CH 12: Transformations

Scaling Predictors

We have seen that centering predictors makes interpreting the intercept more intuitive. *This is especially useful when including interactions in the model.*

Beyond just centering variables, *standarding predictors can also be helpful. This puts them on the same scale, so it is easier to identify the most important variables.*

A common standardization approach is using a z-score, *substracting the mean and dividing by the standard deviation.*

Standardization can use a conventional center point that makes intuitive sense, *say 3 bedrooms rather than 2.7 bedrooms.*

Logarithmic Transformations

When additivity and linearity are not reasonable, we often need to consider nonlinear transformations. Transformation can be applied to predictors, the outcome variable, or both.

With outcomes that are necessarily positive, *butterfly distance or housing prices, taking a logarithmic transformation of the response variable can be useful. This restricts predictions (of distance) to be positive.*

Our general linear model framework results in additive models, but consider:

$$\begin{aligned} \log(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\ y &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon\right) \\ y &= \exp(\beta_0) \times \exp(\beta_1 x_1) \times \exp(\beta_2 x_2) \times \exp(\epsilon) \end{aligned}$$

thus the relationship is now multiplicative. Model coefficient correspond to the log outcome.

Building and comparing regression models for prediction

Before moving forward to think about logistic regression and generalized linear models, it is important to present approaches for building and comparing models.

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 *total score*

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.

LOO (or AIC) can be used for model comparisons.

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.