# BIOSTAT 650 SEC 002 HW6

Jaehoon Kim

2024-10-28

```
mySASData = read.sas7bdat("completedata.sas7bdat")

var = c(
"Depression",
"Age",
"Sex",
"R_E",
"Fatalism"
)

df <- mySASData[ ,var]
typeof(df$Depression)
```

```
## [1] "double"
```

```
#(a)
df <- df |>
  mutate(Age_Cat = cut(Age,breaks = c(-Inf, quantile(Age,c(0.25, 0.5, 0.75)), Inf), labels = c(0, 1, 2,


#df$Age_Cat

#(b)
#include age variable as well
#figure this code out later
#total <- nrow(df)
#age_summary <- df %>%
#  group_by(Age_Cat) %>%
#  summarize(
#    Percentage = n()/total*100,
#    mean = mean(Depression),
#    Max = max(Depression),
#    Min = min(Depression),
#    sd = sd(Depression)
#  )
#age_summary
dfage0<- df[(df$Age_Cat==0)|(df$Age_Cat==1),]
dfage0$Age_Cat=as.numeric(dfage0$Age_Cat)-1
library("gtsummary")
dfage0 |> select("Depression", "Age_Cat", "Age")|>
  tbl_summary(by = Age_Cat,
              type = list(Depression ~ "continuous", Age ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({sd}); {median}, {min}, {max}"),
              digits = list(all_continuous() ~ c(2,2,0,0)))
```

| **Characteristic** | **0** N = 156[1] | **1** N = 158[1] |
|---|---|---|
| Depression | 7.01 (5.69); 6, 0, 25.00 | 5.41 (5.24); 3, 0, 27.00 |
| Age | 51.49 (3.39); 52, 45, 56.00 | 60.04 (2.12); 60, 57, 64.00 |

[1] Mean (SD); Median, Min, Max

```
#Arguments ' ' single quotes for list not working for some reason.
```

```
dfage23<- df[(df$Age_Cat==2)|(df$Age_Cat==3),]
dfage23$Age_Cat=as.numeric(dfage23$Age_Cat)-1
library("gtsummary")
dfage23 |> select("Depression", "Age_Cat", "Age")|>
  tbl_summary(by = Age_Cat,
              type = list(Depression ~ "continuous", Age ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({sd}); {median}, {min}, {max}"),
              digits = list(all_continuous() ~ c(2,2,0,0)))
```

| **Characteristic** | **2** N = 159[1] | **3** N = 139[1] |
|---|---|---|
| Depression | 5.95 (5.86); 5, 0, 26.00 | 3.71 (3.39); 3, 0, 14.00 |
| Age | 71.06 (3.53); 72, 65, 76.00 | 82.76 (4.63); 82, 77, 97.00 |

[1] Mean (SD); Median, Min, Max

```
total <- nrow(df)
age_summary <- df |>
  group_by(Age_Cat) |>
  summarize(
    Percentage = n()/total*100
  )
age_summary
```

```
## # A tibble: 4 x 2
##   Age_Cat Percentage
##   <fct>        <dbl>
## 1 0             25.5
## 2 1             25.8
## 3 2             26.0
## 4 3             22.7
```

**(C)**

```
#(c)
df<- df |>
  mutate(
    Fatalism_c = (df$Fatalism-mean(df$Fatalism))/IQR(df$Fatalism)
  )



fitmodel <- lm(Depression~Fatalism_c+Sex+R_E+factor(Age_Cat), data=df)
summary(fitmodel)
```

```
##
## Call:
## lm(formula = Depression ~ Fatalism_c + Sex + R_E + factor(Age_Cat),
##     data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.905 -3.368 -1.278  2.076 20.837
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.6002     0.5341  12.358  < 2e-16 ***
## Fatalism_c         2.2117     0.3193   6.927 1.10e-11 ***
## Sex                0.5206     0.4075   1.278  0.20189
## R_E                0.2789     0.4250   0.656  0.51193
## factor(Age_Cat)1  -1.6394     0.5650  -2.901  0.00385 **
## factor(Age_Cat)2  -0.7899     0.5709  -1.384  0.16698
## factor(Age_Cat)3  -3.5238     0.6051  -5.824 9.35e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.987 on 605 degrees of freedom
## Multiple R-squared:  0.1236, Adjusted R-squared:  0.1149
## F-statistic: 14.22 on 6 and 605 DF,  p-value: 3.433e-15
```

```
#Q:Unsure how to interpret results, why does one categorical not show up?
#R automatically uses Age cat as multiple dummy variables when set as factor var.
#group 0 as a reference (compared to the age category 0 on avg age 1 lower depression score on average 
#same increase in mean depression for every unit point increase in age category.



#Mean differences in depression seem to get bigger?
```

The reference group NHW males with ages below the 25th quartile with average fatalism score on average have statistically significant depression score of 6.6 at significance level $\alpha = 0.05$.

Those with 8.25 unit points higher fatalism score on average have 2.2117 points statistically significantly higher depression score at significance level $\alpha = 0.05$.

Females on average have statistically insignificant 0.52 unit points higher depression scores than males at significance level $\alpha = 0.05$.

MA on average have have statistically insignificant 0.2789 unit points higher depression scores than males at significance level $\alpha = 0.05$.

Overall there seems to be general decrease in depression scores for all age groups older than the youngest quartile. The differences were 1.6394, 0.7899. 3.5238 points respectively with the difference being statistically insignificant from age group 0 for age group 2 whilst the other groups 1 and 3 had a statistically significant difference in depression scores from age group 0. There is no clear strictly increasing or decreasing trend in depression score as the age group got older.

**(D)**

```
#Alternatively take the difference in SSR using the full model - SSR using the model with only fatalism



#(d)
anova(fitmodel)
```

```
## Analysis of Variance Table
##
## Response: Depression
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## Fatalism_c        1  1026.2 1026.16 41.2626 2.696e-10 ***
## Sex               1    45.1   45.06  1.8118   0.17879
## R_E               1   115.4  115.45  4.6422   0.03159 *
## factor(Age_Cat)   3   935.4  311.79 12.5372 5.817e-08 ***
## Residuals       605 15045.8   24.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the ANOVA table (Type I), we get $F = 12.537$. it is larger than value at the critical value $F_{3,605} = 2.60$, thus its p value is less than 0.05. Thus there is a statistically significant association between age in categories and depression adjusting for fatalism, sex, and race/ethnicity.

### (e)

```
#(e) as numeric??

#For some reason turning it into numeric adds a 1
df$Age_Cat <- as.numeric(df$Age_Cat)-1
typeof(df$Age_Cat)
```

```
## [1] "double"
```

```
fitmodelc <- lm(Depression~Fatalism_c+Sex+R_E+Age_Cat, data=df)
summary(fitmodelc)
```

```
##
## Call:
## lm(formula = Depression ~ Fatalism_c + Sex + R_E + Age_Cat, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.447 -3.532 -1.223  2.079 21.916
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.5408     0.4919  13.296  < 2e-16 ***
## Fatalism_c    2.0823     0.3200   6.507 1.61e-10 ***
## Sex           0.5074     0.4110   1.235    0.217
## R_E           0.2993     0.4288   0.698    0.485
## Age_Cat      -0.9421     0.1925  -4.894 1.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.033 on 607 degrees of freedom
## Multiple R-squared:  0.1045, Adjusted R-squared:  0.09856
## F-statistic:  17.7 on 4 and 607 DF,  p-value: 9.393e-14
```

```
#bakes in a stronger assumption (ordinality and constant effect for each jump) that every jump in age g

#Not really(?) It just gives a mean change in Depression for every jump??
#Does it give mean differences between mean of each group?
```

For 1 level move up in age category, average depression scores are statistically significantly lower by 0.94

points at a significance level $\alpha = 0.05$. The model makes a strong assumption that it is dealing with ordinal variable with equal spacing between levels, but this is not the case with our age groups differ in range by group 0,1,2,3 thus we cannot make a meaningful interpretation of 0.94 when moving between groups. We can at least make a generic interpretation that depression score generally increases with age.

**(F)**

We conclude that higher age has a statistically significant association with lower depression scores after adjusting for sex, fatalism, and race/ethnicity at a significance level $\alpha = 0.05$ Thus, reject the null hypothesis that there is no trend between higher age categories and differences in depression after adjusting for sex, fatalism, and race/ethnicity. This seems to be due to the much stronger association with lower depression from the oldest age group compared to the other age groups when age was used as a categorical variable in part (c)

**(G)** The coefficients for fatalism and race/ethnicity changes the most when age category is used instead of continuous age. The 5% difference is pretty much harmless but could potentially be damaging in other situations. This probably happened because categorizing age introduced a loss of information to the model. Specifically, the information loss likely comes from the fact that the moment different observations with different ages were collapsed into the same age grou,p the variation in Depression explained by the increment in age between those two observations (ex) 25 years to 30 years old) would have been lost and explanation of variation was taken up by the variables centered Fatalism and gender.

```r
df<- df |> mutate(Fatalism_c = (df$Fatalism-mean(df$Fatalism))/IQR(df$Fatalism))

#Ensure all models have the exact same covariate order or the comparison will produce trash!
fitmodelcont <- lm(Depression ~ Fatalism_c + Sex + R_E + Age, data = df)
coef1 = summary(fitmodel)$coefficients
coef2 = summary(fitmodelc)$coefficients
coef3 = summary(fitmodelcont)$coefficients


summaries = function(x) {
  out = data.frame(t(cbind(coef1[x,], coef2[x, ], coef3[x,])))
  out$"Percentage difference from coef where age is continuous" = round((out$Estimate - out$Estimate[3])
  colnames(out)[1:4] = c(dimnames(coef1)[[2]])
  rownames(out) = c("Age cat dummy", "Age ordinal", "Age continous")
  return(out)
}


Fatalism_c.summ = summaries(2)
R_E.summ = summaries(3)
Sex.summ = summaries(4)

knitr::kable(Fatalism_c.summ, caption = "Fatalism_c")
```

Table 3: Fatalism_c

| | Estimate | Std. Error | t value | Pr(>\|t\|) | Percentage difference from coef where age is continuous |
|---|---|---|---|---|---|
| Age cat dummy | 2.211663 | 0.3192949 | 6.926709 | 0 | 5.25 |
| Age ordinal | 2.082275 | 0.3200265 | 6.506571 | 0 | -0.91 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| Age continous | 2.101332 | 0.3198352 | 6.570045 | 0 | 0.00 |

```
knitr::kable(R_E.summ, caption = "R_E")
```

Table 4: R_E

|  | Estimate | Std. Error | t value | Pr(>\|t\|) | Percentage difference from coef where age is continuous |
|---|---|---|---|---|---|
| Age cat dummy | 0.5205844 | 0.4074793 | 1.277572 | 0.2018902 | 1.04 |
| Age ordinal | 0.5074068 | 0.4109512 | 1.234713 | 0.2174151 | -1.51 |
| Age continous | 0.5152018 | 0.4105187 | 1.255002 | 0.2099610 | 0.00 |

```
knitr::kable(Sex.summ, caption = "Sex")
```

Table 5: Sex

|  | Estimate | Std. Error | t value | Pr(>\|t\|) | Percentage difference from coef where age is continuous |
|---|---|---|---|---|---|
| Age cat dummy | 0.2789195 | 0.4250385 | 0.6562218 | 0.5119309 | -5.32 |
| Age ordinal | 0.2992889 | 0.4288257 | 0.6979269 | 0.4854902 | 1.59 |
| Age continous | 0.2945994 | 0.4277096 | 0.6887837 | 0.4912226 | 0.00 |

### (h)

```
coef_agecat <- summary(fitmodelc)$coefficients['Age_Cat',]
df<- df |>
  mutate(Fatalism_c = (df$Fatalism-mean(df$Fatalism))/IQR(df$Fatalism))

df$Age_Cat10 = 10 * (as.numeric(df$Age_Cat))
fit4 <- lm(Depression ~ Fatalism_c + R_E + Sex + Age_Cat10, data = df)
coef_agecat10 = summary(fit4)$coefficients['Age_Cat10',]

df$Age_Cat300 = df$Age_Cat10
df$Age_Cat300[which(df$Age_Cat10 == 30)] = 300
fit5 <- lm(Depression ~ Fatalism_c + R_E + Sex + Age_Cat300, data = df)
coef_agecat300 = summary(fit5)$coefficients['Age_Cat300',]

meanscore <- by(df$Age, df$Age_Cat, mean)
df$AgeMean = meanscore[1]
df$AgeMean[df$Age_Cat==1] = meanscore[2]
df$AgeMean[df$Age_Cat==2] = meanscore[3]
df$AgeMean[df$Age_Cat==3] = meanscore[4]
fit6 <- lm(Depression ~ Fatalism_c + R_E + Sex + AgeMean, data = df)
coef_agemean = summary(fit6)$coefficients['AgeMean', ]
```

```r
age_summary2 <- rbind(coef_agecat, coef_agecat10, coef_agecat300, coef_agemean)
row.names(age_summary2) = c(
  "Original 0,1,2,3",
  "scale by 10",
  "0,10,20,300",
  "Mean age in category"
)

knitr::kable(age_summary2)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---:|---:|---:|---:|
| Original 0,1,2,3 | -0.9421274 | 0.1925099 | -4.893916 | 1.3e-06 |
| scale by 10 | -0.0942127 | 0.0192510 | -4.893916 | 1.3e-06 |
| 0,10,20,300 | -0.0093808 | 0.0017213 | -5.449837 | 1.0e-07 |
| Mean age in category | -0.0902048 | 0.0183707 | -4.910264 | 1.2e-06 |

In all different scores, the p values were statistically significant at a significance level $\alpha = 0.05$ for which the p values were $1.3 * 10^{-6}, 1.0 * 10^{-7}, 1.2 * 10^{-6}$ respectively. The original scores multiplied by 10 reports the exact same t-values whilst the other modified scores do not. This can be attributed to the fact that the scaling increased both beta estimate and standard error by a factor of 10 exclusively for that variable. The modified variables have lower variance than the original. In particular, for the age variable which had a 300 added in the place of 30 we can see that the standard error has decreased dramatically since we have artificially introduced a larger variance which affects the standard error $\sqrt{\sigma^2(X^TX)^{-1}}$, subsequently giving it the highest absolute t-value and the lowest p-value as well.

## 2.

(a)

$\beta_0$: Average SBP of non smoking 55 year olds with Quatelet index 3.5

$\beta_1$: Adjusted age effect among non smokers with Quatelet index 3.5 when $S_i = 0, Q_i^c = 0$.

$\beta_2$: Adjusted effect of Quatelet index for 55 year old non smokers.

$\beta_3$: Adjusted effect of smoking for those aged 55 year old and Quatelet index 3.5

$\beta_4$: Mean change in adjusted age effect for 1 unit higher Quatelet index or mean change in adjusted Quatelet index effect for 1 unit higher age.

$\beta_5$: Mean change in adjusted age effect comparing smokers to non smokers or Mean change in adjusted smoking effect for 1 unit change in age.

(b)

$$SBP_i = \beta_0 + (\beta_1 + \beta_4 Q_i^c + \beta_5 S_i)A_i^c + \beta_2 Q_i^c + \beta_3 S_i + \epsilon_i$$

(c)

$$SBP_i = \beta_0 + \beta_1 A_i + (\beta_2 + \beta_4 A_i)Q^c + \beta_3 S_i + \beta_5 A_i^c S_i + \epsilon_i$$

(d)

$$SBP_i = \beta_0 + \beta_1 A_i + \beta_2 Q_i^c + (\beta_3 + \beta_5 A_i)S_i + \beta_4 A_i^c Q_i^c + \epsilon_i$$

### 3.

(a)

$$Depression_i = \beta_0 + \beta_1 Fatalism_i^c + \beta_2 Sex_i + \beta_3 Sex_i Fatalism_i^c + \beta_4 RE_i + \beta_5 A_i^c + \epsilon$$

Where

$$Fatalism_i^c = \frac{Fatalism_i - 17}{8.5}$$

and

$$Age_i^c = \frac{A_i - 65.88562}{10}.$$

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

**(b)** $\hat{\beta}_3 = 1.325, t = 2.014, df = 606, p = 0.044 < 0.05$ Reject the null hypothesis.

```
df <- mySASData[ ,var]
df$Fatalism_c <- ((df$Fatalism-17)/8.5)
df$Age10 <-(df$Age-mean(df$Age))/10
#scaled age age category numerical

model<-lm(Depression~Fatalism_c+Sex+Sex*Fatalism_c+R_E+Age10, data=df)
summary(model)
```

```
##
## Call:
## lm(formula = Depression ~ Fatalism_c + Sex + Sex * Fatalism_c +
##     R_E + Age10, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.487 -3.433 -1.200  1.920 21.832
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.0506     0.3558  14.197  < 2e-16 ***
## Fatalism_c       1.5459     0.4500   3.435 0.000633 ***
## Sex              0.4351     0.4114   1.058 0.290672
## R_E              0.2995     0.4266   0.702 0.482892
## Age10           -0.8954     0.1751  -5.113 4.26e-07 ***
## Fatalism_c:Sex   1.3251     0.6579   2.014 0.044429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.014 on 606 degrees of freedom
## Multiple R-squared:  0.1124, Adjusted R-squared:  0.1051
## F-statistic: 15.35 on 5 and 606 DF,  p-value: 3.176e-14
```

**(c)** $\hat{\beta}_1 = 1.546, \hat{\beta}_3 = 1.325$

$\text{Var}(\hat{\beta}_1) = 0.2025, \text{Var}(\hat{\beta}_3) = 0.4328$

$CoVar(\hat{\beta}_1, \hat{\beta}_3) = -0.2022$

$SE(\hat{\beta}_1, \hat{\beta}_3) = \sqrt{\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_3) + 2 * CoVar(\hat{\beta}_1, \hat{\beta}_3)} = 0.4805$

$CI_{0.95} = \hat{\beta}_1 + \hat{\beta}_3 \pm t_{0.975,606} * SE(\hat{\beta}_1, \hat{\beta}_3) = (1.9273, 3.8146)$

females 2.871(1.9273,3.8146) and among males 1.5459(1.9273,3.8146).

```
vcov(model)
```

```
##                   (Intercept)    Fatalism_c          Sex           R_E
## (Intercept)      0.1265653169 -0.0018982591 -0.068335901 -0.0948484921
## Fatalism_c      -0.0018982591  0.2025331466  0.001280368  0.0009665405
## Sex             -0.0683359008  0.0012803683  0.169262563 -0.0168964835
## R_E             -0.0948484921  0.0009665405 -0.016896484  0.1820248016
## Age10           -0.0104410427 -0.0014154775 -0.001545174  0.0202702084
## Fatalism_c:Sex  0.0005299621 -0.2022267678 -0.026162379  0.0016146617
##                        Age10 Fatalism_c:Sex
## (Intercept)     -0.010441043   0.0005299621
## Fatalism_c      -0.001415477  -0.2022267678
## Sex             -0.001545174  -0.0261623787
## R_E              0.020270208   0.0016146617
## Age10            0.030671976  -0.0037570956
## Fatalism_c:Sex -0.003757096   0.4328082791
```

### (d)

The association between fatalism and depression among females is significantly different from that among males with p value 0.0444 at a significance level 0.05. Adjusting for age and race/ethnicity, among females, the average depression score is 2.8709 points higher per 8.5 points increase in fatalism score with a confidence interval (1.9273,3.8146) at confidence level 95%.