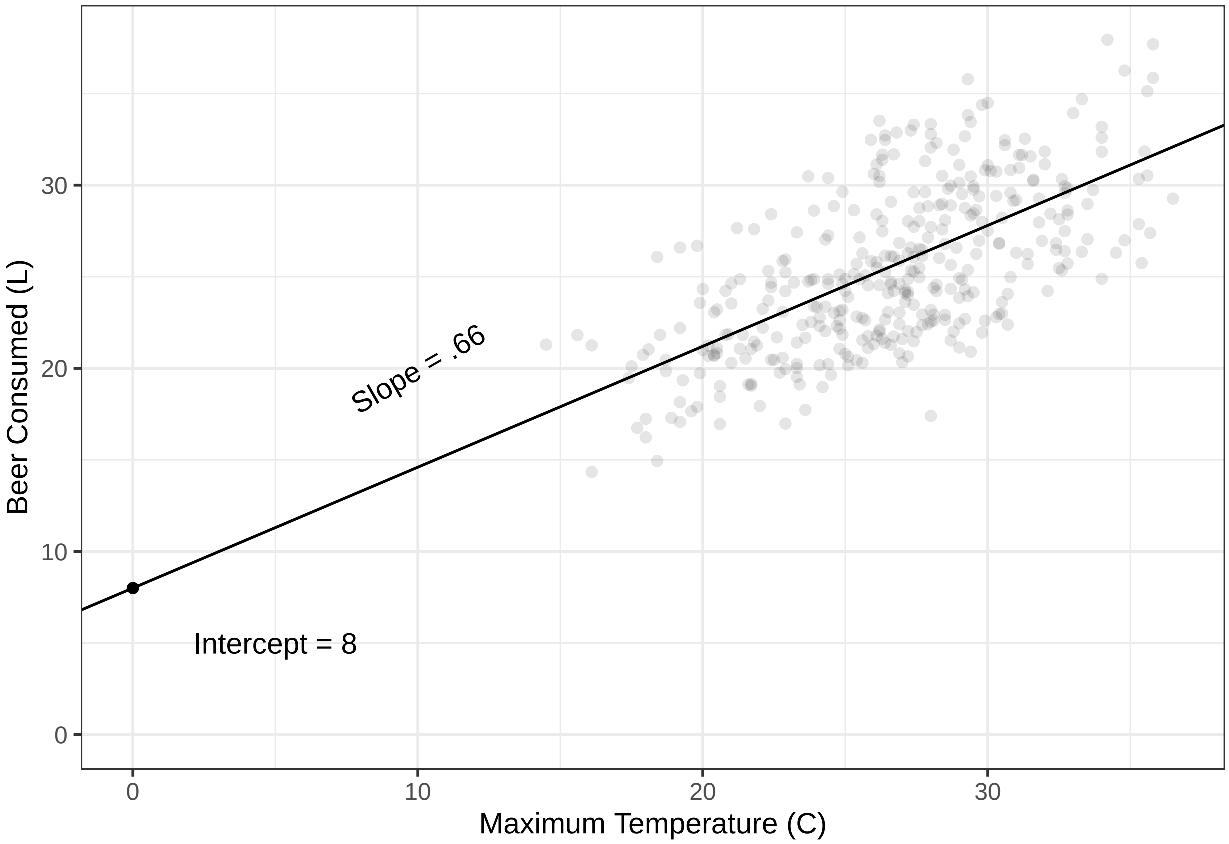
Regression and Other Stories: Ch 3

Chapter 3 focuses on important statistical and mathematical understanding for this class. As the semester progresses, and depending on how 506 will be taught, the treatment of these things might differ.

Recall the Brazilian Beer dataset. Let y_i be the response variable, beer consumption, and x_i be the maximum temperature, both on day i.

Then using scalar notation, we can write out the model as

This model results in a linear relationship between y and x, for the deterministic portion.



Linear Algebra

Using matrix algebra, this model can also be formulated as

$$\mathbf{y} = X\beta + \epsilon, \tag{1}$$

where $\mathbf{y}$ is a $n \times 1$ vector, X is a $n \times 2$ matrix, β is a 2×1 vector, and ϵ is a $n \times 1$ vector.

Non-linear relationships

While linear models assume a linear relationship between y and x, we can also transform covariates.

```
## Simulate logarithmic relationship
dat <- tibble(x = runif(1000, min = 1, max = 10)) %>%
  mutate(log_x = log(x), y = log_x + rnorm(1000, mean = 0, sd =.1))
```

```
fig1 <- dat %>% ggplot(aes(y=y,x=x)) + geom_point(alpha = .2) +
  geom_smooth(method = 'lm', formula = 'y~x', color = 'red', linetype = 2) +
  geom_smooth(method = 'loess', formula = 'y~x') + theme_bw()
fig2 <- dat %>% ggplot(aes(y=y,x=log_x)) + geom_point(alpha = .2) + theme_bw() +
  geom_smooth(method = 'lm', formula = 'y~x')
grid.arrange(fig1, fig2, nrow = 1, ncol = 2)
```

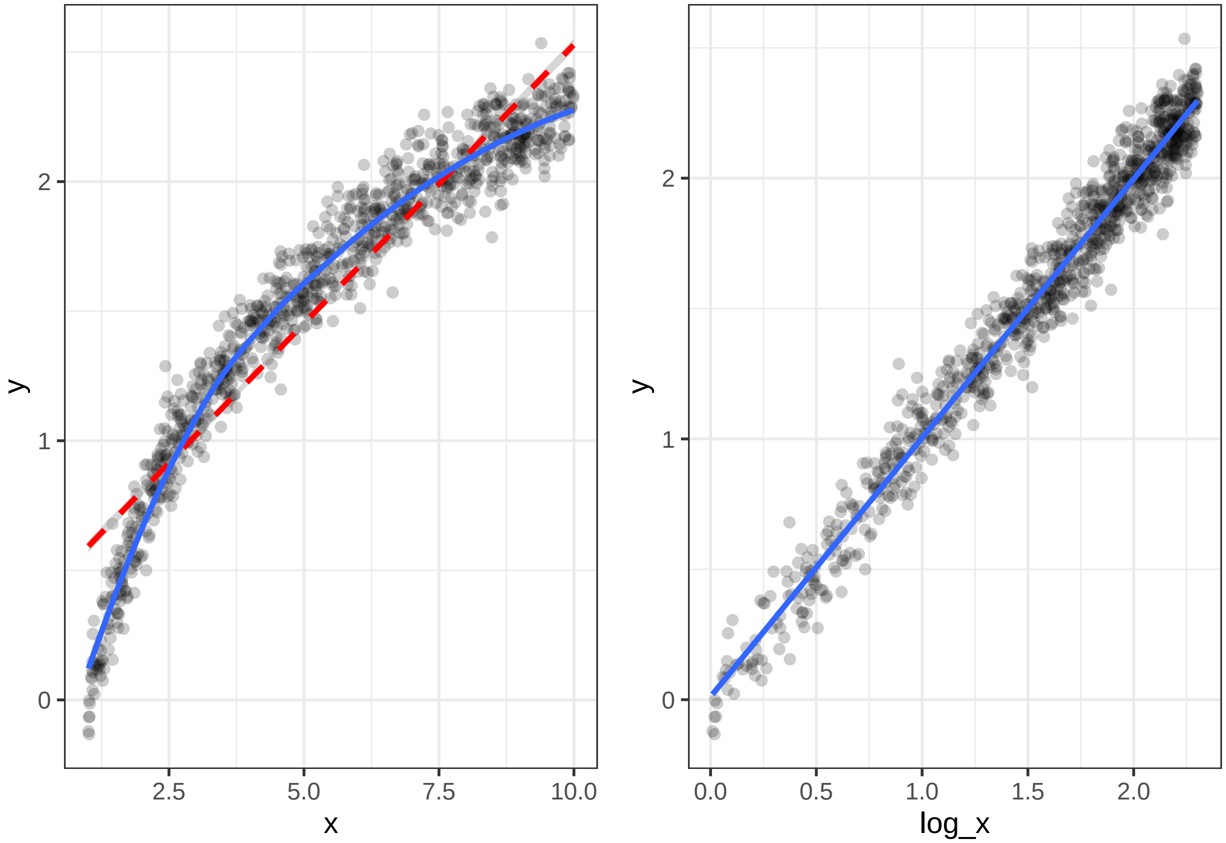


Figure 1: Plots of y vs. x (left panel) and log(x) (right panel). The figure in the left shows a non-linear relationship between y and x. The blue line is a LOESS curve and the red line is a linear relationship. The panal on the right shows a linear relationship between y and log(x).

Probability Distributions

While regression models allow us to fit linear (and non-linear) functional relationships,

The error term (ϵ) controls the randomness fundamental to data and that provides a mechanism to

Understanding uncertainty begins with a discussion about fundamental statistical terms of a probability distribution.

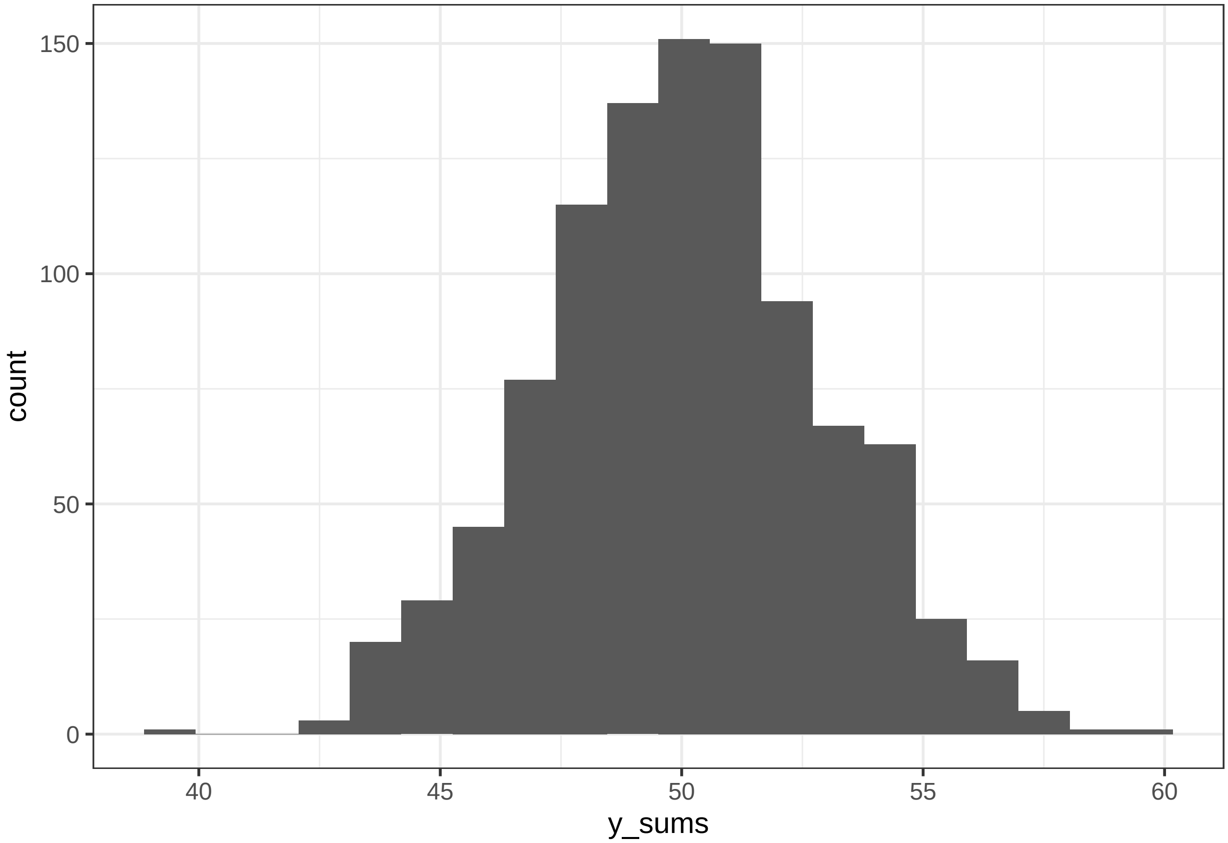
- mean:
- variance:
- standard deviation:

Normal Distribution The Central Limit Theorem (CLT) states

> that the sum of many small, independent random variables will be a random variable that approximates a normal distribution.

The following code demonstrates the CLT. **Q** what does each line of code do? Then complete the caption and/or title on the figure.

```
num_sims <- 1000
num_obs <- 100
y <- matrix(runif(num_obs* num_sims), nrow = num_sims, ncol = num_obs)
y_sums <- tibble(y_sums = rowSums(y))
y_sums %>% ggplot(aes(x=y_sums)) +
  geom_histogram(bins = 20) +
  theme_bw()
```



In R we can:

1. Simulate from distributions

```
rnorm(n = 1, mean = 0 , sd = 1)
```

```
## [1] -0.8838417
```

2. Find quantiles from distributions

```
qnorm(.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

3. We can also evaluate probability densities and cumulative distribution functions with `d` and `p`.

If a random variable, x is normally distributed, then a linear transformation of x is also normally distributed.

Assume $x \sim N(0, 1)$, then

- ax

- x

- ax

Other probability Distributions

More on these later: