

36-463 / 36-663: Multilevel & Hierarchical Models

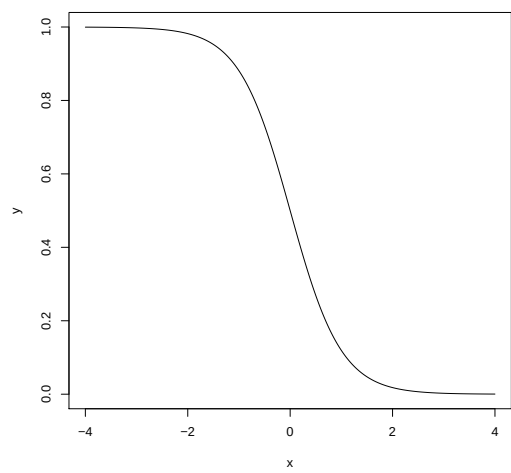
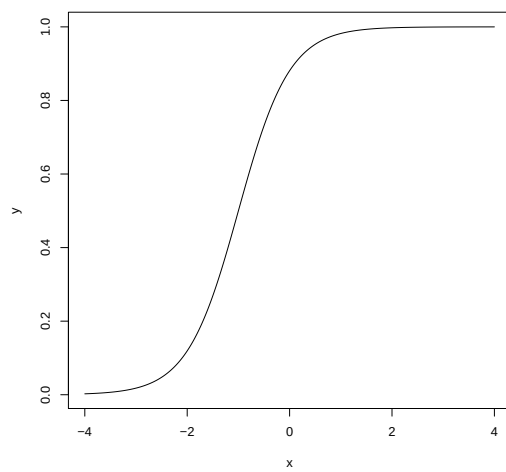
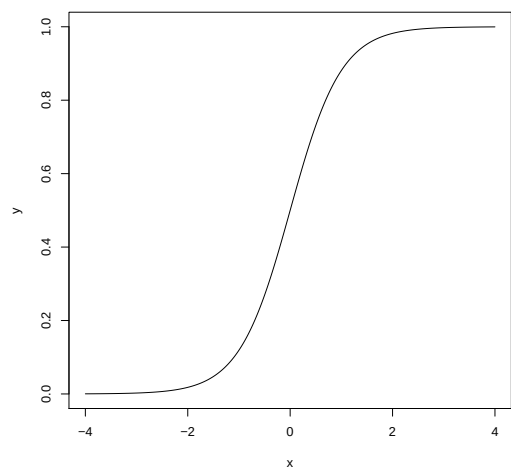
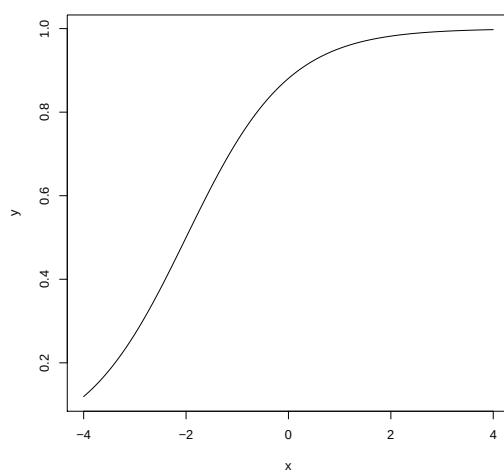
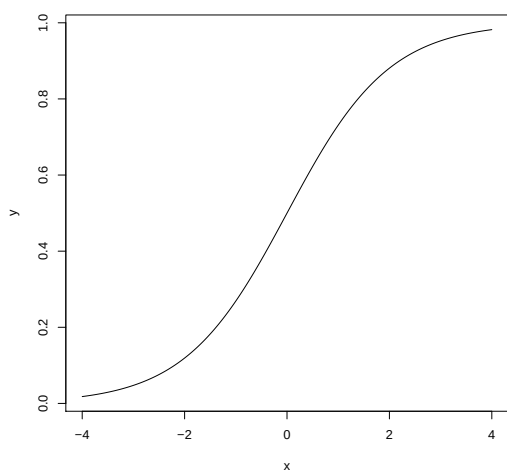
HW04 Solution

September 30, 2016

Exercise 1a

Part a-d

The following from left to right is for a-d part plots. to do it by hand please refer to page 80 in the textbook for instructions.



Exercise 1b)

```
asian <- rodents$race==5 | rodents$race==6 | rodents$race==7
black <- rodents$race==2
hisp <- rodents$race==3 | rodents$race==4

object <- glm(rodent2 ~ asian + black + hisp, family = binomial(link = "logit"),
data = rodents)

summary(object)

Call:
glm(formula = rodent2 ~ asian + black + hisp, family = binomial(link = "logit"),
    data = rodents)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9922  -0.9293  -0.4690  -0.4690   2.1270

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.1521     0.1281 -16.798  <2e-16 ***
asianTRUE      0.5518     0.2665   2.070   0.0384 *
blackTRUE      1.5361     0.1687   9.108  <2e-16 ***
hispTRUE       1.6995     0.1664  10.212  <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1672.2  on 1521  degrees of freedom
Residual deviance: 1526.3  on 1518  degrees of freedom
(225 observations deleted due to missingness)
AIC: 1534.3

Number of Fisher Scoring iterations: 4
```

From the regression output above you can see that the white race is used as a reference level. Now if we divide the coefficients for asian, black and hisp by 4 we get 0.138, 0.384, 0.425. For example, we can say that asian people on average have a 13.8 % more in probability of having rodents compare to white people. Now, for the intercept, we can say that white people on average have a 10 % probability of getting rodents.

Exercise 1b part b)

```
object <- glm(rodent2 ~ asian + black + hisp + inthole2 + intcrack2 + inthole2:intcrack2,
family = binomial(link = "logit"), data = rodents)
summary(object)

Call:
glm(formula = rodent2 ~ asian + black + hisp + inthole2 + intcrack2 +
    inthole2:intcrack2, family = binomial(link = "logit"), data = rodents)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8929	-0.7809	-0.4119	-0.4119	2.2401

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.4242	0.1420	-17.072	< 2e-16 ***
asianTRUE	0.6269	0.2800	2.239	0.025167 *
blackTRUE	1.3929	0.1804	7.719	1.17e-14 ***
hispTRUE	1.6065	0.1782	9.015	< 2e-16 ***
inhole2	1.4109	0.3847	3.667	0.000245 ***
intcrack2	1.2956	0.2074	6.245	4.23e-10 ***
inhole2:intcrack2	-0.2795	0.5154	-0.542	0.587666

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1604.3 on 1456 degrees of freedom
Residual deviance: 1353.4 on 1450 degrees of freedom
(290 observations deleted due to missingness)
AIC: 1367.4

Number of Fisher Scoring iterations: 4

There are many models you can try for this question, here is one example. Notice that by including the inhole2 and intcrack2 variables and their interaction the AIC values went down from 1534.3 to 1367.4. In this particular model, the race indicator variables significance did not change. It is also interesting to see that in this model inhole2 and intcrack2 are independent.

Exercise 2a)

```
object <- glm(round(fupacts) ~ factor(women_alone) + factor(couples),  
family = poisson,data= risky[risky$sex == "woman",])  
summary(object)
```

Call:

```
glm(formula = round(fupacts) ~ factor(women_alone) + factor(couples),  
family = poisson, data = risky[risky$sex == "woman", ])
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-6.6380	-4.8878	-3.5695	0.5786	27.4059

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.09248	0.02684	115.213	< 2e-16 ***
factor(women_alone)1	-0.61215	0.04321	-14.167	< 2e-16 ***

```

factor(couples)1      -0.22360      0.03770  -5.931 3.01e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 7354.4  on 216  degrees of freedom
Residual deviance: 7144.3  on 214  degrees of freedom
AIC: 7796.3

Number of Fisher Scoring iterations: 6

yhat <- predict(object,type = "response")

z <- (risky$fupacts - yhat)/sqrt(yhat)

n <- nrow(risky[risky$sex == "woman",])
k <- 3
cat("overdispersion ratio is ", sum(z^2)/(n-k),"\n")
overdispersion ratio is 86.1367
cat("p-value of overdispersion test is ", 1- pchisq(sum(z^2), n-k),"\n")
p-value of overdispersion test is 0

library(lmtest)
waldtest(object,test = "Chisq")
Wald test

Model 1: round(fupacts) ~ factor(women_alone) + factor(couples)
Model 2: round(fupacts) ~ 1
      Res.Df Df  Chisq Pr(>Chisq)
1        214
2        216 -2 200.83 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

So there is significant dispersion in the model which is reflected by the chisq test above. this model is better than the null model with only the intercept.

part b

```

> object <- glm(round(fupacts) ~ factor(women_alone) + factor(couples)+ factor(bs_hiv),
offset = log(bupacts),family = poisson,data= risky[risky$sex == "woman",],
subset = bupacts >0)
> summary(object)

```

Call:

```

glm(formula = round(fupacts) ~ factor(women_alone) + factor(couples) +
    factor(bs_hiv), family = poisson, data = risky[risky$sex ==
    "woman", ], subset = bupacts > 0, offset = log(bupacts))

```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

-11.161 -3.328 -1.385 1.955 19.503

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.04182	0.02730	-1.532	0.126
factor(women_alone)1	-0.61974	0.04333	-14.303	< 2e-16 ***
factor(couples)1	-0.32625	0.03827	-8.525	< 2e-16 ***
factor(bs_hiv)positive	-0.34982	0.04841	-7.227	4.95e-13 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 5004.6 on 212 degrees of freedom
Residual deviance: 4724.9 on 209 degrees of freedom
AIC: 5372.4

Number of Fisher Scoring iterations: 6

```
yhat <- predict(object,type = "response")
z <- (risky[risky$sex == "woman" & risky$bupacts >0,]$fupacts - yhat)/sqrt(yhat)
cat("overdispersion ratio is ", sum(z^2)/(n-k),"\n")
overdispersion ratio is 32.9315
```

k <- 4

```
cat("p-value of overdispersion test is ", 1- pchisq(sum(z^2), n-k),"\n")
p-value of overdispersion test is 0
```

there is still evidence of overdispersion. The model fits better including the pre-treatment variables.

part c)

```
> object <- glm(round(fupacts) ~ factor(women_alone) + factor(couples)+
factor(bs_hiv), offset = log(bupacts),family = quasipoisson,
data= risky[risky$sex == "woman",],subset = bupacts >0)
> summary(object)
```

Call:

```
glm(formula = round(fupacts) ~ factor(women_alone) + factor(couples) +
    factor(bs_hiv), family = quasipoisson, data = risky[risky$sex ==
    "woman", ], subset = bupacts > 0, offset = log(bupacts))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-11.161	-3.328	-1.385	1.955	19.503

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.04182	0.15865	-0.264	0.7923
factor(women_alone)1	-0.61974	0.25177	-2.462	0.0146 *
factor(couples)1	-0.32625	0.22236	-1.467	0.1438

```

factor(bs_hiv)positive -0.34982    0.28126  -1.244    0.2150
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

(Dispersion parameter for quasipoisson family taken to be 33.76132)

Null deviance: 5004.6  on 212  degrees of freedom
Residual deviance: 4724.9  on 209  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6

```

The model seems to suggest that the group in which just the woman participated is effective compared to control.

part d

I ran my model with only women subjects included. Including men and women in the model is not consistent with the second treatment assignment that only woman participant are allowed.

exercise 2b

Consider the “nes” data set and fit logistic regression and probit regression using the following variables.

```

> m1<-glm(vote~income+south+cd,family=binomial(link="logit"))
> summary(m1)

```

Call:

```
glm(formula = vote ~ income + south + cd, family = binomial(link = "logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9521	-1.3223	0.6631	0.8105	1.1656

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.1795551	0.0383996	4.676	2.93e-06 ***
income	0.3055781	0.0114665	26.650	< 2e-16 ***
south	-0.4580166	0.0287844	-15.912	< 2e-16 ***
cd	0.0008552	0.0014971	0.571	0.568

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

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 36122  on 30316  degrees of freedom
Residual deviance: 35003  on 30313  degrees of freedom
AIC: 35011

```

Number of Fisher Scoring iterations: 4

```
> m2<-glm(vote~income+south+cd,family=binomial(link="probit"))
> summary(m2)
```

Call:

```
glm(formula = vote ~ income + south + cd, family = binomial(link = "probit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9667	-1.3235	0.6617	0.8137	1.1594

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.1189543	0.0232770	5.110	3.22e-07 ***
income	0.1835891	0.0068519	26.794	< 2e-16 ***
south	-0.2764158	0.0174754	-15.817	< 2e-16 ***
cd	0.0005363	0.0008855	0.606	0.545

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 36122 on 30316 degrees of freedom
Residual deviance: 35002 on 30313 degrees of freedom
AIC: 35010

Number of Fisher Scoring iterations: 4

```
m1$coefficient/1.6
(Intercept)      income      south      cd
0.1122219125 0.1909862822 -0.2862603662 0.0005344688
> m2$coefficient
(Intercept)      income      south      cd
0.1189543271 0.1835890776 -0.2764158031 0.0005362871
```

It is seen that the coefficients of logistic regression divided by 1.6 is almost the same as the coefficients of probit regression.

exercise 2c

The following code is used to preprocess the data.

```
vote1986<-read.table("1986.txt")
names(vote1986)[1]="district"
names(vote1986)[2]="location"
names(vote1986)[3]="incumbency86"
names(vote1986)[4]="dem86"
names(vote1986)[5]="rep86"
vote1988<-read.table("1988.txt")
names(vote1988)[1]="district"
names(vote1988)[2]="location"
names(vote1988)[3]="incumbency88"
names(vote1988)[4]="dem88"
```



```
names(vote1988)[5]="rep88"
data<-merge(vote1986, vote1988,by=c("district","location"))
new.data=subset(data,incumbency88!=-9&incumbency86!=-9&dem88!=-9&dem86!=-9)
attach(new.data)
```

Now we fit the model and summarize the linear model and t-model:

```
> m1<-lm(dem88~dem86+rep86+as.factor(incumbency88))
> summary(m1)
```

Call:

```
lm(formula = dem88 ~ dem86 + rep86 + as.factor(incumbency88))
```

Residuals:

Min	1Q	Median	3Q	Max
-84542	-10982	202	12455	67635

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.924e+04	4.733e+03	4.065	5.78e-05	***
dem86	8.014e-01	4.119e-02	19.455	< 2e-16	***
rep86	4.854e-02	3.888e-02	1.249	0.212550	
as.factor(incumbency88)0	1.554e+04	4.570e+03	3.401	0.000738	***
as.factor(incumbency88)1	3.394e+04	3.776e+03	8.989	< 2e-16	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 20760 on 407 degrees of freedom

Multiple R-squared: 0.7718, Adjusted R-squared: 0.7695

F-statistic: 344.1 on 4 and 407 DF, p-value: < 2.2e-16

```
> library(hett)
> m2<-tlm(dem88~dem86+rep86+as.factor(incumbency88))
> summary(m2)
```

Location model :

Call:

```
tlm(lform = dem88 ~ dem86 + rep86 + as.factor(incumbency88))
```

Residuals:

Min	1Q	Median	3Q	Max
-87422.2	-11232.8	-249.4	12071.1	68164.8

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.877e+04	4.050e+03	4.636	4.79e-06	***
dem86	8.002e-01	3.525e-02	22.701	< 2e-16	***
rep86	7.041e-02	3.327e-02	2.116	0.0349	*
as.factor(incumbency88)0	1.707e+04	3.910e+03	4.366	1.60e-05	***
as.factor(incumbency88)1	3.290e+04	3.231e+03	10.183	< 2e-16	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Scale parameter(s) as estimated below)

Scale Model :

Call:

```
t1m(lform = dem88 ~ dem86 + rep86 + as.factor(incumbency88))
```

Residuals:

Min	1Q	Median	3Q	Max
-2.0000	-1.6622	-0.5862	1.3134	5.3897

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	19.16455	0.09853	194.5	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be 2)

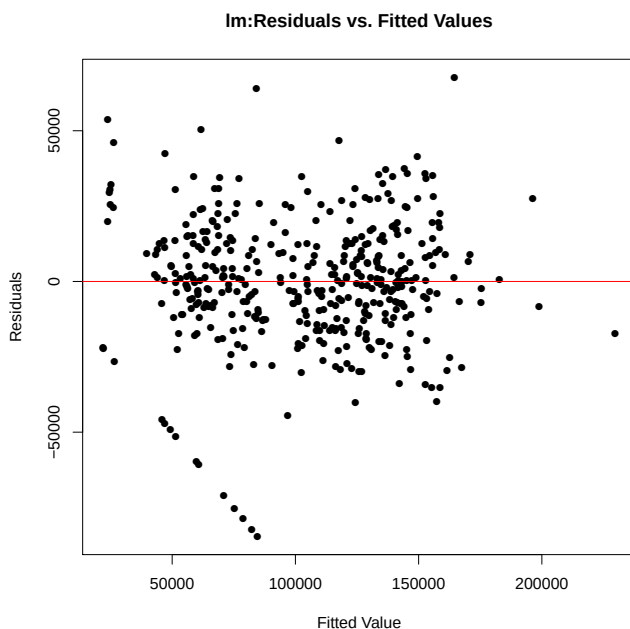
Est. degrees of freedom parameter: 3

Standard error for d.o.f: NA

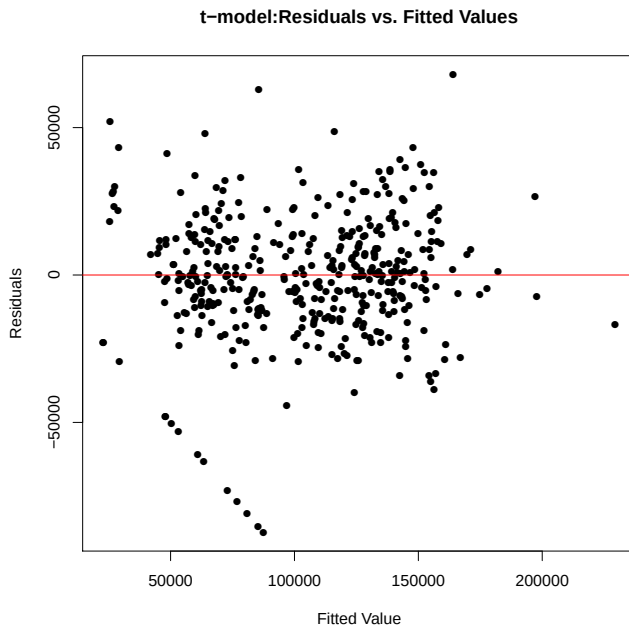
No. of iterations of model : 6 in 0.007

Heteroscedastic t Likelihood : -4662.917

The residual plot for the linear regression is as follows:



The residual plot for the linear regression is as follows:



From the result we see that the coefficients for t-model are all significant, but it is not true for the linear model. The residual plots also show insignificant difference between two methods.