Linear Regression Report

The problem: Here, we consider another clinically relevant problem in our dataset: predicting how many days after the rehabilitation program someone remains sober. This would represent an extra level of granularity from our classification predictions, in that we could hopefully identify people who are at risk of relapsing within 30 days, for instance, and keep them in the clinic for a few extra weeks.

Our outcome measure, `obstime`: As our continuous outcome we used the variable `obstime`, which encodes how many days after the rehabilitation program we observed sobriety in each patient. In our models, we encounted many challenges working with this covariate, since

1) patients are often lost to follow-up (meaning they don't answer our calls to record their latest sobriety data), making it difficult to confirm whether or not they've relapsed or not.
2) `obstime` by its nature is an "observed" time, meaning it's not exactly the number days of sober, but the number of days when we confirmed that the patients were sober. This will be even more of a challenge later on, when we deal with interpretation.

Here, we describe how we've dealt with the challenges of modeling `obstime`.

Modeling Approach: As a general modeling approach, since we were working with just a few subjects, we wanted to start with a simple model using the most salient features, look at our resulting residuals to see if we missed any big trend, and then add more features as needed.

As our evaluation metric we used RMSE, which has particular meaning in this context, since it represents how many days of sobriety on average our prediction would deviate from the true number of days of sobriety. Currently, there are no standard measures used to predict time to relapse, so we used the RMSE for a baseline model that predicts the mean of `obstime`. This baseline model has RMSE = 684.62, which would translate to making predictions that could be off by almost 2 years! To be clinically relevant, we wanted to shoot for a model that could make predictions with RMSE = 30 days, since that could help clinicians realistically assess which of their inpatients are at risk of relapsing within a few weeks of leaving the program and potentially stage interventions as necessary.

When comparing models to each other, we used k-fold cross-validation with $k = 5$, which we used because we were most concerned about overfitting, so wanted to choose a k that allowed us to to leave more data out and see if we were overfitting. When testing our final model, we used LOOCV because [[when testing, we cared about getting the most accurate possible estimate of our generalization error]]

Feature Selection: We had reason to believe a-priori that the brain region called the Nucleus Accumbens (NAcc; essentially, a brain region that is marker of dopamine response and anticipating rewards) would be the most pertinent of the brain regions for predicting `obstime`. There's two main approaches to extract information from a brain region: 1) by extracting the raw time series for each type of trial (for example, visualizing how activity changes at each time interval for all "drugs" trials; see "Time Series" plot in EDA below); and 2) by modeling how much activation in the whole trial is above baseline for each type of trial and extracting beta coefficients (for example, coefficient representing how much "drugs" trials are above baseline; see "Density Plots of Betas" plot in EDA below)

Although you could theoretically include both approaches, it is considered best practice to stick with one approach in the neuroscience field, so we needed to figure out which approach contained the most salient information for modeling `obstime`. To address that problem (and the problem of figuring out our most predictive features generally), we used the Lasso.

From our Lasso results, we found that the most salient features for predicting `obstime` were: 1) `relIn6Mos` (an indicator variable corresponding to whether someone relapsed in 6 months after the program), and 2) `naccR_drugs_beta` (a continuous variable corresponding to the betas extracted from the right side of the NAcc, called the RNAcc). These results were very interesting because we included all of our covariates (including different brain regions using both approaches mentioned above), and the model selected the betas from the RNAcc, which confirmed our a priori hypothesis that the NAcc was the most predictive brain region!

We used those two variables as our starting point for our original model of `obstime`.

Our original model of `obstime`:

Here, our best model is a linear model with an interaction of the two features we found above. So Let Y = `obstime`, X_1 = `naccR_drugs_beta`, and X_2 = `relIn6Mos`.

We found that our average RMSE with k-fold cross-validation where $k = 37$ was 73.54. Although that is better than our baseline model, it's still not quite clinically useful to say someone is going to relapse within a 75-day window.

To diagnose this issue, we carefully looked at our residuals and noticed that the main deviations came from non-relapsers, which makes sense because of the way `obstime` is encoded. [[EXPLANATION – assume there is a real linear relationship between X1 and Y, we would naturally expect that non-relapsers would have more error since `obstime` does not actually represent number of days sober for them]]

After careful consideration, we decided to only use regression in the context of our patient population who relapsed because we thought that would be the most clinically useful. We believe that a nice pipeline would be to assess the likelihood of someone relapsing or not within 6 months (see our classification section), and then use a regression built for potential relapsers to determine how many days they'll remain sober to determine whether they need to stay longer.

Our final model of `obstime` (note: constrained to only relapsers):

Here, our best model was a simple linear model where Y = `obstime` and X_1 = `naccR_drugs_beta`.

We found that our average RMSE with k-fold cross-validation where $k = 15$ was 34.83. Note that, here, our baseline model has an RMSE of 137.5.

These results were very exciting! Not only did we improve our RMSE from the baseline, but these results could also be used clinically to make a prediction that someone is going to relapse which is off by around 30 days (i.e., month). However, we have to be cautious to note that these results are drawn from only 15 subjects, so we would need to replicate in order to really see how clinically relevant these results are.

Conclusions, Learnings, and Thoughts on Bias-Variance:

We were happy to learn that we could predict days of sobriety within an error that we feel is clinically relevant. I believe these results can help clinicians gain an extra degree of clarity in the important question of how long patients will stay sober.

In the process of learning how to deal with `obstime`, we stumbled upon methods known as Survival Analyses, which are essentially. These models are built to deal with variables like `obstime`, where some observations may be lost to follow-up. I would be interested in the future to apply those analyses to our data.

Because we had so few subjects to work with after we filtered our dataset to only the relapsers, we had to think a lot about how testing our model repeatedly (even with LOOCV) might introduce our own Bias into the modeling process. For that reason, I think our final RMSE may be an underestimate of the true error. At the same time, working with such a small dataset forced us to carefully consider every decision we made, and get the most of the data we had.

Load Data

```
df_trials_full <- read_csv("relapse_trials_full.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   subjid = col_character(),
##   group_idx = col_integer(),
##   trial_cond = col_character(),
##   trial_onsetTR = col_integer()
## )
```

```
## See spec(...) for full column specifications.
```

```
df_subjects <- read_csv("relapse_subjects.csv") %>% filter(relIn6Mos %in% c(0, 1))
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   subjid = col_character(),
##   depression_diag = col_integer(),
##   anxiety_diag = col_integer(),
##   ptsd_diag = col_integer(),
##   age = col_integer(),
##   clinical_diag = col_integer(),
##   bis = col_integer()
## )
## See spec(...) for full column specifications.
```

```
df_subjid <- df_subjects %>%
  select(subjid, relapse = relIn6Mos)
```

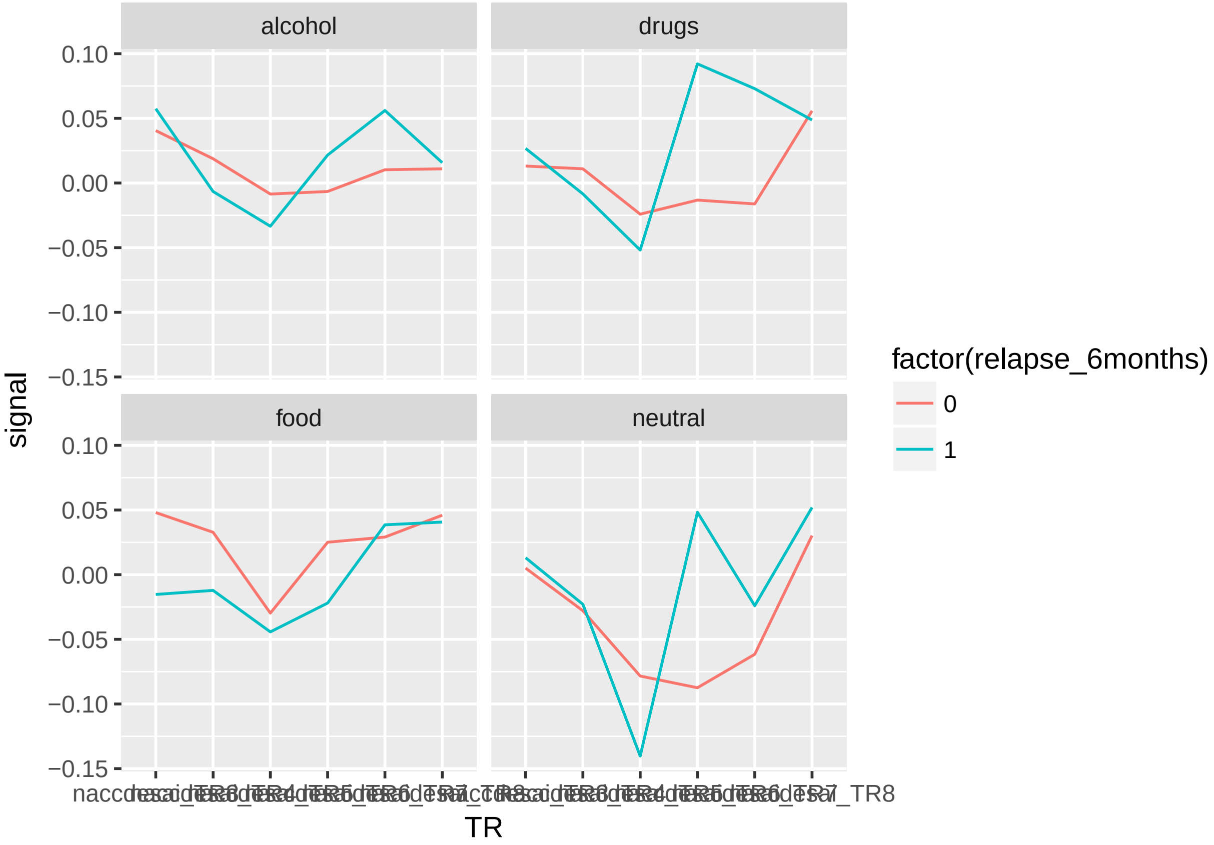
Lasso

```
set.seed(1)
is_zero <- function(x) all(mean(x^2) == 0)

vars_no_nas <-
  df_subjects %>%
  summarise_all(function(x) {mean(is.na(x))}) %>%
  select_if(is_zero) %>%
  select(-contains("relapse"),
         -starts_with("mpfc"),
         -starts_with("vta"),
         -starts_with("acing"),
         -starts_with("ains"),
         -ends_with("mean"),
         -censored
        # -ends_with("beta")
         ) %>%
  colnames()

by_subj_no_nas <-
  df_subjects %>%
  select(vars_no_nas)

#this does cross-validates lasso. alpha = 1 specifies that we want lasso. type.measure = class specifie
lasso_ind <- cv.glmnet(x = by_subj_no_nas %>% select(-subjid, -obstime) %>%  data.matrix,
                y = by_subj_no_nas$obstime,
                alpha = 1)

#this gives us the coefficients for a choice of lambda (lambda.min is the lambda that gives smallest er
lasso_coefs <- coef(lasso_ind, s = lasso_ind$lambda.min)

#this gives cv error for best performing lambda
min(lasso_ind$cvm)

lasso_coefs
```

Exploratory Data Analysis

Time Series (Relapse x Condition)

```r
df_trials_full %>%
  select(subjid, group_idx, relapse_6months, trial_cond, contains("naccdesai")) %>%
  filter(group_idx == 1) %>%
  group_by(relapse_6months, trial_cond) %>%
  summarise_at(vars(contains("TR")), mean, na.rm = TRUE) %>%
  ungroup() %>%
  slice(1:8) %>%
  gather(key = "TR", value = "signal", naccdesai_TR3:naccdesai_TR8) %>%
  ggplot(aes(x = TR, y = signal)) +
  geom_line(aes(group = relapse_6months, color = factor(relapse_6months))) +
  facet_wrap(~trial_cond)
```

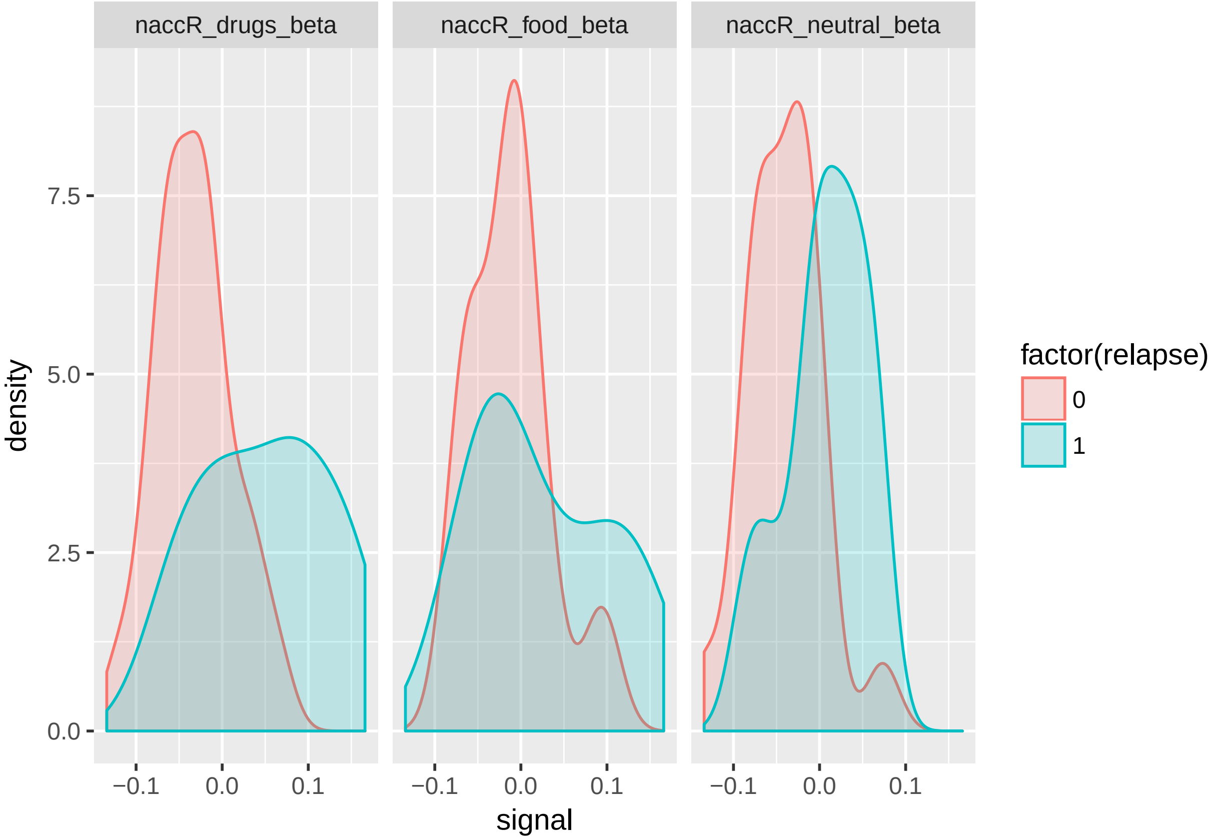


Some thoughts:

- Drugs is significantly higher for future relapsers at TR6
- Neutral is significantly higher for future relapsers at TR6

Density Plot of Betas (Relapse x Condition)

```
df_subjects %>%
  select(subjid, relapse = relIn6Mos, starts_with("nacc")) %>%
  select(relapse, ends_with("beta")) %>%
  gather(key = "condition", value = "signal", naccR_drugs_beta, naccR_food_beta, naccR_neutral_beta) %>%
  ggplot(aes(x = signal, y = ..density.., color = factor(relapse), fill = factor(relapse))) +
  geom_density(alpha = 0.2) +
  facet_wrap(~ condition)
```



This is a plot of the extracted coefficient for nacc_betas, which shows that relapsers should have a higher mean beta than non-relapsers…

- Looks like nacc_drugs_beta captures some differentiating factor between relapsers and non-relapsers! Perhaps we should try a classification using the nacc_betas!
- Do we need any kind of transformation here?
- We should find out what regression these coefficients are extracted from to help interpret…

Original model of `obstime`

```
set.seed(1)

train <- df_subjects %>% sample_frac(0.8)
test <- setdiff(df_subjects, train)
```

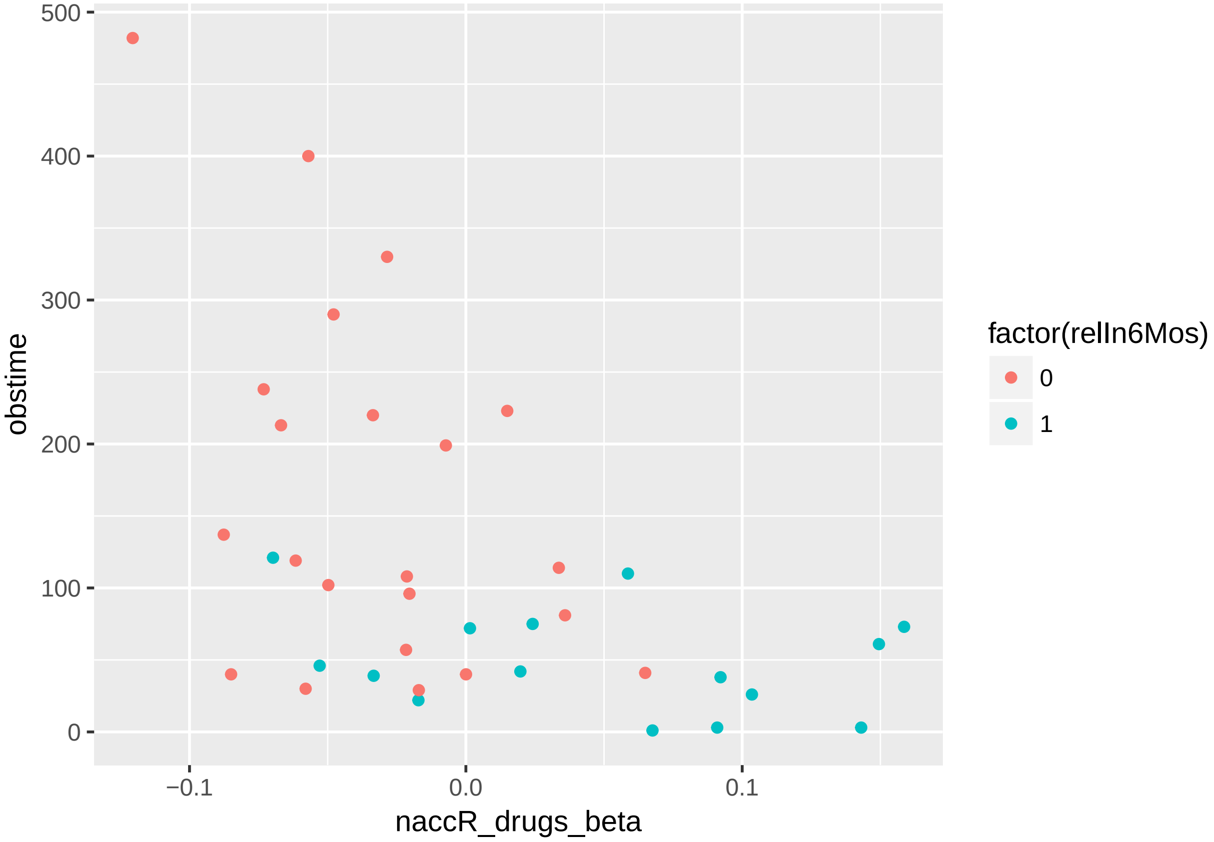
Baseline model

```
mean_obstime <- mean(df_subjects$obstime)


sqrt(sum((df_subjects$obstime - mean_obstime) ^ 2))
```

```
## [1] 684.6242
```

Visualize our data

```
df_subjects %>%
  ggplot(aes(x = naccR_drugs_beta, y = obstime)) +
  geom_point(aes(color = factor(relIn6Mos)))
```



Try different models!

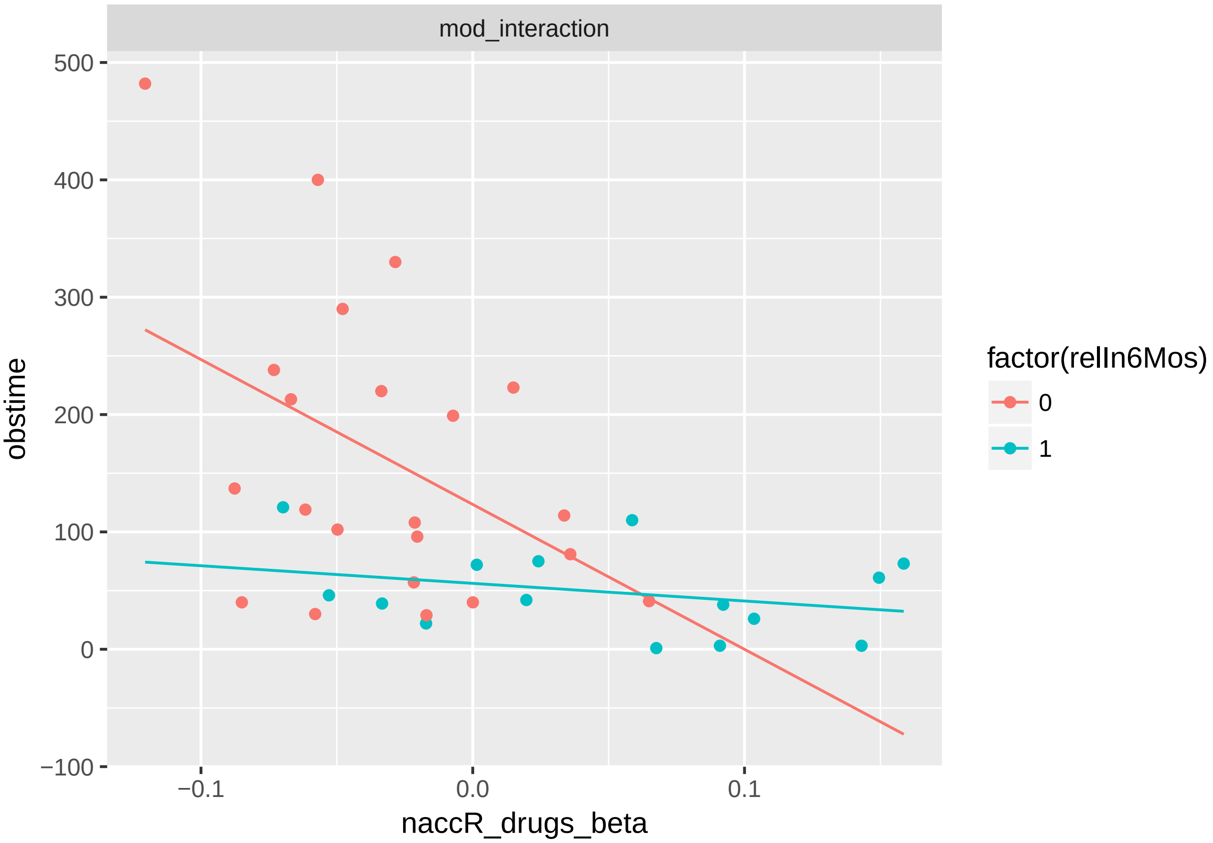
```
mod_interaction <- lm(obstime ~ relIn6Mos * naccR_drugs_beta, data = df_subjects)
mod_linear <- lm(obstime ~ relIn6Mos + naccR_drugs_beta, data = df_subjects)
```

Visualize our models fit on the data

This code can generate predictions for any number of models! Here, we only consider our best model `mod_interaction`

```
grid <-
  df_subjects %>%
  data_grid(relIn6Mos, naccR_drugs_beta) %>%
  gather_predictions(mod_interaction)
```

```
df_subjects %>%
  ggplot(aes(x = naccR_drugs_beta, y = obstime, color = factor(relIn6Mos))) +
  geom_point() +
  geom_line(data = grid, aes(y = pred)) +
  facet_wrap(~ model)
```



Evaluate our best model `mod_interaction`

Here, we use k-fold cross-validation with k = 37 and RMSE as our evaluation metric

```
df_subjects %>%
  crossv_kfold(k = 37) %>%
  mutate(
    mod = map(train, ~lm(obstime ~ relIn6Mos * naccR_drugs_beta, data = .)),
    rmse = map2_dbl(mod, test, rmse)
  ) %>%
  summarise(
    avg_rmse = mean(rmse)
  )
```

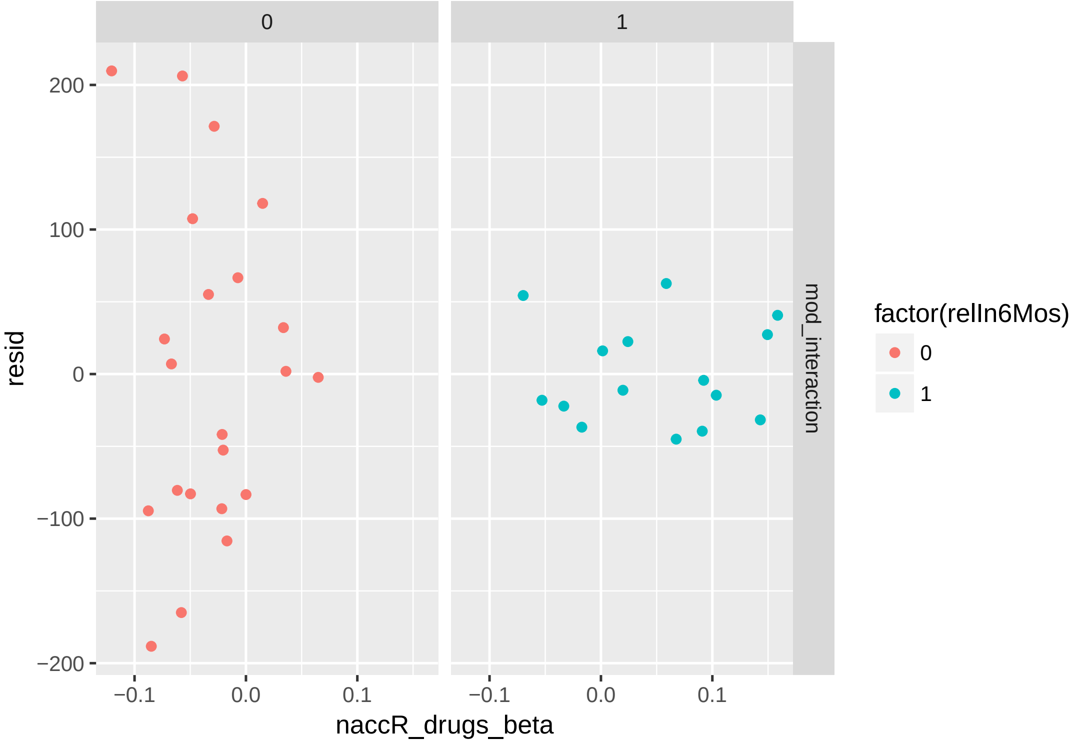
```
## # A tibble: 1 x 1
##   avg_rmse
##      <dbl>
## 1 73.54308
```

Visualize our residuals for our best model `mod_interaction`

Here, we faceted by whether patients relapsed or not, so that we can make a decision about

1) whether to use more features to capture any big effect
2) whether to only use relapsers in our final model!

```
df_subjects %>%
  gather_residuals(mod_interaction) %>%
  ggplot(aes(x = naccR_drugs_beta, y = resid, color = factor(relIn6Mos))) +
  geom_point() +
  facet_grid(model ~ factor(relIn6Mos))
```



Our final model of `obstime` (note: filtered to only relapsers)

```
df_subjects_relapsers <- df_subjects %>% filter(relIn6Mos == 1)
```

Baseline model

```
mean_obstime <- mean(df_subjects_relapsers$obstime)


sqrt(sum((df_subjects_relapsers$obstime - mean_obstime) ^ 2))
```

```
## [1] 137.486
```

Try different models on the filtered dataset!

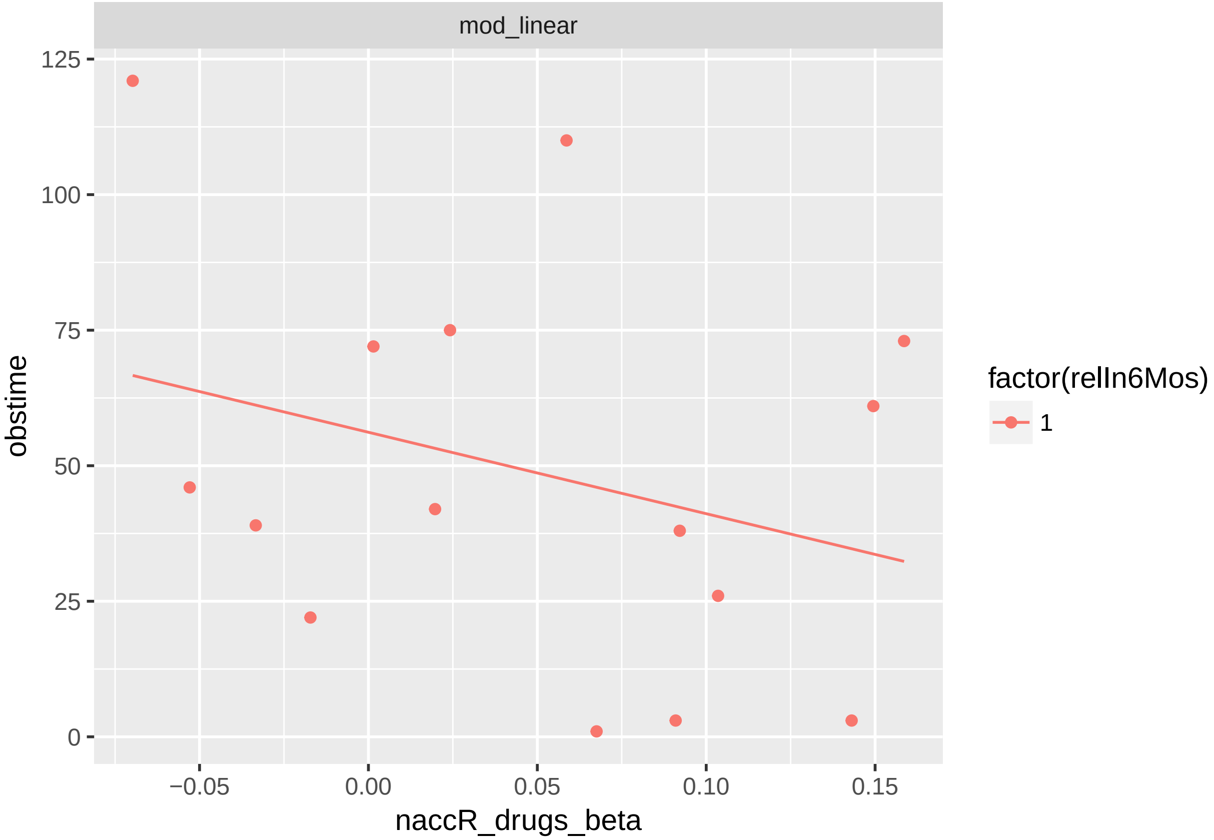
```
mod_linear <- lm(obstime ~ naccR_drugs_beta, data = df_subjects_relapsers)
```

Visualize our models fit on the filtered dataset

This code can generate predictions for any number of models! Here, we only consider our best model `mod_linear`

```
grid <-
  df_subjects_relapsers %>%
  data_grid(relIn6Mos, naccR_drugs_beta) %>%
  gather_predictions(mod_linear)
```

```
df_subjects_relapsers %>%
  ggplot(aes(x = naccR_drugs_beta, y = obstime, color = factor(relIn6Mos))) +
  geom_point() +
  geom_line(data = grid, aes(y = pred)) +
  facet_wrap(~ model)
```



Evaluate our best model `mod_linear`

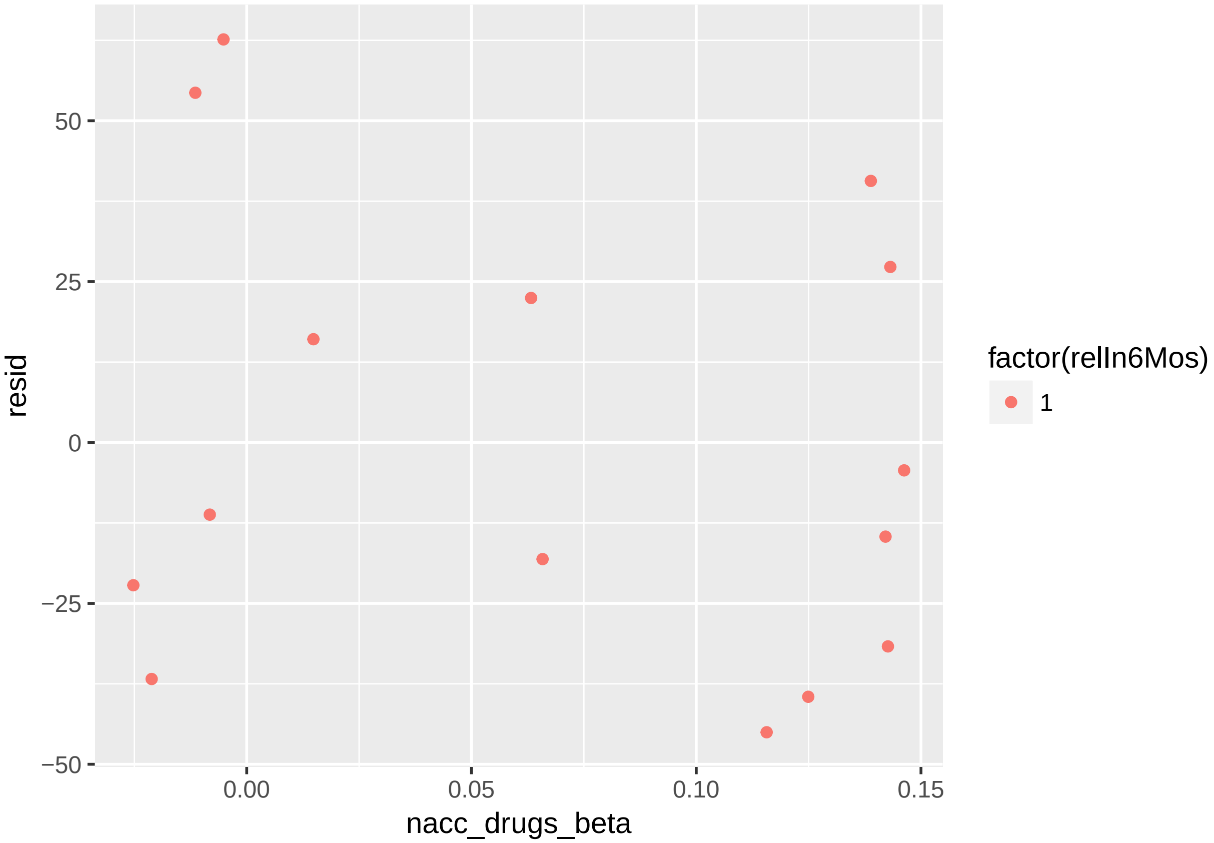
Here, we use k-fold cross-validation with k = 15 and RMSE as our evaluation metric

```
df_subjects_relapsers %>%
  crossv_kfold(k = 15) %>%
  mutate(
    mod = map(train, ~lm(obstime ~ naccR_drugs_beta, data = .)),
    rmse = map2_dbl(mod, test, rmse)
  ) %>%
  summarise(
    avg_rmse = mean(rmse)
  )
```

```
## # A tibble: 1 x 1
##   avg_rmse
##      <dbl>
## 1 34.82884
```

Visualize our residuals for our best model `mod_linear`

```
df_subjects_relapsers %>%
  gather_residuals(mod_interaction) %>%
  ggplot(aes(x = nacc_drugs_beta, y = resid, color = factor(relIn6Mos))) +
  geom_point()
```


resid
50
25
0
−25
−50
0.00
0.05
0.10
0.15
nacc_drugs_beta
factor(relIn6Mos)
● 1