Math650 Homework 11

Yu Huang

2006-11-26

Abstract

Questions: Chap 14. No 9, No 10, No 13.

1 Chap 14 No 9

Code:

```
data1 = read.csv("/usr/local/doc/statistical_sleuth/ASCII/case1402.csv")
data1$LOGWILLIAM = log(data1$WILLIAM)
data1$LOGFORREST = log(data1$FORREST)
data1.glm = glm(LOGWILLIAM~STRESS + SO2 + O3 + O3*STRESS, data=data1)
summary(data1.glm)

diff_of_ozone_slope = data1.glm$coefficients[5]
cat("difference between ozone slope parameters for stressed versus well-watered plots is",
cat("in the scale of seed yield, well-watered is ", exp(diff_of_ozone_slope), " times bett

sd_error_of_diff_of_ozone_slope = summary(data1.glm)$coefficients[5,2]
#got it from the output of summary()

upper_95_bound = diff_of_ozone_slope + qt(0.975, 25)*sd_error_of_diff_of_ozone_slope
lower_95_bound = diff_of_ozone_slope - qt(0.975, 25)*sd_error_of_diff_of_ozone_slope

cat("95% confidence interval in the scale of seed yield, from", exp(lower_95_bound), "to",
```

In the scale of seed yield, the ozone slope parameter for well-watered is 23.52451% of the stressed. The linear model outputted by R regards well-watered as 1 and stressed as 0. The 95% confidence interval is 1.439116% to 384.5432%.

However, the tricky point is that while converting value to the scale of seed yield (take the exponential), the exponent of the product of coefficient and ozone dosage can't be separated into two exponents, (i.e. $e^{x*y} = e^{x^y} ! = e^x * e^y$). So this interpretation is problematic.

In the book(section 14.1.2), the interpretation is reversed in terms of well-watered and stressed.

2 Chap 14, No 10

Code:

```
> data1$SO2 = factor(data1$SO2)
> data1.glm_forrest = glm(LOGFORREST~O3 + SO2 + STRESS + O3*SO2 +
+ O3*STRESS + SO2*STRESS + O3*SO2*STRESS , data=data1)
> data1.glm_william = glm(LOGWILLIAM~O3 + SO2 + STRESS + O3*SO2 +
+ O3*STRESS + SO2*STRESS + O3*SO2*STRESS, data=data1)
> data1.glm_forrest.anova = anova(data1.glm_forrest)
> print(data1.glm_forrest.anova)
Analysis of Deviance Table

Model: gaussian, link: identity

Response: LOGFORREST

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev
NULL                         29    1.34127
O3            1  0.72077        28    0.62050
SO2           2  0.06346        26    0.55703
STRESS        1  0.00804        25    0.54899
O3:SO2        2  0.01731        23    0.53168
O3:STRESS     1  0.01363        22    0.51805
SO2:STRESS    2  0.02854        20    0.48951
O3:SO2:STRESS 2  0.06835        18    0.42116
> data1.glm_william.anova = anova(data1.glm_william)
> print(data1.glm_william.anova)
Analysis of Deviance Table

Model: gaussian, link: identity

Response: LOGWILLIAM

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev
NULL                         29    1.95574
O3            1  1.14959        28    0.80616
SO2           2  0.27798        26    0.52817
STRESS        1  0.23764        25    0.29053
O3:SO2        2  0.00371        23    0.28682
O3:STRESS     1  0.01277        22    0.27405
SO2:STRESS    2  0.02632        20    0.24772
O3:SO2:STRESS 2  0.00933        18    0.23840
>
> #the output of anova is 4-column. The 3rd and 4th columns are complimentary to> #1st and
> #Starting from the 2nd row, the 3rd and 4th column is the df and residual after
> #including the parameters from this row and above.
>
```

```
> print_anova_table = function(anova_result)
+ {
+ no_of_rows = dim(anova_result)[1]
+ no_of_columns = dim(anova_result)[2]
+ rss_full_model = anova_result[no_of_rows, no_of_columns]/anova_result[no_of_rows, no_of_
+ #the 3rd and 4th column of last row is regarded as the full model
+ cat("Source\tDf\tSum of squares\tMean square\tF-stat\tp-value\n")
+ for (i in seq(2, no_of_rows))
+ {
+ mean_sq = anova_result[i, 2]/anova_result[i,1]
+ f_stat = mean_sq/rss_full_model
+ p_value = pf(f_stat, anova_result[i,1], anova_result[no_of_rows, no_of_columns-1], lower
+ source_name = row.names(anova_result)[i]
+ cat(source_name, "\t", anova_result[i,1], "\t", anova_result[i, 2], "\t", mean_sq, "\t",
+ }
+ }
>
> print_anova_table(data1.glm_forrest.anova)
Source  Df      Sum of squares  Mean square     F-stat  p-value
O3      1       0.7207733       0.7207733       30.80501        2.867628e-05
SO2     2       0.06346401      0.03173201      1.356189        0.2827372
STRESS  1       0.008037132     0.008037132     0.3434976       0.5650948
O3:SO2  2       0.01731267      0.008656337     0.3699617       0.6958926
O3:STRESS       1       0.01363068      0.01363068      0.5825595       0.4551988
SO2:STRESS      2       0.02854002      0.01427001      0.6098836       0.5542675
O3:SO2:STRESS   2       0.06834879      0.03417440      1.460574        0.2583349
> print_anova_table(data1.glm_william.anova)
Source  Df      Sum of squares  Mean square     F-stat  p-value
O3      1       1.149587        1.149587        86.79937        2.624323e-08
SO2     2       0.2779832       0.1389916       10.49454        0.0009527212
STRESS  1       0.2376424       0.2376424       17.94315        0.0004969365
O3:SO2  2       0.003711956     0.001855978     0.1401353       0.8701797
O3:STRESS       1       0.01276862      0.01276862      0.9640928       0.3391722
SO2:STRESS      2       0.02632498      0.01316249      0.9938319       0.3895778
O3:SO2:STRESS   2       0.009329642     0.004664821     0.3522166       0.7078668
```

The result is identical.

3 Chap 14, No 13

Code:

```
> t.test(data1$WILLIAM, data1$FORREST, paired=TRUE)

        Paired t-test

data:  data1$WILLIAM and data1$FORREST
t = -0.4976, df = 29, p-value = 0.6225
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -325.1779  197.9113
sample estimates:
mean of the differences
              -63.63333

> wilcox.test(data1$WILLIAM, data1$FORREST, paired=TRUE)

        Wilcoxon signed rank test

data:  data1$WILLIAM and data1$FORREST
V = 177, p-value = 0.2621
alternative hypothesis: true location shift is not equal to 0

>
> data1$LOGRATIO = log(data1$FORREST/data1$WILLIAM)
> data1.glm_logratio = glm(LOGRATIO~O3 + SO2 + STRESS + O3*SO2 +
+ O3*STRESS + SO2*STRESS + O3*SO2*STRESS, data=data1)
> data1.glm_logratio.anova = anova(data1.glm_logratio)
> print(data1.glm_logratio.anova)
Analysis of Deviance Table

Model: gaussian, link: identity

Response: LOGRATIO

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev
NULL                         29    0.95423
O3            1  0.04982        28    0.90441
SO2           2  0.08253        26    0.82189
STRESS        1  0.15827        25    0.66361
O3:SO2        2  0.00503        23    0.65859
O3:STRESS     1  0.05278        22    0.60580
SO2:STRESS    2  0.10564        20    0.50017
O3:SO2:STRESS 2  0.09206        18    0.40811
> print_anova_table(data1.glm_logratio.anova)
Source  Df      Sum of squares  Mean square     F-stat  p-value
O3       1       0.04982009      0.04982009      2.197349        0.1555445
SO2      2       0.08252682      0.04126341      1.819951        0.1906140
STRESS   1       0.1582732       0.1582732       6.98075         0.01656262
O3:SO2   2       0.005026355     0.002513177     0.1108454       0.8956833
O3:STRESS        1       0.05278453      0.05278453      2.328098        0.1444362
SO2:STRESS       2       0.1056379       0.05281897      2.329617        0.1259637
O3:SO2:STRESS    2       0.09205504      0.04602752      2.030076        0.1603157
```

p-value is 0.6225 for t-test and 0.2621 for wilcox test(signed rank test). This means the difference between cultivar is not significant.

A linear regression similar to the one above, with response variable replaced by log(FORREST/WILLIAM) shows only the coefficient of *STRESS of WATER* is sort of significant(p-value=0.016) in interpreting the difference.