Regression: dummy variables

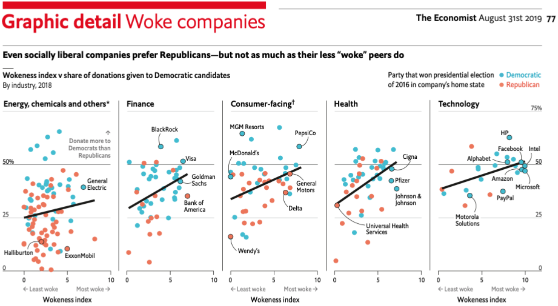
Author: Nicholas G Reich

Outline

- Dummy variables for categorical covariates

Categorical predictors

- Assume X is a categorical / nominal / factor variable with k levels: e.g. ‘Industry’.
- If you use a single predictor with continuous values of $1, 2, \ldots, K$ this assumes that a “one unit increase” has a clear meaning.
- You need to create *indicator* or *dummy* variables so that each level stands on its own and can be estimated separately.



Graphic detail Woke companies

The Economist August 31st 2019 **77**

Even socially liberal companies prefer Republicans—but not as much as their less “woke” peers do

Wokeness index v share of donations given to Democratic candidates

By industry, 2018

Party that won presidential election of 2016 in company’s home state ● Democratic ● Republican

Categorical predictors

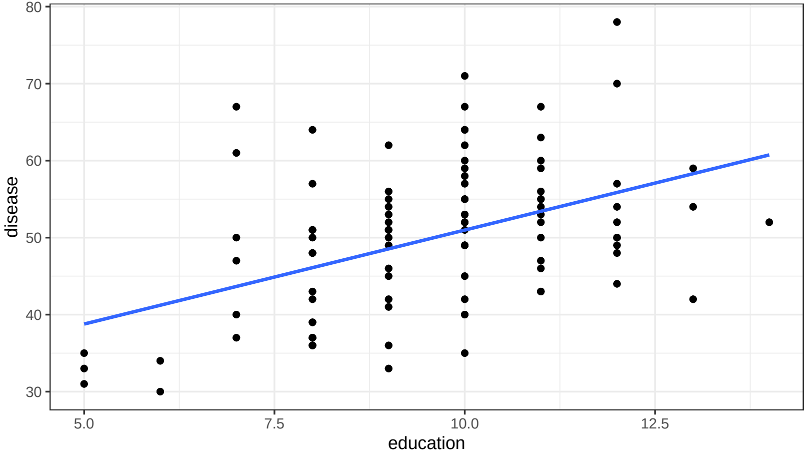
Important to distinguish between...

- “nominal” categorical variables: e.g. ones with no natural ordering, such as Industry, country, etc...
- “ordinal” categorical variables: e.g. ones with a natural ordering, such as education level, or age grouping.

Categorical predictor example: lung data

Education could plausibly be continuous (e.g. you could interpret a one-unit increase), but likely a linear assumption is not great. Thinking of education as a "factor" may be more practical.

```
qplot(education, disease, data=dat) + geom_point() +
  geom_smooth(method="lm", se=FALSE)
```

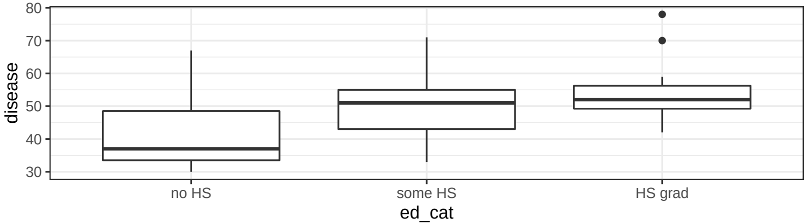


Defining a categorical variable

We could define educational level relative to high-school (HS) achievement.

$$\text{ed\_cat}_i = \begin{cases} \text{no HS}, & \text{if education}_i < 5 \\ \text{some HS}, & \text{if } 5 \leq \text{education}_i < 8 \\ \text{HS grad}, & \text{if } 8 \leq \text{education}_i \end{cases}$$

```
dat$ed_cat <- cut(dat$education, breaks = c(-Inf, 8, 12, Inf),
                  right=FALSE,  ## intervals "open" on the right
                  labels=c("no HS", "some HS", "HS grad"))
qplot(ed_cat, disease, geom="boxplot", data=dat)
```



Indicator variables

- An indicator variable is a binary variable. Multiple indicator variables can be used to encode which of multiple categories an observation belongs to. When constructed as below, these are referred to as ‘dummy variables’.
- Let x be a categorical variable with k levels .
- Choose one group as the baseline (e.g. “no HS”).
- Create $(k-1)$ binary variables to encode the information about which group each observation belongs to.

```
dat$someHS <- as.numeric(dat$ed_cat=="some HS")
dat$HSgrad <- as.numeric(dat$ed_cat=="HS grad")
dat[8:13, c("disease", "education", "ed_cat", "someHS", "HSgrad")]

##    disease education  ed_cat someHS HSgrad
## 8       58        10 some HS      1      0
## 9       52        14 HS grad      0      1
## 10      57        12 HS grad      0      1
## 11      43        11 some HS      1      0
## 12      48         8 some HS      1      0
## 13      34         6   no HS      0      0
```

Standard model interpretation

```r
## note that R doesn't need the two indicator variables we created by hand
## the lm() function will create them for us, saving us work.
mod1 <- lm(disease ~ crowding + ed_cat, data=dat)
```

Interpret:

$\text{dis}_i = \beta_0 + \beta_1 \cdot \text{crowding}_i + \beta_2 \cdot \text{someHS}_i + \beta_3 \cdot \text{HSgrad}_i + \epsilon_i$.

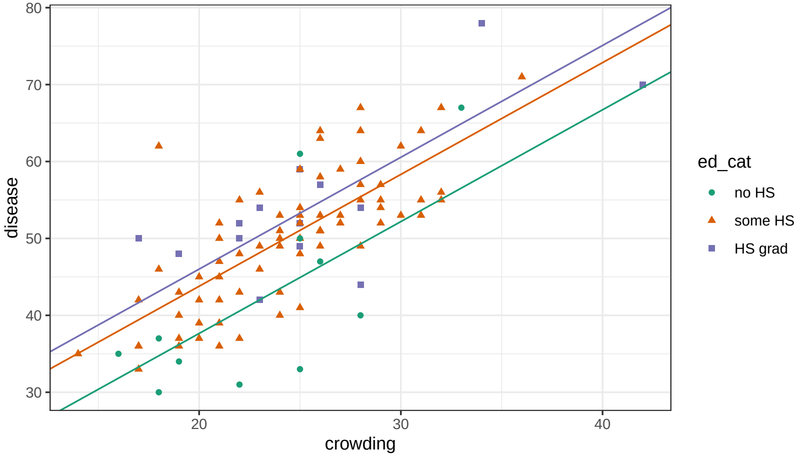
$\beta_0 =$

$\beta_1 =$

$\beta_2 =$

Categorical predictor example: lung data

```
coefs <- coef(mod1)
ggplot(dat, aes(x=crowding, y=disease, color=ed_cat, shape=ed_cat)) +
  geom_point() + scale_color_manual(values=c("#1b9e77", "#d95f02", "#7570b3"))+
  geom_abline(intercept = coefs[1], slope = coefs[2], color="#1b9e77")+
  geom_abline(intercept = coefs[1]+coefs[3], slope = coefs[2], color="#d95f02")
  geom_abline(intercept = coefs[1]+coefs[4], slope = coefs[2], color="#7570b3")
```



Categorical predictor example: lung data

$$\text{dis}_i = \beta_0 + \beta_1 \cdot \text{crowding}_i + \beta_2 \cdot \text{someHS}_i + \beta_3 \cdot \text{HSgrad}_i + \epsilon_i$$

```
mod1 <- lm(disease ~ crowding + ed_cat, data=dat)
round(summary(mod1)$coef, 2)

##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.57       3.64    2.35     0.02
## crowding            1.45       0.13   10.85     0.00
## ed_catsome HS       6.13       2.04    3.00     0.00
## ed_catHS grad       8.36       2.56    3.27     0.00
```

Categorical predictor example: interaction

$$\widehat{\text{dis}}_i = \beta_0+\beta_1\cdot c_i+\beta_2\cdot\text{someHS}_i+\beta_3\cdot\text{HSgrad}_i+\beta_4\cdot c_i\cdot\text{someHS}_i+\beta_5\cdot c_i\cdot\text{HSgrad}_i$$

In terms of the betas, what are the equations of the regression lines for predicted disease value for a hypothetical individual in the ‘no HS’, ‘some HS’ and ‘HS grad’ categories?

```
mod1 <- lm(disease ~ crowding*ed_cat, data=dat)
round(summary(mod1)$coef, 2)

##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.85       9.21    0.09     0.93
## crowding                     1.79       0.39    4.60     0.00
## ed_catsome HS               12.42      10.09    1.23     0.22
## ed_catHS grad               24.70      11.77    2.10     0.04
## crowding:ed_catsome HS      -0.27       0.42   -0.65     0.52
## crowding:ed_catHS grad      -0.67       0.48   -1.40     0.17
```