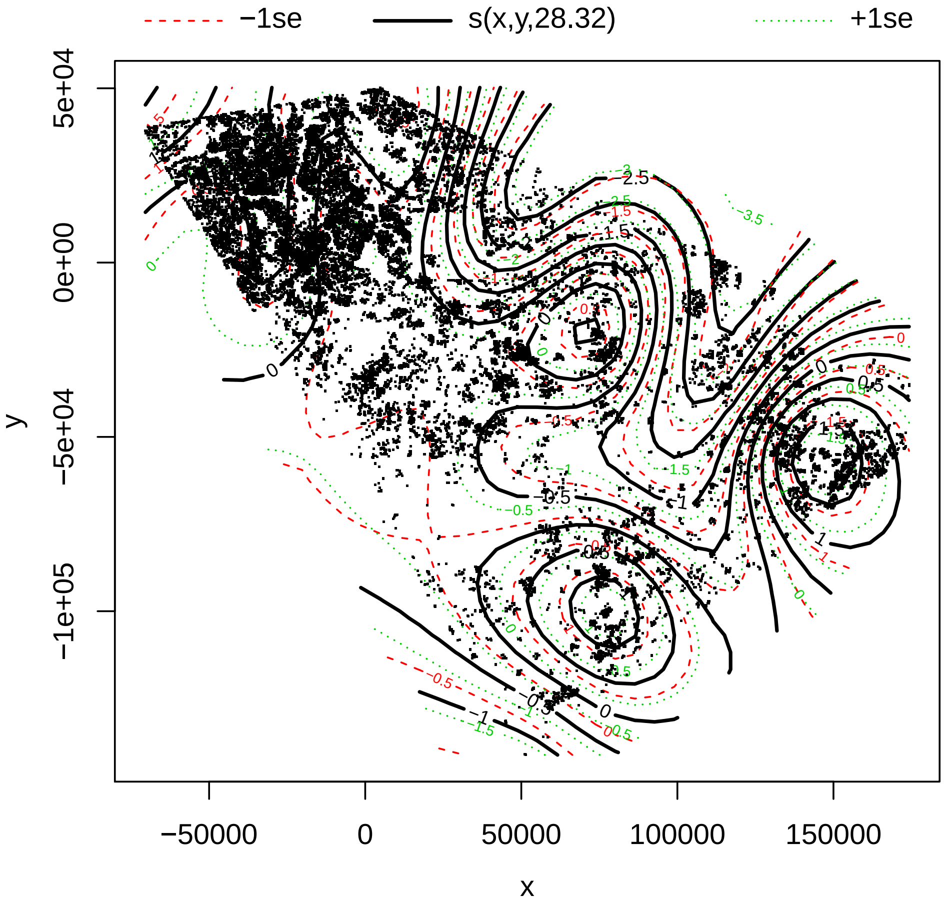
Homework 9 - STAT 511
Amal Agarwal

Answer 1

(a) Model: infected.mndnr$_i \sim Bern(p_i)$, with $g(E(Y_i)) = \eta_i + f(x_i, y_i) = X_i\beta + f(x_i, y_i)$ (g being the canonical logit link in bernoulli model)
$log(\dfrac{p_i}{1-p_i}) = \beta_0 + \beta_1 mortal_i + \beta_2 phys_i + \beta_3 si_i + \beta_4 usize_i + \beta_5 height_i + \beta_6 dbh_i + f(x_i, y_i)$
where $f(x, y)$ is the smooth function of the (x, y) location of the stand. In R, this can be fitted using gam as follows:

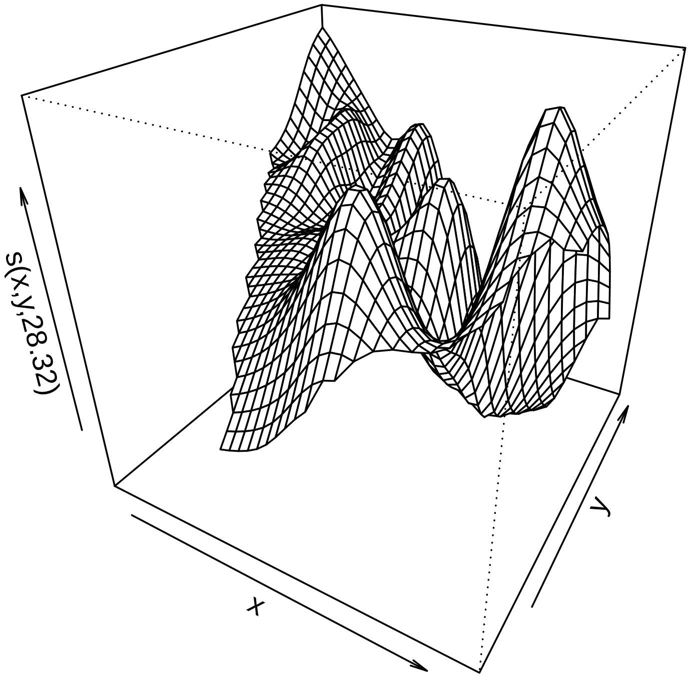
```
> library(mgcv)
> mist=read.csv("mistletoe.csv",sep=",")
> fit.gm=gam(infected.mndnr~mortal+phys+si+usize+height+dbh+s(x,y),data=mist,family
> summary(fit.gm)

Family: binomial
Link function: logit

Formula:
infected.mndnr ~ mortal + phys + si + usize + height + dbh +
    s(x, y)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.245676   0.229233 -14.159  < 2e-16 ***
mortal       1.129592   0.032270  35.004  < 2e-16 ***
phys         0.239472   0.043452   5.511 3.56e-08 ***
si          -0.046674   0.002842 -16.423  < 2e-16 ***
usize        0.326035   0.027580  11.821  < 2e-16 ***
height       0.072098   0.003142  22.946  < 2e-16 ***
dbh         -0.360025   0.017732 -20.304  < 2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(x,y) 28.32  28.96  740.4  <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

R-sq.(adj) =  0.216   Deviance explained = 25.7%
UBRE = -0.4732  Scale est. = 1         n = 25431
```

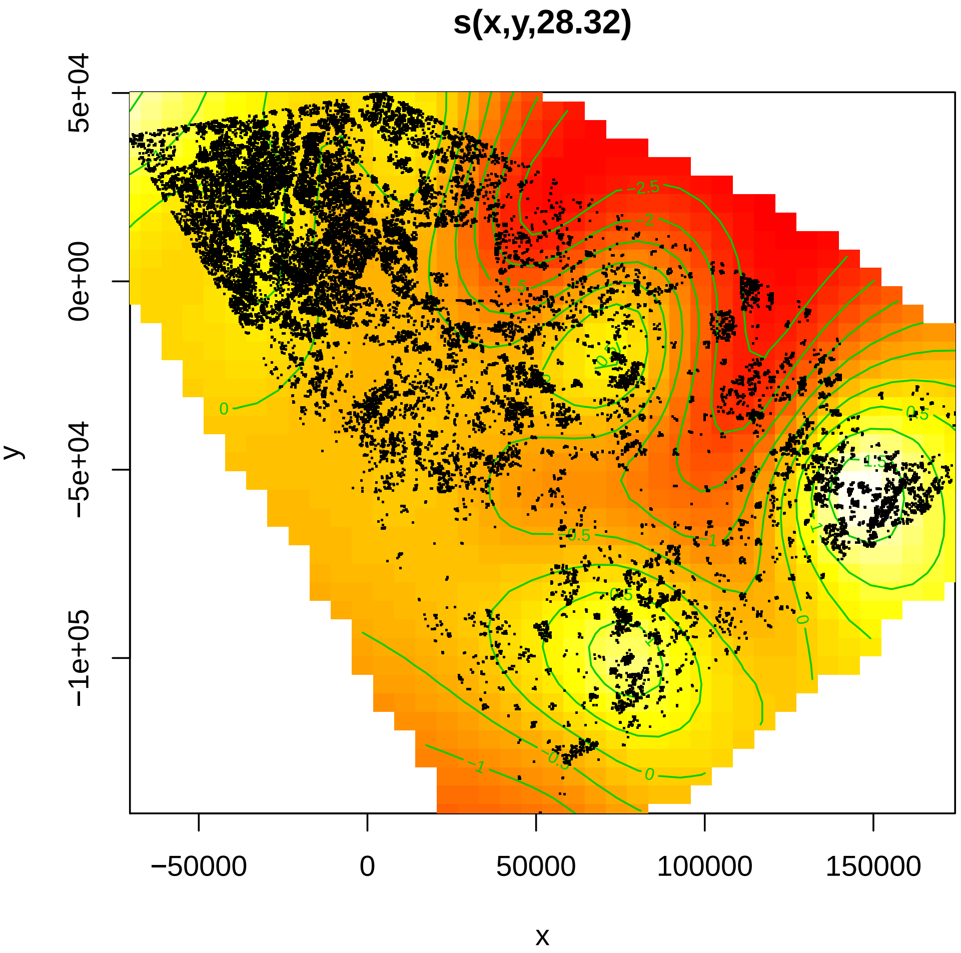
2-D contour plot of the smooth function $f(x, y)$ is given as:



Clearly there are 4 regions indicating high values of the predicted log(odds) given that other predictors (except the spatial coordinates x and y) are held at zero. Since $log(odds_i) \propto p_i$, these regions have elevated risk of mistletoe presence.

This can be more easily interpreted with the following 3D plot showing 4 peaks and the colour plot (scheme 2) showing 4 lighter regions corresponding to high values.


s(x,y,28.32)
x
y



This information might be very useful to the forest managers because if they know what forest stands are more prone to mistletoe presence they can take preventive measures to reduce the growth of mistletoe in these areas. Mistletoe is a parasitic plant and kills trees and so these preventive measures might help in conservation of natural habitats.

Answer 2

(a)

- Fitting Model 1: admit$_i$ $\sim$ $Bern(p_i)$, with $l(E(Y_i)) = \eta_i + f(gre_i) + g(gpa_i) = X_i\beta + f(gre_i) + g(gpa_i)$ (l being the canonical logit link in bernoulli model)
  $log(\dfrac{p_i}{1-p_i}) = \beta_0 + \beta_1 rank_i + f(gre_i) + g(gpa_i)$

```
> library(mgcv)
> admissions <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
> fit1=gam(admit~rank+s(gre)+s(gpa),data=admissions,family=binomial)
> summary(fit1)

Family: binomial
Link function: logit

Formula:
admit ~ rank + s(gre) + s(gpa)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5386     0.3164   1.702   0.0887 .
rank         -0.5631     0.1276  -4.413 1.02e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Approximate significance of smooth terms:
        edf Ref.df Chi.sq p-value
s(gre) 1.00  1.000  4.318  0.0377 *
s(gpa) 2.97  3.713  8.273  0.0686 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

R-sq.(adj) =  0.096   Deviance explained = 8.82%
UBRE = 0.16949  Scale est. = 1         n = 400
```

- Fitting Model 2: admit$_i \sim Bern(p_i)$, with $l(E(Y_i)) = \eta_i + g(gpa_i) = X_i\beta + g(gpa_i)$ (l being the canonical logit link in bernoulli model)
  $log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 rank_i + \beta_2 gre_i + g(gpa_i)$

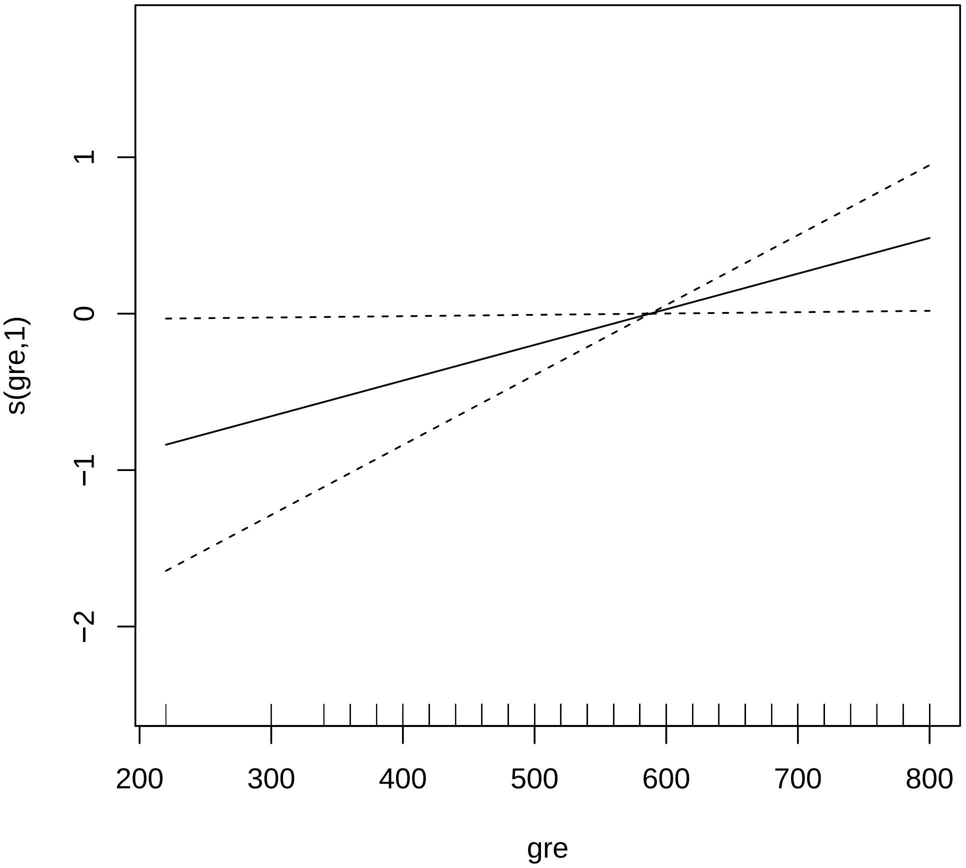
```
> library(mgcv)
> admissions <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
> fit2=gam(admit~rank+gre+s(gpa),data=admissions,family=binomial)
> summary(fit2)

Family: binomial
Link function: logit

Formula:
admit ~ rank + gre + s(gpa)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.800502   0.743319  -1.077   0.2815
rank        -0.563112   0.127610  -4.413 1.02e-05 ***
```

```
gre          0.002279   0.001097   2.078   0.0377 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(gpa) 2.971  3.714  8.275  0.0686 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

R-sq.(adj) =  0.096   Deviance explained = 8.83%
UBRE = 0.16949  Scale est. = 1         n = 400
```

- AIC values of the above models can be computed as:

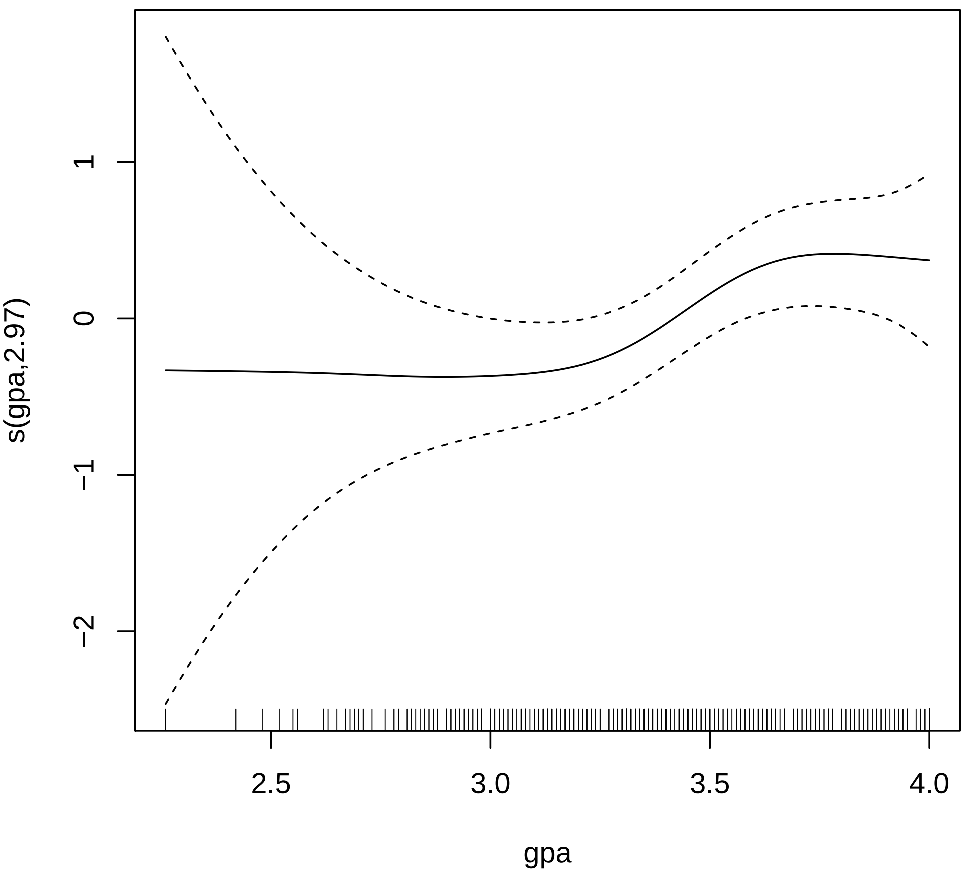
  ```
  > AIC(fit1)
  [1] 467.7949
  > AIC(fit2)
  [1] 467.7948
  ```

  Clearly the difference is of the order of $10^{-4} << 2$ which is insignificant and does not help in making any model selection decision. Plotting the smooth function $f(gre)$ in Model 1 as follows, we can notice that it looks very much linear.



This linear dependence of the f(gre) with respect to gre nullifies the effect of smoothing. Due to the absence of any non-linear dependence of f(gre) on gre, the two models are equivalent and this also accounts for the little difference in AIC values.

- Now the $log(odds)$ depends on gre linearly in Model 1 as can be seen from the previous graph. Thus the mean of admiitance is directly proportional to the gre score. Model 2 gives the same result. Futher the smooth function of gpa can be plotted as:



This clearly shows that the log(odds) of admittance is constant below 3.3 gpa, varies approximately linearly with gpa for $3.3 \leq gpa \leq 3.6$ and again becomes roughly constant for $gpa \geq 3.6$. This implies that there is greater probability of admittance for $gpa \geq 3.5$ than for $gpa \leq 3.5$. Again this holds for both the models.