PSTAT 126

Regression Analysis

Laura Baracaldo & Rodrigo Targino

Lecture 12 & 13
Categorical Predictors

Categorical Predictors

We have studied multiple regression models with quantitative predictors only, but what if we want to include predictors that are qualitative in nature, such as: *eye color*, *treatment*, *location* or *type of business*?

Factors: *Factor Variables* allow the inclusion of qualitative predictors in the mean function of a multiple linear regression model. The different categories of a factor variable are called *levels*.

Examples of ***Two-Level Factors*** are: Sex (Male/Female), Treatment (Treated/Untreated), Health status (Sick/Healthy) etc; whereas ***Multiple-Level Factors*** include: Eye color (green/black/brown/blue), party affiliation (Democrat/Republican/Independent), product quality (bad, medium, good), among others

Example - Categorical predictors

High-School Data Set: Data was collected as a subset of 200 students from the “High School and Beyond” study conducted by the National Education Longitudinal Studies (NELS) program of the National Center for Education Statistics (NCES).

```
data(hsb);head(hsb,10)
```

```
##     id gender         race    ses schtyp     prog read write math science socst
## 1   70   male        white    low public  general   57    52   41      47    57
## 2  121 female        white middle public vocation   68    59   53      63    61
## 3   86   male        white   high public  general   44    33   54      58    31
## 4  141   male        white   high public vocation   63    44   47      53    56
## 5  172   male        white middle public academic   47    52   57      53    61
## 6  113   male        white middle public academic   44    52   51      63    61
## 7   50   male african-amer middle public  general   50    59   42      53    61
## 8   11   male     hispanic middle public academic   34    46   45      39    36
## 9   84   male        white middle public  general   63    57   54      58    51
## 10  48   male african-amer middle public academic   57    55   52      50    51
```

Example - Categorical predictors

- Gender: Female/Male
- Race: African-American/Asian/Hispanic/White
- Socioeconomic class: High/Low/Middle
- School type(schtyp): Private/Public
- High school program: Academic/General/Vocation

```r
summary(hsb[,-1])
```

```
##     gender             race            ses          schtyp          prog          read      
##  female:109   african-amer: 20   high  :58   private: 32   academic:105   Min.   :28.00  
##  male  : 91   asian       : 11   low   :47   public :168   general : 45   1st Qu.:44.00  
##               hispanic    : 24   middle:95                 vocation: 50   Median :50.00  
##               white       :145                                            Mean   :52.23  
##                                                                           3rd Qu.:60.00  
##                                                                           Max.   :76.00  
##      write            math          science          socst      
##  Min.   :31.00   Min.   :33.00   Min.   :26.00   Min.   :26.00  
##  1st Qu.:45.75   1st Qu.:45.00   1st Qu.:44.00   1st Qu.:46.00  
##  Median :54.00   Median :52.00   Median :53.00   Median :52.00  
##  Mean   :52.77   Mean   :52.65   Mean   :51.85   Mean   :52.41  
##  3rd Qu.:60.00   3rd Qu.:59.00   3rd Qu.:58.00   3rd Qu.:61.00  
##  Max.   :67.00   Max.   :75.00   Max.   :74.00   Max.   :71.00  
```

Two-Level Factors

We aim to incorporate qualitative predictors within the MLR framework, so that we can extend estimation, inferential and diagnostics techniques more easily. In order to include *factors* in the model $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$ we need to codify the categorical variables by using *dummy variables*.

For a **Two-Level Factor** with levels *A* and *B*, we define dummy variables for individual ith as:

$$z_i = \begin{cases} 1 & \text{if } ith \in \text{Level A} \\ 0 & \text{if } ith \notin \text{Level A} \end{cases}$$

So that the model at the individual level is written as:

$$y_i = \beta_0 + \beta_A z_i + \epsilon_i = \begin{cases} \beta_0 + \beta_A + \epsilon_i & \text{if } ith \in \text{Level A} \\ \beta_0 + \epsilon_i & \text{if } ith \notin \text{Level A} \end{cases}$$

High School Data Example

Suppose we want to study the response y: *Science Score* as a function of *School Type* (private/public). We define the dummy variable with respect to Level public:

$$z_i = \begin{cases} 1 & \text{if } ith \in \text{Public} \\ 0 & \text{if } ith \notin \text{Public} \end{cases}$$

Thus the Linear model is:

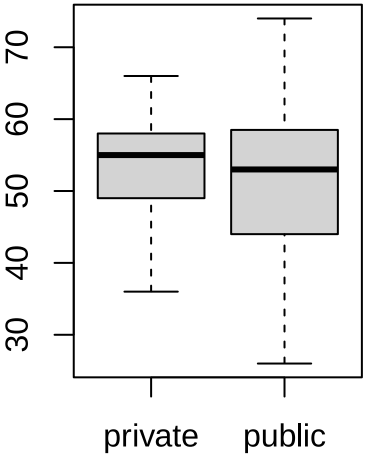
$$y_i = \beta_0 + \beta_{public} z_i + \epsilon_i$$

The interpretation of β_{public}: Average difference in science scores for students in private schools with respect science scores in public schools: $\beta_{public} = \bar{y}_{public} - \bar{y}_{private}$.

Private Schools vs Public Schools

Research question: Is there a statistically significant difference in the average science scores of public and private schools?

```
par( mar = c(2, 2, 0.8, 0.5));plot(science~schtyp, hsb)
```



Private Schools vs Public Schools

```
lmod <- lm(science~schtyp, hsb) # R automatically recognizes schtyp as a factor
summary(lmod)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)   53.312500   1.750993 30.4470164 1.257740e-76
## schtyppublic -1.741071   1.910490 -0.9113221 3.632338e-01
```

```
#R creates dummy var associated to b_public
lmod2 <- lm(science~as.factor(schtyp), hsb)
summary(lmod2)$coefficients
```

```
##                          Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)             53.312500   1.750993 30.4470164 1.257740e-76
## as.factor(schtyp)public -1.741071   1.910490 -0.9113221 3.632338e-01
```

Private Schools vs Public Schools

What if we want to construct a dummy variable with respect to the level private?, i.e:

$$z_i = \begin{cases} 1 & \text{if } ith \in \text{Private} \\ 0 & \text{if } ith \notin \text{Private} \end{cases}$$

Thus the Linear model is:

$$y_i = \beta_0 + \beta_{private} z_i + \epsilon_i$$

$\beta_{private} = \bar{y}_{private} - \bar{y}_{public}$

```
private<- ifelse(hsb$schtyp=="private", 1, 0)
lmod3 <- lm(science~private, hsb) ;summary(lmod3)$coefficients
```

```
##              Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 51.571429  0.7641958 67.4845733 1.317579e-138
## private      1.741071  1.9104895  0.9113221  3.632338e-01
```

Factors and Quantitative predictors

Suppose we want to include a quantitative variable x and a two-level factor z in the model. There are two possibilities:

1. Separate regression lines for each level with the same slope:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i = \begin{cases} \beta_0 + \beta_2 + \beta_1 x_i + \epsilon_i & ith \in \mathsf{A} \\ \beta_0 + \beta_1 x_i + \epsilon_i & ith \notin \mathsf{A} \end{cases}$$

2. Separate regression lines for each level with different slopes:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + \epsilon_i & ith \in \mathsf{A} \\ \beta_0 + \beta_1 x_i + \epsilon_i & ith \notin \mathsf{A} \end{cases}$$

High School Example

1. Separate regression lines with common slope and different intercepts.

```
lmod4 <- lm(science~math+schtyp, hsb) ;summary(lmod4)$coefficients
```

```
##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)  16.83230818 3.49241733  4.81967262 2.867077e-06
## math          0.66630487 0.05871397 11.34831838 2.714636e-23
## schtyppublic -0.07134314 1.49665267 -0.04766847 9.620288e-01
```

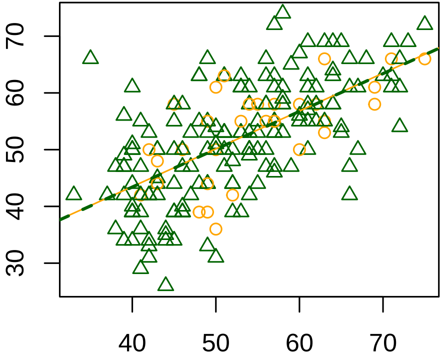
2. Separate regression lines with different slopes and different intercepts.

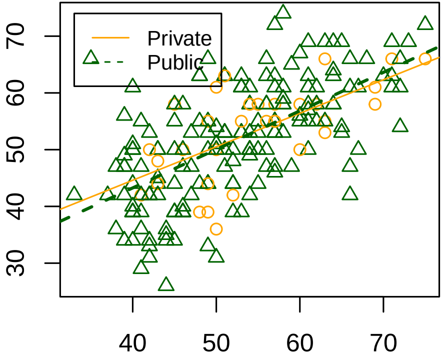
```
lmod5 <- lm(science~math+schtyp + math:schtyp  , hsb) ;summary(lmod5)$coefficients
```

```
##                      Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       21.00194195  8.6726120  2.4216397 0.0163611963
## math               0.59014718  0.1564221  3.7727875 0.0002137695
## schtyppublic      -4.89629406  9.3042936 -0.5262403 0.5993162301
## math:schtyppublic  0.08870108  0.1688129  0.5254403 0.5998710777
```

High School Example

```
par(mar = c(2, 2, 0.8, 0.5), mfrow=c(1,2)); colors<- c("orange", "darkgreen")
plot(science~math, hsb, pch=as.numeric(schtyp), col=colors[hsb$schtyp])
abline(lmod4$coefficients[1],lmod4$coefficients[2],col="orange"  )
abline(lmod4$coefficients[1] + lmod4$coefficients[3],lmod4$coefficients[2], col="darkgreen" , lty=2, lwd=2)
plot(science~math, hsb, pch=as.numeric(schtyp), col=colors[hsb$schtyp])
abline(lmod5$coefficients[1],lmod5$coefficients[2],col="orange"  )
abline(lmod5$coefficients[1] + lmod5$coefficients[3],lmod5$coefficients[2]
       +lmod5$coefficients[4], col="darkgreen" , lty=2, lwd=2)
legend(min(hsb$math),max(hsb$science),legend=c( "Private", "Public"),col=c("orange", "darkgreen"), lty=1:2, cex
```





Junior School Project Example

Data set: Junior School Project collected from primary (U.S. term is elementary) schools in inner London. y: English Score, x: Math Score, z: Girl=1/Boy=0.

1. Separate regression lines with common slope and different intercepts.

```
lmod6 <- lm(english~math+gender, jsp) ;summary(lmod6)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  0.8612033  5.9239913 0.1453755 8.845630e-01
## math         1.6328342  0.2007725 8.1327583 4.612838e-14
## gendergirl  11.9531480  3.0095840 3.9716944 1.000162e-04
```

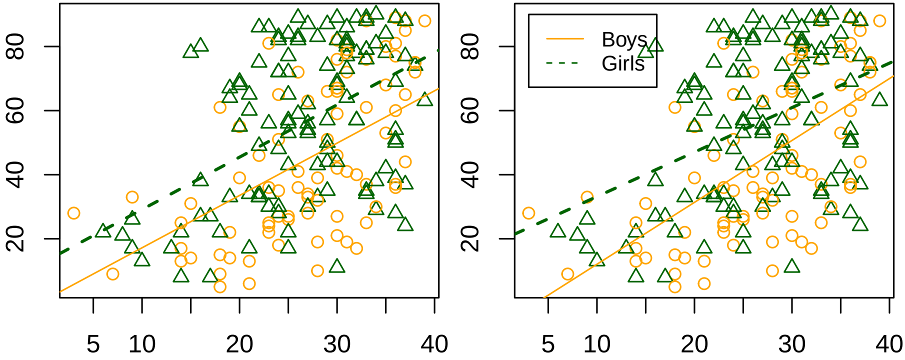
2. Separate regression lines with different slopes and different intercepts.

```
lmod7 <- lm(english~math+gender+math:gender, jsp) ;summary(lmod7)$coefficients
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     -7.2523204  8.4378332 -0.8595003 3.911146e-01
## math             1.9309689  0.2984735  6.4694817 7.681545e-10
## gendergirl      26.5340946 11.2288999  2.3630182 1.910586e-02
## math:gendergirl -0.5426657  0.4026852 -1.3476176 1.793371e-01
```

High School Example

```
par(mar = c(2, 2, 0.8, 0.5), mfrow=c(1,2))
plot(english~math, jsp, pch=as.numeric(gender), col= colors[jsp$gender])
abline(lmod6$coefficients[1],lmod6$coefficients[2], col="orange"  )
abline(lmod6$coefficients[1] + lmod6$coefficients[3],lmod6$coefficients[2],  col="darkgreen" , lty=2, lwd=2)
plot(english~math, jsp, pch=as.numeric(gender), col= colors[jsp$gender])
abline(lmod7$coefficients[1],lmod7$coefficients[2], col="orange" )
abline(lmod7$coefficients[1] + lmod7$coefficients[3],lmod7$coefficients[2]
       +lmod7$coefficients[4], col="darkgreen" , lty=2, lwd=2)
legend(min(jsp$math),max(jsp$english),legend=c("Boys", "Girls"),col=c("orange", "darkgreen"), lty=1:2, cex=0.8)
```



Factors With More Than Two Levels

Suppose we have a factor with more than m levels, then we create $m - 1$ dummy variables $z_2, \dots, z_m$ for subjects $1, \dots, n$ where:

$$z_{ij} = \begin{cases} 1 & \text{if } ith \in \text{Level } j \\ 0 & \text{if } ith \notin \text{Level } j \end{cases}$$

So that level 1 is the reference level. Why do we create $m - 1$ and not m dummy variables? *Answer:* To make $\boldsymbol{X}^T\boldsymbol{X}$ non-singular. Note that if we created m dummy variables, the design matrix $\boldsymbol{X}$ would have m linearly independent columns out of $m + 1$ columns $\Rightarrow \boldsymbol{X}^T\boldsymbol{X}$ would not be invertible.

$$\begin{bmatrix} \mathbf{1}_{g_1} & \mathbf{1}_{g_1} & \mathbf{0}_{g_1} & \dots & \mathbf{0}_{g_1} \\ \mathbf{1}_{g_2} & \mathbf{0}_{g_2} & \mathbf{1}_{g_2} & \dots & \mathbf{0}_{g_2} \\ \vdots & \vdots & \dots & \ddots & \vdots \\ \mathbf{1}_{g_m} & \mathbf{0}_{g_m} & \mathbf{0}_{g_m} & \dots & \mathbf{1}_{g_m} \end{bmatrix}$$

HS Example: Multiple-Level Factor

y: Science Score; *Factor*: Socioeconomic class (ses), *Levels*: High, low, middle.

```
attach(hsb)
contrasts(ses) # To identify reference level in R
```

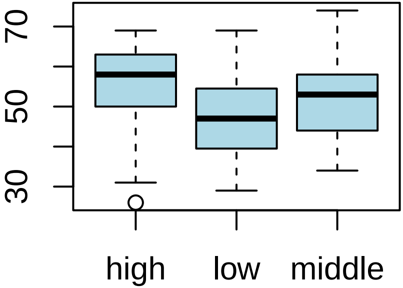
```
##        low middle
## high     0      0
## low      1      0
## middle   0      1
```

$$z_{2i} = \begin{cases} 1 & \text{if } ith \in \text{Low} \\ 0 & \text{if } ith \notin \text{Low} \end{cases} \qquad z_{3i} = \begin{cases} 1 & \text{if } ith \in \text{Middle} \\ 0 & \text{if } ith \notin \text{Middle} \end{cases}$$

$$y_i = \beta_0 + \beta_L z_{2i} + \beta_M z_{3i} + \epsilon_i = \begin{cases} \beta_0 + \beta_L + \epsilon_i & ith \in \text{Low} \\ \beta_0 + \beta_M + \epsilon_i & ith \in \text{Middle} \\ \beta_0 + \epsilon_i & ith \in \text{High} \end{cases}$$

HS Example: Multiple-Level Factor

```
par(mar = c(3, 2, 0.1, 2))
plot(science~ses, hsb, col="lightblue")
```




```
lmod <- lm(science ~ ses, hsb)
summary(lmod)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 55.448276   1.253244 44.243785 2.784866e-104
## seslow      -7.746148   1.873189 -4.135274  5.245239e-05
## sesmiddle   -3.743013   1.590449 -2.353432  1.958629e-02
```

HS Example: Factor with Quantitative Predictor

We start by a model that considers three separate regression lines with different intercepts and different slopes:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 x z_{2i} + \beta_5 x z_{3i} + \epsilon_i \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x + \epsilon_i & ith \in \text{Low} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x + \epsilon_i & ith \in \text{Middle} \\ \beta_0 + \beta_1 x + \epsilon_i & ith \in \text{High} \end{cases}
\end{aligned}
$$

HS Example: Factor with Quantitative Predictor

```
lmod2 <- lm(science ~ ses+math+math:ses, hsb)
summary(lmod2)$coefficients
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    13.93452219  6.5917409  2.1139366 3.579831e-02
## seslow         -4.37035144  9.1284650 -0.4787608 6.326479e-01
## sesmiddle      10.93573010  7.9533586  1.3749826 1.707227e-01
## math            0.73904166  0.1159916  6.3715080 1.330549e-09
## seslow:math     0.03658966  0.1715758  0.2132566 8.313508e-01
## sesmiddle:math -0.22506463  0.1431631 -1.5720855 1.175602e-01
```

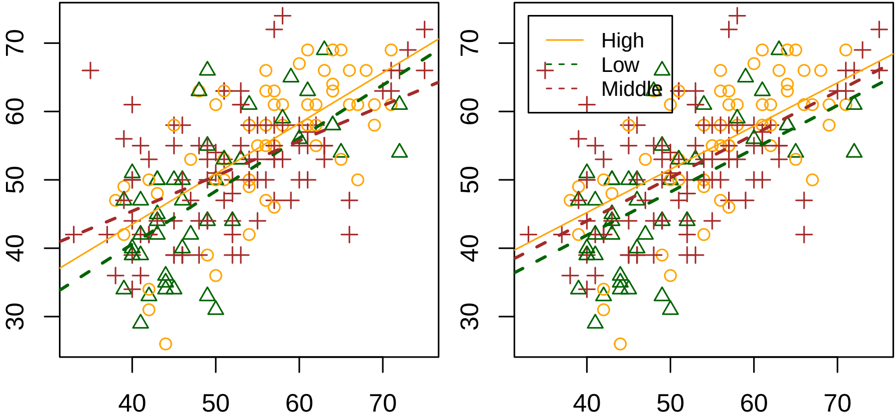
HS Example: Factor with Quantitative Predictor

If we remove the interaction between *math* and *ses* we introduce the model that considers three separate regression lines with different intercepts but common slope:

```
lmod3 <- lm(science ~ ses+math, hsb)
summary(lmod3)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 19.9108292 3.52782538  5.6439384 5.763019e-08
## seslow      -3.3162096 1.55995346 -2.1258388 3.476833e-02
## sesmiddle   -1.2365268 1.29733049 -0.9531317 3.416973e-01
## math         0.6326494 0.06020199 10.5087801 8.975825e-21
```

HS Example: Factor with Quantitative Predictor



ANOVA: Analysis of Variance

We can run a sequential ANOVA in order to decide on which predictors we should include in the model. Starting from a null model, we add the factor variable, then the quantitative variable and finally we add the interaction between them:

```
anova(lmod2)
```

```
## Analysis of Variance Table
##
## Response: science
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## ses         2  1561.6   780.8  13.4765 3.309e-06 ***
## math        1  6467.4  6467.4 111.6291 < 2.2e-16 ***
## ses:math    2   238.7   119.4   2.0602    0.1302
## Residuals 194 11239.8    57.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```