Biostat 200C Homework 5

Due 11:59PM June 2nd

Jiahao Tian

Q1. Doctor visits in Australia, ELMR Exercise 13.4

```
data(dvisits)
help("dvisits")
```

The `dvisits` data comes from the Australian Health Survey of 1977–78 and consist of 5190 single adults where young and old have been oversampled. Use `help("dvisits")` to check the variables.

(a) Build a generalized additive model with `doctorco` as the response and `sex`, `age`, `agesq`, `income`, `levyplus`, `freepoor`, `freerepa`, `illness`, `actdays`, `hscore`, `chcond1` and `chcond2` as possible predictor variables. Select an appropriate size for your model. (Hint. fit a simpler model first and check some marginal plots.)

Solution:

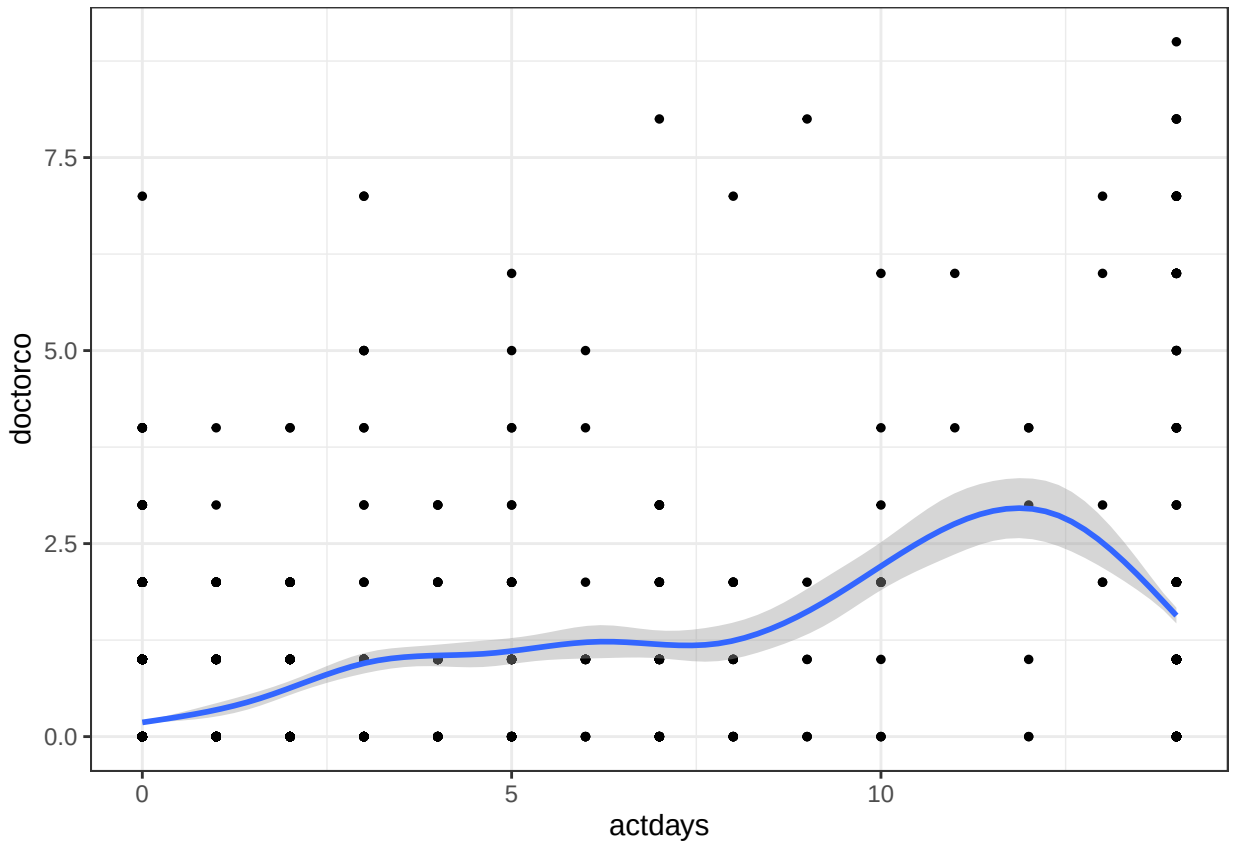
```
summary(dvisits)
```

```
##       sex              age              agesq            income      
##  Min.   :0.0000   Min.   :0.1900   Min.   :0.0361   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.2200   1st Qu.:0.0484   1st Qu.:0.2500  
##  Median :1.0000   Median :0.3200   Median :0.1024   Median :0.5500  
##  Mean   :0.5206   Mean   :0.4064   Mean   :0.2071   Mean   :0.5832  
##  3rd Qu.:1.0000   3rd Qu.:0.6200   3rd Qu.:0.3844   3rd Qu.:0.9000  
##  Max.   :1.0000   Max.   :0.7200   Max.   :0.5184   Max.   :1.5000  
##     levyplus         freepoor          freerepa         illness      
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000  
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :1.000  
##  Mean   :0.4428   Mean   :0.04277   Mean   :0.2102   Mean   :1.432  
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:2.000  
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :5.000  
##     actdays           hscore          chcond1          chcond2      
##  Min.   : 0.0000   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median : 0.0000   Median : 0.000   Median :0.0000   Median :0.0000  
##  Mean   : 0.8619   Mean   : 1.218   Mean   :0.4031   Mean   :0.1166  
##  3rd Qu.: 0.0000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :14.0000   Max.   :12.000   Max.   :1.0000   Max.   :1.0000  
##     doctorco         nondocco          hospadmi         hospdays     
##  Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.000  
```

```
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median :0.0000   Median : 0.0000   Median :0.0000   Median : 0.000
##  Mean   :0.3017   Mean   : 0.2146   Mean   :0.1736   Mean   : 1.334
##  3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.000
##  Max.   :9.0000   Max.   :11.0000   Max.   :5.0000   Max.   :80.000
##     medicine        prescrib         nonpresc
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.0000   Median :0.0000
##  Mean   :1.218   Mean   :0.8626   Mean   :0.3557
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :8.000   Max.   :8.0000   Max.   :8.0000
```

```
dvisits %>%
  ggplot(mapping = aes(x = actdays, y = doctorco)) +
  geom_point(size = 1) +
  geom_smooth() +
  theme_bw()
```

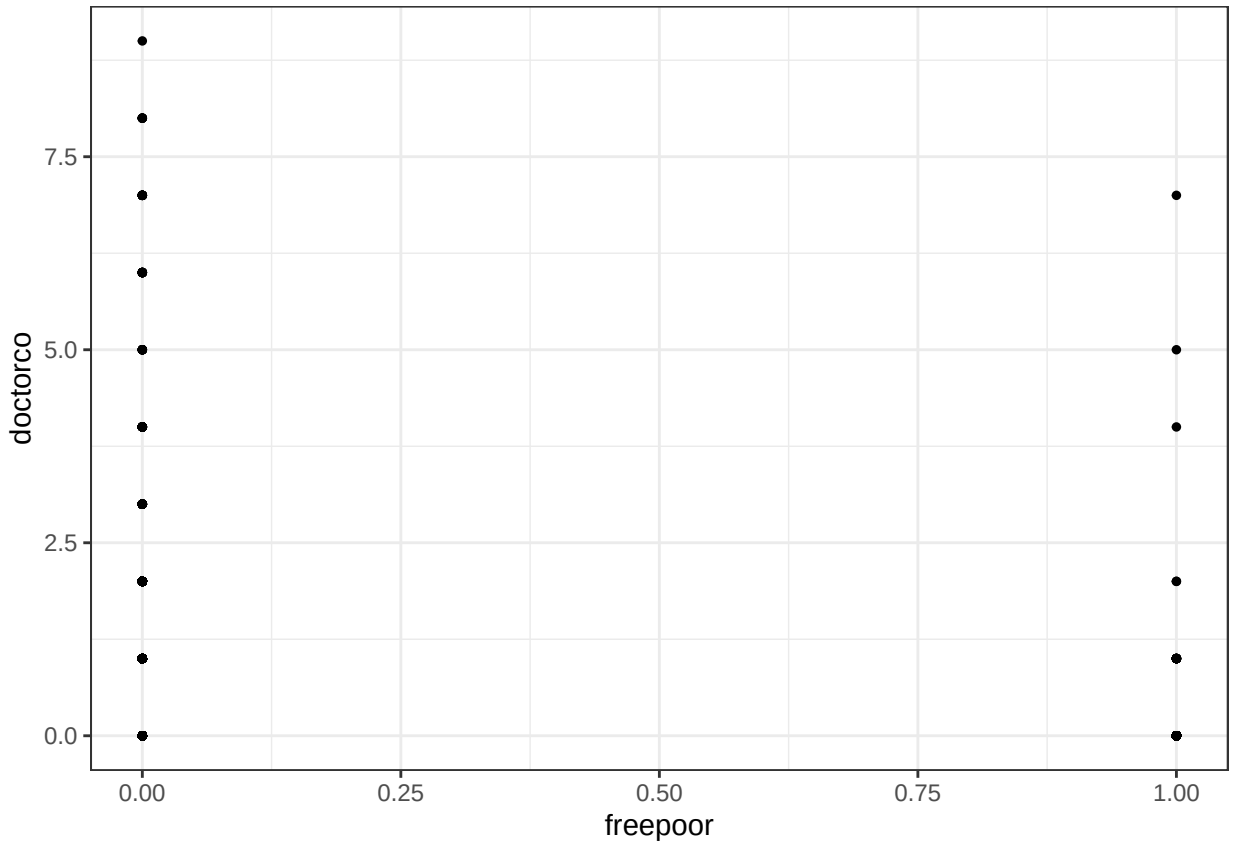
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
dvisits %>%
  ggplot(mapping = aes(x = freepoor, y = doctorco)) +
  geom_point(size = 1) +
  geom_smooth() +
  theme_bw()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```




```
#Lots of binary, no need to do transformation.
```

```
#It is not necessary to use age squared term, when we used age term and transformed it.

dvgam <- gam(doctorco ~ sex + s(age) + s(income) + levyplus + freepoor +
               freerepa + factor(illness) + s(actdays) + s(hscore) +
               chcond1 + chcond2, family = poisson, scale = -1, data = dvisits)
summary(dvgam)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## doctorco ~ sex + s(age) + s(income) + levyplus + freepoor + freerepa +
##     factor(illness) + s(actdays) + s(hscore) + chcond1 + chcond2
##
## Parametric coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.46949    0.12168 -20.295  < 2e-16 ***
```

```
## sex                0.13038    0.06461   2.018   0.0437 *
## levyplus           0.09819    0.08293   1.184   0.2365
## freepoor          -0.46847    0.20785  -2.254   0.0242 *
## freerepa           0.10384    0.10759   0.965   0.3345
## factor(illness)1   0.88936    0.12011   7.405 1.53e-13 ***
## factor(illness)2   1.09731    0.12551   8.743  < 2e-16 ***
## factor(illness)3   0.99274    0.13766   7.212 6.33e-13 ***
## factor(illness)4   1.10434    0.14937   7.393 1.66e-13 ***
## factor(illness)5   1.19746    0.14881   8.047 1.04e-15 ***
## chcond1            0.04913    0.07638   0.643   0.5201
## chcond2            0.10329    0.09280   1.113   0.2658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df      F p-value
## s(age)     1.000  1.001  8.837 0.00296 **
## s(income)  2.291  2.875  2.297 0.08168 .
## s(actdays) 6.087  7.132 82.310 < 2e-16 ***
## s(hscore)  1.001  1.003  9.537 0.00201 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =    0.2   Deviance explained = 27.5%
## GCV = 0.79375  Scale est. = 1.315     n = 5190
```

The variables which are not significant are `levyplus, chcond1, chcond2, freerepa`

The final model; `doctorco ~ sex + age + s(income) + freepoor + factor(illness) + s(actdays) + s(hscore)`

(b) Check the diagnostics.

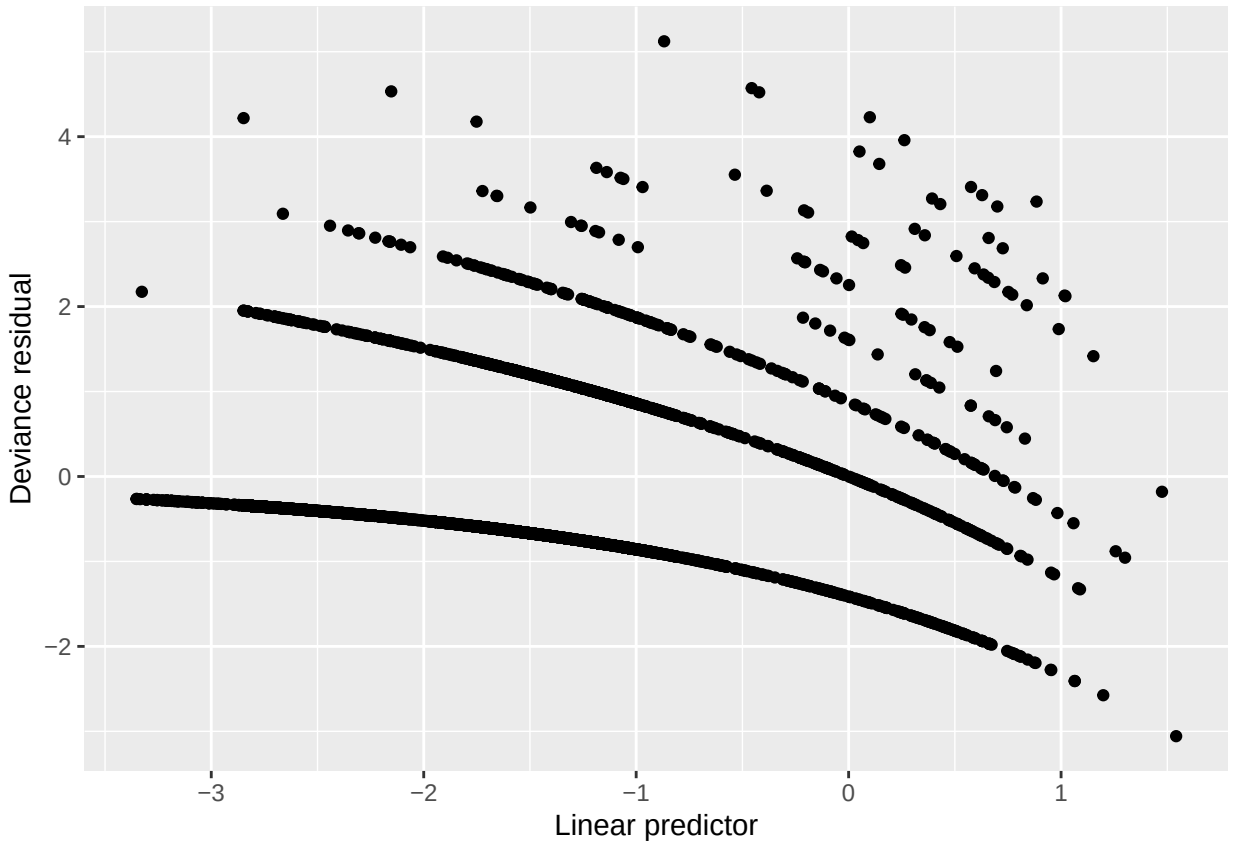
Solution:

```
dvgam2 <- gam(doctorco ~ sex + age + s(income) + factor(illness) +
                s(actdays) + s(hscore) + freepoor,
              family = poisson, scale = -1, data = dvisits)
anova(dvgam2, dvgam, test = "Chisq")
```

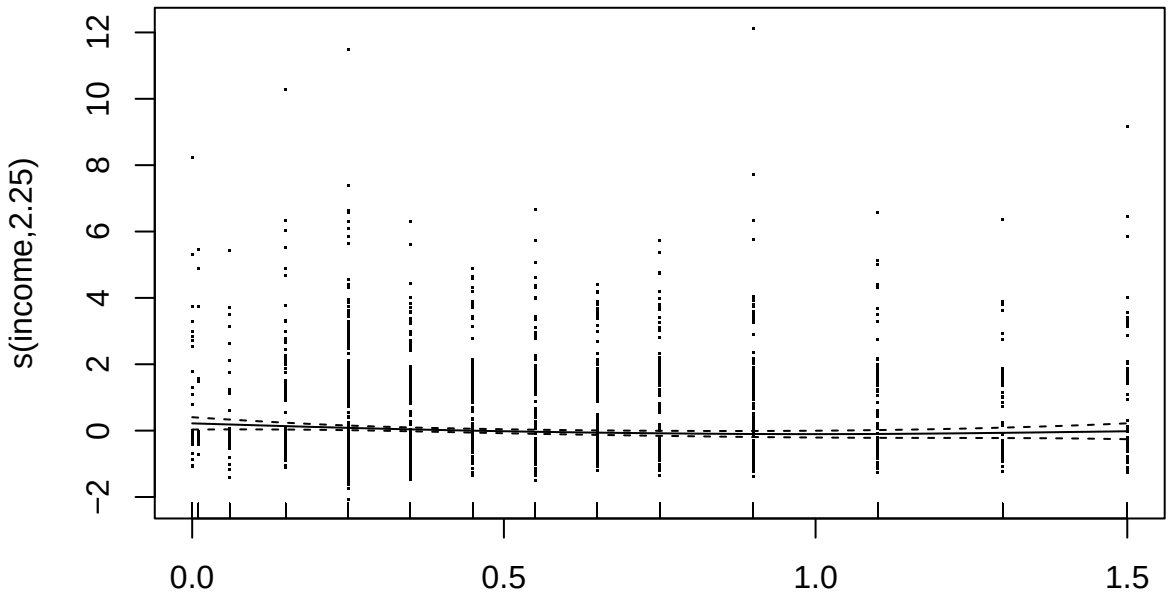
```
## Analysis of Deviance Table
##
## Model 1: doctorco ~ sex + age + s(income) + factor(illness) + s(actdays) +
##     s(hscore) + freepoor
## Model 2: doctorco ~ sex + s(age) + s(income) + levyplus + freepoor + freerepa +
##     factor(illness) + s(actdays) + s(hscore) + chcond1 + chcond2
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    5170.1     4088.4
## 2    5166.0     4084.1 4.0838   4.2202   0.5365
```

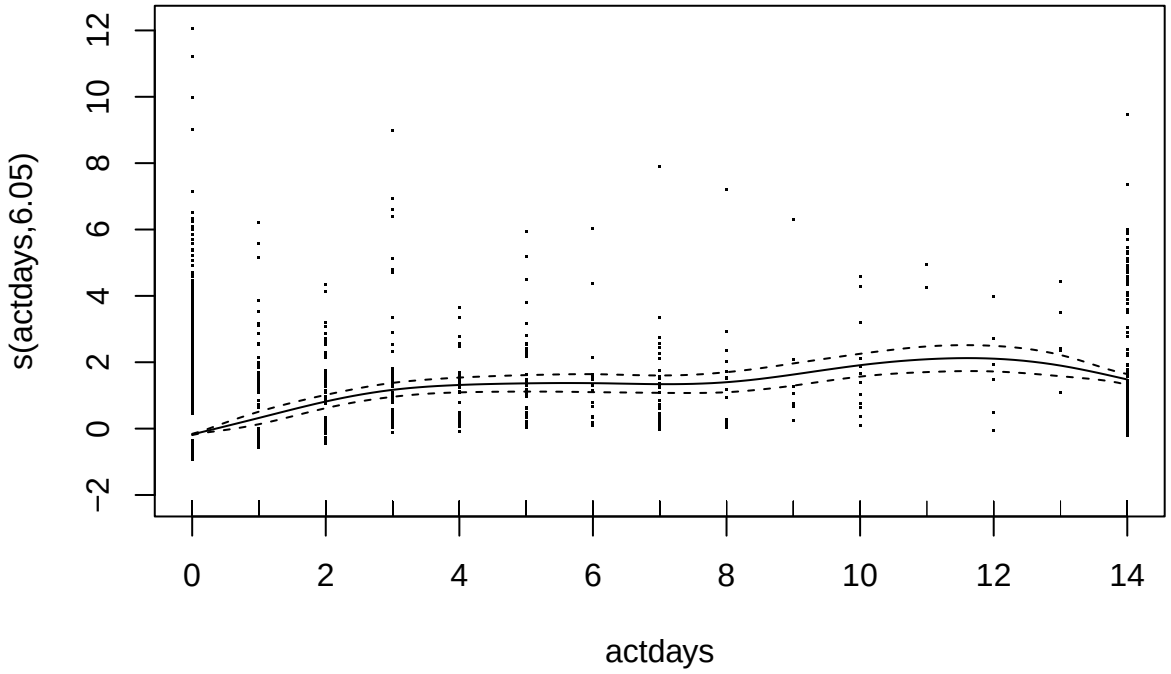
From the p-value 0.5365, which tells that there is not a significant differences between these two model, both of them explain the same variances.

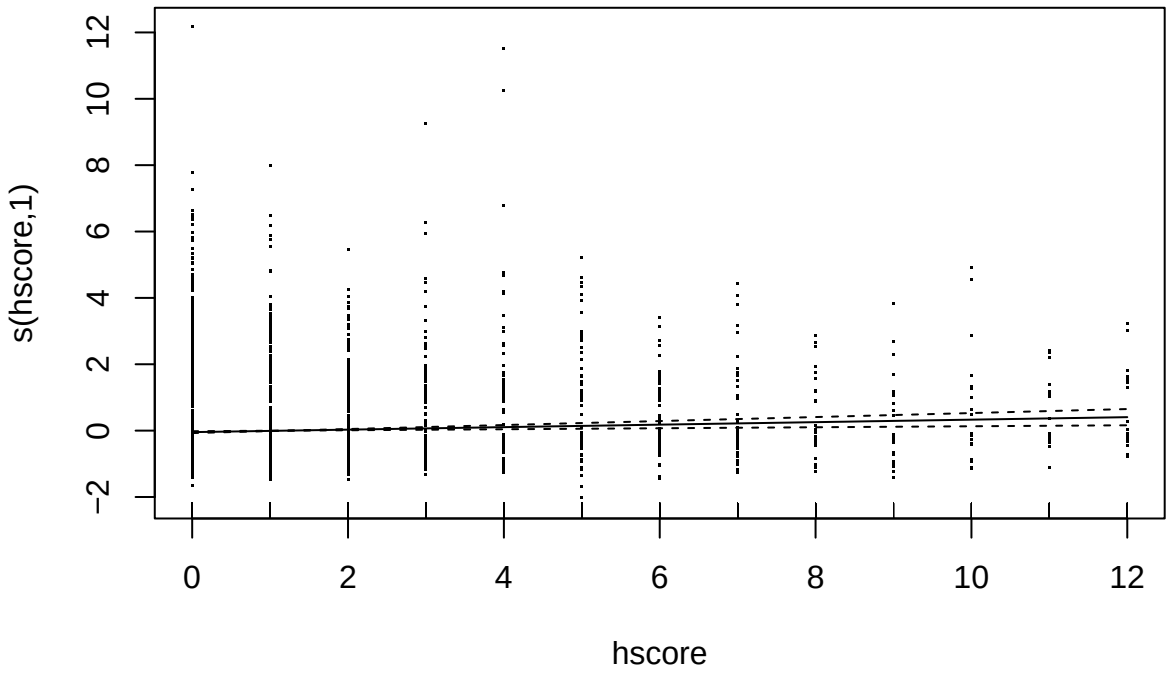
```
dvisits %>%
mutate(devres = residuals(dvgam2, type = "deviance"),
linpred = predict(dvgam2, type = "link")) %>%
ggplot + geom_point(mapping = aes(x = linpred, y = devres)) +
  labs(x = "Linear predictor", y = "Deviance residual")
```




```
plot(dvgam2, residuals = TRUE)
```


s(income,2.25)
12
10
8
6
4
2
0
−2
0.0
0.5
1.0
1.5
income


s(actdays,6.05)
12
10
8
6
4
2
0
−2
0
2
4
6
8
10
12
14
actdays



(c) What sort of person would be predicted to visit the doctor the most under your selected model?

Solution:

```
summary(dvgam2)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## doctorco ~ sex + age + s(income) + factor(illness) + s(actdays) +
##     s(hscore) + freepoor
##
## Parametric coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.68307    0.12625 -21.253  < 2e-16 ***
## sex               0.14539    0.06361   2.286  0.02230 *
## age               0.71855    0.16279   4.414 1.04e-05 ***
## factor(illness)1  0.90184    0.11900   7.578 4.13e-14 ***
## factor(illness)2  1.12330    0.12311   9.124  < 2e-16 ***
## factor(illness)3  1.01974    0.13422   7.597 3.56e-14 ***
## factor(illness)4  1.13035    0.14644   7.719 1.40e-14 ***
## factor(illness)5  1.22750    0.14534   8.446  < 2e-16 ***
## freepoor         -0.52418    0.20208  -2.594  0.00952 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F  p-value
## s(income)  2.251  2.827  2.64 0.053921 .
```

```
## s(actdays) 6.046  7.091 86.61  < 2e-16 ***
## s(hscore)  1.005  1.009 10.94 0.000926 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.202   Deviance explained = 27.4%
## GCV = 0.79332  Scale est. = 1.3122    n = 5190
```

We see that the type of patient most likely to visit a doctor:

- female
- older people
- low income
- has recent illness
- has reduced recent days of activity due to injury or illness
- has bad health according to the Goldberg scale
- is covered free by government because of old-age or disability pension
- is not covered by government because of lower income

(d) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1, 2, etc. times.

Solution:

```
(lastp <- predict(dvgam2, type = "response")[nrow(dvisits)])
```

```
##       5190
## 0.09990163
```

```
lp <- round(dpois(0 : 4, lastp), 3)
names(lp) <- 0 : 4
lp
```

```
##     0     1     2     3     4
## 0.905 0.090 0.005 0.000 0.000
```