APPENDIX B

Solution Outlines for Selected Exercises

Note: This appendix contains brief outlines of solutions and hints of solutions for at least a few exercises from each chapter. Many of these are extracts of solutions that were kindly prepared by Jon Hennessy for Statistics 244 at Harvard University in 2013.

Chapter 1

1.1 In the random component, set $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $\phi = \sigma^2$, $a(\phi) = \phi$, and $c(y_i, \phi) = -y_i^2/2\phi - \log(2\pi\phi)$. Use the identity link function.

1.2 **b.** *Hint*: What is the range for a linear predictor, and what is the range of the identity link applied to a binomial probability or to a Poisson mean?

1.5 The predicted number of standard deviation change in y for a standard deviation change in x_i, adjusting for the other explanatory variables.

1.11 Taking $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{c-1})^{\mathrm{T}}$,

$$X = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_{r-1}} & \mathbf{0}_{n_{r-1}} & \mathbf{0}_{n_{r-1}} & \cdots & \mathbf{1}_{n_{r-1}} \\ \mathbf{1}_{n_r} & -\mathbf{1}_{n_r} & -\mathbf{1}_{n_r} & \cdots & -\mathbf{1}_{n_r} \end{pmatrix}$$

1.18 **b.** Let $\boldsymbol{G} = \boldsymbol{X}_1(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^{\mathrm{T}}$. Then $\boldsymbol{G}\boldsymbol{X}_c = \boldsymbol{X}_1(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c = \boldsymbol{X}_1$.

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.

1.19 **a.** For the model $E(y_{ij}) = \beta_0 + \beta_i + \gamma x_{ij}$, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_r, \gamma)^{\mathrm{T}}$, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{in_i})^{\mathrm{T}}$, and

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} & \boldsymbol{x}_1 \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} & \boldsymbol{x}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_{n_r} & \mathbf{0}_{n_r} & \mathbf{0}_{n_r} & \cdots & \mathbf{1}_{n_r} & \boldsymbol{x}_r \end{pmatrix}$$

b. (i) γ, because for each group it is a difference of means at x values one-unit apart. (ii) β_i

c. We can construct $\boldsymbol{X}$ to have identifiable parameters by imposing the constraint $\beta_1 = 0$. For a fixed x, β_i then represents the difference between $E(y)$ in group i and in group 1, and γ represents the change in $E(y)$ per unit increase in x for each group.

Chapter 2

2.3 **a.** The normal equation for β_j is $\sum_i y_i x_{ij} = \sum_i \mu_i x_{ij}$. For β_0, $\sum_i y_i = \sum_i (\beta_0 + \beta_1 x_i)$, so $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. For β_1, $\sum_i y_i x_i = n\bar{x}\beta_0 + \beta_1 \sum_i x_i^2$, so $\hat{\beta}_1 = [\sum_i (x_i - \bar{x})(y_i - \bar{y})]/[\sum_i (x_i - \bar{x})^2]$.

2.8 *Hint*: See McCullagh and Nelder (1989, p. 85) or Wood (2006, p. 13).

2.9 *Hint*: For simplification, express the model with all variables centered, so there is no intercept term, $\hat{\beta}_1$ and $\hat{\beta}_2$ are unchanged, and $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is 2×2. Since $\mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2) > 0$, the off-main-diagonal elements of $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$ are positive.

2.12 $\mathrm{rank}(\boldsymbol{H}) = \mathrm{tr}(\boldsymbol{H}) = \mathrm{tr}[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}] = \mathrm{tr}(\boldsymbol{I}_p) = p = \mathrm{rank}(\boldsymbol{X})$.

2.17 *Hint*: Need **0** be a solution?

2.19 **a.** Following an example in Rodgers et al. (1984), let

$$\boldsymbol{X} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{W} = \begin{pmatrix} -1 & 5 \\ -5 & 1 \\ 3 & 1 \\ -1 & 3 \end{pmatrix}.$$

The columns of $\boldsymbol{X}$ are uncorrelated and linearly independent. The columns of $\boldsymbol{W}$ are orthogonal and linearly independent. However, the columns of $\boldsymbol{X}$ are not orthogonal and the columns of $\boldsymbol{W}$ are not uncorrelated.

b. If $\mathrm{corr}(\boldsymbol{u}, \boldsymbol{v}) = 0$, then $\sum_i (u_i - \bar{u})(v_i - \bar{v}) = 0$. This implies that $\boldsymbol{u}^* = (\boldsymbol{u} - \bar{\boldsymbol{u}})$ and $\boldsymbol{v}^* = (\boldsymbol{v} - \bar{\boldsymbol{v}})$ are orthogonal. If $\bar{u} = 0$ (or equivalently $\bar{v} = 0$), then

$\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal since $\sum_i (u_i - \bar{u})(v_i - \bar{v}) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} = 0$. If $\bar{\boldsymbol{u}} \neq \boldsymbol{0}$ and $\bar{\boldsymbol{v}} \neq \boldsymbol{0}$, then there is no guarantee that $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal.

c. If $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} = 0$, then the numerator of corr$(\boldsymbol{u}, \boldsymbol{v})$ is $\sum_i (u_i - \bar{u})(v_i - \bar{v}) = (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} - n\bar{u}\bar{v}) = -n\bar{u}\bar{v}$. This can only equal 0 if at least one of $\bar{u}$ and $\bar{v}$ equals 0.

2.23 **a.** For the saturated model, $E(y_i) = \beta_i$, $i = 1, \dots, n$, $\boldsymbol{X} = \boldsymbol{I}_n$ and $C(\boldsymbol{X}) = \mathbb{R}^n$. $C(\boldsymbol{X})^{\perp} = \boldsymbol{0}$. $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{I}_n$ and $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{0}_{n\times n}$, the matrix of 0's.

b. $\hat{\boldsymbol{\beta}} = \boldsymbol{y}$ and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\beta}} = \boldsymbol{y}$. $s^2 = \sum_i [(y_i - y_i)^2]/(n - n) = 0/0$. The model is not sensible in practice because the prediction $\hat{\mu}_i$ only considers observation y_i and ignores the others. Also, the model is not useful for predicting new values of y.

2.25 $\boldsymbol{a}^{\mathrm{T}} = \boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$

2.29 From (2.6), the leverage for an observation in group i is $1/n_i$.

2.30 **a.** The mean of the leverages is $\mathrm{tr}(\boldsymbol{H})/n = \mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]/n = \mathrm{tr}[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}]/n = \mathrm{tr}(\boldsymbol{I}_p)/n = p/n$.

2.34 Model is not identifiable unless each factor has a constraint such as setting the parameter = 0 at the first or last level.

2.36 *Hint*: $\hat{\mu}_{ij} = \bar{\boldsymbol{y}}_{i.} + \bar{\boldsymbol{y}}_{.j} - \bar{\boldsymbol{y}}_{..}$. See Hoaglin and Welsch (1978).

Chapter 3

3.5 Use $\boldsymbol{I} = \boldsymbol{P}_0 + (\boldsymbol{I} - \boldsymbol{P}_0)$, with $\boldsymbol{P}_0$ the projection matrix for the null model. With $\boldsymbol{\mu}_0 = \mu_0\boldsymbol{1}$ and $\bar{\boldsymbol{y}} = \bar{y}\boldsymbol{1}$, we have $\boldsymbol{P}_0\boldsymbol{\mu}_0 = \boldsymbol{I}\boldsymbol{\mu}_0 = \boldsymbol{\mu}_0$ and $\boldsymbol{P}_0\boldsymbol{y} = \bar{\boldsymbol{y}}$,

$$(\boldsymbol{y} - \boldsymbol{\mu}_0)^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}_0) = (\boldsymbol{y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{P}_0(\boldsymbol{y} - \boldsymbol{\mu}_0) + (\boldsymbol{y} - \boldsymbol{\mu}_0)^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_0)(\boldsymbol{y} - \boldsymbol{\mu}_0)$$
$$= n(\bar{y} - \mu_0)^2 + (\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \bar{\boldsymbol{y}}).$$

By Cochran's Theorem, $(1/\sigma^2)n(\bar{y} - \mu_0)^2 \sim \chi^2_{1,\lambda}$, where $\lambda = (n/\sigma^2)(\mu - \mu_0)^2$, and $(1/\sigma^2)(\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \bar{\boldsymbol{y}}) = \frac{1}{\sigma^2}(n-1)s^2 \sim \chi^2_{n-1}$, and the two quantities are independent. Thus, the test statistic

$$\frac{n(\bar{y} - \mu_0)^2/\sigma^2}{s^2/\sigma^2} \sim \frac{\chi^2_{1,\lambda}}{\chi^2_{n-1}/(n-1)} \sim F_{1,n-1,\lambda}.$$

The equivalent t test uses the signed square root of this test statistic,

$$\frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim \pm\sqrt{F_{1,n-1,\lambda}} \sim t_{n-1,\lambda}.$$

Under H_0, $\lambda = 0$ and the null distributions are $F_{1,n-1}$ and t_{n-1}.

3.7 **a.** For $\lambda = (1/\sigma^2)\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{\mu}$,

$$(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{\mu} = \begin{pmatrix} (\mu_1 - \bar{\mu})\mathbf{1}_{n_1} \\ \vdots \\ (\mu_c - \bar{\mu})\mathbf{1}_{n_c} \end{pmatrix}.$$

Thus, $\lambda = (1/\sigma^2)\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{\mu} = (1/\sigma^2)\sum_i n_i(\mu_i - \bar{\mu})^2$.

b. $\lambda = (2n/\sigma^2)0.25\sigma^2 = n/2$. $P(F_{2,3n-3,\lambda} > F_{2,3n-3,0.05})$ equals 0.46 for $n = 10$, 0.94 for $n = 30$, and 0.99 for $n = 50$.

c. $\lambda = 2n\Delta^2$. The powers are 0.05, 0.46, and 0.97 for $\Delta = 0, 0.5, 1.0$.

3.9 **a.** Let $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}$, $\boldsymbol{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$. Then $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ and $\mu_1 = \beta_0 + \beta_1$ and $\mu_2 = \beta_0$. Thus $\mu_1 - \mu_2 = \beta_1$.

c. $H_0 : \mu_1 = \mu_2$ is $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$. We can use the decomposition

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_X - \boldsymbol{P}_0)\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y},$$

finding that $(1/\sigma^2)\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{P}_X - \boldsymbol{P}_0)\boldsymbol{y} = (1/\sigma^2)(\bar{y}_1 - \bar{y}_2)^2(\frac{1}{n_1} + \frac{1}{n_2})^{-1} \sim \chi^2_{1,\lambda}$, where $\lambda = 0$ under H_0, and $(1/\sigma^2)\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} = (1/\sigma^2)(n_1 + n_2 - 2)s^2 \sim \chi^2_{n_1+n_2-2}$, where s^2 is the pooled variance estimate. The test statistic is thus

$$\frac{(\bar{y}_1 - \bar{y}_2)^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{s^2} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \sim F_{1,n_1+n_2-2}.$$

d. The square root of F gives the t statistic, which has $df = n_1 + n_2 - 2$.

3.10 $(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,n_1+n_2-2}\, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

3.13 **d.** df values are $r - 1$, $c - 1$, $(r - 1)(c - 1)$, $(N - rc)$. Each mean square = SS/df. For H_0: no interaction, test statistic F = (interaction MS)/(residual MS) has $df_1 = (r - 1)(c - 1)$, $df_2 = N - rc$.

3.15 For models $k = 0$ and 1, substitute $1 - R_k^2 = (1 - R^2_{adj,k})[(n - p_k)/(n - 1)]$ and simplify.

3.17 **a.** Under the general linear hypothesis framework, let $\boldsymbol{\Lambda}$ be a single row of 0s except for 1 in position j and -1 in position k. The F test of H_0: $\boldsymbol{\Lambda\beta} = 0$ gives

$$F = \frac{(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}})^{\mathrm{T}}[\boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}]^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}}{s^2} = \frac{(\hat{\beta}_j - \hat{\beta}_k)^2}{s^2\boldsymbol{\Lambda}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}} \sim F_{1,n-p}.$$

The denominator simplifies to $[SE_j^2 + SE_k^2 - 2\widehat{\mathrm{cov}}(\hat{\beta}_j, \hat{\beta}_k)]$.

b. Equivalently, the F test compares the model M_1 with p parameters and the simpler model M_0 that replaces columns for $\boldsymbol{x}_{*j}$ and $\boldsymbol{x}_{*k}$ in the model matrix with a single column for $\boldsymbol{x}_{*j} + \boldsymbol{x}_{*k}$.

3.19 **a.** $\boldsymbol{\Lambda} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix}$.

b. $df_1 = c - 2$, $df_2 = n - c$.

c. Quantitative model has disadvantage that true relationship may be far from linear in the chosen scores (e.g., nonmonotonic), and qualitative model has disadvantage of lower power for detecting effect if true relationship is close to linear.

3.20 $\boldsymbol{\ell}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-p} s\sqrt{\boldsymbol{\ell}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{\ell}^{\mathrm{T}}}$.
Special case $(\hat{\beta}_j - \hat{\beta}_k) \pm t_{\alpha/2,n-p}\sqrt{SE_j^2 + SE_k^2 - 2\widehat{\mathrm{cov}}(\hat{\beta}_j, \hat{\beta}_k)}$.

3.21 *Hint*: For the null model, $\bar{y}$ is the estimated linear predictor.

3.22 The actual $P[|y - \bar{y}|/\sigma\sqrt{1 + \frac{1}{n}} \leq 1.96]$

$$= \Phi\left(1.96\sqrt{1 + \frac{1}{n}} + z_o/\sqrt{n}\right) - \Phi\left(-1.96\sqrt{1 + \frac{1}{n}} + z_o/\sqrt{n}\right).$$

3.23 Squared partial correlation = $(\mathrm{SSE}_0 - \mathrm{SSE}_1)/\mathrm{SSE}_0$.

3.26 With $\alpha = 0.05$, over most values of c and n, the ratio is on the order 0.96–0.98. For example, you can find the ratio in R using

```
> qtukey(1-alpha, c, c*(n-1))/(sqrt(2)*qt(1 - alpha/
(c*(c-1)), c*(n-1)))
```

3.30 When x_i is uniformly distributed over $(2.0, 4.0)$, $R^2 \approx 0.34$. When x_i is uniformly distributed over $(3.5, 4.0)$, $R^2 \approx 0.05$. The wider the range sampled for the explanatory variable, the larger R^2 tends to be, because in $R^2 = 1 - \mathrm{SSE/TSS}$, SSE tends to be unaffected but TSS tends to increase.

3.28 Expect same $E(\hat{\beta}_1)$ and σ^2 but larger SE for $\hat{\beta}_1$, since from Section 2.1.3, $\mathrm{var}(\hat{\beta}_1) = \sigma^2/[\sum_{i=1}^{n}(x_i - \bar{x})^2]$.

Chapter 4

4.7 **a.** For observations in group A, since $\partial\mu_A/\partial\eta_i$ is constant, the likelihood equation corresponding to β_1 sets $\sum_A (y_i - \mu_A)/\mu_A = 0$, so $\hat{\mu}_A = \bar{y}_A$. The likelihood equation corresponding to β_0 gives

$$\sum_A \frac{(y_i - \mu_A)}{\mu_A}\left(\frac{\partial\mu_A}{\partial\eta_i}\right) + \sum_B \frac{(y_i - \mu_B)}{\mu_B}\left(\frac{\partial\mu_B}{\partial\eta_i}\right) = 0.$$

The first sum is 0 from the first likelihood equation, and for observations in group B, $\partial\mu_B/\partial\eta_i$ is constant, so the second sum sets $\sum_B(y_i - \mu_B)/\mu_B = 0$, and $\hat{\mu}_B = \bar{y}_B$.

4.9 $\boldsymbol{W} = \sigma^{-2}\boldsymbol{I}$, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}$

4.11 *Hint*: If you have difficulty with this exercise, see Section 1.4 of Agresti (2013).

4.12 *Hint*: Construct a large-sample normal interval for $\boldsymbol{x}_0\boldsymbol{\beta}$ and then apply the inverse link function to the endpoints.

4.13

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1) = 2\sum_i y_i(\hat{\mu}_{1i} - \hat{\mu}_{0i}) + \frac{\hat{\mu}_{0i}^2}{2} - \frac{\hat{\mu}_{1i}^2}{2}$$
$$= \sum_i (y_i - \hat{\mu}_{0i})^2 - \sum_i (y_i - \hat{\mu}_{1i})^2.$$

4.14 Note that $\sum_i \hat{\mu}_i = \sum_i y_i$ is the likelihood equation generated by the intercept when the link function is canonical.

4.16 **a.** $d_i = 2[n_i y_i \log(y_i/\hat{\pi}_i) + n_i(1 - y_i)\log[(1 - y_i)/(1 - \hat{\pi}_i)]]$.

4.20 For log likelihood $L(\mu) = -n\mu + (\sum_i y_i)\log(\mu)$, the score is $u = (\sum_i y_i - n\mu)/\mu$, $H = -(\sum_i y_i)/\mu^2$, and the information is n/μ. It follows that the adjustment to $\mu^{(t)}$ in Fisher scoring is $[\mu^{(t)}/n][(\sum_i y_i - n\mu^{(t)})/\mu^{(t)}] = \bar{y} - \mu^{(t)}$, and hence $\mu^{(t+1)} = \bar{y}$. For Newton–Raphson, the adjustment to $\mu^{(t)}$ is $\mu^{(t)} - (\mu^{(t)})^2/\bar{y}$, so that $\mu^{(t+1)} = 2\mu^{(t)} - (\mu^{(t)})^2/\bar{y}$. Note that if $\mu^{(t)} = \bar{y}$, then also $\mu^{(t+1)} = \bar{y}$.

4.22 If the link is not canonical,

$$\frac{\partial\mu_i}{\partial\eta_i} = \frac{\partial\mu_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\eta_i} = b''(\theta_i)\frac{\partial\theta_i}{\partial\eta_i}$$

$$\frac{\partial L_i}{\partial\beta_j} = \frac{(y_i - \mu_i)}{\text{var}(y_i)}b''(\theta_i)\frac{\partial\theta_i}{\partial\eta_i}x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{a(\phi)}\frac{\partial\theta_i}{\partial\eta_i}.$$

Then, $\partial^2 L_i/\partial\beta_j\partial\beta_k$ depends on y_i, so $\partial^2 L/\partial\beta_j\partial\beta_k \neq E(\partial^2 L/\partial\beta_j\partial\beta_k)$.

4.25 *Hint*: Apply Jensen's inequality to $E[-\log(x)]$, where $P[x = (p_{Mj}/p_j)] = p_j$.

4.27 **a.** If y has standard deviation $\sigma = c\mu$, then using $\log(y) \approx \log(\mu) + (y - \mu)/\mu$, we have $\text{var}[\log(y)] \approx \text{var}(y)/\mu^2 = c^2$.

b. If $\log(y_i) \sim N(\mu_i, \sigma^2)$, then $E(y_i) = e^{\mu_i+\sigma^2/2}$ by the mgf of a normal. Thus, $\log[E(y_i)] = \mu_i + \sigma^2/2 = E[\log(y_i)] + \sigma^2/2$.

c. L_i is also the log-normal fitted median. If $f(\cdot)$ is a monotonic function, median$[f(x)] = f$[median(x)]. Then, $e^{\text{median}(\log(y))} = \text{median}(y)$ and e^{L_i} is the ML estimate of the median of the conditional distribution of y_i. The median would often be more relevant because y_i has a skewed distribution.

Chapter 5

5.5 **a.** Let $P(y = 1) = p$ and $P(y = 0) = 1 - p$. By Bayes' Theorem,

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)p}{P(x \mid y = 1)p + P(x \mid y = 0)(1 - p)}$$

$$= \frac{p(1/\sqrt{2\pi}\sigma)\exp[-(x-\mu_1)^2/2\sigma^2]}{p(1/\sqrt{2\pi}\sigma)\exp[-(x-\mu_1)^2/2\sigma^2] + (1-p)(1/\sqrt{2\pi}\sigma)\exp[-(x-\mu_0)^2/2\sigma^2]}$$

$$= \frac{\exp\{\log(\frac{p}{1-p}) - \frac{1}{2\sigma^2}[(x-\mu_1)^2 - (x-\mu_0)^2]\}}{1 + \exp\{\log(\frac{p}{1-p}) - \frac{1}{2\sigma^2}[(x-\mu_1)^2 - (x-\mu_0)^2]\}}$$

$$= \frac{\exp[\text{logit}(p) - \frac{1}{2\sigma^2}(\mu_1^2 - \mu_0^2) + \frac{\mu_1-\mu_0}{\sigma^2}x]}{1 + \exp[\text{logit}(p) - \frac{1}{2\sigma^2}(\mu_1^2 - \mu_0^2) + \frac{\mu_1-\mu_0}{\sigma^2}x]}.$$

So, set $\alpha = \text{logit}(p) - (1/2\sigma^2)(\mu_1^2 - \mu_0^2)$ and $\beta = (\mu_1 - \mu_0)/\sigma^2$.

b. When x has a $N(\mu_j, \sigma_j^2)$ distribution, then

$$P(y = 1 \mid x) = \frac{\exp[\text{logit}(p) + \log(\sigma_0/\sigma_1) - (x-\mu_1)^2/2\sigma^2 + (x-\mu_0)^2/2\sigma^2]}{1 + \exp[\text{logit}(p) + \log(\sigma_0/\sigma_1) - (x-\mu_1)^2/2\sigma^2 + (x-\mu_0)^2/2\sigma^2]}$$

Then $\alpha = \text{logit}(p) + \log(\sigma_0/\sigma_1) - \mu_1^2/2\sigma_1^2 - \mu_0^2/2\sigma_0^2$, $\beta = (\mu_1/\sigma_1^2) - (\mu_0/\sigma_0^2)$, and $\gamma = -(1/2)[(1/\sigma_1^2) - (1/\sigma_0^2)]$, where γ is the coefficient for the quadratic term.

5.3
```
> x <- c(1,2,3,4,5,6,7,8)
> y <- c(1,1,0,0,0,0,1,1)
> fit.toy <- glm(y ~ x, family = binomial)
> library(ROCR)
> pred.toy <- prediction(fitted(fit.toy), y)
> perf.toy <- performance(pred.toy,"tpr","fpr")
> plot(perf. toy)
> performance(pred.toy, "auc")
[1] 0.5
```

5.7 In terms of the probability π that $y = 1$ at x_0, $\beta = [\text{logit}(\pi)]/x_0$, $\text{var}(\hat{\beta}) \approx [n\pi(1-\pi)x_0^2]^{-1}$, and the noncentrality goes to 0 as $\beta \to \infty$ (i.e., as $\pi \to 1$). So, the Wald test loses power.

5.8 Condition on the margins of each 2×2 stratum. Let T be total number of successes for treatment 1, summed over strata. P-value is $P(T \geq t_{obs})$ for tables with the given margins, based on hypergeometric probabilities in each stratum. For details, see Agresti (1992 or 2013, Section 7.3.5).

5.9 *Hint*: There are $\binom{6}{3}$ possible data configurations with 3 successes, all equally-likely under H_0. Exact P-value = 0.05.

5.14 **a.** Assuming $\pi_1 = \cdots = \pi_N = \pi$, we can maximize

$$L(\pi) = \sum_{i=1}^{N} y_i \log(\pi) + (n_i - y_i)\log(1-\pi)$$

to show that $\hat{\pi} = (\sum_i y_i)(\sum_i n_i)$. The Pearson statistic for ungrouped data is

$$\begin{aligned} X^2 &= \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} \frac{(y_{ij}-\hat{\pi})^2}{\hat{\pi}} + \frac{[1-y_{ij}-(1-\hat{\pi})]^2}{1-\hat{\pi}} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} \frac{(y_{ij}-\hat{\pi})^2}{\hat{\pi}(1-\hat{\pi})} = \frac{N\hat{\pi}(1-\hat{\pi})}{\hat{\pi}(1-\hat{\pi})} = N, \end{aligned}$$

Because $X^2 = N$, the statistic is completely uninformative.

5.16 **a.** Treating the data as N binomial observations and letting $s_i = \sum_{j=1}^{n_i} y_{ij}$, the kernel of the log likelihood (ignoring the binomial coefficients) is

$$L(\boldsymbol{\pi}) = \sum_{i=1}^{N} s_i \log(\pi_i) + (n_i - s_i)\log(1-\pi_i).$$

Treating the data as $n = \sum_{i=1}^{N} n_i$ Bernoulli observations, the log likelihood is

$$\begin{aligned} L(\boldsymbol{\pi}) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} y_{ij}\log(\pi_i) + (1-y_{ij})\log(1-\pi_i) \\ &= \sum_{i=1}^{N} s_i \log(\pi_i) + (n_i - s_i)\log(1-\pi_i). \end{aligned}$$

b. For the saturated model case, the two data forms differ. Treating the data as N binomial observations, there are N parameters $\pi_1, \dots, \pi_N$. Treating the data as n Bernoulli observations, there are n parameters, $\{\pi_{ij}\}$.

c. The difference between deviances of two unsaturated models does not depend on the form of data entry because the log likelihood of the saturated model cancels out when taking the difference between deviances. It depends only on the log likelihoods of the unsaturated models, which from (a) do not depend on the form of data entry.

5.17 **a.** For the ungrouped case, the deviance for M_0 is 16.3 and the deviance for M_1 is 11.0. For the ungrouped case, the deviance for M_0 is 6.3 and the deviance for M_1 is 1.0. The saturated model in the ungrouped case has 12 parameters and the log likelihood of the saturated model is 0 while the saturated model in the grouped case has three parameters.

b. The differences between the deviances is the same ($16.3 - 11.0 = 6.3 - 1.0 = 5.3$). The log likelihoods for the grouped and ungrouped cases only differ by the binomial coefficients. The difference between deviances is double the difference in the log likelihoods. The difference between the log likelihoods for either case is

$$\begin{aligned} D_0 - D_1 &= -2[L(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y})] + 2[L(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y})] \\ &= 2[L(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}) - L(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y})]. \end{aligned}$$

For the grouped case, the binomial coefficients cancel out.

5.22 **a.**

$$\begin{aligned} P(y=1) &= P(U_1 > U_0) = P(\beta_{10} + \beta_{11}x + \epsilon_1 > \beta_{00} + \beta_{01}x + \epsilon_0) \\ &= P\left\{(1/\sqrt{2})(\epsilon_0 - \epsilon_1) < (1/\sqrt{2})[\beta_{10} - \beta_{00} + (\beta_{11} - \beta_{01})x]\right\} \end{aligned} \tag{11.2}$$

$= \Phi(\beta_0' + \beta_1' x)$ with $\beta_0' = (1/\sqrt{2})(\beta_{10} - \beta_{00})$ and $\beta_1' = (1/\sqrt{2})(\beta_{11} - \beta_{01})$.

Chapter 6

6.4

$$\frac{\partial \pi_3(x)}{\partial x} = -\frac{[\beta_1 \exp(\alpha_1 + \beta_1 x) + \beta_2 \exp(\alpha_2 + \beta_2 x)]}{[1 + \exp(\alpha_1 + \beta_1 x) + \exp(\alpha_2 + \beta_2 x)]^2}.$$

a. Denominator > 0 and numerator < 0 when $\beta_1 > 0$ and $\beta_2 > 0$.

6.6 **b.**

$$\begin{aligned}\pi_{ij} &= P(u_{ij} > u_{ik}\ \forall j \neq k) = E[P(u_{ij} > u_{ik}\ \forall j \neq k \mid u_{ij})] \\ &= E\left[\prod_{k\neq j} P(\alpha_k + x_i\beta_k + \epsilon_{ik} < u_{ij} \mid u_{ij})\right] = E\left[\prod_{k\neq j} \Phi(u_{ij} - \alpha_k - x_i\beta_k) \mid u_{ij}\right] \\ &= \int \phi(u_{ij} - \alpha_j - x_i\beta_j) \prod_{k\neq j} \Phi(u_{ij} - \alpha_k - x_i\beta_k) du_{ij}\end{aligned}$$

We can form the likelihood by noting that $\ell = \prod_{i=1}^{N} \prod_{j=1}^{c} \pi_{ij}^{y_{ij}}$.

6.8 For a baseline-category logit model with $\beta_j = j\beta$,

$$\frac{P(y_i = j+1 \mid x_i = u)}{P(y_i = j \mid x_i = u)} = e^{u(j+1)\beta}/e^{uj\beta} = e^{u\beta}.$$

Thus, the odds ratio comparing $x_i = u$ versus $x_i = v$ is

$$\frac{P(y_i = j+1 \mid x_i = u)}{P(y_i = j \mid x_i = u)} \Bigg/ \frac{P(y_i = j+1 \mid x_i = v)}{P(y_i = j \mid x_i = v)} = e^{(u-v)\beta}.$$

Note that the odds ratio does not depend on j (i.e., proportional odds structure for adjacent-category logits).

6.9 See Agresti (2013, Section 16.5) for details.

6.13 **a.** For $j < k$, $\text{logit}[P(y \leq j \mid x = x_i)] - \text{logit}[P(y \leq k \mid x = x_i)] = (\alpha_j - \alpha_k) + (\beta_j - \beta_k)x_i$. This difference of logits cannot be positive since $P(y \leq j) \leq P(y \leq k)$; however, if $\beta_j > \beta_k$ then the difference is positive for large positive x_i, and if $\beta_j < \beta_k$ then the difference is positive for large negative x_i.

b. You need monotone increasing $\{\alpha_j + \beta_j\}$.

6.16 **d.** *Hint*: Show that $\text{var}(T)/\text{var}(S)$ is a squared correlation between two random variables, where with probability π_j the first equals b_j and the second equals $f_j'(\theta)/f_j(\theta)$.

Chapter 7

7.4 $2\sum_i y_i \log(y_i/\bar{y})$, chi-squared with $df = 1$ when μ_1 and μ_2 are large.

7.7 **a.** Use that under H_0, conditional on n the data have a multinomial distribution with equal probabilities.

7.9 For this model, $\partial \mu_i / \partial x_{ij} = \beta_j \mu_i$ and the likelihood equation for the intercept is $\sum_i \mu_i = \sum_i y_i$.

7.11 $\hat{\mu}_{ijk} = n_{i+k} n_{+jk} / n_{++k}$, the same as applying the ordinary independence model to each partial table. The residual $df = \ell(r-1)(c-1)$.

7.16 Baseline-category logit model with additive factor effects for B and C.

7.21 See Greenwood and Yule (1920).

7.23 **a.**

$$f(y;\mu,k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k$$
$$= \exp\left[y \log \frac{\mu}{\mu+k} + k \log \frac{k}{\mu+k} + \log\Gamma(y+k) - \log\Gamma(k) + \log\Gamma(y+1)\right]$$

Let $\theta = \log[\mu/(\mu+k)]$, $b(\theta) = -\log(1-e^{\theta})$, and $a(\phi) = 1/k$.

b. Letting $x = ya(\phi)$,

$$f(x;\mu,k) = \exp\left\{ \frac{x \log[\mu/(\mu+k)] + \log[k/(\mu+k)]}{1/k} \right.$$
$$\left. + \log\Gamma(y+k) - \log\Gamma(k) + \log\Gamma(y+1) + \log k \right\}$$

where the $\log k$ at the end of the equation is the Jacobian.

7.25 The likelihood is proportional to $[k/(\mu+k)]^{nk}[\mu/(\mu+k)]^{\sum_i y_i}$. The log likelihood depends on μ through

$$-nk \log(\mu+k) + \sum_i y_i[\log \mu - \log(\mu+k)].$$

Differentiating with respect to μ, setting equal to 0, and solving for μ yields $\hat{\mu} = \bar{y}$.

7.27 From including the GR term, the likelihood equations imply that the fitted GR marginal totals equal the sample values. For example, the sample had 1040 white females, and necessarily the fitted model will have 1040 white females. The model with AC, AM, CM, AG, AR, GM, GR two-factor terms and no three-factor interaction terms fits well ($G^2 = 19.9$, $df = 19$).

Chapter 8

8.2 For a beta-binomial random variable s, $\text{var}(s) = E[n\pi(1-\pi)] + \text{var}(n\pi) = nE(\pi) - nE(\pi^2) + n[(E\pi)^2 - (E\pi)^2] + n^2\text{var}(\pi) = nE(\pi)[1 - E(\pi)] + n(n-1)\text{var}(\pi) = n\mu(1-\mu) + n(n-1)\mu(1-\mu)\theta/(1+\theta) = n\mu(1-\mu)[1 + (n-1)\theta/(1+\theta)]$.

8.5 **a.** If $\text{logit}(y_i) = \beta_i + \sigma z$, then $y_i = (e^{\beta_i+\sigma z})/(1 + e^{\beta_i+\sigma z})$. Taking $f(\sigma) = (e^{\beta_i+\sigma z})/(1 + e^{\beta_i+\sigma z})$ and expanding $f(\sigma)$ around $f(0)$ by Taylor approximation,

$$y_i = \frac{e^{\beta_i}}{1+e^{\beta_i}} + \frac{e^{\beta_i}}{1+e^{\beta_i}}\frac{1}{1+e^{\beta_i}}\sigma z + \frac{e^{\beta_i}(1-e^{\beta_i})}{2(1+e^{\beta_i})^3}\sigma^2 z^2 + \cdots$$

b. Using this approximation and the fact that $E(z) = 0$ and $\text{var}(z) = 1$, we have $E(y_i) \approx \mu_i$ and $\text{var}(y_i) \approx [\mu_i(1-\mu_i)]^2\sigma^2$.

c. The binomial approximation would imply that for a single region $v(\mu_i) = \phi\mu_i(1-\mu_i)$. This approach is inappropriate when $n_i = 1$ since in that case $\phi = 1$. Regardless of n_i, the binomial distribution assumes the small regions are independent, but contiguous regions would likely have dependent results.

8.8 For the null model $\mu_i = \beta$ and $v(\mu_i) = \sigma^2$,

$$u(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} v(\mu_i)^{-1}(y_i - \mu_i) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\sigma^2}.$$

Thus $\hat{\beta} = \bar{y}$ and the variance of $\hat{\beta}$ is $V = \sigma^2/n$. A sensible model-based estimate of V is $\hat{V} = (1/n^2)\sum_{i=1}^{n}(y_i - \bar{y})^2$. The actual asymptotic variance of $\hat{\beta}$ is

$$V\left[\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{\text{var}(y_i)}{v(\mu_i)^2} \frac{\partial \mu_i}{\partial \beta}\right] V = \frac{\sigma^2}{n}\left(\sum_{i=1}^{n} \frac{\beta}{\sigma^4}\right)\frac{\sigma^2}{n} = \frac{\beta}{n}.$$

To find the robust estimate of the variance that adjusts for model misspecification, we replace $\text{var}(y_i)$ in the expression above with $(y_i - \bar{y})^2$, leading to $[\sum_{i=1}^{n}(y_i - \bar{y})^2]/n^2$.

8.11 The model-based estimator tends to be better when the model holds, and the robust estimator tends to be better when there is severe overdispersion so that the model-based estimator tends to underestimate the actual *SE*.

8.17 **b.** *Hint*: Is it realistic to treat the success probability as identical from shot to shot?

Chapter 9

9.3 *Hint*: Use (2.7) with $\boldsymbol{A} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$, for which $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu} = 0$.

9.8 *Hint*: See Exercise 9.8 for correlations for the autoregressive structure.

9.21 *Hints:* The covariance is the same for any pair of cells in the same row, and $\mathrm{var}(\sum_j y_{ij}) = 0$ since y_{i+} is fixed. If $(x_1, \dots, x_d)$ is multivariate normal with common mean and common variance σ^2 and common correlation ρ for pairs (x_j, x_k), then $[\sum_j (x_j - \bar{x})^2]/\sigma^2(1-\rho)$ is chi-squared with $df = (d-1)$.

9.19 **b.** Given $S_i = 0$, $P(y_{i1} = y_{i2} = 0) = 1$. Given $S_i = 2$, $P(y_{i1} = y_{i2} = 1) = 1$. Given $y_{i1} + y_{i2} = 1$,

$$\begin{aligned} P(y_{i1}, y_{i2} \mid S_i = 1) &= \exp(\beta_1)/[1+\exp(\beta_1)], \quad y_{i1} = 0, \quad y_{i2} = 1 \\ &= 1/[1+\exp(\beta_1)], \quad y_{i1} = 1, \quad y_{i2} = 0. \end{aligned}$$

9.20 **a.** *Hint:* Apply the law of large numbers due to A. A. Markov for independent but not identically distributed random variables, or use Chebyshev's inequality.

9.22 $P(y_{ij} = 1 \mid \boldsymbol{u}_i) = \Phi(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i)$, so

$$P(y_{ij} = 1) = \int P(z \le \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i) f(\boldsymbol{u}; \boldsymbol{\Sigma}) d\boldsymbol{u}_i,$$

where z is a standard normal variate that is independent of $\boldsymbol{u}_i$. Since $z - \boldsymbol{z}_{ij}\boldsymbol{u}_i$ has an $N(0, 1 + \boldsymbol{z}_{ij}\boldsymbol{\Sigma}\boldsymbol{z}_{ij}^{\mathrm{T}})$ distribution, the probability in the integrand is $\Phi(\boldsymbol{x}_{ij}\boldsymbol{\beta}[1 + \boldsymbol{z}_{ij}\boldsymbol{\Sigma}\boldsymbol{z}_{ij}^{\mathrm{T}}]^{-1/2})$, which does not depend on u_i, so the integral is the same. The parameters in the marginal model equal those in the GLMM divided by $[1 + \boldsymbol{z}_{ij}\boldsymbol{\Sigma}\boldsymbol{z}_{ij}^{\mathrm{T}}]^{1/2}$, which in the univariate case is $\sqrt{1+\sigma^2}$.

9.24

$$\begin{aligned} \mathrm{cov}(y_{ij}, y_{ik}) &= E[\mathrm{cov}(y_{ij}, y_{ik} \mid u_i)] + \mathrm{cov}[E(y_{ij} \mid u_i), E(y_{ik} \mid u_i)] \\ &= 0 + \mathrm{cov}[\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + u_i), \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + u_i)]. \end{aligned}$$

The functions in the last covariance term are both monotone increasing functions of u_i, and hence are nonnegatively correlated.

9.31 *Hint*: See Diggle et al. (2002, Sec. 4.6).

Chapter 10

10.1 Given $\sum y_i = n$, $(y_1, \ldots, y_c)$ have a multinomial distribution for n trials with probabilities $\{\pi_i = 1/c\}$, and y_i has a binomial distribution with index n and parameter $\pi = 1/c$. The $\{y_i\}$ are exchangeable but not independent.

10.4 Normal with mean

$$\left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}}\right)$$

and covariance matrix $(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}$.

10.5 $E(\sigma^2 \mid \boldsymbol{y}) = ws^2 + (1-w)E(\sigma^2)$, where $w = (n-p)/(n-p+\nu_0-2)$.

10.9

$$E(\tilde{\pi} - \pi)^2 = \left(\frac{n}{n+n^*}\right)^2 \frac{\pi(1-\pi)}{n} + \left(\frac{n^*}{n+n^*}\right)^2 (\mu - \pi)^2.$$

10.11 **a.** ML estimate = 0, confidence interval = (0.0, 0.074).

b. Posterior mean = 1/27 = 0.037, posterior 95% equal-tail interval is (0.001, 0.132), 95% HPD interval is (0, 0.109) where 0.109 is 95th percentile of beta(1, 26) density.

Chapter 11

11.6 $\boldsymbol{X} = \boldsymbol{I}_n$, so

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{y}/(1+\lambda).$$

This has greater shrinkage of the ML estimate $\hat{\boldsymbol{\beta}} = \boldsymbol{y}$ as λ increases.

11.9 *Hint*: Can the Dirichlet recognize ordered categories, such as higher correlation between probabilities closer together? Can it recognize hierarchical structure?

11.10 **a.** $\tilde{\pi}(x)$ converges to the overall sample proportion, $\hat{\pi} = (\sum_i y_i)/n$, and the estimated asymptotic variance is approximately $\hat{\pi}(1-\hat{\pi})/n$.

11.11 *Hint*: At the first stage, can you fit the model with ordinary least squares?