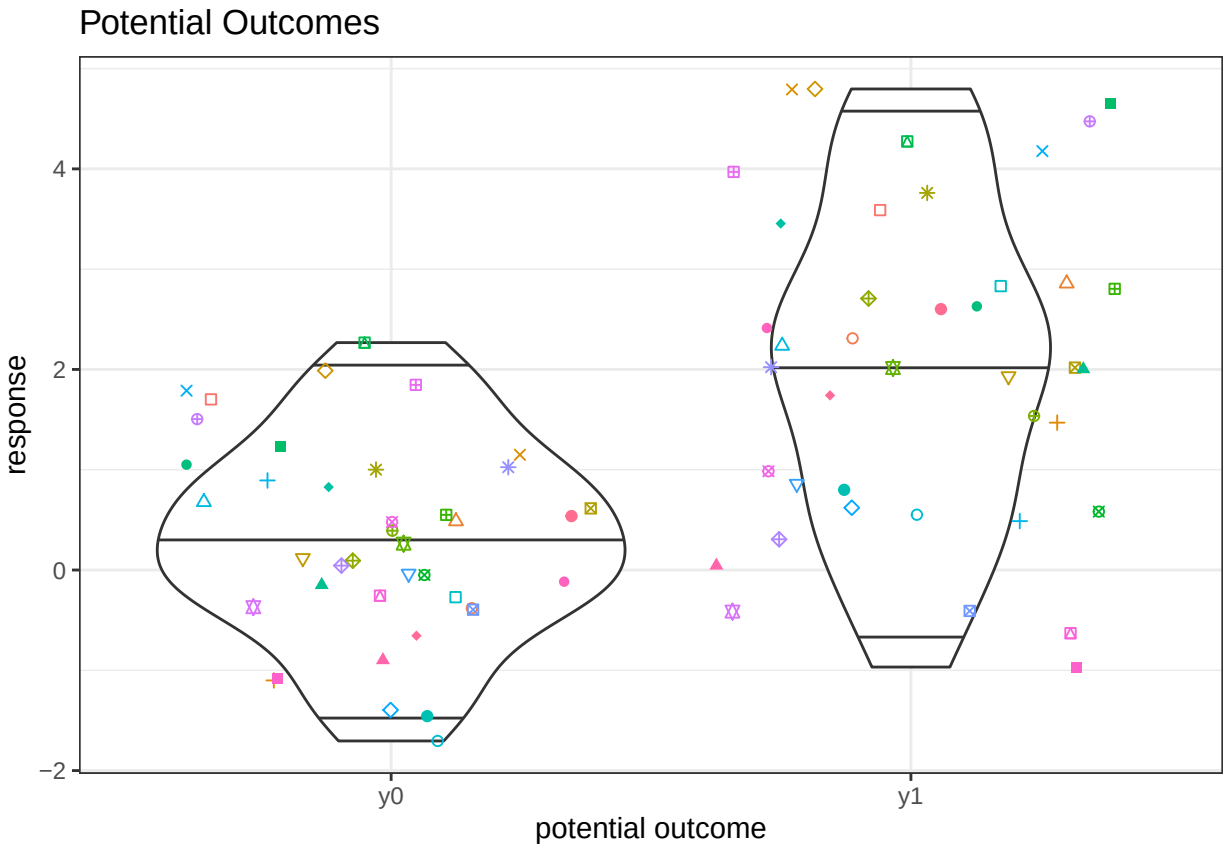
Causal Inference and Regression

Causal inference can be characterized as a predictive problem, where the question is *what would have happened under different circumstances.*

Simulate and visualize data with two potential outcomes

```
n <- 40
y0 <- rnorm(n)
y1 <- y0 + rnorm(n, mean = 2)
```

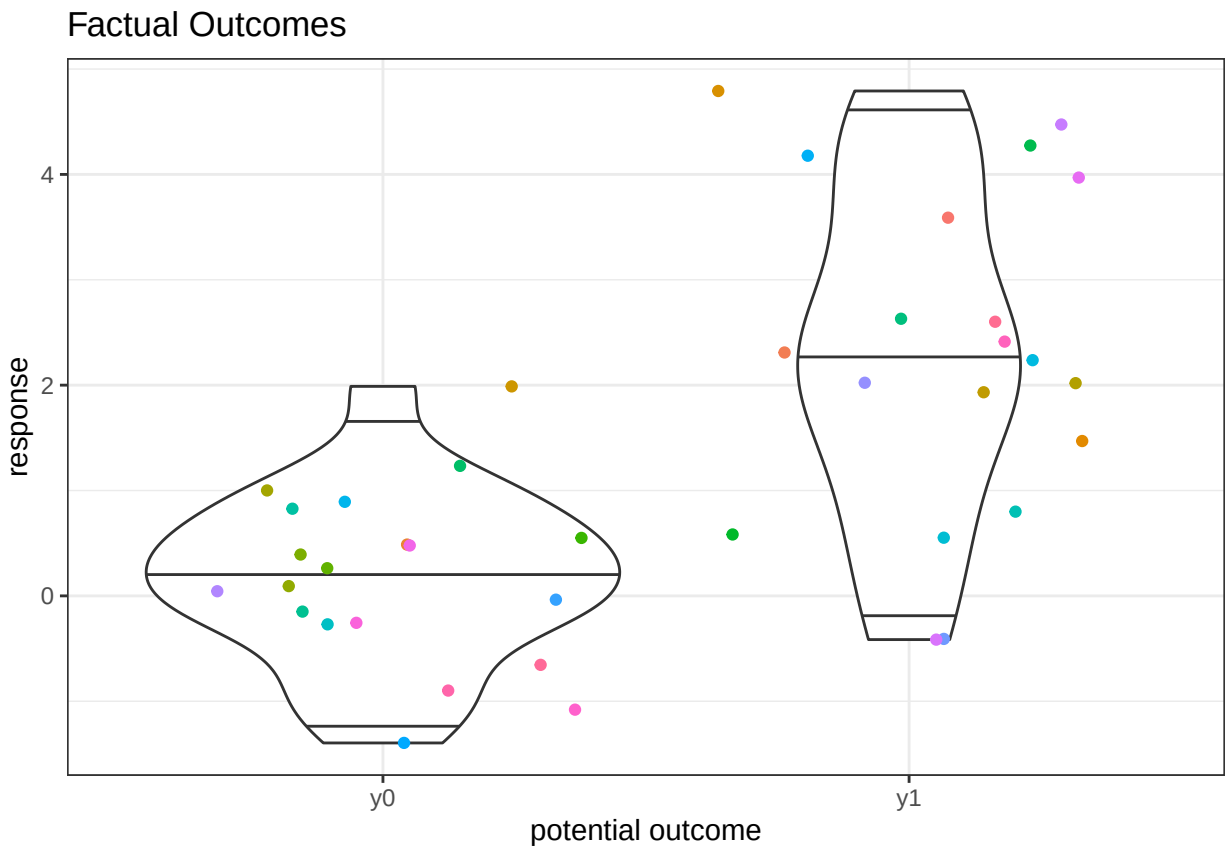
```
dat <- tibble(id = factor(rep(1:n, 2)), response = c(y0, y1),
              `potential outcome` = rep(c('y0','y1'), each = n))

dat %>% ggplot(aes(y = response, x = `potential outcome`)) +
  geom_violin(draw_quantiles = c(.025, .5, .975)) +
  geom_jitter(aes(color = id),  shape = rep(c(0:19),4)) +
  theme_bw() + theme(legend.position = 'none') +
  ggtitle('Potential Outcomes')
```



Then randomly assign treatments to each unit and visualize differences

```
treatment <- sample(rep(0:1, each = n/2))
sample_dat <- tibble(id = factor(1:n), response = y0,
                     `potential outcome` = rep('y0', n)) %>% filter(treatment == 0) %>% bind_rows(
                    tibble(id = factor(1:n), response = y1,
                     `potential outcome` = rep('y1', n)) %>% filter(treatment == 1)
                     )
```

```
sample_dat %>% ggplot(aes(y = response, x = `potential outcome`)) +
  geom_violin(draw_quantiles = c(.025, .5, .975)) +
  geom_jitter(aes(color = id)) + theme_bw() + theme(legend.position = 'none') +
    ggtitle('Factual Outcomes')
```



Pre-treatment covariates Consider a setting with:

- pre-treatment covariates for sampling units (either continuous or categorical)

- treatments (control + treatment) applied to the sampling units.

Begin with a randomized block design (paired comparisons, is a special case with blocks of size 2). Analysis can

- use differences between treatment and control

- use differences between treatment and control for each (block)

- Adjusting for pre-treatment variables (categorical/blocks or continuous)

$$y_i = \tau z_i + X_i\beta + \epsilon_i,$$

where z_i is an indicator for treatment and τ is the average treatment in the grade.

Note this assumes the average treatment is the same for each block. How would we allow for varying treatment effects?