

Regression and Other Stories: Ch 2

Data and measurement

This linear models course focuses on:

- fitting lines (functional relationships),
- making comparisons and predictions, and
- assessing uncertainties in inferences.

A major component of these steps requires understanding and assessing assumptions. However, before talking about that and more, it is necessary to understand the data. There is a common adage,

Validity and Reliability

As you start working with datasets, I'd encourage you to think about where the data came from

Measurement Consider the data collection process for vegetation cover class data.



Figure 8. 1 m² quadrat frame with dashed lines showing 25% and 5% coverages

Figure 1: Cover class data, source: Upper Columbia Basin Network Sagebrush Steppe Vegetation Monitoring Protocol by Yeo, Rodhouse, Dicus, Irvine, Garrett

When designing studies and collecting data, precision of the measurements is an important consideration

Measurement can be challenging in many situations, a few include:

Taking multiple measurements can be advantageous.

Validity The **validity** of measurement is the degree to which it represents what you are trying to measure.

ROS defines validity as

Reliability The **reliability** of a measurement is the degree to which it is

Sample Selection A fundamental goal in statistics is to extend inferences from a sample to a study population.

This extension requires understanding how the data was selected. In particular,

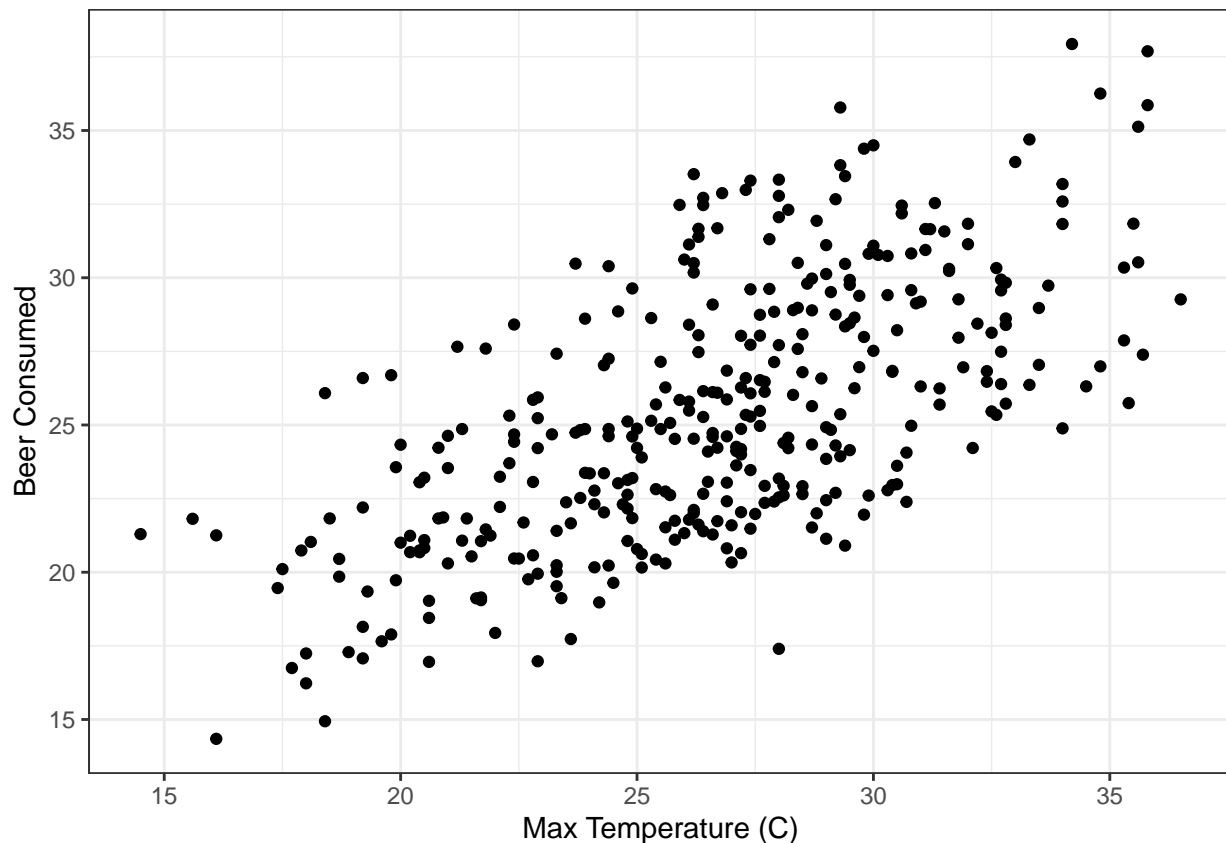
Data Visualization

ROS suggests that all graphs are comparisons.

```
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()
## )
```

Q: What kind of graph is this and what is the comparison?



Graphical displays of data enable viewing several pieces of data. Specifically, a different variable can be associated with all of the following elements.

Graphical principles

```
fig1 <- beer %>% mutate(weekend_fact = factor(weekend, labels = c('weekday', 'weekend'))) %>%
  ggplot(aes(y = consumed, x = max_tmp, color = weekend_fact, shape = weekend_fact)) +
  geom_point() + theme_bw() + ylab('Beer Consumed (l)') +
  xlab('Max Temperature (C)') + labs(color = 'Day of Week') +
  geom_smooth(method = 'loess', formula = 'y~x') + guides(shape = FALSE) +
  ggtitle('Sample of Beer Consumption in Sao Paolo, Brazil')

fig2 <- beer %>% mutate(weekend_fact = factor(weekend, labels = c('weekday', 'weekend'))) %>%
  ggplot(aes(y = consumed, x = weekend_fact, color = weekend_fact, shape = weekend_fact)) +
  geom_violin() + geom_jitter() + ylab('Beer Consumed (l)') +
  ylim(0, NA) + theme_bw() + theme(legend.position = 'none') + xlab('Day of Week') +
  labs(caption = 'source:https://www.kaggle.com/dongearge/beer-consumption-sao-paulo')

grid.arrange(fig1, fig2)
```

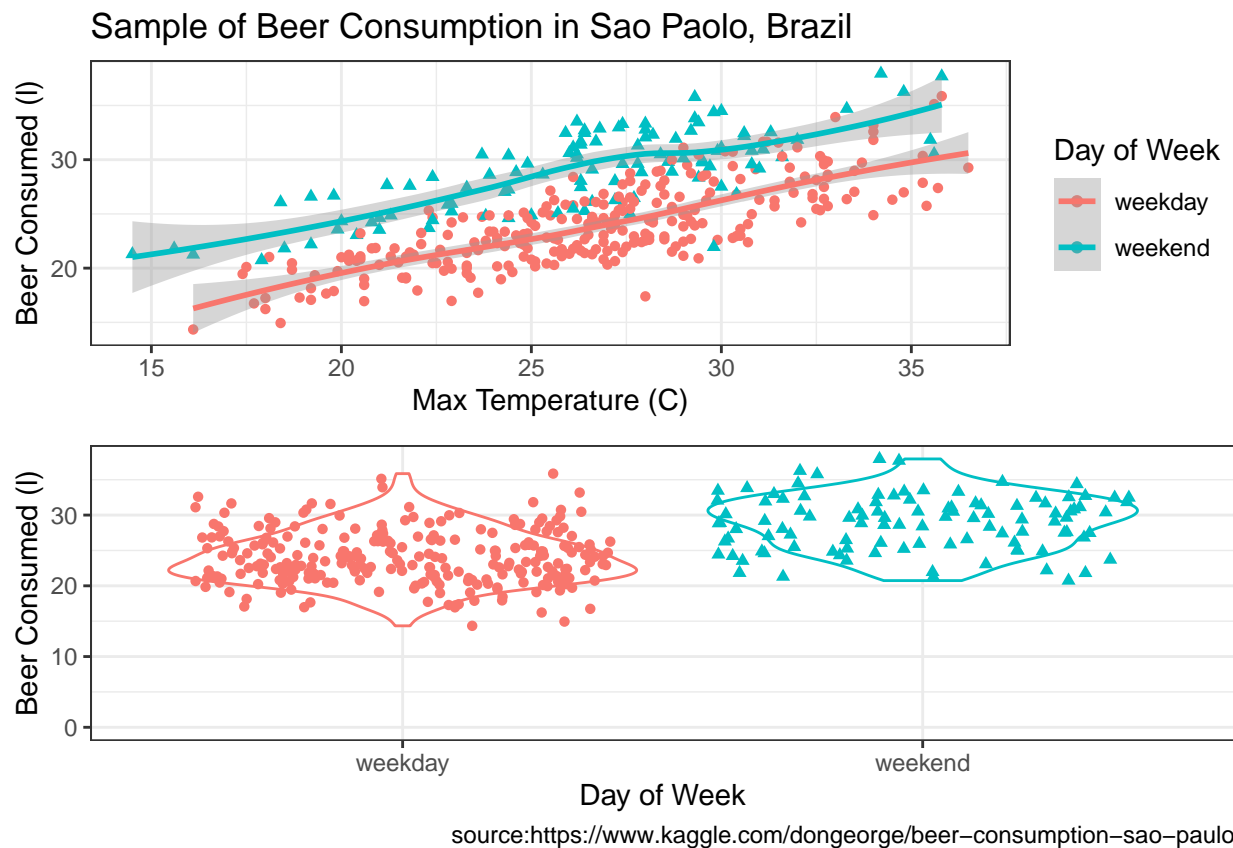


Figure 2: Exploration of beer consumption by temperature and weekend / weekday. The figures suggest higher temperatures and weekends are associated with predicted increases in beer consumption.

There are three main types of graphics:

1. Exploratory Data Analysis:

2. Graphs of fitted models and inferences:

3. Graphs of Final Results: