Biostat 200C Homework 5

Due 11:59PM June 2nd

Yenlin Lai

Q1. Doctor visits in Australia, ELMR Exercise 13.4

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

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

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.)

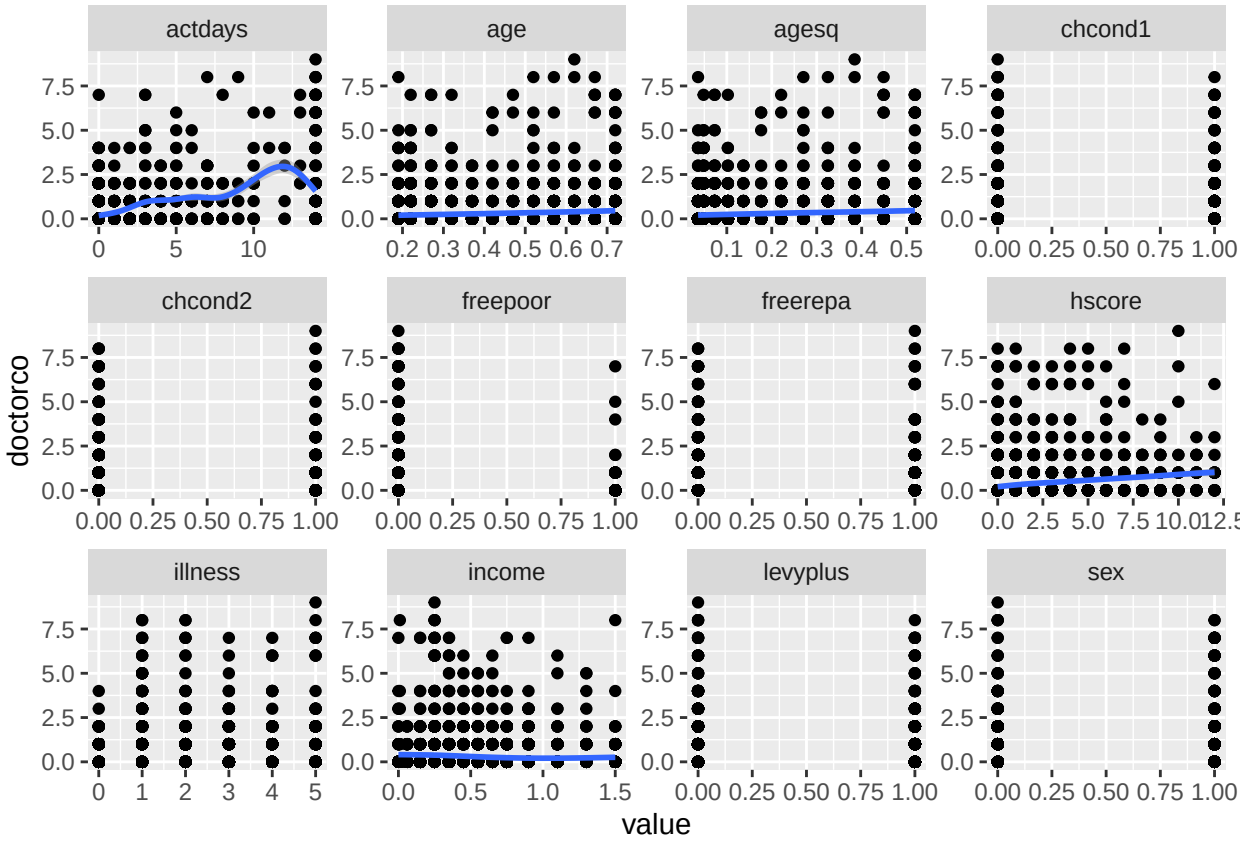
Solution: Check the marginal plots first. From the marginal plots below, we can see that some of the variables (`sex`, `levyplus`, `freepoor`, `freerepa`, `chcond1` and `chcond2`) cannot fit a spline/smooth in the generalized additive model.

```
predlist <- setdiff(colnames(dvisits),
                    c("doctorco", "nondocco", "hospadmi", "hospdays",
                      "medicine", "prescrib", "nonpresc"))
dvisits %>%
  pivot_longer(all_of(predlist),
               names_to = "predictor", values_to = "value") %>%
  ggplot() +
  geom_point(aes(x = value, y = doctorco)) +
  facet_wrap(~ predictor, scales = "free") +
  geom_smooth(aes(x = value, y = doctorco))
```

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

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



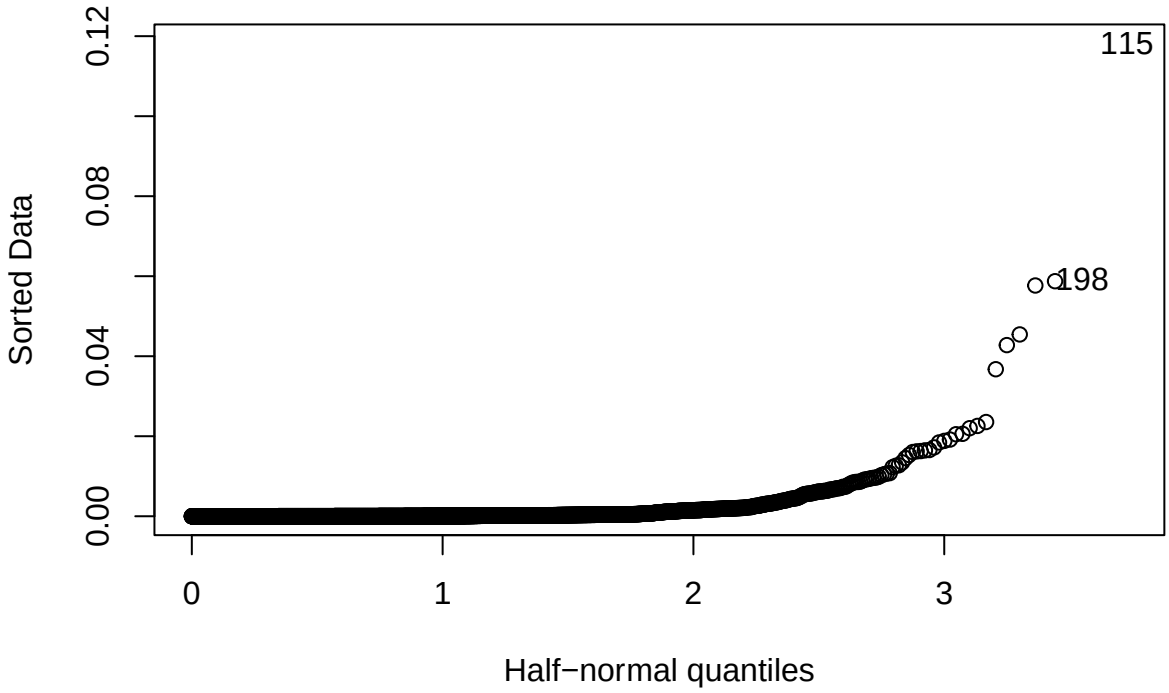
We then fit a simpler model - generalized linear model (poisson regression model).

```
lmod = glm(doctorco ~ sex + age + agesq + income + levyplus + freepoor +
             freerepa + illness + actdays + hscore + chcond1 + chcond2
           ,data = dvisits, family = "poisson")
summary(lmod)
```

```
##
## Call:
## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = "poisson", data = dvisits)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9170  -0.6862  -0.5743  -0.4839   5.7005
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.223848   0.189816 -11.716   <2e-16 ***
## sex          0.156882   0.056137   2.795   0.0052 **
## age          1.056299   1.000780   1.055   0.2912
## agesq       -0.848704   1.077784  -0.787   0.4310
## income      -0.205321   0.088379  -2.323   0.0202 *
## levyplus     0.123185   0.071640   1.720   0.0855 .
## freepoor    -0.440061   0.179811  -2.447   0.0144 *
## freerepa     0.079798   0.092060   0.867   0.3860
## illness      0.186948   0.018281  10.227   <2e-16 ***
## actdays      0.126846   0.005034  25.198   <2e-16 ***
## hscore       0.030081   0.010099   2.979   0.0029 **
## chcond1      0.114085   0.066640   1.712   0.0869 .
## chcond2      0.141158   0.083145   1.698   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4379.5  on 5177  degrees of freedom
## AIC: 6737.1
##
## Number of Fisher Scoring iterations: 6
```

```
halfnorm(cooks.distance(lmod))
```



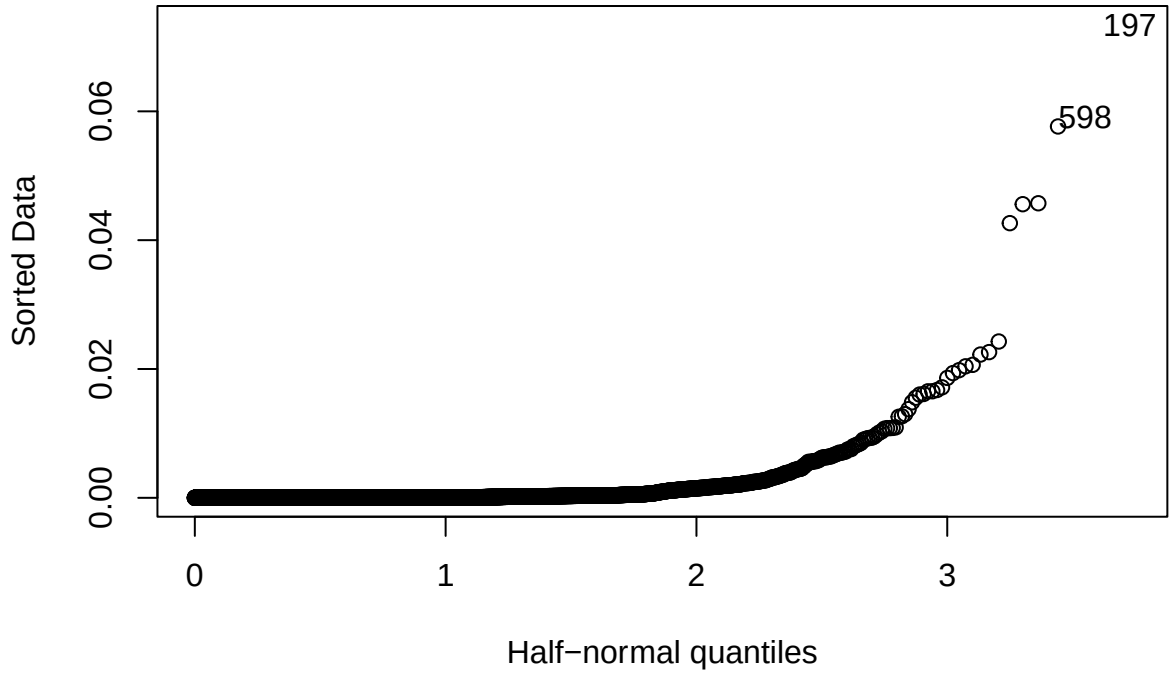
The influential plot suggests that case 115 is an outlier. We remove this observation for the following analyses.

```
dvisits2 <- dvisits[-115,]
lmod2 = glm(doctorco ~ sex + age + agesq + income + levyplus + freepoor +
            freerepa + illness + actdays + hscore + chcond1 + chcond2
          ,data = dvisits2, family = "poisson")
summary(lmod2)
```

```
##
## Call:
## glm(formula = doctorco ~ sex + age + agesq + income + levyplus +
##     freepoor + freerepa + illness + actdays + hscore + chcond1 +
##     chcond2, family = "poisson", data = dvisits2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9317  -0.6848  -0.5741  -0.4818   5.6916
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.243162   0.190305 -11.787  < 2e-16 ***
## sex          0.167834   0.056345   2.979  0.00289 **
## age          1.072515   1.002848   1.069  0.28486
## agesq       -0.837427   1.079697  -0.776  0.43798
## income      -0.185515   0.088435  -2.098  0.03593 *
## levyplus     0.119757   0.071663   1.671  0.09470 .
```

```
## freepoor    -0.638875   0.198630  -3.216  0.00130 **
## freerepa     0.078109   0.092119   0.848  0.39649
## illness      0.189802   0.018309  10.367  < 2e-16 ***
## actdays      0.124679   0.005067  24.605  < 2e-16 ***
## hscore       0.032067   0.010112   3.171  0.00152 **
## chcond1      0.096470   0.066814   1.444  0.14877
## chcond2      0.138767   0.083208   1.668  0.09537 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5604.2  on 5188  degrees of freedom
## Residual deviance: 4358.2  on 5176  degrees of freedom
## AIC: 6711.9
##
## Number of Fisher Scoring iterations: 6
```

```
halfnorm(cooks.distance(lmod2))
```



It seems that there is no outliers in the sub-setting data. Use the package `gam` to fit the generalized additive model, with poisson distribution.

```
library(gam)
```

```
## Warning: package 'gam' was built under R version 4.1.3
```

```
## Loading required package: splines

## Loading required package: foreach

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##     accumulate, when

## Loaded gam 1.20.1
```

```
amod <- gam(doctorco ~ sex + s(age) + s(agesq) + s(income) + levyplus + freepoor
            + freerepa + s(illness, 6) + s(actdays) + s(hscore) + chcond1
            + chcond2, data = dvisits2, scale = -1, family = "poisson")
summary(amod)
```

```
##
## Call: gam(formula = doctorco ~ sex + s(age) + s(agesq) + s(income) +
##     levyplus + freepoor + freerepa + s(illness, 6) + s(actdays) +
##     s(hscore) + chcond1 + chcond2, family = "poisson", data = dvisits2,
##     scale = -1)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7646 -0.6952 -0.5290 -0.3505  5.0050
##
## (Dispersion Parameter for poisson family taken to be 1)
##
##     Null Deviance: 5604.171 on 5188 degrees of freedom
## Residual Deviance: 4062.566 on 5157 degrees of freedom
## AIC: 6454.326
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                Df Sum Sq Mean Sq  F value    Pr(>F)
## sex             1   58.0   58.02  47.2492 6.993e-12 ***
## s(age)          1  208.0  208.05 169.4132 < 2.2e-16 ***
## s(agesq)        1    0.4    0.40   0.3227 0.5700375
## s(income)       1   15.6   15.65  12.7418 0.0003608 ***
## levyplus        1    0.0    0.00   0.0026 0.9590847
## freepoor        1   14.0   14.01  11.4072 0.0007370 ***
## freerepa        1    5.7    5.72   4.6541 0.0310262 *
## s(illness, 6)   1  230.5  230.49 187.6904 < 2.2e-16 ***
## s(actdays)      1  743.9  743.89 605.7457 < 2.2e-16 ***
## s(hscore)       1   14.9   14.85  12.0939 0.0005101 ***
## chcond1         1    0.1    0.09   0.0723 0.7880005
## chcond2         1    1.2    1.18   0.9615 0.3268630
## Residuals    5157 6333.1    1.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Anova for Nonparametric Effects
##                Npar Df Npar Chisq    P(Chi)
## (Intercept)
## sex
## s(age)              3      1.210    0.7507
## s(agesq)            3      1.082    0.7815
## s(income)           3      5.908    0.1162
## levyplus
## freepoor
## freerepa
## s(illness, 6)       4     71.222 1.255e-14 ***
## s(actdays)          3    200.565 < 2.2e-16 ***
## s(hscore)           3      3.745    0.2904
## chcond1
## chcond2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use stepwise method to choose the optimal model.

```
scope_list = list(
  "age" = ~1 + s(age) + s(age, 2) + s(age, 3),
  "agesq" = ~1 + s(agesq) + s(agesq, 2) + s(agesq, 3),
  "income" = ~1 + s(income) + s(income, 2) + s(income, 3),
  "illness" = ~1 + s(illness) + s(illness, 2) + s(illness, 3) + s(illness, 4),
  "actdays" = ~1 + s(actdays) + s(actdays, 2) + s(actdays, 3),
  "hscore" = ~1 + s(hscore) + s(hscore, 2) + s(hscore, 3),
  "sex" = ~1 + sex,
  "levyplus" = ~1 + levyplus,
  "freepoor" = ~1 + freepoor,
  "freerepa" = ~1 + freerepa,
  "chcond1" = ~1 + chcond1,
  "chcond2" = ~1 + chcond2)
step.Gam(amod, scope_list)
```

```
## Start:  doctorco ~ sex + s(age) + s(agesq) + s(income) + levyplus + freepoor +      freerepa + s(illn
## Step:1 doctorco ~ s(illness, 6) + s(age) + s(income) + s(actdays) +      s(hscore) + sex + levyplus +
## Step:2 doctorco ~ s(illness, 6) + s(age, 2) + s(income) + s(actdays) +      s(hscore) + sex + levyplu
## Step:3 doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) +      s(hscore) + sex + levy
## Step:4 doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) +      s(hscore) + sex + levy
## Step:5 doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) +      s(hscore, 2) + sex + l
## Step:6 doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) +      s(hscore, 2) + sex + l
## Step:7 doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) +      s(hscore, 2) + sex + i
## Step:8 doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) +      s(hscore, 2) + sex + i
```

```
## Call:
## gam(formula = doctorco ~ s(illness, 6) + s(age, 2) + s(income,
##     2) + s(actdays) + s(hscore, 2) + sex + freepoor, family = "poisson",
##     data = dvisits2, scale = -1, trace = FALSE)
##
## Degrees of Freedom: 5188 total; 5171 Residual
## Residual Deviance: 4073.089
```

```
amod2 <- gam(doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) +
               s(hscore, 2) + sex + freepoor, family = "poisson",
             data = dvisits2, scale = -1)
summary(amod2)
```

```
##
## Call: gam(formula = doctorco ~ s(illness, 6) + s(age, 2) + s(income,
##     2) + s(actdays) + s(hscore, 2) + sex + freepoor, family = "poisson",
##     data = dvisits2, scale = -1)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6602 -0.6944 -0.5379 -0.3496  5.0987
##
## (Dispersion Parameter for poisson family taken to be 1)
##
##     Null Deviance: 5604.171 on 5188 degrees of freedom
## Residual Deviance: 4073.089 on 5171 degrees of freedom
## AIC: 6436.85
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                 Df Sum Sq Mean Sq  F value    Pr(>F)
## s(illness, 6)    1  372.0  371.99 301.2873 < 2.2e-16 ***
## s(age, 2)        1  130.1  130.06 105.3424 < 2.2e-16 ***
## s(income, 2)     1    8.1    8.14   6.5962 0.0102477 *
## s(actdays)       1  742.2  742.19 601.1225 < 2.2e-16 ***
## s(hscore, 2)     1   14.2   14.21  11.5098 0.0006975 ***
## sex              1    8.2    8.22   6.6582 0.0098973 **
## freepoor         1   13.5   13.46  10.8979 0.0009693 ***
## Residuals     5171 6384.5    1.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                Npar Df Npar Chisq    P(Chi)
## (Intercept)
## s(illness, 6)       4     74.790  2.22e-15 ***
## s(age, 2)           1      0.316   0.57426
## s(income, 2)        1      3.974   0.04623 *
## s(actdays)          3    201.980 < 2.2e-16 ***
## s(hscore, 2)        1      0.937   0.33302
## sex
## freepoor
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The final model is `doctorco ~ s(illness, 6) + s(age, 2) + s(income, 2) + s(actdays) + s(hscore, 2) + sex + freepoor`, and all the variables are significant. We then fit the final model again to the package `mgcv` to plot the transformations on the predictors identified by the additive model.

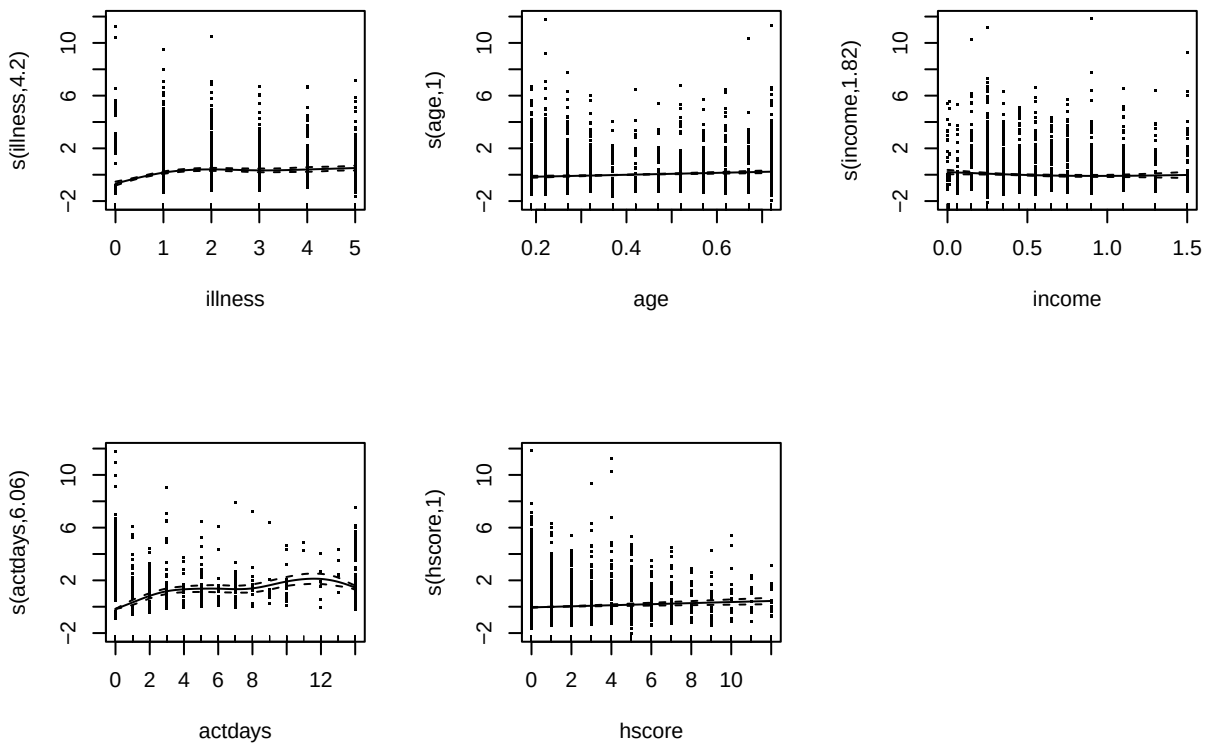
```r
amod3 <- mgcv::gam(doctorco ~ s(illness, k = 6) + s(age, k = 2) +
                     s(income, k = 2) + s(actdays) + s(hscore, k = 2) + sex +
                     freepoor, family = "poisson", data = dvisits2, scale = -1)
```

```
## Warning in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots): basis dimension, k, increased

## Warning in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots): basis dimension, k, increased

## Warning in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots): basis dimension, k, increased
```

```r
summary(amod3)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## doctorco ~ s(illness, k = 6) + s(age, k = 2) + s(income, k = 2) +
##     s(actdays) + s(hscore, k = 2) + sex + freepoor
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.67621    0.05415 -30.953  < 2e-16 ***
## sex          0.15611    0.06362   2.454  0.01417 *
## freepoor    -0.73013    0.22344  -3.268  0.00109 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df     F  p-value
## s(illness) 4.195  4.720 19.78  < 2e-16 ***
## s(age)     1.001  1.001 20.50 6.24e-06 ***
## s(income)  1.818  1.967  3.48 0.047074 *
## s(actdays) 6.065  7.108 85.34  < 2e-16 ***
## s(hscore)  1.001  1.002 12.74 0.000358 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.202   Deviance explained = 27.4%
## GCV = 0.78888  Scale est. = 1.305     n = 5189
```

```r
plot(amod3, residuals = TRUE, page = 1)
```



We can see that the confidence bands do not highly overlap with `s() = 0` for those variables, thus, these variables are significant.

(b) Check the diagnostics.

Solution:

```
library(mgcv)
```

```
## Warning: package 'mgcv' was built under R version 4.1.2
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

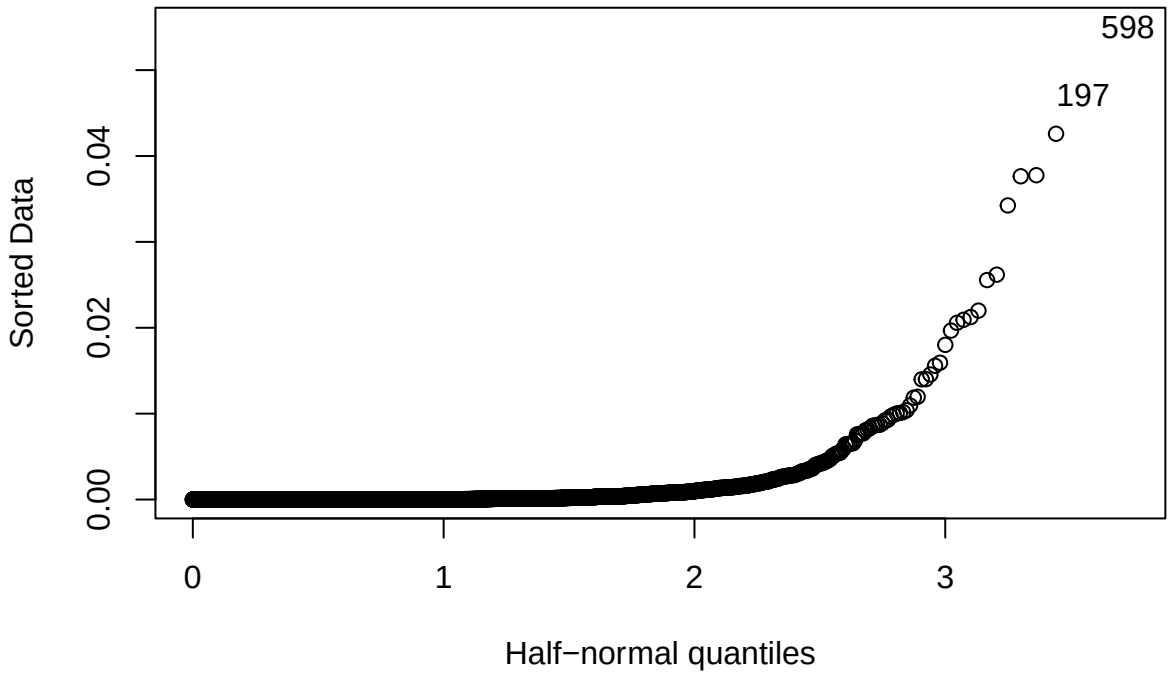
```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-38. For overview type 'help("mgcv-package")'.
```

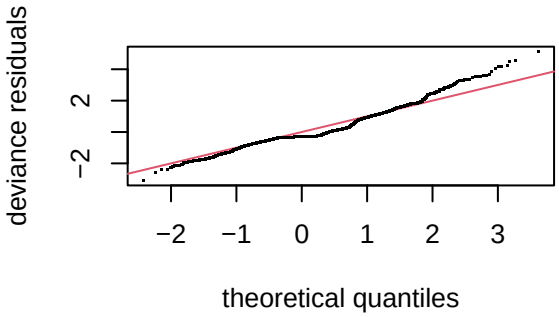
```
##
## Attaching package: 'mgcv'
```

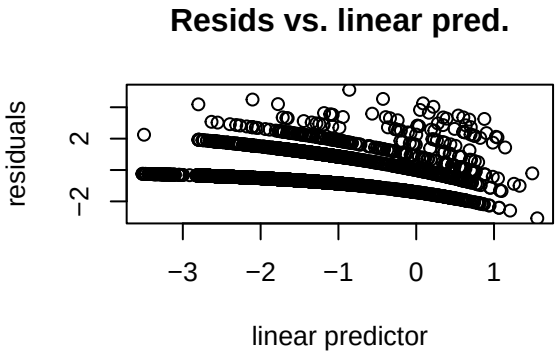
```
## The following objects are masked from 'package:gam':
##
##     gam, gam.control, gam.fit, s
```

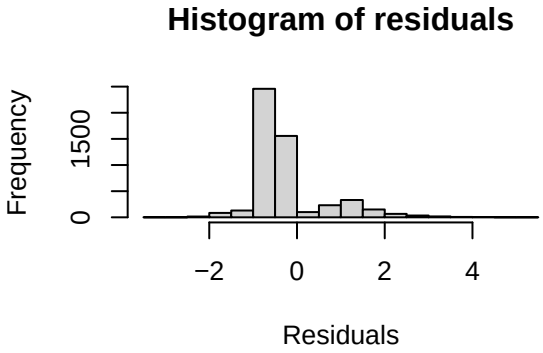
```
halfnorm(cooks.distance(amod3))
```

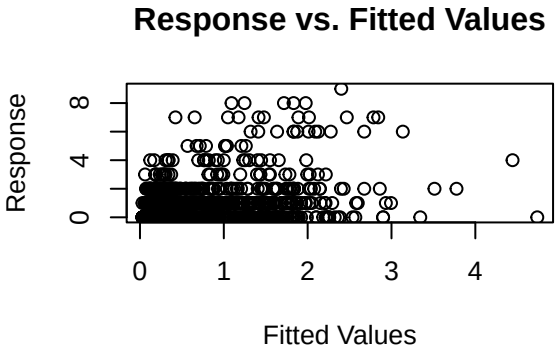



```
gam.check(amod3)
```










```
##
## Method: GCV   Optimizer: outer newton
## full convergence after 7 iterations.
## Gradient range [-2.397902e-07,1.576588e-06]
## (score 0.788883 & scale 1.305021).
## Hessian positive definite, eigenvalue range [1.422204e-07,0.0001909849].
## Model rank =  23 / 23
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##              k'  edf k-index p-value
## s(illness) 5.00 4.20    0.91   0.430
## s(age)     2.00 1.00    0.91   0.510
## s(income)  2.00 1.82    0.90   0.355
## s(actdays) 9.00 6.06    0.91   0.465
## s(hscore)  2.00 1.00    0.89   0.095 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output, the small p-value indicates the residuals are not randomly distributed. Since we have all the p-values greater than 0.05, we can say the residuals for each variables are randomly distributed.

From the diagnostics plots above, the Q-Q plot do not fit well, which means the data may not meet the normality. For the `histogram of residuals`, the distribution should be bell-shaped and centered at 0, and our model seems to not be the best. We also expect to see the dots are evenly distributed and centered

at 0 in the `resid vs. linear pred.` plot, and the dots should be lined up in the `response vs. fitted values` plot; therefore, the model does not fit really well.

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

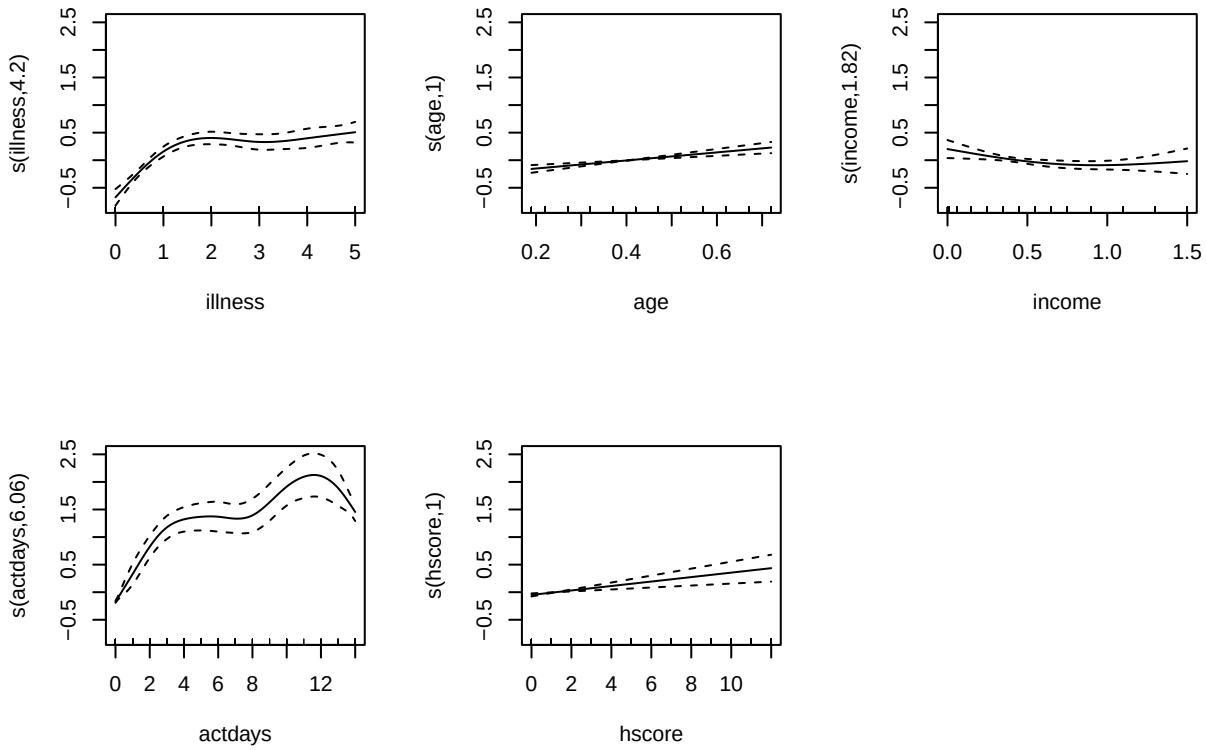
Solution:

```
summary(amod3)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## doctorco ~ s(illness, k = 6) + s(age, k = 2) + s(income, k = 2) +
##     s(actdays) + s(hscore, k = 2) + sex + freepoor
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.67621    0.05415 -30.953  < 2e-16 ***
## sex          0.15611    0.06362   2.454  0.01417 *
## freepoor    -0.73013    0.22344  -3.268  0.00109 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F  p-value
## s(illness) 4.195  4.720 19.78  < 2e-16 ***
## s(age)     1.001  1.001 20.50 6.24e-06 ***
## s(income)  1.818  1.967  3.48 0.047074 *
## s(actdays) 6.065  7.108 85.34  < 2e-16 ***
## s(hscore)  1.001  1.002 12.74 0.000358 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.202   Deviance explained = 27.4%
## GCV = 0.78888  Scale est. = 1.305     n = 5189
```

The estimated coefficients for `sex` is 0.156, which means female would be predicted to visit the doctor more than male. The estimated coefficients for `freepoor` is -0.73, which means people who covered by government because low income, recent immigrant, unemployed would be predicted to visit the doctor more.

```
plot(amod3, page = 1)
```



From the plots shown above, people whose `illness` are 2 to 5, `age` are 72, `income` are less than 200, `actdays` are 11.5, and `hscore` are 12 have a highest probability to visit the doctor the most.

Combine the two results, and we get people who is female and covered by government because low income, recent immigrant, unemployed, and whose `illness` are 2 to 5, `age` are 72, `income` are less than 200, `actdays` are 11.5, and `hscore` are 12 would be predicted to visit the doctor the most.

(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:

```
predict(amod3, dvisits2[5189,], type = "response")
```

```
##      5190
## 0.1037571
```

The parameter of poisson distribution for the last person in the dataset is 0.1037571.

```
dpois(0, 0.1037571)
```

```
## [1] 0.9014442
```

```
dpois(1, 0.1037571)
```

```
## [1] 0.09353124
```

```
dpois(2, 0.1037571)
```

```
## [1] 0.004852265
```

```
dpois(3, 0.1037571)
```

```
## [1] 0.000167819
```

```
dpois(4, 0.1037571)
```

```
## [1] 4.353103e-06
```

```
dpois(5, 0.1037571)
```

```
## [1] 9.033306e-08
```

```
dpois(6, 0.1037571)
```

```
## [1] 1.562116e-09
```

```
dpois(7, 0.1037571)
```

```
## [1] 2.315438e-11
```

```
dpois(8, 0.1037571)
```

```
## [1] 3.003039e-13
```

```
dpois(9, 0.1037571)
```

```
## [1] 3.462073e-15
```

The probability they visit 0 time is 0.9014442; the probability they visit 1 time is 0.09353124; the probability they visit 2 times is 0.004852265; the probability they visit 3 times is 0.000167819; the probabilities they visit 4,5,6,7,8, and 9 times are almost 0.