Regression with categorical predictors

"qualitative

Categorical variable: Its values are categories (or attributes) with no particular order, e.g., race, OS, etc. The values should **not** be coded as 1, 2, 3, .., unless R knows to treat the variable as a `factor`.

indicator variable.

Dummy variable: A binary variable z with value 0 or 1

Key idea: Represent a categorical variable with C categories using $C-1$ dummy variables, $z_1, \ldots, z_{C-1}$. The model is

$$\boxed{E(Y|\mathbf{z}) = \beta_0 + \gamma_1 z_1 + \ldots + \gamma_{C-1} z_{C-1}.}$$

usual regression

Base (or reference) category: $z_1 = \ldots = z_{C-1} = 0$.

X

Ex 1: OS with two categories — Windows and Mac.

$C=2$ Need one dummy — z_1

$z_1 \Rightarrow$ indicator of windows OS

$E[Y|OS] = \beta_0 + \gamma_1 z_1$

		z_1
	windows	1
"baseline" $\rightarrow$	mac	0

$E[Y|OS = win] = \beta_0 + \gamma_1(1) = \beta_0 + \gamma_1$

$E[Y|OS = mac] = \beta_0 + \gamma_1(0) = \beta_0$ = mean response for base category.

X

Ex 2: Race with three categories — White, Black and other.

$C = 3 \Rightarrow$ Need two dummies – Z_1, Z_2

$Z_1 \rightarrow$ indicator of white

$Z_2 \longrightarrow$ indicator of black

	Z_1	Z_2
White	1	0
Black	0	1
baseline $\rightarrow$ Other	0	0

model:

$E[Y \mid \text{race}] = \beta_0 + \gamma_1 Z_1 + \gamma_2 Z_2$

$E[Y \mid \text{race} = \text{other}] = \beta_0 + \gamma_1(0) + \gamma_2(0) = \beta_0$

mean response for baseline

- In general, $\beta_0 =$ mean for base category, and $\cancel{\beta_j}\ \gamma_j =$ difference in means for category j and base category
- The regression model may have both numerical as well as categorical predictors.
- The model may have several categorical predictors.
- To test whether a categorical variable is significant, simultaneously test **all** corresponding slopes. In other words, the hypotheses are $H_0 : \gamma_1 = \ldots = \gamma_{C-1} = 0$, vs. H_1 : at least one non-zero slope, and they should be tested using an F-test with $C - 1$ numerator d.f.

OS Example

$$\gamma_1 = (\beta_0 + \gamma_1) - \beta_0$$
$$= E[Y/\text{win}] - E[Y/\text{mac}]$$
= change in mean response over the base category.

Race Example

$$E[Y|\text{race}=\text{white}] = \beta_0 + \gamma_1(1) + \gamma_2(0) = \beta_0 + \gamma_1$$
$$\Rightarrow \gamma_1 = E[Y|\text{race}=\text{white}] - E[Y|\text{race}=\text{other}]$$
= change in mean response over baseline

$$E[Y/\text{race}=\text{Black}] = \beta_0 + \gamma_2$$
$$\Rightarrow \gamma_2 = E[Y|\text{race}=\text{black}] - E[Y/\text{race}=\text{other}]$$
= change in mean response over baseline.

Example: Jane data.

```
# Read the Jane data

jane <- read.table("jane.csv", sep=",", header=T)

> str(jane)
'data.frame': 150 obs. of  3 variables:
 $ x     : int  1 1 1 2 2 2 3 3 3 4 ...
 $ color: Factor w/ 3 levels "blue","green",..: 3 1 2 3 1 :
 $ y     : num  24.9 12.3 16.6 25.2 12.1 ...
>
```

quantitative pred.

qualitative predictor

response

If 'color' is not already a factor,
jane$color ← as.factor(jane$color)

```
attach(jane)

> table(color)
color
 blue green   red
```

```
   50     50     50
>

# Include both x and color as predictors
```

make sure that 'color' is coded as factor.

```
fit1 <- lm(y~x+color)

# Note: color is already a factor variable. If this is
# numeric, then we need to write:
# fit1 <- lm(y~ x + factor(color))

> summary(fit1)

Call:
lm(formula = y ~ x + color)

Residuals:
     Min       1Q   Median       3Q      Max
```

Model: $E[Y|x, \text{color}] = \beta_0 + \beta_1 x + \gamma_1 z_1 + \gamma_2 z_2$,

-14.2398 -2.9939 0.1725 3.5555 11.9747

indicator for green z_1

indicator for red

(keep all the other predictor)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	13.16989	1.01710	12.948	< 2e-16	***
x	1.00344	0.02848	35.227	< 2e-16	***
colorgreen	2.12586	1.00688	2.111	0.0364	*
colorred	6.60586	1.00688	6.561	8.7e-10	***

$\hat\beta_0$, $\hat\beta_1$, $\hat\gamma_1$, $\hat\gamma_2$

z_1, z_2

(⟹ blue = base)

$H_0: \beta_1 = 0$

$H_0: \gamma_1 = 0$

$H_0: \gamma_2 = 0$

Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1

$\hat\sigma$

Residual standard error: 5.034 on 146 degrees of freedom

Multiple R-squared: 0.898, Adjusted R-squared: 0.8959

F-statistic: 428.6 on 3 and 146 DF, p-value: < 2.2e-16

$\underbrace{\text{p-value: < 2.2e-16}}_{}$ ⟹ At least one of x and color is useful.

H_0: both x and color are not needed

>

Test of model significance.

Is color significant?

```
fit2 <- lm(y~x)
```

'reduced' model without color

fit1 — full model

```
> anova(fit2, fit1)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + color
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    148 4837.5
2    146 3700.4  2    1137.1 22.433 3.197e-09 ***
---
Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1
>
```

F-statistic

$H_0: \gamma_1 = \gamma_2 = 0$ vs. H_1: At least γ_i is not zero.

Reject $H_0 \Rightarrow$ color is useful.

Q: What is the predicted response for a subject with color=blue and x=2?

Fitted model:

$$E[Y|x, \text{color}] = 13.17 + (1.00)x + 2.13 z_1 + 6.61 z_2$$

z_1: "green" z_2: "red"

$$\Rightarrow \hat{Y} = 13.17 + (1.00)(2) + 2.13(0) + 6.61(0)$$

$$= 15.17.$$

Prediction at color = green and $x = 2$:

$$\hat{Y} = 13.17 + (1.00)(2) + (2.13)(1) + 6.61(0)$$

$$= \; ?$$

Note: $\hat{\beta}_0 = 13.17 = E[Y \mid x = 0, \text{color} = \text{blue}]$.

Model building

- Need a model that provides a good fit and accurate prediction, without over fitting.

- Adjusted R^2
  - Compare models, choose the one with highest adjusted R^2
  - Best subset selection
    - suppose there are K predictors to be considered.

      $\Rightarrow$ A total of 2^K possible models.
    - Can use adjusted R^2
    - Impractical when K is large.

- Stepwise selection or Adjusted R^2
  - Forward (partial F-test, p. 392)
  - Backward (" " , p. 393) elimination
  - Both

- Stepwise selection with AIC or Adjusted R^2 to find the model of the best size

AIC: Akaike Information criterion

$$AIC = 2k - 2\ln(L)$$

$2k$: # parameters (= # regression coeff)

$\ln(L)$: log likelihood fn. at MLE (i.e., max. loglikelihood)

- "Smaller is better"
- As # parameters ↑, $\ln(L)$ ↑, but AIC may ↑ or ↓
- Penalized criterion. [First term = penalty for complexity, second term = goodness of fit].

- stepwise selection with AIC or with adjusted R^2.
  - use ~~stepAIC~~ step function in R

- See the R handout about model building for implementation of these ideas.