Multiple Linear Regression: Categorical Predictors

Author: Nicholas G Reich, Jeff Goldsmith

*This material is part of the **statsTeachR** project*

Multiple Linear Regression: recapping model definition

In matrix notation...

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $Cov(\boldsymbol{\epsilon}) = \sigma^2 I$

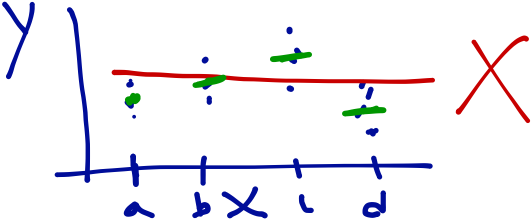
In individual observation notation...

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$

Categorical predictors

- Assume X is a categorical / nominal / factor variable with k levels
- With only one categorical X, we have classic one-way ANOVA design
- Can't use a single predictor with levels $1, 2, \ldots, K$ – this has the wrong interpretation
- Need to create *indicator* or *dummy* variables



Indicator variables

- Let x be a categorical variable with k levels (e.g. with $k = 3$ "red", "green", "blue").
- Choose one group as the baseline (e.g. "red")
- Create $(k-1)$ binary terms to include in the model:

$$
\begin{aligned}
x_{1,i} &= \mathbb{1}(x_i = \text{"green"}) = \begin{cases} 1, & \text{if green} \\ 0, & \text{o.w.} \end{cases} \\
x_{2,i} &= \underline{\mathbb{1}(x_i = \text{"blue"})}
\end{aligned}
$$

- For a model with no additional predictors, pose the model

$$
y_i = \beta_0 + \beta_1 \underline{x_{1,i}} + \ldots + \beta_{k-1} \underline{x_{k-1,i}} + \epsilon_i
$$

  and estimate parameters using least squares
- Note distinction between *predictors* and *terms*

  predictors: x

  terms: x_1, x_2

dummy

x	x_1	x_2
"g"	1	0
"r"	0	0
"b"	0	1

Categorical predictor design matrix

$$y = \underbrace{\underline{X}\beta}_{} + \varepsilon$$

Which of the following is a "correct" design matrix for a categorical predictor with 3 levels?

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{or} \quad \mathbf{X}_2 = \begin{array}{c} \begin{matrix} \beta_0 & \beta_1 & \beta_2 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} \text{red} \\ \\ \\ \text{gr} \\ \\ \\ \text{bl} \\ \\ \\ \end{matrix} \quad \text{or} \quad \mathbf{X}_3 = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix}$$

$$\cancel{\beta_0}$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

ANOVA model interpretation

X has categories
1, ..., k

Using the model $y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_{k-1} x_{k-1,i} + \epsilon_i$, interpret

$\beta_0 =$ $E(y \mid X = k \text{ or } X_1 = 0, X_2 = 0, \ldots)$

$\beta_1 =$ difference in expected value of y between the reference group (k) and group 1

Equivalent model   design matrix X_3

Define the model $y_i = \beta_1 x_{i1} + \ldots + \beta_k x_{i,k} + \epsilon_i$ where there are indicators for each possible group

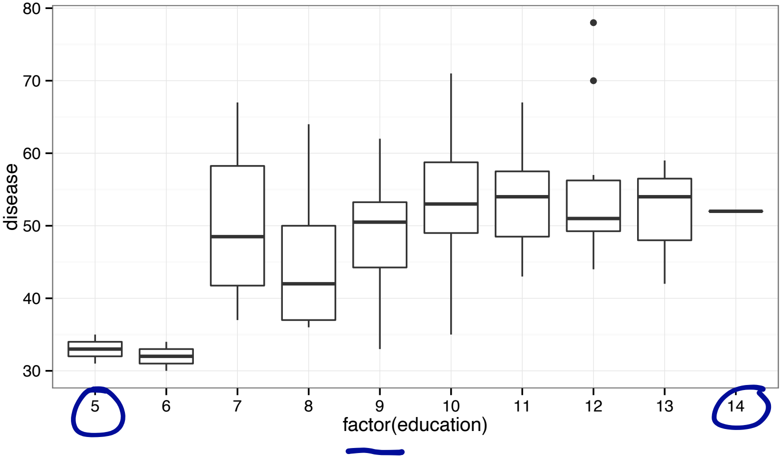
$\beta_1 =$ E(y) when x is in group 1

$\beta_2 =$

$$
\begin{aligned}
E(y|x=1) &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
&= \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 0 \\
&= \beta_1
\end{aligned}
$$

Categorical predictor example: lung data

```
qplot(factor(education), disease, geom="boxplot", data=dat)
```



Categorical predictor example: lung data

$$dis_i = \beta_0 + \beta_1 educ_{6,i} + \beta_2 educ_{7,i} + \cdots + \beta_{\cancel{14}\,9} educ_{14,i}$$

```
mlr7 <- lm(disease ~ factor(education), data=dat)
summary(mlr7)$coef

##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)             33.00      4.913  6.7173 1.689e-09
## factor(education)6      -1.00      7.768 -0.1287 8.979e-01
## factor(education)7      17.33      6.017  2.8808 4.969e-03
## factor(education)8      11.18      5.329  2.0975 3.879e-02
## factor(education)9      15.50      5.353  2.8953 4.765e-03
## factor(education)10     20.38      5.188  3.9289 1.683e-04
## factor(education)11     20.53      5.382  3.8155 2.505e-04
## factor(education)12     22.20      5.601  3.9633 1.489e-04
## factor(education)13     18.67      6.948  2.6868 8.609e-03
## factor(education)14     19.00      9.825  1.9338 5.632e-02
```

expected value of disease for 5^{th} graders is 33.0.

expected value of disease for 10^{th} graders = 33.0 + 20.38

Categorical predictor releveling

$$dis_i = \beta_0 + \beta_1 educ_{5,i} + \beta_2 educ_{6,i} + \beta_1 educ_{7,i} + \beta_2 educ_{9,i} + \cdots + \beta_{1\cancel{4}} educ_{14,i}$$

9

```
dat$educ_new <- relevel(factor(dat$education), ref="8")
mlr8 <- lm(disease ~ educ_new, data=dat)
summary(mlr8)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   44.176      2.064 21.4059 7.303e-37
## educ_new5    -11.176      5.329 -2.0975 3.879e-02
## educ_new6    -12.176      6.361 -1.9143 5.880e-02
## educ_new7      6.157      4.041  1.5238 1.311e-01
## educ_new9      4.324      2.964  1.4588 1.482e-01
## educ_new10     9.208      2.654  3.4695 8.059e-04
## educ_new11     9.357      3.014  3.1042 2.559e-03
## educ_new12    11.024      3.391  3.2507 1.626e-03
## educ_new13     7.490      5.329  1.4057 1.633e-01
## educ_new14     7.824      8.756  0.8935 3.740e-01
```

$= 33$

$E(y|educ = 5) = 44.176 + -11.176$

Categorical predictor: no baseline group

$$dis_i = \beta_1 educ_{5,i} + \beta_2 educ_{6,i} + \cdots + \beta_{\cancel{14}\,10} educ_{14,i}$$

```
mlr9 <- lm(disease ~ factor(education) - 1, data=dat)
summary(mlr9)$coef

##                      Estimate Std. Error t value  Pr(>|t|)
## factor(education)5      33.00      4.913   6.717 1.689e-09
## factor(education)6      32.00      6.017   5.318 7.716e-07
## factor(education)7      50.33      3.474  14.489 3.846e-25
## factor(education)8      44.18      2.064  21.406 7.303e-37
## factor(education)9      48.50      2.127  22.799 6.282e-39
## factor(education)10     53.38      1.669  31.991 1.359e-50
## factor(education)11     53.53      2.197  24.366 3.801e-41
## factor(education)12     55.20      2.691  20.514 1.713e-35
## factor(education)13     51.67      4.913  10.517 2.758e-17
## factor(education)14     52.00      8.509   6.111 2.561e-08
```

Creating categories using cut()

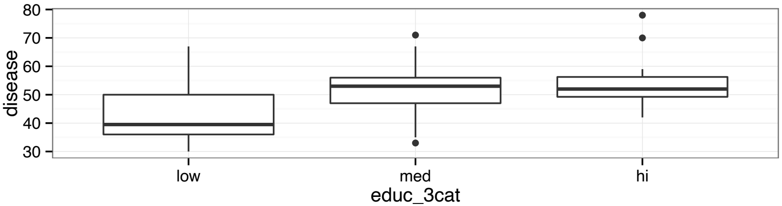
$$dis_i = \beta_1 educ_{low,i} + \beta_2 educ_{med,i} + \cdots + \beta_{14} educ_{hi,i}$$

```r
dat$educ_3cat <- cut(dat$education, breaks=3,
                     labels=c("low", "med", "hi"))
mlr10 <- lm(disease ~ educ_3cat - 1, data=dat)
coef(mlr10)
```

Factor

```
## educ_3catlow educ_3catmed  educ_3cathi
##        43.43        52.05        54.21
```

```r
qplot(educ_3cat, disease, geom="boxplot", data=dat)
```



Today's big ideas

- Multiple linear regression: categorical variables