Biostat 200C Homework 3

Due May 14 @ 11:59PM

Q1.

The **log-logistic** distribution with the probability density function

$$f(y) = \frac{e^{\theta}\lambda y^{\lambda-1}}{(1+e^{\theta}y^{\lambda})^2}$$

is sometimes used for modelling survival times.

- (a) Find the survivor function $S(y)$, the hazard function $h(y)$ and the cumulative hazard function $H(y)$.
- (b) Show that the median survival time is $\exp(-\theta/\lambda)$.
- (c) Use R to plot the hazard function for $\lambda = 1$ and $\lambda = 5$ with $\theta = -5$, $\theta = -2$, and $\theta = 1/2$, in one figure.

Q2. ELMR Exercise 7.5

The data arise from a large postal survey on the psychology of debt. The frequency of credit card use `ccarduse` is a three-level factor ranging from never, occasionally to regularly.

```
data(debt)
help(debt)
```

- (a) Declare the response as an ordered factor and make a plot showing the relationship to `prodebt`. Comment on the plot. Use a table or plot to display the relationship between the response and the income group.
- (b) Fit a proportional odds model for credit card use with all the other variables as predictors. What are the two most significant predictors and what is their qualitative effect on the response? What is the least significant predictor?
- (c) Use stepwise AIC to select a smaller model than the full set of predictors. You will need to handle the missing values carefully. Report on the qualitative effect of the predictors in your chosen model. Can we conclude that the predictors that were dropped from the model have no relation to the response?
- (d) Compute the median values of the predictors in your selected model. At these median values, compare the predicted outcome probabilities for both smokers and nonsmokers.
- (e) Fit a proportional hazards model to the same set of predictors and recompute the two sets of probabilities from the previous question. Does it make a difference to use this type of model?

Q1.

The **log-logistic** distribution with the probability density function

$$f(y) = \frac{e^{\theta}\lambda y^{\lambda-1}}{(1+e^{\theta}y^{\lambda})^2}$$

is sometimes used for modelling survival times.

- (a) Find the survivor function $S(y)$, the hazard function $h(y)$ and the cumulative hazard function $H(y)$.

$$S(y) = 1 - F(y) \qquad h(y) = \frac{f(y)}{S(y)} \qquad H(y) = -\log S(y)$$

$$F(y) = \int_0^y \frac{e^{\theta}\lambda y^{\lambda-1}}{(1+e^{\theta}y^{\lambda})^2}\,dy \qquad u = e^{\theta}y^{\lambda}+1 \qquad \frac{du}{dy} = e^{\theta}\lambda y^{\lambda-1} \qquad dy = \frac{e^{-\theta}y^{1-\lambda}}{\lambda}\,du$$

$$\Rightarrow \int_1^u \frac{1}{u^2}\,du = -\frac{1}{u}\Big|_1^u \Rightarrow -\frac{1}{u}+1 \Rightarrow F(y) = \frac{e^{\theta}y^{\lambda}}{e^{\theta}y^{\lambda}+1}$$

$$S(y) = 1 - \frac{e^{\theta}y^{\lambda}}{e^{\theta}y^{\lambda}+1} = \frac{1}{e^{\theta}y^{\lambda}+1}$$

$$h(y) = \frac{e^{\theta}\lambda y^{\lambda-1}}{(1+e^{\theta}y^{\lambda})^2} \cdot \cancel{e^{\theta}y^{\lambda}+1} = \frac{e^{\theta}\lambda y^{\lambda-1}}{1+e^{\theta}y^{\lambda}}$$

$$H(y) = -\log(Sy) = \theta + \lambda\log(y) \text{ or } (\log(e^{\theta}y^{\lambda}+1))$$

- (b) Show that the median survival time is $\exp(-\theta/\lambda)$. $e^{-\theta/\lambda}$

$$F(y) = \frac{1}{2} = \frac{e^{\theta} y^{\lambda}}{e^{\theta} y^{\lambda}+1}$$

$$S(y) = 1 - \frac{1}{2} = \frac{e^{\theta} y^{\lambda}}{e^{\theta} y^{\lambda}+1} = \frac{1}{2}$$

$$2(e^{\theta} y^{\lambda}) = e^{\theta} y^{\lambda} + 1$$

$$1 = e^{\theta} y^{\lambda}$$

$$y^{\lambda} = e^{-\theta}$$

$$\lambda \log y = -\theta$$

$$\log y = -\frac{\theta}{\lambda}$$

$$y = e^{-\theta/\lambda}$$

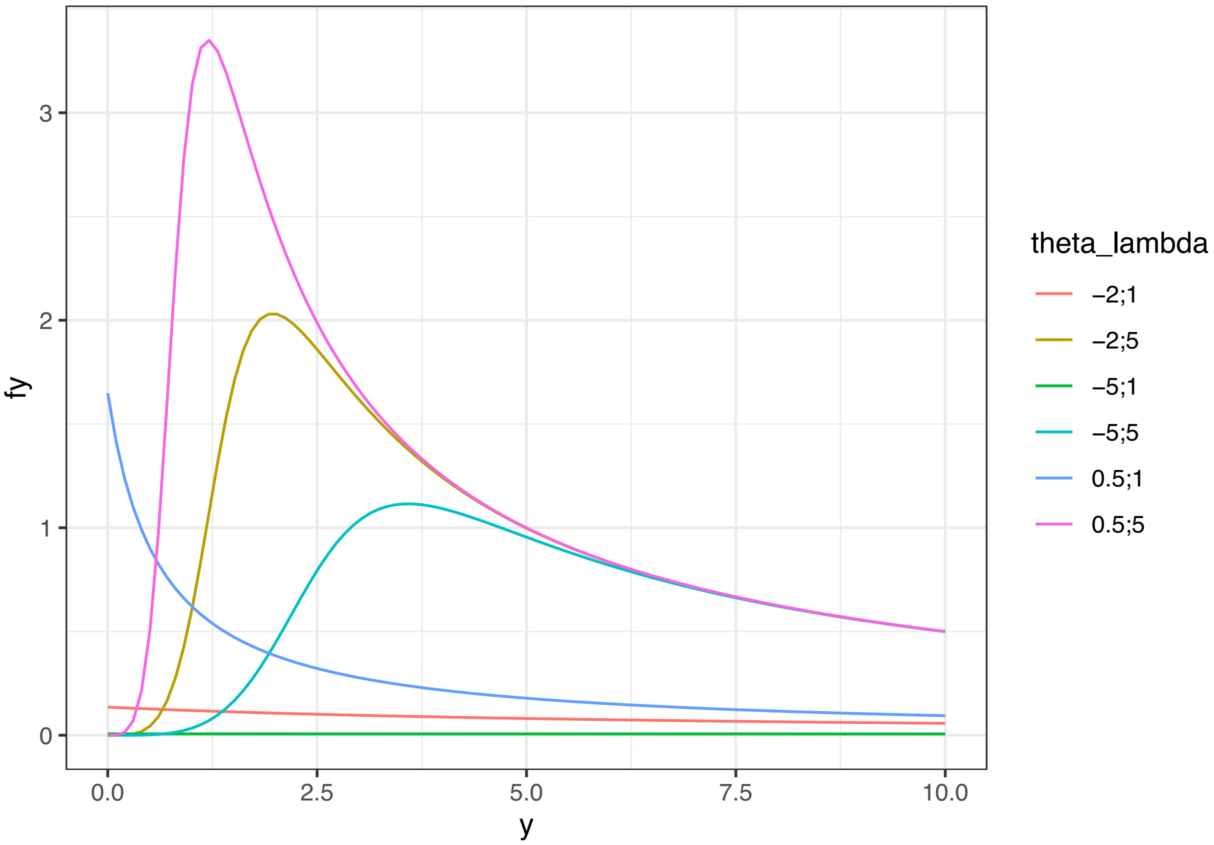
- (c) Use R to plot the hazard function for $\lambda = 1$ and $\lambda = 5$ with $\theta = -5$, $\theta = -2$, and $\theta = 1/2$, in one figure.

Solution:

```
.hf <- function(y, theta, lambda){
   fy <- (exp(theta) * lambda * y^(lambda - 1)) / (1 + exp(theta) * y^(lambda))
  return(fy)
}
plot_data <- NULL
```

```
for(theta in c(-5, -2, 0.5)){
  for(lambda in c(1, 5)){
    plot_data <- rbind(plot_data, data.frame(
      fy = .hf(y = seq(0, 10, length = 100), theta, lambda),
      y = seq(0, 10, length = 100), theta_lambda = paste0(theta, ";", lambda)
      )
    )
  }
  }

ggplot(
  data = plot_data, aes
  (x = y, y = fy, group = theta_lambda, color = theta_lambda)) +
  geom_line() + theme_bw()
```



Q2. ELMR Exercise 7.5

The data arise from a large postal survey on the psychology of debt. The frequency of credit card use `ccarduse` is a three-level factor ranging from never, occasionally to regularly.

```
data(debt)
help(debt)
```

- (a) Declare the response as an ordered factor and make a plot showing the relationship to `prodebt`. Comment on the plot. Use a table or plot to display the relationship between the response and the income group.

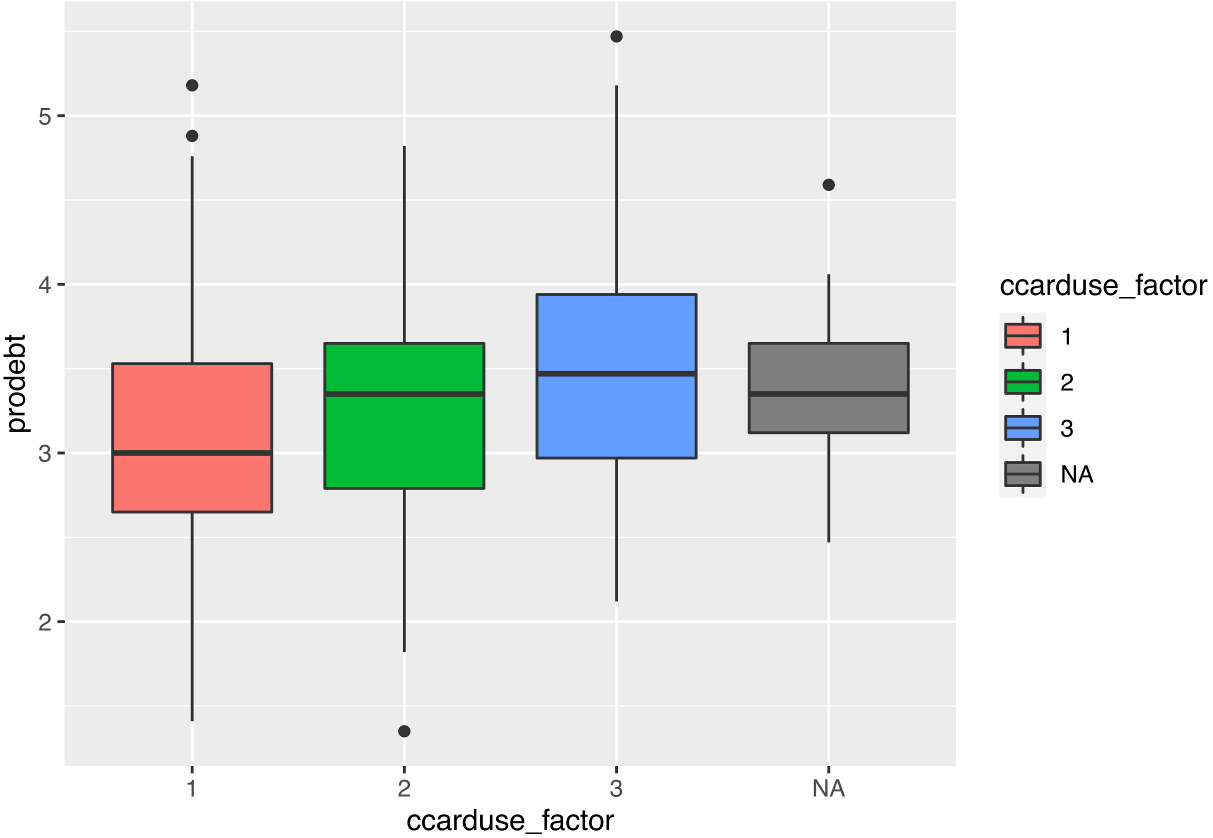
Solution:

```
debt0 <- debt %>%
  mutate(ccarduse_factor = as.factor(ccarduse)) %>%
  mutate_at(vars(incomegp,house,agegp),ordered)
summary(debt0$ccarduse_factor)
```

```
##    1    2    3 NA's
##  232  105   93   34
```

```
debt0 %>%
  ggplot() +
  geom_boxplot(mapping = aes(
    x = ccarduse_factor, y=prodebt, fill=ccarduse_factor))
```

```
## Warning: Removed 45 rows containing non-finite values (stat_boxplot).
```

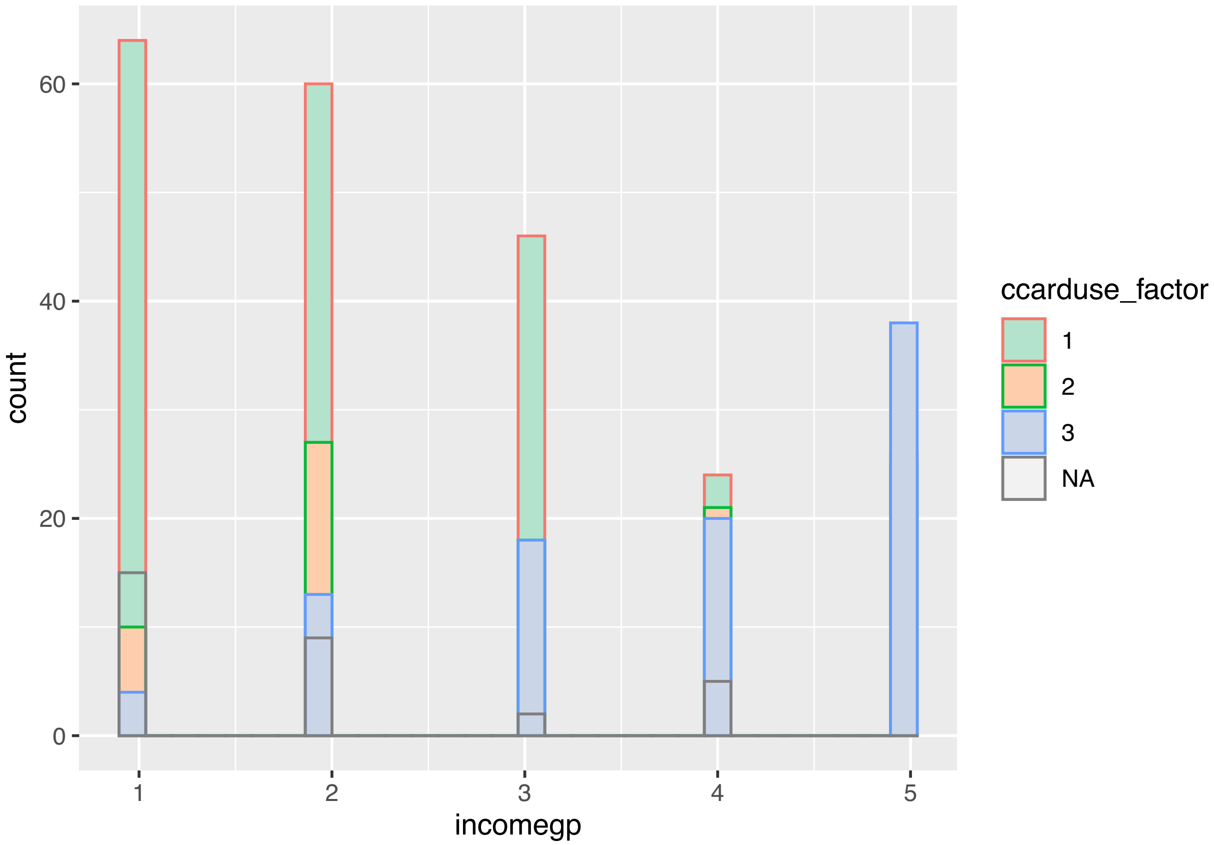


From the boxplot, a higher frequency of a person use credit cards are associated with a higher attitudes to debt.

```
ggplot(debt0 %>%
         mutate_at(vars(incomegp), unclass) %>%
         mutate_at(vars(ccarduse_factor), as.factor),
       aes(x = incomegp, color = ccarduse_factor, fill = ccarduse_factor)) +
  geom_histogram(position = 'identity') + scale_fill_brewer(palette = "Pastel2")
```

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

```
## Warning: Removed 19 rows containing non-finite values (stat_bin).
```



From the plot, a higher frequency of a person use credit cards are associated with a higher level of income.

- (b) Fit a proportional odds model for credit card use with all the other variables as predictors. What are the two most significant predictors and what is their qualitative effect on the response? What is the least significant predictor?

Solution:

```
library(MASS)
debt$ccarduse <- as.factor(debt$ccarduse)
pomod <- polr(ccarduse ~ incomegp + house + children +
                singpar + agegp + bankacc + bsocacc + manage +
                cigbuy + xmasbuy + locintrn + prodebt , data = debt, Hess = TRUE)
coeftest(pomod)
```

```
##
## t test of coefficients:
##
##            Estimate Std. Error t value  Pr(>|t|)
## incomegp   0.471313   0.106097  4.4423 1.267e-05 ***
## house      0.116001   0.232363  0.4992 0.6179993
## children -0.078724   0.125033 -0.6296 0.5294330
## singpar    0.881718   0.597114  1.4766 0.1408591
## agegp      0.205684   0.157610  1.3050 0.1929224
## bankacc    2.102696   0.593392  3.5435 0.0004600 ***
## bsocacc    0.473216   0.267133  1.7715 0.0775337 .
## manage    0.181792   0.165290  1.0998 0.2723170
## cigbuy   -0.735459   0.298068 -2.4674 0.0141867 *
## xmasbuy   0.470143   0.412963  1.1385 0.2558672
## locintrn  0.118812   0.142398  0.8344 0.4047602
## prodebt   0.610464   0.182247  3.3497 0.0009162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result, `incomegp` and `bankacc` are most significant predictors. And `house` is the least significant predictor.

```
exp(pomod$coefficients)
```

```
##  incomegp     house  children   singpar     agegp   bankacc   bsocacc    manage
## 1.6020964 1.1229975 0.9242949 2.4150459 1.2283646 8.1882137 1.6051485 1.1993643
##    cigbuy   xmasbuy  locintrn   prodebt
## 0.4792856 1.6002228 1.1261586 1.8412851
```

Interpret: Increase level in income group is associated with a higher odds of using credit cards more often. Interpret: People have have a bank account are associated with a higher odds of using credit cards more often.

- (c) Use stepwise AIC to select a smaller model than the full set of predictors. You will need to handle the missing values carefully. Report on the qualitative effect of the predictors in your chosen model. Can we conclude that the predictors that were dropped from the model have no relation to the response?

Solution:

```
debt1 <- debt %>%
  na.exclude() %>%
  mutate_at(vars(incomegp, house, agegp), ordered)
pomod1 <- polr(ccarduse ~ incomegp + house + children +
                singpar + agegp + bankacc + bsocacc + manage +
                cigbuy + xmasbuy + locintrn + prodebt , data = debt1, Hess = TRUE)
summary(pomod1)
```

```
## Call:
## polr(formula = ccarduse ~ incomegp + house + children + singpar +
##     agegp + bankacc + bsocacc + manage + cigbuy + xmasbuy + locintrn +
##     prodebt, data = debt1, Hess = TRUE)
##
## Coefficients:
##                Value Std. Error t value
## incomegp.L  1.38901     0.4100  3.3881
## incomegp.Q -0.37146     0.3412 -1.0886
## incomegp.C  0.22825     0.3025  0.7546
## incomegp^4 -0.12451     0.2763 -0.4507
## house.L     0.38695     0.3573  1.0831
## house.Q    -0.31067     0.2686 -1.1565
## children   -0.21106     0.1422 -1.4840
## singpar     0.80560     0.6327  1.2732
## agegp.L     0.59576     0.4095  1.4549
## agegp.Q    -0.84090     0.3187 -2.6386
## agegp.C     0.03403     0.2415  0.1409
## bankacc     1.89918     0.6013  3.1584
## bsocacc     0.46969     0.2799  1.6780
## manage      0.16846     0.1683  1.0007
## cigbuy     -0.78520     0.3125 -2.5125
## xmasbuy     0.47726     0.4267  1.1184
## locintrn    0.15973     0.1460  1.0938
## prodebt     0.70373     0.1882  3.7399
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  6.2133  1.3973     4.4468
## 2|3  7.6853  1.4234     5.3991
##
## Residual Deviance: 499.4114
## AIC: 539.4114
```

```
#stepwise
pomod2 <- step(pomod1)
```

```
## Start:  AIC=539.41
## ccarduse ~ incomegp + house + children + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + xmasbuy + locintrn + prodebt
##
##            Df    AIC
## - house     2 537.66
## - manage    1 538.42
## - locintrn  1 538.62
## - xmasbuy   1 538.69
## - singpar   1 538.99
## <none>        539.41
## - children  1 539.65
## - bsocacc   1 540.24
## - agegp     3 543.62
## - cigbuy    1 543.93
## - incomegp  4 544.41
## - bankacc   1 550.29
## - prodebt   1 552.08
##
## Step:  AIC=537.66
## ccarduse ~ incomegp + children + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + xmasbuy + locintrn + prodebt
##
##            Df    AIC
## - locintrn  1 536.54
## - manage    1 536.90
## - xmasbuy   1 537.25
## - singpar   1 537.38
## <none>        537.66
## - children  1 538.33
## - bsocacc   1 539.59
## - cigbuy    1 542.31
## - agegp     3 542.56
## - incomegp  4 545.98
## - bankacc   1 550.69
## - prodebt   1 550.99
##
## Step:  AIC=536.54
## ccarduse ~ incomegp + children + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + xmasbuy + prodebt
##
##            Df    AIC
## - manage    1 535.98
## - singpar   1 536.23
## - xmasbuy   1 536.35
## <none>        536.54
## - children  1 537.30
## - bsocacc   1 538.74
## - agegp     3 541.04
## - cigbuy    1 541.15
## - incomegp  4 545.82
## - prodebt   1 549.07
## - bankacc   1 550.68
##
## Step:  AIC=535.98
## ccarduse ~ incomegp + children + singpar + agegp + bankacc +
##     bsocacc + cigbuy + xmasbuy + prodebt
##
##            Df    AIC
## - singpar   1 535.26
## <none>        535.98
## - xmasbuy   1 536.01
## - children  1 537.00
## - bsocacc   1 539.37
## - agegp     3 540.98
## - cigbuy    1 541.40
## - incomegp  4 544.24
## - prodebt   1 547.33
## - bankacc   1 551.37
##
## Step:  AIC=535.26
## ccarduse ~ incomegp + children + agegp + bankacc + bsocacc +
##     cigbuy + xmasbuy + prodebt
##
##            Df    AIC
## <none>        535.26
## - xmasbuy   1 535.40
## - children  1 536.07
## - bsocacc   1 538.38
## - cigbuy    1 540.85
## - agegp     3 541.03
## - incomegp  4 542.24
## - prodebt   1 547.06
## - bankacc   1 549.99
```

```
summary(pomod2)
```

```
## Call:
## polr(formula = ccarduse ~ incomegp + children + agegp + bankacc +
##     bsocacc + cigbuy + xmasbuy + prodebt, data = debt1, Hess = TRUE)
##
## Coefficients:
##               Value Std. Error t value
## incomegp.L  1.36179     0.3753  3.6287
## incomegp.Q -0.30367     0.3297 -0.9210
## incomegp.C  0.17033     0.2980  0.5717
## incomegp^4 -0.11255     0.2669 -0.4217
## children   -0.22971     0.1384 -1.6594
## agegp.L     0.54780     0.3534  1.5499
## agegp.Q    -0.91670     0.3034 -3.0217
## agegp.C     0.03771     0.2393  0.1576
## bankacc     2.05036     0.5840  3.5110
## bsocacc     0.59832     0.2661  2.2487
## cigbuy     -0.82847     0.3061 -2.7069
## xmasbuy     0.59538     0.4132  1.4408
## prodebt     0.65625     0.1804  3.6378
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  4.7492  0.9753     4.8695
## 2|3  6.2018  1.0026     6.1855
##
## Residual Deviance: 505.2628
## AIC: 535.2628
```

```
coeftest(pomod2)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## incomegp.L  1.361787   0.375278  3.6287 0.0003367 ***
## incomegp.Q -0.303668   0.329699 -0.9210 0.3577943
## incomegp.C  0.170334   0.297953  0.5717 0.5679814
## incomegp^4 -0.112554   0.266909 -0.4217 0.6735626
## children   -0.229714   0.138433 -1.6594 0.0981211 .
## agegp.L     0.547802   0.353447  1.5499 0.1222642
## agegp.Q    -0.916705   0.303375 -3.0217 0.0027387 **
## agegp.C     0.037712   0.239329  0.1576 0.8749019
## bankacc     2.050364   0.583977  3.5110 0.0005177 ***
## bsocacc     0.598318   0.266069  2.2487 0.0252830 *
## cigbuy     -0.828473   0.306058 -2.7069 0.0071953 **
## xmasbuy     0.595378   0.413232  1.4408 0.1507279
## prodebt     0.656253   0.180400  3.6378 0.0003256 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
exp(pomod2$coefficients)
```

```
## incomegp.L incomegp.Q incomegp.C incomegp^4   children    agegp.L    agegp.Q 
##  3.9031634  0.7381060  1.1857014  0.8935494  0.7947607  1.7294467  0.3998344 
##    agegp.C    bankacc    bsocacc     cigbuy    xmasbuy    prodebt 
##  1.0384323  7.7707321  1.8190563  0.4367157  1.8137169  1.9275570
```

Smaller model: `ccarduse ~ incomegp + children + agegp + bankacc + bsocacc + cigbuy + xmasbuy + prodebt`

Interpret:

1. Increase level in income group is associated with a higher odds of raise one level of credit cards use.
2. Increase number of children in household is associated with a lower odds of increase one level of credit cards use.
3. Increase level in age group is associated with a higher odds of raise one level of credit cards use.
4. Compared to people have no bank account, people have a bank account are associated with a higher odds of raise one level of credit cards use.
5. Compared to people have no building society account, people have a building society account are associated with a higher odds of raise one level of credit cards use.
6. Compared to people do not buy cigarettes, people who buy cigarettes are associated with a lower odds of raise one level of credit cards use.
7. Compared to people do not buy Christmas presents for children, people who buy Christmas presents for children are associated with a higher odds of raise one level of credit cards use.
8. Increase level in attitudes to debt is associated with a higher odds of raise one level of credit cards use.

No, even we dropped those predictor, but its still can relate to the response variable.

- (d) Compute the median values of the predictors in your selected model. At these median values, compare the predicted outcome probabilities for both smokers and nonsmokers.

Solution:

```
debt2 <- data.frame(i = 0)
for(i in c("incomegp", "children", "agegp", "bankacc", "bsocacc", "cigbuy",
           "xmasbuy", "prodebt")){
  temp <- data.frame(i = quantile(debt1[,i], 0.5, type = 3))
  debt2 <- cbind(debt2, temp)}
debt2 <- debt2[,-1]
names(debt2) <- c("incomegp", "children", "agegp", "bankacc", "bsocacc",
                  "cigbuy", "xmasbuy", "prodebt")
```

Smokers

```
debt2[6] = 1
predict(pomod2, debt2, type = "probs")
```

```
##         1         2         3
## 0.5084236 0.3071005 0.1844760
```

Nonsmokers

```
debt2[6] = 0
predict(pomod2, debt2, type = "probs")
```

```
##         1         2         3
## 0.3111442 0.3476305 0.3412253
```

- (e) Fit a proportional hazards model to the same set of predictors and recompute the two sets of probabilities from the previous question. Does it make a difference to use this type of model?

Solution:

```
pomod3 = polr(ccarduse ~ incomegp + children + agegp + bankacc +  bsocacc +
                cigbuy + xmasbuy + prodebt,
              method = "cloglog", data = debt1, Hess = TRUE)
summary(pomod3)
```

```
## Call:
## polr(formula = ccarduse ~ incomegp + children + agegp + bankacc +
##     bsocacc + cigbuy + xmasbuy + prodebt, data = debt1, Hess = TRUE,
##     method = "cloglog")
##
## Coefficients:
##                Value Std. Error t value
## incomegp.L  0.68409    0.19851  3.4461
## incomegp.Q -0.06668    0.17171 -0.3883
## incomegp.C  0.13240    0.17039  0.7770
## incomegp^4 -0.05261    0.15764 -0.3337
## children   -0.15850    0.08048 -1.9694
## agegp.L     0.34171    0.19949  1.7129
## agegp.Q    -0.53383    0.16983 -3.1433
## agegp.C     0.01662    0.14784  0.1124
## bankacc     1.00245    0.25444  3.9399
## bsocacc     0.36193    0.16242  2.2284
## cigbuy     -0.40915    0.16315 -2.5078
## xmasbuy     0.14582    0.23781  0.6132
## prodebt     0.42334    0.11563  3.6612
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  2.1228  0.5391     3.9377
## 2|3  2.9816  0.5469     5.4513
##
## Residual Deviance: 515.1315
## AIC: 545.1315
```

Smokers

```
debt2[6] = 1
predict(pomod3, debt2, type = "probs")
```

```
##         1         2         3 
## 0.5003375 0.3052067 0.1944558
```

Nonsmokers

```
debt2[6] = 0
predict(pomod3, debt2, type = "probs")
```

```
##         1         2         3 
## 0.3692502 0.2937532 0.3369966
```

No, the results compare to the previous question are very similar. So does not make a big difference to use this type of model.