Overview of Linear Models

Building and comparing regression models for prediction

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.

2. It is not always necessary to include these inputs as separate predictors

3. For inputs that have large effects, consider including their interactions as well.

4. Use standard errors to get a sense of uncertainties in parameter estimates. Know these will change if new predictors are added to the model.

5. Make decisions about including or excluding predictors based on a combination of contextual understanding (prior knowledge), data, and the uses of the regression model
   a. If the coefficient of a predictor is estimated precisely, generally makes sense to keep it in the model

   b. If the standard error is large and there seems to be no substantive reason to include it in the model, it can make sense to remove it.

   c. If the predictor is important for the problem at hand (groups interested in comparing or controlling for), generally recommend keeping it in the model.

   d. If a coefficient does not make sense (unexpected sign), try to understand how this could happen.

10 tips to improve your regression modeling

From appendix B

1. Think about variation and replication

2. Forget about statistical significance

3. Graph the relevant and not the irrelevant
    a. Graph the fitted model
    b. Make many graphs
    c. Don't graph the irrelevant

4. Interpret regression coefficients as comparisons

5. Understand statistical methods using fake-data simulations

6. Fit many models

7. Set up a computational workflow
    a. Data subsetting
    b. Fake-data and predictive simulation

8. Use transformations

9. Do causal inference in a targeted way

10. Learn methods through live examples.

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*

Regression with Multiple Predictors

More beer…

Now consider jointly considering both weekend/weekday and maximum temperature. The model can now be written as

$$y = \beta_0 + \beta_1 x_{weekend=1} + \beta_2 x_{tmp} + \epsilon,$$

where:

- y is the beer consumption,

- β_0 *is the consumption on a weekday with maximum temperature of 0*

- β_1 *is the expected difference in consumption between a weekend day and a weekend, holding maximum temperature constant*

- β_2 *is the expected difference in consumption for a 1 degree change in maximum temperature, holding the day of week constant.*

```
ml_regression <- beer %>% stan_glm(consumed ~ weekend + max_tmp, data = ., refresh = 0)
ml_regression
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      consumed ~ weekend + max_tmp
##  observations: 365
##  predictors:   3
## ------
##             Median MAD_SD
## (Intercept) 5.9    0.8
## weekend     5.2    0.3
## max_tmp     0.7    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 2.4    0.1
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Coefficient Interpretation

- When interpreting coefficients in a multiple regression model, it is important to understand that these values control for other predictors in the model! *The values will change with the inclusion/exclusion of other predictors.*

- Sometimes we cannot necessarily hold all other predictors constant in a model. *A simple example would be* $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

The textbook puts an emphasis on differentiating predictive and counterfactual interpretations:

- The predictive interpretation focuses on how the outcome differs, on average, when *comparing* two groups of items that differ by 1 unit (and all other predictors are the same). *The coefficient is the expected difference in y between these two items.*

- The counterfactual interpretation focuses on how the outcome would differ with an individual, rather than between individuals. *The coefficient is the expected change in y caused by adding one to the predictor (holding all other predictors the same)*

The counterfactual interpretation should be reserved for a situation where causal inferences are reasonable, such as a completely randomized experimental design.

It is easy to get careless with wording and say things like "a change in temperature is associated with a change in consumption," but *the safest interpretation focuses on comparisons between units rather than changes within units.*