Regression and Other Stories: Ch 1.4 - 1.6

Building, interpreting, and checking regression models

The authors present four cycles for an iterative data analysis process:

1. Model Building: *start with simple linear regression and expanding to include additional predictors, transformations, and interactions.*

2. Model Fitting: *writing code to estimate regression coefficients and uncertainties*

3. Understanding model fits: *data visualization, investigation of connection between data and model fits*

4. Criticism: *finding flaws, questionable assumptions and considering improvements to the model or summarizing the limitations and claims that can be made from the model.*

Classical and Bayesian Inference

Model fitting can be done in different ways... With any approach there are three considerations:

1. *information: what is used for estimation*

2. *assumptions*

3. *interpretation*

Information Information pertains to what *data is used to estimate the model, how that data was collected, and whether prior knowledge exists about the data.*

Assumptions The authors discuss three basic assumptions that underlay a regression model

1. *functional form of the relationship between x and y, for instance $y = x\beta + \epsilon$*

2. *where the data comes from: sample/observational study, non-response, etc..*

3. *real world relevance of the measured data: are responses accurate, can responses be generalized to other settings, places, times...*

I'd probably add a fourth assumption about the distributional nature of the responses – more later.

Interpretation **Classical (or frequentist) Inference:** This approach summarize the data *(not including prior opinions) to get estimates with well understood statistical properties, low bias and low variance.*

The results and interpretation are based long-run expectations of the methods that are correct on average (unbiased) and confidence intervals that contain the true parameter the appropriate percent of the time (coverage). *However, the interpretation about a single study can be tricky (see STAT 216).*

Classical methods do tend to be conservative, in that strong statements are not make with *weak* data. *Classical methods do have a clear objective path, assuming assumptions are checked and frequency properties are a reasonable solution.*

Inference is largely driven by Null Hypothesis Significance Testing (NHST) and p-values.

Bayesian Inference: This approach summarize the data *and includes existing prior information.*

Results and interpretations are probabilistic *(e.g. The probability that the parameter is in the interval is 95 %.) can be summarized by simulation*

Bayesian inference uses additional information which can potentially give more reasonable results (using the prior to regularize the model), *but specifying the prior information requires additional assumptions and can be subjective.*

Inference is largely summarized using posterior distributions of parameters.

Computing

Classical methods tend to use least-squares estimation (or maximum likelihood).

```r
beer <- read_csv('http://math.montana.edu/ahoegh/Data/Brazil_cerveja.csv')
```

```
## Parsed with column specification:
## cols(
##   consumed = col_double(),
##   precip = col_double(),
##   max_tmp = col_double(),
##   weekend = col_double()
## )
```

```r
lm_beer <- lm(consumed ~ max_tmp, data = beer)
summary(lm_beer)
```

```
##
## Call:
## lm(formula = consumed ~ max_tmp, data = beer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9116 -2.8451 -0.3342  2.3929  8.6191
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.97494    1.10459    7.22 3.07e-12 ***
## max_tmp      0.65485    0.04097   15.98  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.375 on 363 degrees of freedom
## Multiple R-squared:  0.413,  Adjusted R-squared:  0.4114
## F-statistic: 255.4 on 1 and 363 DF,  p-value: < 2.2e-16
```

The textbook authors (and your instructor), recommend using Bayesian inference for regression. *Others here, and elsewhere, may be more familiar with the classical methods. So we will still consider both throughout the class.*

Furthermore, using Bayesian methods with *weakly informative* prior information enables stable estimates and simulation based inference, but also can result (or approximately result) in frequentist solutions.

```r
stan_glm(consumed ~ max_tmp, data = beer, refresh = 0) %>% print()
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      consumed ~ max_tmp
##  observations: 365
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 8.0    1.1
## max_tmp     0.7    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 3.4    0.1
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```