

Lec 04: Indicator variables and Splines

MATH 456 - Spring 2016

Navbar: [\[Home\]](#) [\[Schedule\]](#) [\[Data\]](#) [\[Week 5 Overview\]](#) [\[HW Rubric\]](#) [\[Google Group\]](#) [\[Google Drive\]](#)

Assigned Reading

Affi: Chapter 9.3, Harrel Ch 2

Factor variable coding

- Better used term: Indicator variable
- Math notation: **I(gender == “Female”)**.
- A.k.a reference coding
- For a nominal X with K categories, define K indicator variables.
 - Choose a reference (referent) category:
 - Leave it out
 - Use remaining K-1 in the regression.
 - Often, the largest category is chosen as the reference category.

Example: Binary indicator for gender

Consider the linear model of FEV on gender(x_1), height(x_2) and age(x_3) where gender interacts with both age and height. In other words, gender changes the relationship between height and FEV1, and the relationship between age and FEV1.

$$FEV1 \sim \beta_0 + \beta_1 * gender + \beta_2 * height + \beta_3 * age + \beta_4 * gender * height + \beta_5 * gender * age$$

If we let gender = 0 if the record is on a male, and gender = 1 if the record is on a female, then the model for males would be:

$$FEV1 \sim \beta_0 + \beta_2 * height + \beta_3 * age$$

and the model for females would be:

$$FEV1 \sim (\beta_0 + \beta_1) + (\beta_2 + \beta_4) * height + (\beta_3 + \beta_5) * age$$

Example: Religion against income and depression

Consider a log-linear model for the effect of marital status (X_2) on log income while controlling for age(X_1). This is called a log-linear model because the outcome has been log transformed.

$$\log(Y_i) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$$

```
dep <- read.table("C:/GitHub/MATH456/data/Depress_020916.txt", sep="\t", header=TRUE)
names(dep) <- tolower(names(dep)) # I hate all capital variable names
levels(dep$marital)
```

```
## [1] "Divorced"      "Married"      "Never Married" "Separated"
## [5] "Widowed"
```

Marital status has 5 levels, so we would need 4 indicator variables. R always uses the first level of a factor variable as the reference level.

- Let $x_2 = 1$ when `marital='Married'`, and 0 otherwise,
- let $x_3 = 1$ when `marital='Never Married'`, and 0 otherwise,
- let $x_4 = 1$ when `marital='Separated'`, and 0 otherwise,
- let $x_5 = 1$ when `marital='Widowed'`, and 0 otherwise.

The mathematical model would look like:

$$\log(Y)|X \sim \beta_0 + \beta_1 * x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

Two levels of interpretation here.

1. The outcome is log transformed, so the interpretation has to be back-transformed.
2. The coefficients for the other levels of the categorical variable are in *comparison* to the reference level.

Interpretation of log-linear models Calculate the change in Y that corresponds to a one unit change in x_1 . Since marital status is remaining constant, I will exclude it from the calculations below to save space and not to detract from the main point.

Write each equation down

$$\begin{aligned} \log(Y)|x_1 &= \beta_0 + \beta_1 x_1 \\ \log(Y)|(x_1 + 1) &= \beta_0 + \beta_1 (x_1 + 1) \end{aligned}$$

Find the difference

$$(\log(Y)|x_1) - (\log(Y)|(x_1 + 1)) = (\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 (x_1 + 1))$$

and simplify.

$$\begin{aligned} \log\left(\frac{Y|x_1}{Y|x_1 + 1}\right) &= \beta_1 \\ \frac{Y|x_1}{Y|x_1 + 1} &= e^{\beta_1} \end{aligned}$$

Each 1-unit increase in x_j multiplies the expected value of Y by $e^{\hat{\beta}_j}$.

Interpretation: $100\hat{\beta}_j$ is the expected **percentage** change in Y for a unit increase in x_j .

The nice thing about factor variables in R, is that the appropriate indicator variables are automatically created for you by the linear model (`lm()`) function.

```
summary(lm(log(income) ~ age + marital,data=dep))
```

```
##
## Call:
## lm(formula = log(income) ~ age + marital, data = dep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62643 -0.46829  0.01535  0.45280  1.48175
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.065696   0.166760  18.384 < 2e-16 ***
## age           -0.009043   0.002919  -3.098  0.002143 **
## maritalMarried    0.416653   0.124698   3.341  0.000944 ***
## maritalNever Married -0.183156   0.141354  -1.296  0.196109
## maritalSeparated  -0.394544   0.223431  -1.766  0.078482 .
## maritalWidowed    -0.278352   0.173159  -1.607  0.109042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7059 on 288 degrees of freedom
## Multiple R-squared:  0.1971, Adjusted R-squared:  0.1832
## F-statistic: 14.14 on 5 and 288 DF,  p-value: 2.236e-12
```

- For every year older, a persons income decreases by 1%. ($\exp(-0.009) = 0.99$)
- Married individuals have a 52% higher income compared to those who are divorced. ($\exp(-0.417) = 1.52$)
- Those who have never been married have 16% lower income compared to those who are divorced. ($\exp(-0.183) = 0.83$)
- Separated individuals have 32% lower income compared to those who are divorced. ($\exp(-0.394) = 0.67$)
- Widowed individuals have 24% lower income compared to those who are divorced. ($\exp(-0.278) = 0.76$)

Other references on how to interpret regression parameters when they have been log transformed:

- http://www.ats.ucla.edu/stat/mult_pkg/faq/general/log_transformed_regression.htm
- <http://www.kenbenoit.net/courses/ME104/logmodels2.pdf>

On Your Own

On Your Own Create a model to analyze the relationship of education status to depression level as measured by CESD after controlling for age. Combine all education levels below a HS graduate into one reference category called “Up to HS” prior to analysis.

This is a seemingly simple request, but there are a lot of steps you must do to correctly analyze this question.

1. Ensure that you are using the analyzable version of the depression data set. It may be helpful to confirm that your recodes are correct by comparing your data mangement code file to mine [dm_depress](#) located on our course website.
2. Reference your Ch3 homework (or the [solutions](#)) if you need help collapsing educational categories.
3. Ensure that R is treating “Up to HS” as the reference category for education level. If it is not, use the `levels` argument of the `factor()` function to reorder your factor levels. This is also presented in the Ch3 solutions.
4. Consider a transformation of `CESD`. Explain and justify using graphical measures why you chose to, or chose not to, transform `CESD` prior to modeling.
5. Check the model fit by examining the residuals to see if the assumption that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is upheld.
6. Identify any potential outliers. Explain why you think they are outliers. Examine their standardized residuals and leverage values. If any seem to stand out or have high values for either measure, exclude them from the analysis and re-run the model.
7. Once you have finalized your model, interpret ALL coefficients in context of the problem. State if any are significantly predictive of the outcome, provide p-values in your conclusion.
8. Does this model do well at all in predicting `CESD`? Answer this question using both the coefficient of determination and the ANOVA test of overall global fit (testing that all β 's are 0)

Splines

References

- Afifi Section 9.4
- Harrell 2.4.3, pg 39 <http://biostat.mc.vanderbilt.edu/tmp/course.pdf>
- https://www.youtube.com/watch?v=o_d4hmKhmsQ
- <http://www.r-bloggers.com/thats-smooth/>