Stats 401 Lab 10

Ed Wu

11/9/2018

Outline

- Factors
- Double Subscript Notation
- Factors in the Linear Model
- Over-specified Models
- Prediction with Factors

Factors

- Recall: factors are explanatory variables with discrete levels
- Factors are also called categorical variables
- For example, sex could be a factor with two levels: male and female

Example

The iris data set was collected by Edgar Anderson. It contains measurements from 150 samples of irises (50 of each of three species: setosa, versicolor, and virginica). In this lab we will consider the petal length and petal width measurements.

```r
data(iris)
iris = iris[,3:5]
head(iris)
```

```
##   Petal.Length Petal.Width Species
## 1          1.4         0.2  setosa
## 2          1.4         0.2  setosa
## 3          1.3         0.2  setosa
## 4          1.5         0.2  setosa
## 5          1.4         0.2  setosa
## 6          1.7         0.4  setosa
```

Suppose we want to study whether petal length varies by species.

Double Subscript Notation

- Let y_{ij} represent the petal length of the j-th iris sample of species i, where $i = 1, 2, 3$ corresponds to the three species, and $j = 1, \ldots, 50$
- We have the following probability model for this experiment:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, 2, 3$ and $j = 1, \ldots, 50$, where $\epsilon_{ij} \sim \text{iid normal}(0, \sigma)$

Dummy Variables

- In order to convert our model from double subscript notation to the linear model, we need to use “dummy” (or “indicator”) variables
- A dummy variable for a factor level is equal to 1 if the observation equals that level, and 0 otherwise
- If we look at the iris data set, we can see the factor “Species” is 50 setosa, then 50 versicolor, then 50 virginica
- A dummy variable for versicolor would be the column vector of 50 0’s, then 50 1’s, then 50 0’s: $(\underbrace{0,\dots,0}_{\text{50 times}},\underbrace{1,\dots,1}_{\text{50 times}},\underbrace{0,\dots,0}_{\text{50 times}},)$

Lab Activity (Part 1)

Suppose we have 3 observations, and a factor variable for each observation's sex: ("Male","Female","Male")

1. What is the dummy variable for "Male"?
2. What is the dummy variable for "Female"?

Converting to a Linear Model

- Now we can write the model in the form $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Let $\mathbf{x}_1 = (\underbrace{1, \ldots, 1}_{\text{50 times}}, \underbrace{0, \ldots, 0}_{\text{100 times}})$ be the dummy variable corresponding to setosa
- Let $\mathbf{x}_2 = (\underbrace{0, \ldots, 0}_{\text{50 times}}, \underbrace{1, \ldots, 1}_{\text{50 times}}, \underbrace{0, \ldots, 0}_{\text{50 times}},)$ be the dummy variable corresponding to versicolor
- Let $\mathbf{x}_3 = (\underbrace{0, \ldots, 0}_{\text{100 times}}, \underbrace{1, \ldots, 1}_{\text{50 times}})$ be the dummy variable corresponding to virginica
- Let $\mathbf{y} = (y_1, \ldots, y_{150}) = (y_{1,1}, \ldots, y_{1,50}, y_{2,1}, \ldots, y_{2,50}, y_{3,1}, \ldots, y_{3,50})$ be the concatenated petal length measurements
- Let $\mathbf{e} = (e_1, \ldots, e_{150}) = (e_{1,1}, \ldots, e_{1,50}, e_{2,1}, \ldots, e_{2,50}, e_{3,1}, \ldots, e_{3,50})$ be the concatenated residuals

Linear Model

- One way to write the probability model is $Y_k = \mu_1 x_{k,1} + \mu_2 x_{k,2} + \mu_3 x_{k,3} + \epsilon_k$ for $k = 1, \ldots, 150$
- This is equivalent to $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbb{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix}$ and $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)$
- In this case, the interpretation of the parameters (μ_1, μ_2, μ_3) are the means for each factor level

Iris Example Continued

We can obtain the sample linear model in R:

```
lm1 = lm(Petal.Length ~ Species - 1, data = iris)
summary(lm1)$coefficients[,1:2]
```

```
##                   Estimate Std. Error
## Speciessetosa        1.462 0.06085848
## Speciesversicolor    4.260 0.06085848
## Speciesvirginica     5.552 0.06085848
```

We can see, for example, that the coefficient for “setosa” corresponds to the mean of the setosa samples:

```
mean(iris$Petal.Length[iris$Species == "setosa"])
```

```
## [1] 1.462
```

No Intercept vs. Intercept

- In the above model, we didn't include an intercept
- We could also write the model with an intercept: $Y_k = \mu + \alpha_2 x_{k,2} + \alpha_3 x_{k,3} + \epsilon_k$ for $k = 1, \ldots, 150$
- In this case, we would have the following interpretations
  - μ would be the mean petal length of setosa
  - α_2 would be the difference between the mean of setosa and the mean of versicolor
  - α_3 would be the difference between the mean of setosa and the mean of virginica

Iris Example Continued

We can fit the sample linear model (with an intercept) in R:

```r
lm2 = lm(Petal.Length ~ Species, data = iris)
summary(lm2)$coefficients[,1:2]
```

```
##                   Estimate Std. Error
## (Intercept)          1.462 0.06085848
## Speciesversicolor    2.798 0.08606689
## Speciesvirginica     4.090 0.08606689
```

Let’s check how these coefficients compare to our previous model

Over-specified Models

- In the model with the intercept, we had to drop one of the dummy variables
- Suppose we had written the model as: $Y_k = \mu + \alpha_3 x_{k,3} + \alpha_2 x_{k,2} + \alpha_3 x_{k,3} + \epsilon_k$ for $k = 1, \ldots, 150$
- Why does this model not work?

R Warnings

- By default, R uses the intercept version. If we wish to switch to the no intercept version, we need to specify that
- You may be working with R data in which factors are coded as characters instead. This can cause issues with your code so it is a good idea to convert these variables to factors prior to your analysis

Lab Activity (Part 2)

Suppose we are interested in studying the relationship between undergraduate major and salary. We collect a sample of size 7. We collect the salary in 1000s, as well as the major (engineering, computer science, or underwater basket weaving)

```
##   salary occupation
## 1    112        eng
## 2     90        eng
## 3     75         cs
## 4     90         cs
## 5     80        ubw
## 6    157        ubw
## 7     69        ubw
```

1. What is the probability model in double subscript form? Make sure to define all notation appropriately.
2. Suppose we wish to write out the sample linear model in the form $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$. What is the full $\mathbb{X}$ matrix?

Lab Activity (Part 3)

Returning to our iris example. We now include petal width in our linear model:

```
lm3 = lm(Petal.Length ~ . -1, data = iris)
summary(lm3)$coefficients[,1:2]
```

```
##                   Estimate Std. Error
## Petal.Width       1.018712 0.15224171
## Speciessetosa     1.211397 0.06524192
## Speciesversicolor 2.909188 0.20882146
## Speciesvirginica  3.488090 0.31303383
```

Suppose we have a new observation that is of species virginica and has a petal width of 1.

1. By hand, obtain the predicted value for this observation
2. In R, obtain a 95% prediction interval for this observation

Lab Ticket

We are studying whether the weights of red and pink grapefruit differ. We collect 5 grapefruits and measure their weight in grams:

```
##   weight type
## 1    8.3  red
## 2    7.0  red
## 3    7.5 pink
## 4    9.0  red
## 5    6.0 pink
```

We also fit a linear model:

```
##             Estimate Std. Error
## (Intercept)     6.75  0.7285831
## typered         1.35  0.9405967
```

1. Write out the design matrix $\mathbb{X}$ used in this linear model
2. We have a new grapefruit that is pink. What is its predicted weight?