Biostat 200C Homework 3

Due May 14 @ 11:59PM

Yenlin Lai

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 $H(y)$.

Solution: See the pdf file attached.

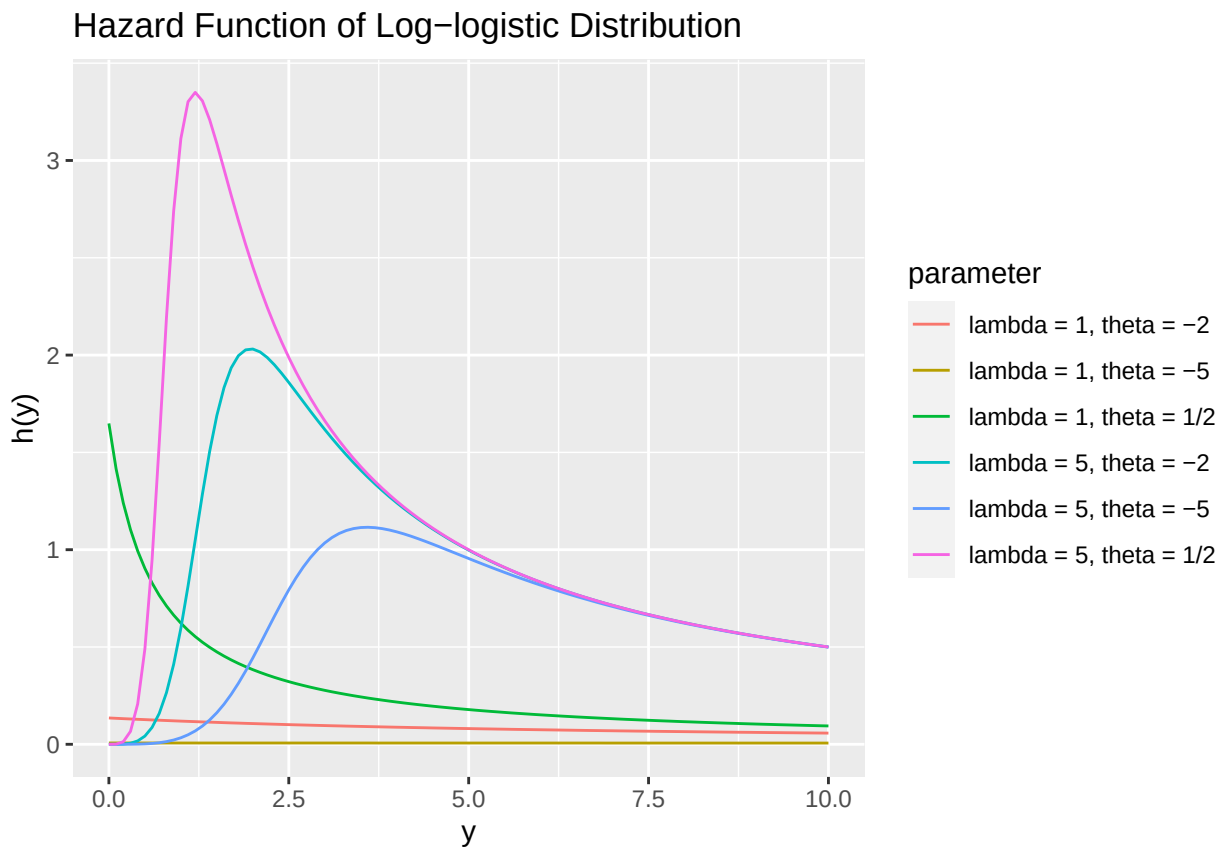
- (b) Show that the median survival time is $\exp(-\theta/\lambda)$.

Solution: See the pdf file attached.

- (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:

```
y <- seq(0, 10, by = 0.1)
loglog <- tibble(x = rep(y,6),
       pdf = c(
         hy1 = (exp(1/2) * 1 * y^(1-1)) / (1 + exp(1/2) * y^1),
         hy2 = (exp(-2) * 1 * y^(1-1)) / (1 + exp(-2) * y^1),
         hy3 = (exp(-5) * 1 * y^(1-1)) / (1 + exp(-5) * y^1),
         hy4 = (exp(1/2) * 5 * y^(5-1)) / (1 + exp(1/2) * y^5),
         hy5 = (exp(-2) * 5 * y^(5-1)) / (1 + exp(-2) * y^5),
         hy6 = (exp(-5) * 5 * y^(5-1)) / (1 + exp(-5) * y^5)),
       parameter = as.factor(rep(c("lambda = 1, theta = 1/2",
                                   "lambda = 1, theta = -2",
                                   "lambda = 1, theta = -5",
                                   "lambda = 5, theta = 1/2",
                                   "lambda = 5, theta = -2",
                                   "lambda = 5, theta = -5"),
                                 each = 101)))
ggplot(loglog) +
  geom_line(mapping = aes(x = x, y = pdf,
                          col = parameter)) +
  labs(x = "y", y = "h(y)", title = "Hazard Function of Log-logistic Distribution")
```



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)
```

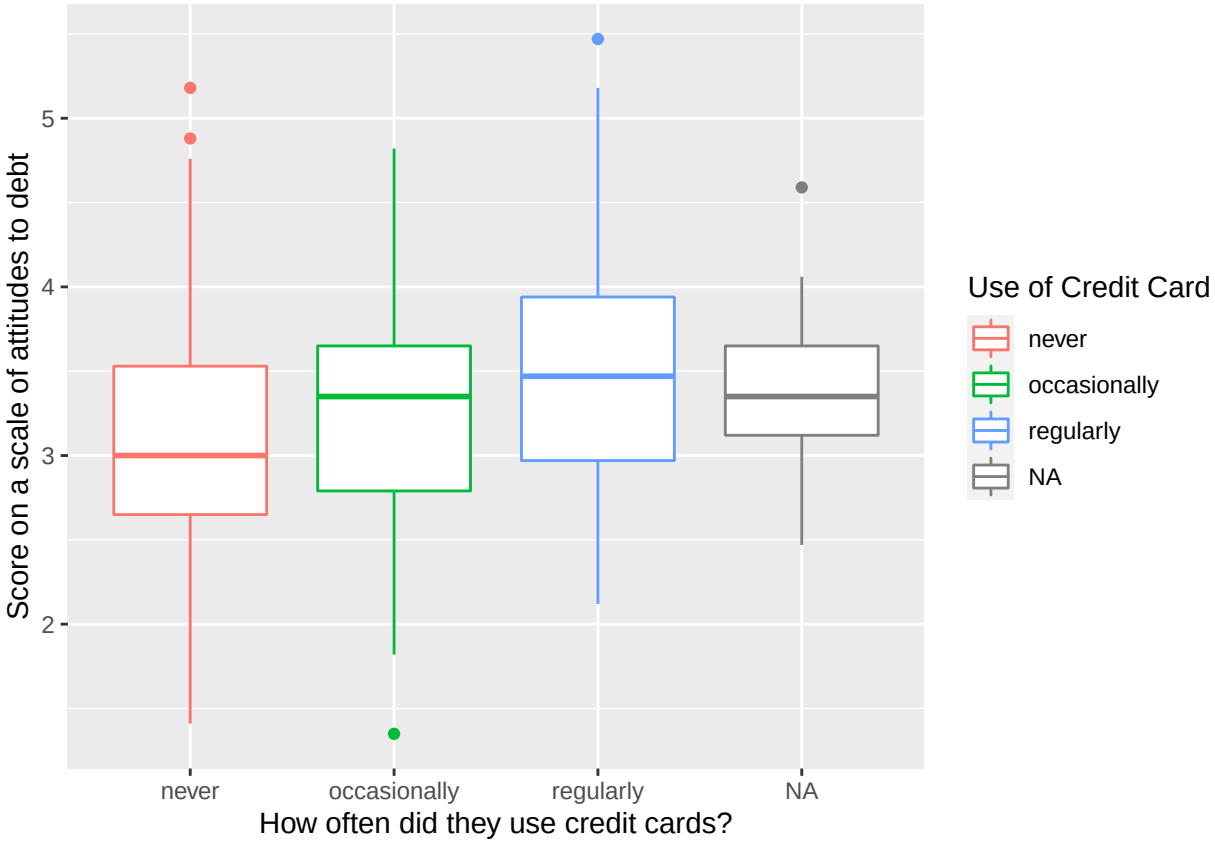
```
## starting httpd help server ... done
```

- (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:

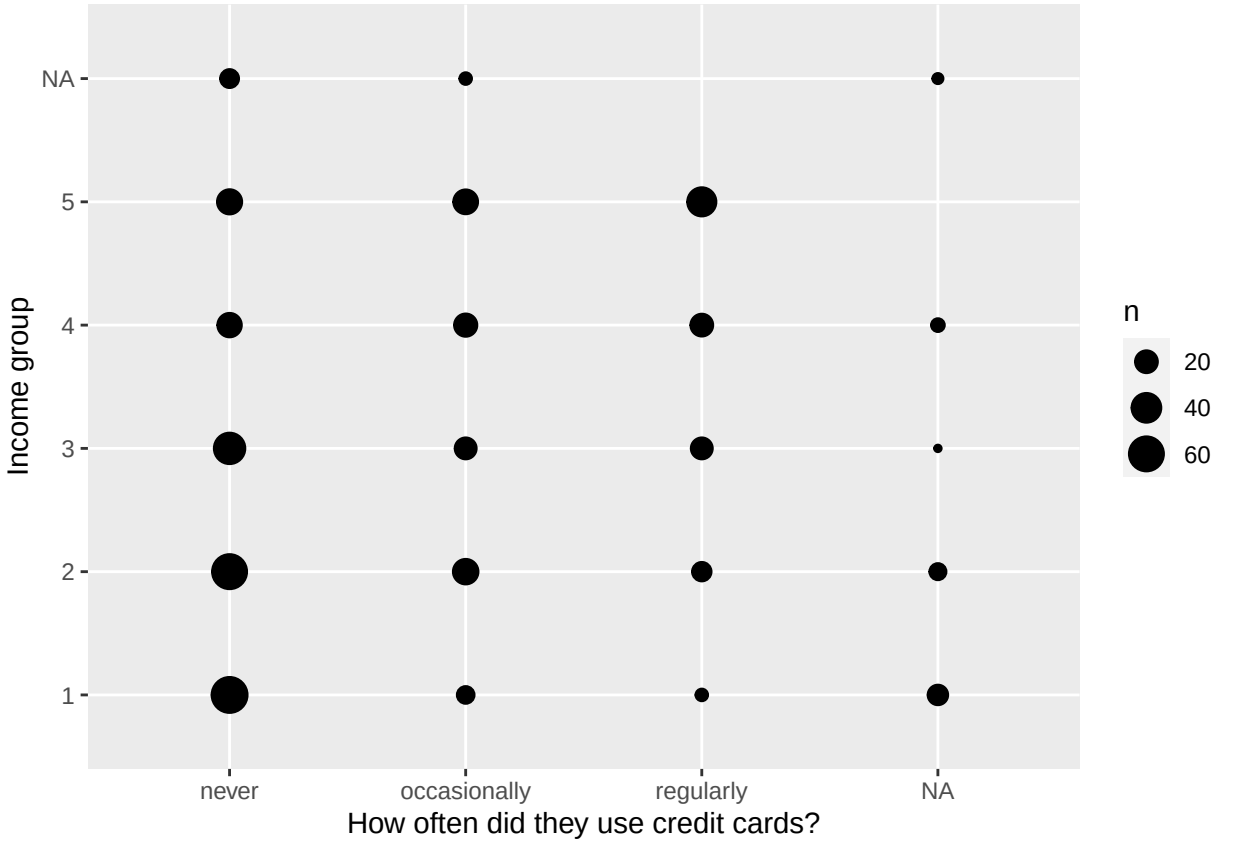
```
library(ggplot2)
ggplot(data = debt) +
    geom_boxplot(mapping = aes(y = prodebt, x = as.factor(ccarduse),
                               colour = as.factor(ccarduse))) +
  labs(x = "How often did they use credit cards?",
       y = "Score on a scale of attitudes to debt",
       colour = "Use of Credit Card") +
  scale_x_discrete(labels = c("never","occasionally","regularly","NA")) +
  scale_colour_discrete(labels = c("never","occasionally","regularly","NA"))
```

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



The relationship between `ccarduse` and `prodebt` is shown above in the box plots. We can observe that the more often of using credit cards, the higher scores on a scale of attitudes to debt.

```
ggplot(data = debt) +
  geom_count(mapping = aes(x = as.factor(ccarduse),
                           y = as.factor(incomegp))) +
  scale_x_discrete(labels = c("never","occasionally","regularly","NA")) +
  labs(y = "Income group", x = "How often did they use credit cards?")
```



The relationship between `ccarduse` and `incomegp` is shown above. In y-axis, income group = 1 had the lowest income, while income group = 5 had the highest income. From the graph above, we can observe that the higher income group tended to use credit cards more often; in the other hand, the lower income group tended to never use the credit cards.

- (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)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
pb <- polr(as.factor(ccarduse) ~ ., data = na.omit(debt))
summary(pb)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = as.factor(ccarduse) ~ ., data = na.omit(debt))
##
## Coefficients:
##             Value Std. Error t value
## incomegp  0.47131     0.1061  4.4423
## house     0.11600     0.2324  0.4992
## children -0.07872     0.1250 -0.6296
## singpar   0.88172     0.5971  1.4766
## agegp     0.20568     0.1576  1.3050
## bankacc   2.10270     0.5934  3.5435
## bsocacc   0.47322     0.2671  1.7715
## manage    0.18179     0.1653  1.0998
## cigbuy   -0.73546     0.2981 -2.4674
## xmasbuy   0.47014     0.4130  1.1385
## locintrn  0.11881     0.1424  0.8344
## prodebt   0.61046     0.1822  3.3497
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  7.9694  1.4752     5.4023
## 2|3  9.3944  1.5051     6.2417
##
## Residual Deviance: 511.673
## AIC: 539.673
```

From the results above, we can see that the two most significant predictors are `incomegp`(income group, 1=lowest and 5=highest) and `bankacc` (1 = the respondent have a bank account, 0 = the respondent do not have a bank account) since these two predictors have the highest absolute t-value.

When `incomegp` increases one unit, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{0.47131} = 1.60$. For the respondent who have a bank account, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{2.10270} = 8.19$ comparing to those who do not have a bank account.

The least significant predictor in this model is `house` because its corresponding t-value is closet to zero.

- (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:

```
pbsw <- step(pb)
```

```
## Start:  AIC=539.67
## as.factor(ccarduse) ~ incomegp + house + children + singpar +
##     agegp + bankacc + bsocacc + manage + cigbuy + xmasbuy + locintrn +
##     prodebt
##
##            Df    AIC
## - house     1 537.92
```

```
## - children  1 538.07
## - locintrn  1 538.37
## - manage    1 538.89
## - xmasbuy   1 539.00
## - agegp     1 539.38
## <none>        539.67
## - singpar   1 539.75
## - bsocacc   1 540.83
## - cigbuy    1 543.94
## - prodebt   1 549.30
## - bankacc   1 554.83
## - incomegp  1 558.37
##
## Step:  AIC=537.92
## as.factor(ccarduse) ~ incomegp + children + singpar + agegp +
##     bankacc + bsocacc + manage + cigbuy + xmasbuy + locintrn +
##     prodebt
##
##            Df    AIC
## - children  1 536.32
## - locintrn  1 536.57
## - xmasbuy   1 537.19
## - manage    1 537.23
## <none>        537.92
## - singpar   1 538.01
## - agegp     1 538.54
## - bsocacc   1 539.14
## - cigbuy    1 542.55
## - prodebt   1 547.61
## - bankacc   1 553.79
## - incomegp  1 557.55
##
## Step:  AIC=536.32
## as.factor(ccarduse) ~ incomegp + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + xmasbuy + locintrn + prodebt
##
##            Df    AIC
## - locintrn  1 535.01
## - xmasbuy   1 535.34
## - manage    1 535.71
## - singpar   1 536.23
## <none>        536.32
## - bsocacc   1 537.47
## - agegp     1 538.12
## - cigbuy    1 541.09
## - prodebt   1 545.83
## - bankacc   1 551.97
## - incomegp  1 556.19
##
## Step:  AIC=535.01
## as.factor(ccarduse) ~ incomegp + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + xmasbuy + prodebt
##
##            Df    AIC
```

```
## - xmasbuy   1 534.19
## - manage    1 534.58
## - singpar   1 534.90
## <none>        535.01
## - bsocacc   1 536.40
## - agegp     1 536.66
## - cigbuy    1 539.71
## - prodebt   1 543.93
## - bankacc   1 551.87
## - incomegp  1 555.76
##
## Step:  AIC=534.19
## as.factor(ccarduse) ~ incomegp + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + prodebt
##
##            Df    AIC
## - manage    1 533.90
## <none>        534.19
## - singpar   1 534.32
## - bsocacc   1 535.32
## - agegp     1 536.11
## - cigbuy    1 538.62
## - prodebt   1 543.71
## - bankacc   1 550.24
## - incomegp  1 556.78
##
## Step:  AIC=533.9
## as.factor(ccarduse) ~ incomegp + singpar + agegp + bankacc +
##     bsocacc + cigbuy + prodebt
##
##            Df    AIC
## - singpar   1 533.59
## <none>        533.90
## - bsocacc   1 536.04
## - agegp     1 536.27
## - cigbuy    1 539.16
## - prodebt   1 542.16
## - bankacc   1 551.27
## - incomegp  1 555.32
##
## Step:  AIC=533.59
## as.factor(ccarduse) ~ incomegp + agegp + bankacc + bsocacc +
##     cigbuy + prodebt
##
##            Df    AIC
## <none>        533.59
## - bsocacc   1 535.42
## - agegp     1 535.60
## - cigbuy    1 538.72
## - prodebt   1 542.25
## - bankacc   1 549.99
## - incomegp  1 553.43
```

```
summary(pbsw)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = as.factor(ccarduse) ~ incomegp + agegp + bankacc +
##     bsocacc + cigbuy + prodebt, data = na.omit(debt))
##
## Coefficients:
##            Value Std. Error t value
## incomegp  0.4589     0.1007   4.555
## agegp     0.2696     0.1352   1.993
## bankacc   2.0816     0.5753   3.618
## bsocacc   0.5048     0.2591   1.949
## cigbuy   -0.7677     0.2922  -2.627
## prodebt   0.5635     0.1755   3.211
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  5.9944  0.9961     6.0178
## 2|3  7.3948  1.0276     7.1961
##
## Residual Deviance: 517.5895
## AIC: 533.5895
```

The final model we selected using stepwise AIC has six predictors: `incomegp`, `agegp`, `bankacc`, `bsocacc`, `cigbuy`, `prodebt`.

When `incomegp` increases one unit, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{0.4589} = 1.58$.

Similarly, When `agegp` increases one unit, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{0.2696} = 1.31$.

For the respondent who have a bank account, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{2.0816} = 8.02$ comparing to those who do not have a bank account.

Similarly, for the respondent who have a building society account, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{0.5048} = 1.66$ comparing to those who do not have a building society account.

For the respondent who buys cigarettes, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{-0.7677} = 0.46$ comparing to those who does not buy cigarettes.

Finally, when `prodebt` increases one unit, the odds of the frequency of using credit cards going up by one unit (from normal to occasionally, or from occasionally to regularly) is $e^{0.5635} = 1.76$.

We cannot conclude that the predictors that were dropped from the model have no relation to the response. They might still have some relation to the response, but can be explained by other predictors. That is to say, the predictors might be correlated, so we only need some of them to explain the full model.

- (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:

```
newdata_smoker <- data.frame(
  incomegp = median(debt$incomegp, na.rm = TRUE),
  agegp = median(debt$agegp, na.rm = TRUE),
  bankacc = median(debt$bankacc, na.rm = TRUE),
  bsocacc = median(debt$bsocacc, na.rm = TRUE),
  prodebt = median(debt$prodebt, na.rm = TRUE),
  cigbuy = 1)
newdata_nonsmoker <- data.frame(
  incomegp = median(debt$incomegp, na.rm = TRUE),
  agegp = median(debt$agegp, na.rm = TRUE),
  bankacc = median(debt$bankacc, na.rm = TRUE),
  bsocacc = median(debt$bsocacc, na.rm = TRUE),
  prodebt = median(debt$prodebt, na.rm = TRUE),
  cigbuy = 0)
predict(pbsw, newdata_smoker, type = "prob")
```

```
##         1         2         3
## 0.6068704 0.2554381 0.1376915
```

```
predict(pbsw, newdata_nonsmoker, type = "prob")
```

```
##         1         2         3
## 0.4173804 0.3266237 0.2559959
```

After computing the median values of the predictors in the selected model, we got the probability of the smoking respondent had never used credit cards (0.61) is higher than the non-smoking respondent (0.42). The non-smoking respondent has a higher probability of using credit cards occasionally or regularly comparing to the smoking one.

- (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:

```
pe <- polr(as.factor(ccarduse) ~ incomegp + agegp + bankacc + bsocacc
           + cigbuy + prodebt , data = na.omit(debt), method = "cloglog")
summary(pe)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = as.factor(ccarduse) ~ incomegp + agegp + bankacc +
##     bsocacc + cigbuy + prodebt, data = na.omit(debt), method = "cloglog")
##
## Coefficients:
##            Value Std. Error t value
## incomegp  0.2454    0.05950   4.125
## agegp     0.1936    0.08224   2.354
```

```
## bankacc   0.9984    0.23658   4.220
## bsocacc   0.3087    0.15704   1.966
## cigbuy   -0.3120    0.15789  -1.976
## prodebt   0.3418    0.10872   3.143
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  3.0002  0.5307     5.6536
## 2|3  3.8261  0.5424     7.0541
##
## Residual Deviance: 527.372
## AIC: 543.372
```

```
predict(pe, newdata_smoker, type = "prob")
```

```
##         1         2         3
## 0.5497646 0.2886116 0.1616238
```

```
predict(pe, newdata_nonsmoker, type = "prob")
```

```
##         1         2         3
## 0.4424002 0.2941871 0.2634127
```

From the results above, we still got the same conclusion that the probability of the smoking respondent had never used credit cards (0.55) is higher than the non-smoking respondent (0.44), and the non-smoking respondent has a higher probability of using credit cards occasionally or regularly comparing to the smoking one. (For the respondent using credit cards occasionally, the probability are closer for the smokers and non-smokers.) Overall, it does not make a difference to use a proportional hazards model comparing to a proportional odds model.