EPI 809: Multiple Linear Regression

OUTLINE

- Review of simple linear regression with a quantitative and a binary predictor.

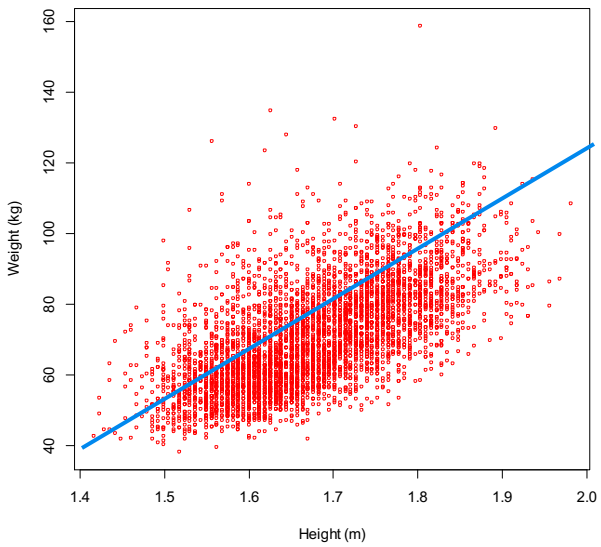
- Linear models for factors with multiple levels
    - Types of contrasts and parameter interpretation

- Multiple Linear regression with categorical and continuous predictors
    - Interpretation of parameters associated to dummy variables and to quantitative predictors

- OLS estimation in multiple linear regression

- Analysis of variance in multiple linear regression
    - Degrees of freedom
    - Variance decomposition
    - Multiple R-sq.
    - F-test
    - Partition of the model variance into sub-components
    - Type-I error control and sequential testing

Linear Regression When X is Quantitative



Model: $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$

Response

Intercept

Predictor

Regression Coef.

Residual

Interpretation:

=> β_1 : Linear rate of change of y with respect to X

=> $\beta_0 + x_i\beta_1$: Prediction equation (linear approx. to the conditional expectation)

What if X is Categorical, How do We Handle This?

Case 1: Categorical Variable With 2 Levels (e.g., Male/Female or South/North)

- 1st We recode X into a Dummy Variable e.g., M_i={ 1 if X_i=Male ; 0 if X_i=Female }

- Then we fit a linear model using M_i as predictor: $y_i = \beta_0 + M_i\beta_1 + \varepsilon_i$

- How do we interpret regression coefficients here?

- We look at the prediction equations both for male and female.

$$E[y_i \mid M_i = 0] = \mu_{y|Female} = \beta_0 \qquad E[y_i \mid M_i = 1] = \mu_{y|Male} = \beta_0 + \beta_1$$

$$\mu_{y|Male} - \mu_{y|Female} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

- Therefore, the regression coefficient represent the difference between the expected value of Male and Female.

- NOTE: The interpretation of the regression coefficient depends on what level was used as reference category.

What if Our Predictor Has More than Two Levels (e.g., Ethnicity: Black/White/Hispanic)?

What if our categorical variable has more than two levels?

Suppose our categorical variable can get on 3 possible values (White, Black, Hispanic). Then one possibility is to create 2 dummy variables:

	W	B
White	1	0
Black	0	1
Hispanic	0	0

$$y_i = \beta_0 + W_{1i}\beta_1 + B_{2i}\beta_2 + \varepsilon_i$$

$$\mu_H = \beta_0 \qquad \mu_W = \beta_0 + \beta_1$$

$$\mu_B = \beta_0 + \beta_2$$

How are these regression coef. Interpreted?

Example in R

https://github.com/gdlc/EPI809/blob/master/MULTIPLE_LINEAR_REGRESSION.md

Multiple Linear Regression Model

$\Rightarrow$ So far we have discussed cases where we have only one predictor (e.g., regression of weight on height or wage on sex).

$\Rightarrow$ However, in most cases, more than one predictor affect the response.

$\Rightarrow$ For example, one may think that weight is not only affected by height but also by sex, age, etc.

$\Rightarrow$ A multiple regression model is a regression model that includes more than one predictor.

$\Rightarrow$ For example, let's consider regressing wages on education and ethnicity (W/B/H).

Model Definition & Interpretation

Let:

- ED= Years of Education
- W={ 1 if White; 0 0therwise } B={ 1 if Black; 0 0therwise }

$$y_i = \beta_0 + \beta_1 ED_i + W_i \beta_2 + B_i \beta_3 + \varepsilon_i$$

Interpretation:

$$Conitnous \text{ predictors}: \frac{\partial E[y_i | ED_i, S_i]}{\partial ED_i} = \beta_1$$

$$Dummy \text{ variables}:$$

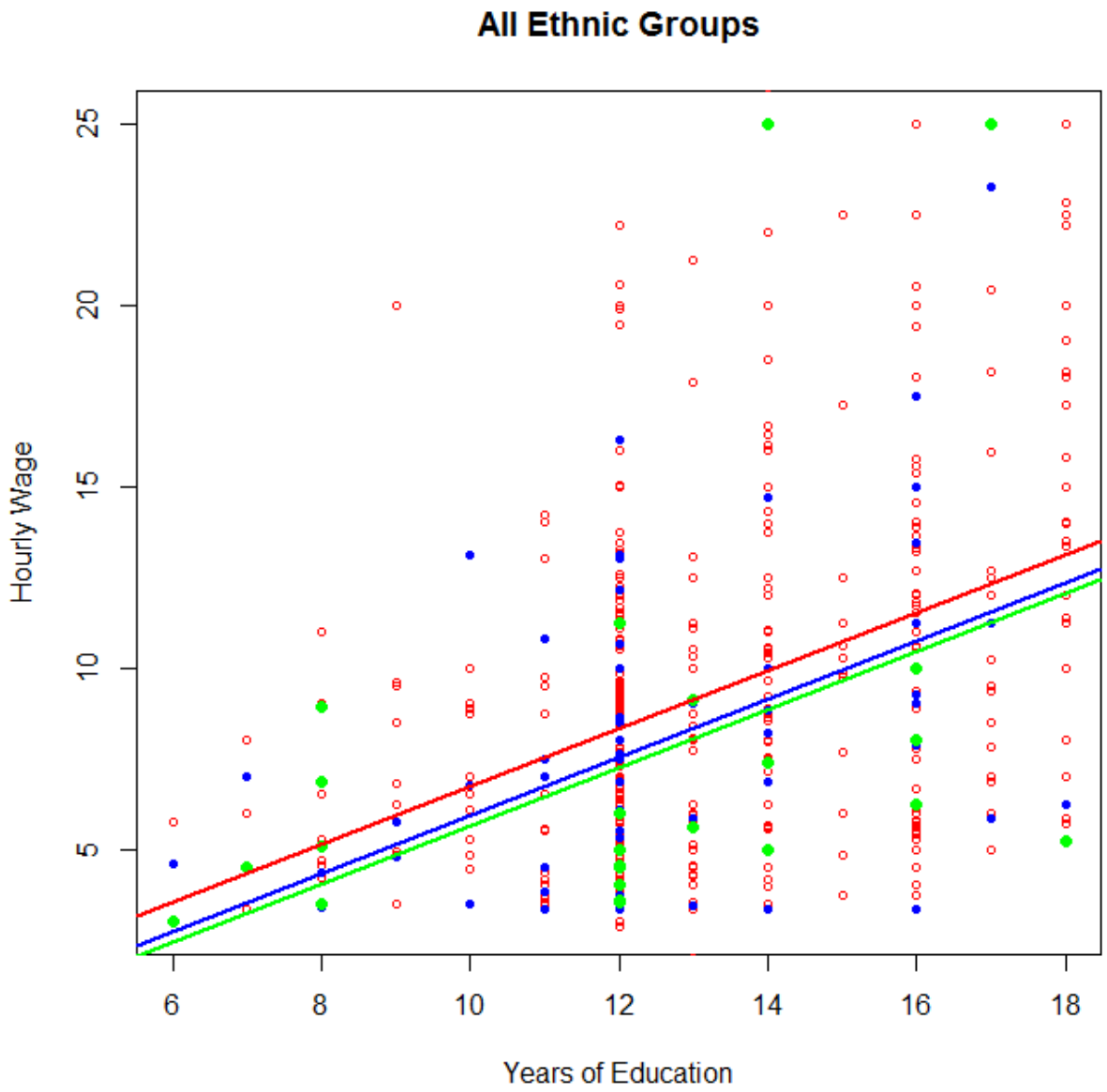
$$E[y_i | White, ED_i] - E[y_i | Hisp, ED_i] = \beta_2$$

$$E[y_i | Black, ED_i] - E[y_i | Hisp, ED_i] = \beta_3$$

Therefore:

- For continuous predictors the reg. coefficients represent the slope of the line.
- For dummy variables, reg. coefficients represent differences.

Graphical Representation



The Slope is the same for all groups.

The dummy variables induce group-specific intercepts.

Estimation (OLS)

$$y_i = \beta_0 + \beta_1 ED_i + \beta_2 W_i + \beta_3 B_i + \varepsilon_i \qquad \varepsilon_i = y_i - \beta_0 - \beta_1 ED_i - \beta_2 W_i - \beta_3 B_i$$

$$RSS(\beta_0, \beta_1, \beta_2, \beta_3) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 ED_i - \beta_2 W_i - \beta_3 B_i)^2$$

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = \underset{\text{argmin}}{} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 ED_i - \beta_2 W_i - \beta_3 B_i)^2$$

Stationary point: set all partial derivatives equal to zero.

Estimation (OLS)

$$\frac{\partial \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)^2}{\partial \beta_0}=-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)$$

$$\frac{\partial \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)^2}{\partial \beta_1}=-2\sum_{i=1}^{n}\left(-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)\right)ED_i$$

$$\frac{\partial \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)^2}{\partial \beta_2}=-2\sum_{i=1}^{n}\left(-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)\right)W_i$$

$$\frac{\partial \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)^2}{\partial \beta_3}=-2\sum_{i=1}^{n}\left(-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 ED_i-\beta_2 W_i-\beta_3 B_i\right)\right)B_i$$

- First Order Conditions: set all these derivatives equal to zero.

- This gives as many equation as coefficients.

- OLS estimates are then obtained by solving all these equations.

- JMP, SAS, R, etc. will give us OLS estiamtes.

Example
(Regression of Wages on Education and Ethnicity)

https://github.com/gdlc/EPI809/blob/master/MULTIPLE_LINEAR_REGRESSION.md

ANOVA in the Multiple Regression Model

ANOVA in the Multiple Regression Model

Model equation…

$$y_i = \beta_0 + \beta_1 ED_i + \beta_2 W_i + \beta_3 B_i + \varepsilon_i$$

Predictions: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 ED_i + \hat{\beta}_2 W_i + \hat{\beta}_3 B_i$

Residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i = \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 ED_i + \hat{\beta}_2 W_i + \hat{\beta}_3 B_i\right)$

Variance partition:

$$\text{Total}: \quad SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

$$Un - Explained: \quad \text{SSE} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

$$Explained\ (\text{regression}): \quad \text{SSR} = SST - SSE$$

$$R^2 = \frac{SSR}{SSE + SSR}$$

Summary

1. How do we include categorical predictors in a linear model

   - we code our predictor variables (e.g., ethnicity) using dummy variables and include these dummy variables into the model.
   - the regression coefficients associated to these dummy variables are interpretable as difference between the mean of the category that the dummy variable represent and the reference group.

2. Multiple Linear Regression Model

   (a) Def. A regression model that includes more than one predictor.

   (b) Parameters. The interpretation of regression coefficients depend on whether the associated covariate (predictor) is continuous or a dummy variable.

   (c) Estimation. The regression coefficients of the multiple linear regression can be jointly estimated using Ordinary Least Squares.

   (d) Model assessment. We can asses goodness of fit of the model based on the ANOVA table from where we can compute R-squared.