Lecture 11: Transformation of Variables

Ailin Zhang

2023-06-06

Recap: Factor Analysis (ANOVA)

- Partition of the response-variable sum of squares into "explained" and "unexplained"
- Procedures for fitting and testing linear models in which the explanatory variables are categorical.
- Suppose that there are no quantitative explanatory variables—only a single factor in your model:

$$Y_i = \beta_0 + \gamma_2 E_{i2} + \gamma_3 E_{i3} + \epsilon_i$$

- $E(\hat{Y}_i)$ in each category (group, level of factor) is the population group mean, can be denoted by μ_j

Recap: One-way ANOVA

Education(E)	Indicator	E(Y)
1 (HS)	$E_2 = E_3 = 0$	$\mu_1 = \beta_0$
2 (B.S.)	$E_2 = 1, E_3 = 0$	$\mu_2 = \beta_0 + \gamma_2$
3 (Advanced)	$E_2 = 0, E_3 = 1$	$\mu_3 = \beta_0 + \gamma_3$

- One-way ANOVA focuses on testing for differences among group means.
- One-way ANOVA examines the relationship between a quantitative response variable and a factor.
- H_0: $\gamma_2 = \gamma_3 = 0$, which implies $\mu_1 = \mu_2 = \mu_3$
- F-statistic for the regression of the response variable on $0/1$ dummy regressors constructed from the factor tests for differences in the response means across levels of the factor.

Recap: One-way ANOVA Table

Source	df	Sum of Squares	Mean Squares	F
Treatment	c-1	SSR	MSR	$F = \frac{MSR}{MSE}$
Error	n-c	SSE	MSE	
Total	n-1	SST		

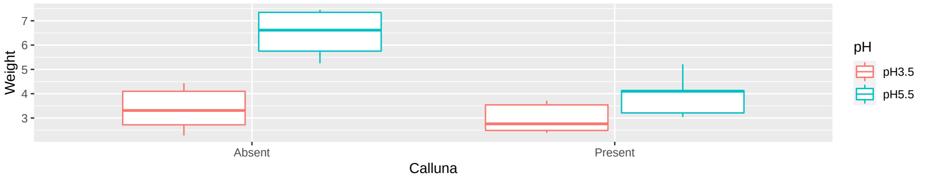
- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{j=1}^{c} n_j(\hat{\mu}_j - \bar{y})^2$
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{j=1}^{c}(n_j - 1)s_j^2$

Where sample variance: $s_j^2 = \dfrac{\sum_{i \in \text{group}_j}(y_i - \mu_j)^2}{n_j - 1}$

Recap: Two-way ANOVA

- Two-way ANOVA allows us to determine whether there are significant differences between the effects of two categorical variables.
- Change of response variable may depend on
  - the level of the categorical variable (additive model)
  - the level of the interaction model.

```
festuca <- read.table(file = "festuca.txt",header = T)
ggplot(data = festuca, aes(x = Calluna, y = Weight,
                           colour = pH)) +
  geom_boxplot()
```



Recap: ANOVA and Linear Regression

Model	Terms	Regression Sum of Squares
1	pH, Calluna, pH*Calluna	SSR_1
2	pH, Calluna	SSR_2
3	pH	SSR_3
4	Calluna	SSR_4

The four models will produce the two-way ANOVA table

Source	Model Contrasted	df	Sum of Squares	Mean Squares	F
pH	2-4	1	$SSR_2 - SSR_4$	MSR	$F = \frac{MSR}{MSE}$
Calluna	2-3	1	$SSR_2 - SSR_3$	MSR	$F = \frac{MSR}{MSE}$
Interaction	1-2	1	$SSR_1 - SSR_2$	MSR	$F = \frac{MSR}{MSE}$
Error	SSE from Model 1	n-4	SSE	MSE	
Total		n-1	SST		

Transformation of Variables

- When and Why?

  Original variables violates one or more of the standard regression assumptions.

  - Linearity of the model
  - Constant variance for the error

Transformation of Variables

- When and Why?

  Original variables violates one or more of the standard regression assumptions.

  - Linearity of the model
  - Constant variance for the error

- Transformation to achieve linearity

- Transformation to stabilize variance

Transformation of Variables

- When and Why?

  Original variables violates one or more of the standard regression assumptions.

  - Linearity of the model
  - Constant variance for the error

- Transformation to achieve linearity

- Transformation to stabilize variance

Our primary focus:

- Polynomial Models
- Ordinal Categorical Predictors

Data: Smoking and FEV (Lung Capacity)

Sample of 654 youths, aged 3 to 19, in the area of East Boston during middle to late 1970's. The variables are

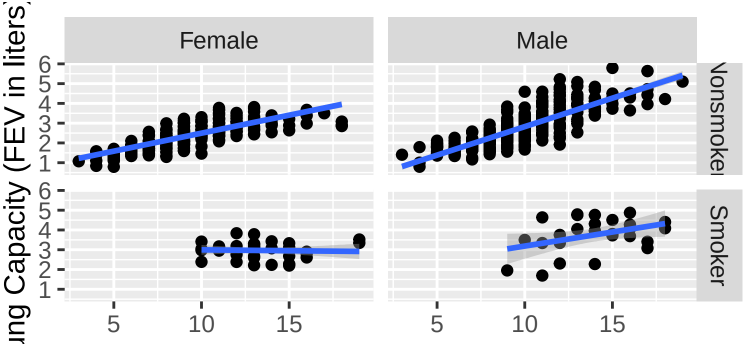
- age: Subject's age in years
- fev: Lung capacity of subject, measured by forced expiratory volume (abbreviated as FEV), the amount of air an individual can exhale in the first second of forceful breath in liters
- ht: Subject's height in inches
- sex: Gender of the subject coded as: 0 = Female, 1 = Male
- smoke:Smoking status coded as: 0 = Nonsmoker, 1 = Smoker

```
fevdata = read.table("fevdata.txt", header=TRUE)
fevdata$sex = factor(fevdata$sex, labels=c("Female","Male"))
fevdata$smoke = factor(fevdata$smoke, labels=c("Nonsmoker","Sm
```

Lung Capacity Dataset

```
ggplot(fevdata, aes(x = age, y = fev)) +
geom_point() + facet_grid(smoke~sex) +
geom_smooth(method='lm') + xlab("Age (years)") +
ylab("Lung Capacity (FEV in liters)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Test Non-linearity: Female Nonsmokers

- Children stop growing after they turn adults.
- FEV might not grow linearly with age, at least for female nonsmokers in the plots above.

Test Non-linearity: Female Nonsmokers

- Children stop growing after they turn adults.
- FEV might not grow linearly with age, at least for female nonsmokers in the plots above.
- To test non-linearity, one can add some nonlinear function of age, e.g. age^2 and see if the nonlinear term is significant.

```
f.nonsmokers = subset(fevdata, sex == "Female" & smoke == "Nonsmoker")
summary(lm(fev ~ age + I(age^2), data=f.nonsmokers))$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -0.50745967 0.210390388 -2.411991 1.651843e-02
## age          0.43979072 0.042969068 10.235054 4.714677e-21
## I(age^2)    -0.01297867 0.002120258 -6.121267 3.176433e-09
```

Test Non-linearity: Female Nonsmokers

- Children stop growing after they turn adults.
- FEV might not grow linearly with age, at least for female nonsmokers in the plots above.
- To test non-linearity, one can add some nonlinear function of age, e.g. age^2 and see if the nonlinear term is significant.

```
f.nonsmokers = subset(fevdata, sex == "Female" & smoke == "Nonsmoker")
summary(lm(fev ~ age + I(age^2), data=f.nonsmokers))$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -0.50745967 0.210390388 -2.411991 1.651843e-02
## age          0.43979072 0.042969068 10.235054 4.714677e-21
## I(age^2)    -0.01297867 0.002120258 -6.121267 3.176433e-09
```

- There is significant evidence of non-linearity.

Polynomial Models

- Fitting the polynomial model:

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \epsilon$$

doesn't mean we believe it is correct. It is just a decent approximation to the true underlying nonlinear model:

$$\text{fev} = f(\text{age}) + \epsilon$$

Polynomial Models

- Fitting the polynomial model:

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \epsilon$$

doesn't mean we believe it is correct. It is just a decent approximation to the true underlying nonlinear model:

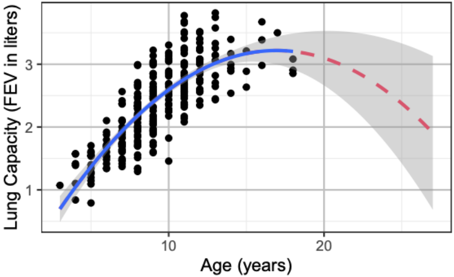
$$\text{fev} = f(\text{age}) + \epsilon$$

- One can try higher-order polynomials

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \cdots + \beta_k (\text{age})^k + \epsilon$$

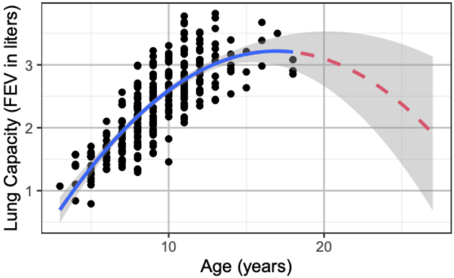
if lower-order ones don't capture the nonlinear pattern well.

Caution: Don't trust extrapolation!



Does lung capacity decrease after children turn adults?

Caution: Don't trust extrapolation!



Does lung capacity decrease after children turn adults?

- We are not sure whether the nonlinear relations is a polynomial (it's just an approximation!).
- Extrapolating the model beyond the range of data is dangerous.

Test of Non-linearity: Male Nonsmokers

```
m.nonsmokers = subset(fevdata, sex == "Male" & smoke == "Nonsmoker")
summary(lm(fev ~ age + I(age^2), data=m.nonsmokers))$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 0.143622429 0.298683403 0.4808517 6.309644e-01
## age         0.245874713 0.059137388 4.1576864 4.175341e-05
## I(age^2)    0.002057928 0.002820682 0.7295854 4.662000e-01
```

- The large P-value 0.466 for the age^2 means little evidence of non-linearity

Test of Non-linearity: Male Nonsmokers

```
m.nonsmokers = subset(fevdata, sex == "Male" & smoke == "Nonsmoker")
summary(lm(fev ~ age + I(age^2), data=m.nonsmokers))$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 0.143622429 0.298683403 0.4808517 6.309644e-01
## age         0.245874713 0.059137388 4.1576864 4.175341e-05
## I(age^2)    0.002057928 0.002820682 0.7295854 4.662000e-01
```

- The large P-value 0.466 for the age^2 means little evidence of non-linearity
- This just means fev is approximaltely linear in age in the range of data for male smokers. Extrapolating the line beyond of the range of data remain dangerous

Test of Non-linearity: Male Nonsmokers

```
m.nonsmokers = subset(fevdata, sex == "Male" & smoke == "Nonsmoker")
summary(lm(fev ~ age + I(age^2), data=m.nonsmokers))$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 0.143622429 0.298683403 0.4808517 6.309644e-01
## age         0.245874713 0.059137388 4.1576864 4.175341e-05
## I(age^2)    0.002057928 0.002820682 0.7295854 4.662000e-01
```

- The large P-value 0.466 for the age^2 means little evidence of non-linearity
- This just means fev is approximaltely linear in age in the range of data for male smokers. Extrapolating the line beyond of the range of data remain dangerous
- The discrepancy in the significance of age^2 between boys and girls is an evidence of age:sex interaction — lung capacities of girls stop growing earlier than boys.

Interactions

```r
nonsmokers = subset(fevdata, smoke == "Nonsmoker")
summary(lm(fev ~ (age + I(age^2))*sex, data=nonsmokers))$coef
```

```
##                      Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)       -0.50745967 0.263273891 -1.927497 5.440320e-02
## age                0.43979072 0.053769726  8.179151 1.795959e-15
## I(age^2)          -0.01297867 0.002653204 -4.891696 1.295007e-06
## sexMale            0.65108210 0.369659312  1.761303 7.871126e-02
## age:sexMale       -0.19391600 0.074369400 -2.607470 9.354961e-03
## I(age^2):sexMale  0.01503659 0.003611741  4.163254 3.611388e-05
```

- For girls: $\hat{\text{fev}} = -0.507 + 0.44\text{age} - 0.013(\text{age})^2$
- For boys: $\hat{\text{fev}} = (-0.507 + 0.651) + (0.44 - 0.194)\text{age} + (0.015 - 0.013)\text{age}^2 = 0.144 + 0.246\text{age} + 0.002(\text{age})^2$

Interpretation of Coefficients in a Polynomial Model

- Recall in MLR, we said β_j is the mean change in the response Y when X_j is increased by one unit holding other X_i's constant.
- For a model that involves polynomial terms like:

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_1^2}_{\text{Polynomial of } X_1} + \beta_3 X_3 + \cdots + \beta_p X_p + \epsilon$$

Interpretation of Coefficients in a Polynomial Model

- Recall in MLR, we said β_j is the mean change in the response Y when X_j is increased by one unit holding other X_i's constant.
- For a model that involves polynomial terms like:

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_1^2}_{\text{Polynomial of } X_1} + \beta_3 X_3 + \cdots + \beta_p X_p + \epsilon$$

- it makes no sense to interpret a single coefficient for a polynomial like β_1 or β_2 since it's impossible to change X_1 while holding X_1^2 constant

Interpretation of Coefficients in a Polynomial Model

- Recall in MLR, we said β_j is the mean change in the response Y when X_j is increased by one unit holding other X_i's constant.
- For a model that involves polynomial terms like:

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_1^2}_{\text{Polynomial of } X_1} + \beta_3 X_3 + \cdots + \beta_p X_p + \epsilon$$

- it makes no sense to interpret a single coefficient for a polynomial like β_1 or β_2 since it's impossible to change X_1 while holding X_1^2 constant
- Interpret the polynomials all together: the mean of Y change with X_1 following the curve $\beta_1 X_1 + \beta_2 X_1^2$ holding other X_i's constant.

Recap: Interpretation of Coefficients of Indicator Variables

$S = \beta_0 + \beta_1 X + \gamma_2 E_2 + \gamma_3 E_3 + \alpha M_1 + \epsilon$

- Interpret γ_2 as the mean difference in salary S between HS graduates and those with a Bachelor's degree if they were at the same management status and had the same years of experience.

Test of Non-linearity: Smokers

```r
m.smokers = subset(fevdata, sex == "Male" & smoke == "Smoker")
summary(lm(fev ~ age + I(age^2), data=m.smokers))$coef
```

```
##                Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) -6.05764029 4.77931185 -1.267471 0.21766965
## age          1.31395132 0.70557307  1.862247 0.07539297
## I(age^2)    -0.04253231 0.02550048 -1.667902 0.10889460
```

```r
f.smokers = subset(fevdata, sex == "Female" & smoke == "Smoker")
summary(lm(fev ~ age + I(age^2), data=f.smokers))$coef
```

```
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  5.98969790  1.9596750  3.056475 0.004204225
## age         -0.43746338  0.2843111 -1.538679 0.132627095
## I(age^2)     0.01536442  0.0101347  1.516022 0.138246407
```

- age^2 is insignificant for male smokers or female smokers, which might be just due to the small sample size that makes it difficult to detect the non-linearity.

Interactions

```
smokers = subset(fevdata, smoke == "Smoker")
summary(lm(fev ~ (age + I(age^2))*sex, data=smokers))$coef
```

```
##                        Estimate Std. Error   t value   Pr(>|t
## (Intercept)          5.98969790 2.79718654  2.141329 0.036388
## age                 -0.43746338 0.40581786 -1.077980 0.285430
## I(age^2)             0.01536442 0.01446599  1.062106 0.292515
## sexMale            -12.04733819 4.53161519 -2.658509 0.010086
## age:sexMale          1.75141470 0.66462640  2.635187 0.010726
## I(age^2):sexMale    -0.05789673 0.02389848 -2.422612 0.018498
```

Male smokers still have significantly larger lung capacities than female nonsmokers, though neither show significant non-linearity

Question: Can we remove the square term age^2 for smokers?

Question: Can we remove the square term age^2 for smokers?

```
anova(lm(fev ~ (age + I(age^2))*sex, data=smokers),
  lm(fev ~ age*sex, data=smokers))
```

```
## Analysis of Variance Table
##
## Model 1: fev ~ (age + I(age^2)) * sex
## Model 2: fev ~ age * sex
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     59 21.279
## 2     61 23.489 -2   -2.2098 3.0635 0.05422 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Ordinal Categorical Predictors; Salary Data

- An ordinal variable is a categorical variable with ordered categories.
  - e.g., E = education (HS only, BA or BS, advance degree) in the salary survey data is an ordinal variable

Ordinal Categorical Predictors; Salary Data

- An ordinal variable is a categorical variable with ordered categories.
  - e.g., E = education (HS only, BA or BS, advance degree) in the salary survey data is an ordinal variable
- When we create indicators for E and include them in a model, we ignore the fact that the 3 education levels are ordered. Therefore, the estimated salary might not be ordered by education levels.

Ordinal Categorical Predictors; Salary Data

- An ordinal variable is a categorical variable with ordered categories.
  - e.g., E = education (HS only, BA or BS, advance degree) in the salary survey data is an ordinal variable
- When we create indicators for E and include them in a model, we ignore the fact that the 3 education levels are ordered. Therefore, the estimated salary might not be ordered by education levels.
- We can incorporate the ordinal info of a ordinal predictor by assigning a score to each its category like

$$E = \begin{cases} 1 & \text{if HS only} \\ 2 & \text{if Bachelor's degree} \\ 3 & \text{if Advanced degree} \end{cases}$$

Ordinal Categorical Predictors: Salary Data

$S = \beta_0 + \beta_1 X + \beta_2 E + \alpha M_1$

This way, the fitted salary will always be ordered by education levels:

- a BS increases salary by β_1 from HS
- an advanced degree increases salary by another β_1 from BS

```
salary = read.table("salary.txt", header=TRUE)
salary$Man = factor(salary$M,
                    labels=c("No manager","Manager"))
lmsalary = lm(S ~ X+E+Man, salary)
summary(lmsalary)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 6963.4777  665.69473 10.460467 2.876500e-13
## X            570.0874   38.55905 14.784789 3.000130e-18
## E           1578.7503  262.32162  6.018377 3.737405e-07
## ManManager  6688.1299  398.27563 16.792717 3.043277e-20
```

Flexibilty with Oridinal Predictors

- If one believe the salary gap between a Bachelor's deg. and HS diploma is greater than that between a Bachelor's deg. and an adv. deg., one may try a different scoring (1, 2.5, 3), i.e.,

$$E = \begin{cases} 1 & \text{if HS only} \\ 2.5 & \text{if Bachelor's degree} \\ 3 & \text{if Advanced degree} \end{cases}$$

Flexibilty with Oridinal Predictors

- If one believe the salary gap between a Bachelor's deg. and HS diploma is greater than that between a Bachelor's deg. and an adv. deg., one may try a different scoring (1, 2.5, 3), i.e.,

$$E = \begin{cases} 1 & \text{if HS only} \\ 2.5 & \text{if Bachelor's degree} \\ 3 & \text{if Advanced degree} \end{cases}$$

$S = \beta_0 + \beta_1 X + \beta_2 E + \alpha M_1$

This way, the fitted salary will always be ordered by education levels:

- a BS increases salary by $1.5\beta_1$ from HS
- an advanced degree increases salary by another $0.5\beta_1$ from BS

R Example

```r
salary$E.score1 = ifelse(salary$E == 2, 2.5, salary$E)
summary(lm(S ~ M + E.score1 + X, data=salary))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 6401.1019  561.45563 11.400904 1.943394e-14
## M           6741.1167  330.93922 20.369652 2.183795e-23
## E.score1    1703.2956  204.22077  8.340462 1.882827e-10
## X            562.2457   32.00685 17.566421 5.778756e-21
```

- A BS increases mean salary by $1.5 \times 1703 = 2554.5$ than HS
- An advanced degree increases mean salary by another $0.5 \times 1703 = 851.5$ than BS

Another scoring scheme

```r
salary$E.score2 = ifelse(salary$E >= 2, salary$E + 1, salary$E)
summary(lm(S ~ M + E.score2 + X, data=salary))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 7072.138  535.55418 13.205270 1.516893e-16
## M           6711.605  349.58912 19.198552 2.074099e-22
## E.score2    1135.099  148.43586  7.647068 1.750709e-09
## X            565.975   33.81701 16.736401 3.442138e-20
```

Another scoring scheme

```
salary$E.score2 = ifelse(salary$E >= 2, salary$E + 1, salary$E)
summary(lm(S ~ M + E.score2 + X, data=salary))$coef
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 7072.138  535.55418 13.205270 1.516893e-16
## M           6711.605  349.58912 19.198552 2.074099e-22
## E.score2    1135.099  148.43586  7.647068 1.750709e-09
## X            565.975   33.81701 16.736401 3.442138e-20
```

- Which model fits better? How to compare?

Another scoring scheme

```r
salary$E.score2 = ifelse(salary$E >= 2, salary$E + 1, salary$E)
summary(lm(S ~ M + E.score2 + X, data=salary))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 7072.138  535.55418 13.205270 1.516893e-16
## M           6711.605  349.58912 19.198552 2.074099e-22
## E.score2    1135.099  148.43586  7.647068 1.750709e-09
## X            565.975   33.81701 16.736401 3.442138e-20
```

- Which model fits better? How to compare?

```r
summary(lm(S ~ M + E.score1 + X, data=salary))$r.squared
```

```
## [1] 0.9493052
```

```r
summary(lm(S ~ M + E.score2 + X, data=salary))$r.squared
```

```
## [1] 0.943712
```

Comparison of Models with Ordinal and Nominal Predictors

- Whatever scoring one uses for E, the ordinal model is always a nested model of that treats E as nominal

Comparison of Models with Ordinal and Nominal Predictors

- Whatever scoring one uses for E, the ordinal model is always a nested model of that treats E as nominal

```
anova(lm(S ~ M + E.score1 + X, data=salary),
lm(S ~ M + as.factor(E) + X, data=salary))
```

```
## Analysis of Variance Table
##
## Model 1: S ~ M + E.score1 + X
## Model 2: S ~ M + as.factor(E) + X
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     42 50750487
## 2     41 43280719  1   7469768 7.0761 0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Pros and Cons of Using Ordinal Predictors

Pros: Simplicity, fewer parameters

Pros and Cons of Using Ordinal Predictors

Pros: Simplicity, fewer parameters

- A categorical predictor with c categories requires $c - 1$ parameters if regarded as nominal since each indicator variable needs 1 parameter
- An ordinal predictor that uses the scores as the numerical values for its categories of need only 1 parameter

Cons:

Pros and Cons of Using Ordinal Predictors

Pros: Simplicity, fewer parameters

- A categorical predictor with c categories requires $c - 1$ parameters if regarded as nominal since each indicator variable needs 1 parameter
- An ordinal predictor that uses the scores as the numerical values for its categories of need only 1 parameter

Cons:

- The choice of scores seems arbitrary
- If the scores are not chosen properly, the model might not be reasonable.