Regression and Other Stories: Ch 1.1 - 1.4

Regression and Other Stories

Why this book? This book, and the precursor (ARM), are on the applied end of the spectrum, but they "focus on understanding regression models and applying them to real problems."

The first few weeks of the course will parallel Part I in ROS, which focuses on key tools and concepts in mathematics, *statistics*, and *computing*.

There will be additional examples and details in the textbook. Furthermore, the textbook web page has code and data (https://avehtari.github.io/ROS-Examples/) to replicate examples in the textbook.

Initially we will focus on challenges pertaining to statistical inference and regression modeling with an emphasis on predictive models (as opposed to casual models). ROS lists four key skills:

Why Statistics

Q: Why are we here? Why study Statistics?

ROS details three challenges of statistical inference. **Q**: What does inference mean?

Each of these challenges can be formulated through the lens of prediction (new observations, future outcomes, etc..) although the third challenge does require valid measurements pertaining to the construct of interest.

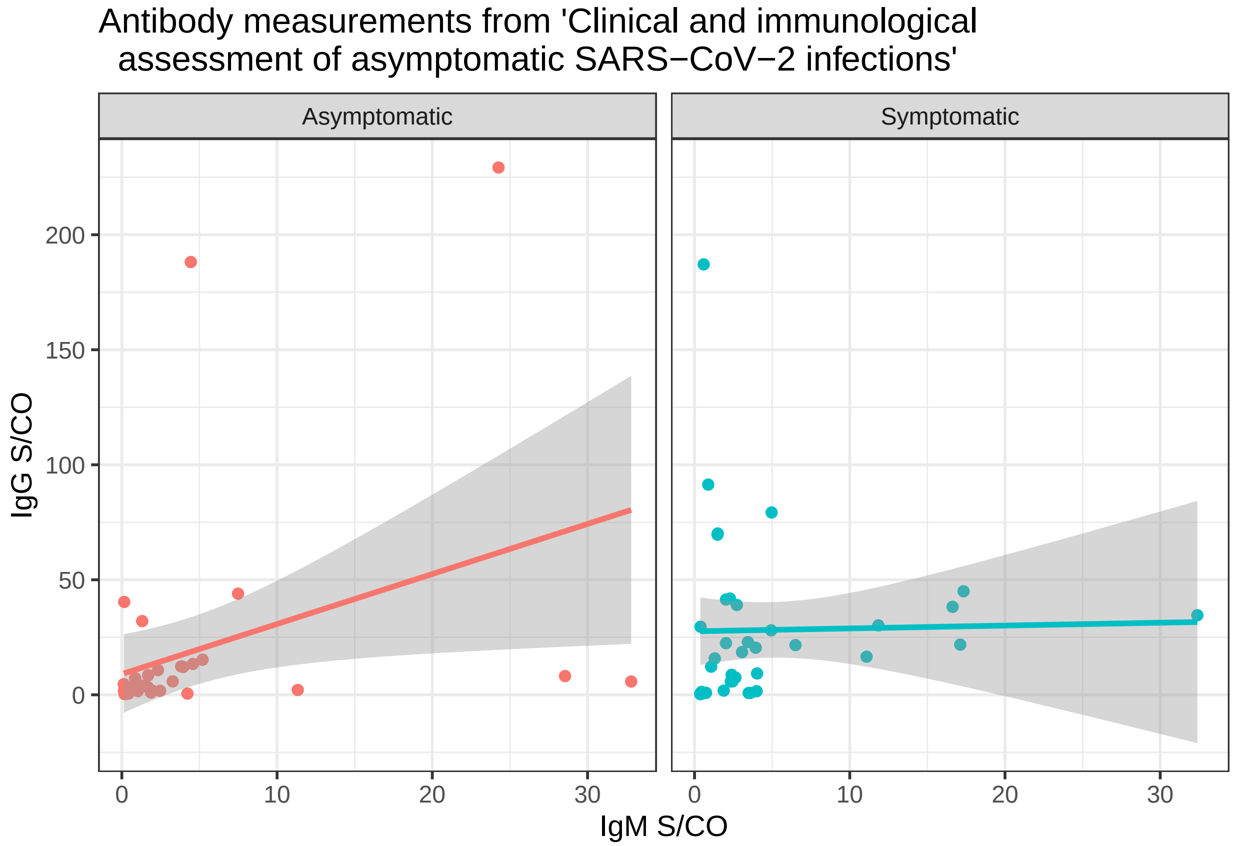
Regression

Using ROS language,

Note this is different than casual language, which might be summarized as “regression is a method that allows researchers to summarize how changing a value of a predictor causes an outcome to change.”

Using the antibody data that we have previously seen, a regression model can be visualized as:

```r
read_csv("http://math.montana.edu/ahoegh/Data/Covid_3a.csv") %>%
  ggplot(aes(y = `IgG S/CO`, x = `IgM S/CO`, color = Group)) +
  geom_point() + theme_bw() + facet_wrap(.~ Group) +
  geom_smooth(method = 'lm', formula = "y ~ x") +
  ggtitle("Antibody measurements from 'Clinical and immunological
  assessment of asymptomatic SARS-CoV-2 infections'") +
  theme(legend.position = "none")
```



This will be one of the first of many figures that we create.

A formal regression model associated with the asymptomatic patients can also be specified in R.

```
covid <- read_csv("http://math.montana.edu/ahoegh/Data/Covid_3a.csv")
asymptomatic <- covid %>% filter(Group == "Asymptomatic") %>%
  rename(IgG = `IgG S/CO`, IgM = `IgM S/CO`)
lm_fit <- lm(IgG ~ IgM, data =asymptomatic)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = IgG ~ IgM, data = asymptomatic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.605  -9.387  -8.348  -4.987 169.464
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.0114     8.4822   1.062   0.2953
## IgM           2.1752     0.9706   2.241   0.0315 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.1 on 35 degrees of freedom
## Multiple R-squared:  0.1255, Adjusted R-squared:  0.1005
## F-statistic: 5.023 on 1 and 35 DF,  p-value: 0.03146
```

Q: Interpret the output here.

```
stan_fit <- stan_glm(IgG ~ IgM, data =asymptomatic, refresh = 0)
print(stan_fit)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      IgG ~ IgM
##  observations: 37
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 9.1    8.6
## IgM         2.2    0.9
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 45.4    5.2
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Q: Now, how do we interpret this output? Where are the p-values?

What can we do with regression?

- Prediction:
- Exploring Associations:
- Extrapolation:
- Causal Inference:

For any of the settings, the model needs to be a reasonable approximation of reality. To summarize the George Box quote, "all models are wrong, but some are useful." In other words, the model needs to have enough complexity to capture all of the necessary information.

Regression Interpretations

There are two common ways regression can be used for causal inference:

Estimating a relationship

For causal inference,

The easiest way to establish comparable groups is to use

Given a treatment x and an outcome y, which can be continuous or categorical,

One way to model this with an

Controlling for Differences

In many scenarios, the units that receive different treatments may vary.

So the goal is to adjust for the differences in the experimental units before assigning and applying the treatments. This difference is often referred to as

A regression model can be used to adjust for pre-treatment differences.

Coefficients in Predictive Models

Even if the goal is just prediction, interpreting regression coefficients can be tricky. Consider a dataset consisting of the volume of beer consumed in Sao Paulo, Brazil. For more information about the data, see https://www.kaggle.com/dongeorge/beer-consumption-sao-paulo. We will work on a cleaned dataset that contains:

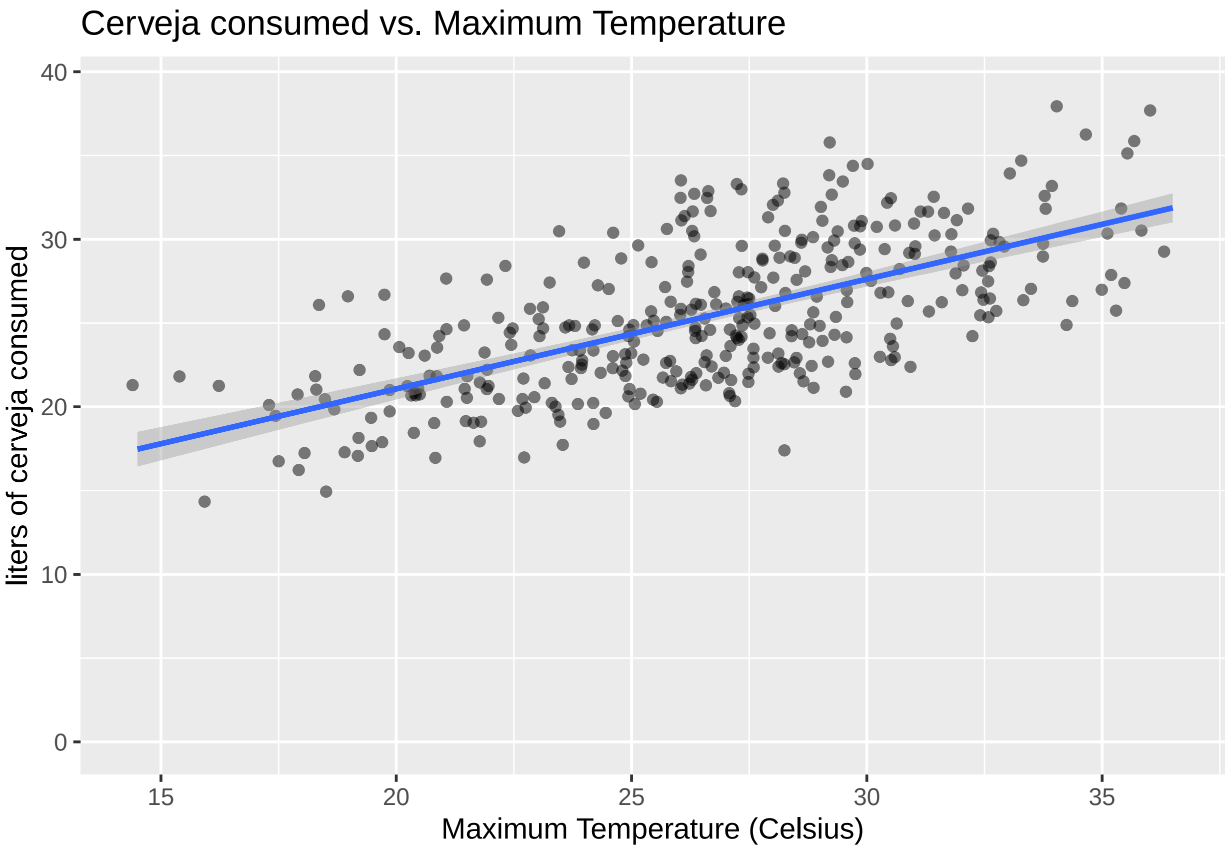
- consumed: daily volume of beer consumed in liters
- precip: daily precipitation in (mm)
- max_tmp: daily maximum temperature in C
- weekend: indicator variable for if the day is a weekend.

It is not obvious how the data was collected, but here is a note from the data provider: "The data (sample) were collected in São Paulo, Brazil, in a university area, where there are some parties with groups of students from 18 to 28 years of age (average)."

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

Let's explore the relationship between maximum daily temperature and liters of beer consumed.

Q: What is the association between temperature and beer consumption?



```r
stan_glm(consumed ~ max_tmp, data = beer, refresh = 0) %>% print(digits = 2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      consumed ~ max_tmp
##  observations: 365
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 7.94   1.11
## max_tmp     0.66   0.04
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 3.38   0.12
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Q: can we say that each degree *causes* 0.66 (±.8) more liters of beer to be consumed?