Inference 2: Fitting models to data (cont)

(Updated with interaction terms).
Last updated: November 2, 2017

Numerical vs categorical: the indicator trick

Key: linear model still makes sense for categorical inputs.

$$Y = \beta_0 + \beta_1 X,$$

X takes values in $0, 1$. **What is the meaning of β_1?**

.

Numerical vs categorical: the indicator trick

Key: linear model still makes sense for categorical inputs.

$$Y = \beta_0 + \beta_1 X,$$

X takes values in $0, 1$. **What is the meaning of β_1**?
how much Y increases if $X = 1$ over $X = 0$.

Numerical vs categorical: the indicator trick

Key: linear model still makes sense for categorical inputs.

$$Y = \beta_0 + \beta_1 X,$$

X takes values in $0, 1$. **What is the meaning of β_1?**
how much Y increases if $X = 1$ over $X = 0$.

Give interpretations to β_0 in the following examples

- teacher: salary vs degree (BA $=$ 0, MA $=$ 1).
- hsb2: math score vs school type (public $=$ 0, private $=$ 1)

Numerical vs categorical: the indicator trick

Key: linear model still makes sense for categorical inputs.

$$Y = \beta_0 + \beta_1 X,$$

X takes values in $0, 1$. **What is the meaning of β_1?**
how much Y increases if $X = 1$ over $X = 0$.

Give interpretations to β_0 in the following examples

- teacher: salary vs degree (BA $=$ 0, MA $=$ 1).
- hsb2: math score vs school type (public $=$ 0, private $=$ 1)

t-test: is there a significant difference? (yes/no)
linear model: how much is the difference?

Numerical vs categorical: more than one categories

X takes values in $0, 1, \ldots, k$. **What to do?**

Numerical vs categorical: more than one categories

X takes values in $0, 1, \ldots, k$. **What to do?**
Recode into k indicator variables! Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

What is the meaning of $\beta_1, \ldots, \beta_k$?

Numerical vs categorical: more than one categories

X takes values in $0, 1, \ldots, k$. **What to do?**
Recode into k indicator variables! Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

What is the meaning of $\beta_1, \ldots, \beta_k$?
β_j = how much Y increases if $X = j$ over $X = 0$

Numerical vs categorical: more than one categories

X takes values in $0, 1, \ldots, k$. **What to do?**
Recode into k indicator variables! Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

What is the meaning of $\beta_1, \ldots, \beta_k$?
β_j = how much Y increases if $X = j$ over $X = 0$

Give interpretations to β_j's in the following examples

- hsb2: math score vs socioeconomic background (low = 0, middle = 1, high = 2)

Numerical vs categorical: interaction term

Let X, Z be two categorical variables with k_1 and k_2 categories. The interaction term $X * Z$ (or $X : Z$) is another categorical variable with $k_1 k_2$ many categories, obtained as all possible combinations (interactions) of X and Z.

Example:

- hsb2: $X =$ socioeconomic (low, middle, high); $Z =$ school type (public, private). **What are the categories of $X : Z$?**

Numerical vs categorical: interaction term

Let X, Z be two categorical variables with k_1 and k_2 categories. The interaction term $X * Z$ (or $X : Z$) is another categorical variable with $k_1 k_2$ many categories, obtained as all possible combinations (interactions) of X and Z.
Example:

- hsb2: $X =$ socioeconomic (low, middle, high); $Z =$ school type (public, private).
- $X : Z =$ interaction of socioeconomic and school type: 6 categories (low ses and private school; low ses and public school; middle and private; middle and public; high and private; high and public)

Why interaction term?

What are the meanings of the coefficients in the following three models?

Model 1:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} + \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

Model 2:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} : \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

Model 3:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} + \texttt{ses} : \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

.

Why interaction term?

What are the meanings of the coefficients in the following three models?

Model 1:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} + \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

Model 2:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} : \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

Model 3:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} + \texttt{ses} : \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

What is the predicted mean math score for each of the 6 possible combinations of socioeconomic and school types in each model?

.

Why interaction term?

What are the meanings of the coefficients in the following three models?

Model 1:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} + \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

Model 2:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} : \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

Model 3:

$$\texttt{lm}(\texttt{math} \sim \texttt{ses} + \texttt{ses} : \texttt{schtype}, \texttt{data} = \texttt{hsb2})$$

What is the predicted mean math score for each of the 6 possible combinations of socioeconomic and school types in each model?

For answers see lecture R codes, November 2nd.

Numerical vs categorical: diagnostics

Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$.
Assumptions on the ϵ_i's:

- Independence
- Normally distributed
- Constant variance

Examples: hsb2: math vs ses; teacher: salary vs degree; hsb2: math vs schooltype.

F-test and t-test

Suppose X has two categories.
Model:

$$Y = \beta_0 + \beta_1 X.$$

Hypothesis test: are all the group means equal?

- H0: group means are equal ($\beta_1 = 0$)
- HA: group means are not equal ($\beta_1 \neq 0$)
- Test: two-sided t-test .

F-test and t-test

Suppose X has ~~two~~ $k+1$ categories.
Model:

$$\cancel{Y = \beta_0 + \beta_1 X} \quad Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Hypothesis test: are all the group means equal?

- H0: group means are equal ($\cancel{\beta_1 = 0}$ $\beta_1 = \cdots = \beta_k = 0$)
- HA: group means are not equal ($\cancel{\beta_1 \neq 0}$ one of the $\beta_1, \ldots, \beta_k$ is non-zero)
- Test: ~~two-sided t-test~~ F-test.

Multiple regression

Multiple regression = more than one input variables.
Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

$X_1, \ldots, X_k$: input variables. Can be either numerical or indicator.
Examples

- We saw: one categorical with $k+1$ categories recoded as k binaries
- teacher: salary vs degree + fulltime + years
- marioKart: totalPrice vs everything else