ASSIGNMENT 4

1. Eitas Rimkus (u184503)
2. Nikodem Baehr (u459229)
3. Samuel Friedlaender (u848264)

Question 1

1

In order to get the Likelihood function for sample $x_1, ...x_n$ as a function of parameters α and x_m we need to take take the derivative of the $F(x)$ with respect to x. Thus we get the pdf of the Pareto as: $\frac{\delta F(x)}{\delta x} = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & \text{if } x \geq x_m \\ 0 & \text{if } x < x_m \end{cases}$.

So, then our Likelihood function is $L(x; \alpha, x_m) = \prod_{i=1}^n \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$

2

Since $x_m \leq min(x)$ and since the objecrive function which is the Likelihood function is monotonically increasing in x_m we can maxizimize the likelihood function by the $X_{1:n}$ so $\hat{x_m} = X_{1:n}$.

3

When $x \geq x_m$, we have that the log-likelihood function is $\log(L(x_i; \alpha, x_m)) = \sum_{i=1}^n \log(\frac{\alpha x_m^\alpha}{x^{\alpha+1}}) = n\log(\alpha) + n\alpha\log(x_m) - (\alpha+1)\sum_{i=1}^n \log(x_i)$.

The FOC is $\frac{\delta}{\delta\alpha}\log(L(x_i; \alpha, x_m)) = \frac{n}{\alpha} + n\log(x_m) - \sum_{i=1}^n \log(x_i) = 0$.

So, we get that maximum likelihood estimator for α is $\hat{\alpha} = \frac{n}{\sum_{i=1}^n \log(\frac{x_i}{x_m})}$

4

Using the distribution of $\hat{\alpha}_n$ which is $\hat{\alpha} = N(\alpha, \frac{1}{nI(\alpha)})$. We firstly calculate the Fisher Information $I(\alpha)$. From Question 3 we know that the $\frac{\delta}{\delta\alpha}\log(L(x_i; \alpha, x_m)) = \frac{n}{\alpha} + n\log(x_m) - \sum_{i=1}^n \log(x_i)$ Now if we only consider one x and we take the second derivative, so we obtain: $\frac{\delta^2}{\delta\alpha^2}\log(f(x; \alpha, x_m)) = -\frac{1}{\alpha^2}$. So using one of the definitions of Fisher information, $I(\alpha) = -E[\frac{\delta^2}{\delta\alpha^2}\log(f(x; \alpha, x_m))]$ we get that $I(\alpha) = \frac{1}{\alpha^2}$. So then the asymptotic varience of the MLE estimator $\hat{\alpha}_n$ is α^2/n. Thus the standard error is the square root so $\sigma(\hat{\alpha}_n) = \frac{\hat{\alpha}}{\sqrt{n}}$.

Question 2

1

```r
summary(docvis)
```

```
##      docvis             age            income            female      
##  Min.   :  0.000   Min.   :2.500   Min.   :-50.00   Min.   :0.0000  
##  1st Qu.:  0.000   1st Qu.:3.200   1st Qu.: 16.00   1st Qu.:0.0000  
##  Median :  1.000   Median :4.000   Median : 27.00   Median :0.0000  
##  Mean   :  3.957   Mean   :4.083   Mean   : 34.34   Mean   :0.4719  
##  3rd Qu.:  5.000   3rd Qu.:4.800   3rd Qu.: 43.17   3rd Qu.:1.0000  
##  Max.   :134.000   Max.   :6.400   Max.   :280.78   Max.   :1.0000  
##      black            hispanic         married          physlim      
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.00000   Median :0.0000   Median :1.0000   Median :0.0000  
##  Mean   :0.05281   Mean   :0.2452   Mean   :0.6358   Mean   :0.1657  
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##     private          chronic      
##  Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :1.0000   Median :0.0000  
##  Mean   :0.7854   Mean   :0.3264  
##  3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000
```

```r
a1=sum(docvis$docvis==0)
a1
```

```
## [1] 1606
```

1606 people from the sample have not been to a doctor.

2

```r
Any = I(docvis$docvis>0)
Any[Any == "True"] = 1
Age = docvis$age
Female = docvis$female
```

```r
Chronic = docvis$chronic
Married = docvis$married
a2.1 = glm(formula=Any~Age+Female+Chronic+Married+Age*Chronic+Age*Married+Age*Female+Female*Married+Fema
summary(a2.1)
```

```
##
## Call:
## glm(formula = Any ~ Age + Female + Chronic + Married + Age *
##     Chronic + Age * Married + Age * Female + Female * Married +
##     Female * Chronic + Chronic * Married, family = binomial(link = "probit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4206  -1.0291   0.4560   0.8905   1.5689
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.97088    0.16239  -5.979 2.25e-09 ***
## Age              0.16941    0.04197   4.036 5.43e-05 ***
## Female           1.20128    0.17729   6.776 1.24e-11 ***
## Chronic          0.86029    0.22294   3.859 0.000114 ***
## Married          0.05125    0.17649   0.290 0.771523
## Age:Chronic      0.04413    0.04936   0.894 0.371325
## Age:Married      0.03500    0.04344   0.806 0.420411
## Age:Female      -0.17852    0.04289  -4.162 3.15e-05 ***
## Female:Married   0.09053    0.08774   1.032 0.302173
## Female:Chronic   0.01770    0.10182   0.174 0.861962
## Chronic:Married -0.04366    0.10452  -0.418 0.676180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5785.8  on 4411  degrees of freedom
## Residual deviance: 4936.4  on 4401  degrees of freedom
## AIC: 4958.4
##
## Number of Fisher Scoring iterations: 5
```

```r
a2.2 = glm(Any~Age+Female+Chronic+Age*Female+Married,family=binomial(link="probit"))
summary(a2.2)
```

```
##
```

```
## Call:
## glm(formula = Any ~ Age + Female + Chronic + Age * Female + Married,
##     family = binomial(link = "probit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3436  -1.0388   0.4590   0.9019   1.6151
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.10393    0.11451  -9.640  < 2e-16 ***
## Age          0.19810    0.02798   7.081 1.43e-12 ***
## Female       1.24530    0.17265   7.213 5.47e-13 ***
## Chronic      1.02603    0.05018  20.449  < 2e-16 ***
## Married      0.22079    0.04359   5.065 4.08e-07 ***
## Age:Female  -0.17524    0.04158  -4.215 2.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5785.8  on 4411  degrees of freedom
## Residual deviance: 4939.0  on 4406  degrees of freedom
## AIC: 4951
##
## Number of Fisher Scoring iterations: 4
```

We assume that older people people and people with a chronic disease are more likely to visit the doctor's (positive correlation). The older people get, the more likely they have to visit the doctor. Same thing holds for people visit a chronic disease due to mandatory check-ups.

The regression confirms it with coefficients for age and chronic having positive signs (0.19810,1.02603 respectively). The explanatory variables are significant as the p-values are less than 0.05.

3

```
library('glmx')
help("hetglm")
a3 = hetglm(Any~Age+Female+Chronic+Married+Age*Female|Age+Female+Chronic+Married, family=binomial(link=
summary(a3)
```

```
##
## Call:
```

```
## hetglm(formula = Any ~ Age + Female + Chronic + Married + Age * Female |
##     Age + Female + Chronic + Married, family = binomial(link = "probit"))
##
## Deviance residuals:
##     Min      1Q  Median      3Q     Max
## -2.4631 -1.0190  0.4517  0.8751  1.5625
##
## Coefficients (binomial model with probit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.82335    0.16530  -4.981 6.32e-07 ***
## Age          0.14996    0.03345   4.484 7.34e-06 ***
## Female       1.00412    0.18006   5.577 2.45e-08 ***
## Chronic      1.26573    0.54139   2.338   0.0194 *
## Married      0.14250    0.04488   3.175   0.0015 **
## Age:Female  -0.16672    0.03310  -5.037 4.73e-07 ***
##
## Latent scale model coefficients (with log link):
##         Estimate Std. Error z value Pr(>|z|)
## Age     -0.07338    0.04221  -1.738   0.0821 .
## Female  -0.23939    0.13017  -1.839   0.0659 .
## Chronic  0.64759    0.51489   1.258   0.2085
## Married -0.09125    0.08288  -1.101   0.2709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-likelihood: -2467 on 10 Df
## LR test for homoscedasticity: 4.044 on 4 Df, p-value: 0.4001
## Dispersion: 1
## Number of iterations in nlminb optimization: 8
```

Given that a p-value is greater than 0.05, at $\alpha = 0.05$ we don't reject H_0 s.t. the model is homoskedastic. Hence, we should continue with the homoskedastic model.

4

```r
a4=glm(Any~Age+Female+Chronic+Married+Age*Female, family=binomial(link="logit"))
summary(a4)
```

```
##
## Call:
## glm(formula = Any ~ Age + Female + Chronic + Married + Age *
##     Female, family = binomial(link = "logit"))
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2963  -1.0349   0.4652   0.8952   1.6211
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.80781    0.19077  -9.477  < 2e-16 ***
## Age          0.32276    0.04652   6.939 3.96e-12 ***
## Female       2.04973    0.28978   7.073 1.51e-12 ***
## Chronic      1.75470    0.09085  19.314  < 2e-16 ***
## Married      0.36560    0.07284   5.019 5.19e-07 ***
## Age:Female  -0.28828    0.07012  -4.111 3.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5785.8  on 4411  degrees of freedom
## Residual deviance: 4940.3  on 4406  degrees of freedom
## AIC: 4952.3
##
## Number of Fisher Scoring iterations: 4
```

The signs of explanatory variables are the same as if in model from *Question 2.2*. The magnitude of estimates is different.

5

```r
library("mfx")
probitmfx(Any~Age+Female+Chronic+Married+Age*Female, docvis)
```

```
## Call:
## probitmfx(formula = Any ~ Age + Female + Chronic + Married +
##     Age * Female, data = docvis)
##
## Marginal Effects:
##                dF/dx Std. Err.       z     P>|z|
## Age         0.072135  0.010207  7.0673 1.580e-12 ***
## Female      0.424732  0.052393  8.1067 5.202e-16 ***
## Chronic     0.331499  0.013379 24.7779 < 2.2e-16 ***
## Married     0.081259  0.016172  5.0248 5.040e-07 ***
```

```
## Age:Female -0.063813  0.015159 -4.2096 2.559e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "Female"  "Chronic" "Married"
```

```
logitmfx(Any~Age+Female+Chronic+Married+Age*Female,docvis)
```

```
## Call:
## logitmfx(formula = Any ~ Age + Female + Chronic + Married + Age *
##     Female, data = docvis)
##
## Marginal Effects:
##                dF/dx Std. Err.       z     P>|z|
## Age         0.071020  0.010272  6.9142 4.704e-12 ***
## Female      0.419672  0.052417  8.0064 1.182e-15 ***
## Chronic     0.333056  0.013358 24.9332 < 2.2e-16 ***
## Married     0.081640  0.016453  4.9620 6.977e-07 ***
## Age:Female -0.063433  0.015464 -4.1020 4.095e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "Female"  "Chronic" "Married"
```

Marginal effects at average for probit and logit for age and chronic have the same signs and similar magnitude (max. difference of 0.008). The marginal effects for female have the same sign and differ by 0.01 in magnitude.

The marginal effects for age and chronic are computed in a similar fashion, but a marginal effect for each explanatory variable is found by taking a derivative w.r.t to the variable itself.

6

The marginal effect of age for men, using the probit model, is that on average if a man gets one year older, the probability that he will visit a doctor at least once increases by 0.072135 (7.2pp), ceteris paribus.

Given that a man has a chronic disease, using the probit model, on average, the probability that he will visit a doctor at least once increases by 0.331499 (33.1pp), ceteris paribus. #The marginal effect of age for men, using the logit model, is that on average if a man gets one year older, the probability that he will visit a doctor at least once increases by 0.071020 (7.1pp), ceteris paribus.

Given that a man has a chronic disease, using the logit model, on average, the probability that he will visit a doctor at least once increases by 0.333056 (33.3pp).

7

```r
predict=a2.2$fitted.values>0.5
table(Any,predict)
```

```
##    predict
## Any FALSE TRUE
##   0   810  796
##   1   515 2291
```

```r
mean(Any==predict)
```

```
## [1] 0.7028558
```

Model correctly classified 70.3% of observations. It is better than the cut-off point of 50%.