Lab Assignment 5 BIS 505b

Wenxin Xu

4/11/2021

Instructions

This Lab Assignment uses data from a study conducted to investigate factors related to insecticide-treated bed net (ITN) use in sub-Saharan Africa. You may keep the sections on **Public Health Application**, **Data Background** and **Data Key** in your submission if you wish. Perform your work in the **Assignment** section below. In this assignment, report any p-values that are less than 0.001 as **<0.001** and round values reported in your narrative text to **3** decimal places. **Be sure to clearly state the reference category when interpreting the effects of categorical variables in any regression model.** Perform all hypothesis testing at the α = 0.05-level.

Public Health Application

Malaria represents about 1.4% of the global burden of disease, and in Africa, it is the primary cause of disease burden as measured by Disability Adjusted Life Years (DALY) lost. The continent bears over 90% of the global burden of about 2.7 million deaths attributable to malaria and houses over 300 million people who suffer from this disease yearly, the worst hit being young children and pregnant women. More than three quarters of global malaria deaths occur in children under 5-years of age living in malarious countries in sub-Saharan Africa, where 25% of all childhood mortality below the age of 5 is attributable to malaria.

Data Background

Insecticide-treated bed nets (ITNs) are a form of personal protection that has been shown to reduce malaria illness, severe disease, and death due to malaria in endemic regions. In community-wide trials in several African settings, ITNs were shown to reduce the death of children under 5 years from all causes by about 20%.[1] To investigate factors related to ITN use, a cross-sectional study was conducted. In the study, a survey was given to the head of the household in a tropical region where malaria is a public health problem. Households (n = 1418) in a tropical region owning an insecticide-treated net were sampled and asked whether the ITN had been used the previous night. The head of the household answered a questionnaire that asked about the age of the head of household [`age`], household wealth [`wealth`], whether there was a child younger than 5 years old residing in the household [`child`], whether the household was located in a rural or urban area [`rural`], the family size in the household [`famsize`], the type of roof for the household [`roof`], and the distance to health care facility

[hcdist]. All questionnaires were completed with an interviewer and were conducted during the rainy season. The primary goal of this study was to understand the factors associated with using insecticide-treated nets. The primary outcome is whether an ITN was used on the previous night [net]. A CSV file [nets.csv] is provided which contains data from the households in the study.

Data Key – nets.csv

Variable Name	Definition
ID	Unique identifier for each household
famsize	Size of the family (including head of household)
age	Age of head of household
rural	Rural household
	0 = Urban (reference)
	1 = Rural
hcdist	Distance to health care facility
	1 = <15 miles (reference)
	2 = 15-50 miles
	3 = 50+ miles
child	Household has children under the age of 5 years
	0 = No (reference)
	1 = Yes
wealth	Wealth index of household
	1 = Lowest quartile (reference)
	2 = Second quartile
	3 = Third quartile
	4 = Highest quartile
roof	Type of roof
	0 = Corrugated metal (reference)
	1 = Thatched
net	ITN use the previous night
	0 = No ITN use
	1 = ITN use (event of interest)

Assignment

1. [5 points] Import the CSV file `nets.csv` in the third code chunk above. Name your data frame `nets` and create the factor variables `rural_factor` (reference = “Urban”), `hcdist_factor` (reference = “<15 miles”), `child_factor` (reference = “No”), `wealth_factor` (reference = “Lowest quartile”), `roof_factor` (reference = “Corrugated metal”), and `net_factor` (“success” = “ITN use”). After these steps, `nets` should contain 15 variables. [**Note:** When creating factor variables, **do not** use the `ordered=TRUE` option to create ordinal variables. No written response is required for this question. Display the code chunk(s) that perform the requested data management steps for this question.]

```
# creat 6 factor variables
nets <- mutate(nets,
               rural_factor = factor(rural,
                                     levels = c(0,1),
                                     labels = c("Urban", "Rural")),

               hcdist_factor = factor(hcdist,
                                      levels = c(1,2,3),
                                      labels = c("<15 miles", "15-50 mile
s", "50+ miles")),

               child_factor = factor(child,
                                     levels = c(0,1),
                                     labels = c("No","Yes")),

               wealth_factor = factor(wealth,
                                      levels = c(1,2,3,4),
                                      labels = c("Lowest quartile", "Secon
d quartile","Third quartile","Highest quartile")),

               roof_factor = factor(roof,
                                    levels = c(0,1),
                                    labels = c("Corrugated metal", "Thatch
ed")),

               net_factor = factor(net,
                                   levels = c(0,1),
                                   labels = c("No ITN use", "ITN use")))
```

```
# check number of variables in dataset
ncol(nets)
```

```
## [1] 15
```

2. The **research question** is: What characteristics are associated with ITN use in a household [`net`], where ITN use the previous night is the event of interest (i.e., `1` =“success”)? By the end of this question, you will complete **Table 1**, giving a snapshot of the characteristics of the sample and each characteristic’s unadjusted association with ITN use.

Table 1. Characteristics of the Sample and Unadjusted Associations with ITN Use

Used ITN the Previous Night	Yes (N=384)	No (N=1034)	Unadjusted OR (95% CI)[2]	P-value[3]
Family size, mean (SD)	xx.xx (xx.xx)	xx.xx (xx.xx)	x.xx (x.xx, x.xx)	x.xxx
Head of household age, mean (SD)	xx.xx (xx.xx)	xx.xx (xx.xx)	x.xx (x.xx, x.xx)	x.xxx
Rural household, n (%)				
Rural	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx
Urban (ref)	xxx (xx.x%)	xxx (xx.x%)	-	-
Distance to health care facility, n (%)				
50+ miles	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx
15-50 miles	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx
<15 miles (ref)	xxx (xx.x%)	xxx (xx.x%)	-	-
Children under 5 in household, n (%)				
Yes	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx
No (ref)	xxx (xx.x%)	xxx (xx.x%)	-	-
Household wealth index, n (%)				
Highest quartile	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx
Third quartile	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx
Second quartile	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx
Lowest quartile (ref)	xxx (xx.x%)	xxx (xx.x%)	-	-
Roof Type, n (%)				
Thatched	xxx (xx.x%)	xxx (xx.x%)	x.xx (x.xx, x.xx)	x.xxx

Used ITN the Previous Night	Yes (N=384)	No (N=1034)	Unadjusted OR (95% CI)[2]	P-value[3]
Corrugated metal (ref)	xxx (xx.x%)	xxx (xx.x%)	-	-

a. In this question, you will investigate the relationship between the **quantitative predictor variables** listed below and ITN use.

Quantitative predictor variables:

- Family size
- Head of household age

(i). [6 points] Report the mean and standard deviation of the quantitative variables by ITN use category. Fill in these values in **Table 1**. [**Note:** Other than filling in the requested information in **Table 1**, no written response is required for **(i)**. Display the code chunk(s) that perform the requested analyses and the **R** output.]

```
# mean and standard deviation of family size and head of household age
library(arsenal)

library(knitr)

my_labels <- list(famsize = "Family size, mean (SD)",
                  age = "Head of household age, mean (SD)",
                  net_factor = "Used ITN the Previous Night, mean (SD)")

my_controls <- tableby.control(numeric.stats = c("meansd"))

table1 <- tableby(net_factor ~ famsize + age,
                  data = nets,
                  control = my_controls
                  )


kable(summary(table1,
        labelTranslations = my_labels,
        title = "Table1. Characteristics of the Sample and Unadjusted Asso
ciations with ITN Use",
        term.name = TRUE))
```

Used ITN the Previous Night, mean (SD)	No ITN use (N=1034)	ITN use (N=384)	Total (N=1418)	p value
Family size, mean (SD)				< 0.001
Mean (SD)	5.118 (1.506)	7.195 (1.608)	5.681 (1.790)	
Head of household age, mean (SD)				< 0.001

Used ITN the Previous Night, mean (SD)	No ITN use (N=1034)	ITN use (N=384)	Total (N=1418)	p value
Mean (SD)	50.072 (6.864)	44.398 (7.427)	48.535 (7.457)	

(ii). [10 points] Run a logistic regression model for each quantitative predictor variable (i.e., *2 separate models*). For each model, estimate the unadjusted odds ratio of ITN use and the 95% confidence interval of the odds ratio. For each odds ratio, report the p-value from the Wald test of $H_0 : OR = 1$ vs. $H_1 : OR \neq 1$. Fill in these values in **Table 1**. For your written response: Interpret each odds ratio and interpret the conclusion of the hypothesis test performed for each predictor variable analyzed in this question **in the context of this application**.

```
# logistic regression for family size
mod.famsize <- glm(net ~ famsize, data = nets, family = binomial(link = "l
ogit"))

summary(mod.famsize)
```

```
##
## Call:
## glm(formula = net ~ famsize, family = binomial(link = "logit"),
##     data = nets)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8386  -0.7537  -0.3360   0.6387   2.7339
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.31247    0.34400  -18.35   <2e-16
## famsize      0.86652    0.05245   16.52   <2e-16
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1656.4  on 1417  degrees of freedom
## Residual deviance: 1228.8  on 1416  degrees of freedom
## AIC: 1232.8
##
## Number of Fisher Scoring iterations: 5
```

```
# 95% confidence interval
confint.default(mod.famsize)
```

```
##                  2.5 %     97.5 %
## (Intercept) -6.9866894 -5.6382424
## famsize      0.7637253  0.9693152
```

The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -6.312 + 0.867 Family size.

The unadjusted odds ratio is $\hat{OR} = e^b$ = 2.379 [95% CI (2.146, 2.636)], which means a 1-unit increase in family size increases the odds of ITN use by 137.862%.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly associated with family size (p-value <.001).

```
# logistic regression for head of household age
mod.age <- glm(net ~ age, data = nets, family = binomial(link = "logit"))

summary(mod.age)
```

```
##
## Call:
## glm(formula = net ~ age, family = binomial(link = "logit"), data = net
s)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7047  -0.8059  -0.5968   0.9593   2.3321
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.347070   0.443748   9.796   <2e-16
## age         -0.112873   0.009483 -11.903   <2e-16
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1656.4  on 1417  degrees of freedom
## Residual deviance: 1487.2  on 1416  degrees of freedom
## AIC: 1491.2
##
## Number of Fisher Scoring iterations: 4
```

The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = 4.347 − 0.113 Head of household age.

The unadjusted odds ratio is $\hat{OR} = e^b$ = 0.893 [95% CI (0.877, 0.91)], which means a 1-unit increase in head of household age decreases the odds of ITN use by 10.674%.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly associated with head of household age (p-value <.001).

(iii). [10 points] Using the appropriate simple logistic regression model, report the following fitted probabilities of ITN use:

- Report the fitted probabilities of ITN use for households with family sizes of 2, 5, and 10
- Do the estimated probabilities increase or decrease as expected based on the odds ratios estimated in **(ii)**? Explain.

```
# family sizes of 2, 5, and 10 are in the range
range(nets$famsize, na.rm = TRUE)
```

```
## [1]  1 12
```

```
# values of x to estimate p
pred.x_famsize <- data.frame(famsize = c(2,5,10))

# returns fitted probabilities
phat <- predict(mod.famsize, newdata = pred.x_famsize, type = "response")

cbind(pred.x_famsize, phat)
```

```
##   famsize       phat
## 1       2 0.01015658
## 2       5 0.12133327
## 3      10 0.91315151
```

- The estimated probability of ITN use for households with family sizes of 2 is equal to $\hat{p} =$ 0.01.
- The estimated probability of ITN use for households with family sizes of 5 is equal to $\hat{p} =$ 0.121.
- The estimated probability of ITN use for households with family sizes of 10 is equal to $\hat{p} =$ 0.913.
- The estimated probability of ITN use increases with households with family sizes. This is expected based on the estimated \$ = \$ 2.379 > 1.

Report the fitted probabilities of ITN use for households where the age of the head of the household is 20, 40, and 60 years

Do the estimated probabilities increase or decrease as expected based on the odds ratios estimated in **(ii)**? Explain.

```
# age of the head of the household is 20, 40, and 60 years are in the rang
e
range(nets$age, na.rm = TRUE)
```

```
## [1] 20 73
```

```
# values of x to estimate p
pred.x_age <- data.frame(age = c(20,40,60))

# returns fitted probabilities
phat <- predict(mod.age, newdata = pred.x_age, type = "response")

cbind(pred.x_age, phat)
```

```
##   age      phat
## 1  20 0.8898901
## 2  40 0.4581403
## 3  60 0.0812650
```

- The estimated probability of ITN use for households where the age of the head of the household is 20 years is equal to $\hat{p} =$ 0.89.
- The estimated probability of ITN use for households where the age of the head of the household is 40 years is equal to $\hat{p} =$ 0.458.
- The estimated probability of ITN use for households where the age of the head of the household is 60 years is equal to $\hat{p} =$ 0.081.
- The estimated probability of ITN use decreases with households with family sizes. This is expected based on the estimated $ = $ 0.893 < 1.

b. In this question, you will investigate the relationship between the **categorical predictor variables** listed below and ITN use.

Categorical predictor variables:

- Rural household indicator
- Distance to health care facility
- Presence of children under age of 5 years in household
- Wealth index of household
- Type of roof

(i). [10 points] Report the frequency and relative frequency (%) of each level of the categorical variable by ITN use category. When reporting percentages, report **row percentages** (e.g., out of households in a rural setting, what percentage used the ITN the previous night). Fill in these values in **Table 1**. [**Note:** Other than filling in the requested information in **Table 1**, no written response is required for **(i)**. Display the code chunk(s) that perform the requested analyses and the **R** output.]

```
# frequency and relative frequency of categorical variable by ITN use cate
gory
my_labels <- list(rural_factor = "Rural household, n (%)",
                  hcdist_factor = "Distance to health care facility, n
 (%)",
                  child_factor = "Children under 5 in household, n (%)",
                  wealth_factor = "Household wealth index, n (%)",
                  roof_factor = "Roof Type, , n (%)")

my_controls <- tableby.control(cat.stats = c("countrowpct"))

tab2 <- tableby(net_factor ~ rural_factor + hcdist_factor + child_factor +
wealth_factor + roof_factor,
                data = nets,
                control = my_controls)

kable(summary(tab2,
        labelTranslations = my_labels,
        title = "Table1. Characteristics of the Sample and Unadjusted Asso
ciations with ITN Use",
        term.name = TRUE))
```

net_factor	No ITN use (N=1034)	ITN use (N=384)	Total (N=1418)	p value
Rural household, n (%)				< 0.001
Urban	578 (79.4%)	150 (20.6%)	728 (100.0%)	
Rural	456 (66.1%)	234 (33.9%)	690 (100.0%)	
Distance to health care facility, n (%)				< 0.001
<15 miles	265 (64.0%)	149 (36.0%)	414 (100.0%)	
15-50 miles	188 (70.9%)	77 (29.1%)	265 (100.0%)	
50+ miles	581 (78.6%)	158 (21.4%)	739 (100.0%)	
Children under 5 in household, n (%)				< 0.001
No	611 (78.5%)	167 (21.5%)	778 (100.0%)	
Yes	423 (66.1%)	217 (33.9%)	640 (100.0%)	

net_factor	No ITN use (N=1034)	ITN use (N=384)	Total (N=1418)	p value
Household wealth index, n (%)				< 0.001
Lowest quartile	303 (83.7%)	59 (16.3%)	362 (100.0%)	
Second quartile	310 (74.2%)	108 (25.8%)	418 (100.0%)	
Third quartile	275 (69.6%)	120 (30.4%)	395 (100.0%)	
Highest quartile	146 (60.1%)	97 (39.9%)	243 (100.0%)	
Roof Type, , n (%)				< 0.001
Corrugated metal	640 (80.5%)	155 (19.5%)	795 (100.0%)	
Thatched	394 (63.2%)	229 (36.8%)	623 (100.0%)	

(ii). [25 points] Run a logistic regression model for each categorical predictor variable (i.e., *5 separate models*). For each model, estimate the unadjusted odds ratio(s) of ITN use and the 95% confidence interval of the odds ratio(s). [**Note:** Odds ratios should be computed assuming the reference levels specified in question **1**. Models containing categorical predictors with 3 or more levels will have more than one odds ratio]. For each odds ratio, report the p-value from the Wald test of $H_0 : OR = 1$ vs. $H_1 : OR \neq 1$. Fill in these values in **Table 1**. For your written response: Interpret each odds ratio and interpret the conclusion of the hypothesis test performed for each odds ratio **in the context of this application**.

- Rural household indicator

```
# check dummy variable coding
contrasts(nets$rural_factor)
```

```
##       Rural
## Urban     0
## Rural     1
```

```
mod.rural <- glm(net ~ rural_factor, data = nets,
                 family = binomial(link = "logit"))

summary(mod.rural)
```

```
##
## Call:
## glm(formula = net ~ rural_factor, family = binomial(link = "logit"),
##     data = nets)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9102  -0.9102  -0.6793   1.4706   1.7774
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.34894    0.09163 -14.721  < 2e-16
## rural_factorRural  0.68177    0.12191   5.592 2.24e-08
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1656.4  on 1417  degrees of freedom
## Residual deviance: 1624.5  on 1416  degrees of freedom
## AIC: 1628.5
##
## Number of Fisher Scoring iterations: 4
```

The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -1.349 + 0.682 Rural.

The unadjusted odds ratio is $\hat{OR} = e^b$ = 1.977 [95% CI (1.557, 2.511)], which means the odds of ITN use is 97.737% higher in households in rural compared to households in urban.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly different in rural and urban (p-value <.001).

- Distance to health care facility

```
# check dummy variable coding
contrasts(nets$hcdist_factor)
```

```
##             15-50 miles 50+ miles
## <15 miles             0         0
## 15-50 miles           1         0
## 50+ miles             0         1
```

```
mod.hcdist <- glm(net ~ hcdist_factor, data = nets,
                  family = binomial(link = "logit"))

summary(mod.hcdist)
```

```
##
## Call:
## glm(formula = net ~ hcdist_factor, family = binomial(link = "logit"),
##     data = nets)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9446  -0.8286  -0.6936   1.4296   1.7565
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.5758     0.1024  -5.623 1.88e-08
## hcdist_factor15-50 miles  -0.3169     0.1697  -1.867   0.0619
## hcdist_factor50+ miles    -0.7264     0.1361  -5.335 9.54e-08
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1656.4  on 1417  degrees of freedom
## Residual deviance: 1627.4  on 1415  degrees of freedom
## AIC: 1633.4
##
## Number of Fisher Scoring iterations: 4
```

The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -0.576 − 0.317 15-50 miles − 0.726 50+ miles.

The unadjusted odds ratio of the first dummy variable 15-50 miles is $\hat{OR} = e^{b1}$ = 0.728 [95% CI (0.522, 1.016)], which means the odds of ITN use is 27.156% lower in household located 15-50 miles from health care facility compared to household located <15 miles.

We fail to reject H_0 of the Wald test and can't conclude that the odds of ITN use is significantly different in household located 15-50 miles from health care facility and household located <15 miles (p-value = 0.062).

The unadjusted odds ratio of the second dummy variable 50+ miles is $\hat{OR} = e^{b2}$ = 0.484 [95% CI (0.37, 0.632)], which means the odds of ITN use is 51.634% lower in household located 50+ miles from health care facility compared to household located <15 miles.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly different in household located 50+ miles from health care facility and household located <15 miles (p-value <.001).

- Presence of children under age of 5 years in household

```
# check dummy variable coding
contrasts(nets$child_factor)
```

```
##     Yes
## No    0
## Yes   1
```

```
mod.child <- glm(net ~ child_factor, data = nets,
                 family = binomial(link = "logit"))

summary(mod.child)
```

```
##
## Call:
## glm(formula = net ~ child_factor, family = binomial(link = "logit"),
##     data = nets)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9101  -0.9101  -0.6952   1.4708   1.7543
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.29710    0.08732 -14.855  < 2e-16
## child_factorYes   0.62963    0.12082   5.211 1.87e-07
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1656.4  on 1417  degrees of freedom
## Residual deviance: 1628.9  on 1416  degrees of freedom
## AIC: 1632.9
##
## Number of Fisher Scoring iterations: 4
```

The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -1.297 + 0.63 Children.

The unadjusted odds ratio is $\hat{OR} = e^{b}$ = 1.877 [95% CI (1.481, 2.378)], which means the odds of ITN use is 87.691% higher in households which have children under the age of 5 years compared to households which not.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly different in households which have children under the age of 5 years and households which not (p-value <.001).

- Wealth index of household

```
# check dummy variable coding
contrasts(nets$wealth_factor)
```

```
##                   Second quartile Third quartile Highest quartile
## Lowest quartile                 0              0                0
## Second quartile                 1              0                0
## Third quartile                  0              1                0
## Highest quartile                0              0                1
```

```
mod.wealth <- glm(net ~ wealth_factor, data = nets,
                  family = binomial(link = "logit"))

summary(mod.wealth)
```

```
##
## Call:
## glm(formula = net ~ wealth_factor, family = binomial(link = "logit"),
##     data = nets)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0094  -0.8510  -0.7732   1.3553   1.9048
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.6362     0.1423 -11.498  < 2e-16
## wealth_factorSecond quartile    0.5818     0.1809   3.215   0.0013
## wealth_factorThird quartile     0.8069     0.1795   4.495 6.94e-06
## wealth_factorHighest quartile   1.2273     0.1934   6.346 2.22e-10
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1656.4  on 1417  degrees of freedom
## Residual deviance: 1611.5  on 1414  degrees of freedom
## AIC: 1619.5
##
## Number of Fisher Scoring iterations: 4
```

The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -1.636 + 0.582 Second quartile + 0.807 Third quartile + 1.227 Highest quartile.

The unadjusted odds ratio of the first dummy variable Second quartile is $\hat{OR} = e^{b1}$ = 1.789 [95% CI (1.255, 2.551)], which means the odds of ITN use is 78.917% higher in household within the second quartile of wealth index compared to household within the lowest quartile of wealth index.

We fail to reject H_0 of the Wald test and can’t conclude that the odds of ITN use is significantly different in household within the second quartile of wealth index and household within the lowest quartile of wealth index (p-value = 0.001).

The unadjusted odds ratio of the second dummy variable Third quartile is $\hat{OR} = e^{b2}$ = 2.241 [95% CI (1.576, 3.186)], which means the odds of ITN use is 124.099% higher in household within the third quartile of wealth index compared to household within the lowest quartile of wealth index.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly different in household within the second quartile of wealth index and household within the lowest quartile of wealth index. (p-value <.001).

The unadjusted odds ratio of the third dummy variable Highest quartile is $\hat{OR} = e^{b3}$ = 3.412 [95% CI (2.335, 4.985)], which means the odds of ITN use is 241.2% higher in household within the highest quartile of wealth index compared to household within the lowest quartile of wealth index.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly different in household within the highest quartile of wealth index and household within the lowest quartile of wealth index. (p-value <.001).

- Type of roof

```
# check dummy variable coding
contrasts(nets$roof_factor)
```

```
##                  Thatched
## Corrugated metal        0
## Thatched                1
```

```
mod.roof<- glm(net ~ roof_factor, data = nets,
                 family = binomial(link = "logit"))

summary(mod.roof)
```

```
## 
## Call:
## glm(formula = net ~ roof_factor, family = binomial(link = "logit"), 
##     data = nets)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.9573  -0.9573  -0.6586   1.4148   1.8083  
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         -1.41804    0.08952 -15.840  < 2e-16 ***
## roof_factorThatched  0.87541    0.12214   7.167 7.66e-13 ***
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1656.4  on 1417  degrees of freedom
## Residual deviance: 1603.9  on 1416  degrees of freedom
## AIC: 1607.9
## 
## Number of Fisher Scoring iterations: 4
```

The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -1.418 + 0.875 Thatched.

The unadjusted odds ratio is $\hat{OR} = e^b$ = 2.4 [95% CI (1.889, 3.049)], which means the odds of ITN use is 139.987% higher in households which have thatched roof compared to households which have corrugated metal roof.

We reject H_0 of the Wald test and conclude that the odds of ITN use is significantly different in households which have thatched roof and households which have corrugated metal roof (p-value <.001).

3. The **research question** would like to investigate the effect of each predictor variable while controlling or adjusting for the other predictors.

a. [4 points] Build a multiple logistic regression model that includes the variables: (1) family size, (2) age of head of household, (3) living in a rural area, (4) distance to health care facility, (5) having a child under age 5, (6) wealth index and (7) roof type. Except for family size and age, which are continuous variables, the other predictor variables in the model should be treated as categorical variables and their factor versions should be included in the model using the reference levels specified in question **1**. Report the fitted multiple logistic regression model.

```
mod.mul <- glm(net ~ famsize + age + rural_factor + hcdist_factor + child_
factor + wealth_factor + roof_factor,
               data = nets,
               family = binomial(link = "logit"))

summary(mod.mul)
```

```
##
## Call:
## glm(formula = net ~ famsize + age + rural_factor + hcdist_factor +
##     child_factor + wealth_factor + roof_factor, family = binomial(link
= "logit"),
##     data = nets)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6500  -0.5448  -0.2449   0.2763   2.9188
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.60068    0.70824  -3.672 0.000241
## famsize                         0.88754    0.05983  14.834  < 2e-16
## age                            -0.11190    0.01210  -9.250  < 2e-16
## rural_factorRural               0.77607    0.16309   4.759 1.95e-06
## hcdist_factor15-50 miles       -0.39336    0.23499  -1.674 0.094148
## hcdist_factor50+ miles         -0.85479    0.18508  -4.618 3.87e-06
## child_factorYes                 0.74270    0.16314   4.553 5.30e-06
## wealth_factorSecond quartile    0.57660    0.23439   2.460 0.013894
## wealth_factorThird quartile     0.68047    0.23525   2.892 0.003822
## wealth_factorHighest quartile  1.42797    0.25794   5.536 3.09e-08
## roof_factorThatched             1.09797    0.16525   6.644 3.05e-11
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1656.39  on 1417  degrees of freedom
## Residual deviance:  980.84  on 1407  degrees of freedom
## AIC: 1002.8
##
## Number of Fisher Scoring iterations: 6
```

- The fitted model is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -2.6007 + 0.888 Family size − 0.112 Age + 0.776 Rural − 0.393 15-50 miles − 0.855 50+ miles + 0.743 Children + 0.577 Second quartile + 0.68 Third quartile + 1.428 Highest quartile + 1.098 Thatched.

b. [5 points] Report the fitted probability of ITN use for households of size 5, with a head of household that is 50 years of age, without children under the age of 5, in the second quartile of wealth, and assume the house has a thatched roof, is in a rural setting, and is 30 miles from the nearest health care facility.

```
# values of x used to estimate p
pred.x <- data.frame(famsize = 5, age = 50, rural_factor = "Rural", hcdist
_factor = "15-50 miles", child_factor = "No", wealth_factor = "Second quar
tile", roof_factor = "Thatched")

# return the predicted probability
phat <-predict(mod.mul, newdata = pred.x, type = "response")
cbind(pred.x, phat)
```

```
##   famsize age rural_factor hcdist_factor child_factor   wealth_factor
## 1       5  50        Rural   15-50 miles           No Second quartile
##   roof_factor      phat
## 1    Thatched 0.1543882
```

The fitted probability is 0.154.

c. [4 points] Interpret the odds ratio associated with family size in the multiple logistic regression model. Is family size significantly associated with ITN use, adjusted for all the other variables included in the model? If you were volunteering in Africa with an organization that wanted to increase ITN utilization, which households would you approach first: smaller sized households or larger sized households?

While controlling for the other predictors, the odds ratio associated with family size is $\hat{OR} = e^{b1}$ = 2.43 [95% CI (2.16, 2.73)], which means 1-unit increase in family size increases the odds of ITN use by 142.916%. I would approach smaller sized households first because smaller sized households have a lower odds of ITN use compared to larger sized households.

d. [6 points] Interpret the odds ratios associated with the distance to health care facility dummy variables in the multiple logistic regression model. As a volunteer, which households would you approach first: households closer to health care facilities or households further away?

While controlling for the other predictors, the odds ratio associated with the first dummy variable 15-50 miles of distance to health care facility is $\hat{OR} = e^{b4}$ = 0.675 [95% CI (0.426, 1.07)]. The odds of ITN use is 32.521% lower in households 15-50 miles from the nearest health care facility compared to households <15 miles from the nearest health care facility. While controlling for the other predictors, the odds ratio associated with the second dummy variable 50+ miles of distance to health care facility is $\hat{OR} = e^{b5}$ = 0.425 [95% CI (0.296, 0.611)]. The odds of ITN use is 57.463% lower in households >50 miles from the nearest health care facility compared to households <15 miles from the nearest health care facility. I would approach households further away to health care facilities first because smaller they have a lower odds of ITN use compared to households closer to health care facilities.

e. [10 points] Perform a likelihood ratio test to determine if distance to health care facility is an overall important predictor in the multiple regression model. (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

```
# reduced model
mod.red <- glm(net ~ famsize + age + rural_factor + child_factor + wealth_
factor + roof_factor,
               data = nets,
               family = binomial(link = "logit"))

# LRT comparing full and reduced models
anova(mod.red, mod.mul, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: net ~ famsize + age + rural_factor + child_factor + wealth_fac
tor +
##     roof_factor
## Model 2: net ~ famsize + age + rural_factor + hcdist_factor + child_fac
tor +
##     wealth_factor + roof_factor
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1409    1002.86
## 2      1407     980.84  2   22.022 1.652e-05
```

i. State the null and alternative hypotheses;

$H_0 : \beta_4 = \beta_5 = 0$ vs. $H_1 : \beta_4, \beta_5$ not all 0.

ii. From your **R** output, report the value of the test statistic and p-value;

Test statistic G = 22.022 and p-value <.001.

iii. State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that at least one β_4 or β_5 is not equal to 0. The overall effect of distance to health care facility is statistically significant in the full model that controls for other predictors.

f. [5 points] In a few sentences, summarize your findings. In particular, list which factors are associated with an increased odds of ITN utilization and which factors are associated with a decreased odds of ITN utilization based on the adjusted odds ratios estimated in your multiple logistic regression model.

Based on the adjusted odds ratios estimated in your multiple logistic regression model, larger family size, smaller age of a head of household, household in a rural setting, households closer to health care facilities, household with children under the age of 5, household with higher wealth index and household has a thatched roof are associated with an increased odds of ITN utilization. In contrast, smaller family size, larger age of a head of household, household in a urban setting, households further from health care facilities, household without children under the age of 5, household with lower wealth index and household has a corrugated metal roof are associated with a decreased odds of ITN utilization.

Table 1. Characteristics of the Sample and Unadjusted Associations with ITN Use

Used ITN the Previous Night	**Yes (N=384)**	**No (N=1034)**	**Unadjusted OR (95% CI)[4]**	**P-value[5]**
Family size, mean (SD)	7.2(1.61)	5.12 (5.12)	2.38 (2.15, 2.64)	<.001
Head of household age, mean (SD)	44.4(7.43)	50.07(6.86)	0.89 (0.88, 0.91)	<.001
Rural household, n (%)				

Used ITN the Previous Night	Yes (N=384)	No (N=1034)	Unadjusted OR (95% CI)[4]	P-value[5]
Rural	234 (16.5%)	456 (32.2%)	1.98 (1.56, 2.51)	<.001
Urban (ref)	150 (10.6%)	578 (40.8%)	-	-
Distance to health care facility, n (%)				
50+ miles	158 (11.1%)	581 (41%)	0.48 (0.37, 0.63)	<.001
15-50 miles	77 (5.4%)	188 (13.3%)	0.73 (0.52, 1.02	0.062
<15 miles (ref)	149 (10.5%)	265 (18.7%)	-	-
Children under 5 in household, n (%)				
Yes	217 (15.3%)	423 (29.8%)	1.877 (1.481, 2.378)	<.001
No (ref)	167 (11.8%)	611 (43.1%)	-	-
Household wealth index, n (%)				
Highest quartile	97 (6.8%)	146 (10.3%)	3.41 (2.34, 4.98)	<.001
Third quartile	120 (8.5%)	275 (19.4%)	2.24 (1.58, 3.19)	<.001
Second quartile	108 (7.6%)	310 (21.9%)	1.79 (1.26, 2.55)	0.001
Lowest quartile (ref)	59 (4.2%)	303 (21.4%)	-	-
Roof Type, n (%)				
Thatched	229 (16.1%)	394 (27.8%)	2.4 (1.89, 3.05)	<.001
Corrugated metal (ref)	155 (10.9%)	640 (45.1%)	-	-

1. https://www.cdc.gov/malaria/malaria_worldwide/reduction/itn.html (https://www.cdc.gov/malaria/malaria_worldwide/reduction/itn.html)↩

2. Categorical variable comparisons relative to reference cateogry↩

3. From unadjusted logistic regression model↩

4. Categorical variable comparisons relative to reference cateogry↩

5. From unadjusted logistic regression model↩