CH 11: Assumptions - Part I

Assumptions for Regression Models

The assumptions described in *Regression and Other Stories,* are more broad than many textbooks. In order of importance,

1. **Validity:** *data should map to the research question: outcome measure reflects phenomenon of interest, model includes all relevant predictors. May require iteration between the data and the research question that can be answered*

2. **Representativeness:** *the typical goal of a regression model is to make inferences about a population from a sample, thus this is an implicit assumption of the model. Formally, the assumption is that conditional on the predictors X, the distribution of the outcome (y) is representative – post stratification.*

3. **Additivity and linearity:** *the functional form of the relationship between X and y must be accurately captured. Interactions, basis functions, transformations, and even other models (Gaussian Process, see 534)*

4. **Independence of Errors:** *The errors of the model are independent, is violated when to sampling units are "similar:" time series, spatial, repeated measures. Can result in uncertainty measurements that are too small.*

5. **Equal Variance of Errors:** *unequal variance - heteroscedasticity. most problematic when making probabilistic predictions. Has minimal impact on regression line. Weighted least squares, or including this information in a hierarchical model can mitigate this problem.*

6. **Normality of Errors:** *the distribution of the errors has minimal importance with fitting regression line –think about least squares. Again, it is important with probabilistic prediction. They (ROS) don't recommend looking at QQ-plots but this is not a conventional view*

What if the assumptions are violated??

- *Add more model complexity.*

- *change the data or model*

- *change or restrict the research questions*

Plots of fitted model

For simple models with one continuous predictor and/or one categorical predictor, we have see how to fit the model with `geom_smooth`.

With additional covariates in the model this becomes more challenging. Consider the candy dataset and a model

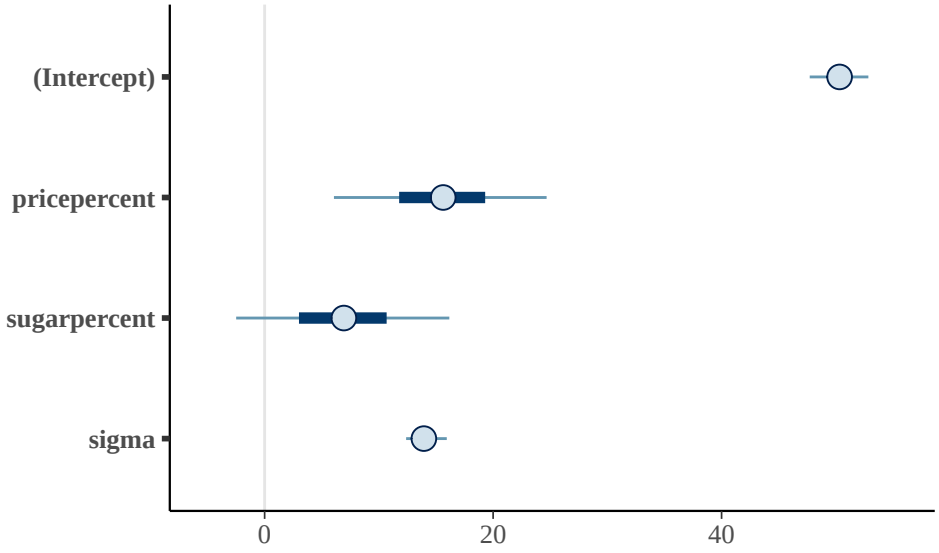
```
candy <- read_csv("https://math.montana.edu/ahoegh/teaching/stat446/candy-data.csv") %>%
  mutate(pricepercent = pricepercent - mean(pricepercent),
         sugarpercent = sugarpercent - mean(sugarpercent))
```

```
## Parsed with column specification:
## cols(
##   competitorname = col_character(),
##   chocolate = col_double(),
##   fruity = col_double(),
##   caramel = col_double(),
##   peanutyalmondy = col_double(),
##   nougat = col_double(),
##   crispedricewafer = col_double(),
##   hard = col_double(),
##   bar = col_double(),
##   pluribus = col_double(),
##   sugarpercent = col_double(),
##   pricepercent = col_double(),
##   winpercent = col_double()
## )
```

```
candy_model <- stan_glm(winpercent ~ pricepercent + sugarpercent, data = candy, refresh = 0)
print(candy_model)
```

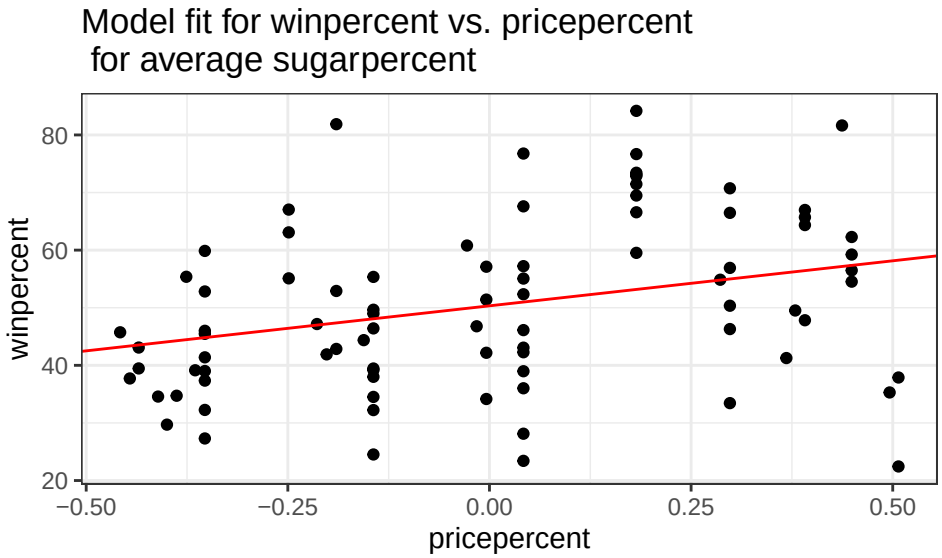
```
## stan_glm
##  family:       gaussian [identity]
##  formula:      winpercent ~ pricepercent + sugarpercent
##  observations: 85
##  predictors:   3
## ------
##              Median MAD_SD
## (Intercept)  50.3    1.5
## pricepercent 15.6    5.6
## sugarpercent  6.9    5.7
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 13.9    1.1
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
plot(candy_model)
```



- One option is to plot the response against each predictor holding the other continuous predictors constant and setting levels of categorical predictors. *Note this is different than just using* `geom_smooth` *and ignoring the other predictors.*

```
candy %>%
  ggplot(aes(y = winpercent, x = pricepercent)) +
  geom_point() +
  geom_abline(intercept = candy_model$coefficients['(Intercept)'],
              slope = candy_model$coefficients['pricepercent'],
              color = 'red') +
  labs(title = 'Model fit for winpercent vs. pricepercent \n for average sugarpercent') +
  theme_bw()
```

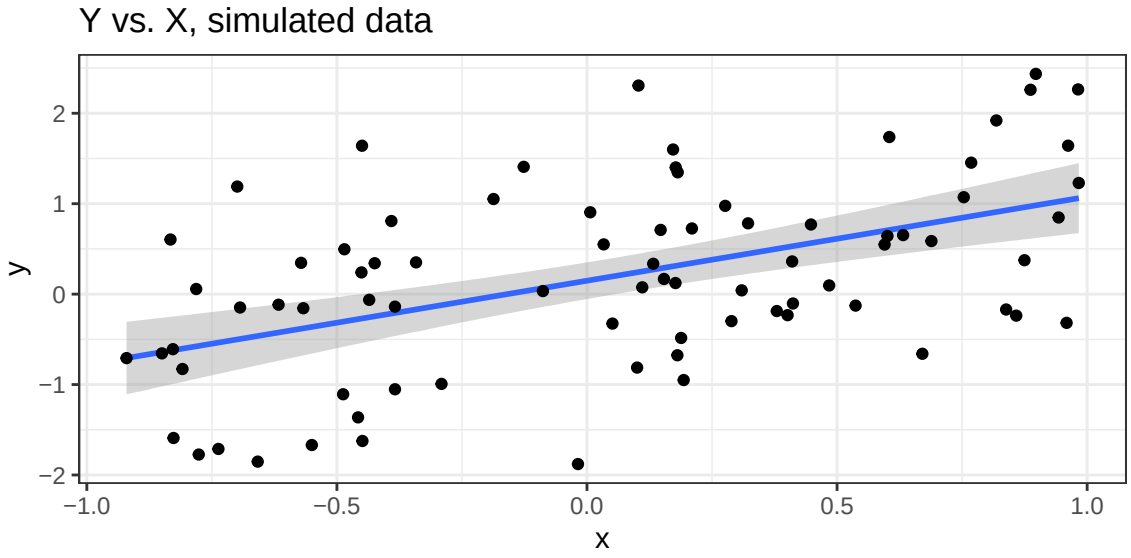


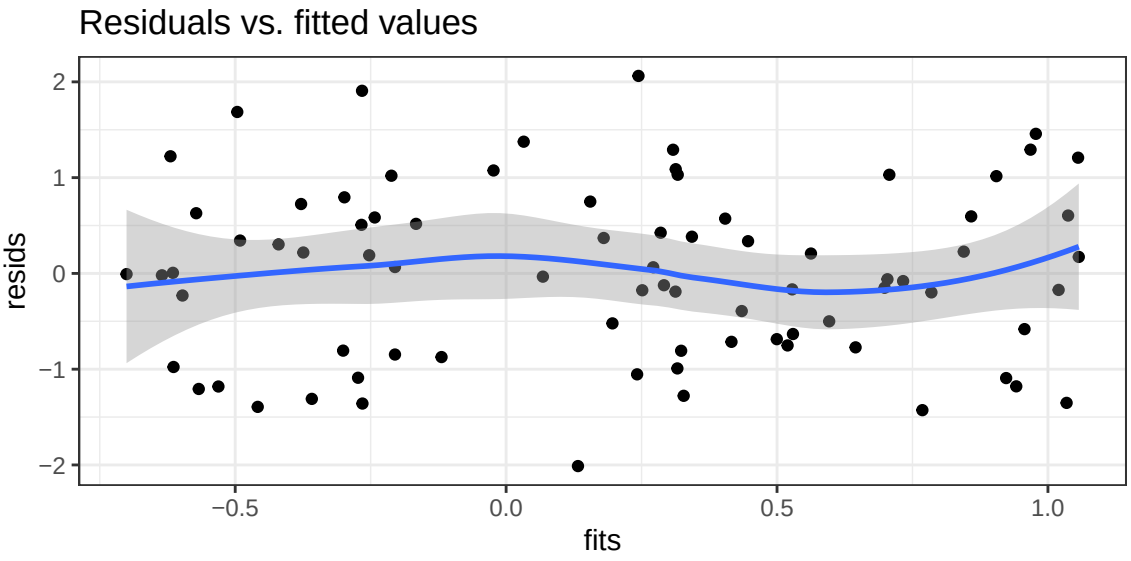
Residual Plots

Model fit can also be evaluated looking at residuals plots. Recall, residuals are defined as $r_i = y_i - X_i\hat{\beta}$.

These plots should result in absence of patterns.

Residual Plots from Fake Data It is not always obvious (at least initially) what residual plots should look like and what variations could be expected when the model is indeed true.





It can also be useful to create a panel of figures to explore residuals vs. each covariate.