Homework 3

Allan Kimaina

Question 1: Do Problem 3 in chapter 5 of Gelman and Hill.

You are interested in how well the combined earnings of the parents in a child's family predicts high school graduation. You are told that the probability a child graduates from high school is 27% for children whose parents earn no income and is 88% for children whose parents earn $60,000. Determine the logistic regression model that is consistent with this information. (For simplicity you may want to assume that income is measured in units of $10,000).

$$logit(0.88) = logit(0.27) + 6 * x$$
$$1.99243 = -0.9946226 + 6 * x$$
$$x = \frac{1.99243 + 0.9946226}{6} = 0.4978421$$

Substituiting X we get:

$$P(y = 1) = logit^{-1}(-0.9946226 + 0.4978421 * x)$$

We can simulate x1 using our Regression Model:

Table 1:

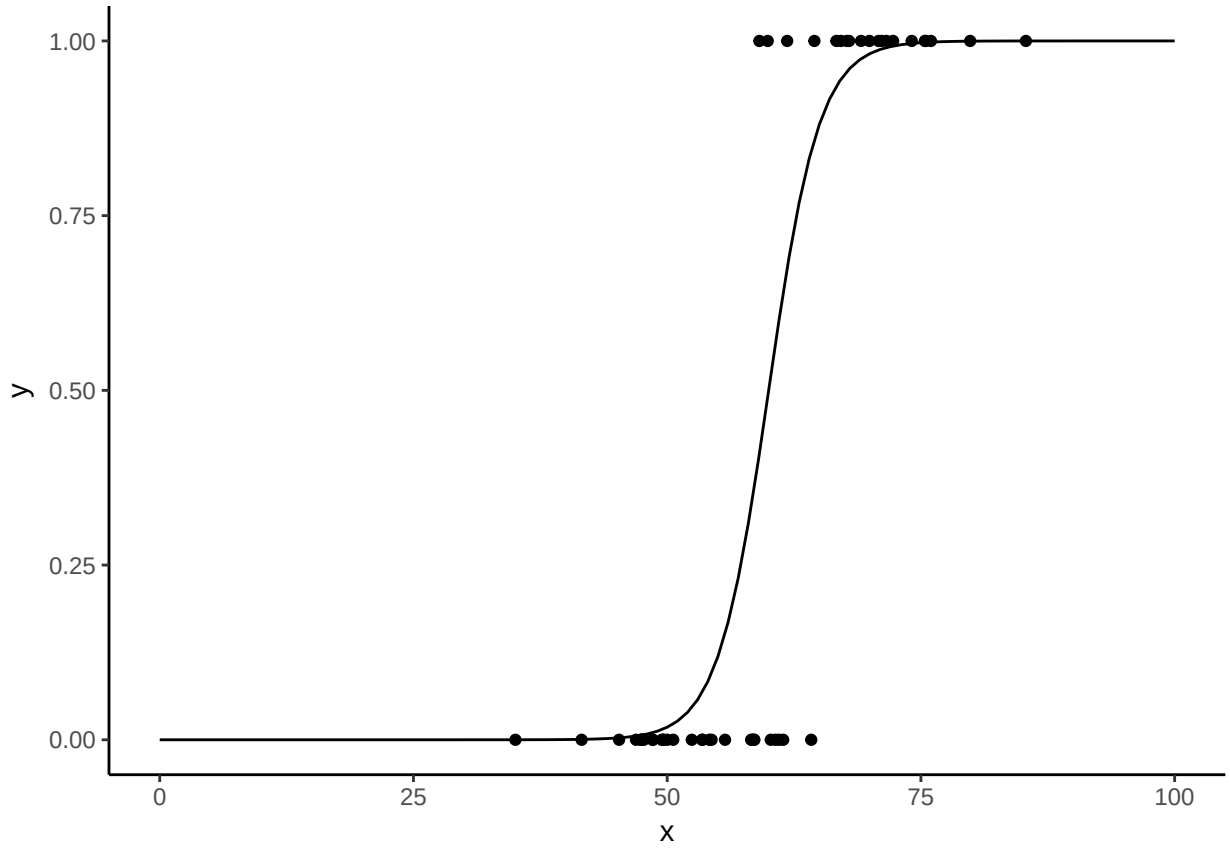
	Dependent variable:
	y1
x1	0.496^{***} (0.002)
Constant	-0.990^{***} (0.002)
Observations	1,000,000
Log Likelihood	−570,739.400
Akaike Inf. Crit.	1,141,483.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 2: Do Problem 5 in chapter 5 of Gelman and Hill.

In a class of 50 students, a logistic regression is performed of course grade (pass or fail) on midterm exam score (continuous values with mean 60 and standard deviation 15). The fitted model is $Pr(pass) = logit^{-1}(-24 + 0.4x)$.

Part A

Graph the fitted model. Also on this graph put a scatter plot of hypothetical data consistent with the information given.



Part B

Suppose the midterm scores were transformed to have a mean of 0 and standard deviation of 1. What would be the equation of the logistic regression using these transformed scores as a predictor?

$$P(pass) = logit^{-1}(\beta_0 + \beta_1 x) => original$$

$$P(pass) = logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 \frac{x - \bar{x}}{S_x}) => standardized(x_i)$$

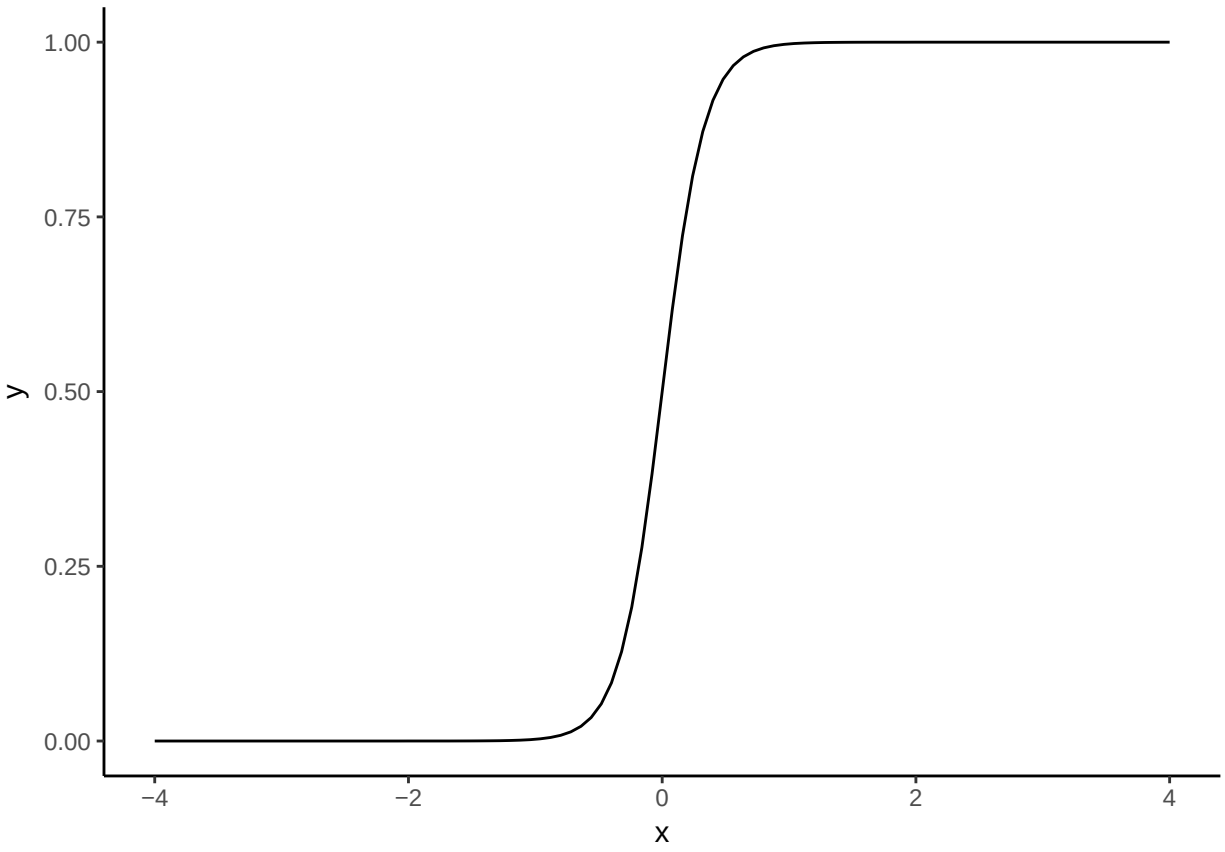
From above we find that:

$$\hat{\beta}_1 = \beta_1 * S_x = .4 * 15 = 6$$

$$\hat{\beta}_0 = \beta_0 + \frac{6 * 60}{15} = -24 + 24 = 0$$

Substituting 6 and 0 we get:

$$P(pass) = logit^{-1}(0 + 6x)$$



Part C

Create a new predictor that is pure noise (for example, in R you can create `newpred <- rnorm(n,0,1)`). Add it to your model. How much does the deviance decrease?

	Model 1	Model 2
(Intercept)	0.57 (0.51)	0.56 (0.53)
x1	4.42 (1.39)**	4.43 (1.41)**
randomNoise		0.04 (0.50)
AIC	29.06	31.06
BIC	32.89	36.79
Log Likelihood	-12.53	-12.53
Deviance	25.06	25.06
Num. obs.	50	50

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Statistical models

If we add a predictor that is pure noise deviance barely decreases. It only decreases by 0.006 i.e from 25.064 to 25.058

Question 3: Do Problem 8 in chapter 5 of Gelman and Hill.

Building a logistic regression model: the folder `rodents` contains data on rodents in a sample of New York City apartments.

Part A

Build a logistic regression model to predict the presence of rodents (the variable `rodent2` in the dataset) given indicators for the ethnic groups (race). Combine categories as appropriate. Discuss the estimated coefficients in the model.

	Model 1
(Intercept)	-1.70 (0.17)***
raceBlack	1.31 (0.23)***
racePuerto Rican	1.15 (0.27)***
raceOther Hispanic	1.46 (0.24)***
raceAsian/Pacific Islander	0.55 (0.39)
raceAmer-Indian/Native Alaskan	2.11 (0.93)*
raceTwo or more races	0.79 (0.85)
AIC	877.00
BIC	909.28
Log Likelihood	-431.50
Deviance	863.00
Num. obs.	744

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Statistical models

We find that:

- `White Race (Intercept)`: The odds of having rodents infestation in an apartment primarily occupied by White people is $\exp(-1.70) = 0.18$. In other words, an apartment occupied by white people has a $logit^{-1}(-1.70) = 0.1539 = 15.39\%$ probability of having rodents infestation
- `Black Race`: The odds of having rodents infestation in an apartment primarily occupied by black people is $\exp(1.312) = 3.71$ times the odds of having rodents infestation in an apartment occupied by white people. In other words, apartment occupied by black people was 3.71 times more likely to be infested by rodents compared to apartment occupied by white.
- `Puerto Rican Race`: The odds of having rodents infestation in an apartment primarily occupied by Puerto Rican people is $\exp(1.312) = 3.1428571$ times the odds of having rodents infestation in an apartment occupied by white people.
- `Other Hispanic Race`: The odds of having rodents infestation in an apartment primarily occupied by Puerto Rican people is $\exp(1.4624) = 4.3164557$ times the odds of having rodents infestation in an apartment occupied by white people.
- `Amer-Indian/Native Alaska`: The odds of having rodents infestation in an apartment primarily occupied by Amer-Indian/Native Alaska people is $\exp(2.1102) = 8.25$ times the odds of having rodents infestation in an apartment occupied by white people.

The other races like Asian/Pacific Islander were insignificant implying that the odds of rodents infestation were not different from the comparison group (white race).

In summary, Hispanic and black presence in an apartment are positively associated with rodents infestation.

Part B

Add to your model some other potentially relevant predictors describing the apartment, building, and community district. Build your model using the general principles explained in Section 4.6. Discuss the coefficients for the ethnicity indicators in your model.

We performed stepwise AIC regression selection before assessing these predictors using guidelines highlighted in section 4.6. After ensuring that all sign of predictors and significance of predictors were OK, we went ahead and assessed effect modifiers (interactions). Interactions between race and Hispanic, and race and black presence were statistically insignificant. Putting all these factors into consideration We ended up creating a model shown below:

	Model 1
(Intercept)	-9.77 (2.32)***
personrm	0.82 (0.24)***
unitflr22	1.29 (0.91)
unitflr23	0.99 (0.90)
unitflr24	1.57 (0.92)
unitflr25	1.57 (0.93)
unitflr26	1.22 (0.94)
unitflr27	1.24 (0.94)
unitflr28	-1.63 (1.42)
unitflr29	2.29 (1.49)
regext1	-0.50 (0.21)*
povertyx21	0.37 (0.21)
extflr5_21	0.92 (0.48)
intcrack21	0.94 (0.27)***
inthole21	0.58 (0.34)
intleak21	0.50 (0.23)*
struct	-0.98 (0.21)***
help1	-0.51 (0.23)*
black_Mean	2.35 (0.65)***
board2_Mean	-2.63 (1.29)*
help_Mean	5.38 (2.06)**
hispanic_Mean	2.74 (1.02)**
old_Mean	1.54 (0.71)*
poverty_Mean	3.53 (2.20)
regext_Mean	1.70 (1.03)
extwin4_2_Mean	10.78 (7.31)
intcrack2_Mean	-20.16 (5.75)***
inthole2_Mean	26.21 (7.90)***
vacrate	8.99 (4.91)
AIC	741.70
BIC	875.45
Log Likelihood	-341.85
Deviance	683.70
Num. obs.	744

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Statistical models

Make sure to carefully explain what your final model means.

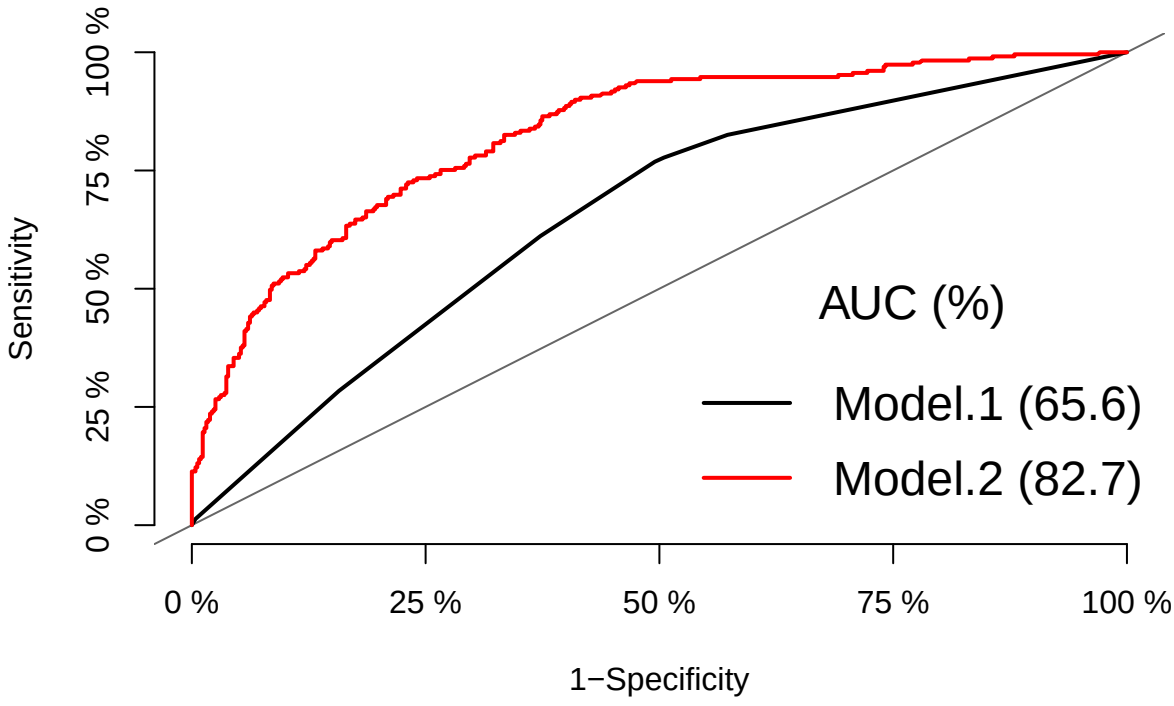
In general:

- `black_Mean`: Holding the other predictors constant at the base level or 0, a 1% increase in black people in the district changes the odds of rodents infestation by 9.41. Intuitively, holding the other predictors constant at the mean level, a 1% increase in black occupants in the district increases the estimated probability of rodents infestation by (0.904019/4)*100 = 22.6%
- `hispanic_Mean`: Holding the other predictors constant, a 1% increase in Hispanic people in the district changes the odds of rodents infestation in an apartment by 11.85. In other words, holding the other predictors constant at the mean level, a 1% increase in Hispanic occupants in the district increases the estimated probability of rodents infestation by (2.4727/4)*100 = 23.1%
- `race`: Given the fact that the householder race was statistically insignificant, we did not include it in our model. Also the effect of this predictor on the response was not substantial. Contextually, householder race doesn't really matter, the proportion of race around you is much more relevant.

In summary, black and Hispanic population in that district were associated with higher chances of rodents infestation with Hispanic population leading.

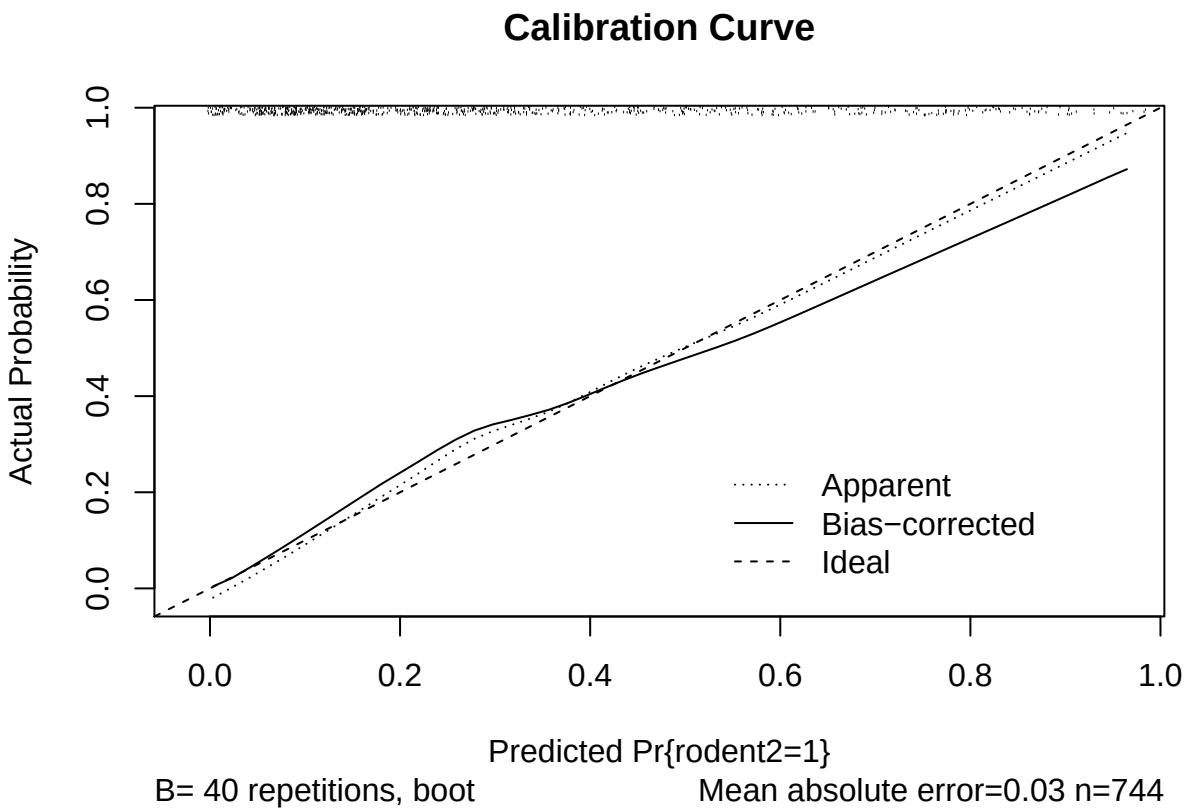
Check its fit using diagnostics. Make an ROC plot and a calibration curve.

ROC



The curve rises steeply indicating that the % of true positives accurately predicted by this logit model rose faster than the false positive.In fact we have a good acceptable AUC of about 82.7%.

Calibration Curve



n=744 Mean absolute error=0.03 Mean squared error=0.00142 0.9 Quantile of absolute error=0.061

In general, the calibration plot did not provide substantial evidence that our models was overfitted. However, the calibration plot depicted that the model had little evidence of underestimation of high probabilities.

Confusion Matrix

From the confusion matrix below we have a good truth detection rate i.e few false positive and few false negative

```
##                 actual_values
## predicted_values   0   1
##                0 466 111
##                1  49 118
```

Multicollinearity Diagnosis

Most of the predictors in the mode had Variable Inflation Factor of less than 2. intcrack and inhole indicated elevated level of VIF implying presence of collinearity. This was not alarming because we are building a predictive model.

```
  personrm      unitflr22      unitflr23      unitflr24      unitflr25
  1.075627      14.095042      16.097583      14.134223      11.031508
 unitflr26      unitflr27      unitflr28      unitflr29        regext1
  9.151704      11.042414       1.743298       1.722500       1.170401
```

```
povertyx21     extflr5_21     intcrack21      inthole21      intleak21
  1.170990       1.114750       1.393961       1.269603       1.218002
    struct          help1     black_Mean    board2_Mean      help_Mean
  1.214216       1.134305       3.261601       4.723903       5.270872
```

hispanic_Mean old_Mean poverty_Mean regext_Mean extwin4_2_Mean 4.879821 1.486010 6.283520 3.350785 4.226038 intcrack2_Mean inthole2_Mean vacrate 25.007303 17.375821 2.583700

Question 4: The full dehydration outcome for the Dhaka study that we looked at in class is a three-level variable (none, some or severe). Build two proportional odds regression models for this three-level variable. In the first use the clinical signs only. In the second, add in the additional predictors. Keep variables that you feel are important to prediction and to interpretation. Interpret your results clearly.

Model with all Clinical Variables:

We treated the following variables as clinical predictor variables. - genapp - tears - skin - resp - thirst - eyes - heart - mucous - pulse - urine

We then went ahead and threw all these variables into the model.

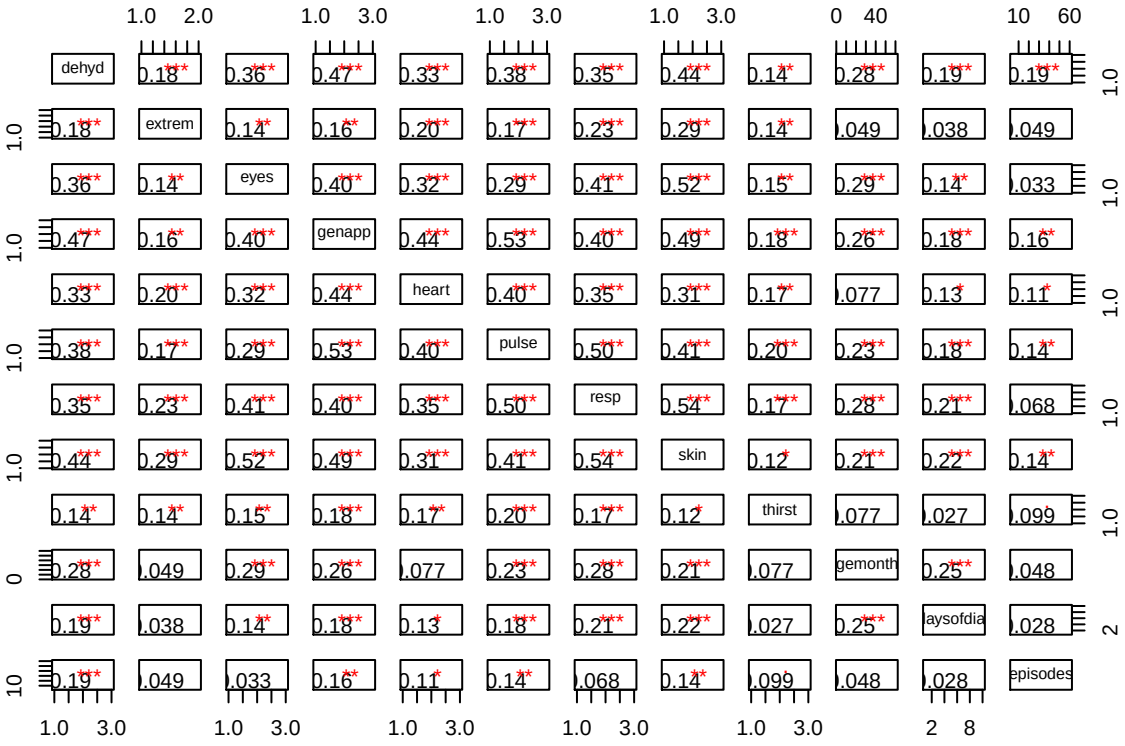
```
Call:
vglm(formula = ordered(dehyd) ~ genapp + tears + skin + resp +
    thirst + eyes + heart + mucous + muac + pulse + urine, family = cumulative(parallel = TRUE),
    data = dhaka)


Pearson residuals:
                  Min      1Q  Median     3Q   Max
logit(P[Y<=1]) -2.018 -0.5798 -0.2227 0.6829 4.561
logit(P[Y<=2]) -9.291  0.1024  0.1855 0.3396 1.533

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept):1  1.301e-01  1.508e+00   0.086  0.93124
(Intercept):2  3.283e+00  1.521e+00   2.159  0.03086 *
genapp1       -5.018e-01  2.933e-01  -1.711  0.08711 .
genapp2       -1.089e+00  3.426e-01  -3.178  0.00148 **
tears1        -2.431e-01  2.620e-01  -0.928  0.35351
tears2        -9.323e-01  4.587e-01  -2.033  0.04207 *
skin1         -8.580e-01  2.733e-01  -3.139  0.00170 **
skin2         -7.688e-01  5.765e-01  -1.333  0.18238
resp1         -1.735e-02  3.014e-01  -0.058  0.95410
resp2         -6.958e-01  9.298e-01  -0.748  0.45425
thirst1        5.063e-01  9.202e-01   0.550  0.58221
thirst2        2.603e-01  9.917e-01   0.262  0.79296
eyes1         -2.302e-01  3.530e-01  -0.652  0.51423
eyes2         -1.042e+00  5.134e-01  -2.030  0.04239 *
heart1        -3.550e-01  2.644e-01  -1.343  0.17942
heart2        -1.634e+01  9.088e+02      NA       NA
mucous1        1.473e-02  2.732e-01   0.054  0.95699
mucous2        9.942e-01  1.285e+03   0.001  0.99938
muac           4.473e-03  8.132e-03   0.550  0.58228
pulse1        -6.043e-01  3.073e-01  -1.966  0.04924 *
pulse2        -4.337e-01  4.685e-01  -0.926  0.35453
urine1        -3.212e-01  2.570e-01  -1.250  0.21134
urine2        -3.258e-02  3.733e-01  -0.087  0.93044
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Residual deviance: 600.5693 on 737 degrees of freedom

Log-likelihood: -300.2846 on 737 degrees of freedom

Number of iterations: 14
```

From this model, we observed that mucous, muac, thirst and urine were statistically insignificant therefore we removed them from our final model.

Best Model with both clinical and non clinical variables:

We then went ahead and created a correlation matrix in order to assess correlation in potential predictors.



- 3 non clinical predictors had significant correlation with the response, therefore we included it in our final model. The correlation matrix also established that most of the clinical predictors included in the previous model had significant and strong correlation.
- We also checked for interaction between several covariates but none were significant. Potential covariates we assessed for interactions included: tears and eyes, thirst and urine, heart and pulse, daysOfDia and skin

- After checking for interactions we came up with our final model shown below:

Interpretation

- In general only 4 predictors were statistically significant: genapp, skin, eyes, episodes
- `genapp`: General appearance has 3 level: Normal (0), Restless/Irritable (1), Lethargic/Unconscious (2). Normal is the reference in this model.
  - `genapp1` Holding the other predictor constant, the odds of having no dehydration in children with "Restless/Irritable" general appearance is 0.5399427 times the odds of having no dehydration in children with "normal" general appearance.
  - `genapp2` Holding the other predictor constant, the odds of having no dehydration in children with "Lethargic/Unconscious" general appearance is 0.2624490 times the odds of having no dehydration in children with "normal" general appearance.
- `skin`: Skin has 3 level: Normal Skin Pinch (0), Slow Skin Pinch (1), Very Slow Skin Pinch (2). Normal is the reference in this model.
  - `skin1` Holding the other predictor constant, the odds of having no dehydration in children with "slow" skin pinch is 0.4272413 times the odds of having no dehydration in children with "normal" skin pinch.
  - `skin2` Holding the other predictor constant, the odds of having no dehydration in children with "very slow" skin pinch is 0.4339146 times the odds of having no dehydration in children with "normal" skin pinch.
- `eyes`: Eyes has 3 level: Normal (0), Sunken (1), Very Sunken (2). Normal is the reference in this model.
  - `eyes1` "Sunken" eyes was statistically insignificant.
  - `eyes2` Holding the other predictor constant, the odds of having no dehidration in children with "Very Sunken" eyes is 0.3023435 times the odds of having no dehydration in children with "normal" eyes.

Source Code

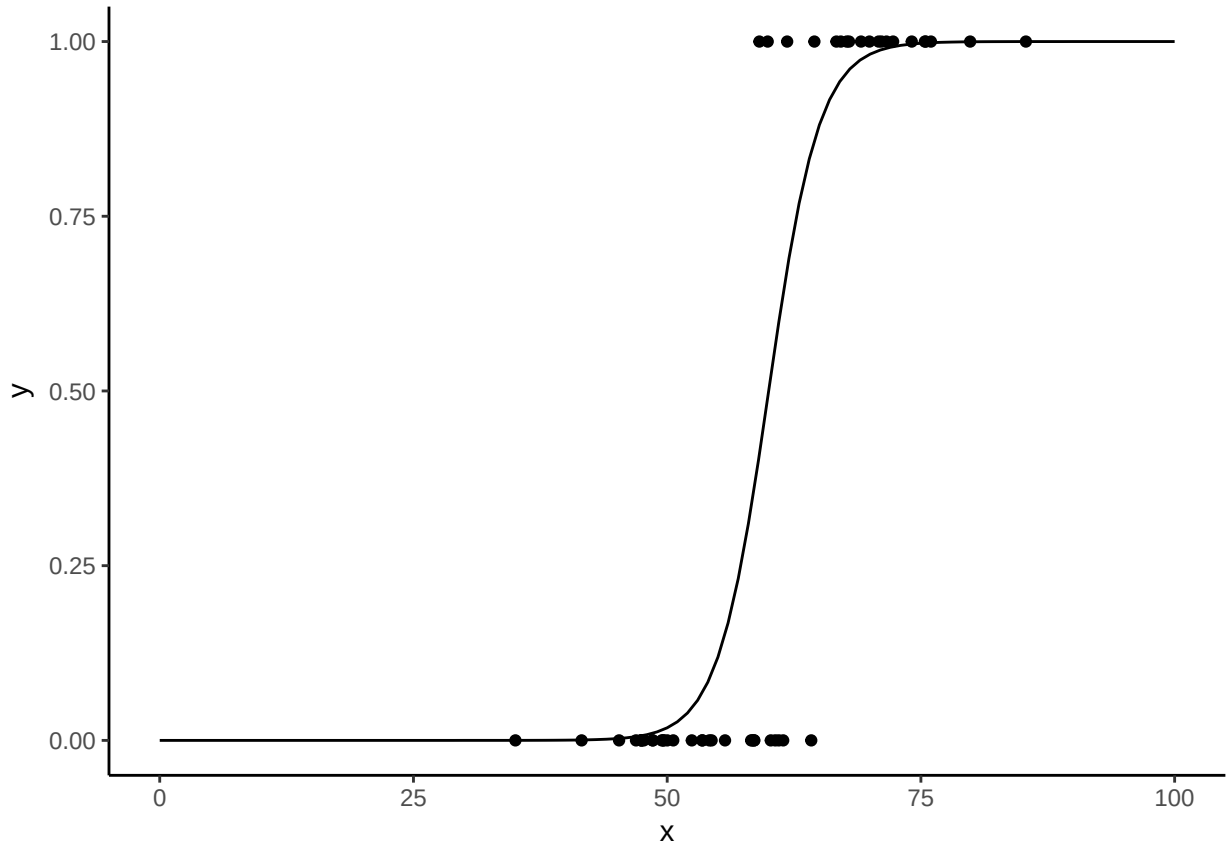
```
#' ---
#' title: "Homework 3"
#' author: "Allan Kimaina"
#' header-includes:
#' - \usepackage{pdflscape}
#' - \newcommand{\blandscape}{\begin{landscape}}
#' - \newcommand{\elandscape}{\end{landscape}}
#' output:
#'   pdf_document: default
#'   html_document: default
#' ---
#'
#'
knitr::opts_chunk$set(echo = F)

# loadl lm  package
library(dplyr)
library(car)
library(sjPlot)
library(sjmisc)
library(sjlabelled)
library(ggpubr)
library(ggpmisc)
library(gridExtra)
library(stargazer)
library(e1071)
library(jtools)
library(effects)
library(multcompView)
library(ggplot2)
library(ggrepel)
library(MASS)
library(broom)
library(ggcorrplot)
library(leaps)
library(relaimpo)
library(olsrr)

# load GLM packages
library(ROCR)
library(arm)
library(foreign)
library(nnet)
library(VGAM)
library(ordinal)
library(ModelGood)
library(InformationValue)
library(rms)
library(texreg)
```

```
#'
#' \onecolumn
#'
#' # Question 1: Do Problem 3 in chapter 5 of Gelman and Hill.
#' You are interested in how well the combined earnings of the parents in a child's family predicts hig
#'
#'
#' $$logit(0.88) = logit(0.27) + 6*x$$
#' $$1.99243 = -0.9946226 + 6*x$$
#' $$x = \frac{1.99243 + 0.9946226}{6} =  0.4978421$$
#'
#' ## Substituiting X we get:
#' $$P(y=1) = logit^{-1}(-0.9946226 +  0.4978421*x)$$
#'
#'
#'
#'
#' ## We can simulate x1 using our Regression Model:
#'
#'
#'
#'

x1=rnorm(1000000,0,1)
pr1=invlogit(-0.9946226 +  0.4978421*x1)
y1<-rbinom(1000000,1,pr1)

df <- data.frame(x1=x1,y1=y1)


logit.model <- glm(y1 ~ x1, data=df, family=binomial(link="logit"))
#summary(logit.model)
stargazer(logit.model,
          header=F,
          type = "latex",
          no.space = T,
          summary = F,
          single.row = T
          )
```

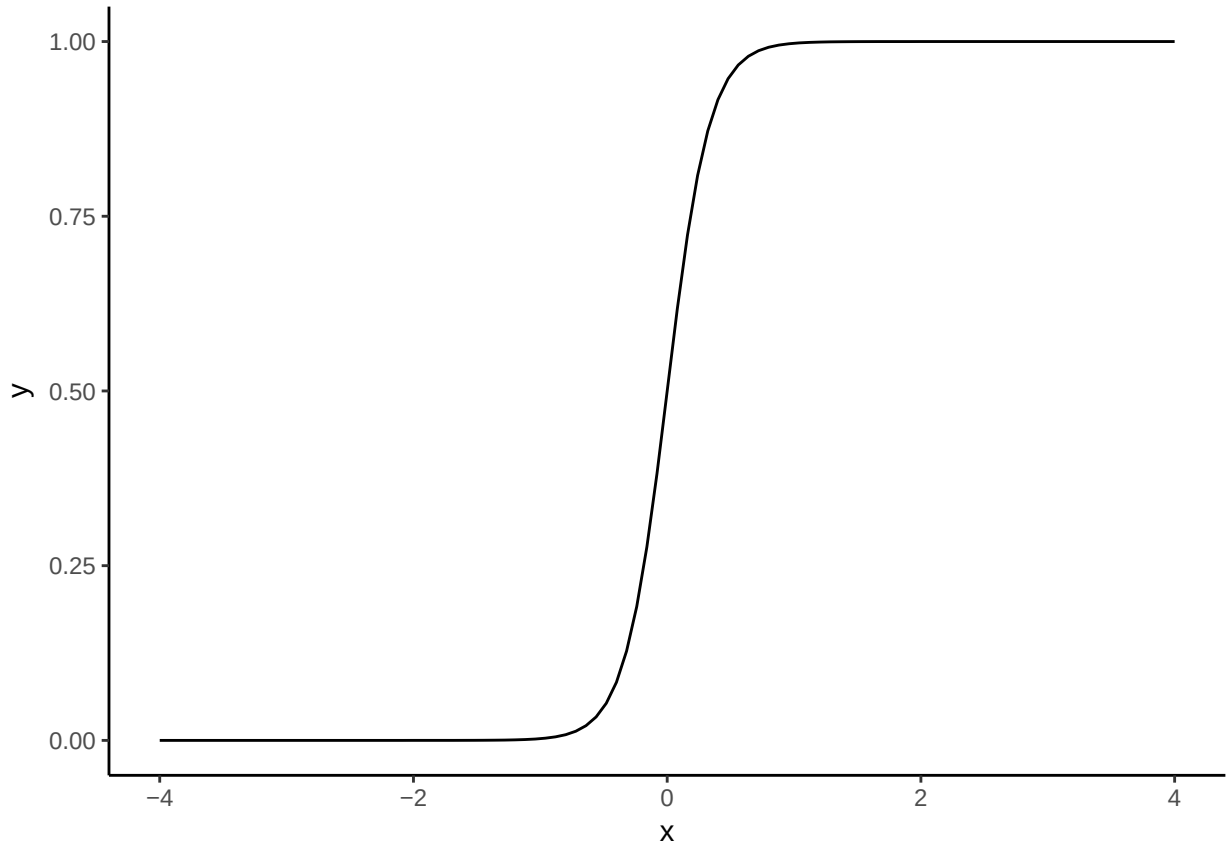
```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & y1 \\
## \hline \\[-1.8ex]
##  x1 & 0.501$^{***}$ (0.002) \\
##   Constant & $-$1.000$^{***}$ (0.002) \\
##  \hline \\[-1.8ex]
```

```
## Observations & 1,000,000 \\
## Log Likelihood & $-$569,068.200 \\
## Akaike Inf. Crit. & 1,138,140.000 \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
#'
#'
#' \onecolumn
#'
#' # Question 2: Do Problem 5 in chapter 5 of Gelman and Hill.
#' In a class of 50 students, a logistic regression is performed of course grade (pass or fail) on midt
#'
#'
#'
#'
#' ## Part A
#' Graph the fitted model. Also on this graph put a scatter plot of hypothetical data consistent with t
#'
#'
set.seed(654)
x1=rnorm(50,60,15)
pr=invlogit(-24+0.4*x1)
y1<-rbinom(50,1,pr)
df2 <- data.frame(
  x1=x1,
  y1,y1,
  pr,pr
)
ggplot(data=data.frame(x=c(0,100)), aes(x=x)) + stat_function(fun=function(x) invlogit(-24 + 0.4*x))+ge
    theme_classic()+ guides(fill=FALSE)
```




```
#'
#'
#' ## Part B
#' Suppose the midterm scores were transformed to have a mean of 0 and standard deviation of 1. What wo
#' $$P(pass) = logit^{-1}( \beta_0 + \beta_1x) =>original$$
#' $$P(pass) = logit^{-1}( \hat{\beta_0} +\hat{\beta_1}\frac{x-\bar{x}}{S_{x} }) =>standardized (x_i)$$
#'
#'
#'
#' From above we find that:
#' $$\hat{\beta_1} =\beta_1 * S_{x} = .4*15 =6$$
#' $$\hat{\beta_0} =\beta_0 + \frac{6*60}{15 } = -24+24 =0$$
#' Substituting 6 and 0 we get:
#' $$P(pass) = logit^{-1}(0 + 6x )$$
#'
#'
#'

ggplot(data=data.frame(x=c(-4,4)), aes(x=x)) + stat_function(fun=function(x) invlogit(0 + 6*x))+theme_cl
```




```
#'
#'
#' \onecolumn
#'
#' ## Part C
#'
#' Create a new predictor that is pure noise (for example, in R you can create `newpred <- rnorm(n,0,1)
#'
#'
set.seed(555)
randomNoise <- rnorm(50,0,1) # noide

x1=rnorm(50,0,1)
pr1=invlogit(6*x1)
y1<-rbinom(50,1,pr1)

df3 <- data.frame(x1=x1,y1=y1)


logit.model <- glm(y1 ~ x1, data=df3, family=binomial(link="logit"))
logit.model2 <- glm(y1 ~ x1+randomNoise, data=df3, family=binomial(link="logit"))

texreg(list(logit.model, logit.model2), single.row=TRUE,  float.pos = "h")
```

```
##
## \begin{table}[h]
```

```
## \begin{center}
## \begin{tabular}{l c c }
## \hline
##  & Model 1 & Model 2 \\
## \hline
## (Intercept)    & $0.57 \; (0.51)$      & $0.56 \; (0.53)$      \\
## x1             & $4.42 \; (1.39)^{**}$ & $4.43 \; (1.41)^{**}$ \\
## randomNoise    &                       & $0.04 \; (0.50)$      \\
## \hline
## AIC            & 29.06                 & 31.06                 \\
## BIC            & 32.89                 & 36.79                 \\
## Log Likelihood & -12.53                & -12.53                \\
## Deviance       & 25.06                 & 25.06                 \\
## Num. obs.      & 50                    & 50                    \\
## \hline
## \multicolumn{3}{l}{\scriptsize{$^{***}p<0.001$, $^{**}p<0.01$, $^*p<0.05$}}
## \end{tabular}
## \caption{Statistical models}
## \label{table:coefficients}
## \end{center}
## \end{table}
```

```
#'
#'
#' If we add a predictor that is pure noise deviance barely decreases. It only decreases by 0.006 i.e f
#'
#' \onecolumn
#'
#' # Question 3: Do Problem 8 in chapter 5 of Gelman and Hill.
#'
#' Building a logistic regression model: the folder `rodents` contains data on rodents in a sample of N
#'
#'

rodents.df <- read.table("data/rodents.txt")
rodents.df$race <- factor(rodents.df$race,
                          labels=c("White",
                                   "Black",
                                   "Puerto Rican",
                                   "Other Hispanic",
                                   "Asian/Pacific Islander",
                                   "Amer-Indian/Native Alaskan",
                                   "Two or more races"))
rodents.df$numunits <- as.factor(rodents.df$numunits)
rodents.df$extwin4_2 <- as.factor(rodents.df$extwin4_2)
rodents.df$unitflr2 <- as.factor(rodents.df$unitflr2)
rodents.df$stories <- as.factor(rodents.df$stories)
rodents.df$intcrack2 <- as.factor(rodents.df$intcrack2)
rodents.df$inthole2 <- as.factor(rodents.df$inthole2)
rodents.df$intleak2 <- as.factor(rodents.df$intleak2)
rodents.df$extflr5_2 <- as.factor(rodents.df$extflr5_2)
rodents.df$borough <-  as.factor(rodents.df$borough)
rodents.df$cd <- as.factor(rodents.df$cd)
rodents.df$help <- as.factor(rodents.df$help)
```

```r
rodents.df$old <- as.factor(rodents.df$old)
rodents.df$dilap <- as.factor(rodents.df$dilap)
rodents.df$povertyx2 <- as.factor(rodents.df$povertyx2)
rodents.df$housing <-  as.factor(rodents.df$housing)
rodents.df$regext <- as.factor(rodents.df$regext)
rodents.df$poverty <- as.factor(rodents.df$poverty)
rodents.df$intpeel_cat <- as.factor(rodents.df$intpeel_cat)
rodents.df$board2 <- as.factor(rodents.df$board2)
rodents.df$subsidy <- as.factor(rodents.df$subsidy)
rodents.df$under6 <- as.factor(rodents.df$under6)

# Missing values
missingNA <- sapply(rodents.df, function(x) sum(is.na(x)))
rodents.df <- na.omit(rodents.df)


#'
#'
#' ## Part A
#'
#' Build a logistic regression model to predict the presence of rodents (the variable `rodent2` in the
#'
#'
#'
#'

rodents.logit1 <- glm(rodent2 ~ race, data=rodents.df, family=binomial(link="logit"))
#summary(rodents.logit1)
texreg(list(rodents.logit1), single.row=TRUE,  float.pos = "h")
```

```
##
## \begin{table}[h]
## \begin{center}
## \begin{tabular}{l c }
## \hline
##  & Model 1 \\
## \hline
## (Intercept)                    & $-1.70 \; (0.17)^{***}$ \\
## raceBlack                      & $1.31 \; (0.23)^{***}$  \\
## racePuerto Rican               & $1.15 \; (0.27)^{***}$  \\
## raceOther Hispanic             & $1.46 \; (0.24)^{***}$  \\
## raceAsian/Pacific Islander     & $0.55 \; (0.39)$        \\
## raceAmer-Indian/Native Alaskan & $2.11 \; (0.93)^{*}$    \\
## raceTwo or more races          & $0.79 \; (0.85)$        \\
## \hline
## AIC                            & 877.00                  \\
## BIC                            & 909.28                  \\
## Log Likelihood                 & -431.50                 \\
## Deviance                       & 863.00                  \\
## Num. obs.                      & 744                     \\
## \hline
## \multicolumn{2}{l}{\scriptsize{$^{***}p<0.001$, $^{**}p<0.01$, $^*p<0.05$}}
## \end{tabular}
## \caption{Statistical models}
```

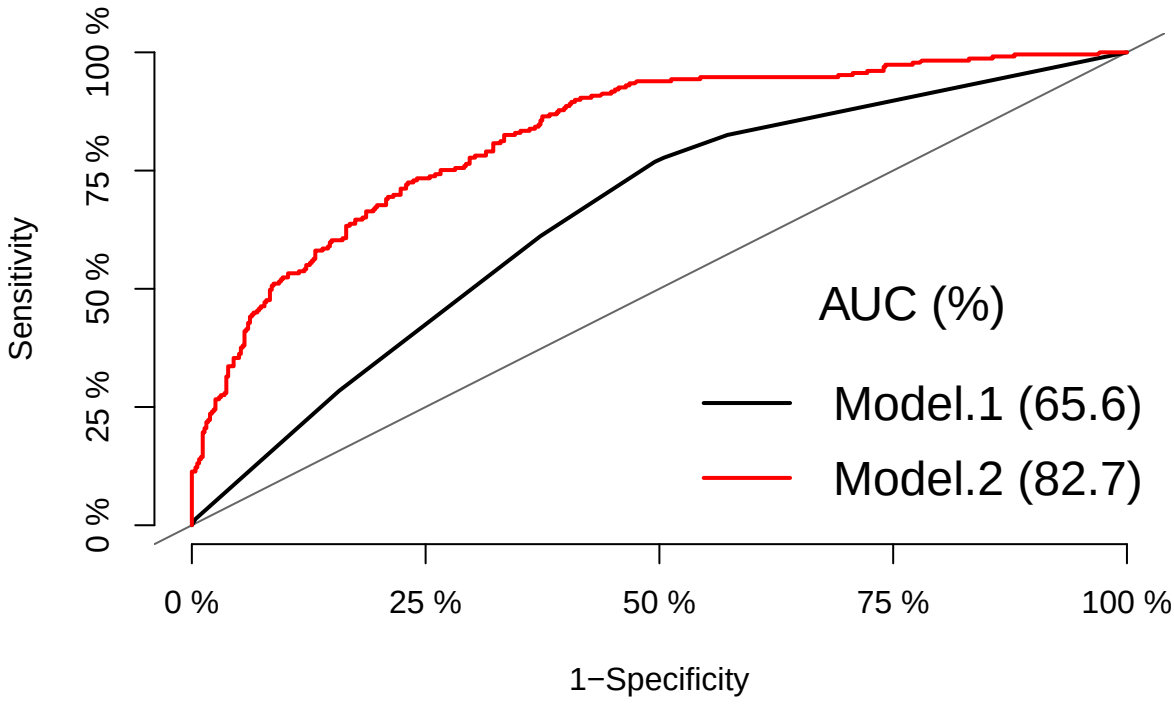
```
## \label{table:coefficients}
## \end{center}
## \end{table}
```

```
#'
#'
#'
#'
#' We find that:
#'
#' * `White Race (Intercept)`: The odds of having rodents infestation in an apartment primarily occupie
#'
#' * `Black Race`:
#'  The odds of having rodents infestation in an apartment primarily occupied by black people is $\exp(
#'
#'  * `Puerto Rican Race`:
#'  The odds of having rodents infestation in an apartment primarily occupied by Puerto Rican people is
#'
#'   * `Other Hispanic Race`:
#'  The odds of having rodents infestation in an apartment primarily occupied by Puerto Rican people is
#'
#'  * `Amer-Indian/Native Alaska`:
#'  The odds of having rodents infestation in an apartment primarily occupied by Amer-Indian/Native Ala
#'
#'
#'  The other races like Asian/Pacific Islander  were insignificant implying that the odds of rodents i
#'
#'
#'
#' .
#'
#' ##  Part B
#'
#' ### Add to your model some other potentially relevant predictors describing the apartment, building,
#'
#' We performed stepwise AIC regression selection before assessing these predictors using guidelines hi
#'
#'
#'
rodents.df$rodent2 <- as.factor(rodents.df$rodent2)
#rodents.df$hispanic_Mean10 <- rodents.df$hispanic_Mean * 10
#rodents.df$black_Mean10 <- rodents.df$black_Mean * 10
#rodents.logit2 <- glm(rodent2 ~ race + hispanic_Mean10 + black_Mean10 + borough + old + housing + pers

rodentLogitModel.all <- glm(formula=rodent2~.-sequenceno-cd, family=binomial(link="logit"),data=rodents

#summary(rodentLogitModel.all)
#rodentLogitModel.all.stepAIC  <- stepAIC(rodentLogitModel.all, direction="both")
#rodentLogitModel.all.stepAIC$anova

rodents.logit2 <- glm( rodent2 ~ personrm + unitflr2 + regext +
povertyx2 + extflr5_2 + intcrack2 + inthole2 + intleak2 +
struct + help + black_Mean + board2_Mean + help_Mean + hispanic_Mean +
 old_Mean + poverty_Mean + regext_Mean + extwin4_2_Mean +
```

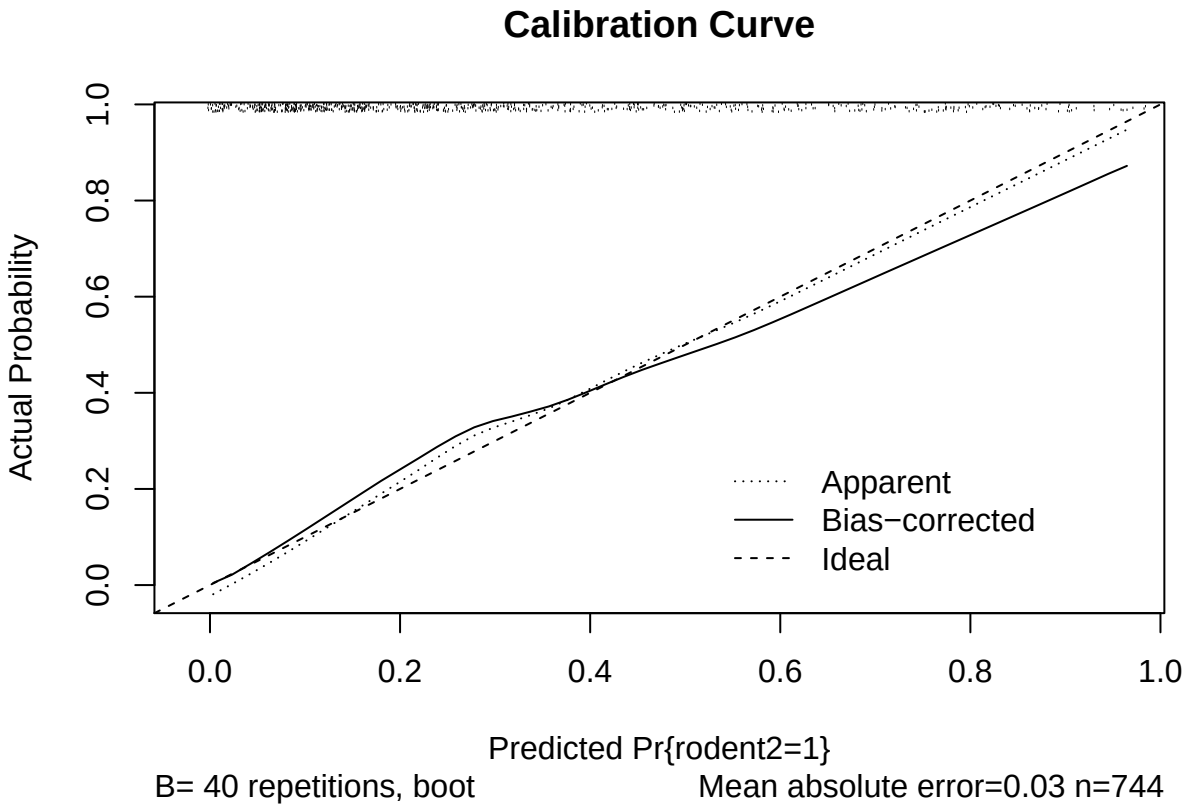
```
intcrack2_Mean + inthole2_Mean + vacrate, data=rodents.df, family=binomial(link="logit"))


#'
#'
#'

texreg(list(rodents.logit2), single.row=TRUE,  float.pos = "h")
```

```
##
## \begin{table}[h]
## \begin{center}
## \begin{tabular}{l c }
## \hline
##  & Model 1 \\
## \hline
## (Intercept)      & $-9.77 \; (2.32)^{***}$  \\
## personrm         & $0.82 \; (0.24)^{***}$   \\
## unitflr22        & $1.29 \; (0.91)$         \\
## unitflr23        & $0.99 \; (0.90)$         \\
## unitflr24        & $1.57 \; (0.92)$         \\
## unitflr25        & $1.57 \; (0.93)$         \\
## unitflr26        & $1.22 \; (0.94)$         \\
## unitflr27        & $1.24 \; (0.94)$         \\
## unitflr28        & $-1.63 \; (1.42)$        \\
## unitflr29        & $2.29 \; (1.49)$         \\
## regext1          & $-0.50 \; (0.21)^{*}$    \\
## povertyx21       & $0.37 \; (0.21)$         \\
## extflr5\_21      & $0.92 \; (0.48)$         \\
## intcrack21       & $0.94 \; (0.27)^{***}$   \\
## inthole21        & $0.58 \; (0.34)$         \\
## intleak21        & $0.50 \; (0.23)^{*}$     \\
## struct           & $-0.98 \; (0.21)^{***}$  \\
## help1            & $-0.51 \; (0.23)^{*}$    \\
## black\_Mean      & $2.35 \; (0.65)^{***}$   \\
## board2\_Mean     & $-2.63 \; (1.29)^{*}$    \\
## help\_Mean       & $5.38 \; (2.06)^{**}$    \\
## hispanic\_Mean   & $2.74 \; (1.02)^{**}$    \\
## old\_Mean        & $1.54 \; (0.71)^{*}$     \\
## poverty\_Mean    & $3.53 \; (2.20)$         \\
## regext\_Mean     & $1.70 \; (1.03)$         \\
## extwin4\_2\_Mean & $10.78 \; (7.31)$        \\
## intcrack2\_Mean  & $-20.16 \; (5.75)^{***}$ \\
## inthole2\_Mean   & $26.21 \; (7.90)^{***}$  \\
## vacrate          & $8.99 \; (4.91)$         \\
## \hline
## AIC              & 741.70                   \\
## BIC              & 875.45                   \\
## Log Likelihood   & -341.85                  \\
## Deviance         & 683.70                   \\
## Num. obs.        & 744                      \\
## \hline
## \multicolumn{2}{l}{\scriptsize{$^{***}p<0.001$, $^{**}p<0.01$, $^*p<0.05$}}
## \end{tabular}
```

```
## \caption{Statistical models}
## \label{table:coefficients}
## \end{center}
## \end{table}
#'
#'
#'
#' ## Make sure to carefully explain what your final model means.
#'
#' In general:
#'
#' * `black_Mean`: Holding the other predictors constant at the base level or 0, a 1% increase in black
#' * `hispanic_Mean`: Holding the other predictors constant, a 1% increase in Hispanic people in the di
#' * `race`: Given the fact that the householder race was statistically insignificant, we did not inclu
#'
#'
#' In summary, black and Hispanic population in that district were associated with higher chances of ro
#'
#'
#'
#' ## Check its fit using diagnostics. Make an ROC plot and a calibration curve.
#'
#' ### ROC
#'
#'
#'


plot(Roc(list("Model 1"=rodents.logit1,"Model 2"=rodents.logit2)),legend=TRUE,auc=TRUE)
```




```
#'
#'
#' The curve rises steeply indicating that the % of true positives accurately predicted by this logit m
#'
#' ### Calibration Curve
#'
#'

#Concordance(rodents.df$rodent2, predict(rodents.logit2,type="response"))

lgfit <- lrm( rodent2 ~personrm + unitflr2 + regext +
povertyx2 + extflr5_2 + intcrack2 + inthole2 + intleak2 +
struct + help + black_Mean + board2_Mean + help_Mean + hispanic_Mean +
 old_Mean + poverty_Mean + regext_Mean + extwin4_2_Mean +
intcrack2_Mean + inthole2_Mean + vacrate, data=rodents.df, x=TRUE, y=TRUE)



# exp(final.fit$coefficients)
plot(calibrate(lgfit), main="Calibration Curve")
```




```
##
## n=744   Mean absolute error=0.03   Mean squared error=0.00142
## 0.9 Quantile of absolute error=0.061
```

```
#'
#'
#' In general, the calibration plot did not provide substantial evidence that our models was overfitted
#'
#' ### Confusion Matrix
#' From the confusion matrix below we have a good truth detection rate i.e few false positive and few f
#'
threshold=0.5
predicted_values<-ifelse(predict(rodents.logit2,rodents.df,type="response")>threshold,1,0)
actual_values<-rodents.df$rodent2
conf_matrix<-table(predicted_values,actual_values)
conf_matrix
```

```
##                 actual_values
## predicted_values   0   1
##                0 466 111
##                1  49 118
```

```
#Sensitivity(actual_values,predicted_values,threshold)
#Specificity(actual_values,predicted_values,threshold)


#'
```

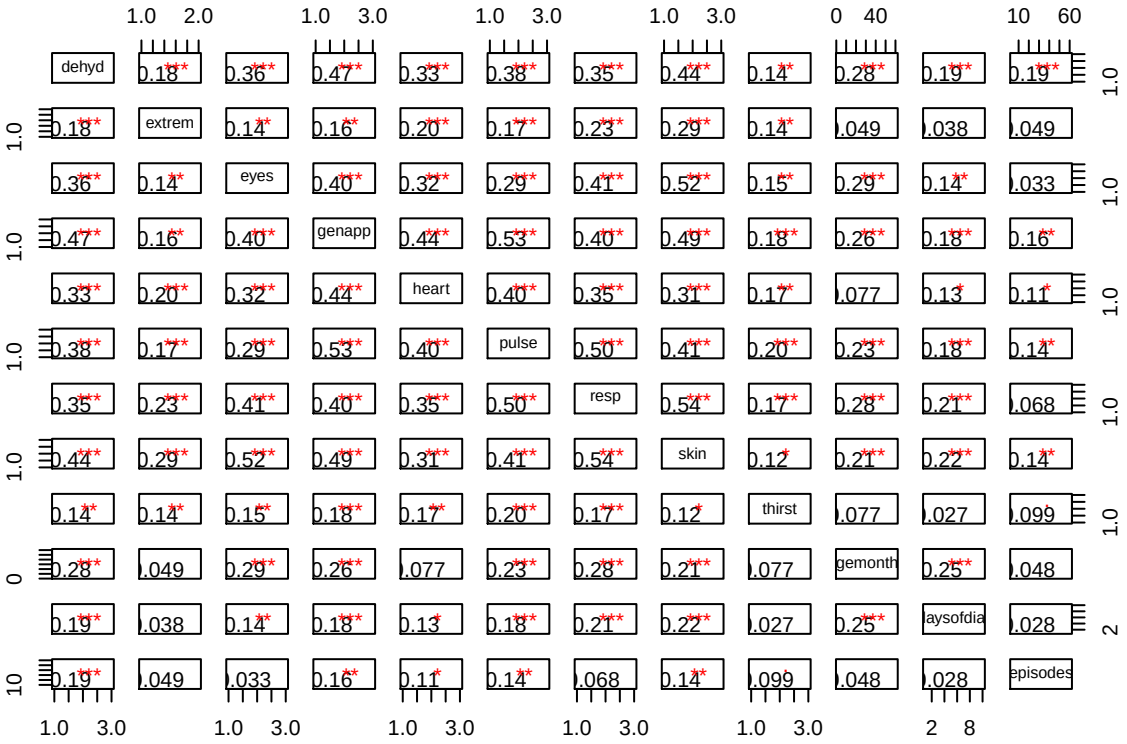
```
#'
#'
#' ### Multicollinearity Diagnosis
#' Most of the predictors in the mode had Variable Inflation Factor of less than 2. intcrack and inhole
#'
#'

vif(rodents.logit2)
```

```
##       personrm      unitflr22      unitflr23      unitflr24      unitflr25
##       1.075627      14.095042      16.097583      14.134223      11.031508
##      unitflr26      unitflr27      unitflr28      unitflr29        regext1
##       9.151704      11.042414       1.743298       1.722500       1.170401
##     povertyx21     extflr5_21     intcrack21      inthole21      intleak21
##       1.170990       1.114750       1.393961       1.269603       1.218002
##          struct          help1     black_Mean    board2_Mean      help_Mean
##       1.214216       1.134305       3.261601       4.723903       5.270872
##  hispanic_Mean       old_Mean   poverty_Mean    regext_Mean extwin4_2_Mean
##       4.879821       1.486010       6.283520       3.350785       4.226038
## intcrack2_Mean  inthole2_Mean        vacrate
##      25.007303      17.375821       2.583700
```

```
#'
#'
#'
#'
#' \onecolumn
#'
#' # Question 4: The full dehydration outcome for the Dhaka study that we looked at in class is a three
#'
#' ### Model with all Clinical Variables:
#'
#' We treated the following variables as clinical predictor variables.
#'   - genapp
#'   - tears
#'   - skin
#'   - resp
#'   - thirst
#'   - eyes
#'   - heart
#'   - mucous
#'   - pulse
#'   - urine
#'
#' We then went ahead and threw all these variables into the model.
#'
#'
dhaka=read.csv("data/dhaka.csv")
dhaka$dehyd = factor(dhaka$dehyd)
dhaka$genapp = factor(dhaka$genapp)
dhaka$tears = factor(dhaka$tears)
dhaka$skin = factor(dhaka$skin)
dhaka$resp = factor(dhaka$resp)
dhaka$thirst = factor(dhaka$thirst)
```

```r
dhaka$eyes = factor(dhaka$eyes)
dhaka$capref=factor(dhaka$capref)
dhaka$extrem=factor(dhaka$extrem)
dhaka$heart=factor(dhaka$heart)
dhaka$mucous=factor(dhaka$mucous)
dhaka$pulse=factor(dhaka$pulse)
dhaka$urine=factor(dhaka$urine)


dhaka.clinical.model<- vglm(ordered(dehyd) ~ genapp+tears+skin+resp+thirst+eyes+heart+mucous+pulse+urine



#'
#'
#' ```
#' Call:
#' vglm(formula = ordered(dehyd) ~ genapp + tears + skin + resp +
#'     thirst + eyes + heart + mucous + muac + pulse + urine, family = cumulative(parallel = TRUE),
#'     data = dhaka)
#'
#'
#' Pearson residuals:
#'                      Min      1Q  Median     3Q   Max
#' logit(P[Y<=1]) -2.018 -0.5798 -0.2227 0.6829 4.561
#' logit(P[Y<=2]) -9.291  0.1024  0.1855 0.3396 1.533
#'
#' Coefficients:
#'                  Estimate Std. Error z value Pr(>|z|)
#' (Intercept):1  1.301e-01  1.508e+00   0.086  0.93124
#' (Intercept):2  3.283e+00  1.521e+00   2.159  0.03086 *
#' genapp1       -5.018e-01  2.933e-01  -1.711  0.08711 .
#' genapp2       -1.089e+00  3.426e-01  -3.178  0.00148 **
#' tears1        -2.431e-01  2.620e-01  -0.928  0.35351
#' tears2        -9.323e-01  4.587e-01  -2.033  0.04207 *
#' skin1         -8.580e-01  2.733e-01  -3.139  0.00170 **
#' skin2         -7.688e-01  5.765e-01  -1.333  0.18238
#' resp1         -1.735e-02  3.014e-01  -0.058  0.95410
#' resp2         -6.958e-01  9.298e-01  -0.748  0.45425
#' thirst1        5.063e-01  9.202e-01   0.550  0.58221
#' thirst2        2.603e-01  9.917e-01   0.262  0.79296
#' eyes1         -2.302e-01  3.530e-01  -0.652  0.51423
#' eyes2         -1.042e+00  5.134e-01  -2.030  0.04239 *
#' heart1        -3.550e-01  2.644e-01  -1.343  0.17942
#' heart2        -1.634e+01  9.088e+02      NA       NA
#' mucous1        1.473e-02  2.732e-01   0.054  0.95699
#' mucous2        9.942e-01  1.285e+03   0.001  0.99938
#' muac           4.473e-03  8.132e-03   0.550  0.58228
#' pulse1        -6.043e-01  3.073e-01  -1.966  0.04924 *
#' pulse2        -4.337e-01  4.685e-01  -0.926  0.35453
#' urine1        -3.212e-01  2.570e-01  -1.250  0.21134
#' urine2        -3.258e-02  3.733e-01  -0.087  0.93044
#' ---
```

```
#' Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#'
#' Number of linear predictors:  2
#'
#' Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
#'
#' Residual deviance: 600.5693 on 737 degrees of freedom
#'
#' Log-likelihood: -300.2846 on 737 degrees of freedom
#'
#' Number of iterations: 14
#'
#' ```
#'
#' From this model, we observed that mucous, muac, thirst and urine were statistically insignificant th
#'
#' ### Best Model with both clinical and non clinical variables:
#'
#' We then went ahead and created a correlation matrix in order to assess correlation in potential pred
#'
#'


corr.panel <- function(x, y, digits=2, prefix="", cex.cor)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

    test <- cor.test(x,y)
    # borrowed from printCoefmat
    Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
                  cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                  symbols = c("***", "**", "*", ".", " "))

    text(0.3, 0.3, txt, cex = 1)
    text(.6, .6, Signif, cex=1, col=2)
}



pairs(dhaka[c(1,3:6,8:10,12,16:18)], lower.panel=corr.panel, upper.panel=corr.panel)
```




```
#'
#'
#'  - 3 non clinical predictors had significant correlation with the response, therefore we included it
#'
#' - We also checked for interaction between several covariates but none were significant. Potential co
#'
#' - After checking for interactions we came up  with our final model shown below:
#'
#'

#none some server
dhaka.best.model<- vglm(ordered(dehyd)~genapp+resp+skin+eyes+ pulse+daysofdiar+episodes, family=cumulati

#summary(dhaka.best.model)


#'
#'
#' ### Interpretation
#'
#' - In general only 4 predictors were statistically significant: genapp, skin, eyes, episodes
#'
#' * `genapp`:
#' General appearance  has 3 level: Normal (0), Restless/Irritable (1), Lethargic/Unconscious (2). Norm
#'     - `genapp1` Holding the other predictor constant, the odds of having no dehydration in children
#'     - `genapp2` Holding the other predictor constant, the odds of having no dehydration in children
```

```
#'
#' * `skin`:
#' Skin  has 3 level: Normal Skin Pinch (0), Slow Skin Pinch (1), Very Slow Skin Pinch (2). Normal is t
#'     - `skin1` Holding the other predictor constant, the odds of having no dehydration in children wi
#'     - `skin2` Holding the other predictor constant, the odds of having no dehydration in children wi
#'
#'
#' * `eyes`:
#' Eyes has 3 level: Normal (0), Sunken (1), Very Sunken (2). Normal is the reference in this model.
#'     - `eyes1` "Sunken" eyes was statistically insignificant.
#'     - `eyes2` Holding the other predictor constant, the odds of having no dehidration in children wi
#'
#'
#' \onecolumn
#'
#' # Source Code
#'
#'

#'
```