1. (a) [4 points] 
\[ h(t) = h_0(t) e^{1.0672 \text{ P27} + .7256 \text{ CYCLINE} + 1.2040 \text{ Nodes} + \ldots + 0.1980 \text{ Year}} \]

(b) [4 points] 
The key assumptions are:
(i) Proportional hazards as described in part (a)
(ii) Event times are mutually independent
(iii) Censoring is uninformative. The censoring process is independent of the time to event process

(c) [8 points] 
The estimated coefficient is 0.8961 with standard error 0.372. To eliminate the effect of the NODES x CYCLINE interaction we must assume that cancer has not spread to the lymph nodes (NODES = 0). Then also holding tumor size, age, year and P27 status constant, \( e^{0.8961} = 2.45 \) is the increase in the hazard of dying from breast cancer for women with abnormal levels of CYCLINE. Given any particular tumor size category, age category, year of diagnosis category and P27 level category, women with an abnormal level of CYCLINE have about a 2.45 greater risk of dying from breast cancer than women with normal levels of CYCLINE, provided that the cancer has not spread to the lymph nodes. An approximate 95% confidence interval for this hazard ratio is obtained as follows:
\[ .8961 \pm (1.96)(.372) \Rightarrow (0.167, 1.625) \]
Apply the exponential function to each end of the interval to obtain (1.18, 5.08).

(d) [4 points] 
At any point in time, breast cancer patients with abnormal levels of P27 have approximately 3 times greater hazard of dying \( e^{1.0884} = 2.97 \) than breast cancer patients with normal levels of P27, if the cancer has not spread to lymph nodes. If the cancer has spread to lymph nodes, an abnormal P27 level increases the hazard by a factor of about \( e^{1.0884 - 0.0274} = e^{1.061} = 2.89 \). Since the coefficient on NODES * P27 is not significantly different from zero, the data suggest that risk of dying is increased by about a factor of 3 when the P27 level is abnormal, regardless of whether or not the cancer has spread to the lymph nodes. Of course, the hazard of dying is further increased by a factor of \( e^{1.3577} = 3.89 \) if the cancer spreads to the lymph nodes.

(e) [4 points] 
To test the null hypothesis that the coefficients are zero for both the NODES * P27 and NODES * CYCLINE interactions, compare the values of the logarithms of the partial likelihoods. Do not reject the null hypothesis because
\[ d^2 = (836.111 - 835.797) = 0.314 < \chi^2_{2,0.5} = 5.99 \]
Alternatively, one could perform a Wald tests by rejecting the null hypothesis if
Many students used the p-values listed in the second table on page 2 of the exam to provide a separate test for each parameter. One should comment on controlling the overall type I error level when performing two separate tests. Colinearity is another potential complication. If there is a strong correlation between \( \hat{\beta}_{\text{nodes} \times \text{P27}} \) and \( \hat{\beta}_{\text{nodes} \times \text{cycline}} \), p-values taken from the second table on page 2 of the exam might indicate that neither coefficient was significantly different from zero, but one of the coefficients may become “significant” if the other interaction is deleted from the model.

(f) [4 points] Since \( \hat{S}_0(5) = 0.93 \), then the probability that a female breast cancer patient with tumor size less than 2 cm, age between 25 and 35, who was diagnosed between 1983 and 1988, and who has abnormal levels of both CYCLINE and P27 is

\[
\hat{S}(5) = \left[ \hat{S}_0(5) \right]^{\exp(1.0884+0.8961)} = (.93)^{\exp(1.9845)} = (.93)^{7.2754} = .59
\]

(g) [4 points]

(i) Plot the scaled Schoenfeld residuals for P27 against time, or some monotone function of time. Display 95% confidence limits on the plot. Deviations from a horizontal line provide evidence against the proportional hazards assumption.

(ii) Estimate the cumulative hazard function \( \hat{H}(t) \) and plot \( \log(\hat{H}(t)) \) against time for each level of the P27 variable. Non-parallel curves provide evidence against the proportional hazards assumption.

(h) [4 points] Include a P27 * log(time) term in the model and obtain the partial likelihood estimate of its coefficient, \( \hat{B} \), and its standard error \( S_B \).

\[
p - \text{value} = 2 \left[ 1 - \Phi \left( \frac{\hat{B}}{S_B} \right) \right]
\]

Alternatively, you could compute \( d^2 = -2 \) (difference in the natural logarithms for the maximized partial likelihoods) and compute

\[
p - \text{value} = \Pr \left\{ x^2_{(1)} > d^2 \right\}
\]

2. (a) [4 points] \( e^{\hat{B}_{0.2}} = e^{-0.8668} = .4203 \) is an estimate of the odds of infection at 4 months for subjects using diet 2. It is not an odds ratio. The estimated probability of infection is .296. An approximate 95% confidence for the
log odds is 
\[-.8668 \pm (1.96)(.3345) \Rightarrow (-1.5224, -.2112).\]
A corresponding 95% confidence interval for the probability of infection after 4 months on diet 2 is (.179, .449).

(b) [8 points] The relative risk can be approximated by an odds ratio
\[
\frac{\text{odds of infection at 12 months}}{\text{odds of infection at 8 months}} = e^{2305(12-4)-.0111(12-4)2-.2305(8-4)+.0111(8-4)2} = e^{3892} = 1.4768
\]

Using diet 2, the risk of infection is about 50% greater at 12 months than at 8 months. An approximate 95% confidence interval is constructed as 
\((e^L, e^U)\) where 
\[L = .3892 - (1.96)S\] and 
\[U = .3892 + (1.96)S\] and 
\[S = \sqrt{\text{a}^T \text{V} \text{a}} = 0.038\] with 
\[\text{a}^T = (0, 0, 0, 4, 0, 0, 48, 0, 0)\] and \(V\) is the covariance matrix shown on page 6 of the exam. Then an approximate 95% confidence interval is \((1.37, 1.59)\).

(c) [4 points] You could construct a Wald test \(X^2 = (A\hat{\beta})^T (AVA)^{-1} (A\hat{\beta})\) where \(V\) is the covariance matrix shown on the bottom of page 6 and 
\[
A = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0
\end{bmatrix}.
\]
There are many choices of \(A\) that yield the same value of \(X^2 = (A\hat{\beta})^T (AVA)^{-1} (A\hat{\beta})\). Reject the null hypothesis if \(X^2 > X^2_{(2), .05} = 5.99\). Alternatively, you could describe a likelihood ratio test.

(d) [6 points] \(\hat{\beta}\) is still a consistent estimator of \(\beta\). Since there are 210 subjects in this study, \(\hat{\beta}\) would be approximately distributed as a normal random vector with expectation \(\hat{\beta}\) and covariance matrix estimated as

\[
\left[\sum_{i,j} D_{ik}^T V_{ik}^{-1} D_{ik}\right]^{-1} = \left[\sum_{i,j} D_{ik}^T V_{ik}^{-1} (Y_{ik} - \hat{\pi}_{ik}) (Y_{ik} - \hat{\pi}_{ik})^T V_{ik}^{-1} D_{ik}\right]^{-1} \left[\sum_{i,j} D_{ik}^T V_{ik}^{-1} D_{ik}\right]^{-1}
\]

(e) [6 points] The estimates on page 6 are obtained by solving the GEE equations 
\[0 = \sum_{i,j} D_{ik}^T V_{ik}^{-1} (Y_{ik} - \pi_{ik})\]
where 
\[V_{ik} = V_{ik}^{1/2} R V_{ik}^{1/2}\]
\(V_{ik}\) is defined in part (d) and \(R\) is a 6 x 6 correlation matrix. Since there are 210 subjects in this study, \(R\) should be well estimated and the GEE
estimates on page 6 should have smaller standard errors than the initial estimates for $\beta$ listed on page 5 of the exam.

3. (a) [6 points] Let $Y_{ij} = [Y_{ij1} \ Y_{ij2} \ Y_{ij3} \ Y_{ij4} \ Y_{ij5}]^T$ denote the vector of measured increases in blood pressure for the $j$-th rabbit in the $i$-th treatment group. The model can be written as

$$Y = \begin{bmatrix} Y_{1,1} \\ \vdots \\ Y_{1,20} \\ Y_{2,1} \\ \vdots \\ Y_{2,20} \end{bmatrix} = \begin{bmatrix} X^* & 0 \\ \vdots & \vdots \\ 0 & X^* \\ \vdots & \vdots \\ 0 & X^* \end{bmatrix} \begin{bmatrix} \beta_{0,1} \\ \vdots \\ \beta_{1,2} \end{bmatrix} + \begin{bmatrix} Z^* & 0 & 0 & \ldots & 0 \\ 0 & Z^* & 0 & \ldots & 0 \\ 0 & 0 & Z^* & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & Z^* \end{bmatrix} \begin{bmatrix} \gamma_{0,1,1} \\ \gamma_{1,1,1} \\ \gamma_{0,1,2} \\ \gamma_{1,1,2} \\ \gamma_{0,2,20} \\ \gamma_{1,2,20} \end{bmatrix} + \varepsilon$$

where $X^* = Z^* = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$

(b) [6 points] Note that $\Sigma = \text{Var}(Y) = Z^T Q Z + R$, where $R = \sigma^2 \mathbf{I}$ and $Q$ is a block diagonal matrix with the block $\begin{bmatrix} \sigma^2 \gamma_0 & \sigma \gamma_0 \gamma_1 \\ \sigma \gamma_0 \gamma_1 & \sigma^2 \gamma_1 \end{bmatrix}$ for each rabbit.

then $\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$

(c) [6 points] Describe REML estimation

(d) [6 points] This estimator is not a linear function of the observed responses, so it is not a best linear unbiased estimator (blue). If the model is correctly specified, this estimator will be consistent as the number of rabbits is increased in each treatment group, and it has a limiting normal distribution with expectation $\beta$ and an estimator of the covariance matrix is
\[
\frac{1}{n_1 + n_2} \left[ X^T \hat{\Sigma}^{-1} X \right]^{-1}
\]
where \( n_i \) is the number of rabbits in the \( i \)-th group,

\[
X_\bullet = \begin{bmatrix}
1 & 1 \\
1 & 2 \\
1 & 3 \\
1 & 4 \\
1 & 5
\end{bmatrix}
\quad \text{and} \quad
\hat{\Sigma} = \begin{bmatrix}
1 & 1 \\
1 & 2 \\
1 & 3 \\
1 & 4 \\
1 & 5
\end{bmatrix}
\begin{bmatrix}
\hat{\sigma}^2_{\gamma_0} & \hat{\sigma}^2_{\gamma_0 \gamma_1} \\
\hat{\sigma}^2_{\gamma_0 \gamma_1} & \hat{\sigma}^2_{\gamma_1}
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 2 & 3 & 4 & 5
\end{bmatrix} + \hat{\sigma}^2_{\varepsilon}
\]

(e) [6 points] AIC = -2 log(REML likelihood) + 2(number of covariance parameters)

Model A: \quad 243.15 + 2(4) = 259.15

Model B; CS \quad 236.14 + 2(2) = 240.14

AR(1) \quad 218.88 + 2(2) = 222.88

Unstructured \quad 214.37 + 2(15) = 244.37

The AIC results for model B suggest that the AR(1) model is adequate. You cannot compare the AIC value for model A with the AIC values for model B because the non-random parts of these two models are not the same.

(f) [4 points] Do not reject the null hypothesis because \( 218.88 - 214.37 = 4.51 \) is smaller than \( \chi^2_{13,0.05} = 22.36 \)

(g) [4 points] You can use a likelihood ratio test because the non-random part of model A is nested in the non-random part of model B and the covariance structure for the random part of model A is nested within the covariance structure for the random part of model B, when the unstructured covariance matrix is used for model B. Model B is shown to provide a significant improvement over model A:

\[
267.35 - 232.18 = 35.17 > \chi^2_{(25-8),0.05} = 27.59
\]
## Final Exam scores

<table>
<thead>
<tr>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>