Below are model solutions and additional comments discussing common mistakes. In terms of revision, it would be a very good return on your to time track down lost marks and streamline your solutions.

The PCA question was generally answered quite poorly.

QUESTION 1

a. The model being fitted is written as follows

$$Y_{ij} \sim \text{Bin}(\pi_{ij}, n_{ij}), \tag{1}$$

where Y_{ij} is the number of cases of disease for children of gender type $i \in \{1, 2\}$ and feeding status $j \in \{1, 2, 3\}$. π_{ij} is the probability of disease and is mapped to a linear predictor η_{ij} according to the logit link function, which implies

$$\pi_{ij} = e^{\eta_{ij}}/(1 + e^{\eta_{ij}}). \tag{2}$$

The linear predictor takes the form

$$\eta_{ij} = \mu + \alpha_i + \gamma_j, \tag{3}$$

where $\alpha_1 = \gamma_1 = 0$.

Some comments: It is also possible to get full marks using a single subscript $i \in \{1, 2, \ldots, 6\}$ and using x variables in the linear predictor: $\eta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$. If you go down this route, you need to define the x variables. See how in both formulations it is possible to count the number of paramters $p = 4$ and the number of datapoints $n = 6$.

b. Using the R printout of the estimated coefficients and their standard errors: in the presence of feeding status, gender is significant at the 5% level; in the presence of gender, breast feeding is significantly different from bottle feeding (at the 0.1% level); and in the presence of gender, supplement feeding isn't significantly different - at the 10% level - to bottle feeding.

Comments: It is possible to use the anova table to show that the variable describing feeding status is significant in the presence of the gender variable. But you can't show that gender is still significant in the presence of feeding, nor can you drill down to the individual significance of the parameters.

c. $D/\phi \sim \chi^2_2$ under the assumption the model fits. Hence, if the model fits, $\mathbb{E}(D/\phi) = 2$. Consequently since $D/\phi \approx 0.72 < 2$ this model will pass a goodness of fit test at the 95% level. (Comment: To understand why, run this code:

```
df = 1:100
critical.point = qchisq(0.95,df)
plot(df,critical.point)
abline(a=0,b=1)
```

.)

We also note that the percentage of deviance $(100 \times (26.4 - 0.72)/26.4)$ explained is about 97%. This is a strong hint the model fits. We would want to check the residuals to be certain the model does indeed fit however.

Note: If you run the above code you can see why in cases where D/ϕ is only a little bit more than $n-p$ we may need to carry out a goodness of fit on the computer (or using statistics tables) just to see what side of the quantile our observed test statistics lies. If D/ϕ far exceeds $n - p$ then the plot shows that we can be almost certain the model doesn't fit (at the 95% confidence level).

d.

```
> 1-pchisq(0.72192,2)
[1] 0.6970069
```

Hence the model does indeed fit (subject to residual checking).

Comments: It is not appropriate to test whether the change in deviance between the Null and the fitted model $26.4 - 0.72$ is significant. This only tells us whether we can safely remove *all* the variables without significantly altering the fit. But this doesn't tell us if the model fits, only whether the Null model is as good as the fitted model.

e. Using the notation in question **a**, $\hat{\gamma}_2 \sim \text{Normal}(\gamma_2, \sigma^2)$, hence $\hat{\gamma}_2$ plus/minus $1.96\hat{\sigma}$ is a 95% CI for γ_2 ($\hat{\sigma} = 0.153$). To interpret note, $\hat{\gamma}_2 = -0.67$, thus suggesting that breast feeding reduces the probability of disease relative to bottle feeding.

Note: Answers based on odds are also good - e.g. $\exp(\hat{\gamma}_2$ plus/minus $1.96\hat{\sigma})$ being a 95% CI for the odds multiplier.

f. $D/(n-p) \approx 0.7/2 = 0.35$ or 1 since $\phi = 1$ for the Binomial.

My comments: note that since in general (i.e. for all glms) $\mathbb{E}(D/(n-p)) = \phi$, for the Binomial $\mathbb{E}(D/(n-p)) = 1$.

g. The probit has the advantage of adhering to the constraints on the range of values for the mean ($[0,1]$). Whereas the identity link could produce estimated probabilities outside $[0,1]$.

My comments: It might be that the identity link is actually better though. This could be described by saying that *locally* the probability of success is roughly linear in the inputs x.

h. This model is saturated hence $D = 0$ and $d = 0$. Justification: the deviance is defined as $D = -2\phi \log \frac{L(\hat{\beta})}{L(\beta_{\text{sat}})}$. Hence when fitting a saturated model $\hat{\beta} = \beta_{\text{sat}}$ and so $L(\hat{\beta}) = L(\beta_{\text{sat}})$; consequently $D = 0$. And, $d = n - p = n - n = 0$ since the saturated model has as many parameters as data points.

Question 2

a. Let $T = L^T(X - (\mu_1 + \mu_2)/2)$ and let $k = \log \frac{\pi_2 C(1,2)}{\pi_1 C(2,1)}$ then the rule is: if $T > k$ alloc to pop 1 else pop 2. See chapter 11 of notes for the derivation.

b. Using the computer $D^2(\bar{x_1}, \bar{x_2}) = \Delta^T Q \Delta \approx 2.425$, where $\Delta = \bar{x_1} - \bar{x_2}$ is the difference in the sample mean vectors and $Q = S_U^{-1}$ is the inverse of the sample correlation matrix. Now,

$$\mathbb{P}(\text{error}) = \mathbb{P}(T < k \mid X \text{ is from pop 1})\pi_1 + \mathbb{P}(T \geq k \mid X \text{ is from pop 2})\pi_2,$$

and we can estimate π_1 as 37/49 and π_2 as 12/49 from the table given in the question. For k we also need the costs $C(.,.)$, since these arn't given in the question, we assume that they are equal, hence $\hat{k} = \log(12/37)$. If X comes from pop 1, $T \sim N(\alpha/2, \alpha)$ else $T \sim N(-\alpha/2, \alpha)$, where $\alpha = D^2(\mu_1, \mu_2)$. Therefore we can use as our estimator $\hat{\alpha} = 2.425$. Putting all this together we can see that:

$$\mathbb{P}(\text{error}) = \Phi((k - \alpha/2)/\sqrt{\alpha})\pi_1 + \left(1 - \Phi((k + \alpha/2)/\sqrt{\alpha}\right)\pi_2.$$

Plugging in the estimates for α, k, π_1, π_2, it turns out $\mathbb{P}(\text{error})$ is about 17%:

```
> k = log(12/37)
> a = 2.425
> p1 = 37/49
> p2 = 12/49
> pnorm((k-a/2)/sqrt(a))*p1 + (1-pnorm((k+a/2)/sqrt(a)))*p2
```

```
[1] 0.1673056
```

Note: It turns out the question contains a copy/paiste error: if you used `Sp` you get $D^2 = 2.4$, but using `invSp` you get $D^2 = 1.6$. Either answer obtains full scores.

c.

i. When one of the eigenvalues in PCA is precisely 0 it means that the variables are linearly related. Hence we can remove any variable without loss of information.

ii. Recalling the formula: $\text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y)+2\text{Cov}(X,Y)$, and looking at S_U, since $52+73+2\times 36 = 197$, it follows that $X_1 = X_2 + X_3 + \text{constant}$. Hence removing X_1 is the sensible choice as it seems X_2 and X_3 are the natural variables for the problem and X_1 is merely a derived variable. Furthermore X_2 and X_3 have comparable variances, this means we would not have to re-scale the variables when performing pca. Had we carried say X_1 and X_2 instead, we would need to rescale.

Comment: It is also possible to derive that $X_1 = X_2 + X_3 + \text{constant}$, by examining the third principle component loadings. But note the constant term, we can't prove that $X_1 = X_2 + X_3$!

Since we now only have two variables, there are broadly three possible scenarios. If $\text{Cov}(X_2, X_3) \approx 0$, then the Normal distribution is spherical, thus the orthogonal directions of maximal variation will be the standard basis directions $(1, 0)$ and $(0, 1)$ (approx), the order being determined by which of X_2 and X_3 has the larger variance. If $\text{Cov}(X_2, X_3) > 0$, then the Normal dist. is elliptical and is orientated towards the positive half of the $x-$axis, hence the first component will be of the form $(+, +)$ describing performance across both tests together and the second component will be of the form $(+, -)$ contrasting performance across the tests. Finally, if $\text{Cov}(X_2, X_3) < 0$ then the situation is as in the previous one, but the rank order of components reversed. We are in the $\text{Cov}(X_2, X_3) > 0$ scenario in this question. (The above guesses for the pattern in the signs of the components are given only up to multiplication by minus 1.)

Uses for the principle components could be: data visualisation, pre-processing for a regression or to improve our understanding of the ways in which the people differed on the tests (there are others). Regarding the latter and given what we have said above, we expect to be able to reinterpret score 2 and score 3 in terms of *overall attitude across both scores* and *contrasting attitude between the scores.*

Notes: It may help to draw a picture of the $x - y$ plane and sketch what the Normal dist. would roughly look like in the above three scenarios.