SMM634 Answers - Model for Binary Data

1. (a)
```
install.packages("AER")
library(AER)
data(SwissLabor)
```

   (b) The variable `participation` must be converted to a numeric variable using `as.numeric()` as `glm()` does not accepts the dependent variable to be of class factor. This will turn `participation` = no into `participation` = 1 and `participation` = yes into `participation` = 2, so using -1 we obtain the values 0 and 1.

   (c)
```
swiss_lpm <- lm(participation.b ~ income + age + education, data = SwissLabor)
summary(swiss_lpm)

Call:
lm(formula = participation.b ~ income + age + education, data = SwissLabor)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7650 -0.4577 -0.2814  0.5013  0.8699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.691614   0.442119   6.088 1.72e-09 ***
income      -0.181068   0.042572  -4.253 2.34e-05 ***
age         -0.048036   0.015904  -3.020   0.0026 **
education   -0.011283   0.005849  -1.929   0.0541 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4888 on 868 degrees of freedom
Multiple R-squared:  0.04232,   Adjusted R-squared:  0.03901
F-statistic: 12.79 on 3 and 868 DF,  p-value: 3.507e-08
```

   (d) All coefficients are statistically significant, except for that associated with education which is marginally significant. As for the interpretation, the probability of participation decreases by 0.18 when income increase by 1% (keeping constant all other regressors); the probability of participation decreases by 0.05 when age increase by one unit (which in this case is a decade); finally, the probability of participation decreases by 0.01 if education increases by one year.

   (e)
```
swiss_probit <- glm(participation.b ~ income + age + education,
+                     family = binomial(link = "probit"), data = SwissLabor)
> summary(swiss_probit)

Call:
glm(formula = participation.b ~ income + age + education,
    family = binomial(link = "probit"), data = SwissLabor)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6995  -1.1041  -0.8071   1.1810   1.9059

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.96528    1.22159   4.883 1.04e-06 ***
income      -0.49631    0.11731  -4.231 2.33e-05 ***
age         -0.12588    0.04163  -3.024   0.0025 **
```

```
education   -0.02855    0.01533  -1.862   0.0625 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1164.8  on 868  degrees of freedom
AIC: 1172.8

Number of Fisher Scoring iterations: 4

>
> swiss_logit <- glm(participation.b ~ income + age + education,
+                    family = binomial,data = SwissLabor)
> summary(swiss_logit)

Call:
glm(formula = participation.b ~ income + age + education, family = binomial,
    data = SwissLabor)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6918  -1.1044  -0.8097   1.1800   1.8826

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.61688    2.02506   4.749 2.04e-06 ***
income      -0.80014    0.19405  -4.123 3.73e-05 ***
age         -0.20171    0.06731  -2.997  0.00273 **
education   -0.04636    0.02477  -1.871  0.06130 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1165.0  on 868  degrees of freedom
AIC: 1173

Number of Fisher Scoring iterations: 4
```

(f) They are not comparable because the effects are measured on a different scale.

(g)
```
predictions_p <- predict(swiss_probit,
                         newdata = data.frame("income" = c(10.70, 10.70),
 "age" = c(4,4), "education" = c(11, 12)),
                         type = "response")

diff(predictions_p)
          2
-0.01121074


predictions_l <- predict(swiss_logit,
                         newdata = data.frame("income" = c(10.70, 10.70),
                         "age" = c(4,4),
```

```
                          "education" = c(11, 12)),
                          type = "response")

diff(predictions_l)
          2
-0.01135786
```

The effects on the probability are the same for the logit and probit model; that is when education increases from 11 to 12 years, the probability of participation decreases by 0.01.

(h)
```
predictions_p <- predict(swiss_probit,
+                          newdata = data.frame("income" = c(10.70, 10.70),
+                          "age" = c(4,4), "education" = c(1, 2)),
+                          type = "response")
>
> diff(predictions_p)
          2
-0.01132117
>
>
> predictions_l <- predict(swiss_logit,
+                          newdata = data.frame("income" = c(10.70, 10.70),
+                          "age" = c(4,4),
+                          "education" = c(1, 2)),
+                           type = "response")
>
> diff(predictions_l)
          2
-0.01149688
```

The effects do not change. They seem to be constant.

(i) The effect of education on the probability of `participation` is -0.011283 in the linear probability model which is the same as those obtained using a logit and a probit.