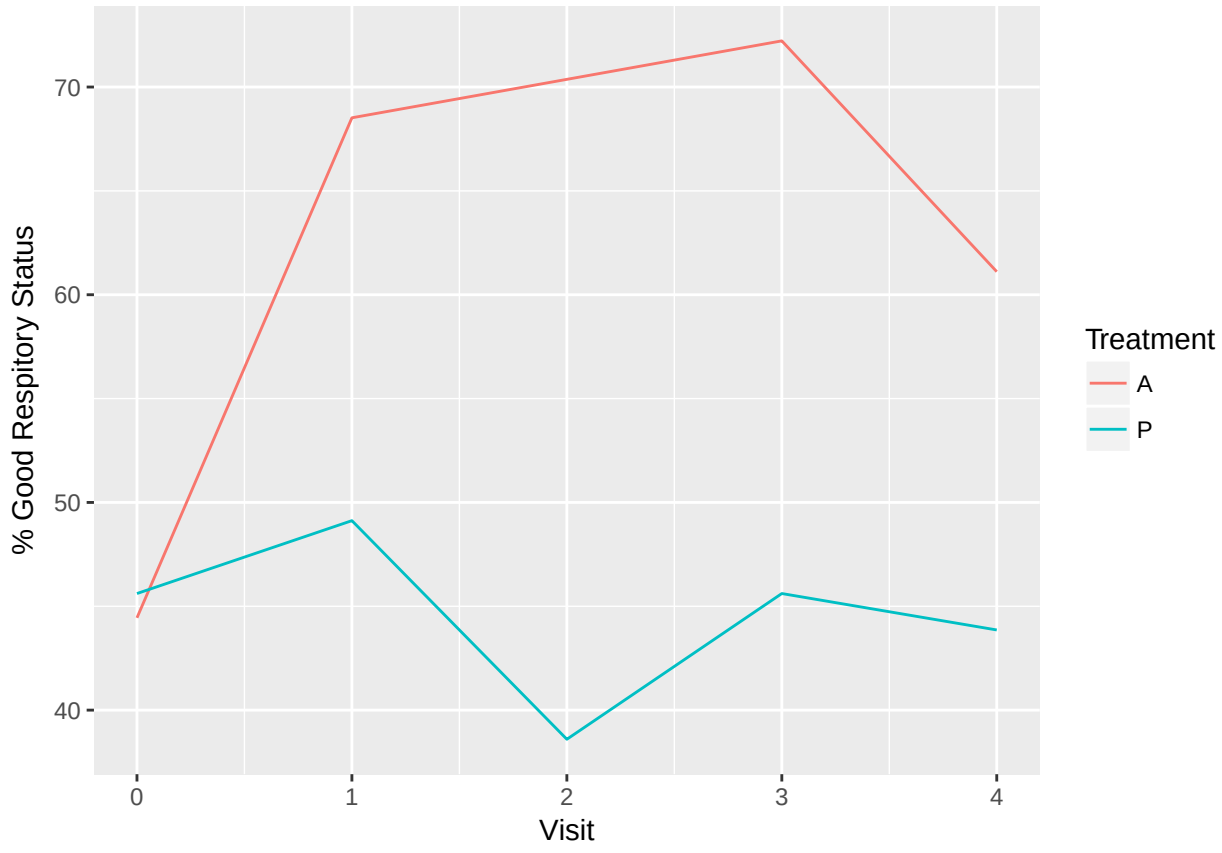
Groups assigned to active and placebo treatments have approximately equal proportions of good and poor

respiratory status at baseline and of subjects at the two study centers. The age distribution seems to be similar for active treatment and placebo groups, and for every other variable represented. There seem to be more female subjects assigned to placebo than active treatment, which makes gender a possible confounder.

At the intersection of 'baseline' and 'status', the rectangles in the top left and bottom right represent visits at which an individual's current status is the same as their baseline status, whereas the top right and bottom left rectangles represent visits at which an individual's current status is different from their baseline status. The first two rectangles are larger, indicating that individuals with good respiratory status at baseline are more likely to have good respiratory status at a follow-up visit than those with poor respiratory status at baseline, and vice versa, as one would expect.

Female subjects are somewhat more likely to have poor respiratory status than male subjects, both at baseline and at follow-up. Subjects at center 1 are more likely to have poor respiratory status at baseline and at follow-up than those at center 2. Individuals receiving the active treatment are more likely to have good respiratory status at follow-up visits, but not at baseline, than those in the placebo group. This suggests the treatment may be working, although I will return to this question.

It is also not known how respiratory status changes over time in each treatment group. I have shown that subjects with poor respiratory status outnumber those with good respiratory status at baseline, and that the reverse is true at post-treatment visits overall, but so far this is all that is known about the pattern of respiratory status over time. The plot below shows the mean respiratory status at baseline (denoted as visit zero) and at each of the four post-treatment visits among subjects assigned to active and placebo treatments (visits 1 through 4).



The percentage of subjects with good respiratory status starts around 45 for both treatment groups. For the active treatment group, the share of patients with good respiratory status goes above 60% starting with the first post-treatment visit and stays there. For the placebo group, the percentage with good respiratory status fluctuates in a range between about 38% and 48% but never goes any higher. This suggests that individuals

who receive the active treatment are more likely to have good respiratory status than those who do not.

However, this does not tell us whether differences in post-treatment respiratory status between treatment groups could have occurred due to chance or be explained by other factors. In order to answer these questions, I will fit a multivariable model of respiratory status. Because the outcome is binary, I will choose a logistic regression model, one of several models that can be used to model the probability of a binary outcome as the function of several predictor variables. A probit model would also be a reasonable choice.

The basic logistic regression model assumes that observations are independent, meaning that the outcome for each observation is unrelated to every other observation. The data set contains multiple observations for each individual. One would expect that two observations from the same individual would be more likely to have the same outcome than two observations from different individuals, violating the assumption of independence. Therefore, I will choose a model that accounts for different observations from the same individual being correlated with one another, which is known as autocorrelation.

There are two major types of models that can be used to model a binary outcome while accounting for autocorrelation. A marginal model predicts the mean response at any given combination of predictor levels, whereas a mixed model predicts the individual response. The coefficient estimates in a marginal model have a between-subject interpretation. This type of model can tell us, for example, the difference between the expected odds of good respiratory status between individuals assigned to different treatments with all other things being equal (i.e. the difference in the averages). The coefficient estimates in a mixed model have a within-subject interpretation. This type of model can tell us, for example, the expected difference in the odds of good respiratory status for the same individual if they are assigned to different treatments (i.e. the average of the differences).

An unfortunate feature of the mixed model is that the form of the dependence between different observations in the same subject over time must be specified. This is a problem for us, because I do not even know how far apart the visits occur in time, or whether the time intervals between them are the same. With a marginal model, one can use an unstructured correlation matrix, separately estimating the correlation among each pair of timepoints without making any assumptions about how this changes over time.

I will fit a marginal model with an unstructured correlation using the GEE (generalized estimating equations) method. The outcome will be the log odds of good respiratory status at a given post-treatment visit. This will allow us to measure the association of each covariate with the odds of good respiratory status when all others are held constant.

My procedure will be as follows: 0) Check for multicollinearity 1) Select the appropriate scale in which to include age based on the performance of polynomial terms in a main effects model. 2) Evaluate interaction effects in a model containing all main effects and all interactions with treatment. 3) Evaluate main effects in a model containing interactions selected in the previous step. 4) Fit a final model containing only main effects and interactions selected in previous steps. 5) Evaluate the fit of the final model. 6) Interpret the final model.

Multicollinearity occurs when covariates are strongly correlated with one another, such that the value of one covariate can be strongly predicted from other covariates. When this occurs, coefficient estimates become numerically unstable and their standard errors become much larger. A generalized variance inflation factor (GVIF) is used to quantify the extent to which each covariate is correlated with others, with GVIF of 10 traditionally used as a cutoff to identify covariates which are so strongly correlated with others as to seriously interfere with accurately estimating model coefficients.

```
## [1] "GVIF:"
```

```
## Treatment   Visit.f    Center    Gender       Age      Base
##  1.228449  1.099265  1.240562  1.542364  1.283392  1.202996
```

Fortunately, multicolinearlity does not seem to be a problem for this set of variables. The function used to generate GVIFs comes from the public Github repository 'RCode_Master_UPO'of user PedroJ.

Next, I need to select the appropriate scale in which to include subject age. I don't know the form of the relationship between subject age and the odds of good respiratory status. It could be linear, it could

be something else, or there might not be any relationship. I will evaluate polynomial terms of age in a main-effects only model (also containing treatment, visit number, center, gender, baseline status) up to the third order (i.e. linear, quadratic and cubic terms), because a third order polynomial can approximate a wide range of nonlinear relationships well. If the cubic term is significant (i.e. $p < .05$ for a test of the hypothesis that including this coefficient significantly adds to model fit), I will include all three terms. If the quadratic but not the cubic terms are significant, I will include only quadratic and linear terms. If quadratic and cubic terms are not significant, I will include only the linear term.

The ANOVA table below shows tests of the hypothesis that various polynomial terms for age improve model fit. Age1, 2 and 3 refer to linear, quadratic and cubic terms for age, respectively.

	Df	X2	P(>|Chi|)
Age1	1	2.3163066	0.1280237
Age2	1	9.7715478	0.0017723
Age3	1	0.3161189	0.5739494

The quadratic (chisq(1) = 9.77, p = .0018) but not the cubic term (chisq(1) = 0.32, p = 0.5739) significantly add to model fit. Therefore, I will model the quadratic relationship between age and odds of good respiratory status in the full model containing main effects and interactions.

Besides the possibility of confounding, another reason to consider other variables when understanding the relationship between the treatment condition and outcome is that this relationship may differ according to these other variables. When the effect of the treatment is different for subjects with different values of some other variable, one would say that interaction or effect modification is present. For example, if the treatment has a stronger effect on post-treatment respiratory status for individuals with PRS at baseline than those with GRS at baseline, one would say that there is a baseline status by treatment interaction, or that baseline status is an effect modifier.

I will now consider the interactions of treatment with visit, center, gender, age and baseline status in a model containing the main effects listed above with linear and quadratic terms for age. The ANOVA table below shows tests of the hypothesis that various interactions improve model fit. The age by treatment interactions are evaluated in a separate test because I wish to test the two terms (linear and quadratic) together.

	Df	X2	P(>|Chi|)
Treatment:Visit.f	3	3.151801	0.3688074
Treatment:Center	1	2.201967	0.1378347
Treatment:Gender	1	4.015291	0.0450894
Treatment:Base	1	1.453977	0.2278909

```
## [1] "Hypothesis: no age by treatment interaction:"
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.23, df = 2, P(> X2) = 0.89
```

The visit by treatment interaction does not significantly improve model fit (chisq(3) = 3.15, p = 0.3688). I conclude that the effect of the treatment is the same at each visit. In other words, the pattern of change in respiratory status over time (over the post-treatment visits) does not differ for the two treatments. The treatment by center interaction does not significantly improve model fit (chisq(1) = 2.20, p = 0.1378). I conclude that the treatment effect does not differ by center, meaning that the treatment is equally effective when administered at either center. The treatment by age interactions for the linear and quadratic terms, considered together, do not significantly improve model fit (chisq(2) = 2.20, p = 0.1378). I conclude that the

treatment effect does not depend on age, meaning that the treatment is equally effective at all ages. The treatment by baseline status interaction does not significantly improve model fit (chisq(1) = 1.45, p = 0.2279). I conclude that the treatment effect does not differ by baseline status, meaning that the treatment is equally effective for individuals with good and poor respiratory status before initiating treatment. The treatment by gender interaction significantly improves model fit (chisq(4.02), p = .0451). I conclude that the treatment is not equally effective for male and female patients. Based on these results, I will include the treatment by gender interaction and no others.

I will now consider the main effects of visit, center, age and baseline status in a model which also contains the main effects of treatment and gender and their interaction. The main effects of treatment and gender should be retained in the model regardless of whether they significantly add to model fit and are not interpretable because there exists an interaction between treatment and gender. The ANOVA table below shows tests of the hypothesis that various main effects improve model fit. The main effects of age are evaluated in a separate test because I wish to test the two terms (linear and quadratic) together.

	Df	X2	P(>|Chi|)
Visit.f	3	3.559233	0.3131607
Center	1	8.427808	0.0036953
Base	1	26.285694	0.0000003

```
## [1] "Hypothesis: age has no effect:"
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 12.4, df = 2, P(> X2) = 0.0021
```

The main effect of visit does not significantly improve model fit (chisq(3) = 3.56, p = 0.3132). I conclude that the odds of good respiratory status are the same at every post-treatment visit and do not change over time. The main effect of center significantly improves model fit (chisq(1) = 8.43, p = 0.0037). I conclude that the odds of good respiratory status are different for subjects at different centers. Together, the linear and quadratic main effects of age significantly improve model fit (chisq(2) = 12.4, p = 0.0021). I conclude that there is a quadratic relationship between age and the odds of good respiratory status, and that individuals at different ages have different odds of good respiratory status. The main effect of baseline status significantly improves model fit (chisq(1) = 26.29, p < .0001). I conclude that the odds of good respiratory status post-treatment are different for subjects who had good and poor respiratory status at baseline, respectively.

Based on these results, I retain the main effects of age, center and baseline status and remove the main effect of visit. The preliminary final model is shown below:

```
##
## Call:
## geeglm(formula = Status.n ~ Treatment + Center + Gender + Age1 +
##     Age2 + Base + Treatment:Gender, family = "binomial", data = datalong,
##     id = ID, corstr = "unstructured")
##
##  Coefficients:
##                     Estimate   Std.err   Wald Pr(>|W|)
## (Intercept)        -1.963566  0.473873 17.170 3.42e-05 ***
## TreatmentA          2.898459  0.805797 12.938 0.000322 ***
## Center2             0.570835  0.381640  2.237 0.134721
## GenderM             0.104160  0.564657  0.034 0.853648
## Age1               -0.037565  0.014618  6.603 0.010179 *
## Age2                0.002642  0.000776 11.592 0.000662 ***
```

```
## BaseGood            1.973026  0.363245 29.503 5.58e-08 ***
## TreatmentA:GenderM -1.861301  0.882723  4.446 0.034980 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)    1.039  0.4333
##
## Correlation: Structure = unstructured  Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2   0.3911  0.2032
## alpha.1:3   0.1670  0.1106
## alpha.1:4   0.2853  0.1509
## alpha.2:3   0.2790  0.1499
## alpha.2:4   0.3189  0.1649
## alpha.3:4   0.3025  0.1662
## Number of clusters:   111   Maximum cluster size: 4
```

Before interpreting this model, I must evaluate the model fit. The assumptions of the generalized logistic regression model are relatively unrestrictive compared to some other types of regression models. There is no assumption of independence; on the contrary, I assume that the observations are correlated and explicitly account for this correlation. The choice of an unstructured covariance matrix means I make no assumption about the specific form of the dependence between observations in the same subject over time. There is no assumption that errors follow a particular distribution as in the linear model.

There are two fit-related issues that should be explored:

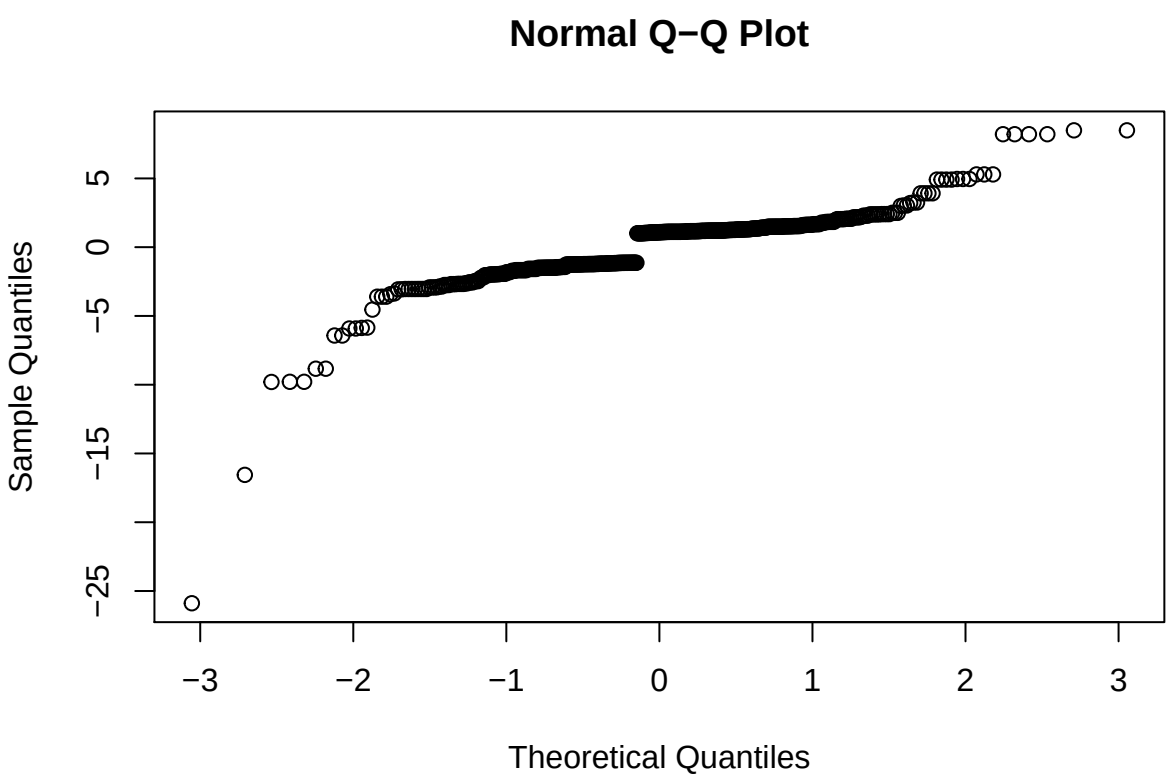
1) If the model is correctly specified, it should fit equally well at all levels of predicted odds. I can test this assumption with the Hosmer-Lemeshow test.

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  Status.n, mod.4.gee$fitted.values
## X-squared = 11, df = 8, p-value = 0.2
```

The result of the Hosmer-Lemeshow test does not provide evidence for lack of fit.

2) Model fit should not depend to a great extent on any one observation or covariate pattern.

Although the model does not rely on the assumption of normally distributed errors, a normal QQ plot of residuals is a good way to identify observations which are poorly fit compared to all others.



There are two observations with extreme negative residuals. Complete data for the two individuals who produced these observations are shown below

	Observation #	Residual	Subject ID
216	216	-25.89	54
330	330	-16.56	83

	ID	Center	Treatment	Gender	Age	Base	V1	V2	V3	V4
54	54	1	A	M	11	Good	Good	Good	Good	Poor
83	83	2	A	M	19	Good	Good	Poor	Good	Good

These observations come from visit 2 in subject 83 and visit 4 in subject 54. Both these individuals had good respiratory status at baseline and at every other visit, and they received the active treatment, so I would predict that they have high odds of good respiratory status at those visits. The fact that I have a few poorly fit observations is not a problem in and of itself, but it would be worthwhile to see what happens to the coefficient estimates when the model is refit without them. If the coefficient estimates change a great deal, this would suggest that these estimates reflect sampling error more than trends in the population from which I am sampling. In that case, one might be unable to replicate these results with an independent sample.

```
## [1] "% change in coefficients:"
```

```
##        (Intercept)         TreatmentA            Center2
##               3.06               2.70              -1.66
##            GenderM               Age1               Age2
```

```
##             -35.78              12.25              10.90
##           BaseGood TreatmentA:GenderM
##               5.57               0.34
```

The proportional change in the coefficient estimates after removing these observations is relatively small. The one exception to this is the main effect of gender which, as I will show later, is not statistically significant anyway. There is no cause for alarm here.

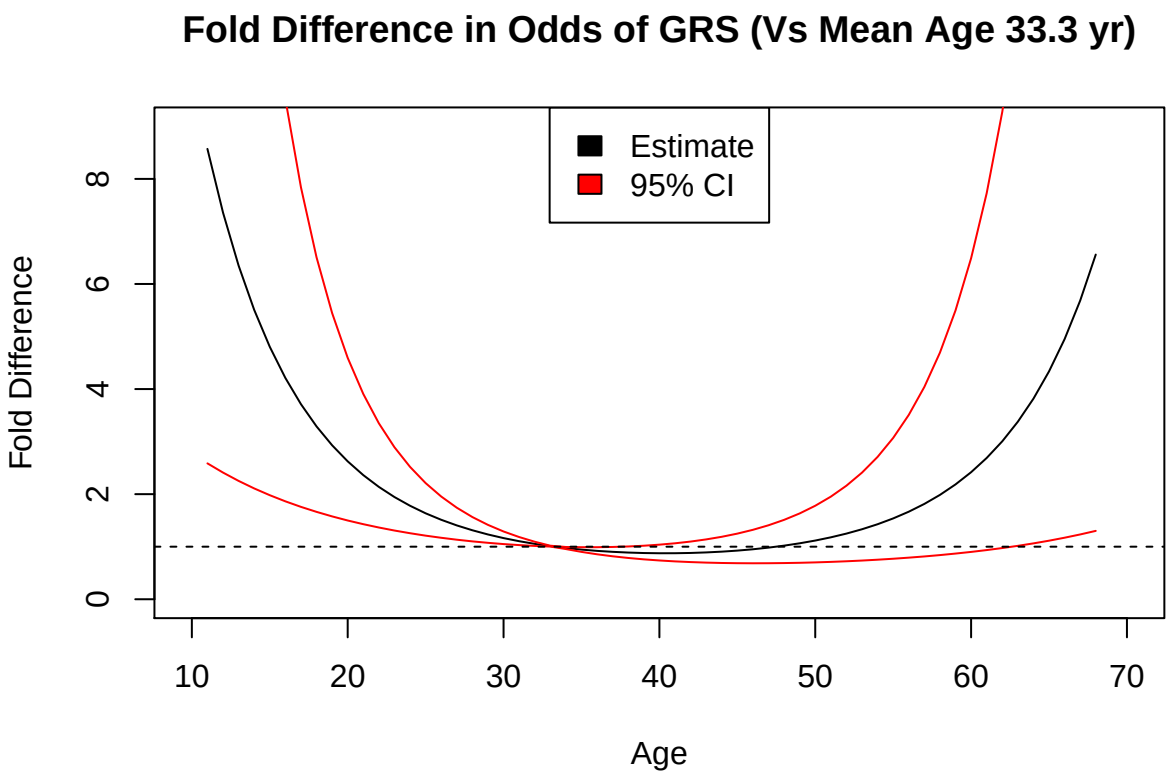
Now I will begin to interpret the model. The table below gives point and range estimates for the fold difference in odds associated with various conditions.

	Estimate	95% Lower Bound	95% Upper Bound
Active vs Placebo - Male	2.821	1.2783	6.226
Active vs Placebo - Female	18.146	3.7401	88.040
Center 2 (vs 1)	1.770	0.8376	3.739
Male vs Female - Placebo	1.110	0.3669	3.356
GRS at Baseline	7.192	3.5292	14.658

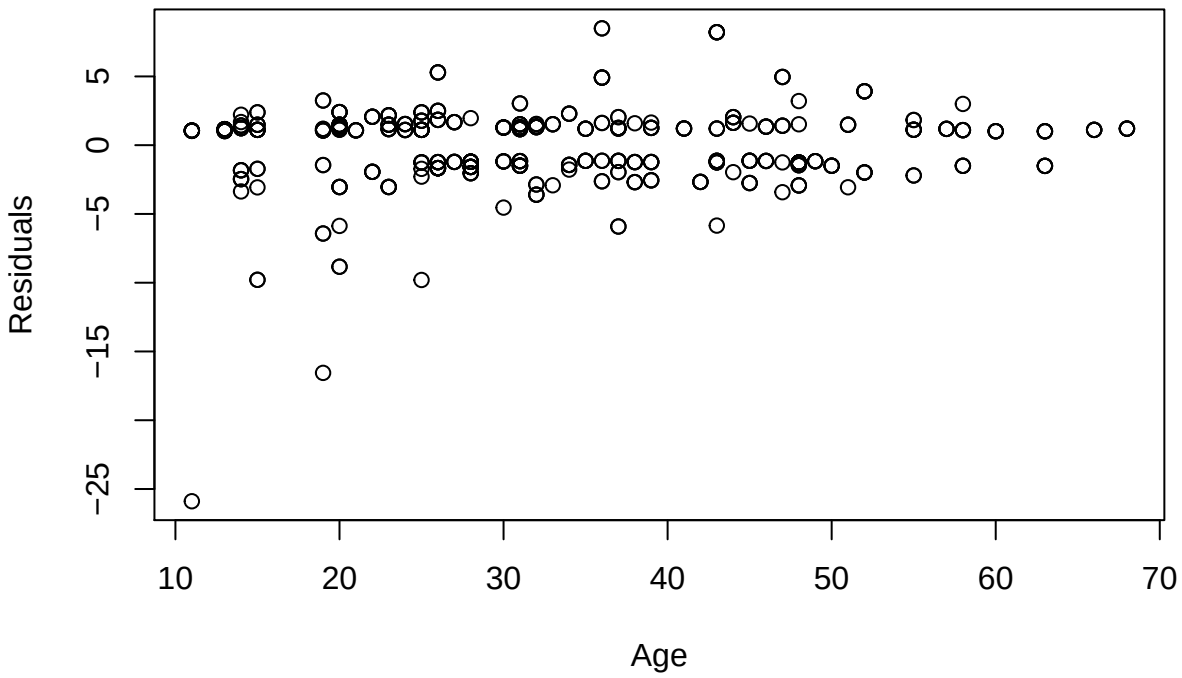
The expected odds of GRS in female subjects receiving the active treatment is 18.15 times the expected odds in female subjects receiving placebo (95% CI from 3.74 to 88.04). One could also represent this as a percent difference and say that the odds are % 1715 higher in female subjects receiving the active treatment, but it can be hard to make sense of percentages that large. I use fold difference rather than percent difference for the other conditions to maintain consistency. The expected odds of GRS in male subjects receiving the active treatment is 2.82 times the expected odds in male subjects receiving placebo (95% CI from 1.278 to 6.23). The treatment appears to be effective for both male and female subjects, but more so for female subjects.

The expected odds of GRS at a post-treatment visit in subjects with GRS at baseline is 7.19 times that of subjects with PRS at baseline (95% CI from 3.529 to 14.66). The expected odds of GRS for subjects at center 2 are 1.77 times those at center 1 (95% CI from 0.838 to 3.74). The expected odds of GRS for male subjects receiving placebo is 1.11 times the expected odds in female subjects receiving placebo (95% CI from 0.367 to 3.36). The confidence interval for these last two include the null value of 1, meaning that I can not say for sure whether the subjects at center 2 actually have higher or lower odds than those at center 1 or whether males receiving placebo actually have higher or lower odds than females receiving placebo.

The nonlinear relationship between age and odds of GRS is depicted in the plot below. The black line indicates the difference in estimated odds of GRS for individuals at a given age as a multiple of the estimated odds for an individual at the mean age in the sample (33.3 years old). The dotted line represents no difference. Naturally, the difference in the estimates at age 33.3 is none at all. The red lines represent 95% upper and lower confidence limits for the difference in the estimated odds.



From about ages 11 to 30, the odds of GRS are significantly higher than for subjects at the mean age of 33.3. Estimated odds of GRS reach their nadir around age 40. I'm skeptical of a model that says that eleven-year-olds have eight times the odds of GRS than 33-year-olds, so I will examine a plot of residuals by age to make sure that the model fits well at all ages.



The model seems to fit well across the range of ages in the sample and not systematically misclassify observations in one direction or another at any age.

In conclusion, the active treatment is effective compared to placebo, but has a greater effect in female than male patients. It appears to be equally effective at all ages, study centers, and for patients with good or poor respiratory status before treatment begins. Patients with GRS before treatment are more likely to have GRS in the future than those who start with PRS, all other things being equal. Patients at different ages have different odds of GRS, with patients around age 40 having the lowest odds of GRS and younger and older patients having greater odds of GRS. The odds of GRS do not appear to change over time for patients receiving either treatment regimen.

References

Stokes, M.E., Davis, C.S. and Koch, G.G. (1995). Categorical Data Analysis using the SAS System. Cary, NC: SAS Institute, Inc.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). Applied Longitudinal Analysis. Hoboken: Wiley.