There are totally 38 points in the exam. The students with score higher than or equal to 35 points will receive 35 points. Please write down your name and student ID number below.

NAME: 

ID: 
1. (8 points) The following data reported a study about the number of a kind of birds with a few explanatory variables in 20 locations in Northern Canada. The count of birds are given by the number of birds in a hectare area in each location. The explanatory variables include the resource of food (food), the resource of water (water), and local temperature (temp). Values of explanatory variables are relative to their long term standards. The R output is given below.

```r
> summary(mod1)
Call:
glm(formula = count ~ (food + water + temp)^2, family = poisson, 
data = birds)
Coefficients:
                Estimate Std. Error z value Pr(>|z|) 
(Intercept)    3.02120   0.06543   46.17 <2e-16 *** 
food           1.23264   0.10387   11.87 <2e-16 ***
water          0.39001   0.20996    1.86  0.0632 .  
temp           -0.20774   0.19279   -1.08  0.2812
food:water     0.53913   0.35853    1.50  0.1326
food:temp      0.40270   0.31042    1.29  0.1945
water:temp     0.85460   0.91502    0.93  0.3503
```

Null deviance: 283.986 on 19 degrees of freedom
Residual deviance: 12.485 on 13 degrees of freedom

```r
> mod2 <- glm(count~food+water+temp,family=poisson,data=birds)
> summary(mod2)
Call:
glm(formula = count ~ food + water + temp, family = poisson, 
data = birds)
Coefficients:
                Estimate Std. Error t value Pr(>|t|) 
(Intercept)   3.00887   0.06415  47.54 <2e-16 ***
food          1.22372   0.08997  13.64 <2e-16 ***
water         0.64840   0.13893   4.67 3.06e-06 ***
temp          -0.05684   0.16077  -0.35  0.724
```

Null deviance: 283.986 on 19 degrees of freedom
Residual deviance: 18.708 on 16 degrees of freedom

```r
> summary(mod3)
Call:
glm(formula = count ~ food + water, family = poisson, data = birds)
Coefficients:
                Estimate Std. Error t value Pr(>|t|) 
(Intercept)   3.00255   0.06186  48.96 <2e-16 ***
```
null deviance: 283.986 on 19 degrees of freedom
Residual deviance: 18.833 on 17 degrees of freedom

> summary(mod3)$cov.unscaled

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>food</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.003826726</td>
<td>-0.003640855</td>
<td>-0.0033240180</td>
</tr>
<tr>
<td>food</td>
<td>-0.003640855</td>
<td>0.0067838256</td>
<td>0.0007683732</td>
</tr>
<tr>
<td>water</td>
<td>-0.003324018</td>
<td>0.0007683732</td>
<td>0.0192638199</td>
</tr>
</tbody>
</table>

> round(qchisq(0.95,1:6),2)

[1] 3.84 5.99 7.81 9.49 11.07 12.59

(a) (2 points). Justify whether overdispersion is a concern. Provide a test about whether interaction effects can be ignored.

(b) (2 points). Define a statistic similar to $R^2$ in regression. Provide values of the statistic in the three fitted models, respectively.
(c) (2 points). Provide two tests for the significance of temperature based on the main effect model.

(d) (2 points). Provide the 95% confidence interval of the count when food = 1.0 and water = 0.8.
2. (10 points) The data reports the impurity of ironstone collected by four different methods and five different operators. Each operator used the four methods to collect five samples, and the impurity is reported. The R output is given below.

> g <- lm(yy~method*operator)
> anova(g)
Analysis of Variance Table
Response: yy
Df  Sum Sq Mean Sq  F value Pr(>F)
method 3 422.14 140.712 6.5459 0.0005148 ***
operator 4 895.97 223.992 10.4200 7.624e-07 ***
method:operator 12 394.41 32.867 1.5290 0.1310873
Residuals 80 1719.70 21.496

> gg <- lme(yy~method,random=~1|operator)
> summary(gg)
Linear mixed-effects model fit by REML
Random effects:
  Formula: ~1 | operator
   (Intercept) Residual
   StdDev: 3.17027 4.793686
Fixed effects: yy ~ method
  Value   Std.Error DF  t-value p-value
  (Intercept)          23.535     1.7115 92 13.7509  0.0000
  methodB              1.344     1.3559 92  0.9910  0.3243
  methodC              3.749     1.3559 92  2.7651  0.0069
  methodD             -1.893     1.3559 92 -1.3962  0.1660

> g0 <- lme(yy~1,random=~1|operator,method="ML")
> g1 <- lme(yy~method,random=~1|operator,method="ML")
> anova(g1,g0)

Model df AIC    BIC  logLik Test L.Ratio p-value
g1  1 6 616.4686 632.0996 -302.2343
  g0 2 3 627.7635 635.5790 -310.8818 1 vs 2 17.29492 6e-04

(a) (2 points) Is the dataset nested or not? Explain.
(b) (2 points) Write down the linear mixed-effect model that was fitted in the R output. Specify model assumptions.

(c) (2 points) Write down estimates of model parameters in the previous part. If zero-sum constraint is used, what are the estimates of fixed effects.

(d) (2 points) Explain the method and the result of the test in the R output.

(e) (2 points) Can the testing method in the previous part be used for significance of the random effect? If not, provide a test and simply describe the method.
3. (10 points) The data reports the growth of trees in a certain forest. The size of trees was measured by diameter in a ten year period for 50 trees. The R output is given.

```r
> summary(gg)
Linear mixed-effects model fit by REML
Random effects:
  Formula: ~Year | Tree
  Structure: General positive-definite, Log-Cholesky parametrization
  StdDev  Corr
(Intercept) 4.0616 (Intr)
Year 0.1810 -0.008
Residual 0.6708
Fixed effects: Diameter ~ Year
  Value Std.Error DF t-value p-value
(Intercept) 19.6076 0.5780 449 33.92097 0
Year 1.9585 0.0276 449 70.85176 0
Correlation:
  (Intr)
Year -0.045
> g0 <- lme(Diameter~Year,random=~Year|Tree,data=TreeSize,method="ML")
> g1 <- lme(Diameter~Year,random=~1|Tree,data=TreeSize,method="ML")
> anova(g0,g1)
    Model df      AIC      BIC    logLik Test L.Ratio p-value
g0     1   6 1422.727 1448.014 -705.3633
g1     2   4 1550.520 1567.379 -771.2602  1 vs 2 131.7937 <.0001
```

(a) (2 points). Specify the statistical model fitted by the code.
(b) (2 points). Write estimates of parameters in the model given in the previous part.

(c) (2 points). Compute the predicted value of the mean diameter and its 95% confidence interval when “year” is 5.

(d) (2 points). Compute the variation of tree diameters when “year” is 5.

(e) (2 points). Compute the 95% confidence interval of the observed value when “year” is 5.
4. (10 points). In the study of the effect of a treatment for late stage lung cancer, the survival weeks of two groups of patients were observed. The observed survival weeks for placebo group were: 1, 3+, 4, 4, 4, 5, 10+, 11, 12+, 12, 23+, 24, 26, 30, 32, 32, 35+, 37+, 39, 40+, and the observed survival weeks for treatment group were: 5, 12+, 13, 14, 16, 21, 23, 29+, 31, 32, 54, 54+, 61+, 73+, 76, 100, 134, 151, 188, 232+. The R output is given.

Call:
survdiff(formula = Surv(time, censor) ~ trt)

          N  Observed  Expected (O-E)^2/V
trt=placebo 20       13       8.22         4.8
trt=treatment 20     14      18.78         4.8

Call:
survreg(formula = Surv(time, censor) ~ trt, dist = "weibull")

           Value     Std. Error    z      p
(Intercept) 3.3771    0.261     12.952  2.30e-38
trtreatment 1.1724    0.362      3.234  1.22e-03
Log(scale)  -0.0643    0.155     -0.413  6.79e-01

Call:
coxph(formula = Surv(time, censor) ~ trt)

          coef  exp(coef)    se(coef)    z     p
trtreatment -0.97  0.379       0.455    -2.13  0.033

Call: survfit.coxph(object = gcox, newdata = data.frame(trt = placebo))

        time n.risk n.event survival std.err lower 95% CI upper 95% CI
      14     27       1     0.62481  0.1050    4.49e-01    0.869
      16     26       1     0.58709  0.1080    4.09e-01    0.842
      21     25       1     0.55082  0.1099    3.73e-01    0.814
      23     24       1     0.51598  0.1108    3.39e-01    0.786
      24     22       1     0.48023  0.1115    3.05e-01    0.757
      26     21       1     0.44448  0.1117    2.72e-01    0.727
      30     19       1     0.40760  0.1116    2.38e-01    0.697
      31     18       1     0.37073  0.1111    2.06e-01    0.667
      32     17       3     0.26879  0.1039    1.26e-01    0.573

(a) (2 points). Compute the K-M estimator for the placebo group when $t < 20$. 

(b) (2 points). Test whether the two survival functions are the same.

(c) (2 points). Suppose the death follows exponential distribution with density \( f(t) = \lambda e^{-\lambda t} \).
Compute the estimate values of \( \lambda \) for both groups.

(d) (2 points). Assume the death follows the Weibull distribution, do you accept the null hypothesis that the death also follows exponential distribution.

(e) (2 points). Write down the assumption of Cox proportional hazard model and propose a method to diagnose it. Compute the estimate of \( S(24) \) for both groups.