STAT 408
Applied Regression Analysis

Nan Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Categorical Predictors

Categorical Predictors

- Predictors that are qualitative in nature are described as categorical variables or factors
  - Eye color/gender

- The different categories are called levels
  - Suppose we recognize eye colors of “blue”, “green”, “brown” and “black”, then eye color is a factor with four levels

- In linear model, if we simply code the categories into numerical values, we assume the impact of increasing from one level to another is same
  - Also artificially define an “order” for different levels

- Both are unrealistic constraints added on our linear model – we need an appropriate coding method for categorical variable

A Two-Level Factor

- We start from an example to check the simplest case: a categorical predictor with two levels
- Dataset “sexab” describes the effects of childhood sexual abuse on 76 adult females
- Three variables are included: childhood sexual abuse (csa, categorical), childhood physical abuse (cpa, continuous), and post-traumatic stress disorder (ptsd, continues, response)
- Among 76 females, 45 experienced childhood sexually

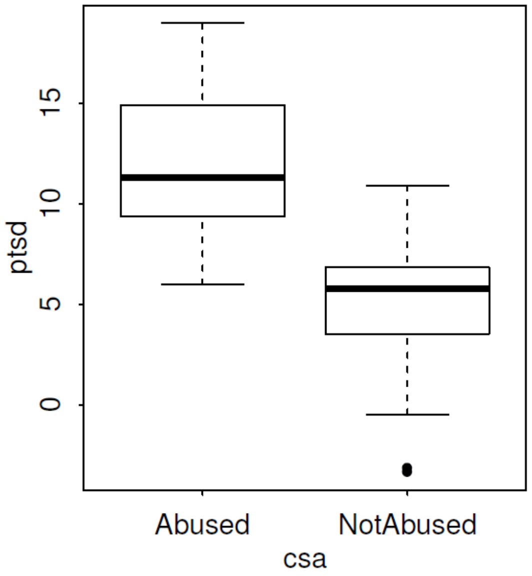
abused and 31 did not

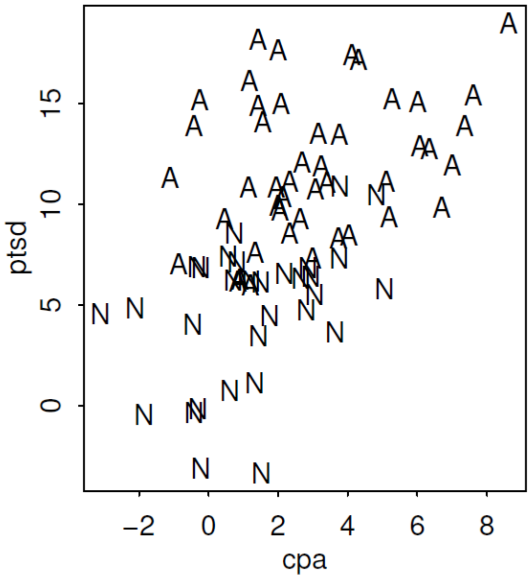
cpa	ptsd	csa
2.04786	9.71365	Abused
0.83895	6.16933	Abused
-0.24139	15.15926	Abused
-1.11461	11.31277	Abused
2.01468	9.95384	Abused
6.71131	9.83884	Abused

A Two-Level Factor

- Let’s visualize the data based on the categorical variable csa (abused/not abused)

```
plot(ptsd ~ csa, sexab)
plot(ptsd ~ cpa, pch=as.character(csa), sexab)
```





- We can see that the ptsd is much higher in the abused group

- Also, the ptsd is positively related to cpa

A Two-Level Factor

- In R, the lm function automatically encodes the categorical variable to perform linear regression, but here we will manually implement those operations

- To add categorical variables into regression model, we use <u>dummy variable</u>
  - For a categorical predictor with two levels, we define dummy variables d1 and d2

$$d_i = \begin{cases} 0 & \text{is not level i} \\ 1 & \text{is level i} \end{cases}$$

```
d1 <- ifelse(sexab$csa == "Abused",1,0)
d2 <- ifelse(sexab$csa == "NotAbused",1,0)
```

- Next, we regress ptsd on d1 and d2

A Two-Level Factor

```
> lmod <- lm(ptsd ~ d1 + d2, sexab)
> sumary(lmod)
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.696      0.624    7.53  1.0e-10
d1             7.245      0.811    8.94  2.2e-13

n = 76, p = 2, Residual SE = 3.473, R-Squared = 0.52
```

- We see a warning about singularities and that the parameter for d2 has not been estimated

```
> model.matrix(lmod)
   (Intercept) d1 d2
1            1  1  0
2            1  1  0
...
44           1  1  0
45           1  1  0
46           1  0  1
47           1  0  1
...
76           1  0  1
```

- The reason is intercept = d1 + d2, so we have perfect linear relation among predictors

- To obtain unique solution for model estimation, R automatically dropped one predictor d2

Interpretation of Two-Level Factor Model

- The fitted model is ptsd = 4.696 + 7.245 * d1
- The intercept 4.696 is the ptsd when d1=0 (not abused), which is the average ptsd in <u>not-abused group</u>

```
> mean(sexab[sexab$csa=="NotAbused", 'ptsd'])
[1] 4.695874
```

- The slope 7.245 is the <u>average increase</u> of ptsd if d1 changes from 0 to 1 (not abused to abused)
- So 4.696 + 7.245 = 11.941 would be the <u>average ptsd in abused group</u>

```
> mean(sexab[sexab$csa=="Abused", 'ptsd'])
[1] 11.94109
```

- Therefore, the linear regression on one two-level factor is to calculate the average responses in these two groups

The Interpretation of Two-Level Factor Model

- The dropped level d2 (not-abused) is called reference level (baseline), which can be understood as the baseline category
- The reference level is usually the "non-treatment" category, in this example, the not-abused
- R can automatically encode the categorical variable into dummy variable, but the reference level is based on the alphabetical sequence (abused)

```
lmod <- lm(ptsd~csa, sexab)
summary(lmod)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   11.9411     0.5177  23.067  < 2e-16 ***
csaNotAbused  -7.2452     0.8105  -8.939 2.17e-13 ***
```

- The fitted model is ptsd = 11.9411 - 7.245 * d2

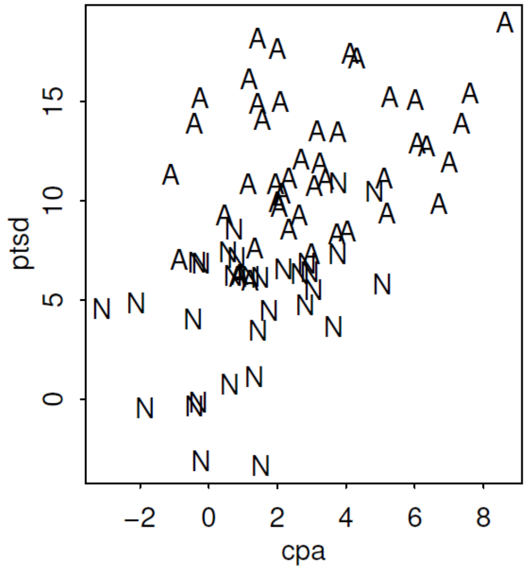
Factors and Quantitative Predictors

- What if the predictors include both categorical and numerical variables?

- In the sexab dataset, the ptsd not only relates to csa (categorical), but also cpa (numerical) – we need to include both in our linear model

- Suppose we have a response Y, a quantitative predictor X and a two-level factor variable represented by a dummy variable d:

$$d = \begin{cases} 0 & \text{reference level} \\ 1 & \text{treatment level} \end{cases}$$



Factors and Quantitative Predictors

- We could build a linear model with both x and d as:

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x.d + \varepsilon$$

- This model separates regression lines for each group with the different slopes
  - When d = 0, the parameter (slope) of x is β_1
  - When d = 1, the parameter (slope) of x is $\beta_1 + \beta_3$

- This model can also be understood as adjusting the covariate X for the effect of d

Factors and Quantitative Predictors

- We use this method to regress ptsd on cpa and csa

```
lmod <- lm(ptsd~cpa+csa+cpa:csa,sexab)
summary(lmod)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       10.5571     0.8063  13.094  < 2e-16 ***
cpa                0.4500     0.2085   2.159   0.0342 *  
csaNotAbused      -6.8612     1.0747  -6.384 1.48e-08 ***
cpa:csaNotAbused   0.3140     0.3685   0.852   0.3970    
```

- The fitted model is

  ptsd = 10.5571 + 0.45*cpa – 6.8612*NotAbused + 0.314*cps*NotAbused

- If no abuse

  ptsd = 10.5571 + 0.45*cpa – 6.8612*1 + 0.314*cpa*1 = 3.7 + 0.764*cpa

- If abuse

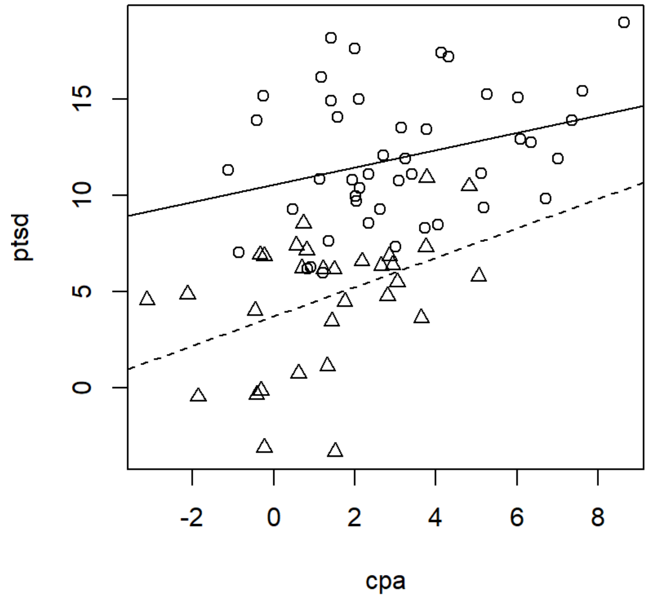
  ptsd = 10.5571 + 0.45*cpa

Factors and Quantitative Predictors

- The interaction term implies that we believe the sexual abuse variable not only affects the average stress disorder (distance between two lines), but also the relation between stress disorder and physical abuse (slope)

- However, the interaction term is not significant, indicating sexual abuse dose not change the relation between stress disorder and physical abuse (slope), and the two lines should be parallel

```
plot(ptsd~cpa, sexab, pch=as.numeric(csa))
abline(3.7, 0.764, lty=2)
abline(10.5571, 0.45)
```



Factors and Quantitative Predictors

- We refit the model without the interaction term

```
lmod <- lm(ptsd~cpa+csa,sexab)
summary(lmod)

              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    10.2480     0.7187  14.260  < 2e-16 ***
cpa             0.5506     0.1716   3.209  0.00198 ** 
csaNotAbused   -6.2728     0.8219  -7.632 6.91e-11 ***
```

- The fitted model is

ptsd = 10.248 + 0.5506*cpa – 6.2728*NotAbused

- If no abuse

ptsd = 10.248 + 0.5506*cpa – 6.2728*1 = 3.9752 + 0.5506 *cpa

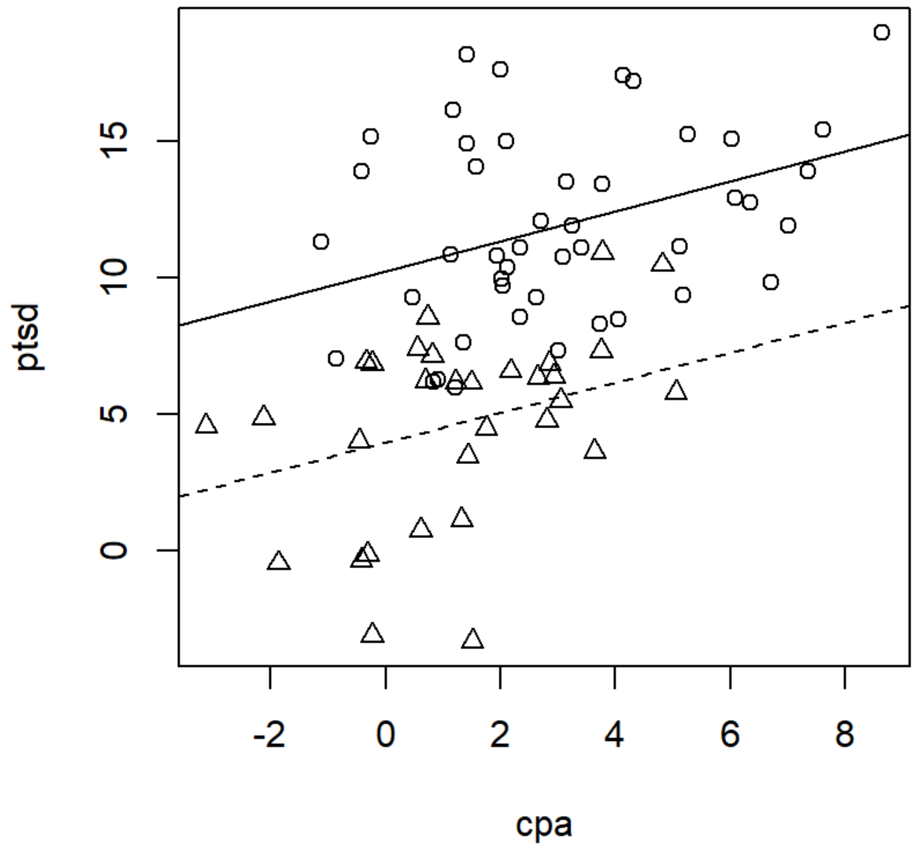
- If abuse

ptsd = 10.248 + 0.5506*cpa

Factors and Quantitative Predictors

- In this model, cpa and csa will affect ptsd independently and they are significant

```r
plot(ptsd~cpa, sexab, pch=as.numeric(csa))
abline(3.9752, 0.5506, lty=2)
abline(10.248, 0.5506)
```



Factors With More Than Two Levels

- Let's generalize the two-level factors to multi-level factor
- Suppose we have a factor with *f* levels, then we create *f* − 1 dummy variables d_2, …, d_f

$$d_i = \begin{cases} 0 & \text{is not level i} \\ 1 & \text{is level i} \end{cases}$$

where level one d_1 is the reference level

- We demonstrate the use of multilevel factors with a study on the happiness and social life

Factors With More Than Two Levels

- Dataset "happy" contains 5 variables on 39 MBA students
  - Happy: happiness on a 10-point scale where 10 is most happy (numerical)
  - Money: family income in thousands of dollars (numerical)
  - Sex: 1 = satisfactory sexual activity, 0 = not (binary)
  - Love: 1 = lonely, 2 = secure relationships, 3 = deep feeling of belonging and caring (3-level categorical)
  - Work: 5-point scale where 1 = no job, 3 = OK job, 5 = great job (5-level categorical)

- Let's regress happy on other variables

Factors With More Than Two Levels

summary(happy)

lmod <- lm(happy~money+sex+love+work, data = happy)

summary(lmod)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749 .
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07 ***
work         0.476079   0.199389   2.388   0.0227 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.058 on 34 degrees of freedom
Multiple R-squared:  0.7102,    Adjusted R-squared:  0.6761
F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

- By default, the model treats love and work as numerical variables
- It forces the same effects of love/work same when moving across different levels

Factors With More Than Two Levels

- We encode love and work to dummy variables and redo the regression

```
happy$love <- as.factor(happy$love)
happy$work <- as.factor(happy$work)
summary(happy)
lmod <- lm(happy~money+sex+love+work, data = happy)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.370241   0.993288   3.393  0.00196 **
money        0.008374   0.005464   1.533  0.13587
sex         -0.345443   0.470171  -0.735  0.46822
love2        1.850241   0.766964   2.412  0.02217 *
love3        3.845091   0.722507   5.322 9.39e-06 ***
work2       -0.792463   0.920846  -0.861  0.39629
work3        0.113597   0.899973   0.126  0.90040
work4        0.808892   0.857931   0.943  0.35329
work5        0.382735   1.128814   0.339  0.73693
```

- How can we interpret the output?

Factors With More Than Two Levels

- Let's check the design matrix after transforming to dummy variables

  X.factor <- model.matrix(lmod)

(Intercept)	money	sex	love2	love3	work2	work3	work4	work5
1	36	0	0	1	0	0	1	0
1	47	1	0	1	0	0	0	0
1	53	0	0	1	0	0	0	1
1	35	1	0	1	0	1	0	0
1	88	1	0	0	1	0	0	0
1	175	1	0	1	0	0	1	0
1	175	1	0	1	0	0	1	0
1	45	0	1	0	0	1	0	0

- We have two more columns for "love", and four more columns for "work"
- Those columns serve as the indicator for certain levels in love and work
- The baseline (love=1 and work=1) are incorporated in intercept

Factors With More Than Two Levels

- The impact of love on happy are different among different levels
    - The benefits of “deep feeling of belonging and caring” is much larger than “secure relationships”

- Compare model with or without dummy variables, we find work variable no longer significant

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.370241   0.993288   3.393  0.00196 **
money        0.008374   0.005464   1.533  0.13587
sex         -0.345443   0.470171  -0.735  0.46822
love2        1.850241   0.766964   2.412  0.02217 *
love3        3.845091   0.722507   5.322 9.39e-06 ***
work2       -0.792463   0.920846  -0.861  0.39629
work3        0.113597   0.899973   0.126  0.90040
work4        0.808892   0.857931   0.943  0.35329
work5        0.382735   1.128814   0.339  0.73693
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749 .
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07 ***
work         0.476079   0.199389   2.388   0.0227 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.058 on 34 degrees of freedom
Multiple R-squared:  0.7102,    Adjusted R-squared:  0.6761
F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```