Solution: Sample Final

1. I.

(a)

$$\begin{aligned} E(Y_{ij}) &= \mu + \gamma t_j, Var(Y_{ij}) = Var(\rho_i) + Var(\varepsilon_{ij}) = \sigma_\rho^2 + \sigma^2, \\ Cov(Y_{ij}, Y_{ij'}) &= \sigma_\rho^2, j \neq j', \\ Corr(Y_{ij}, Y_{ij'}) &= \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma^2}, j \neq j'. \end{aligned}$$

(b) Here

$$\begin{aligned} E(Y_{ij}) &= \mu + \gamma t_j, \\ Var(Y_{ij}) &= Var(\rho_i) + Var(\gamma_{1i} t_j) + Var(\varepsilon_{ij}) \\ &= \sigma_\rho^2 + t_j^2 \sigma_1^2 + \sigma^2. \\ Cov(Y_{ij}, Y_{ij'}) &= Var(\rho_i) + Cov(\gamma_{1i} t_j, \gamma_{1i} t_{j'}) \\ &= \sigma_\rho^2 + t_j t_{j'} \sigma_1^2, j \neq j', \\ Corr(Y_{ij}, Y_{ij'}) &= \frac{\sigma_\rho^2 + t_j t_{j'} \sigma_1^2}{\sqrt{(\sigma_\rho^2 + t_j^2 \sigma_1^2 + \sigma^2)(\sigma_\rho^2 + t_{j'}^2 \sigma_1^2 + \sigma^2)}}, j \neq j'. \end{aligned}$$

II. Solution not given since Poisson Regression has not been covered in the class.
III. Note that

$$L(k) = E\left[||X\hat{\beta}(k) - X\beta||^2 \right] = L_1(k) + L_2(k).$$

The most appropriate value of k is the one at which $L(\cdot)$ achieves its minimum. Thus we have

k	0	2.5	5	7.5	10
$L(k)$	12	5.88	5.60	5.97	6.56

The minimum is achieved at $k = 5$. Thus among the given choices, $k = 5$ is most appropriate penalty. Since $k = 0$ corresponds to the least squares estimate, the ridge estimate at $k = 5$ is better than the one given by the least squares method.

2.
(a) This is a repeated measures study with $s = 4$ women (subjects) and $r = 5$ weeks. If Y_{ij} is the observation (hair gain) for the i^{th} subject in the j^{th} week, then the model is

$$Y_{ij} = \mu + \rho_i + \tau_j + \varepsilon_{ij}, j = 1, \ldots, r, i = 1, \ldots, s,$$

where μ is the overall mean, $\{\rho_i\}$ are the subject effects (random), $\{\tau_j\}$ are the week effects (fixed) satisfying the constraint $\sum \tau_j = 0$, and $\{\varepsilon_{ij}\}$ are the iid $N(0, \sigma^2)$ errors. Here $\{\rho_i\}$ are iid $N(0, \sigma_\rho^2)$, and $\{\rho_i\}$ and $\{\varepsilon_{ij}\}$ are mutually independent.

(b) ANOVA table

Source	df	SS	MS	F	p-val
Subject	$s-1=3$	54091	18030.333	34.674	< 0.001
Week	$r-1=4$	11683	2920.75	5.617	between 0.005 and 0.01
Error	$(s-1)(r-1)=12$	6240	520		
Total	$rs-1=19$	72014			

(c) Estimate of σ^2_ρ is

$$s^2_\rho = \frac{MSA - MSE}{r} = \frac{18030.333 - 520}{5} = 3502.0667$$

Estimate of the proportion variability in hair gain that can be explained by variability among subjects is

$$\frac{s^2_\rho}{s^2_\rho + MSE} = \frac{3502.0667}{3502.0667 + 520} = 0.8707$$

(d) Estimates of $D_1 = \mu_2 - \mu_1, D_2 = \mu_3 - \mu_2, D_3 = \mu_4 - \mu_3$ are

$$\begin{aligned}\hat{D}_1 &= \bar{Y}_{\cdot 2} - \bar{Y}_{\cdot 1} = 189.75 - 164.50 = 25.25, \\ \hat{D}_2 &= \bar{Y}_{\cdot 3} - \bar{Y}_{\cdot 2} = 227.00 - 189.75 = 37.25 \\ \hat{D}_3 &= \bar{Y}_{\cdot 4} - \bar{Y}_{\cdot 3} = 217.00 - 227.00 = -10.00.\end{aligned}$$

We also have

$$s^2(\hat{D}_j) = \frac{2}{s} MSE = \frac{2}{4}(520) = 260, s(\hat{D}_j) = 16.1245.$$

Since $t(0.98; 12) = 2.303$ and $t(0.095; 12) = 2.461$, Bonferroni multiplier is (by interpolation)

$$B = t\left(1 - \frac{0.1}{(2)(3)}; 12\right) = t(0.9833; 12) \approx 2.407.$$

Simultaneous 95% confidence intervals for D_j's are $\hat{D}_j \pm Bs(\hat{D}_j), i.e., \hat{D}_j \pm 38.812, j = 1, 2, 3$, i.e.,

$$\begin{aligned}D_1 &: \quad 25.25 \pm 38.812, i.e., (-13.562, 64.062), \\ D_2 &: \quad 37.25 \pm 38.812, i.e., (-1.562, 76.062), \\ D_1 &: \quad -10.00 \pm 38.812, i.e., (-48.812, 28.812)..\end{aligned}$$

Since all the three intervals include zero. there is no convincing evidence that the mean hair gain in weeks j and $j+1$ are different, $j = 1, 2, 3$.

(e) Here $H_0 : \theta = 0$ against $H_1 : \theta > 0$, $\alpha = 0.05$.

We first note that θ is a contrast and thus calculation of $s^2(\hat{\theta})$ requires MSE and $\hat{\theta}/s(\hat{\theta}) \sim t_{12}$ under H_0.

Decision rule: reject H_0 if $t^* = \hat{\theta}/s(\hat{\theta}) > t(0.95; 12) = 1.782$.

Note that

$$\hat{\theta} = (\bar{Y}_{\cdot 2} + \bar{Y}_{\cdot 3} + \bar{Y}_{\cdot 4} + \bar{Y}_{\cdot 5})/4 - (\bar{Y}_{\cdot 2} + \bar{Y}_{\cdot 3})/2 = 6.4375.$$

It has been noted already that θ is a contrast in $\{\mu_j\}$ and it can be written as

$$\theta = (-1/4)\mu_2 + (-1/4)\mu_3 + (1/4)\mu_4 + (1/4)\mu_5.$$

Thus

$$\begin{aligned} s^2(\hat{\theta}) &= \left[(-1/4)^2 + (-1/4)^2 + (1/4)^2 + (1/4)^2\right] \frac{MSE}{s} = 32.5 \\ s(\hat{\theta}) &= 5.7009. \end{aligned}$$

Thus $t^* = \hat{\theta}/s(\hat{\theta}) = 1.129$. Since $t^* < t(0.95; 12)$, we cannot reject H_0. Thus there is no strong evidence to conclude that the mean hair gains in weeks 8 through 32 is larger than the mean hair gain in weeks 8 and 16.

(f) This is a three factor model in which there are $s = 8$ subjects, $r = 5$ weeks and two treatments (placebo(1) and Rogaine(2)). Let Y_{ijk} be the hair gain for the i^{th} woman in the jth week under the k^{th} treatment. Then the model is

$$Y_{ijk} = \mu + \rho_i + \tau_j + \gamma_k + (\tau\gamma)_{jk} + \varepsilon_{ijk}, k = 1, 2, j = 1, \ldots, 5, i = 1, \ldots, 8,$$

where μ is the overall mean, $\{\rho_i\}$ are the subject effects (random), $\{\tau_j\}$ are the week effects (fixed), $\{\gamma_h\}$ are the treatment effects (fixed), $\{(\tau\gamma)_{jk}\}$ are the interaction effects (between week and treatment) and $\{\varepsilon_{ijk}\}$ are iid $N(0, \sigma^2)$ errors. Here, $\sum \tau_j = 0$, $\sum \gamma_k = 0$, $\sum_j (\tau\gamma)_{jk} = 0$ for all k, $\sum_k (\tau\gamma)_{ijk} = 0$ for all j, $\{\rho_i\}$ are iid $N(0, \sigma_\rho^2\}$, and $\{\rho_i\}$ and $\{\varepsilon_{ijk}\}$ are mutually independent.

3.

(a) To test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$, the z-statistic is $z^* = b_j/s(b_j)$. The following are the z-statistics

	X_1	X_2	X_3	X_4	X_5
z^*	1.5884	1.2776	0.2467	-0.3500	0.3581

Variable X_3 is the bast candidate for deletion since the magnitude of its associated z-value is the smallest.

(b) $H_0 : \beta_4 = \beta_5 = 0$ vs H$_1$: not both β_4 and β_5 are 0.

The full model has $\log[\pi_i/(1 - \pi_i)] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_5 X_{i5}$ and the reduced model has $\log[\pi_i/(1 - \pi_i)] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$.

The test statistic is

$$\begin{aligned} G^2 &= -2[\log L(R) - \log L(F)] = [-2 \log L(R)] - [-2 \log L(F)] \\ &= 463.37 - 463.19 = 0.18. \end{aligned}$$

Degrees of freedom for this chi-square test is

$$\begin{aligned} & \text{\# of estimated parameters in the full model} \\ & \text{- \# of estimated parameters in the reduced model} \\ = \ & 6 - 4 = 2. \end{aligned}$$

Decision rule: reject H_0 if $G^2 > \chi^2(0.95; 2) = 5.99$

Since $G^2 < \chi^2(0.95; 2)$, we cannot reject H_0. Conclusion: we may drop variables X_4 and X_5 from the full model.

0.1<p-value<0.95. $[\chi^2(0.050; 2) = 0.103, \chi^2(0.90; 2) = 0.211.]$

(c) The Bonferroni multiplier is

$$B = z\left(1 - \frac{0.05}{(2)(3)}\right) = z(0.9917) = 2.395$$

Simultaneous 95% confidence intervals are $b_j \pm Bs(b_j), j = 1, 2, 3,$

$$\begin{aligned}
\beta_1 &: 0.2280 \pm (2.395)(0.1085), i.e, 0.2280 \pm 0.2599, i.e.(-0.032, 0.488),\\
\beta_2 &: 0.8641 \pm (2.395)(0.3276), i.e, 0.8641 \pm 0.7845. i.e.(0.080, 1.649),\\
\beta_3 &: 0.9383 \pm (2.395)(0.2230), i.e, 0.9383 \pm 0.5579. i.e., (0.380, 1.496).
\end{aligned}$$

(d) When $(X_1, X_2, X_3) = (6.5, 3.2, 1)$,then

$$\begin{aligned}
& b_0 + b_1X_1 + b_2X_2 + b_3X_3\\
= & -5.61799 + (0.2280)(6.5) + (0.86406)(3.2) + (0.93832)(1)\\
= & -0.432678.\\
\hat{\pi} = & \frac{\exp(-0.432678)}{1 + \exp(-0.432678)} = 0.39349.
\end{aligned}$$

Odds of admission at (X_1, X_2, X_3) is

$$\theta = \frac{\pi}{1 - \pi} = \exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3).$$

Thus

$$\hat{\theta} = \exp(-0.432678) = 0.6487694.$$

If $[L, U]$ is a 95% confidence interval for π, then a 95% confidence interval for $\theta = \pi/(1 - \pi)$ is $[L/(1 - L), U/(1 - U)]$.

Since $s(\hat{\pi}) = 0.04122$, a 95% confidence interval for π is

$$\begin{aligned}
& \hat{\pi} \pm 1.96s(\hat{\pi}), i.e., 0.39349 \pm (1.96)(0.04122), i.e., 0.39349 \pm 0.08079,\\
& i.e., (0.31270, 0.47428).
\end{aligned}$$

Thus a 95% confidence interval for θ is

$$[L/(1 - L), U/(1 - U)] = [0.4550, 0.9022].$$

(e) Note that the odds ratio is $\theta_1/\theta_0 = \exp(\beta_3)$. Estimate for the odds ratio is $\exp(b_3) = \exp(0.93832)$ =2.555684. A 95% confidence interval for β_3 is

$$\begin{aligned}
& b_3 \pm 1.95s(b_3), i.e., 0.93832 \pm (1.96)(0.23295)\\
& i.e. 0.93832 \pm 0.456582, i.e., (0.481738, 1.394902)
\end{aligned}$$

Thus a 95% confidence interval for the odds ratio θ_1/θ_0 is

$$[\exp(0.481738), \exp(1.394902)] = [1.6189, 4.0346].$$

(f) When GRE=X_1, GPA=X_2 and $X_3 = 1$,
$X_4 = X_1X_3 = X_1$ and $X_5 = X_2X_3 = X_2$, and hence

$$\begin{aligned}\theta_1 &= \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 + \beta_4 X_1 + \beta_5 X_2)\\ &= \exp(\beta_0 + \beta_3 + (\beta_1 + \beta_4)X_1 + (\beta_2 + \beta_5)X_2).\end{aligned}$$

Similarly, when GRE=X_1, GPA=X_2 and $X_3 = 0$,
$X_4 = X_1X_3 = 0$ and $X_5 = X_2X_3 = 0$, and hence

$$\theta_0 = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2).$$

Thus the odds ratio is

$$\theta_1/\theta_0 = \exp(\beta_3 + \beta_4 X_1 + \beta_5 X_2).$$

In order for this ratio not to depend on X_1, X_2 we must have $\beta_4 = \beta_5 = 0$.