

## STAT 262 PROBLEM SET 2 SOLUTIONS

### 1. PROBLEM 1

Part (a):

Execute the following R code:

```
> drugdata <- read.csv("drugdata.csv")
> attach(drugdata)
> mean(chol[isMale==1])
[1] 245.225
> mean(chol[isMale==0])
[1] 253.5125
> sd(chol[isMale==1])
[1] 24.74885
> sd(chol[isMale==0])
[1] 20.07958
> mean(chol[grp=="cont"])
[1] 253.1
> mean(chol[grp=="drug"])
[1] 245.6375
> sd(chol[grp=="cont"])
[1] 17.76399
> sd(chol[grp=="drug"])
[1] 26.58411
```

Part (b):

People interpreted this question in different ways. One way to do this analysis is to compare the cholesterol levels of individuals in the drug group at time 4 to their cholesterol levels at time 1. The R code for doing this is as follows:

```
> t.test(chol[test==1&grp=="drug"&isMale==1],
+ chol[test==4&grp=="drug"&isMale==1])
```

Welch Two Sample t-test

```
t = 3.4095, df = 17.993, p-value = 0.003127
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.47080 56.72920
sample estimates:
mean of x mean of y
 253.7      218.6
```

```
> t.test(chol[test==1&grp=="drug"&isMale==0],
```

```
+ chol[test==4&grp=="drug"&isMale==0])
```

```
Welch Two Sample t-test
```

```
t = -1.5127, df = 17.997, p-value = 0.1477
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -32.010744  5.210744
sample estimates:
mean of x mean of y
  254.3     267.7
```

We see that the drug is effective among men but not among women.

Another way to interpret this question is to test the hypothesis that the difference between the measurements at time 4 and time 1 is greater among the drug group than the control group:

```
> t.test(chol[test==4&grp=="drug"&isMale==1]-
+ chol[test==1&grp=="drug"&isMale==1],
+ chol[test==4&grp=="cont"&isMale==1]-chol[test==1&grp=="cont"&isMale==1])
```

```
Welch Two Sample t-test
```

```
t = -7.4217, df = 18, p-value = 7.015e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -38.23573 -21.36427
sample estimates:
mean of x mean of y
  -35.1     -5.3
```

```
> t.test(chol[test==4&grp=="drug"&isMale==0]-
+ chol[test==1&grp=="drug"&isMale==0],
+ chol[test==4&grp=="cont"&isMale==0]-chol[test==1&grp=="cont"&isMale==0])
```

```
Welch Two Sample t-test
```

```
t = 9.1344, df = 17.725, p-value = 4.027e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 15.77975 25.22025
sample estimates:
mean of x mean of y
  13.4     -7.1
```

Again, we see that the drug is effective among men but not among women. (Although the t-statistic for women is significant, note that this is a two-sided test. We want to test the hypothesis that the drug *reduces* cholesterol. Since the reduction in cholesterol is *greater* for control patients than it is for patients taking the drug, we fail to reject the hypothesis that the drug has no effect.)

Part (c):

I spent way too long trying to reproduce the SAS output for this problem in R.

Here is a way that one can do it:

```
> summary(aov(chol~grp*factor(test)+grp:factor(subject)+Error(factor(subject)),
+ data=drugdata, subset=isMale==1))
```

```
Error: factor(subject)
              Df  Sum Sq Mean Sq
grp              1 14311.2 14311.2
grp:factor(subject) 18 25068.7  1392.7
```

Error: Within

```
              Df Sum Sq Mean Sq F value    Pr(>F)
factor(test)   3 5029.7  1676.6  64.775 < 2.2e-16 ***
grp:factor(test) 3 2580.6   860.2  33.233 2.623e-12 ***
Residuals     54 1397.7    25.9
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> summary(aov(chol~grp*factor(test)+grp:factor(subject)+Error(factor(subject)),
+ data=drugdata, subset=isMale==0))
```

```
Error: factor(subject)
              Df  Sum Sq Mean Sq
grp              1  2796.6  2796.6
grp:factor(subject) 18 26876.1  1493.1
```

Error: Within

```
              Df  Sum Sq Mean Sq F value    Pr(>F)
factor(test)   3  138.44   46.15  2.8253  0.04717 *
grp:factor(test) 3 1158.84  386.28 23.6504 6.601e-10 ***
Residuals     54  881.97   16.33
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

I think the point here is that the “group” term is not significant for women. If you figured out that much, I gave you credit.

Part (d):

It appears that the drug is effective for men but not for women.

## 2. PROBLEM 2

Part (a):

See the R output below:

```
> mutations <- read.csv("mutations.csv")
> summary(glm(mutations~., data=mutations, family=poisson))
```

Call:

```
glm(formula = mutations ~ ., family = poisson, data = mutations)
```

Deviance Residuals:

|  | Min     | 1Q      | Median  | 3Q     | Max    |
|--|---------|---------|---------|--------|--------|
|  | -2.6202 | -0.8573 | -0.2120 | 0.6841 | 2.0033 |

Coefficients:

|             | Estimate   | Std. Error | z value | Pr(> z )     |
|-------------|------------|------------|---------|--------------|
| (Intercept) | -0.0337412 | 0.6195096  | -0.054  | 0.956565     |
| months      | 0.0361526  | 0.0093492  | 3.867   | 0.000110 *** |
| gss         | 0.0467784  | 0.0298100  | 1.569   | 0.116597     |
| CD4         | -0.0004856 | 0.0004124  | -1.177  | 0.239038     |
| VL          | -0.1762405 | 0.1012796  | -1.740  | 0.081835 .   |
| drugs       | 0.0672856  | 0.0376341  | 1.788   | 0.073794 .   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 126.54 on 86 degrees of freedom  
 Residual deviance: 109.10 on 81 degrees of freedom  
 AIC: 295.12

Number of Fisher Scoring iterations: 5

Part (b):

In my mind, simply using the p-values from the output of the glm function in R is an acceptable solution to this problem. (If I remember correctly, these p-values are based on Wald's test, which is very similar to the likelihood ratio test.) If it is necessary to use the likelihood ratio test, one can execute the following commands in R:

```
> anova(glm(mutations~CD4+VL+drugs+gss+months, data=mutations, family=poisson))
Analysis of Deviance Table
```

Model: poisson, link: log

Response: mutations

Terms added sequentially (first to last)

|       | Df | Deviance | Resid. Df | Resid. Dev |
|-------|----|----------|-----------|------------|
| NULL  |    |          | 86        | 126.545    |
| CD4   | 1  | 0.071    | 85        | 126.474    |
| VL    | 1  | 1.609    | 84        | 124.865    |
| drugs | 1  | 0.012    | 83        | 124.853    |
| gss   | 1  | 1.310    | 82        | 123.543    |

```

months 1 14.448 81 109.095
> 1-pchisq(14.448,1)
[1] 0.0001440828
> anova(glm(mutations~CD4+VL+drugs+months+gss, data=mutations, family=poisson))
Analysis of Deviance Table

```

Model: poisson, link: log

Response: mutations

Terms added sequentially (first to last)

|        | Df | Deviance | Resid. Df | Resid. Dev |
|--------|----|----------|-----------|------------|
| NULL   |    |          | 86        | 126.545    |
| CD4    | 1  | 0.071    | 85        | 126.474    |
| VL     | 1  | 1.609    | 84        | 124.865    |
| drugs  | 1  | 0.012    | 83        | 124.853    |
| months | 1  | 13.278   | 82        | 111.575    |
| gss    | 1  | 2.480    | 81        | 109.095    |

```
> 1-pchisq(2.480,1)
```

```
[1] 0.1153023
```

We see that GSS can be dropped from the model but MONTHS cannot.

Part (c):

Based on the R output above, the estimate of the coefficient of GSS is about 0.047, and the estimate of the standard error is 0.030. Thus, the upper and lower 95% confidence bounds are:

```

> mutations.glm.sum <- summary(glm(mutations~.,data=mutations,
+ family=poisson))
> mutations.glm.sum$coefficients
      Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -0.0337411934 0.619509610 -0.05446436 0.9565652043
months      0.0361525978 0.009349160  3.86693543 0.0001102116
gss         0.0467783621 0.029810003  1.56921694 0.1165973991
CD4        -0.0004855678 0.000412409 -1.17739394 0.2390383093
VL        -0.1762405128 0.101279599 -1.74013833 0.0818347317
drugs      0.0672855821 0.037634089  1.78788923 0.0737938804
> mut.coef <- mutations.glm.sum$coefficients
> mut.coef[3,1] - mut.coef[3,2]*qnorm(.975)
[1] -0.01164817
> mut.coef[3,1] + mut.coef[3,2]*qnorm(.975)
[1] 0.1052049

```

It appears that this coefficient is not significantly different from zero.

Part (d):

An easy way to test for overdispersion in R is to fit the model setting the “family” argument to be “quasipoisson” instead of “poisson”:

```
> summary(glm(mutations~., data=mutations, family=quasipoisson))
```

Call:

```
glm(formula = mutations ~ ., family = quasipoisson, data = mutations)
```

Deviance Residuals:

|  | Min     | 1Q      | Median  | 3Q     | Max    |
|--|---------|---------|---------|--------|--------|
|  | -2.6202 | -0.8573 | -0.2120 | 0.6841 | 2.0033 |

Coefficients:

|             | Estimate   | Std. Error | t value | Pr(> t )    |
|-------------|------------|------------|---------|-------------|
| (Intercept) | -0.0337412 | 0.6697609  | -0.050  | 0.95995     |
| months      | 0.0361526  | 0.0101075  | 3.577   | 0.00059 *** |
| gss         | 0.0467784  | 0.0322280  | 1.451   | 0.15051     |
| CD4         | -0.0004856 | 0.0004459  | -1.089  | 0.27936     |
| VL          | -0.1762405 | 0.1094949  | -1.610  | 0.11138     |
| drugs       | 0.0672856  | 0.0406868  | 1.654   | 0.10205     |

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.168809)

Null deviance: 126.54 on 86 degrees of freedom  
 Residual deviance: 109.10 on 81 degrees of freedom  
 AIC: NA

Number of Fisher Scoring iterations: 5

Since the dispersion parameter is very close to 1, overdispersion does not appear to be a problem.

### 3. PROBLEM 3

Part (a):

One can do this in R as follows:

```
> assay <- read.csv("assay.csv")
> summary(aov(SQV~1+Error(factor(ref)),data=assay))
```

Error: factor(ref)

|           | Df | Sum Sq  | Mean Sq | F value | Pr(>F) |
|-----------|----|---------|---------|---------|--------|
| Residuals | 5  | 13.0591 | 2.6118  |         |        |

Error: Within

|           | Df  | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|-----|--------|---------|---------|--------|
| Residuals | 138 | 53.977 | 0.391   |         |        |

Part (b):

The appropriate R commands are the following:

```
> 1-pf(2.6118/.391,5,138)
[1] 1.329986e-05
```

We may reject the hypothesis that there is no “center” effect.

Part (c):

Using the MSE we found in part (a), we may obtain the upper and lower confidence bounds as follows:

```
> mean(assay[,1]) - qnorm(.975)*.391/12
[1] -0.5644589
> mean(assay[,1]) + qnorm(.975)*.391/12
[1] -0.4367346
```