University of North Carolina at Chapel Hill - Dept. of Biostatistics
Qualifying Exam Solutions (Summer 2016)

Exercise 1.A

Redefining x_i so that $x_i^T = [1, x_{i1}]$ where x_{i1} follows the original definition, we may write the likelihood for this data in the following way:

$$\begin{aligned} L &= \texttt{exp}\left\{\sum_{i=1}^{n} y_i x_i^T \beta - ln(1 + e^{x_i^T \beta})\right\} \\ &= c(\beta)\texttt{exp}\left\{\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} y_i x_{i1}\right\} \end{aligned}$$

By theorem, we may then structure a UMPU test in the following way:

$$\phi(\sum y_i x_{i1}) = \begin{cases} 1, & \text{if } \sum y_i x_{i1} < c_1(\sum y_i) \text{ or } \sum y_i x_{i1} > c_2(\sum y_i). \\ \gamma_i & \text{if } \sum y_i x_{i1} = c_i(\sum y_i) \\ 0, & \text{if otherwise.} \end{cases}$$

where $c(\sum y_i)$ is defined so that $E[\phi(\sum y_i x_{i1})| \sum y_i] = \alpha$ and $E[\sum y_i x_{i1} \phi(\sum y_i x_{i1})| \sum y_i] = \alpha E[\sum y_i x_{i1}| \sum y_i]$.

Defining $c(\sum y_i)$:

We must find forms for the means presented above before we can define c. Define the following:

- Let Y be an n-dimensional random variable of which y is a realization. This means that Y is a vector of 0's and 1's.
- Let $t = \sum y_i$ for the observed data
- In the following, let y denote an arbitrary n-dimensional vector of 0's and 1's:
  - $A := \{y : \sum y_i = t\}$
  - $B(c_1, c_2) := \{y : \sum y_i x_{i1} \notin [c_1, c_2]\}$
  - $B_1 := \{y : \sum y_i x_{i1} = c_1\}$
  - $B_2 := \{y : \sum y_i x_{i1} = c_2\}$

It is easy to see that:

$$
\begin{aligned}
E\left[\phi(\sum Y_i x_{i1})\middle|\sum Y_i = t\right] &= P\left(\sum Y_i x_{i1} \notin [c_1, c_2]\middle|\sum Y_i = t\right) + \gamma_1 * P\left(\sum Y_i x_{i1} = c_1\middle|\sum Y_i = t\right) + \\
&\qquad \gamma_2 * P\left(\sum Y_i x_{i1} = c_2\middle|\sum Y_i = t\right) \\
&= \frac{P\left(Y \in (A \cap B)\right) + \gamma_1 * P\left(Y \in (A \cap B_1)\right) + \gamma_2 * P\left(Y \in (A \cap B_2)\right)}{P\left(Y \in A\right)} \\
&= \frac{\sum\limits_{y \in A} c(\beta) exp\left\{\beta_0 \sum y_i + \beta_1 \sum y_i x_{i1}\right\} * \left(I\{y \in B\} + \gamma_1 * I\{y \in B_1\} + \gamma_2 * I\{y \in B_2\}\right)}{\sum\limits_{y \in A} c(\beta) exp\left\{\beta_0 \sum y_i + \beta_1 \sum y_i x_{i1}\right\}} \\
&= \frac{\sum\limits_{y \in A} \left(I\{y \in B\} + \gamma_1 * I\{y \in B_1\} + \gamma_2 * I\{y \in B_2\}\right)}{\sum\limits_{y \in A} 1}
\end{aligned}
$$

Note that the cancelling of terms in the above is allowable despite changing and differing y-values because we have conditioned on $\sum y_i = t$ and also mandated that $\beta_1 = 0$ on the boundary of the null.

Now, consider:

$$
\begin{aligned}
E\left[\sum Y_i x_{i1} | \sum Y_i = t\right] &= \sum_{y \in A} \left(\sum y_i x_{i1}\right) * P(Y = y | \sum Y_i = t) \\
&= \frac{\sum\limits_{y \in A} \left(\sum y_i x_{i1}\right)}{\sum\limits_{y \in A} 1}
\end{aligned}
$$

And finally, consider:

$$
\begin{aligned}
E\left[\left(\sum Y_i x_{i1}\right) \phi\left(\sum Y_i x_{i1}\right) \middle| \sum Y_i = t\right] &= \sum_{y \in A} \left(\sum y_i x_{i1}\right) * \phi\left(\sum y_i x_{i1}\right) * P(Y = y | \sum Y_i = t) \\
&= \frac{\sum\limits_{y \in A} \left(\sum y_i x_{i1}\right) * \left(I\{y \in B\} + \gamma_1 * I\{y \in B_1\} + \gamma_2 * I\{y \in B_2\}\right)}{\sum\limits_{y \in A} 1}
\end{aligned}
$$

We may then choose c_1, c_2, γ_1, and γ_2 to satisfy the above requirements and that $(\gamma_1, \gamma_2) \in [0, 1] \times [0, 1]$. We may proceed by enumerating all possible vectors Y (or samples) for our data. Then restrict to only those samples in A. Then we may vary c_1 and c_2 within these samples until we find suitable cutpoints.

Exercise 1.B

We may approximate the conditional distribution of our test statistic in the following way. Define:

- $\mu = E\left[\sum Y_i x_{i1} \middle| \sum Y_i = t\right]$, as found in part (A)
- $\sigma^2 = V\left[\sum Y_i x_{i1} \middle| \sum Y_i = t\right]$.

  To calculate this value, consider $E\left[\left(\sum Y_i x_{i1}\right)^2 \mid \sum Y_i = t\right]$ which is computed by the same procedure as for μ, except we substitute a $\left(\sum y_i x_{i1}\right)^2$ term in the sum. Then consider the difference of this quantity and μ^2.

Once these values have been obtained, we use the same structure of the test as was found in part (A), given here by:

$$\phi\left(\sum y_i x_{i1}\right) = \begin{cases} 1, & \text{if } \sum y_i x_{i1} < c_1(\sum y_i) \text{ or } \sum y_i x_{i1} > c_2(\sum y_i). \\ 0, & \text{if otherwise.} \end{cases}$$

where $c(\sum y_i)$ is defined so that $E[\phi(\sum Y_i x_{i1}) | \sum Y_i] = \alpha$ and $E[\sum Y_i x_{i1} \phi(\sum Y_i x_{i1}) | \sum Y_i] = \alpha E[\sum Y_i x_{i1} | \sum Y_i]$. The randomization terms in the above disappear since we now assume that our test statistic is a continuous random variable.

Now, we may select c_1, c_2 to satisfy the following equations.

$$\alpha = 1 - \int_{c_1}^{c_2} \left(\frac{1}{\sqrt{2\pi}\sigma}\right) \texttt{exp}\{(-1/2\sigma^2)(z-\mu)^2\} \partial z$$

$$\alpha\mu = \mu - \int_{c_1}^{c_2} \left(\frac{z}{\sqrt{2\pi}\sigma}\right) \texttt{exp}\{(-1/2\sigma^2)(z-\mu)^2\} \partial z$$

We can use numerical integration techniques to solves these equations.

Exercise 1.C

First, let's find the gradient and the Hessian of the log-likelihood function $l(\beta)$:

- $\dot{l}(\beta) = X^T\left[Y - P\right]$
  - X is our covariate matrix including an intercept.
  - $p_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$
  - $P = [p_1, p_2, ..., p_n]^T$
  - $\frac{e^{X\beta}}{1+e^{X\beta}} = \left[\,, \cdots,\,\right]$
- $\ddot{l}(\beta) = -X^T\texttt{diag}(P\cdot(1-P))X$ where the $\cdot$ denotes element-wise multiplication.

Under the null, $\beta_1 = 0$ which gives that $p_i = \pi_0 = \frac{e^{\beta_0}}{1+e^{\beta_0}}$. It is trivial that the MLE of π_0 under the null is given by $\hat{\pi_0} = \sum y_i/n$. Thus, we have:

$$\tilde{\beta} = \begin{pmatrix} \texttt{logit}\left(\sum y_i/n\right) \\ 0 \end{pmatrix}$$

Then the score test is given by:

$$\begin{aligned} SC_n &= \left[0, \sum y_i x_{i1} - \hat{\pi_0}\sum x_{i1}\right]\left(X^T\texttt{diag}(\tilde{P}\cdot(1-\tilde{P}))X\right)^{-1}\begin{bmatrix} 0 \\ \sum y_i x_{i1} - \hat{\pi_0}\sum x_{i1} \end{bmatrix} \\ &= \left[\frac{1}{\hat{\pi_0}(1-\hat{\pi_0})}\right]\frac{\left(\sum y_i x_{i1} - \hat{\pi_0}\sum x_{i1}\right)^2}{\left(\sum(x_{i1} - \bar{x_1}^2)\right)} \end{aligned}$$

This statistic is asymptotically chi-squared with 1 degree of freedom, under the null. So reject when greater than 3.84.

As with the test in part (B), we reject when $\sum y_i x_{i1}$ is too small or too large since the numerator term is squared. We also note the connection between use of the unit normal distribution squared and the chi-squared distribution.

Exercise 1.D

Under the new framework established, define the following:

- $l = [l_1, ..., l_p]^T$ is the vector defined by the stated hypothesis.
- $l_{(-1)}$ is the vector l with the first element removed.
- $\beta_{(-1)}$ is the β vector with the first element removed.
- $\theta = l^T\beta - \theta_0$ - a new parameter for testing.
- $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$ - the new hypothesis with respect to θ

Under these definitions, we can see that:

$$\beta_1 = \frac{\theta + \theta_0 - l_{(-1)}^T \beta_{(-1)}}{l_1}$$

Using the work established above, it is easy to see that the log-likelihood under this framework can be written by:

$$\begin{aligned}
L &= \texttt{exp}\left\{\sum_{i=1}^{n} y_i x_i^T\beta - ln(1 + e^{x_i^T\beta})\right\} \\
&= c(\beta)\texttt{exp}\left\{\beta_1 \sum_{i=1}^{n} y_i x_{i1} + \cdots + \beta_p \sum_{i=1}^{n} y_i x_{ip}\right\} \\
&= c(\beta)\texttt{exp}\left\{(\theta + \theta_0) \sum_{i=1}^{n} y_i (x_{i1}/l_1) + \beta_2 \sum_{i=1}^{n} y_i (x_{i2} - (l_2/l_1)\, x_{i1}) + \cdots + \beta_p \sum_{i=1}^{n} y_i (x_{ip} - (l_p/l_1)\, x_{i1})\right\}
\end{aligned}$$

We can now define the following sufficient statistics:

- $S_1 = \sum_{i=1}^{n} y_i (x_{i1}/l_1)$
- $S_k = \sum_{i=1}^{n} y_i (x_{ik} - (l_k/l_1)\, x_{i1})$ for $k = 2, ..., p$
- $S_{(-1)} = (S_2, \cdots, S_p)$

We may then define our UMPU test in the following way:

$$\phi(S_1) = \begin{cases} 1, & \text{if } S_1 < c_1(S_{(-1)}) \text{ or } S_1 > c_2(S_{(-1)}). \\ \gamma_i & \text{if } S_1 = c_i(S_{(-1)}) \\ 0, & \text{if otherwise.} \end{cases}$$

where $E\left[\phi(S_1) \middle| S_{(-1)}\right] = \alpha$ and $E[S_1\phi(S_1) \middle| S_{(-1)}] = \alpha E[S_1 | S_{(-1)}]$ at the boundary of the null space.

As in part (A), we must define the quantities above to assist in the selection of c_1 and c_2.

Notation:
Before we proceed, define the following:

- Let Y denote the random variable of which our observed data vector y is a realization.
- s_k is the observed value of S_k in the sample for $k = 1, 2, ..., p$
- In the following notation for sets, note that y is an n-dimensional vector of 0's and 1's. It is not necessarily equal to the observed data vector y.
  - $A := \{y : S_2(y) = s_2, \cdots, S_p(y) = s_p\}$
  - $B := \{y : S_1(y) \notin [c_1, c_2]\}$
  - $B_1 := \{y : S_1(y) = c_1\}$
  - $B_2 := \{y : S_1(y) = c_2\}$

Using this notation, we see that:

$$
\begin{aligned}
E\left[\phi(S_1) \middle| S_{(-1)}\right] &= P\left(S_1 \notin (c_1, c_2) \middle| S_{(-1)}\right) + \gamma_1 * P\left(S_1 = c_1 \middle| S_{(-1)}\right) + \gamma_2 * P\left(S_1 = c_2 \middle| S_{(-1)}\right) \\
&= \frac{P\left(Y \in (A \cap B)\right) + \gamma_1 * P\left(Y \in (A \cap B_1)\right) + \gamma_2 * P\left(Y \in (A \cap B_2)\right)}{P\left(Y \in A\right)} \\
&= \frac{\sum\limits_{y \in A} \left(I\{y \in B\} + \gamma_1 * I\{y \in B_1\} + \gamma_2 * I\{y \in B_1\}\right) P\left(Y = y\right)}{\sum\limits_{y \in A} P(Y = y)} \\
&= \frac{\sum\limits_{y \in A} \left(I\{y \in B\} + \gamma_1 * I\{y \in B_1\} + \gamma_2 * I\{y \in B_1\}\right) c(\beta) \mathtt{exp}\left\{(\theta + \theta_0) S_1(y) + \cdots + \beta_p S_p(y)\right\}}{\sum\limits_{y \in A} c(\beta) \mathtt{exp}\left\{(\theta + \theta_0) S_1(y) + \beta_2 S_2(y) + \cdots + \beta_p S_p(y)\right\}} \\
&= \frac{\sum\limits_{y \in A} \left(I\{y \in B\} + \gamma_1 * I\{y \in B_1\} + \gamma_2 * I\{y \in B_1\}\right) c(\beta) \mathtt{exp}\left\{(\theta + \theta_0) S_1(y) + \beta_2 s_2 \cdots + \beta_p s_p\right\}}{\sum\limits_{y \in A} c(\beta) \mathtt{exp}\left\{(\theta + \theta_0) S_1(y) + \beta_2 s_2 + \cdots + \beta_p s_p)\right\}} \\
&= \frac{\sum\limits_{y \in A} \left(I\{y \in B\} + \gamma_1 * I\{y \in B_1\} + \gamma_2 * I\{y \in B_1\}\right) \mathtt{exp}\left\{\theta_0 S_1(y)\right\}}{\sum\limits_{y \in A} \mathtt{exp}\left\{\theta_0 S_1(y)\right\}}
\end{aligned}
$$

The above cancellations hold since $\theta = 0$ on the boundary of the null and because in set A, the values of $S_2, ..., S_p$ are fixed.

Now, following the work above, we have that:

$$E\left[S_1\phi(S_1)\middle|S_{(-1)}\right] = \frac{\sum\limits_{y\in A} S_1(y)\left(I\{y\in B\} + \gamma_1 * I\{y\in B_1\} + \gamma_2 * I\{y\in B_1\}\right)\texttt{exp}\left\{\theta_0 S_1(y)\right\}}{\sum\limits_{y\in A}\texttt{exp}\left\{\theta_0 S_1(y)\right\}}$$

Additionally, we can see that:

$$E[S_1|S_{(-1)}] = \frac{\sum\limits_{y\in A} S_1(y)\texttt{exp}\left\{\theta_0 S_1(y)\right\}}{\sum\limits_{y\in A}\texttt{exp}\left\{\theta_0 S_1(y)\right\}}$$

We may compute these values directly by enumerating all possible values of the random variable Y, and restricting to only those in set A. Then it is just a matter of computing each of these quantities for each possible sample.

We select as values c_1, c_2, γ_1, and γ_2 so that the criteria specified above are met in addition to the requirement that $(\gamma_1, \gamma_2) \in [0,1] \times [0,1]$.

Exercise 1.E

Define the following quantities:

- $Y^T = [y_1, \cdots, y_n]$
- $\tilde{X} = [X_1, X_2]$
  - $X_1 = x_1/l_1$ is an $n \times 1$ matrix.
  - $X_2 = [x_1/l_1, \quad x_2 - (l_2/l_1)x_1, \quad \cdots, \quad x_p - (l_p/l_1)x_1]$ is an $n \times p - 1$ matrix.
- $\tilde{\beta} = (\theta_0, \theta, \beta_2, \cdots, \beta_p)$

Under this definition, it is clear that the likelihood is given by:

$$L = c(\beta)\texttt{exp}\left\{Y^T \tilde{X} \tilde{\beta}\right\}$$

We may then conduct the non-parametric bootstrap as follows:

(0) Define a plausible range–described by $[e_1, e_2]$–for the test statistic $S_1 = \sum y_1 \left(x_{i1}/x_1^*\right)$.

(1) Create B independent bootstrap samples by sampling n rows from the matrix $[Y, \tilde{X}]$ with replacement.

(2) For each sample b, perform the following steps:

(2.1) Define $X_b = [X_1, \tilde{X_b}]$ where $\tilde{X_b}$ are the covariate terms that were sampled. X_1 remains in its original ordering.

In this way, we shuffle the X_1 values to break association between Y and θ, but still preserve relationships between Y and the remaining parameters.

(2.2) Compute S_1 for this sample, label it S_b.

(2.3) Compute the following

$$t_b = \min(S_1 - e_1, e_2 - S_1)$$

(3) Compute S_1 for the original data, labeled S_0 and subsequently $t_0 = \min(S_0 - e_1, e_2 - S_0)$.

(4) Compute the exact p-value using the formula:

$$p_{boot} = \frac{\sum_{i=1}^{B} I\{t_b <= t_0\}}{B}$$

Step 0 and use of the t_b allows us to compute a two-sided test which mirrors the UMPU test. Alternatively, we could use the conditional p-values for a UMPU test for each sample and compare those across bootstrap samples and the original sample.

Exercise 2.A

Define the following notation:

- $\theta = (\mu, p)$
- $c(\theta) = \texttt{exp}\left\{-(1/2)\mu^2 + n\texttt{log}(1-p)\right\}$
- $h(s, \vec{x}) = \texttt{exp}\left\{-[(s+1)/2]\texttt{log}(2\pi) - (1/2)\sum_{i=1}^{s+1} x_i^2 + \texttt{log}\binom{n}{s}\right\}$
- $\xi_1 = \mu$
- $\xi_2 = \texttt{logit}(p) - (1/2)\mu^2$
- $T_1 = (s+1)\bar{x}$
- $T_2 = s$

We may then specify the likelihood for the data in the following way:

$$
\begin{aligned}
f_{S,\vec{X}}(s, \vec{x}) &= f_S(s) f_{\vec{X}|S=s}(x_1, \cdots, x_n) \\
&= \binom{n}{s} p^s (1-p)^{n-s} (2\pi)^{-(s+1)/2} \texttt{exp}\left\{\sum_{i=1}^{s+1} -(1/2)(x_i - \mu)^2\right\} \\
&= \texttt{exp}\left\{\texttt{log}\binom{n}{s} - [(s+1)/2]\texttt{log}(2\pi) - (1/2)\sum_{i=1}^{s+1} x_i^2\right\} \texttt{exp}\left\{n\texttt{log}(1-p) - (1/2)\mu^2\right\} \times \\
&\qquad \texttt{exp}\left\{\mu(s+1)\bar{x} + s\texttt{logit}(p) - (s/2)\mu^2\right\} \\
&= h(s, \vec{x}) c(\theta) \texttt{exp}\left\{\sum_{i=1}^{2} \xi_i T_i(s, \vec{x})\right\}
\end{aligned}
$$

Since $\xi_1 \in (-\infty, \infty)$ and $\xi_2 \in (-\infty, \infty)$, a parameter space which contains an open set in $\mathbb{R}^2$, it is clear that we have a full-rank member of the multiparameter exponential family. The sufficient statistic is given by $T = (T_1, T_2)$. For full rank exponential families, the minimal sufficient statistic T is complete.

The above holds for a single dimensional family when $\mu = 0$ by letting $c(\theta) = c(p) = (1-p)^n$, $\xi_1 = 0$, and $\xi_2 = \texttt{logit}(p)$. The complete sufficient statistic is one-dimensional in this case and is given by $T = T_2$ as defined above.

Exercise 2.B

To derive the MLE's, consider the portion of the log-likelihood containing our parameters $\theta = (\mu, p)$.

$$l = -[(s+1)/2]\mu^2 + \mu[(s+1)\bar{x}] + s\,\texttt{logit}(p) + n\,\texttt{log}(1-p)$$

We then have the following gradient and Hessian:

- $\dot{l} = \begin{pmatrix} (s+1)(\bar{x}-\mu) \\ (s-np)[p(1-p)]^{-1} \end{pmatrix}$
- $\ddot{l} = -\begin{pmatrix} s+1 & 0 \\ & \\ 0 & sp^{-2} + (n-s)(1-p)^{-2} \end{pmatrix}$

It is clear from the above that $\ddot{l}$ is negative definite across all θ. Thus, the following point – which satisfies $\dot{l} = 0$ – is the global maximum:

$$(\hat{\mu}, \hat{p}) = (\bar{x}, s/n)$$

Exercise 2.C

As we are in the exponential family, our standard asymptotic MLE theory applies:

$$\sqrt{n}\left(\begin{pmatrix}\hat{\mu}\\ \hat{p}\end{pmatrix} - \begin{pmatrix}\mu\\ p\end{pmatrix}\right) \underset{d}{\rightarrow} N\left(0, I^{-1}\right)$$

Where I is the Fisher's information matrix. Using the Hessian specified in part (B), it is clear that:

$$I^{-1} = E[\ddot{l}]^{-1} = \begin{pmatrix}[np+1]^{-1} & 0\\ 0 & p(1-p)/n\end{pmatrix}$$

***Exercise* 2.D**

By remark 2.3.1 of the BIOS 761 notes [slide 306], multiparameter exponential families have the property that the power function – $\beta_\phi(\theta)$ – for a test ϕ is continuous in θ for all ϕ.

Now suppose that ϕ is unbiased. This implies that:

- $\beta_\phi(\theta) \leq \alpha$ for $\theta \in \Theta_{H0}$
- $\beta_\phi(\theta) \geq \alpha$ for $\theta \in \Theta_{H1}$

Define $\Theta_B := \{\theta = (\mu, p) : \mu = 0\}$. Fix a value of p, call it p^*. By the definition of continuity and the existence of a limit, it must be that as $\theta \to (0, p^*)$ from the left and the right of the boundary, the power approaches a value β^*. This result, in combination with the definition of unbiasedness, guarantees that $\beta^* = \alpha$. Since p^* was arbitrary, this holds across the entire boundary.

To be a little more specific, if $\beta^* < \alpha$ then the test could not also be both continuous and unbiased as is required. We see this since the limit on the right side of the boundary would necessarily approach a value greater than or equal to α. However, the value at the boundary would be less than α. Thus, we have a contradiction to the assumption of continuity.

Were $\beta^* > \alpha$, we would not meet the definition of unbiasedness specified in bullet (1). Again, we have a contradiction.

Thus, by contradiction, it is clear that $\beta(0, p) = \alpha$.

Exercise 2.E

Using the results of problem (2.A), we can define the UMPU test in the following way:

$$\phi(T_1) = \begin{cases} 1, & \text{if } T_1 > c(T_2). \\ 0, & \text{if } T_1 \leq c(T_2). \end{cases} = \begin{cases} 1, & \text{if } \bar{x} > c'(s). \\ 0, & \text{if } \bar{x} \leq c'(s). \end{cases}$$

where $c'(s)$ is defined so that $E_{\Theta_0}[\phi\left([s+1]\bar{x}\right)|S = s] = \alpha$.

Note that, on the boundary, we have the following:

$$(S+1)\bar{x}|S = s \sim N(0, s+1)$$

Which implies:

$$\left(\sqrt{S+1}\right)\bar{x}|S = s \sim N(0,1)$$

Then we may define our test as:

$$\phi(T_1) = \begin{cases} 1, & \text{if } \bar{x} > \Phi^{-1}(1-\alpha)/\sqrt{s+1}. \\ 0, & \text{if } \bar{x} \leq \Phi^{-1}(1-\alpha)/\sqrt{s+1}. \end{cases}$$

where $\Phi^{-1}(x)$ is the inverse CDF of a N(0,1) random variable.