STAT 224 Autumn 2022 HW4

Matthew Zhao

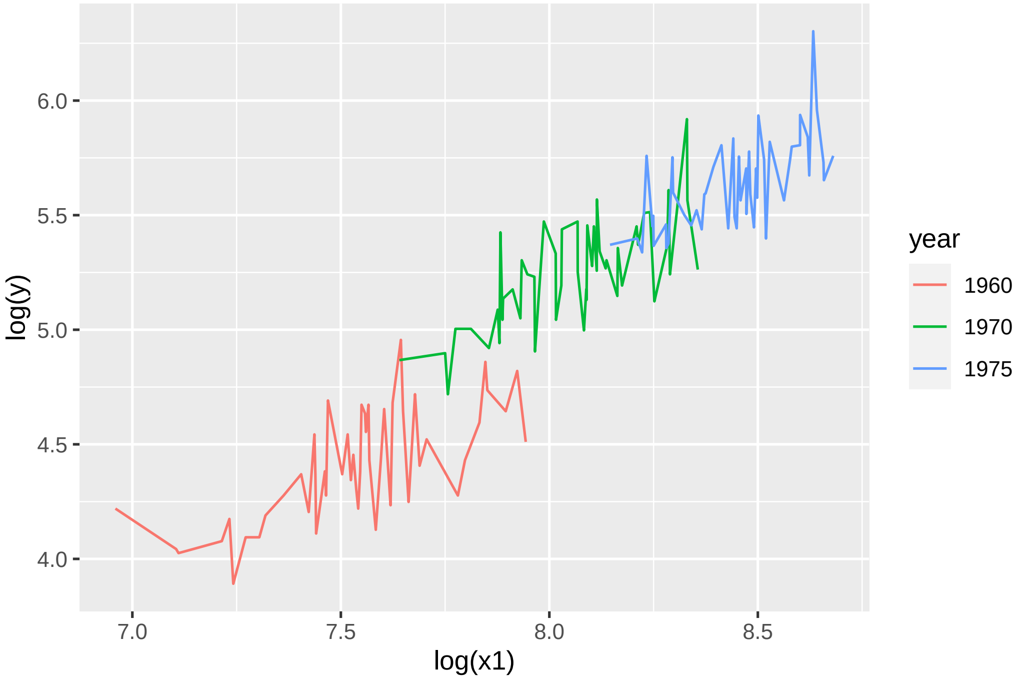
Question 1

http://www.stat.uchicago.edu/~yibi/s224/data/P151-153.txt

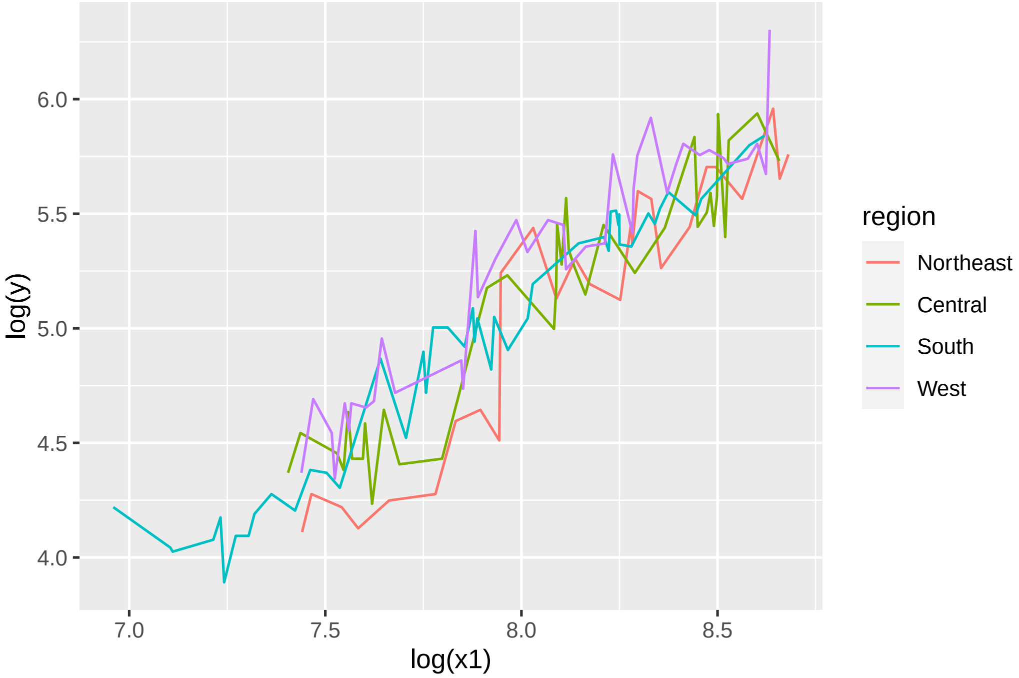
```
eduexp = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/P151-153.txt", header=T)
eduexp$region = factor(eduexp$region, labels=c("Northeast","Central","South","West"))
eduexp$year = as.factor(eduexp$year)
```

Q1a — 4 points

```
ggplot(data = eduexp, mapping = aes(x=log(x1), y=log(y), color = year)) +
  geom_line()
```



```
ggplot(data = eduexp, mapping = aes(x=log(x1), y=log(y), color = region)) +
  geom_line()
```



Q1b — 5 points

```
lm1 = lm(log(y) ~ log(x1) + year + region + log(x1)*year + log(x1)*region ,data = eduexp)
lm2 = lm(log(y) ~ log(x1) + year + region + log(x1)*region,data = eduexp)
anova(lm2,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(y) ~ log(x1) + year + region + log(x1) * region
## Model 2: log(y) ~ log(x1) + year + region + log(x1) * year + log(x1) *
##     region
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    140 2.61
## 2    138 2.58  2     0.031 0.83   0.44
```

F-stat of 0.83 with df 2, 138 and p-value of 0.44. At the 5% significance level, we fail to reject the null and conclude that the slope for log(x1) does not vary with year i.e. the interaction term is insignificant.

Q1c — 5 points

```
lm1 = lm(log(y) ~ log(x1) + year + region + log(x1)*year + log(x1)*region, data = eduexp)
lm2 = lm(log(y) ~ log(x1) + year + region + log(x1)*year, data = eduexp)
anova(lm2,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(y) ~ log(x1) + year + region + log(x1) * year
## Model 2: log(y) ~ log(x1) + year + region + log(x1) * year + log(x1) *
##     region
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    141 2.73
## 2    138 2.58  3     0.149 2.65  0.051 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-stat of 2.65 with df 3, 138 and p-value of 0.051. At the 5% significance level, we fail to reject the null and conclude that the slope for log(x1) does not vary with region i.e. the interaction term is insignificant.

Q1d — 4 points

```r
lm1 = lm(log(y) ~ log(x1) + year + region, data = eduexp)
lm2 = lm(log(y) ~ 1, data = eduexp)
anova(lm2,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(y) ~ 1
## Model 2: log(y) ~ log(x1) + year + region
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1    149 46.1
## 2    143  2.8  6      43.4 375 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-stat of 375 with df 6, 143 and p-value of less than 2e-16. At the 5% significance level, we reject the null and conclude that log(x1) has an effect on log(y) after accounting for year and regoin assuming no interactions.

Question 2

http://www.stat.uchicago.edu/~yibi/s224/data/NLSY.txt

```r
NLSY = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/NLSY.txt", header=T)
```

Q2a — 3 points

```
lm1 = lm(log(Income2005) ~ AFQT + Gender + AFQT*Gender, data = NLSY)
lm2 = lm(log(Income2005) ~ AFQT + Gender, data = NLSY)
anova(lm2,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: log(Income2005) ~ AFQT + Gender
## Model 2: log(Income2005) ~ AFQT + Gender + AFQT * Gender
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   2581 1997
## 2   2580 1996  1      1.21 1.56   0.21
```

F-stat of 1.56 with df 1, 2580 and p-value of 0.21. At the 5% significance level, we fail to reject the null and conclude that the effect of AFQT on log(Income2005) does not vary with Gender i.e. the interaction is insignificant.

Model A:

```
modela = lm(log(Income2005) ~ AFQT + Gender + Edu2006, data=NLSY)
```

Model B:

```
modelb = lm(log(Income2005) ~ AFQT + Gender + as.factor(Edu2006), data=NLSY)
```

Model C:

```
NLSY$edu5grps = cut(NLSY$Edu2006, c(5.5,11.5,12.5,15.5,16.5,20.5))
NLSY$edu5grps =
  factor(NLSY$edu5grps,
         labels=c("6-11", "12", "13-15", "16", "17+"))
modelc = lm(log(Income2005) ~ AFQT + Gender + edu5grps, data=NLSY)
```

Model D

```
NLSY$eduscore1 = as.numeric(NLSY$edu5grps)
modeld = lm(log(Income2005) ~ AFQT + Gender + eduscore1, data=NLSY)
```

Model E

```
conversion = c("6-11"=9, "12"=12, "13-15"=13, "16"=16, "17+"=17)
NLSY$eduscore2 = conversion[NLSY$edu5grps]
modele = lm(log(Income2005) ~ AFQT + Gender + eduscore2, data=NLSY)
```

Q2b — 8 points

i)

Model A is nested inside Model B because ordinal models are always nested inside their nominal versions. This is because we can set the coefficients of the nominal predictors/indicators such that they are the same as those of the ordinal model. Hence, ordinal models are nested inside their respective nominal versions.

ii)

```
print(modela$coefficients)
```

```
## (Intercept)        AFQT  Gendermale     Edu2006
##    1.823456    0.005914    0.624509    0.076951
```

```
print(modelb$coefficients)
```

```
##          (Intercept)                 AFQT           Gendermale
##             2.468925             0.005595             0.628039
##  as.factor(Edu2006)7  as.factor(Edu2006)8  as.factor(Edu2006)9
##             0.362556             0.267368            -0.243692
## as.factor(Edu2006)10 as.factor(Edu2006)11 as.factor(Edu2006)12
##             0.016654             0.068119             0.285207
## as.factor(Edu2006)13 as.factor(Edu2006)14 as.factor(Edu2006)15
##             0.490078             0.392739             0.403398
## as.factor(Edu2006)16 as.factor(Edu2006)17 as.factor(Edu2006)18
##             0.676220             0.678429             0.887308
## as.factor(Edu2006)19 as.factor(Edu2006)20
##             0.540815             0.862429
```

With model a, the positive coefficient on Edu2006 indicates that log(Income2005) increases with years of education. With model b, we see mostly positive coefficients, with Edu(2006) for 9th grade being the only negative coefficient. Additionally, the magnitude of the coefficients generally rises with higher grades, indicating that more years of education is beneficial, similar to model a. The notable exception besides grade 9 is grade 19, which could potentially be related to the duration of different advanced degree programs.

iii)

Yes. The magnitude of the education generally increases with the grade level of the indicator, meaning that more years of education is associated with higher log(Income)

iv)

```
anova(modela,modelb)
```

```
## Analysis of Variance Table
##
## Model 1: log(Income2005) ~ AFQT + Gender + Edu2006
## Model 2: log(Income2005) ~ AFQT + Gender + as.factor(Edu2006)
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   2580 1935
## 2   2567 1912 13      23.1 2.38 0.0035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null at the 5% significance level and can conclude that the nominal model of education is true over the ordinal model i.e. that there are different effects for each year of education.

Q2c — 8 points

i)

Yes because if the coefficients of the indicators for 6-11, 13-15, and 17+ in model b are the same, then model b becomes model c. Since model c is a special case of model b, we can say that model c is nested in model b.

ii)

```
print(modelc$coefficients)
```

```
##   (Intercept)          AFQT    Gendermale    edu5grps12 edu5grps13-15
##      2.488042      0.005447      0.631062      0.270707      0.416167
##    edu5grps16   edu5grps17+
##      0.666523      0.763394
```

```
print(modelb$coefficients)
```

```
##          (Intercept)                 AFQT           Gendermale
##             2.468925             0.005595             0.628039
##  as.factor(Edu2006)7  as.factor(Edu2006)8  as.factor(Edu2006)9
##             0.362556             0.267368            -0.243692
## as.factor(Edu2006)10 as.factor(Edu2006)11 as.factor(Edu2006)12
##             0.016654             0.068119             0.285207
## as.factor(Edu2006)13 as.factor(Edu2006)14 as.factor(Edu2006)15
##             0.490078             0.392739             0.403398
## as.factor(Edu2006)16 as.factor(Edu2006)17 as.factor(Edu2006)18
##             0.676220             0.678429             0.887308
## as.factor(Edu2006)19 as.factor(Edu2006)20
##             0.540815             0.862429
```

The difference is that in model c, higher education category means larger increase in income while for model b some education categories that are higher than others e.g. grade 19 have less of an impact than lower categories e.g. grade 19 vs grade 17.

iii)

```
anova(modelc,modelb)
```

```
## Analysis of Variance Table
##
## Model 1: log(Income2005) ~ AFQT + Gender + edu5grps
## Model 2: log(Income2005) ~ AFQT + Gender + as.factor(Edu2006)
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   2577 1924
## 2   2567 1912 10      11.8 1.58   0.11
```

We fail to reject the null at the 5% significance level and conclude that the full model with indicators for every grade level is not true i.e. the reduced model with grade clusters is true.

Q2d — 9 points

i)

Yes they are since they are ordinal versions of model c. This means that models d and e are special cases of model c where the coefficients of model c are proportional to each other.

ii)

```
print(modelc$coefficients)
```

```
##   (Intercept)          AFQT    Gendermale    edu5grps12 edu5grps13-15
##      2.488042      0.005447      0.631062      0.270707      0.416167
##    edu5grps16   edu5grps17+
##      0.666523      0.763394
```

```
print(modeld$coefficients)
```

```
## (Intercept)        AFQT  Gendermale   eduscore1
##    2.380142    0.005647    0.630915    0.177672
```

```
print(modele$coefficients)
```

```
## (Intercept)        AFQT  Gendermale   eduscore2
##    1.625785    0.005551    0.627988    0.095544
```

Model c predicts increasing returns to education, where higher education buckets earn more than lower education buckets in increasing increments. model d and e predict constant returns to education, with model e predicting smaller returns for each additional unit of education.

iii)

```
anova(modeld,modelc)
```

```
## Analysis of Variance Table
##
## Model 1: log(Income2005) ~ AFQT + Gender + eduscore1
## Model 2: log(Income2005) ~ AFQT + Gender + edu5grps
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   2580 1926
## 2   2577 1924  3      2.64 1.18   0.32
```

```
anova(modele,modelc)
```

```
## Analysis of Variance Table
##
## Model 1: log(Income2005) ~ AFQT + Gender + eduscore2
## Model 2: log(Income2005) ~ AFQT + Gender + edu5grps
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   2580 1925
## 2   2577 1924  3     0.975 0.44   0.73
```

We can conclude that the full model of model c is not true since we fail to reject the null at the 5% significance level against both models.

Q2e — 4 points

```
summary(modelb)
```

```
##
## Call:
## lm(formula = log(Income2005) ~ AFQT + Gender + as.factor(Edu2006),
##     data = NLSY)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.796 -0.332  0.136  0.504  2.540
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.468925   0.498808    4.95  7.9e-07 ***
## AFQT                 0.005595   0.000778    7.19  8.2e-13 ***
## Gendermale           0.628039   0.034417   18.25  < 2e-16 ***
## as.factor(Edu2006)7  0.362556   0.630528    0.58    0.565
## as.factor(Edu2006)8  0.267368   0.543034    0.49    0.623
## as.factor(Edu2006)9 -0.243692   0.522670   -0.47    0.641
## as.factor(Edu2006)10 0.016654   0.518886    0.03    0.974
## as.factor(Edu2006)11 0.068119   0.514488    0.13    0.895
## as.factor(Edu2006)12 0.285207   0.499993    0.57    0.568
## as.factor(Edu2006)13 0.490078   0.503079    0.97    0.330
## as.factor(Edu2006)14 0.392739   0.502811    0.78    0.435
## as.factor(Edu2006)15 0.403398   0.505810    0.80    0.425
## as.factor(Edu2006)16 0.676220   0.503268    1.34    0.179
## as.factor(Edu2006)17 0.678429   0.509681    1.33    0.183
## as.factor(Edu2006)18 0.887308   0.506870    1.75    0.080 .
## as.factor(Edu2006)19 0.540815   0.513983    1.05    0.293
## as.factor(Edu2006)20 0.862429   0.511990    1.68    0.092 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.863 on 2567 degrees of freedom
## Multiple R-squared:  0.219,  Adjusted R-squared:  0.215
## F-statistic: 45.1 on 16 and 2567 DF,  p-value: <2e-16
```

```
lm1 = lm(log(Income2005) ~ AFQT + Gender, data=NLSY)
anova(lm1, modelb)
```

```
## Analysis of Variance Table
##
## Model 1: log(Income2005) ~ AFQT + Gender
## Model 2: log(Income2005) ~ AFQT + Gender + as.factor(Edu2006)
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   2581 1997
## 2   2567 1912 14      84.7 8.12 <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

This is because individually, the education indicators are not significant e.g. if you just added an indicator for grade 7 it would not be significant. However, collectively when you add indicators for each grade level the full model is significantly different than the reduced model without the education indicators. I illustrate this below.

```
temp = NLSY %>%
  mutate(grade7 = ifelse(Edu2006 == 7, 1, 0))
lm7 = lm(log(Income2005) ~ AFQT + Gender + grade7, data=temp)
summary(lm7)
```

```
##
## Call:
## lm(formula = log(Income2005) ~ AFQT + Gender + grade7, data = temp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.753 -0.355  0.130  0.523  2.566
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.676145   0.041576   64.37   <2e-16 ***
## AFQT        0.010070   0.000625   16.10   <2e-16 ***
## Gendermale  0.604210   0.034648   17.44   <2e-16 ***
## grade7      0.115336   0.394906    0.29     0.77
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.88 on 2580 degrees of freedom
## Multiple R-squared:  0.185,  Adjusted R-squared:  0.184
## F-statistic:  195 on 3 and 2580 DF,  p-value: <2e-16
```

```
anova(lm1,lm7)
```

```
## Analysis of Variance Table
##
## Model 1: log(Income2005) ~ AFQT + Gender
## Model 2: log(Income2005) ~ AFQT + Gender + grade7
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   2581 1997
## 2   2580 1997  1     0.066 0.09   0.77
```