# POL 213, Spring 2024
# Problem Set 3 Solution

TA: Yu-Shiuan (Lily) Huang

## 1 Midterm Exam

Go through your midterm exam and identify questions you missed. Using your textbook and notes, submit updated solutions to any questions you were unable to correctly solve in the exam.

***Please see my comments throughout your assignment on Canvas.***

## 2 Practice with Logistic Regression

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure. In a project in the Araihazar region, researchers measured all the wells people used and labeled them as "safe" or "unsafe" according to whether their arsenic levels fell below or above 0.5 hundreds of micrograms per liter. People with unsafe wells were encouraged to swith to nearby private or community wells or construct new private ones. A few years later, the researchers returned to find out who had switched wells.

Download the "`wells.dat`" data. The data set consists $n = 3020$ households encouraged to switch with the following variables:

- `switch` = 1 if household i switched to a new well; 0 otherwise (Y)
- `arsenic` = contaminant level in i's well (X1)
- `distance` = meters to closest known safe well (X2)
- `assoc` = whether any members of the household are active in community organizations (X3)
- `educ` = education level of the head of household (X4)

Use these data to answer the following questions.

(a) Begin with a model in which the probability of a `switch` is a function of `arsenic` and `distance`. Write down the log likelihood function for your model by adapting the Chilean example we covered in class. It will be easier to rescale the distance in 100-meter units:

```
wells$dist100 <- wells$dist/100
```

$\pi_i \equiv Pr(Y_i = 1)$, where $Y$ is the whether a household $i$ switches to a new well or not.

$\pi_i = \Lambda(\alpha + \beta_1 \texttt{arsenic}_i + \beta_2 \texttt{dist100}_i) = \frac{1}{1+exp(-(\alpha+\beta_1 \texttt{arsenic}_i+\beta_2 \texttt{dist100}_i))}$

$Pr(Y|\alpha, \beta) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

$= \prod_{i=1}^{N} (\frac{1}{1+exp(-(\alpha+\beta_1 \texttt{arsenic}_i+\beta_2 \texttt{dist100}_i))})_i^y (1 - \frac{1}{1+exp(-(\alpha+\beta_1 \texttt{arsenic}_i+\beta_2 \texttt{dist100}_i))})^{1-y_i}$

$= \prod_{i=1}^{N} (1+exp(-(\alpha+\beta_1 \texttt{arsenic}_i+\beta_2 \texttt{dist100}_i)))^{-y_i} (1+exp(\alpha+\beta_1 \texttt{arsenic}_i+\beta_2 \texttt{dist100}_i))^{y_i-1}$

Since the likelihood function for the logit model $L(\alpha, \beta|Y)$ is proportional to $Pr(Y|\alpha, \beta)$, the log-likelihood function can be written as:

$lnL(\alpha, \beta|Y) = \Sigma_{i=1}^{N} [-y_i ln[1 + exp(-\alpha - \beta_1 \texttt{arsenic}_i - \beta_2 \texttt{dist100}_i)] - (1 - y_i) ln[1 + exp(\alpha + \beta_1 \texttt{arsenic}_i + \beta_2 \texttt{dist100}_i)]]$

(b) Run a logistic regression with just predictors $X1$ and $X2$. Run a second model with $X1$, $X2$, $X3$, and $X4$.

Table 1: Logistic Regression

| | *Dependent variable:* | |
|---|---|---|
| | Switch to a new well | |
| | (1) | (2) |
| arsenic | 0.46*** | 0.47*** |
| | (0.04) | (0.04) |
| dist100 | −0.90*** | −0.90*** |
| | (0.10) | (0.10) |
| assoc | | −0.12 |
| | | (0.08) |
| educ | | 0.04*** |
| | | (0.01) |
| Constant | 0.003 | −0.16 |
| | (0.08) | (0.10) |
| Observations | 3,020 | 3,020 |
| Log Likelihood | −1,965.33 | −1,953.91 |
| Akaike Inf. Crit. | 3,936.67 | 3,917.83 |
| *Note:* | *p<0.05; **p<0.01; ***p<0.001 | |

```
# import data
wells <- read.table("data/wells.dat", header = TRUE)

# rescale variable
wells$dist100 <- wells$dist/100

# logit models
```

```r
logit1 <- glm(switch ~ arsenic + dist100, data = wells,
              family = binomial(link = "logit"))
logit2 <- glm(switch ~ arsenic + dist100 + assoc + educ,
              data = wells, family = binomial(link = "logit"))
library(stargazer)
stargazer(logit1, logit2,
          dep.var.labels = "Switch to a new well",
          star.cutoffs = c(0.05, 0.01, 0.001), digits = 2)
```

(c) Choose one of the models and explain why you prefer it.

I performed two tests to do model selection: likelihood ratio test and Akaike Information Criterion (AIC).

1. The Likelihood-Ratio Test (LRT) is a statistical test used to compare the goodness of fit of two models (full model v.s. nested model) based on the ratio of their likelihoods. In this case, the full model is the one with predictors $X1$, $X2$, $X3$, and $X4$, whereas the nested model (i.e., the null model) is the one with just $X1$ and $X2$. If the null hypothesis (the null model) is supported by the observed data, there should be no statistically significant difference between the two likelihoods of the null model and the full model. This indicates that the full model, the one that includes more additional predictors overall does not improve the goodness-of-fit.

The result of the LRT below shows that including $X3$ and $X4$ does significantly increase the goodness-of-fit. The Chi-square value demonstrates that the difference between the two models is statistically significant at any p-value level.

```
Likelihood ratio test

Model 1: switch ~ arsenic + dist100
Model 2: switch ~ arsenic + dist100 + assoc + educ
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -1965.3
2   5 -1953.9  2 22.842  1.096e-05 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. Akaike Information Criterion (AIC) judges a model by how close its fitted values are to the true expected values. The optimal model is the one that tends to have fitted values closest to the true outcome probabilities. This is the model that minimizes:

$$AIC = -2(lnL_M - P) = Dev_M + 2P$$

where $P$ is the number of parameters in the model (including $\alpha$)

The smallest the AIC value, the better the fit. Since fit will improve with additional predictors, adding $2P$ is a penalty for adding predictors in the models. According to Table 1, the AIC in the model with four predictors is smaller than the AIC in the model with only two predictors. The absolute difference between them is about 18.84, which is a strong evidence showing that the model with four predictors performs better than the one with only two predictors.

Based on the above discussion, the model with all four predictors will be the preferred model.

```
library(lmtest)
lrtest(logit1, logit2)

AIC(logit1) - AIC(logit2) # [1] 18.84229

```

(d) Interpret the results by answering the following questions / doing the following calculations: Are the parameters significantly different from 0? Give an odds ratio interpretation for at least one of the covariates.

As shown in Table 1, all the predictors are significantly different from 0 except for `assoc`, which measures whether any members of the household are active in community organizations. Table 2 presents the odds ratio for the intercept and each coefficient in model 2 (the model with all four predictors). For this question, I will focus on the effect of the contaminant level in a household's well on its decision to switch to a new well. As one-unit increase in the contaminant level in a household's well, the odds that a household switches to a new well is 1.5952 times that of a household not switch to a new well. Another way to interpret the effect can also be: a one-unit increase in the contaminant level in a household's well is associated with 59.52% ($= (1.5952 - 1) * 100\%$) increase in the odds of a household switches to a new well versus not switch to a new well. This indicates that the higher the contaminant level in a household's well, the more likely the household switches to a new well.

Table 2: Odds Ratio and Confidence Intervals

| | odds ratio | lower bound | upper bound |
|---|---|---|---|
| Intercept | 0.8550 | 0.7031 | 1.0390 |
| arsenic | 1.5952 | 1.4720 | 1.7328 |
| dist100 | 0.4082 | 0.3319 | 0.5002 |
| assoc | 0.8831 | 0.7595 | 1.0269 |
| educ | 1.0434 | 1.0240 | 1.0632 |

```
cbind(`odds ratio` = exp(coef(logit2)),
      exp(confint(logit2)))
```

(e) Using your preferred model, plot the distribution of 1s and 0s in your response variable and the fitted values for your estimates. Label the axes appropriately.

Figure 7 shows the predicted probabilities with 95% confidence intervals under different contaminant levels while setting all the other variables at their mean or median values.

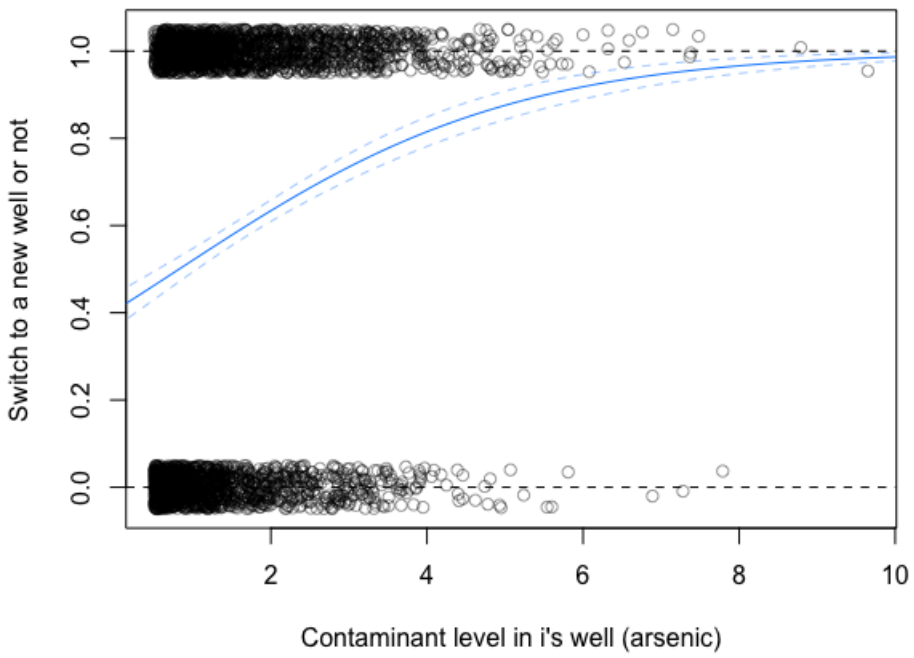


Figure 7: Fitted Probability Line under Different Contaminant Level

```
summary(wells)
x <- seq(0, 10, 0.01)
pdf <- data.frame(arsenic = x,
                  dist100 = rep(0.48332, length(x)),
                  assoc = rep(0, length(x)),
                  educ = rep(4.828, length(x)))
predict.p <- predict(logit2, newdata = pdf,
                     type = "response", se.fit = TRUE)

library(scales)
plot(jitter(switch, .25) ~ arsenic, data = wells,
     xlab="Contaminant level in i's well (arsenic)", # continuous explanatory
    variable
     ylab="Switch to a new well or not", # dichotomous outcome variable
     col = alpha("black", 0.4))
abline(h = 1, lty = 2, col = "black")
abline(h = 0, lty = 2, col = "black")
lines(pdf$arsenic, predict.p$fit, lty = 1, col = "dodgerblue")
lines(pdf$arsenic, predict.p$fit + (1.96*predict.p$se.fit),
      lty = 2, col = alpha("dodgerblue", 0.4))
lines(pdf$arsenic, predict.p$fit - (1.96*predict.p$se.fit),
      lty = 2, col = alpha("dodgerblue", 0.4))

```

# 3 Derivation of Logit Model

Fox textbook, exercise 14.6. (Same across the 2nd and 3rd editions.)

Show that the maximized likelihood for the fitted logit model can be written as

$$log_e L = \Sigma_{i=1}^{n}[y_i log_e P_i + (1 - y_i) log_e(1 - P_i)]$$

where

$$P_i = \frac{1}{1+exp[-(A+B_1X_{i1}+...+B_kX_{ik})]}$$

is the fitted probability that $Y_i = 1$. [*Hint:* Use $p(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$]

In logistic regression, the response variable is modeled with a binomial distribution or its special case Bernoulli distribution. For a response variable $Y_i$ that takes on two values 0, 1 with probability $P_i$ and $(1 - P_i)$, we can summarize the probability distribution of $Y_i$ simply as:

$$p(y_i) \equiv Pr(Y_i = y_i) = P_i^{y_i}(1 - P_i)^{1-y_i}$$

For a sample of n independent observations, the joint probability for the data can be written as:

$$p(y_1, y_2, ..., y_n) = Pr(Y|A, B_k) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i} = L(A, B_k|Y)$$

We can find the parameters that are most likely to have produced the results you observed by maximizing the above joint probability function. Thinking of this equation as a function of the parameters while treating the data $(y_1, y_2, ..., y_n)$ as fixed gives us the likelihood function for the logit model $L(A, B_k|Y)$, which is proportional to $Pr(Y|A, B_k)$. For an easier computation, let's take a log of it to further transform it to a log-likelihood function (the logarithms are good to reduce the numerical value of big numbers; for example, they enable us to simplify calculations by reducing the exponent of a number, and also allow us to transform products into sums). The log-likelihood function can be written as:

$$log_e L(A, B_k|Y) = log_e[\prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}]$$

$$= log_e[P_1^{y_1}(1 - P_1)^{1-y_1} \cdot P_2^{y_2}(1 - P_2)^{1-y_2} \cdot ... \cdot P_n^{y_n}(1 - P_n)^{1-y_n}]$$

$$= log_e(P_1^{y_1}(1 - P_1)^{1-y_1}) + log_e(P_2^{y_2}(1 - P_2)^{1-y_2}) + ... + log_e(P_n^{y_n}(1 - P_n)^{1-y_n})$$

$$= \Sigma_{i=1}^{n}[log_e(P_i^{y_i}(1 - P_i)^{1-y_i})]$$

$$= \Sigma_{i=1}^{n}[log_e P_i^{y_i} + log_e(1 - P_i)^{1-y_i}]$$

$$= \Sigma_{i=1}^{n}[y_i log_e P_i + (1 - y_i) log_e(1 - P_i)]$$