Time: 2 hours

Name (please print): ________________________________

Show all your work and calculations. Partial credit will be given for work that is partially correct. Points will be deducted for false statements, even if the final answer is correct. Please circle your final answer where appropriate.

This exam is closed-book. You may consult two pages of your personal notes. Calculators are permitted.

Honor code: I promise not to cheat on this exam. I will neither give nor receive any unauthorized assistance. I will not to share information about the exam with anyone who may be taking it at a different time. I have not been told anything about the exam by someone who has taken it earlier.

Signature: ________________________________ Date: __________
1. (40 points) The `motors` data frame reports an accelerated life test at each of four temperatures of 10 motorettes.

<table>
<thead>
<tr>
<th>temp = 150</th>
<th>time</th>
<th>8064</th>
<th>8064</th>
<th>8064</th>
<th>8064</th>
<th>8064</th>
<th>8064</th>
<th>8064</th>
<th>8064</th>
<th>8064</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cens</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>temp = 170</th>
<th>time</th>
<th>1764</th>
<th>2772</th>
<th>3444</th>
<th>3542</th>
<th>3780</th>
<th>4860</th>
<th>5196</th>
<th>5448</th>
<th>5448</th>
<th>5448</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cens</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>temp = 190</th>
<th>time</th>
<th>408</th>
<th>408</th>
<th>1344</th>
<th>1344</th>
<th>1440</th>
<th>1680</th>
<th>1680</th>
<th>1680</th>
<th>1680</th>
<th>1680</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cens</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>temp = 220</th>
<th>time</th>
<th>408</th>
<th>408</th>
<th>504</th>
<th>504</th>
<th>504</th>
<th>528</th>
<th>528</th>
<th>528</th>
<th>528</th>
<th>528</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cens</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(a) Provide the definition of the survival function, and calculate its Kaplan-Meier estimate for the group with `temp = 190`. 
(b) Use the output below to test for homogeneity of groups with \text{temp}=190 and \text{temp}=220. State the null and the alternative hypotheses, describe the test that you use, and state the conclusions. Use $\alpha = 0.05$.

\begin{verbatim}
> survdiff(Surv(time,cens)~factor(temp),subset=(temp>=190),data=motors)
...

<table>
<thead>
<tr>
<th>factor(temp)</th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(temp)=190</td>
<td>10</td>
<td>5</td>
<td>6.5</td>
<td>0.346</td>
<td>1.51</td>
</tr>
<tr>
<td>factor(temp)=220</td>
<td>10</td>
<td>5</td>
<td>3.5</td>
<td>0.643</td>
<td>1.51</td>
</tr>
</tbody>
</table>
...
\end{verbatim}
(c) An accelerated failure time model is fitted to the data as shown below. State the model and the assumptions.

```r
> modw <- survreg(Surv(time,cens)~temp,dist="exponential",data=motors)
> summary(modw)
```

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>18.1879</td>
<td>1.76512</td>
<td>10.30</td>
<td>6.75e-25</td>
</tr>
<tr>
<td>temp</td>
<td>-0.0526</td>
<td>0.00917</td>
<td>-5.73</td>
<td>9.98e-09</td>
</tr>
</tbody>
</table>

(d) Using the model above, predict the probability for a motorette to survive past 3000 at temp=180.
(e) A proportional hazards model is fit to the data as shown below. State the model and the assumptions. What is the effect of improving one unit of temp on the survival of a motorette?

```r
> modph <- coxph(Surv(time,cens) ~ temp, data=motors)
> summary(modph)

... coef exp(coef) se(coef)    z     p temp 0.0919    1.10  0.0274  3.36 0.00079

  exp(coef) exp(-coef) lower .95 upper .95 temp  1.10    0.912   1.04   1.16

...```

(f) Discuss diagnostic plots(s) that you’d use to evaluate the appropriateness of the assumptions of the model above.
2. The following table refers to applicants to graduate school at a university during one admission season. It presents admissions decisions by gender of applicants for the six largest graduate departments.

<table>
<thead>
<tr>
<th>Department</th>
<th>Yes</th>
<th>No</th>
<th>Waiting list</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>A</td>
<td>512</td>
<td>89</td>
<td>313</td>
</tr>
<tr>
<td>B</td>
<td>353</td>
<td>17</td>
<td>207</td>
</tr>
<tr>
<td>C</td>
<td>120</td>
<td>202</td>
<td>205</td>
</tr>
<tr>
<td>D</td>
<td>138</td>
<td>131</td>
<td>279</td>
</tr>
<tr>
<td>E</td>
<td>53</td>
<td>94</td>
<td>138</td>
</tr>
<tr>
<td>F</td>
<td>22</td>
<td>24</td>
<td>351</td>
</tr>
<tr>
<td>Total</td>
<td>1198</td>
<td>557</td>
<td>1493</td>
</tr>
</tbody>
</table>

(a) Since the number of applicants was recorded during a fixed period of time, a Poisson sampling scheme is appropriate.

i. Write in mathematical notation the model based on Poisson sampling, which assumes the mutual independence of department, gender and admission status. Explain how the model implies the assumption of independence.

ii. Write in mathematical notation the model that assumes that the department and gender are not independent, but are joint independent of the admission status. Explain how the dependence and independence structure arise from this model.
iii. The residual deviances of the two models above are 2162.2 for the model in (i), and 895.88 for the model in (ii). Test whether the assumption of 3-way independence is appropriate. State the null and the alternative hypothesis, what test you use, and the conclusions. Use $\alpha = 0.05$.

(b) To study whether admissions vary across gender and department type, we would like to model the probability distribution of admission conditional on gender and department.

i. State what sampling distribution is being considered in this case, and why we can continue using Poisson distribution to fit this model.
ii. The output of a candidate model is given below. Write in mathematical notation the model that is being considered, and the assumptions.

```
glm(formula = freq ~ dpt * sex + admit + admit:(dpt + sex), poisson, data = X)
```

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 3.65389  | 0.11157    | 32.750  | < 2e-16  *** |
| dptB           | -1.43742 | 0.22203    | -6.474  | 9.55e-11  *** |
| dptC           | 2.28782  | 0.11779    | 19.423  | < 2e-16  *** |
| dptD           | 1.84179  | 0.12054    | 15.279  | < 2e-16  *** |
| dptE           | 2.01296  | 0.12229    | 16.461  | < 2e-16  *** |
| dptF           | 2.10455  | 0.12288    | 17.127  | < 2e-16  *** |
| sexM           | 2.02756  | 0.10993    | 18.445  | < 2e-16  *** |
| admitW         | -1.66378 | 0.18532    | -8.978  | < 2e-16  *** |
| admitY         | 0.66438  | 0.09816    | 6.768   | 1.30e-11  *** |
| dptB:sexM      | 1.08299  | 0.21636    | 5.005   | 2.57e-07  *** |
| dptC:sexM      | -2.59672 | 0.12008    | -21.625 | < 2e-16  *** |
| dptD:sexM      | -1.89074 | 0.12091    | -15.638 | < 2e-16  *** |
| dptE:sexM      | -2.69802 | 0.13245    | -20.370 | < 2e-16  *** |
| dptF:sexM      | -1.92480 | 0.12948    | -14.865 | < 2e-16  *** |
| dptB:admitW    | 0.01869  | 0.22657    | 0.082   | 0.93426  |
| dptC:admitW    | -0.51976 | 0.20624    | -2.520  | 0.01173  * |
| dptD:admitW    | -0.20387 | 0.19379    | -1.052  | 0.29279  |
| dptE:admitW    | -0.60174 | 0.22570    | -2.666  | 0.00767  ** |
| dptF:admitW    | -1.18225 | 0.22653    | -5.219  | 1.80e-07  *** |
| dptB:admitY    | -0.04445 | 0.10985    | -0.405  | 0.68575  |
| dptC:admitY    | -1.25167 | 0.10614    | -11.792 | < 2e-16  *** |
| dptD:admitY    | -1.28699 | 0.10553    | -12.196 | < 2e-16  *** |
| dptE:admitY    | -1.72731 | 0.12547    | -13.766 | < 2e-16  *** |
| dptF:admitY    | -3.29845 | 0.16973    | -19.434 | < 2e-16  *** |
| sexM:admitW    | -0.09206 | 0.13563    | -0.679  | 0.49727  |
| sexM:admitY    | -0.08065 | 0.07997    | -1.008  | 0.31325  |

(Dispersion parameter for poisson family taken to be 1)
Residual deviance: 27.674
iii. Test whether the probability of admission varies according to the department and gender. State the null and the alternative hypothesis, what test you use, and the conclusions. Use $\alpha = 0.05$. (*Hint:* use results of question (a) above).

iv. Using the model in (ii), calculate the predicted count of admitted male students for department A.
3. Suppose there are observations \((Y_1, X_1), \cdots, (Y_n, X_n)\), where \(Y_i\) are responses and \(X_i\) are predictors. \(Y_1, \cdots, Y_n\) are independent instances from a Geometric distribution with parameters \(p_1, \cdots, p_n\) respectively, \(p_i \in (0, 1)\) for all \(1 \leq i \leq n\), and therefore the probability mass function of \(Y_i\) is

\[
f(y_i) = P[Y_i = y_i] = p_i(1 - p_i)^{y_i},\ y_i = 0, 1, \cdots .
\]

(a) Present the probability mass function of \(Y_i\) in the form of the exponential family, i.e.

\[
f(y_i) = \exp \left[ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi) \right].
\]

Define \(\theta_i, \phi, b(\theta_i), a(\phi),\) and \(c(y, \phi)\).

(b) Compute \(E(Y_i)\) and \(Var(Y_i)\) as functions of \(p_i\).

(c) What is the canonical link for these distributions?

(d) Suppose that under the canonical link the estimate of the linear component is \(\hat{\eta} = 0.2 - 0.1x\). We observe a new value of the predictor is \(x = 5\). Compute the estimate of \(p\) and the predicted value of the mean of the response according to the model.
4. The data reported the counts (denoted by $y$) of choices in a survey with respect to two main questions (denoted by $x_1$ and $x_2$). The levels of the two questions are all classified into “Bad”, “Fair”, “Good”, and “Very Good” and coded by 1, 2, 3 and 4 respectively. The data can be presented in a table as below:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$X_1$</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{13}$</td>
<td>$n_{14}$</td>
</tr>
<tr>
<td>1</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_{23}$</td>
<td>$n_{24}$</td>
</tr>
<tr>
<td>2</td>
<td>$n_{31}$</td>
<td>$n_{32}$</td>
<td>$n_{33}$</td>
<td>$n_{34}$</td>
</tr>
<tr>
<td>3</td>
<td>$n_{41}$</td>
<td>$n_{42}$</td>
<td>$n_{43}$</td>
<td>$n_{44}$</td>
</tr>
</tbody>
</table>

The three following models were fit to these data.

> g1
Call: glm(formula = y ~ factor(x1) + factor(x2), family = poisson)
Coefficients:
(Intercept) factor(x1)2 factor(x1)3 factor(x1)4 factor(x2)2 factor(x2)3 factor(x2)4
 1.6670  -1.5384  2.0984  -0.1885  1.5325  2.9930  1.6494
Degrees of Freedom: 15 Total (i.e. Null); 9 Residual
Null Deviance: 3692
Residual Deviance: 25.94 AIC: 123

> g2
Call: glm(formula=y~factor(x1)+factor(x2)+I(x1*x2),family=poisson)
Coefficients:
(Intercept) factor(x1)2 factor(x1)3 factor(x1)4 factor(x2)2 factor(x2)3 factor(x2)4 I(x1 * x2)
 2.1841  -2.1822  0.7858  -2.1929  0.9177  1.7323  -0.2813  0.2292
Degrees of Freedom: 15 Total (i.e. Null); 8 Residual
Null Deviance: 3692
Residual Deviance: 6.338 AIC: 105.4

> g3
Call: glm(formula=y~factor(x1)+factor(x2)+x1:factor(x2),family=poisson)
Coefficients:
(Intercept) factor(x1)2 factor(x1)3 factor(x1)4 factor(x2)2 factor(x2)3 factor(x2)4 factor(x2)1:x1
 2.9565  -1.2449  2.6602  0.6171  1.1357  1.9052  -0.1769  -0.6452
factor(x2)2:x1 factor(x2)3:x1 factor(x2)4:x1
 1.4981  -0.2523 NA
Degrees of Freedom: 15 Total (i.e. Null); 6 Residual
Null Deviance: 3692
Residual Deviance: 6.085 AIC: 109.1
(a) State the three models above, and find the best-fitting model.

(b) Propose a test for the significance of the linear-by-linear term.

(c) Why does an \( \text{NA} \) appear among the estimates of coefficients in model \( g_2 \)?

(d) Compute \( P(X_1 = 1|X_2 = 1) \) based on model \( g_2 \).
5. The data reports the level \((Y)\) of evaluation with respect to the hardness \((X)\) of matrices.

<table>
<thead>
<tr>
<th>Hardness</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86</td>
<td>88</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>123</td>
<td>170</td>
<td>157</td>
</tr>
<tr>
<td>3</td>
<td>116</td>
<td>242</td>
<td>432</td>
</tr>
<tr>
<td>4</td>
<td>99</td>
<td>248</td>
<td>761</td>
</tr>
</tbody>
</table>

\[
g_1 \leftarrow \text{multinom}(y \sim x, \text{weights}=\text{freq})
\]

\[
\text{summary}(g_1)\$\text{coeff}
\]

\[
\begin{array}{lll}
\text{(Intercept)} & x \\
\text{Median} & 1.373964 & -0.8645553 \\
\text{High} & 1.108437 & -0.5565890 \\
\end{array}
\]

\[
\text{c}(g_1\$\text{dev},g_1\$\text{edf})
\]

\[
[1] \ 4811.248 \ 4.000
\]

\[
g_2 \leftarrow \text{polr}(y \sim x, \text{weights}=\text{freq})
\]

\[
\text{summary}(g_2)\$\text{coeff}
\]

\[
\begin{array}{lllll}
\text{Value} & \text{Std. Error} & \text{t value} \\
x & -0.4937513 & 0.03868508 & -12.763350 \\
\text{Low}\mid\text{Median} & -1.3016567 & 0.12216347 & -10.655040 \\
\text{Median}\mid\text{High} & -0.5465661 & 0.11979129 & -4.562653 \\
\end{array}
\]

\[
\text{c}(g_2\$\text{dev},g_2\$\text{edf})
\]

\[
[1] \ 4918.179 \ 3.000
\]

\[
g_3 \leftarrow \text{multinom}(y \sim \text{factor}(x), \text{weights}=\text{freq})
\]

\[
\text{summary}(g_3)\$\text{coeff}
\]

\[
\begin{array}{llllll}
\text{(Intercept)} & \text{factor}(x)2 & \text{factor}(x)3 & \text{factor}(x)4 \\
\text{Median} & 0.4289912 & -0.6730500 & -1.743826 & -2.468505 \\
\text{High} & 0.4519837 & -0.3724289 & -1.031472 & -1.573189 \\
\end{array}
\]

\[
\text{c}(g_3\$\text{dev},g_3\$\text{edf})
\]

\[
[1] \ 4808.869 \ 8.000
\]

(a) State the three models above, and the corresponding assumptions.
(b) Compute the predicted probabilities in the multinomial model treating $X$ as continuous.

(c) Compute the predicted probabilities in the proportional odds model treating $X$ as continuous.

(d) Does the multinomial model or the proportional model odds fit the data. Explain.
6. Data were collected from a large survey on the psychology of debt. The values of following
variables were recorded,

carduse (Y): frequency of credit card use, coded as 1=never, 2=occasionally, 3=regularly;
house (X_1): security of housing tenure, coded as 1=rent, 2=mortgage, 3=owned outright;
agegp (X_2): age group, 1=youngest, ..., 4=oldest;

> summary(g1)
multinom(formula = carduse ~ house + agegp, data = pdebt)
Coefficients:
   (Intercept)  house2  house3    agegp
2  -1.988939  1.470922  1.462775 0.0004884059
3  -2.812530  1.690832  1.156660 0.2500190108
Std. Errors:
   (Intercept)  house2  house3    agegp
2  0.4807062  0.3722499  0.4414835 0.1506983
3  0.5445655  0.4081272  0.4772643 0.1580799
Residual Deviance: 822.564
AIC: 838.564

> summary(g2)
polr(formula = carduse ~ house + agegp, data = pdebt)
Coefficients:
Value Std. Error t value
house2  1.5625720  0.2924369  5.343279
house3  1.2179735  0.3421232  3.560043
agegp  0.1645859  0.1177110  1.398221
Intercepts:
Value Std. Error t value
1|2  1.7661  0.3872  4.5610
2|3  2.9601  0.4020  7.3628
Residual Deviance: 824.3241
AIC: 834.3241

(a) State the models and the assumptions. Which predictors are treated as continuous, and
which treated as categorical in each case?
(b) State which model you prefer. Why?

(c) Based on the two different models, predict the frequency of credit card use for a person who is in the youngest age group and renting a house.