[STAT 4400] HW-3

Michael Ghattas

2/15/2022

Problem 1

```
require(AER)

## Loading required package: AER

## Loading required package: car

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.2

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

require(arm)

## Loading required package: arm

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 4.1.2

## Loading required package: Matrix

## Loading required package: lme4

## Warning: package 'lme4' was built under R version 4.1.2
```

```
##
## arm (Version 1.12-2, built: 2021-10-15)
```

```
## Working directory is /Users/Home/Desktop
```

```
##
## Attaching package: 'arm'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
require(foreign)
```

```
## Loading required package: foreign
```

```
## Warning: package 'foreign' was built under R version 4.1.2
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
df <- read.dta("/Users/Home/Documents/Michael_Ghattas/School/CU_Boulder/2022/
Spring 2022/STAT - 4400/Data/risky_behaviors.dta", convert.factors = TRUE)
df$fupacts <- round(df$fupacts)
df$couples <- factor(df$couples)
df$women_alone <- factor(df$women_alone)
```

```
summary(df)
```

```
##     sex      couples women_alone      bs_hiv       bupacts
##  woman:217   0:272   0:288       negative:337   Min.   :  0.00
##  man  :217   1:162   1:146       positive: 97   1st Qu.:  5.00
##                                                 Median : 15.00
##                                                 Mean   : 25.91
##                                                 3rd Qu.: 36.00
##                                                 Max.   :300.00
##     fupacts
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :  5.00
##  Mean   : 16.49
##  3rd Qu.: 21.00
##  Max.   :200.00
```

(a)

```
poi.reg <- glm(fupacts ~ women_alone, family=poisson, data = df)
summary(poi.reg)
```

```
## 
## Call:
## glm(formula = fupacts ~ women_alone, family = poisson, data = df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -6.093  -4.979  -3.304   1.237  27.150  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   2.92114    0.01368  213.58   <2e-16 ***
## women_alone1 -0.40367    0.02719  -14.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 13064  on 432  degrees of freedom
## AIC: 14393
## 
## Number of Fisher Scoring iterations: 6
```

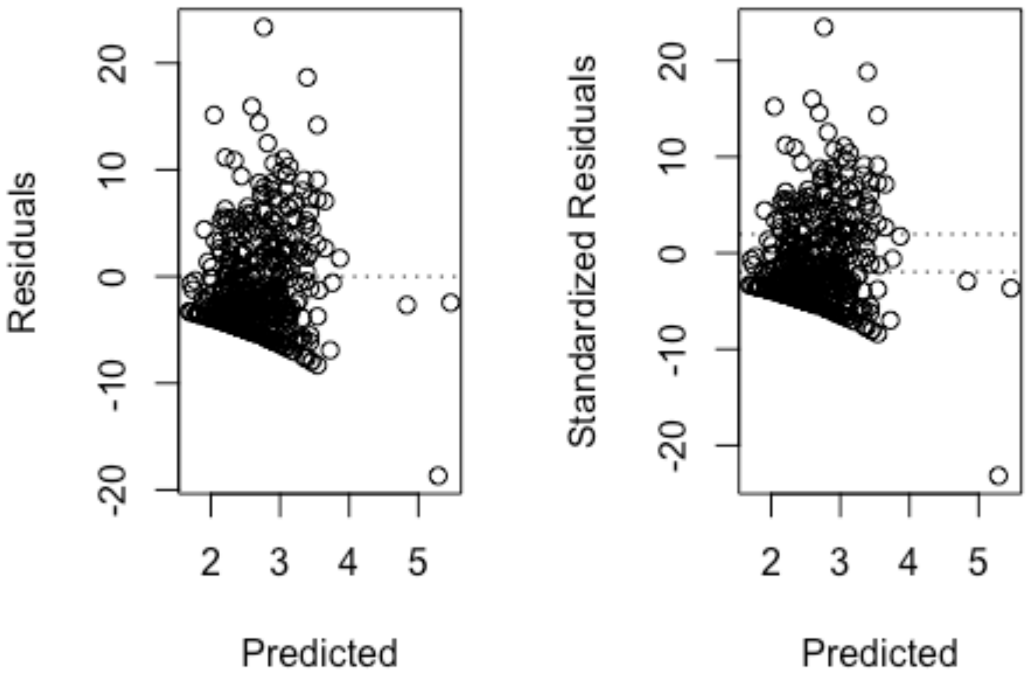
The model is a poor fit, even with the woman_alone factor having a statistical significance.

(b)

```
df$c.bupacts <- (df$bupacts - mean(df$bupacts)) / (2 * sd(df$bupacts))
poi.reg.ext <- glm(fupacts ~ women_alone + sex + c.bupacts + couples + 
bs_hiv, family = poisson, data = df)
summary(poi.reg.ext)
```

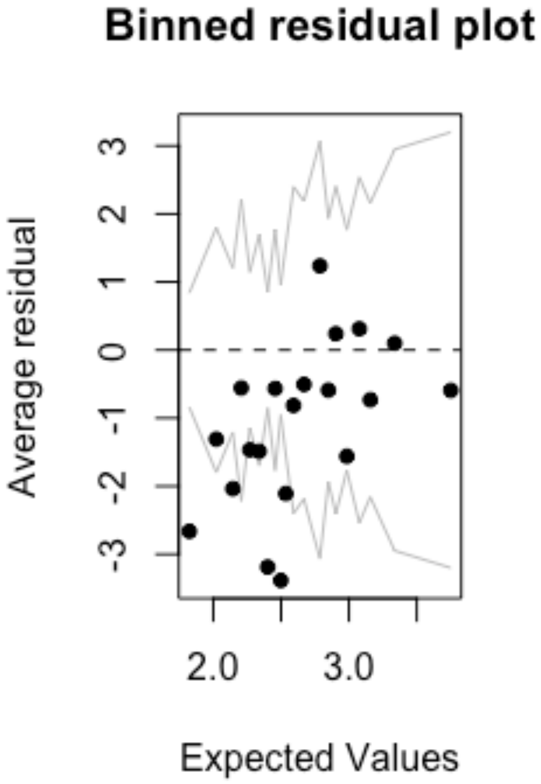
```
## 
## Call:
## glm(formula = fupacts ~ women_alone + sex + c.bupacts + couples + 
##     bs_hiv, family = poisson, data = df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -18.679   -4.305   -2.511    1.368   23.361  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)     3.17508    0.02256 140.721  < 2e-16 ***
## women_alone1   -0.66222    0.03090 -21.434  < 2e-16 ***
## sexman         -0.10867    0.02373  -4.579 4.66e-06 ***
## c.bupacts       0.68808    0.01110  62.013  < 2e-16 ***
## couples1       -0.40998    0.02823 -14.523  < 2e-16 ***
## bs_hivpositive -0.43832    0.03538 -12.389  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 10200  on 428  degrees of freedom
## AIC: 11537
## 
## Number of Fisher Scoring iterations: 6
```

```
par(mfrow = c(1,2))
plot(predict(poi.reg.ext), residuals(poi.reg.ext), xlab = "Predicted", ylab = 
"Residuals")
abline(a = 0, b = 0, lty = 3)
plot(predict(poi.reg.ext), rstandard(poi.reg.ext), xlab = "Predicted", ylab = 
"Standardized Residuals")
abline(a = 1.96, b = 0, lty = 3)
abline(a = -1.96, b = 0, lty = 3)
```



```
binnedplot(predict(poi.reg.ext), rstandard(poi.reg.ext))
dispersiontest(poi.reg.ext, trafo = 1)

## 
##  Overdispersion test
## 
## data:  poi.reg.ext
## z = 5.5689, p-value = 1.282e-08
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha 
## 28.65146
```

```
yhat <- predict (poi.reg.ext, type = "response")
z <- (df$fupacts-yhat) / sqrt(yhat)
n = poi.reg.ext$df.null + 1
k = poi.reg.ext$df.null + 1 - poi.reg.ext$df.residual
cat("overdispersion ratio is ", sum(z^2) / (n-k), "\n")

## overdispersion ratio is  30.00404

cat("p-value of overdispersion test is ", pchisq(sum(z^2), n-k), "\n")

## p-value of overdispersion test is  1
```



The estimated overdispersion is extremely high at 28.65, with over-dispersion ratio at 30.00404, and a p-value of over-dispersion test at 1.

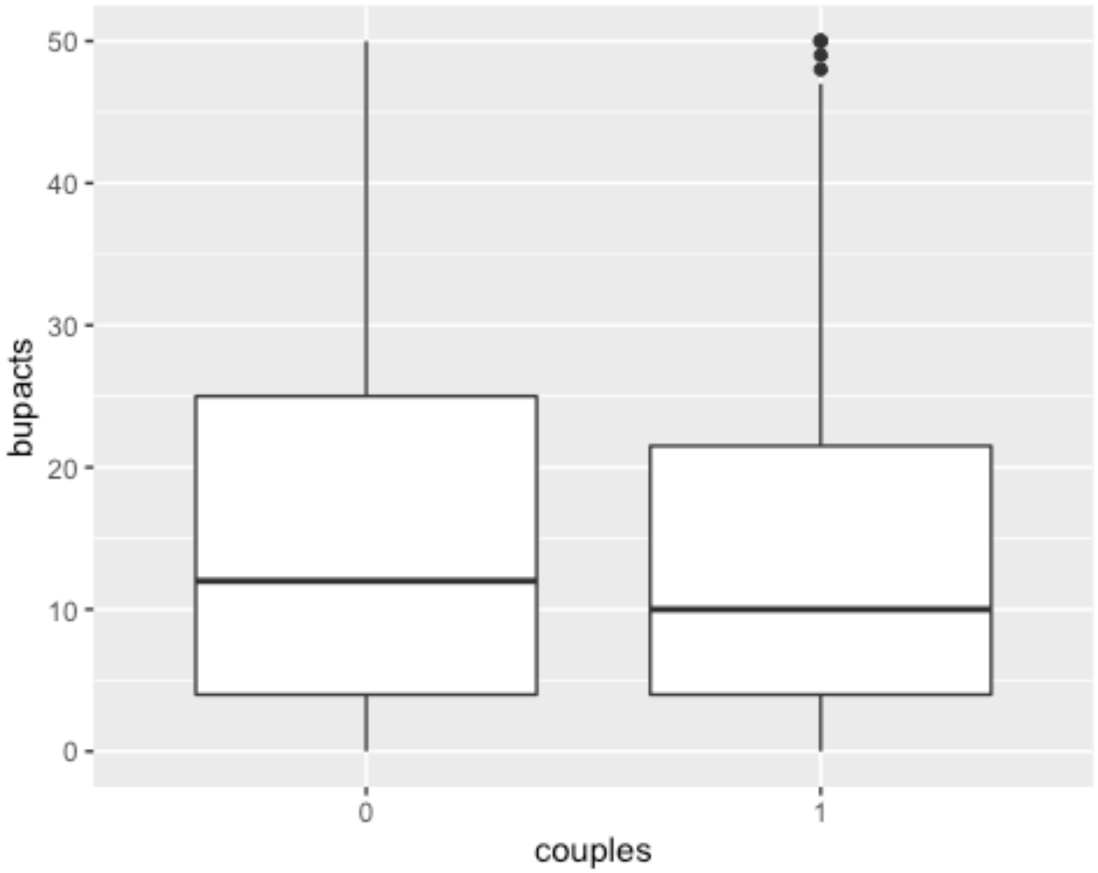
(c)

```
df$c.bupacts <- (df$bupacts - mean(df$bupacts)) / (2 * sd(df$bupacts))
poi.reg.ext <- glm(fupacts ~ women_alone + sex + c.bupacts + couples +
bs_hiv, family = quasipoisson, data = df)
display(poi.reg.ext)
```

```
## glm(formula = fupacts ~ women_alone + sex + c.bupacts + couples +
##     bs_hiv, family = quasipoisson, data = df)
##                  coef.est coef.se
## (Intercept)       3.18     0.12
## women_alone1     -0.66     0.17
## sexman           -0.11     0.13
## c.bupacts         0.69     0.06
## couples1         -0.41     0.15
## bs_hivpositive   -0.44     0.19
## ---
##   n = 434, k = 6
##   residual deviance = 10200.4, null deviance = 13298.6 (difference =
3098.2)
##   overdispersion parameter = 30.0
```

```
ggplot(data=df, aes(x = couples, y = bupacts)) + geom_boxplot() + ylim(0, 50)
```

```
## Warning: Removed 63 rows containing non-finite values (stat_boxplot).
```



```
df$offset <- ifelse(df$bupacts == 0, 1, df$bupacts)
poi.reg.off <- glm(fupacts ~ women_alone + sex + couples + bs_hiv, offset =
log(offset), family = quasipoisson, data = df)
display(poi.reg.off)

## glm(formula = fupacts ~ women_alone + sex + couples + bs_hiv,
##     family = quasipoisson, data = df, offset = log(offset))
##                coef.est coef.se
## (Intercept)    -0.03     0.15
## women_alone1   -0.55     0.21
## sexman         -0.12     0.16
## couples1       -0.41     0.19
## bs_hivpositive -0.31     0.24
## ---
##   n = 434, k = 5
##   residual deviance = 10195.0, null deviance = 10736.5 (difference =
541.5)
##   overdispersion parameter = 46.6
```

Singles tends to have unprotected sex more often than couples. We fit a Poisson model with the number of unprotected sex acts reported at the baseline as an offset.

###(d)

Yes it should! Observations coming from the elements of couples is not i.i.d. THis yields an extremely high positive correlation between the answers of individuals that are a part of a couple.

Problem 2

```
require(arm)
require(foreign)
require(MASS)

df <- read.dta("/Users/Home/Documents/Michael_Ghattas/School/CU_Boulder/2022/
Spring 2022/STAT - 4400/Data/nes5200_processed_voters_realideo.dta")

df$partyid3 <- factor(df$partyid3, labels = c("democrats", "independents",
"republicans", "apolitical"))
df$gender <- factor(df$gender, labels = c("male", "female"))
df$race <- factor(df$race, labels = c("white", "black", "asian", "native
american", "hispanic", "other"))
df$south <- factor(df$south)
df$ideo <- factor(df$ideo, labels = c("liberal", "moderate", "conservative"))

x = df$partyid3
df <- df[!is.na(levels(x)[x]),]

df <- subset(df, partyid3 != "apolitical")
df$partyid3 <- factor(df$partyid3)
```

(a)

```
multi.log <- polr(partyid3 ~ ideo + race + age_10, Hess = TRUE, data = df)
summary(multi.log)
```

```
## Call:
## polr(formula = partyid3 ~ ideo + race + age_10, data = df, Hess = TRUE)
##
## Coefficients:
##                       Value Std. Error  t value
## ideomoderate          1.0923    0.05183  21.0738
## ideoconservative      2.0209    0.04449  45.4226
## raceblack            -2.0887    0.07266 -28.7455
## raceasian             0.2056    0.14655   1.4030
## racenative american -0.4204    0.10648  -3.9483
## racehispanic         -0.9211    0.07610 -12.1030
## raceother            -0.3989    0.48895  -0.8159
## age_10               -0.1147    0.01037 -11.0537
##
## Intercepts:
##                          Value    Std. Error t value
## democrats|independents     0.4669   0.0581     8.0385
## independents|republicans   0.8959   0.0585    15.3225
##
## Residual Deviance: 23593.16
## AIC: 23613.16
## (25245 observations deleted due to missingness)
```

(b)

```
confint(multi.log)
```

```
## Waiting for profiling to be done...
```

```
##                            2.5 %      97.5 %
## ideomoderate          0.99088703  1.19409012
## ideoconservative      1.93404008  2.10845161
## raceblack            -2.23286503 -1.94793720
## raceasian            -0.08077981  0.49403605
## racenative american -0.62975657 -0.21225583
## racehispanic         -1.07107585 -0.77272060
## raceother            -1.37398545  0.56561431
## age_10               -0.13502245 -0.09436586
```

age_10: For a one unit increase in age we expect a -0.11 increase in the expected value of partyid3. ideo: moderates and especially conservatives are more likely to be republicans. In particular. race: whites, and asianes are more likely to identify themselves as republicans, and blacks towards the democrat party.

(c)

```
residuals(multi.log)

## NULL
```

Problem 3

```
require("arm")
require("foreign")
require("ggplot2")
require("VGAM")

## Loading required package: VGAM

## Loading required package: stats4

## Loading required package: splines

##
## Attaching package: 'VGAM'

## The following object is masked from 'package:arm':
##
##     logit

## The following object is masked from 'package:AER':
##
##     tobit

## The following object is masked from 'package:lmtest':
##
##     lrtest

## The following object is masked from 'package:car':
##
##     logit
```

```
require("gridExtra")
```

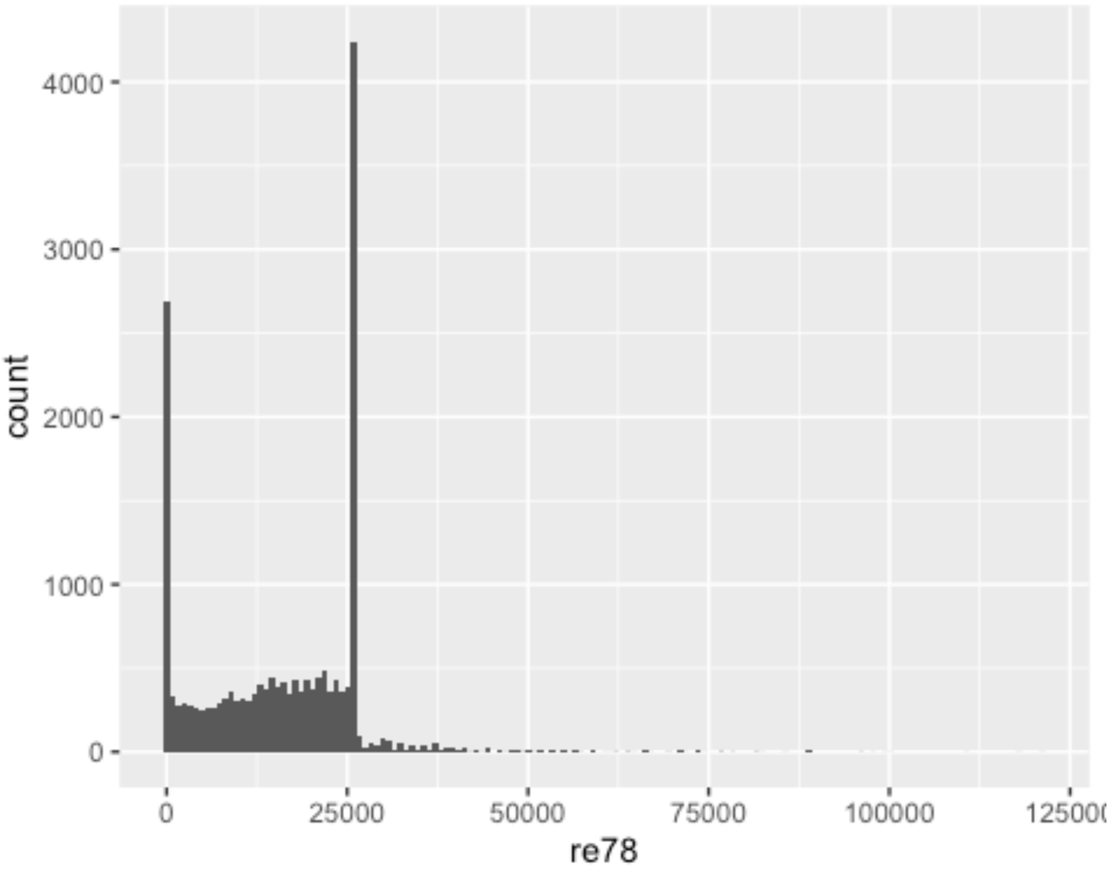
```
## Loading required package: gridExtra
```

```
nsw <- read.dta("/Users/Home/Documents/Michael_Ghattas/School/CU_Boulder/
2022/Spring 2022/STAT - 4400/Data/NSW.dw.obs.dta")

nsw$sample <- factor(nsw$sample, labels = c("NSW", "CPS", "PSID"))
nsw$black <- factor(nsw$black)
nsw$hisp <- factor(nsw$hisp)
nsw$nodegree <- factor(nsw$nodegree)
nsw$married <- factor(nsw$married)
nsw$treat <- factor(nsw$treat)
nsw$educ_cat4 <- factor(nsw$educ_cat4, labels = c("less than high school",
"high school", "sm college", "college"))

standardise <- function(X) {
    cols <- ncol(X)
    for (c in 1:cols) {
        if (is.numeric(X[, c])) {
            start <- ncol(X)
            c.c <- (X[, c] - mean(X[, c], na.rm=TRUE)) / (2 * sd(X[, c],
na.rm = TRUE))
            X[start+1] <- c.c
            colnames(X)[start+1] <- paste0("c.", colnames(X)[c])
        }
    }
    return(X)
}

nsw <- standardise(nsw)
summary(nsw)
```

```
##       age             educ        black      married    nodegree       re74
##  Min.   :16.00   Min.   : 0.00   0:16711   0: 5093   0:13045   Min.   :
0
##  1st Qu.:24.00   1st Qu.:11.00   1: 1956   1:13574   1: 5622   1st Qu.:
4898
```

```
##  Median :31.00   Median :12.00                                 Median :
15525
##  Mean   :33.37   Mean   :12.02                                 Mean   :
14621
##  3rd Qu.:42.00   3rd Qu.:14.00                                 3rd Qu.:
23882
##  Max.   :55.00   Max.   :18.00                                 Max.
:137149
##       re75              re78         hisp       sample      treat
##  Min.   :     0   Min.   :     0   0:17423   NSW :  185   0:18482
##  1st Qu.:  4726   1st Qu.:  6158   1: 1244   CPS :15992   1:  185
##  Median : 14899   Median : 16957             PSID: 2490
##  Mean   : 14253   Mean   : 15657
##  3rd Qu.: 23274   3rd Qu.: 25565
##  Max.   :156653   Max.   :121174
##                  educ_cat4        c.age              c.educ
##  less than high school:5622   Min.   :-0.7913   Min.   :-2.074555
##  high school          :7144   1st Qu.:-0.4269   1st Qu.:-0.176481
##  sm college           :3105   Median :-0.1079   Median :-0.003929
##  college              :2796   Mean   : 0.0000   Mean   : 0.000000
##                               3rd Qu.: 0.3933   3rd Qu.: 0.341176
##                               Max.   : 0.9856   Max.   : 1.031385
##      c.re74             c.re75             c.re78
##  Min.   :-0.7047   Min.   :-0.70089   Min.   :-0.71864
##  1st Qu.:-0.4686   1st Qu.:-0.46850   1st Qu.:-0.43598
##  Median : 0.0436   Median : 0.03179   Median : 0.05966
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.4464   3rd Qu.: 0.44364   3rd Qu.: 0.45474
##  Max.   : 5.9058   Max.   : 7.00266   Max.   : 4.84307
```

```
ggplot(data = nsw, aes(x = re78)) + geom_histogram(binwidth =
(range(nsw$re78)[2] - range(nsw$re78)[1])/150)
```



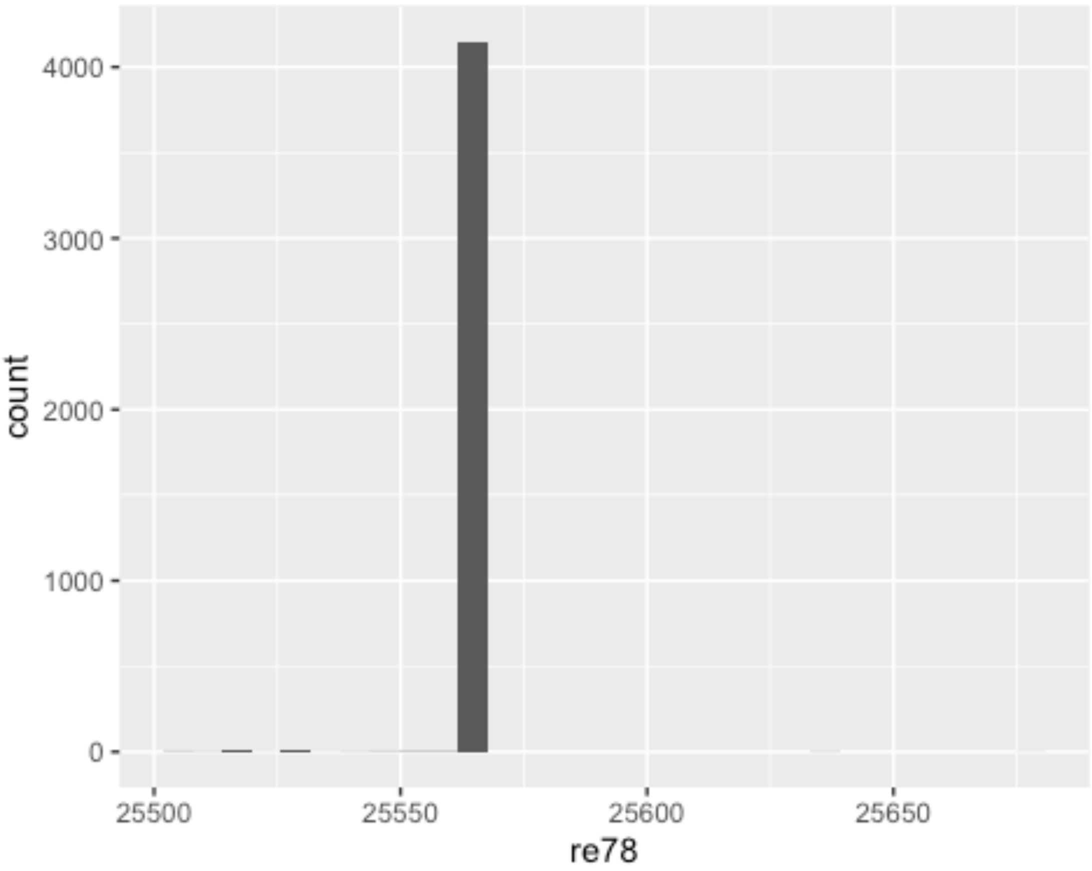
```
modex <- function(x) {
    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
}
```

```
print(paste0("The mode is: ", sprintf("$%3.2f", modex(nsw$re78))))
```

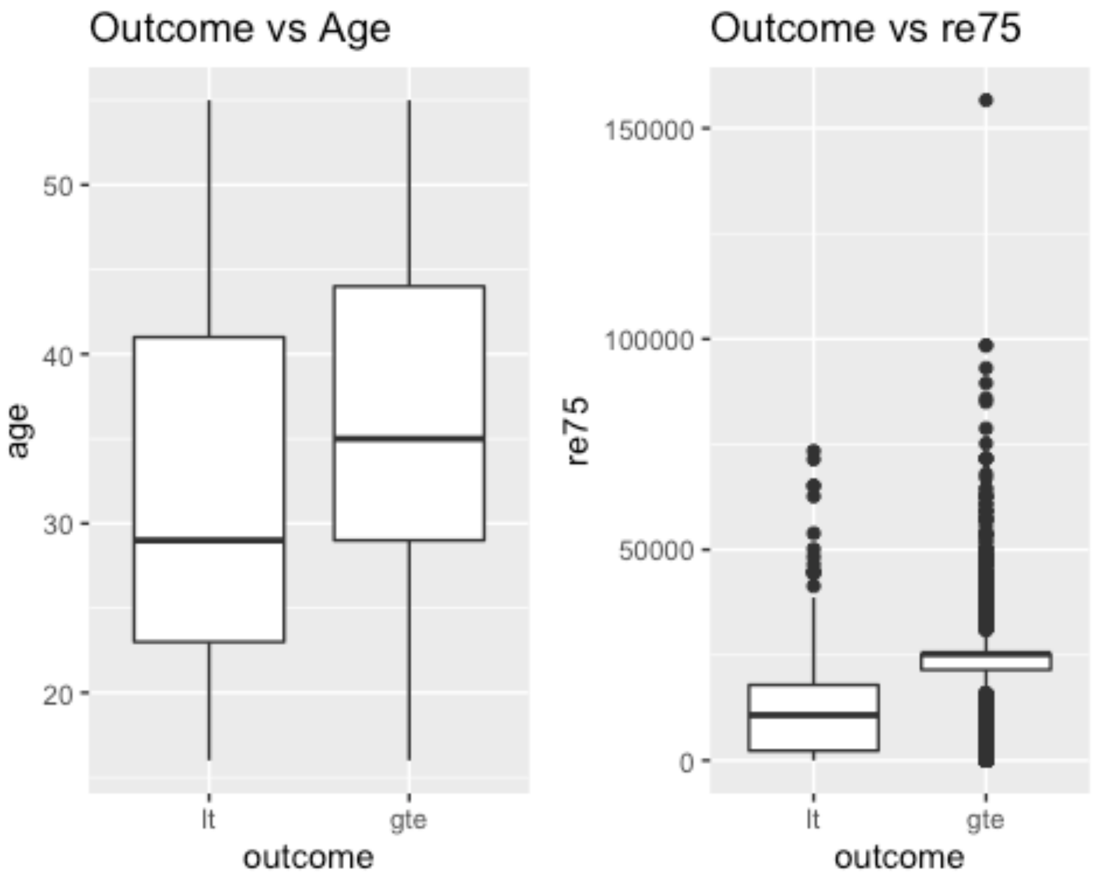
```
## [1] "The mode is: $25564.67"
```

```
ggplot(nsw[nsw$re78 >= 25500 & nsw$re78 < 25700,], aes(x = re78)) +
geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
nsw$outcome <- rep(NA, nrow(nsw))
nsw$outcome <- ifelse(nsw$re78 >= 25564.669921875, 1, 0)
nsw$outcome <- factor(nsw$outcome, labels = c("lt", "gte"))

p1 <- ggplot(data=nsw, aes(x = outcome, y = age)) + geom_boxplot() +
labs(title = "Outcome vs Age")
p2 <- ggplot(data=nsw, aes(x= outcome, y = re75)) + geom_boxplot() +
labs(title = "Outcome vs re75")
grid.arrange(p1, p2, nrow = 1)
```



```
table(nsw$outcome, nsw$black)

##
##            0     1
##   lt   11947  1666
##   gte   4764   290

table(nsw$outcome, nsw$hisp)

##
##            0     1
##   lt   12594  1019
##   gte   4829   225

table(nsw$outcome, nsw$educ)
```

```
##
##          0    1    2    3    4    5    6    7    8    9   10   11   12
13
##    lt    37   12   40   78  107  125  239  293  837  811 1116 1073 5165
867
##    gte    2    2    6   11   13   17   39   45  182  142  203  192 1979
353
##
##         14   15   16   17   18
##    lt   882  383  932  281  335
##    gte  465  155  719  263  266
```

```
fit1 <- glm(outcome ~ c.age + c.educ + c.re75 + black + married, family =
binomial(link = "logit"), data = nsw)
display(fit1)
```

```
## glm(formula = outcome ~ c.age + c.educ + c.re75 + black + married,
##     family = binomial(link = "logit"), data = nsw)
##             coef.est coef.se
## (Intercept) -1.93     0.06
## c.age       -0.05     0.05
## c.educ       0.66     0.05
## c.re75       3.89     0.07
## black1      -0.30     0.08
## married1     0.33     0.06
## ---
##   n = 18667, k = 6
##   residual deviance = 14505.0, null deviance = 21803.0 (difference =
7298.1)
```

```
predicted <- predict(fit1, nsw, type = "response")
y <- ifelse(nsw$re78 >= 25564.669921875, 1, 0)

error.rate <- mean((predicted > 0.5 & y == 0) | (predicted < .5 & y == 1))
print(paste0("Error rate: ", sprintf("%.2f%%", 100*error.rate)))
```

```
## [1] "Error rate: 15.88%"
```

```
fit2.a <- vglm(re78 ~ c.age + c.educ + c.re75, tobit(Lower = 0, Upper =
25563), data = nsw, subset = re78 < 25564)
summary(fit2.a)
```

```
##
## Call:
## vglm(formula = re78 ~ c.age + c.educ + c.re75, family = tobit(Lower = 0,
##     Upper = 25563), data = nsw, subset = re78 < 25564)
##
## Coefficients:
##                 Estimate Std. Error  z value Pr(>|z|)
## (Intercept):1  1.237e+04  7.933e+01  155.976  < 2e-16 ***
## (Intercept):2  9.027e+00  7.283e-03 1239.450  < 2e-16 ***
## c.age         -3.308e+03  1.533e+02  -21.575  < 2e-16 ***
## c.educ        -6.541e+02  1.510e+02   -4.331 1.49e-05 ***
## c.re75         1.362e+04  1.908e+02   71.368  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: mu, loglink(sd)
##
## Log-likelihood: -118527.5 on 27221 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2'
```

```
fit2.b <- vglm(re78 ~ c.age + c.educ + c.re75, tobit(Lower = 25564, Upper =
Inf), data = nsw, subset = re78 >= 25564)
summary(fit2.b)
```

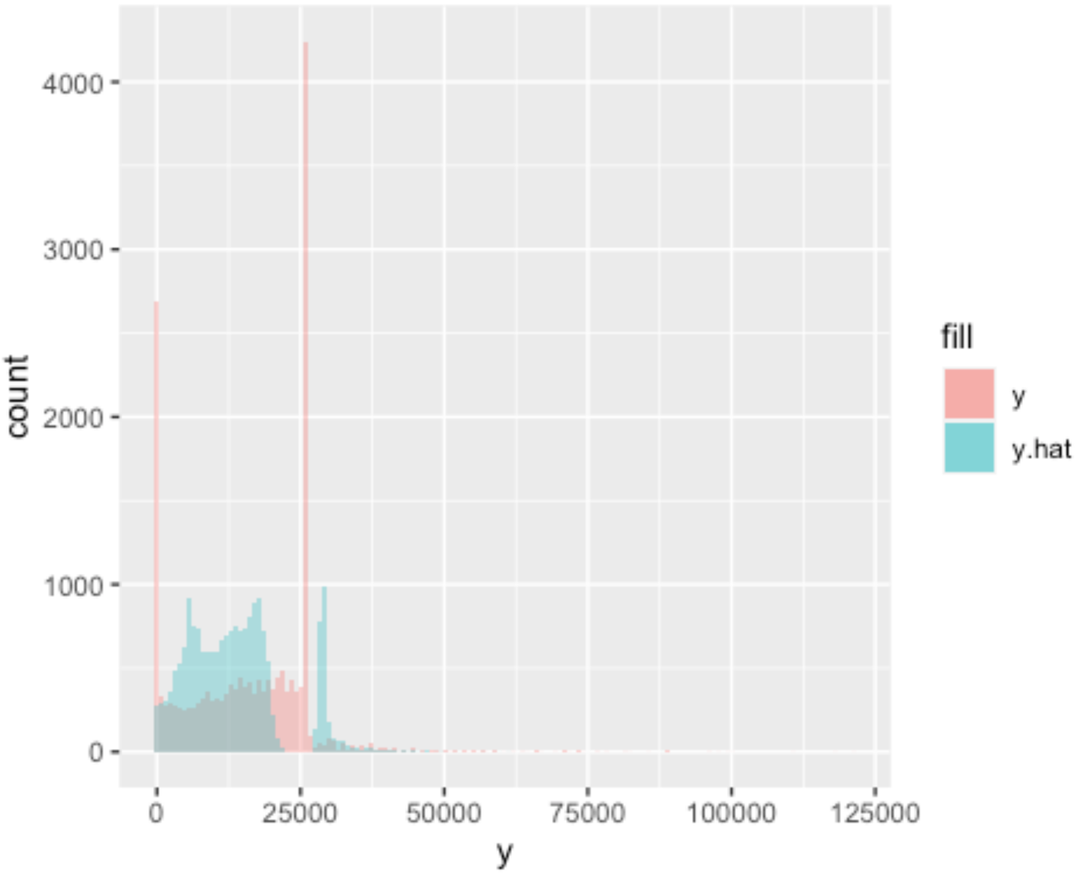
```
##
## Call:
## vglm(formula = re78 ~ c.age + c.educ + c.re75, family = tobit(Lower =
25564,
##     Upper = Inf), data = nsw, subset = re78 >= 25564)
```

```
## 
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept):1  2.343e+04  1.519e+02 154.217  < 2e-16 ***
## (Intercept):2  8.700e+00  1.285e-02 677.246  < 2e-16 ***
## c.age         -1.263e+03  2.294e+02  -5.504 3.71e-08 ***
## c.educ         8.253e+02  2.023e+02   4.080 4.51e-05 ***
## c.re75         9.322e+03  2.511e+02  37.130  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Names of linear predictors: mu, loglink(sd)
## 
## Log-likelihood: -51141.56 on 10103 degrees of freedom
## 
## Number of Fisher scoring iterations: 18 
## 
## No Hauck-Donner effect found in any of the estimates
```

```
y.hat <- ifelse(predict(fit1, nsw) < 0.5, ifelse(predict(fit2.a, nsw) < 0, 0,
predict(fit2.a, nsw)), predict(fit2.b, nsw))
y <- nsw$re78
print(paste0("RMSE: ", sprintf("%.2f", sqrt(mean((y - y.hat) ** 2)))))
```

```
## [1] "RMSE: 8482.20"
```

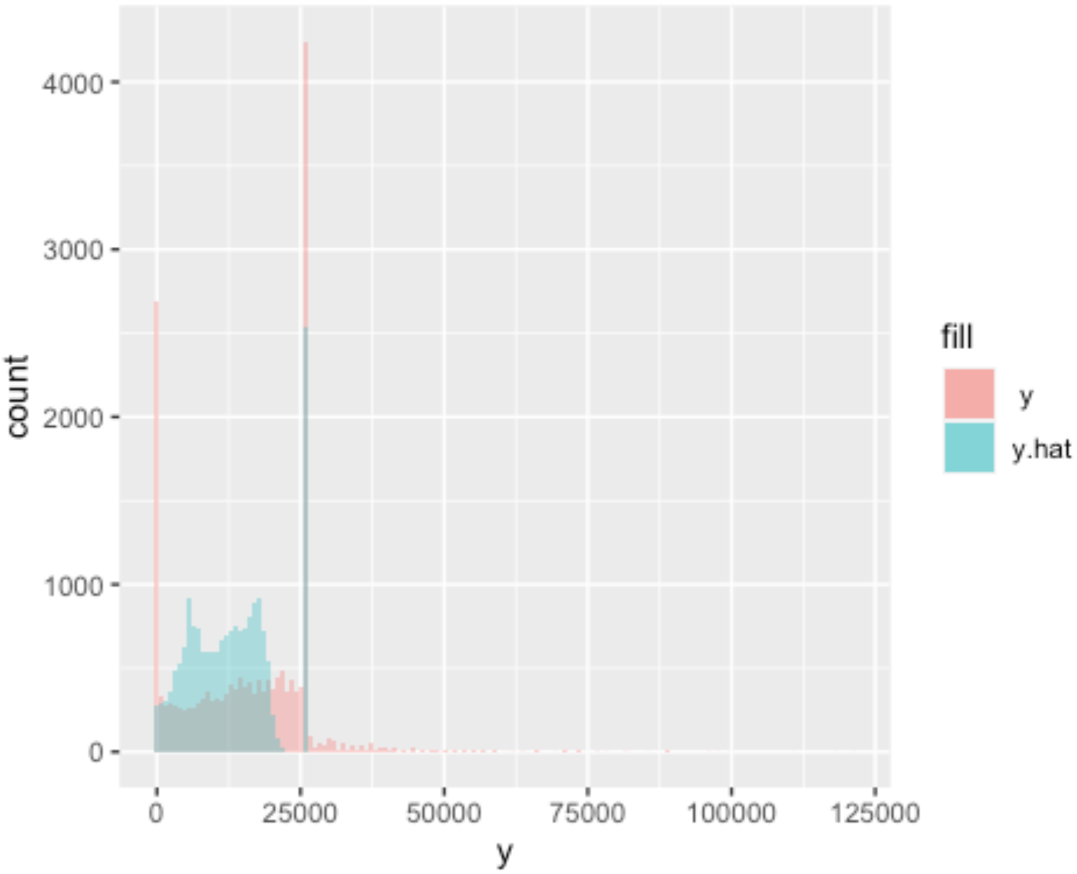
```
ggplot(data = data.frame(cbind(y = y, y.hat = y.hat))) +
    geom_histogram(aes(x = y, fill = "y"), alpha = .35, binwidth = (range(y)
[2] - range(y)[1])/150) +
    geom_histogram(aes(x = y.hat, fill = "y.hat"), alpha = .35, binwidth =
(range(y)[2] - range(y)[1])/150)
```



```
y.hat <- ifelse(predict(fit1, nsw) < 0.5, ifelse(predict(fit2.a, nsw) < 0, 0, predict(fit2.a, nsw)), 25564.669921875)
y <- nsw$re78
print(paste0("RMSE: ", sprintf("%.2f", sqrt(mean((y - y.hat) ** 2)))))
```

```
## [1] "RMSE: 8692.13"
```

```
ggplot(data = data.frame(cbind(y = y, y.hat = y.hat))) +
    geom_histogram(aes(x = y, fill =" y"), alpha = .35, binwidth = (range(y)[2] - range(y)[1])/150) +
    geom_histogram(aes(x = y.hat, fill = "y.hat"), alpha = .35, binwidth = (range(y)[2] - range(y)[1])/150)
```



fit2.a: Underestimate the number of observations where earning in 1978 are zero. fit2.b: Shifts the distribution towards values above $25,564.66. Replaced the prediction of this model with the hard-coded value $25564.67, though further adjustments are needed to improve the fit of the model.

Problem 4

(a)

```
df <- read.csv("/Users/Home/Documents/Michael_Ghattas/School/CU_Boulder/2022/
Spring 2022/STAT - 4400/Data/oscars.csv")
mcl <- polr(factor(Ch) ~ Nom + PrW + Gd + Gdr + DGA, Hess = TRUE, data = df)
summary(mcl)
```

```
## Call:
## polr(formula = factor(Ch) ~ Nom + PrW + Gd + Gdr + DGA, data = df,
##     Hess = TRUE)
##
## Coefficients:
##        Value Std. Error  t value
## Nom -0.11355    0.02119 -5.35872
## PrW  0.14399    0.13420  1.07295
## Gd  -0.01115    0.37425 -0.02978
## Gdr -1.45578    0.26845 -5.42288
## DGA -2.67750    0.35994 -7.43866
##
## Intercepts:
##     Value    Std. Error t value
## 0|1  -5.6997   0.2864   -19.9019
## 1|2  -2.2322   0.1497   -14.9132
##
## Residual Deviance: 1618.15
## AIC: 1632.15
```

(b)

```
Ch = as.numeric(df$Ch)
Nom = as.numeric(df$Nom)
PrW = as.numeric(df$PrW)
Gd = as.numeric(df$Gd)
Gdr = as.numeric(df$Gdr)
DGA = as.numeric(df$DGA)

data = as.matrix(c(Ch, Nom, PrW, Gd, Gdr, DGA))
mcl <- polr(factor(Ch) ~ Nom + PrW + Gd + Gdr + DGA, Hess = TRUE, data =
data)
# plot(Ch, mcl) Error!
# hist(Ch, mcl) Error!
```

Not sure how to resolve this issue!

(c)

```
# hist(residuals(mcl)) Error!
# plot(residuals(mcl)) Error!
```

Not sure how to resolve this issue!

Problem 5

```
require(foreign)
require(nnet)

## Loading required package: nnet

## Warning: package 'nnet' was built under R version 4.1.2

require(ggplot2)
require(reshape2)

## Loading required package: reshape2

ml <- read.dta("/Users/Home/Documents/Michael_Ghattas/School/CU_Boulder/2022/
Spring 2022/STAT - 4400/Data/hsbdemo.dta")
with(ml, table(ses, prog))

##         prog
## ses       general academic vocation
##   low          16       19       12
##   middle       20       44       31
##   high          9       42        7

with(ml, do.call(rbind, tapply(write, prog, function(x) c(M = mean(x), SD =
sd(x)))))

##                 M       SD
## general  51.33333 9.397775
## academic 56.25714 7.943343
## vocation 46.76000 9.318754
```

(a)

```
ml$prog2 <- relevel(ml$prog, ref = "academic")
test <- multinom(prog2 ~ ses + write, data = ml)
```

```
## # weights:  15 (8 variable)
## initial  value 219.722458
## iter  10 value 179.982880
## final  value 179.981726
## converged
```

```
summary(test)
```

```
## Call:
## multinom(formula = prog2 ~ ses + write, data = ml)
##
## Coefficients:
##          (Intercept)  sesmiddle    seshigh      write
## general     2.852198 -0.5332810 -1.1628226 -0.0579287
## vocation    5.218260  0.2913859 -0.9826649 -0.1136037
##
## Std. Errors:
##          (Intercept) sesmiddle   seshigh      write
## general     1.166441 0.4437323 0.5142196 0.02141097
## vocation    1.163552 0.4763739 0.5955665 0.02221996
##
## Residual Deviance: 359.9635
## AIC: 375.9635
```

```
exp(coef(test))
```

```
##          (Intercept) sesmiddle   seshigh     write
## general     17.32582 0.5866769 0.3126026 0.9437172
## vocation   184.61262 1.3382809 0.3743123 0.8926116
```

```
head(pp <- fitted(test))
```

```
##    academic   general  vocation
## 1 0.1482764 0.3382454 0.5134781
## 2 0.1202017 0.1806283 0.6991700
## 3 0.4186747 0.2368082 0.3445171
## 4 0.1726885 0.3508384 0.4764731
## 5 0.1001231 0.1689374 0.7309395
## 6 0.3533566 0.2377976 0.4088458
```

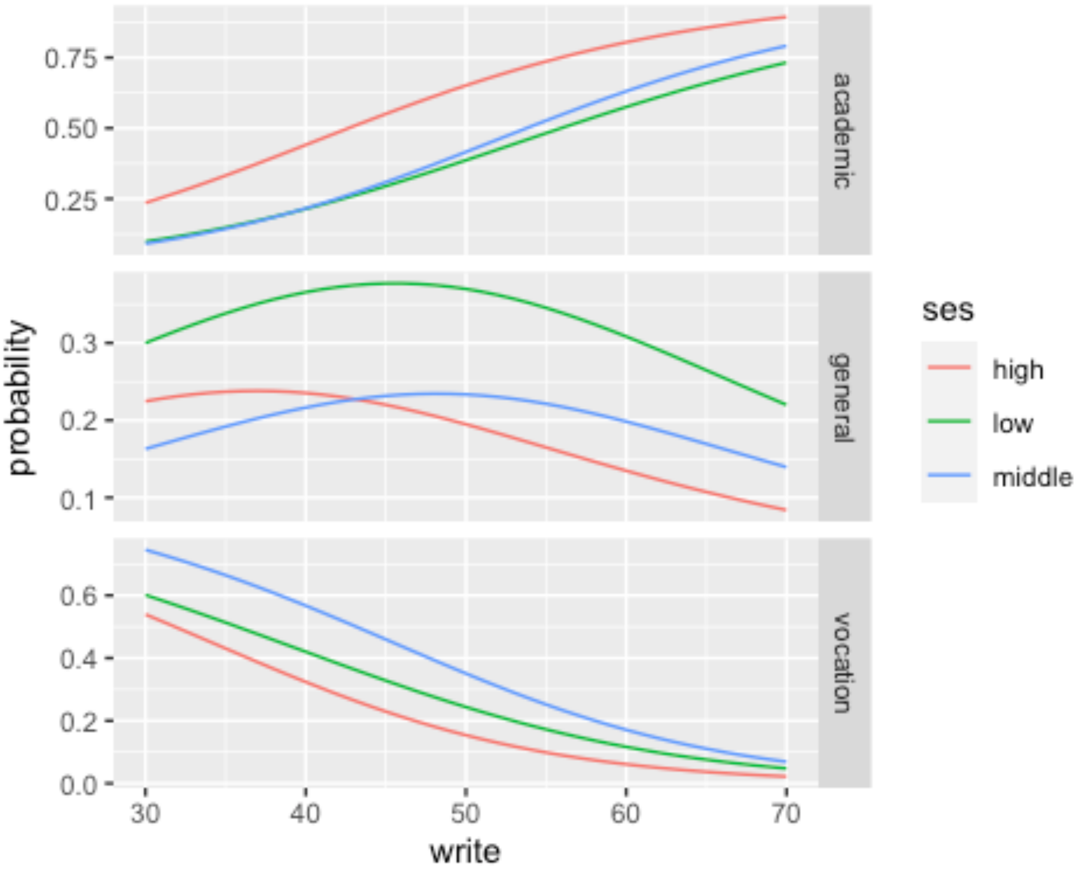
One-unit increase in the variable write is associated with the decrease in the log odds of being in general program vs. academic program in the amount of .058. One-unit increase in the variable write is associated with the decrease in the log odds of being in vocation program vs. academic program. in the amount of .1136. Odds of being in general program vs. in academic program will decrease by 1.163. Odds of being in general program vs. in academic program will decrease by 0.533, although this coefficient is not significant. Odds of being in vocation program vs. in academic program will decrease by 0.983. Odds of being in vocation program vs. in academic program will increase by 0.291 although this coefficient is not significant.

(b)

```
dses <- data.frame(ses = c("low", "middle", "high"), write = mean(ml$write))
predict(test, newdata = dses, "probs")

##    academic   general  vocation
## 1 0.4396845 0.3581917 0.2021238
## 2 0.4777488 0.2283353 0.2939159
## 3 0.7009007 0.1784939 0.1206054

dwrite <- data.frame(ses = rep(c("low", "middle", "high"), each = 41), write
= rep(c(30:70), 3))
pp.write <- cbind(dwrite, predict(test, newdata = dwrite, type = "probs", se
= TRUE))
by(pp.write[, 3:5], pp.write$ses, colMeans)

## pp.write$ses: high
##  academic   general  vocation
## 0.6164315 0.1808037 0.2027648
## ---------------------------------------------------------------
## pp.write$ses: low
##  academic   general  vocation
## 0.3972977 0.3278174 0.2748849
## ---------------------------------------------------------------
## pp.write$ses: middle
##  academic   general  vocation
## 0.4256198 0.2010864 0.3732938
```

(c)

```
lpp <- melt(pp.write, id.vars = c("ses", "write"), value.name =
"probability")
ggplot(lpp, aes(x = write, y = probability, colour = ses)) + geom_line() +
facet_grid(variable ~ ., scales = "free")
```


probability
academic
general
vocation
write
ses
high
low
middle