Categorical Variables - Models

Recall: multiple linear regression model yields plane that minimizes residual sum of squares.

- Simplest: one predictor that is categorical with two levels, ex. Student (T, F)

  - Create a dummy variable: X_i can be 0/1 variables; vars don't need to be normally distributed

$$S_{yes,i} \equiv X_i = \begin{cases} 1 & \text{if person } i \text{ is a student} \\ 0 & \text{otherwise, i.e. person } i \text{ is not a student.} \end{cases}$$

s.t the linear model

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

$$= \begin{cases} \beta_1 + \beta_2 + \varepsilon_i & \text{if person } i \text{ is a student} \\ \beta_1 + \varepsilon_i & \text{if person } i \text{ is not a student} \end{cases}$$

- Interpretation: β_1 is the average balance for non-students
  $(\beta_1 + \beta_2)$ is the average balance for students.

  ie. If you compare i and j, where i is a student and j is not a student, then $E(y_i) - E(y_j) = \beta_2$.

The general model for categorical variable with two levels:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \gamma s_{yes} + \varepsilon_i$$

so

$$y_i = \begin{cases} (\beta_1 + \gamma) + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i & \text{if } i \text{ is student} \\ \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i & \text{if } i \text{ is not a student} \end{cases}$$

- Interpretation:
  - There are two fitted parallel regression planes.
  - The intercepts of the planes are different and are given by β_1 and $\beta_1 + \gamma$, respectively.
  - Comparing two observations i and j, where i is a student and j is not, and the values of all other variables are identical, then $E(y_i) - E(y_j) = \gamma$.

* Note that it does not matter which category is used as baseline, as long as you keep track of what you're doing.

"γ is the effect of student status given all other variables are held constant."

↳ be careful with wording so you don't imply (unproven/unsupported) causality.

See R code:

```
# data(DATASET_VAR)                                   (Credit)
pairs(DATA_VAR[, c(range, #)])                        (Credit[c(2:7, 12)])
attach(DATASET_VAR)                                   (Credit)
fit.cat <- lm(Y ~ Xi + Syes, data = DATA_SET).
fit.student <- lm(Balance ~ Income + Student, data = Credit)
```

* it is difficult to interpret the whole model; be careful about interpretation of γ.

Computational Statistics
Lecture 3: Categorical Variables

02 March 2018

- Different ways to code dummy variables:

(R default) $x_i = \begin{cases} 1 & \text{if } i \text{ is student} \\ 0 & \text{otherwise} \end{cases}$ ⇝ assign different colors.

III

$x_i = \begin{cases} 1 & \text{if } i \text{ is student} \\ -1 & \text{otherwise} \end{cases}$.

- In the model, one can expect that prediction to be off by the value of the residual standard error (distance btwn the lines)

Categorical variables with more than two levels.
ex. ethnicity: levels A, B, C

How to code? Idea 1: $x_i = \begin{cases} 0 & \text{if } i \in A \\ 1 & \text{if } i \in B \\ 2 & \text{if } i \in C \end{cases}$. But then

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

$$= \begin{cases} \beta_1 + \varepsilon_i & \text{if } i \in A \\ \beta_1 + \beta_2 + \varepsilon_i & \text{if } i \in B \\ \beta_1 + 2\beta_2 + \varepsilon_i & \text{if } i \in C \end{cases}$$

* No justification for having the difference between B and C to be the same as A and B, especially if no ordering (ie have a particular value).

Idea 2 For k categories, use k-1 dummy variables (usually); baseline is arbitrarily assigned.
ex. for the above example,

$$x_B, x_C = \begin{cases} 0,0 & \text{if } i \in A \\ 1,0 & \text{if } i \in B \\ 0,1 & \text{if } i \in C \end{cases}$$

Then, in the model, $Y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \gamma_B x_{iB} + \gamma_C x_{iC} + \varepsilon_i$

R: If regression lines ∀ ethnicity (A, B, C) are overlapping, then there is not much effect of ethnicity given income in the model.

- If, however, the model shows an effect / distinguishes between, for example, Asian and non-Asian, we don't need to compare Asian and Caucasian AND Asian and AfAm (similar effect; redundant)

- We can use ANOVA testing to see if ethnicity, in general, is useful.
  Test if it helps to add that variable (ethnicity) at all before assessing differences btwn levels

- Distinguishing category vs. other predictors: different levels may yield/have different relationships with other variables

ie Interaction!

- Categorical Variables, cont'd: Interactions

"Interaction between Income and Student status means that:

- the "effect" of income depends on student status. s.t. the slopes of regression planes differ for student vs. non-student.
- the "effect" of student status depends on income s.t. the vertical distance btwn planes depends on income level.

ie Model:

$$Y_i = \beta_1 + \beta_2 \cdot \text{income}_i + \gamma \cdot x_i + \delta \cdot x_i \cdot \text{income}_i + \varepsilon_i ,$$

where $x_i = 1$ if student $\{ 0$ otherwise
and the new variable δ represents the interaction.

$$\text{ie } y_i = \begin{cases} \beta_1 + \beta_2 \cdot \text{income}_i + \varepsilon_i & \text{if } i \text{ is not student } (x_i = 0) \\ \beta_1 + \beta_2 \cdot \text{income}_i + \gamma + \delta \cdot \text{income}_i & \text{if } i \text{ is student } (x_i = 1) \end{cases}$$

$$\hookrightarrow \beta_1 + (\beta_2 + \gamma) \cdot \text{income}_i + \gamma + \varepsilon_i$$

↳ different slope for student (vs. non-student) (interaction btwn income and student status; ie. the effect of income (on balance) is different btwn/ for students and non-students.)

SECTION[2] for Series 1.

- Regression is obviously generalizable

- Tukey-Anscomb plot (ref. ISLR, p. 62)

If we have a true regression line (red) and the model regression line (blue) is good, then the two lines are similar (p.64) (left fig)
(for right fig, the lines are regression lines of the different simulations / datasets)

Recall To reasonably fit a linear model with least squares, we must assume: (residuals r are estimates of the unknown errors ε)

* ① $E[\varepsilon_i] = 0 \ \forall i$ (the linear regression equation is correct) - unbiased.
$\Rightarrow E[\hat{\beta}] = \beta$ ie $\hat{\beta}$ is unbiased. (for finite sample).
$\Rightarrow E[\hat{Y}] = E[Y] = X\beta$, $E[r] = 0$

② Homoscedasticity (Variance of error is constant over $\forall i$):

$$\text{Var}(\varepsilon_i) = \sigma^2 \ \forall i \equiv \text{Cov}(\varepsilon) = \sigma^2 I_{n \times n} \quad (\Leftarrow E[\varepsilon\varepsilon^T])$$ ie error is uncorrelated.

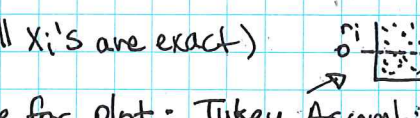
③ Errors are jointly normally distributed

$$\Rightarrow \text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \quad (\text{std. err})$$

$$\Rightarrow \text{Cov}(\hat{Y}) = \sigma^2 P, \quad \text{Cov}(r) = \sigma^2 (I - P)$$

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \Rightarrow \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$ | However residuals are not uncorrelated; ie $\text{Var}(r_i) = \sigma^2 (I - P_{ii})$ | and variance is const.

(And all x_i's are exact)



See R code for plot: Tukey Ascomb: plot residuals (r_i) vs. (to) fitted line ($\hat{Y}_i$). if distribution is not (random) around 0 s.t. sample $\text{corr}(r_i, \hat{Y}_i) \neq 0$, eg varied std. deviation, variance depending on $\hat{Y}_i$, then maybe should use another model, eg. quadratic or cubic (rather than linear).

QQ plot: . std. residuals vs. theoretical quantiles. if good QQ. $\Rightarrow E(\varepsilon) = 0$ is satisfied
- checks if dimensions of the model is correct