STAT 151A Optional Problems 3: Solutions

Billy Fang

These are rough sketches for the solutions. Some computational steps are omitted for brevity.
Beware of mistakes/typos...

1

(a) False. You do not use the coefficient $\widehat{\beta}$ from the full model. Each of the folds has a different OLS coefficient "$\widehat{\beta}$," since you fit the model on different subsets of the data.

(b) True. The null deviance can be interpreted as the residual deviance of the intercept-only model. Thus, the claim follows directly from the more general fact that adding variables to a model (e.g. going from the intercept-only model to a larger model that has an intercept) decreases the residual deviance.

(c) False. Null deviance only depends on the response variable y, and would not give you enough information to compute AIC or BIC for any logistic model. However, AIC and BIC can be computed from the residual deviance and the number of explanatory variables.

2

14.2

The CDF of the uniform distribution on $[0, 1]$ is

$$P(u) = \begin{cases} 0 & u \leq 0, \\ u & 0 < u < 1, \\ 1 & u \geq 1. \end{cases}$$

Plugging in $u = \alpha + \beta X_i$ yields the linear-probability model (Equation 14.3).

14.3

$$\frac{d}{dx}\frac{1}{1+e^{-(\alpha+\beta x)}} = \frac{\beta e^{-(\alpha+\beta x)}}{[1+e^{-(\alpha+\beta x)}]^2} = \beta\frac{1}{1+e^{-(\alpha+\beta x)}}\frac{e^{-(\alpha+\beta x)}}{1+e^{-(\alpha+\beta x)}} = \beta\pi(1-\pi).$$

14.4

$$p(0) = \pi_i^0(1-\pi_i)^{1-0} = 1-\pi_i$$
$$p(1) = \pi_i^1(1-\pi_i)^{1-1} = \pi_i$$

14.9

We are asked to check whether the following holds.

$$\begin{aligned} b &= (X^\top V X)^{-1} X^\top V y^* \\ &= (X^\top V X)^{-1} X^\top V (Xb + V^{-1}(y-p)) \\ &= b + (X^\top V X)^{-1} X^\top (y-p). \end{aligned}$$

Thus we just need to check the second term is zero. [The textbook uses p to denote the vector of fitted values.] This follows from the zero gradient condition $X^\top(y-p) = 0$.

15.6

Since we are given the "answers" in the table already, you can just check that they produce the desired distribution. In principle one could also re-derive the results in the table from scratch.

- **Poisson.** Plugging in the results from the table yields.

$$p(y;\theta,\phi) = \exp\big[y\theta - e^\theta - \log(y!)\big] = e^{-e^\theta}\frac{(e^\theta)^y}{y!} = e^{-\lambda}\frac{\lambda^y}{y!}.$$

  This is the PMF of the Poisson distribution with mean $\lambda = e^\theta$; conversely, the canonical parameter is $\theta = \log\lambda$. As a sanity check, note that $b'(\theta) = e^\theta = \lambda$ which is the mean.

  You can also obtain the results in the table from scratch, as we showed in lab.

- **Gamma.**

$$\begin{aligned} p(y;\theta,\phi) &= \exp\left[\frac{y\theta + \log(-\theta)}{\phi} + \frac{\log(y/\phi)}{\phi} - \log y - \log\Gamma(\phi^{-1})\right] \\ &= \frac{e^{y\theta/\phi}(-y\theta/\phi)^{1/\phi}}{y\Gamma(\phi^{-1})} \\ &= \frac{(-\theta/\phi)^{\phi^{-1}}}{\Gamma(\phi^{-1})} y^{\phi^{-1}-1} e^{-(-\theta/\phi)y} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}. \end{aligned}$$

  This is the PDF of the Gamma distribution with parameters $\alpha = 1/\phi$ and $\beta = -\theta/\phi$; conversely, the canonical parameter is $\theta = -\beta/\alpha$ and the dispersion parameter is $1/\alpha$. As a sanity check, note that $b'(\theta) = -\frac{1}{\theta} = \frac{\alpha}{\beta}$, which is the mean of the distribution.

  If you didn't have the table and wanted to re-derive the results, you can do it from scratch as follows.

$$\begin{aligned} p(y) &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \\ &= \exp[-\beta y + (\alpha-1)\log y + \alpha\log\beta - \log\Gamma(\alpha)] \\ &= \exp\left[\frac{(-\beta/\alpha)y + \log\beta}{1/\alpha} + (\alpha-1)\log y - \log\Gamma(\alpha)\right] \\ &= \exp\left[\frac{(-\beta/\alpha)y + \log(-(-\beta/\alpha))}{1/\alpha} + \alpha\log\alpha + (\alpha-1)\log y - \log\Gamma(\alpha)\right] \\ &= \exp\left[\frac{(-\beta/\alpha)y + \log(-(-\beta/\alpha))}{1/\alpha} + \frac{1}{1/\alpha}\log\frac{y}{1/\alpha} - \log y - \log\Gamma(\frac{1}{1/\alpha})\right]. \end{aligned}$$

  So $\theta = -\beta/\alpha$, $\phi = 1/\alpha$, $b(\theta) = -\log(-\theta)$, and $c(y,\phi) = \frac{1}{\phi}\log\frac{y}{\phi} - \log y - \log\Gamma(\phi^{-1})$.

- **Inverse-Gaussian.** Since there was a typo in the table, we will derive the correct results from scratch.

$$\begin{aligned} p(y) &= \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right] \\ &= \exp\left[-\frac{\lambda}{2\mu^2}(y - 2\mu + \mu^2 y^{-1}) - \frac{1}{2}\log(2\pi y^3/\lambda)\right] \\ &= \exp\left[y\left(-\frac{\lambda}{2\mu^2}\right) - \left(-\frac{\lambda}{\mu}\right) - \frac{1}{2}\left(\log(2\pi y^3/\lambda) + \frac{\lambda}{y}\right)\right] \end{aligned}$$

If we choose $\phi = 1/\lambda$, then $\theta = -\frac{1}{2\mu^2}$, $b(\theta) = -\sqrt{-2\theta}$, and $c(y, \phi) = -\frac{1}{2}\left[\log(2\pi\phi y^3) + \frac{1}{\phi y}\right]$.

Alternate: If we choose $\phi = 2/\lambda$, then $\theta = -\frac{1}{\mu^2}$, $b(\theta) = -2\sqrt{-\theta}$, and $c(y, \phi) = -\frac{1}{2}\left[\log(\pi\phi y^3) + \frac{2}{\phi y}\right]$.

15.7

- **Gaussian.**

$$a(\phi)b''(\theta) = \phi\frac{d^2}{d\theta^2}\frac{\theta^2}{2} = \phi.$$

- **Binomial.**

$$a(\phi)b''(\theta) = \frac{1}{n}\frac{d^2}{d\theta^2}\log(1 + e^\theta) = \frac{1}{n}\frac{d}{d\theta}\frac{1}{1 + e^{-\theta}} = \frac{1}{n}\frac{e^{-\theta}}{[1 + e^{-\theta}]^2} = \frac{1}{n}\mu(1 - \mu),$$

where in the last step we used the fact that $\mu = b'(\theta) = \frac{1}{1+e^{-\theta}}$. (Similar computation to Exercise 14.3.)

- **Poisson.**

$$a(\phi)b''(\theta) = \frac{d^2}{d\theta^2}e^\theta = e^\theta = \mu,$$

where we used $\mu = b'(\theta) = e^\theta$.

- **Gamma.**

$$a(\phi)b''(\theta) = \phi\frac{d^2}{d\theta^2}(-\log(-\theta)) = \phi\frac{d}{d\theta}(-1/\theta) = \phi\frac{1}{\theta^2} = \phi\mu^2,$$

where we used the fact that $\mu = b'(\theta) = -\frac{1}{\theta}$. Recalling $\theta = -\beta/\alpha$ and $\phi = 1/\alpha$, we see that we get the correct variance $\phi\mu^2 = \alpha/\beta^2$.

- **Inverse-Gaussian.**

$$a(\phi)b''(\theta) = \phi\frac{d^2}{d\theta^2}(-\sqrt{-2\theta}) = \phi\frac{d}{d\theta}\frac{1}{\sqrt{-2\theta}} = \phi\frac{1}{\sqrt{2(-\theta)^3}} = \phi\mu^3,$$

where we used the fact that $\mu = b'(\theta) = \sqrt{\frac{1}{2(-\theta)}}$. In this setup we have $\phi = 1/\lambda$, so we get the correct variance μ^3/λ.

Alternate:

$$a(\phi)b''(\theta) = \phi\frac{d^2}{d\theta^2}(-2\sqrt{-\theta}) = \phi\frac{d}{d\theta}\frac{1}{\sqrt{-\theta}} = \phi\frac{1}{2\sqrt{(-\theta)^3}} = \phi\mu^3/2,$$

where we used the fact that $\mu = b'(\theta) = \sqrt{\frac{1}{-\theta}}$. Indeed, in this setup we have $\phi = 2/\lambda$, so we get the correct variance μ^3/λ.

3

To prove $R^2_{\text{adj}} \leq 1$, simply note that $\frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$ is nonnegative.

We now prove the slightly stronger fact that $R^2_{\text{adj}} \leq R^2$. Using the fact that $\frac{n-1}{n-p-1} \geq 1$, we have

$$R^2_{\text{adj}} \coloneqq 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} \leq 1 - \frac{\text{RSS}}{\text{TSS}} =: R^2.$$

4

I do not know how much computation is expected on the exam, but I have included computations of gradients and Hessians in case you are interested.

The link function is $g(\mu_i) = \log \mu_i$, since this is what we choose $\beta_1 + \beta_2 X_i$ to model.

The EDF form of the $\text{Poisson}(\mu_i)$ PMF is

$$p(y_i) = \frac{1}{y_i!} \exp\{y_i \log \mu_i - \mu_i\} = \frac{1}{y_i!} \exp\{y_i \theta_i - e^{\theta_i}\},$$

so this link function is also the canonical link function.

Plugging in $\theta_i = \log \mu_i = \beta_1 + \beta_2 X_i$ yields

$$\begin{aligned}
p(y_i) &= \frac{1}{y_i!} \exp\{y_i(\beta_1 + \beta_2 X_i) - e^{\beta_1 + \beta_2 X_i}\} \\
\ell(\beta) &= \sum_{i=1}^{n} \left[y_i(\beta_1 + \beta_2 X_i) - e^{\beta_1 + \beta_2 X_i} - \log(y_i!)\right] \\
\nabla \ell(\beta) &= \sum_{i=1}^{n} \begin{bmatrix} y_i - e^{\beta_1 + \beta_2 X_i} \\ (y_i - e^{\beta_1 + \beta_2 X_i}) X_i \end{bmatrix}.
\end{aligned}$$

There is no closed-form solution for $\nabla \ell(\beta) = 0$.

If we want to use Newton-Rhapson, we need to compute the Hessian.

$$H(\ell(\beta)) = -\sum_{i=1}^{n} e^{\beta_1 + \beta_2 X_i} \begin{bmatrix} 1 & X_i \\ X_i & X_i^2 \end{bmatrix}.$$

Then we can use the updates

$$\beta^{(m+1)} \leftarrow \beta^{(m)} - [H(\ell(\beta^{(m)}))]^{-1} \nabla \ell(\beta^{(m)}).$$

Alternatively we can use IWLS with the link function $g(\mu) = \log \mu$ (see Fox pp. 446-7). Since this is the canonical link, it will coincide with the above procedure.

5

I do not know how much computation is expected on the exam, but I have included computations of gradients and Hessians in case you are interested.

The link function is $g(\pi_i) = \Phi^{-1}(\pi_i)$, since this is what we choose $\beta_1 + \beta_2 X_i$ to model. [This is probit regression.]

With $\pi_i = \Phi(\beta_1 + \beta_2 X_i)$, we have

$$\begin{aligned}
p(y_i) &= \pi_i^{y_i}(1 - \pi_i)^{1 - y_i} = [\Phi(\beta_1 + \beta_2 X_i)]^{y_i} [1 - \Phi(\beta_1 + \beta_2 X_i)]^{1 - y_i} \\
\ell(\beta) &= \sum_{i=1}^{n} [y_i \log \Phi(\beta_1 + \beta_2 X_i) + (1 - y_i) \log[1 - \Phi(\beta_1 + \beta_2 X_i)]] \\
&= \sum_{i=1}^{n} \left[y_i \log \frac{\Phi(\beta_1 + \beta_2 X_i)}{1 - \Phi(\beta_1 + \beta_2 X_i)} + \log[1 - \Phi(\beta_1 + \beta_2 X_i)] \right].
\end{aligned}$$

Let $\phi := \Phi'$ be the PDF of the standard normal distribution. Then the gradient is

$$\nabla \ell(\beta) = \sum_{i=1}^{n} \left(y_i \frac{\phi(\beta_1 + \beta_2 X_i)}{\Phi(\beta_1 + \beta_2 X_i)} - (1 - y_i) \frac{\phi(\beta_1 + \beta_2 X_i)}{1 - \Phi(\beta_1 + \beta_2 X_i)} \right) \begin{bmatrix} 1 \\ X_i \end{bmatrix}.$$

There is not a closed form solution for the solution to $\nabla \ell(\beta) = 0$, so we will have to resort to some algorithm. If we want to use Newton-Rhapson, we need to compute the Hessian. Note that

$$\phi'(z) = \frac{d}{dz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} = -\frac{1}{\sqrt{2\pi}} z e^{z^2/2} = -z\phi(z).$$

Some tedious computation leads to

$$\begin{aligned} &H(\ell(\beta)) \\ &= -\sum_{i=1}^{n} \phi(\beta_1 + \beta_2 X_i) \left(y_i \frac{\phi(\beta_1 + \beta_2 X_i) + (\beta_1 + \beta_2 X_i)\Phi(\beta_1 + \beta_2 X_i)}{\Phi(\beta_1 + \beta_2 X_i)^2} \right. \\ &\qquad \left. + (1 - y_i) \frac{\phi(\beta_1 + \beta_2 X_i) - (\beta_1 + \beta_2 X_i)(1 - \Phi(\beta_1 + \beta_2 X_i))}{(1 - \Phi(\beta_1 + \beta_2 X_i))^2} \right) \begin{bmatrix} 1 & X_i \\ X_i & X_i^2 \end{bmatrix} \end{aligned}$$

Then, we can do the updates

$$\beta^{(m+1)} \leftarrow \beta^{(m)} - [H(\ell(\beta^{(m)}))]^{-1} \nabla \ell(\beta^{(m)}).$$

Alternatively you can also use IWLS with the link function $g(\pi) = \Phi^{-1}(\pi)$ (see Fox pp. 446-7).

6

Approach 1. Since

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i \in G_1} (y_i - \overline{y})^2 + \sum_{i \in G_2} (y_i - \overline{y})^2,$$

it suffices to show the following two inequalities.

$$\begin{aligned} \sum_{i \in G_1} (y_i - \overline{y}_1)^2 &\le \sum_{i \in G_1} (y_i - \overline{y})^2 \\ \sum_{i \in G_2} (y_i - \overline{y}_1)^2 &\le \sum_{i \in G_2} (y_i - \overline{y})^2. \end{aligned}$$

The first inequality follows directly from the fact that the function $g(z) = \sum_{i \in G_1} (y_i - z)^2$ is minimized at $z = \overline{y}_1$, which can be proved in a number of ways (e.g. setting derivative to zero, variance decomposition, etc.); note also that this is simply the OLS for an intercept-only model. The other inequality can be argued similarly.

Approach 2. Note that $\mathrm{RSS}(j, c)$ is the same as the RSS in one-way ANOVA if you had two groups G_1 and G_2. Then this claim follows immediately from $\mathrm{TSS} = \mathrm{RSS} + \mathrm{RegSS} \ge \mathrm{RSS}$.

7

(a) In this case, X is simply the all-ones vector. Plugging this into the expression for $\widetilde{\beta}$ yields

$$\widetilde{\beta} = (\mathbf{1}^\top \mathbf{1} + \lambda I)^{-1} \mathbf{1}^\top Y = \frac{n}{n + \lambda} \overline{Y}.$$

(b)

$$\begin{aligned} \mathbb{E}[\widetilde{\beta}] &= \frac{n}{n + \lambda} \mathbb{E}[\overline{Y}] = \frac{n}{n + \lambda} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i] = \frac{n}{n + \lambda} \beta, \\ \mathrm{Var}(\widetilde{\beta}) &= \left(\frac{n}{n + \lambda} \right)^2 \mathrm{Var}(\overline{Y}) = \left(\frac{n}{n + \lambda} \right)^2 \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(Y_i) = \left(\frac{n}{n + \lambda} \right)^2 \cdot \frac{\sigma^2}{n}. \end{aligned}$$

(c)

$$\text{MSE}(\widetilde{\beta}) = \text{bias}^2(\widetilde{\beta}) + \text{Var}(\widetilde{\beta}) = \left(\beta - \frac{n}{n+\lambda}\beta\right)^2 + \left(\frac{n}{n+\lambda}\right)^2 \cdot \frac{\sigma^2}{n} = (1-\alpha)^2\beta^2 + \alpha^2\frac{\sigma^2}{n}.$$

For OLS,

$$\text{MSE}(\overline{Y}) = \text{Var}(\overline{Y}) = \frac{\sigma^2}{n}$$

We have

$$\begin{aligned} & \text{MSE}(\widetilde{\beta}) < \text{MSE}(\overline{Y}) \\ \iff & (1-\alpha)^2\beta^2 + \alpha^2\frac{\sigma^2}{n} < \frac{\sigma^2}{n} \\ \iff & (1-\alpha)^2\beta^2 < (1-\alpha^2)\frac{\sigma^2}{n} \\ \iff & \frac{\beta^2}{\sigma^2} < \frac{1+\alpha}{n(1-\alpha)} \end{aligned}$$

Remarks. From part (a), we see that ridge regression simply shrinks the OLS estimate $\overline{Y}$ toward zero.

From part (b) we immediately see that $\widetilde{\beta}$ is a biased estimator, since the expectation is not β. However, we have traded off an increase in bias for a smaller variance $\left(\frac{n}{n+\lambda}\right)^2 \cdot \frac{\sigma^2}{n}$ compared to OLS, since the OLS estimator has variance σ^2/n.

Part (c) describes the exact conditions under which we get "more bang for our buck," that is, when the decrease in variance outweighs the increase in bias to yield a lower MSE overall.

8

(a) The number of errors for node 6 is $0.4417582 \cdot 455 = 201$, i.e. the smaller probability (`yprob`) times the number of datapoints at this node.

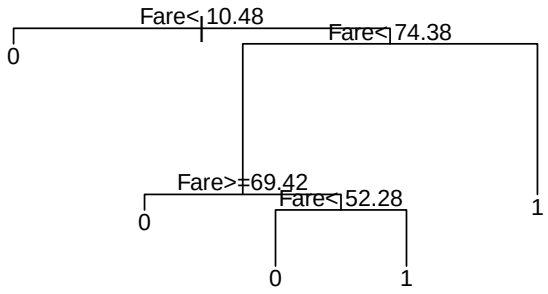
The `yval` for node 6 is 0, because the first `yprob` value (corresponding to 0) is larger than the second `yprob` value (corresponding to 1).

The two probability values for node 26 are $\frac{403-171}{403} \approx 0.576$ and $\frac{171}{403} \approx 0.424$ because there were 171 errors when predicting 0 for the 403 datapoints at this node.

The number of datapoints in node 7 is $552 - 455 = 97$ (number of datapoints in node 3 minus number of datapoints in node 6).

(b) You can use the node numbers in the `rpart` output by imagining a *complete* binary tree, and numbering the nodes in order, starting with the root being labeled 1. That is, the children of the root are $2, 3$, the nodes on the subsequent level are $4, 5, 6, 7$, and so on.

The labeled tree appears below.



(c) This datapoint would end up at the left node on the bottom level (node 26), so the predicted probability of survival is $\frac{171}{403} \approx 0.424$ (see part a).

(d) Note that precision and recall are $\frac{\#\{\text{predict=actual=1}\}}{\#\{\text{predict=1}\}}$ and $\frac{\#\{\text{predict=actual=1}\}}{\#\{\text{actual=1}\}}$ respectively.

From the terminal nodes (marked with $*$), we can focus on nodes 27, and 7 to see that we predict 1 for $37+97 = 134$ datapoints, and among these predictions, $(37-9)+(97-23) = 102$ of them were correct. Thus the precision is $\frac{102}{134} \approx 0.761$.

From the root node, we see that 342 of the datapoints were actually 1, so the recall is $\frac{102}{342} \approx 0.298$.

(e) From the output, we see that this datapoint would go through nodes $3, 6, 13, 26, 53, 107$, which predicts probability of survival 0.8125.

(f) This datapoint would go through nodes $3, 7$, which predicts probability of survival 0.94705882.