HW2

Jinhong Du - 12243476

2020/01/03

Contents

4.14. In a GLM that uses a noncanonical link function, explain why it need not be true that $\sum_i \hat{\mu}_i = \sum_i y_i$. Hence, the residuals need not have a mean of 0. Explain why a canonical link GLM needs an intercept term in order to ensure that this happens.

For noncanonical links, $\hat{\mu}_i = g^{-1}(x_i^\top \hat{\beta})$ where $\hat{\beta}$ is solved from the score equations $X^\top D V^{-1}(y-\mu) = 0$. If we want $\sum_i \hat{\mu}_i = \sum_i y_i$, then at least one row of $X^\top D V^{-1}$ should be $(1, \dots, 1)$. However, this is not hold in general.

For canonical links, the score equations are given by $\sum_{i=1}^n (y_i - \mu_i) x_{ij} = 0$ for $j = 1, \dots, p$, or equivalently, $X^\top (y - \mu) = 0$. To ensure $\sum_{i=1}^n (y_i - \mu_i) = 0$ so that $\sum_i \hat{\mu}_i = \sum_i y_i$, we need $x_{1j_0} = \cdots = x_{nj_0}$ for some j_0, which means that the j_0th covariate is the intercept term.

4.17. Suppose x is uniformly distributed between 0 and 100, and y is binary with $\log[\frac{\pi_i}{(1-\pi_i)}] = -2.0 + 0.04 x_i$. Randomly generate $n = 25$ independent observations from this model. Fit the model, and find corr$(\boldsymbol{y} - \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}})$. Do the same for $n = 100, n = 1000$, and $n = 10,000$, and summarize how the correlation seems to depend on n.

As we can see from the following results, as n increases, the correlation between $\boldsymbol{y} - \hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}$ shrinks to zero.

```
for(n in c(25,100,1e3,1e4)){
    x <- runif(n, 0, 100)
    g_eta <- exp(-2+0.04*x)
    pi <- g_eta/(1+g_eta)
    y <- as.numeric(runif(n)<pi)

    fit <- glm(y ~ x, family=binomial(link = "logit"))
    mu_hat <- fit$fitted.values

    cat(n,'\t',cor(y - mu_hat, mu_hat),'\n')
}
```

```
## 25       0.002795668
## 100      -0.002179554
## 1000     -0.002398261
## 10000    0.0001431459
```

4.20. For n independent observations from a Poisson distribution with parameter μ, show that Fisher scoring gives $\mu^{(t+1)} = \overline{y}$ for all $t > 0$. By contrast, what happens with the Newton-Raphson method?

The density for y is given by $\frac{\mu^y}{y!} e^{-\mu} = e^{y \ln \mu - \mu} \frac{1}{y!}$, and the log likelihood is given by $L(\mu) = (\sum_{i=1}^n y_i) \ln \mu - n\mu - \sum_{i=1}^n \ln(y_i!)$. Then

$$L'(\mu) = \frac{1}{\mu} \sum_{i=1}^n y_i - n, \qquad L''(\mu) = -\frac{1}{\mu^2} \sum_{i=1}^n y_i, \qquad \mathbb{E}[L''(\mu)] = -\frac{n}{\mu}.$$

So

The Fisher scoring gives

$$\mu^{(t+1)} = \mu^{(t)} - \left(J^{(t)}\right)^{-1} L'(\mu^{(t)}) = \mu^{(t)} + \frac{\mu^{(t)}}{n} \left(\frac{1}{\mu^{(t)}} \sum_{i=1}^n y_i - n \right) = \frac{1}{n} \sum_{i=1}^n y_i = \overline{y}.$$

The Newton-Raphson method gives

$$\mu^{(t+1)} = \mu^{(t)} - \left(L''(\mu^{(t)})\right)^{-1} L'(\mu^{(t)}) = \mu^{(t)} + \frac{\mu^{(t)2}}{\sum_{i=1}^n y_i} \left(\frac{1}{\mu^{(t)}} \sum_{i=1}^n y_i - n \right) = 2\mu^{(t)} - \frac{\mu^{(t)2}}{\overline{y}}.$$

4.22. For noncanonical link functions in a GLM, show that the observed information matrix may depend on the data and hence differs from the expected information matrix. Thus, the Newton-Raphson method and Fisher scoring may provide different standard errors.

For noncanonical link functions,

$$\frac{\partial L}{\partial \beta} = X^\top D V^{-1}(y-\mu)$$
$$\frac{\partial^2 L}{\partial \beta_k \partial \beta_j} = \frac{\partial}{\partial \beta_k} \sum_{i=1}^{n} \frac{y_i - \mu_i}{\nu_i} \frac{x_{ij}}{g'(\mu_i)}$$

Since $\frac{y_i-\mu_i}{\nu_i}\frac{x_{ij}}{g'(\mu_i)}$ depends on x_{ij} and y_i, its derivative with respect to β_k also depends on the data. Therefore,

$$\frac{\partial^2 L}{\partial \beta_k \partial \beta_j} \neq \mathbb{E}\left[\frac{\partial^2 L}{\partial \beta_k \partial \beta_j}\right],$$

and the Newton-Raphson method and Fisher scoring may provide different standard errors.

4.27. Section 4.7.2 mentioned that using a gamma GLM with log-link function gives similar results to applying a normal linear model to $\log(\boldsymbol{y})$.

a. Use the delta method to show that when y has standard deviation σ proportional to μ (as does the gamma GLM), $\log(y)$ has approximately constant variance for small σ.

Suppose that σ is proportional to μ, i.e. $\sigma = k\mu$ for some constant k. Since $\log y \approx \log \mu + \frac{1}{\mu}(y-\mu)$, we have $\mathrm{Var}[\log(y)] \approx \frac{\mathrm{Var}(y)}{\mu^2} = \frac{\sigma^2}{\mu^2} = k^2$ is a constant.

b. The gamma GLM with log link refers to $\log[\mathbb{E}(y_i)]$, whereas the ordinary linear model for the transformed response refers to $\mathbb{E}[\log(y_i)]$. Show that if $\log(y_i) \sim N(\mu_i, \sigma^2)$, then $\log[\mathbb{E}(y_i)] = \mathbb{E}[\log(y_i)] + \frac{\sigma^2}{2}$.

Let $z_i = \log y_i$. By the change of variable formula, we have

$$f_{y_i}(y) = f_{z_i}(\log y)\frac{1}{y} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log y - \mu_i)^2}\frac{1}{y}.$$

Then

$$\begin{aligned}
\mathbb{E}(y_i) &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log y-\mu_i)^2}\mathrm{d}y \\
&\overset{t=\log y}{=\!=\!=} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu_i)^2} e^t \mathrm{d}t \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}t^2 + \left(\frac{\mu_i}{\sigma^2}+1\right)t - \frac{1}{2\sigma^2}\mu_i^2}\mathrm{d}t \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(t-\mu_i-\sigma^2\right)^2 + \frac{1}{2\sigma^2}(\sigma_i^4 + 2\mu_i\sigma_i^2)}\mathrm{d}t \\
&= e^{\mu_i+\frac{\sigma^2}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(t-\mu_i-\sigma^2\right)^2}\mathrm{d}t \\
&= e^{\mu_i+\frac{\sigma^2}{2}}.
\end{aligned}$$

Therefore,

$$\log[\mathbb{E}(y_i)] = \mu_i + \frac{\sigma^2}{2} = \mathbb{E}[\log(y_i)] + \frac{\sigma^2}{2}.$$

c. For the lognormal fitted mean L_i for the linear model for $\log(y_i)$, explain why $\exp(L_i)$ is the fitted median for the conditional distribution of y_i. Explain why the fitted median would often be more relevant than the fitted mean of that distribution.

Since for the normal vairable, its fitted mean is also the fitted median, we have L_i is also the lognormal fitted meidan for the linear model for $\log(y_i)$. Also, quantiles are preserved under monotonic transformations. So $\exp(L_i)$ is the fitted mediean for the conditional distribution of y_i. The median would often be more relevant because the lognormal variable y_i has a very skewed distribution.

5.6. Explain how expression (5.6) for $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ in logistic regression suggests that the standard errors of $\{\beta_j\}$ tend to be smaller as you obtain more data. Answer this for (a) grouped data with $\{n_i\}$ increasing, (b) ungrouped data with N increasing.

$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})^{-1}$ where $\hat{\boldsymbol{W}} = \text{diag}(n_1\hat{\pi}_1(1-\hat{\pi}_1), \cdots, n_N\hat{\pi}_N(1-\hat{\pi}_N))$. The (i,j)-th entry of $\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X}$ is $\sum_{k=1}^N n_k\hat{\pi}_k(1-\hat{\pi}_k)X_{ik}X_{kj}$.

(a) Notice that $\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X}$ is semi-positive definite, and the i-th diagonal entry of it inceases as n_i increases. For grouped data, if $n = \sum_{i=1}^N n_i$ increases and p is fixed, then the standard errors of $\{\beta_j\}$ tend to be smaller.

(b) For ungrouped data, if N increases and p is fixed, the terms in the summation $\sum_{j=1}^N n_j\hat{\pi}_j(1-\hat{\pi}_j)X_{ji}^2$ increases. Since every term is non-negative, the standard errors of $\{\beta_j\}$ tend to be smaller.

5.7. Assuming the model $\text{logit}[\mathbb{P}(y_i = 1)] = \beta x_i$, you take all n observations at x_0. Find $\hat{\beta}$ and the large-sample $\text{var}(\hat{\beta})$. For the Wald test, explain why the chisquared noncentrality is $\frac{\beta^2}{\text{var}(\hat{\beta})}$, and evaluate it as $\beta \to \infty$. Explain how this illustrates that the Wald test in logistic regression has poor behavior when the effect is strong.

For grouped data at x_0, $\pi = \mathbb{P}(y_i = 1)$. Then $\text{logit}(\pi) = \beta x_0$ and $\beta = \frac{\text{logit}(\pi)}{x_0}$. Since $\hat{\pi} = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}$, we have $\hat{\beta} = \frac{\text{logit}(\overline{y})}{x_0}$. $\text{var}(\hat{\beta}) \approx (\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})^{-1} = \left(x_0^2 \frac{\partial \mu}{\partial \eta}\frac{1}{\nu}\right)^{-1} = \left(nx_0^2\hat{\pi}(1-\hat{\pi})\right)^{-1} = (n\hat{\pi}(1-\hat{\pi})x_0^2)^{-1}$.

Since the approximate distribution of $\hat{\beta}$ is $N(\beta, \text{var}(\hat{\beta}))$, so by definition, the chi-square noncentrality parameter is $\frac{\beta}{\text{var}(\hat{\beta})}$.

$\lim\limits_{\beta\to\infty} \frac{\beta^2}{\text{var}\hat{\beta}} \approx \lim\limits_{\pi\to 1} n\hat{\pi}(1-\hat{\pi})\text{logit}(\hat{\pi})^2 = \lim\limits_{\pi\to 1} n\hat{\pi}(1-\hat{\pi})\text{logit}[\log^2(\hat{\pi}) - 2\log(\hat{\pi})\log(1+\hat{\pi}) + \log^2(1-\hat{\pi})] = 0$ since

$$\lim_{x\to 1}(1-x)\log^2(1-x) = \lim_{x\to 1}\frac{\log^2(1-x)}{\frac{1}{1-x}} = \lim_{x\to 1}\frac{-\frac{2\log(1-x)}{1-x}}{-\frac{1}{(1-x)^2}} = \lim_{x\to 1}\frac{2\log(1-x)}{\frac{1}{1-x}} = \lim_{x\to 1}\frac{-\frac{2}{1-x}}{-\frac{1}{(1-x)^2}} = \lim_{x\to 1} 2(1-x) = 0.$$

As the true effect in a binary regression model increases, for a given sample size the information decreases so quickly that the standard error grows faster than the effect. So the Wald test has poor behavior.

5.11. Construct the log-likelihood function for the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$ with independent binomial proportions of y_1 successes in n_1 trials at $x_1 = 0$ and y_2 successes in n_2 trials at $x_2 = 1$. Derive the likelihood equations, and show that $\hat{\beta}_1$ is the sample log odds ratio.

$$\begin{aligned}
L(\boldsymbol{\beta}) &= \sum_{i=1}^2 y_i\log(\pi_i) + (n_i - y_i)\log(1-\pi_i) \\
\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^2 \left[\frac{y_i}{\pi_i} - \frac{n_i - y_i}{1-\pi_i}\right]\frac{\partial \pi_i}{\partial \eta_j}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \sum_{i=1}^2 \left[\frac{y_i}{\pi_i} - \frac{n_i - y_i}{1-\pi_i}\right]\pi_i(1-\pi_i)\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^2 \left[(1-\pi_i)y_i - \pi_i(n_i - y_i)\right]\frac{\partial \eta_i}{\partial \beta_j}
\end{aligned}$$

for $j = 0, 1$. So the likelihood equations are given by

$$\begin{cases} \sum_{i=1}^{2} \left[(1-\pi_i)y_i - \pi_i(n_i - y_i)\right] = 0 \\ \sum_{i=1}^{2} \left[(1-\pi_i)y_i - \pi_i(n_i - y_i)\right] x_i = 0 \end{cases}$$

i.e.,

$$\begin{cases} (1-\pi_1)y_1 - \pi_1(n_1 - y_1) = 0 \\ (1-\pi_2)y_2 - \pi_2(n_2 - y_2) = 0 \end{cases}$$

which yields the MLE of π_1 and π_2, $\hat{\pi}_1 = \frac{y_1}{n_1}$ and $\hat{\pi}_2 = \frac{y_2}{n_2}$, respectively. And since there is a one-to-one mapping between π_0, π_1 and β_0, β_1 from $\text{logit}(\pi_0) = \beta_0$ and $\text{logit}(\pi_1) = \beta_0 + \beta_1$, we have $\hat{\beta}_1 = \text{logit}(\hat{\pi}_1) - \text{logit}(\hat{\pi}_0) = \log\left(\frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_0/(1-\hat{\pi}_0)}\right)$. So $\hat{\beta}_1$ is the sample log odds ratio.

5.17. Use the following toy data to illustrate comments in Section 5.5 about grouped versus ungrouped binary data in the effect on the deviance:

x	Number of trials	Number of successes
0	4	1
1	4	2
2	4	4

Denote by M_0 the null model $\text{logit}(\pi_i) = \beta_0$ and by M_1 the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$.

a. Create a data file in two ways, entering the data as (i) ungrouped data: $n_i = 1$, $i = 1, \ldots, 12$, (ii) grouped data: $\tilde{n}_i = 4$, $i = 1, 2, 3$. Fit M_0 and M_1 for each data file. Show that the deviances for M_0 and M_1 differ for the two forms of data entry. Why is this?

From the below results, we can see that the deviances are different.

```
ungrouped_y <- c(0,0,0,1,0,0,1,1,1,1,1,1)
ungrouped_x <- rep(c(0:2),each=4)
fit_ungrouped_M0 <- glm(ungrouped_y~1, family=binomial(link = "logit"))
fit_ungrouped_M1 <- glm(ungrouped_y~ungrouped_x, family=binomial(link = "logit"))
grouped_y <- matrix(append(c(1,2,4),c(3,2,0)),ncol=2)
grouped_x <- c(0:2)
fit_grouped_M0 <- glm(grouped_y~1, family=binomial(link = "logit"))
fit_grouped_M1 <- glm(grouped_y~grouped_x, family=binomial(link = "logit"))

cat(summary(fit_ungrouped_M0)$deviance,',',summary(fit_ungrouped_M1)$deviance,',',
    summary(fit_grouped_M0)$deviance,',',summary(fit_grouped_M1)$deviance)
```

```
## 16.30064 , 11.02826 , 6.25678 , 0.9843993
```

The total deviance for a binomial GLM is of the form,

$$\begin{aligned} D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= \sum_{i=1}^{N} D(y_i, n_i\hat{\pi}_i) = 2\sum_{i=1}^{N} \log\left[\frac{f\left(y_i; \frac{y_i}{n_i}\right)}{f(y_i; \hat{\pi}_i)}\right] \\ &= 2\sum_{i=1}^{N} \log\left[\frac{\binom{n_i}{y_i}\left(\frac{y_i}{n_i}\right)^{y_i}\left(1-\frac{y_i}{n_i}\right)^{n_i-y_i}}{\binom{n_i}{y_i}\hat{\pi}_i^{y_i}\left(1-\hat{\pi}_i\right)^{n_i-y_i}}\right] \\ &= 2\sum_{i=1}^{N} y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + 2\sum_{i=1}^{N}(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right). \end{aligned}$$

Noted that the estimated linear coefficients for the same binomial GLM with ungrouped and grouped data are the same because of the same likelihood equation.

For ungrouped data, the deviance for M_k $(k = 0, 1)$ is given by

$$D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{12} y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + 2\sum_{i=1}^{12}(1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right).$$

For grouped data, the deviance for the same model M_k is given by

$$\tilde{D}_+(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{3}\left[\tilde{y}_i \log\left(\frac{\tilde{y}_i}{4\hat{\pi}_i}\right) + (4-\tilde{y}_i)\log\left(\frac{4-\tilde{y}_i}{4-4\hat{\pi}_i}\right)\right].$$

where $\hat{\tilde{\mu}}_i$ and $\hat{\mu}_j$ are the MLEs of $\tilde{\mu}_i$ and μ_j of the same binary GLM, respectively. Here $\frac{1}{n_i}\hat{\tilde{\mu}}_i = \hat{\mu}_{i_1} = \cdots = \hat{\mu}_{i_{n_i}}$ and $\hat{\tilde{\pi}}_i = \hat{\pi}_{i_1} = \cdots = \hat{\pi}_{i_{n_i}}$ for indexes $i_1, \ldots, i_{n_i} \in \{1, \ldots, N\}$ that are in group i. Then

$$\tilde{y}_i \log\left(\frac{\tilde{y}_i}{4\hat{\pi}_i}\right) + (4-\tilde{y}_i)\log\left(\frac{4-\tilde{y}_i}{4-4\hat{\pi}_i}\right) = \sum_{j=1}^{4}\left[y_{i_j}\log\left(\frac{\tilde{y}_i}{4\hat{\pi}_i}\right) + (1-y_{ij})\log\left(\frac{4-\tilde{y}_i}{4-4\hat{\pi}_i}\right)\right].$$

So

$$\tilde{D}_+(\tilde{\boldsymbol{y}}, \hat{\tilde{\boldsymbol{\mu}}}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{3}\sum_{j=1}^{4}\left[y_{i_j}\log\left(\frac{\tilde{y}_i}{4y_{i_j}}\right) + (1-y_{i_j})\log\left(\frac{4-\tilde{y}_i}{4(1-y_{i_j})}\right)\right].$$

Each term in the above summation does not necessarily equal to zero in general (except when $\frac{1}{4}\tilde{y}_i = y_{i_1} = \cdots = y_{i_4}$ for $i = 1, 2, 3$), i.e., the estimated mean $\frac{1}{4}\tilde{y}_i$ may not equal to the estimated mean $y_{i_1}, \ldots y_{i_4}$ for the two saturated models. So the deviances for M_0 and M_1 differ for the two forms of data entry.

b. Show that the difference between the deviances for M_0 and M_1 is the same for each form of data entry. Why is this? (Thus, the data file format does not matter for inference, but it does matter for goodness-of-fit testing.)

From the following results, we can see that the differences for M_0 and M_1 for the two forms of data are the same.

```
cat(summary(fit_ungrouped_M0)$deviance-summary(fit_ungrouped_M1)$deviance,',',
    summary(fit_grouped_M0)$deviance-summary(fit_grouped_M1)$deviance)
```

```
## 5.27238 , 5.27238
```

Let $\hat{\pi}_i$ and $\hat{\pi}_i'$ be the estimated mean for the i-th sample in the ungrouped data of the model M_0 and M_1, respectively. Then since

$$D(y_i, \hat{\mu}_i) - D(y_i, \hat{\mu}_i') = 2y_i \log\left(\frac{\hat{\pi}_i'}{\hat{\pi}_i}\right) + 2(1-y_i)\log\left(\frac{1-\hat{\pi}_i'}{1-\hat{\pi}_i}\right),$$

we have $D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}') = 2\sum_{i=1}^{12}\left[y_i \log\left(\frac{\hat{\pi}_i'}{\hat{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-\hat{\pi}_i'}{1-\hat{\pi}_i}\right)\right]$. Analogously, we have the difference of the total deviances for the grouped data is

$$\tilde{D}_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) - \tilde{D}_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}') = 2\sum_{i=1}^{3}\left[\tilde{y}_i \log\left(\frac{\hat{\pi}_i'}{\hat{\pi}_i}\right) + (4-\tilde{y}_i)\log\left(\frac{1-\hat{\pi}_i'}{1-\hat{\pi}_i}\right)\right] = D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}'),$$

by noticing that $\tilde{y}_i = \sum_{j=1}^{4} y_{i_j}$.

5.28. For the logistic model (5.7) for a 2×2 table, give an example of cell counts corresponding to (a) complete separation and $\hat{\beta}_1 = \infty$, (b) quasi-complete separation and $\hat{\beta}_1 = \infty$, (c) non-existence of $\hat{\beta}_1$.

		x	
		0	1
y	0	1	0
	1	0	1

		x	
		0	1
y	0	1	1
	1	0	1

		x	
		0	1
y	0	0	1
	1	0	1

5.30. In one of the first studies of the link between lung cancer and smoking, Richard Doll and Austin Bradford Hill collected data from 20 hospitals in London, England. Each patient admitted with lung cancer in the preceding year was queried about their smoking behavior. For each of the 709 patients admitted, they recorded the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year. Of the 709 cases having lung cancer, 688 reported being smokers. Of the 709 controls, 650 reported being smokers. Specify a relevant logistic regression model, explain what can be estimated and what cannot (and why), and conduct a statistical analysis.

		lung cancer		
		0	1	
smoker	0	59	21	80
	1	650	688	1338
		709	709	1418

Let x be the indicator of an individual developed lung cancer or not, and $y \sim \text{Bernoulli}(\pi)$ be the indicator of an individual was a smoker or not. The Logistic model is of the form $\text{logit}(\pi) = \beta_1 x + \beta_0$. We can estimated $\mathbb{E}[y|x=0]$ and $\mathbb{E}[y|x=1]$ but not $\mathbb{E}[x|y=0]$ and $\mathbb{E}[x|y=0]$, since in this case-control study where we fix the number of lung cancer and noncancer patients, the population of x is not the true population.

```
y <- matrix(append(c(650,59),c(688,21)),ncol=2)
x <- c(0,1)
fit <- glm(y~x, family=binomial(link = "logit"))
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "logit"))
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.05682    0.05470  -1.039    0.299
## x            1.08983    0.25992   4.193 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 19.878  on 1  degrees of freedom
## Residual deviance:  0.000  on 0  degrees of freedom
## AIC: 16.237
##
## Number of Fisher Scoring iterations: 3
```

Since the p value of β_0 is not significant, we cannot conclude that $\beta_0 \neq 0$. So we'd better remove the intercept term in the model. Now our model becomes $\text{logit}(\pi) = \beta_1 x$. From the below summary table, the p value of β_1 is extremely significant.

```
fit <- glm(y~x-1, family=binomial(link = "logit"))
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x - 1, family = binomial(link = "logit"))
##
## Deviance Residuals:
##      1       2
## -1.039   0.000
##
## Coefficients:
##   Estimate Std. Error z value Pr(>|z|)
## x   1.0330     0.2541   4.065  4.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19.8780  on 2  degrees of freedom
## Residual deviance:  1.0794  on 1  degrees of freedom
## AIC: 15.316
##
## Number of Fisher Scoring iterations: 3
```

5.32. For the horseshoe crab dataset (`Crabs.dat` at the text website) introduced in Section 4.4.3, let $y = 1$ if a female crab has at least one satellite, and let $y = 0$ if a female crab does not have any satellites. Fit a main-effects logistic model using color and weight as explanatory variables. Interpret and show how to conduct inference about the color and weight effects. Next, allow interaction between color and weight in their effects on y, and test whether this model provides a significantly better fit.

```
Crabs <- read.table("Crabs.dat", header=T)
fit <- glm((y>0)~factor(color)+weight, family=binomial(link = "logit"), data=Crabs)
summary(fit)
```

```
##
## Call:
## glm(formula = (y > 0) ~ factor(color) + weight, family = binomial(link = "logit"),
##     data = Crabs)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.1908  -1.0144   0.5101   0.8683   2.0751
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.2572     1.1985  -2.718  0.00657 **
## factor(color)2   0.1448     0.7365   0.197  0.84410
## factor(color)3  -0.1861     0.7750  -0.240  0.81019
## factor(color)4  -1.2694     0.8488  -1.495  0.13479
## weight           1.6928     0.3888   4.354 1.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 188.54  on 168  degrees of freedom
## AIC: 198.54
##
## Number of Fisher Scoring iterations: 4
```

`weight` are significant in the logistic regression model but `color` are not.

Since $\frac{\pi_i}{x_{ij}} = \beta_j \pi_i (1-\pi_i)$, for the quantitative explanatory variable `weight`, the instantaneous rate of change in π_i is about $1.6928\pi_i(1-\pi_i)$. When `weight` increases by 1 unit, the log odds of having satellites increase by around 443%, adjusting for other explanatory variables. Compared with based color, the odds of having satellites increase by 15% when color is 2; the odds of having satellites decrease by 17% and 72% when color is 3 and 4 respectively, when other explanatory variables are held constant.

```
exp(fit$coefficients)
```

```
##    (Intercept) factor(color)2 factor(color)3 factor(color)4         weight
##     0.03849721     1.15584580     0.83016171     0.28099224     5.43482710
```

```
fit2 <- glm((y>0)~factor(color)+weight+factor(color)*weight, family=binomial(link = "logit"),
        data=Crabs)
anova(fit, fit2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: (y > 0) ~ factor(color) + weight
## Model 2: (y > 0) ~ factor(color) + weight + factor(color) * weight
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       168     188.54
## 2       165     181.66  3    6.886  0.07562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see since the p-value is big, we cannot reject the null hypothesis that two model is different, so big model is not better than simple model.

13. In toxicology, the LD_{50} is the dose that causes a 50% mortality rate (lethal dose 50%). Experiments are often carried out at a sequence of dose levels, $x_0, x_1, \ldots$ each dose being twice the preceding dose. The model most commonly used in toxicology is linear in log dose. Suppose that the following results have been obtained in an experiment at various multiples of the baseline dose.

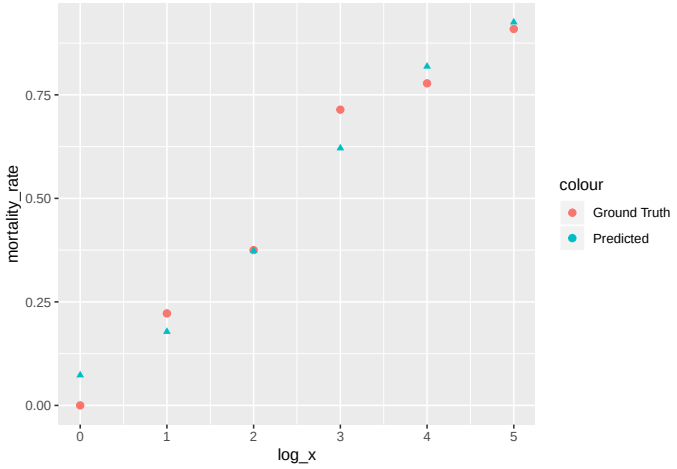
Dose $\log_2(x)$	0	1	2	3	4	5
Mortality y/m	0/7	2/9	3/8	5/7	7/9	10/11

Here y/m is the number of deaths occurring in a sample of m individuals.

1. Plot the data, i.e. the mortality fraction against log dose. Fit the linear logistic model in which the logit of the mortality rate is linear in log dose. Superimpose the fitted probabilities on the plot.

```
death <- c(0, 2, 3, 5, 7, 10)
total <- c(7, 9, 8, 7, 9, 11)
log_x <- c(0:5)
mortality <- matrix(append(death,total-death),ncol=2)

fit <- glm(mortality ~ log_x, family=binomial(link = "logit"))
prob <- predict(fit, data.frame(log_x=c(0:5)), type="response")
require(ggplot2)
df <- data.frame(log_x=c(0:5), mortality_rate=death/total, predicted_probability=prob)
ggplot(df, aes(log_x)) +
    geom_point(aes(y=mortality_rate, colour="Ground Truth"), size=2) +
    geom_point(aes(y=predicted_probability, colour="Predicted"), shape = 17)
```



2. Obtain the estimate of the $\log_2 \mathbf{LD}_{50}$, using the delta method.

Let $y \sim \text{Bernoulli}(\pi)$, $g(x) = \log\left(\frac{x}{1-x}\right)$ and $z = \log_2 x$. The model is given by

$$g(\pi) = \beta_0 + \beta_1 z.$$

Since

$$\pi = \frac{e^{\beta_0+\beta_1 z}}{1+e^{\beta_0+\beta_1 z}} = \frac{1}{2}$$

implies $\beta_0+\beta_1 z = 0$, we have $z = -\frac{\beta_0}{\beta_1} \approx$. Let $f(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\hat{\beta}_0}{\hat{\beta}_1}$, then by delta method, $f(\hat{\beta}_0, \hat{\beta}_1)$ has an approximate

distribution $N\left(f(\beta_0,\beta_1), \mathrm{Var}\left(-\frac{\hat\beta_0}{\hat\beta_1}\right)\right)$ where

$$\begin{aligned}\mathrm{Var}\left(-\frac{\hat\beta_0}{\hat\beta_1}\right) &\approx \nabla f^\top \mathrm{Cov}(\hat\beta_0,\hat\beta_1)\nabla f \\ &= \begin{bmatrix} -\frac{1}{\hat\beta_1} \\ \frac{\hat\beta_0}{\hat\beta_1^2}\end{bmatrix}^\top (\boldsymbol{X}^\top \boldsymbol{V}\boldsymbol{X})^{-1} \begin{bmatrix} -\frac{1}{\hat\beta_1} \\ \frac{\hat\beta_0}{\hat\beta_1^2}\end{bmatrix}\end{aligned}$$

and $\boldsymbol{V} = \mathrm{diag}(n_1\hat\pi_1(1-\hat\pi_1),\ldots,n_N\hat\pi_N(1-\hat\pi_N))$. So the estimated value of $\log_2 \mathrm{LD}_{50}$ is 2.510815 and the approximated 95% confident interval of $\log_2 \mathrm{LD}_{50}$ is $(1.79818, 3.223448)$.

```
beta0 <- fit$coefficients[1]
beta1 <- fit$coefficients[2]
nabla_f <- matrix(c(-1/beta1, beta0/beta1^2), ncol=1)
X <- matrix(cbind(1,log_x), ncol=2)
V <- diag(total*prob*(1-prob))
se <- sqrt(t(nabla_f)%*%solve(t(X)%*%V%*%X)%*%nabla_f)
cat(-beta0/beta1)
```

```
## 2.510815
```

```
cat(-beta0/beta1-se*qnorm(0.975),-beta0/beta1+se*qnorm(0.975))
```

```
## 1.798183 3.223448
```