Modeling HW3 Part 2

Kerui Cao

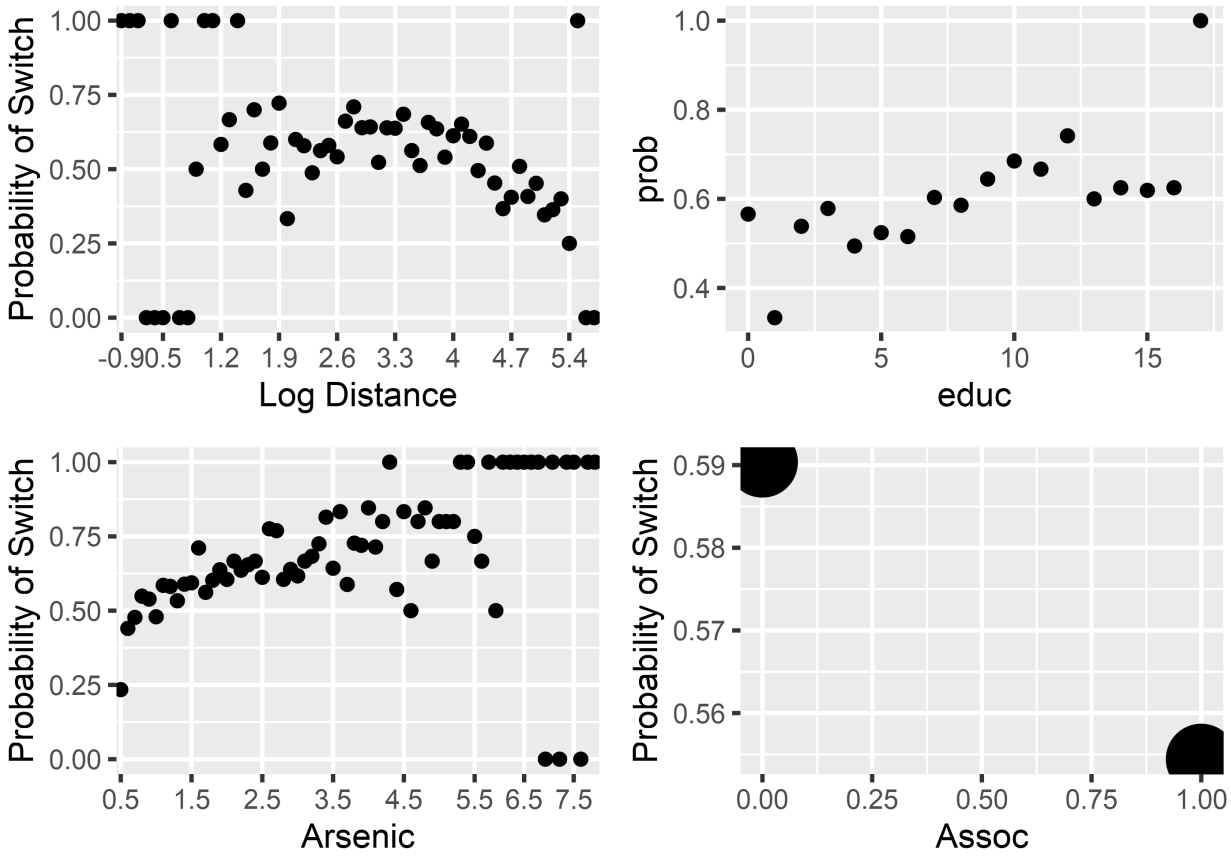
9/29/2019

Graphing logistic regressions:

the well-switching data described in Section 5.4 of the Gelman and Hill are in the folder `arsenic`.

1. Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

Since we know nothing about the real meaning of these variables, we can only try to use information criteria as model selection standard.

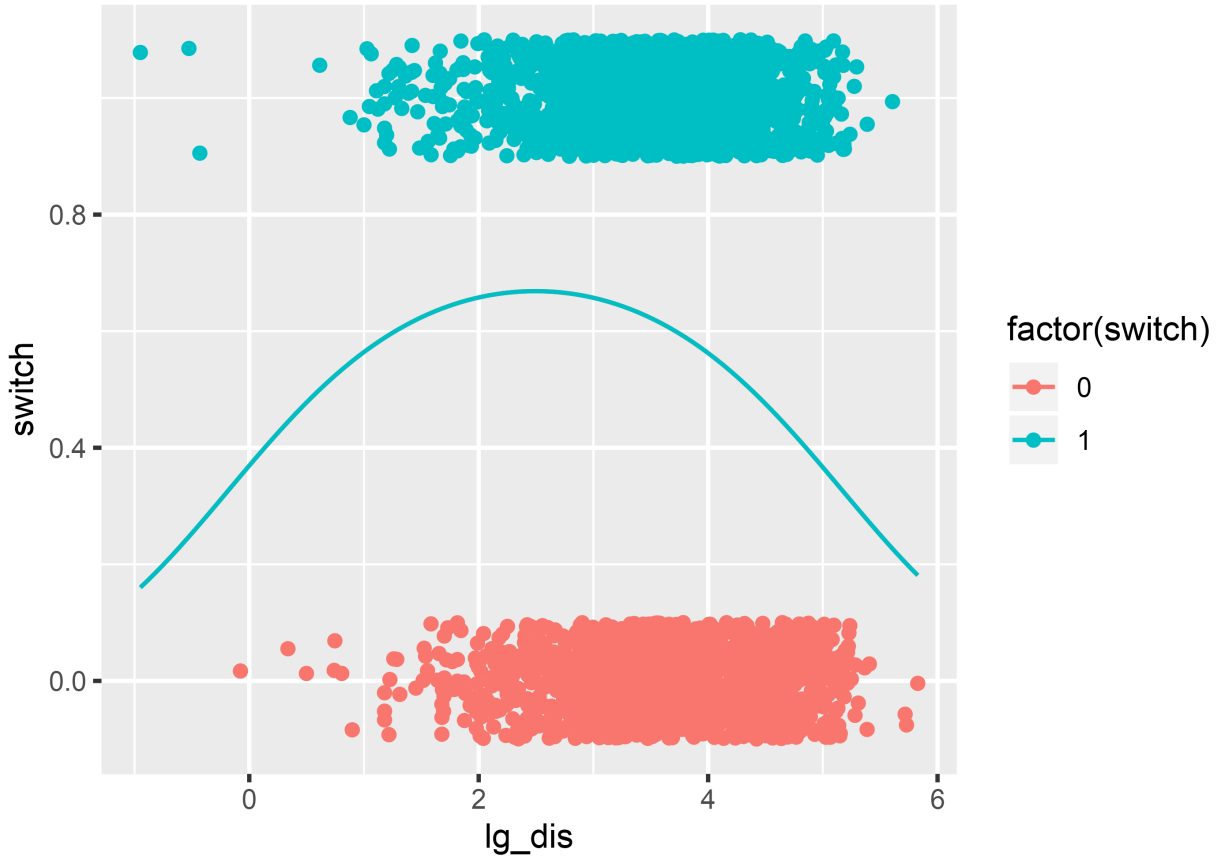


Based on the figures above, we consider the following model:

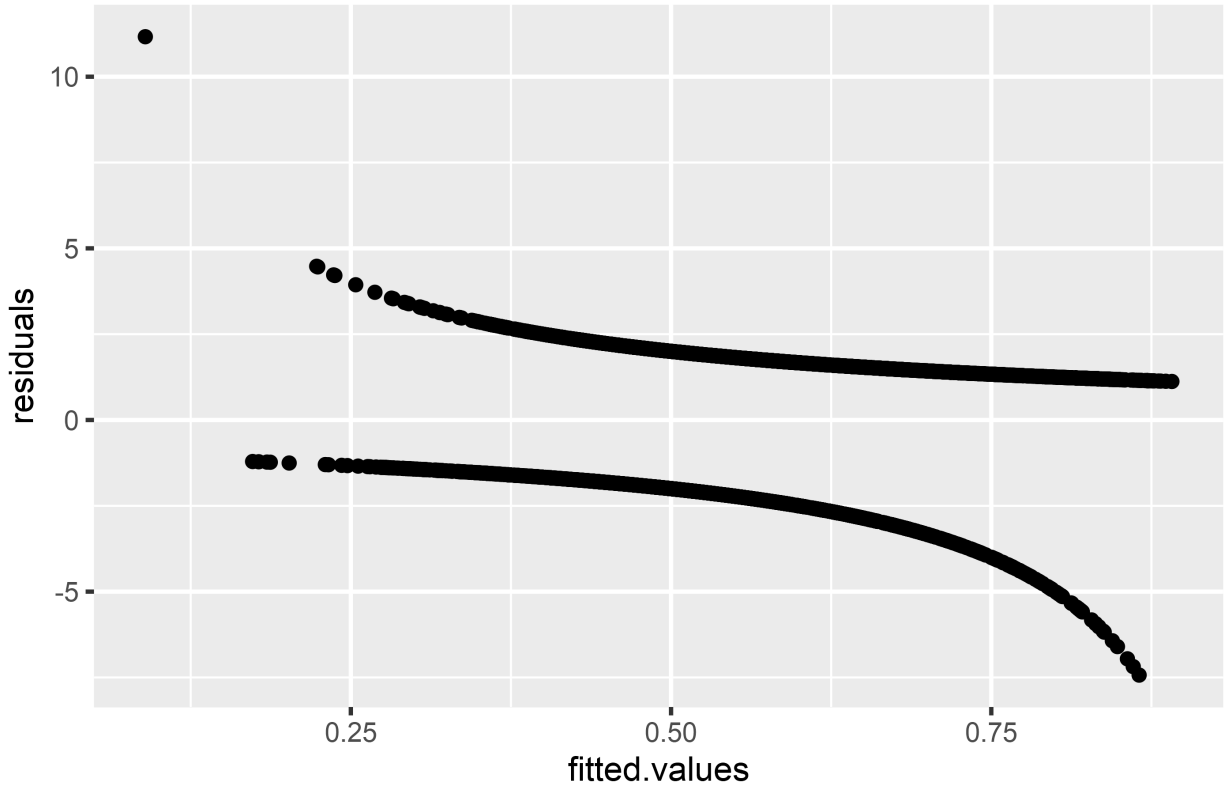
```
##
## Call:
## glm(formula = switch ~ arsenic + I(arsenic^2) + lg_dis + I(lg_dis^2) +
##     educ + assoc, family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0028  -1.1823   0.7349   1.0519   2.1967
```

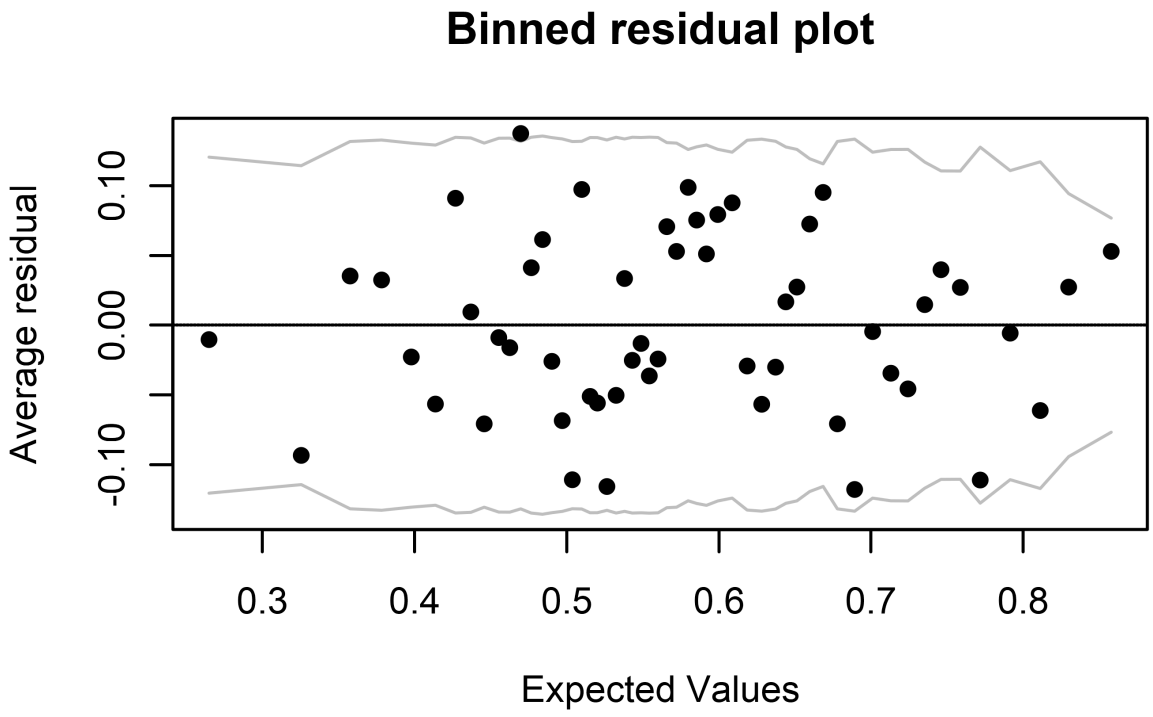
```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.80509    0.42914  -4.206 2.60e-05 ***
## arsenic       0.87840    0.10147   8.656  < 2e-16 ***
## I(arsenic^2) -0.08579    0.01837  -4.671 3.00e-06 ***
## lg_dis        0.99239    0.24474   4.055 5.02e-05 ***
## I(lg_dis^2)  -0.19897    0.03610  -5.511 3.57e-08 ***
## educ          0.04226    0.00963   4.389 1.14e-05 ***
## assoc        -0.11540    0.07724  -1.494    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3892.4  on 3013  degrees of freedom
## AIC: 3906.4
##
## Number of Fisher Scoring iterations: 4
```

2. Make a graph similar to Figure 5.9 of the Gelman and Hill displaying Pr(switch) as a function of distance to nearest safe well, along with the data.



3. Make a residual plot and binned residual plot as in Figure 5.13.



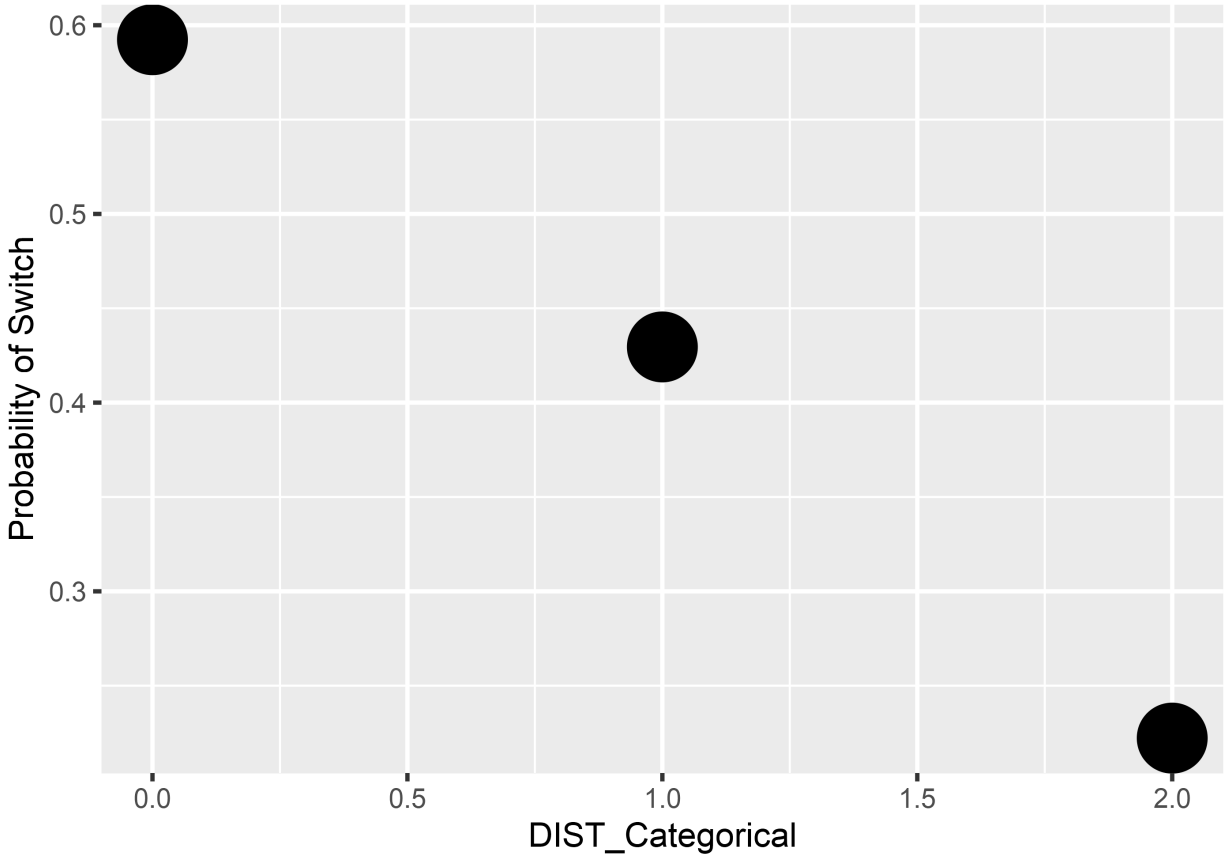


4. Compute the error rate of the fitted model and compare to the error rate of the null model.

Table 1: Error Table

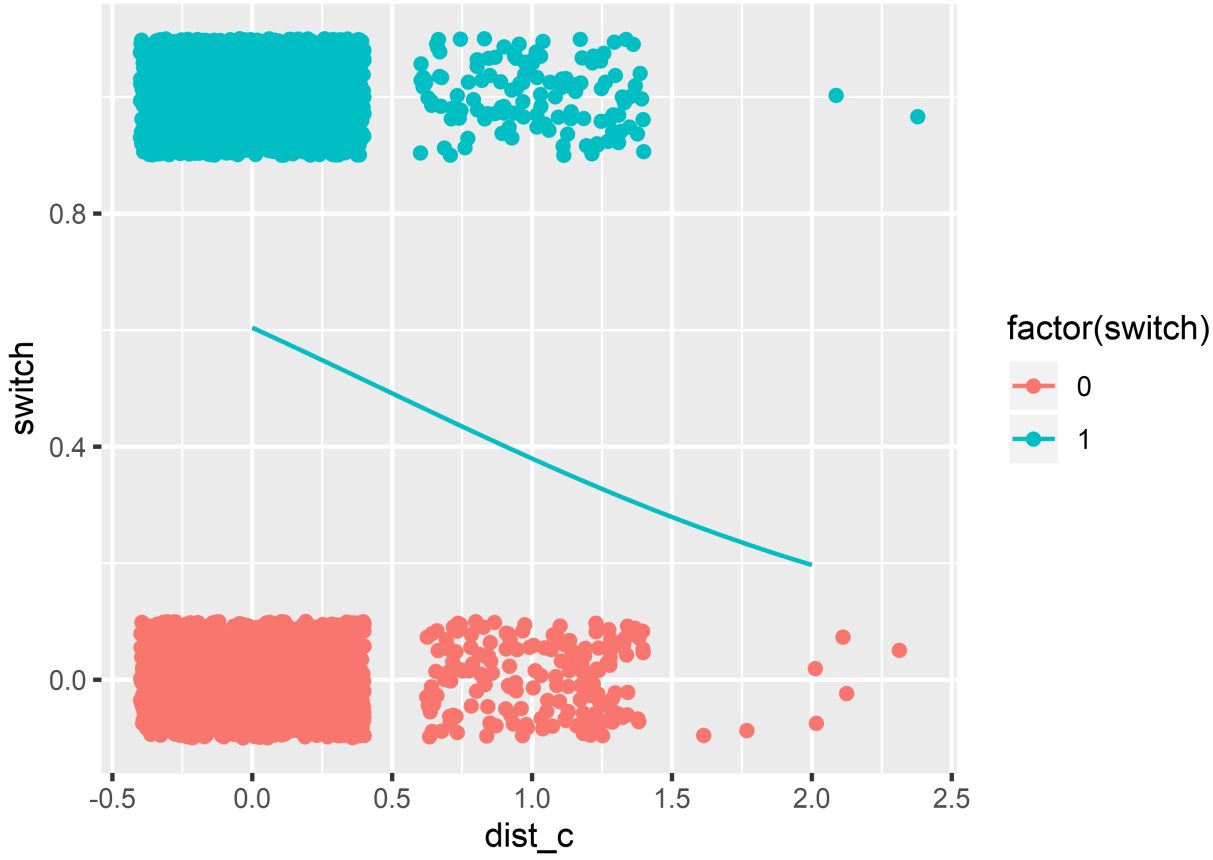
	0	1
No	510	378
Yes	773	1359

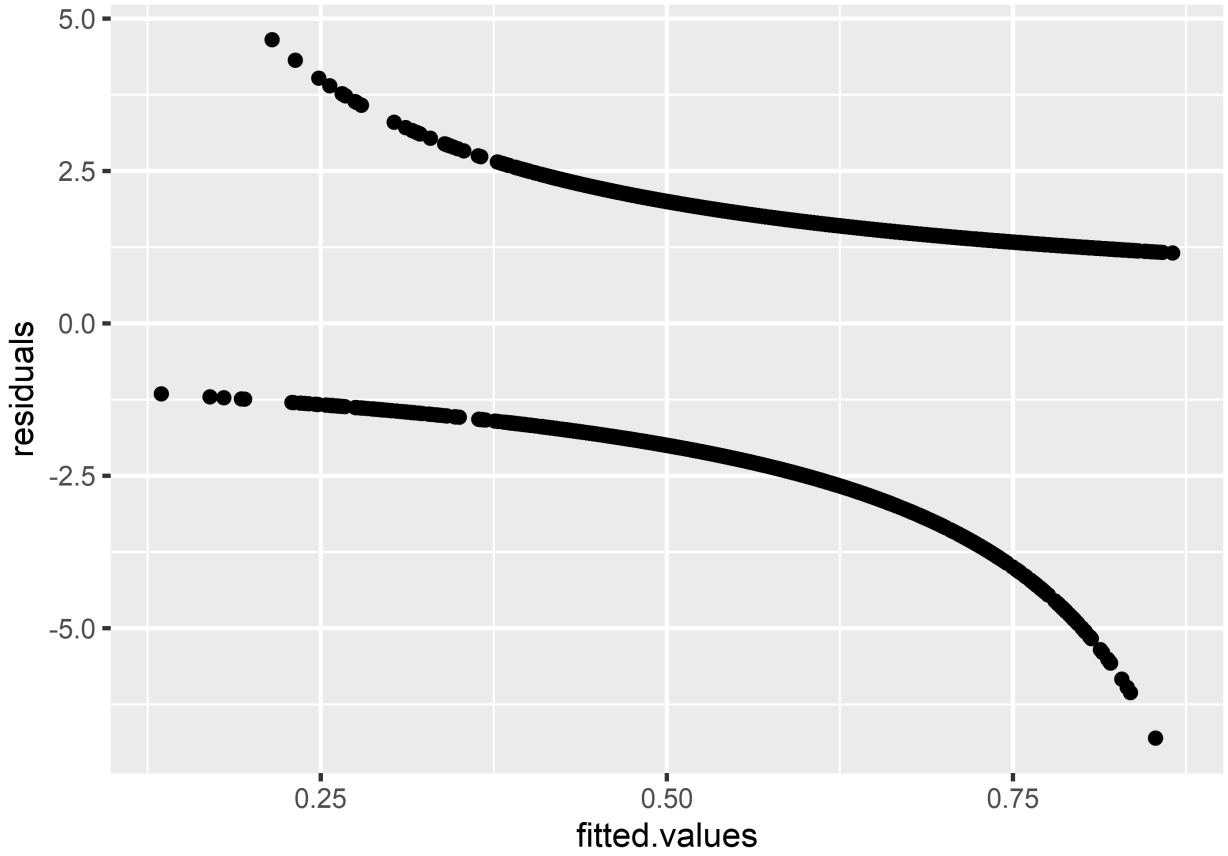
5. Create indicator variables corresponding to `dist < 100`, `100 =< dist < 200`, and `dist > 200`. Fit a logistic regression for Pr(switch) using these indicators. With this new model, repeat the computations and graphs for part (1) of this exercise.




```
##
## Call:
## glm(formula = switch ~ arsenic + I(arsenic^2) + dist_c + educ +
##     assoc, family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9582  -1.1843   0.7388   1.0591   1.7537
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.749198   0.120600  -6.212 5.22e-10 ***
## arsenic       0.783375   0.098736   7.934 2.12e-15 ***
## I(arsenic^2) -0.074200   0.018004  -4.121 3.77e-05 ***
## dist_c       -0.913934   0.123800  -7.382 1.56e-13 ***
```

```
## educ          0.044001   0.009575   4.595 4.32e-06 ***
## assoc        -0.103667   0.076927  -1.348    0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3918.8  on 3014  degrees of freedom
## AIC: 3930.8
##
## Number of Fisher Scoring iterations: 4
```




5.0
2.5
0.0
-2.5
-5.0
residuals
0.25
0.50
0.75
fitted.values

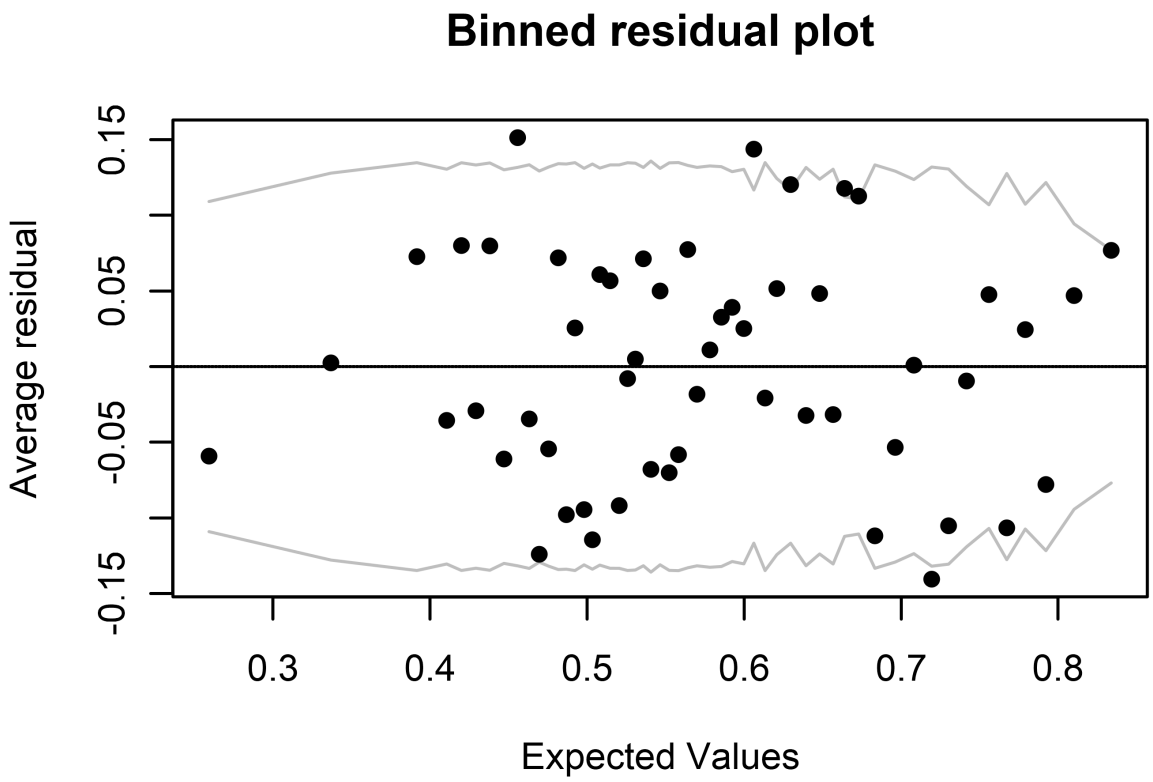


Table 2: Error Table

	0	1
No	506	378
Yes	777	1359

Model building and comparison:

continue with the well-switching data described in the previous exercise.

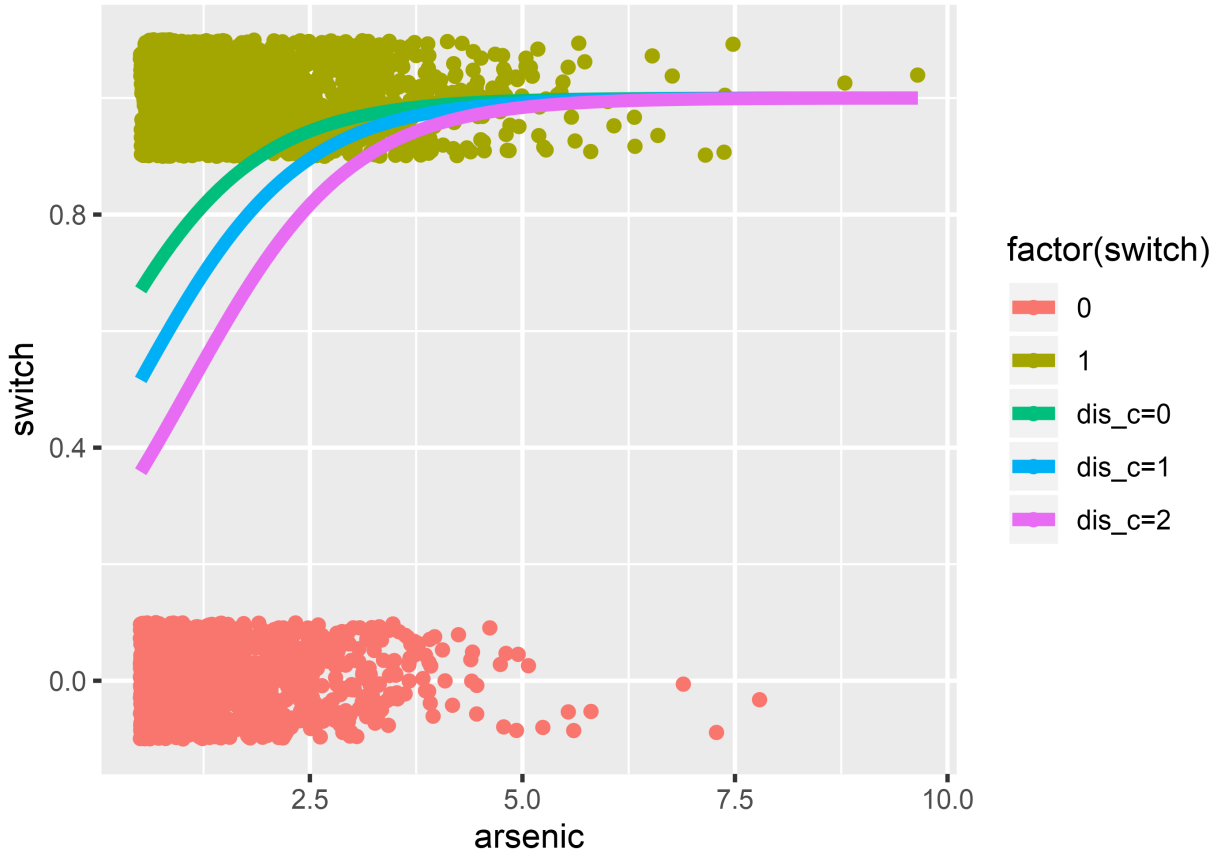
1. Fit a logistic regression for the probability of switching using, as predictors, distance, `log(arsenic)`, and their interaction. Interpret the estimated coefficients and their standard errors.

Because we need to consider the interaction between distance and arsenic, it is not recommended to interact two continuous variables, so we use the categorical version of distance we generated in previous question. Actually here if we simply add dist_c into the model, it is not a proper way adding categorical data, we shouold choose dist_c = 0 as reference, then add two dummy variables representing dist_c =1 and 2, but for simplicity, we just do so.

```
##
## Call:
## glm(formula = switch ~ dist_c + log(arsenic) + dist:log(arsenic),
##     family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.1273  -1.1637   0.7673   1.0465   1.8779
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.173681   0.042815   4.057 4.98e-05 ***
## dist_c            -0.643693   0.147948  -4.351 1.36e-05 ***
## log(arsenic)       1.054962   0.109154   9.665  < 2e-16 ***
## log(arsenic):dist -0.005366   0.001726  -3.109  0.00187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3922.1  on 3016  degrees of freedom
## AIC: 3930.1
##
## Number of Fisher Scoring iterations: 4
```

2. Make graphs as in Figure 5.12 to show the relation between probability of switching, distance, and arsenic level.



3. Following the procedure described in Section 5.7, compute the average predictive differences corresponding to:

i. A comparison of dist = 0 to dist = 100, with arsenic held constant.
ii. A comparison of dist = 100 to dist = 200, with arsenic held constant.

iii. A comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant.
iv. A comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant. Discuss these results.

Comparison of dist = 0 to dist = 100, with arsenic held constant.

```
hi = 1
low = 0
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*hi + b[3]* mo4$model$`log(arsenic)` +
                  b[4]*mo4$model$`log(arsenic)`*hi)
low_p = invlogit(b[1] + b[2]*low + b[3]* mo4$model$`log(arsenic)` +
                   b[4]*mo4$model$`log(arsenic)`*low)
mean(hi_p - low_p)
```

```
## [1] -0.145216
```

Comparison of dist = 100 to dist = 200, with arsenic held constant.

```
hi = 2
low = 1
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*hi + b[3]* mo4$model$`log(arsenic)` +
                  b[4]*mo4$model$`log(arsenic)`*hi)
low_p = invlogit(b[1] + b[2]*low + b[3]* mo4$model$`log(arsenic)` +
                   b[4]*mo4$model$`log(arsenic)`*low)
mean(hi_p - low_p)
```

```
## [1] -0.1399358
```

Comparison of arsenic = 0.5 to arsenic = 1.0, with dist held constant.

```
hi = 1
low = 0.5
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*hi + b[4]*hi*mo4$model$dist_c)
low_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*low + b[4]*low*mo4$model$dist_c)
mean(hi_p - low_p)
```

```
## [1] 0.1073839
```

Comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant.

```
hi = 2
low = 1
b = mo4$coefficients
hi_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*hi + b[4]*hi*mo4$model$dist_c)
low_p = invlogit(b[1] + b[2]*mo4$model$dist_c + b[3]*low + b[4]*low*mo4$model$dist_c)
mean(hi_p - low_p)
```

```
## [1] 0.1402901
```

So according to the result above, we can conclude that:

On average, households which are 100 meter or farther from the nearest safe well are 14.5% less likely to switch, compared to households that are 100 meter or closer from the nearest safe well, at the same arsenic level.

On average, households which are 200 meter farther from the nearest safe well are 14.0% less likely to switch, compared to households that are 200 meter or closer from the nearest safe well, at the same arsenic level.

On average, households whose arsenic level are 1 are 10.7% more likely to switch, compared to households whose arsenic level are 0.5, at the same distant level.

On average, households whose arsenic level are 2 are 14.0% more likely to switch, compared to households whose arsenic level are 1, at the same distant level.