

# STATS 305A H.W.# 5

*Adhitya Venkatesh*

*October 31, 2017*

## Theoretical Problems

1.)

We use the result from the lecture notes that the parameter estimates in the random effects model  $Y_{ij} = \mu + a_i + \epsilon_{ij}$  can be determined from the parameter estimates of  $Y_{ij} = \mu_i + \epsilon_{ij}$ .

The relationship between the expected parameter values  $\mu_i$  and  $a_i$  is given by

$$a_i = \mu_i - \mu$$

The ols estimates from the latter (cell mean) model yields

$$\mu_i = \sum_{j=1}^{n_i} Y_{ij} / n_i$$

and for  $\mu$ :

$$\mu = \sum_{j=1}^{n_i} \sum_{i=1}^K Y_{ij} / \sum_{i=1}^K n_i$$

Thus, we arrive at the final result

$$E(a_i|Y) = \sum_{j=1}^{n_i} Y_{ij} / n_i - \sum_{j=1}^{n_i} \sum_{i=1}^K Y_{ij} / \sum_{i=1}^K n_i$$

2.)

## Computational Problems

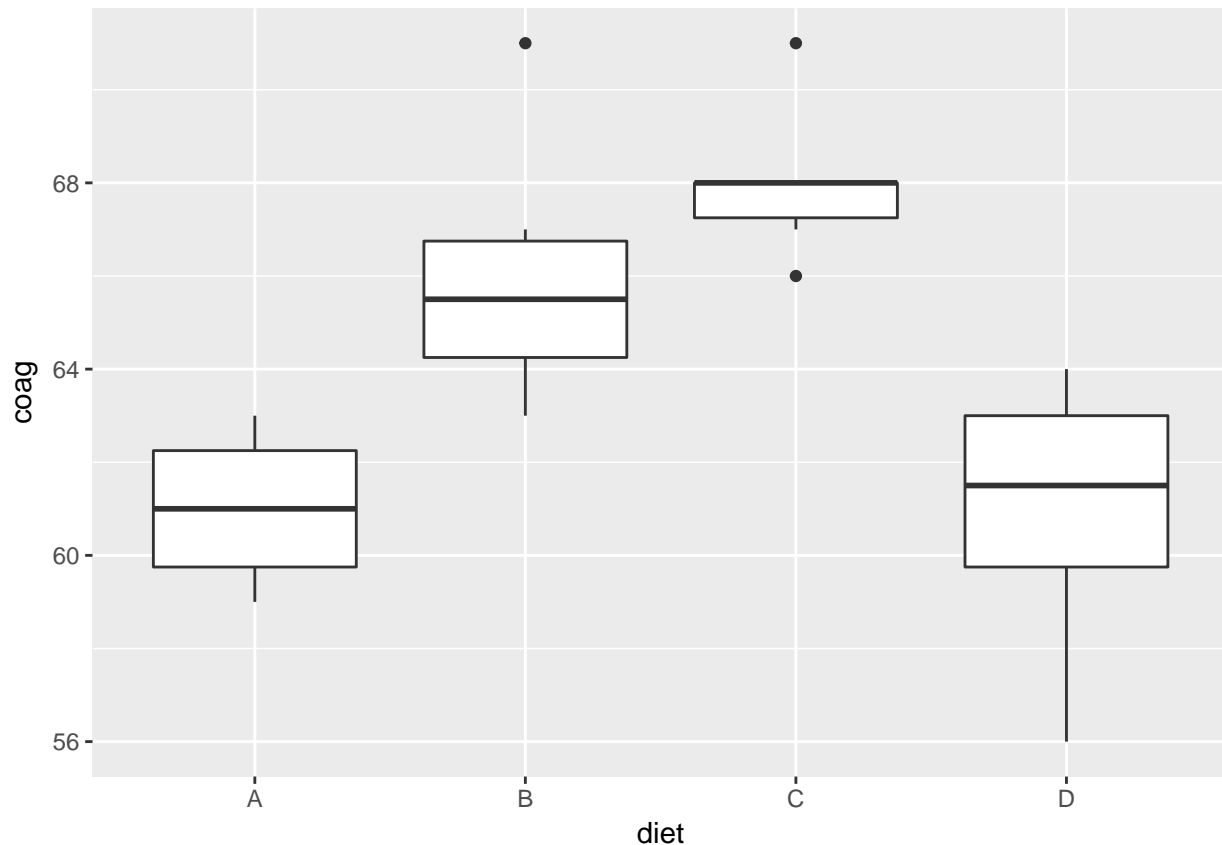
1.)

a.)

```
library(faraway)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
coag_data = faraway::coagulation
ggplot(coag_data, aes(x = diet, y = coag)) + geom_boxplot()
```



We observe from the plots that the median of treatment A and D are about equal, albeit the distribution for the group administered diet A is more symmetrical and has smaller variance than the distribution of the group administered diet D. The highest median coagulation value is yielded by group C, followed by group B. Both of these groups contain outliers but the distribution of B has comparable variance and symmetry to those of distribution A. Distribution C is heavily left skewed with outliers on both extremes (indicated by black dots).

b.)

```
lin_mod_1 = lm(coag_data$coag~coag_data$diet)
summary(lin_mod_1)
```

```
##
## Call:
## lm(formula = coag_data$coag ~ coag_data$diet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##    -5.00    -1.25     0.00     1.25     5.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.100e+01  1.183e+00  51.554 < 2e-16 ***
## coag_data$dietB  5.000e+00  1.528e+00   3.273  0.003803 **
## coag_data$dietC  7.000e+00  1.528e+00   4.583  0.000181 ***
## coag_data$dietD  2.991e-15  1.449e+00   0.000  1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
```

```
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

```
model.matrix(lin_mod_1)
```

```
##      (Intercept) coag_data$dietB coag_data$dietC coag_data$dietD
## 1             1             0             0             0
## 2             1             0             0             0
## 3             1             0             0             0
## 4             1             0             0             0
## 5             1             1             0             0
## 6             1             1             0             0
## 7             1             1             0             0
## 8             1             1             0             0
## 9             1             1             0             0
## 10            1             1             0             0
## 11            1             0             1             0
## 12            1             0             1             0
## 13            1             0             1             0
## 14            1             0             1             0
## 15            1             0             1             0
## 16            1             0             1             0
## 17            1             0             0             1
## 18            1             0             0             1
## 19            1             0             0             1
## 20            1             0             0             1
## 21            1             0             0             1
## 22            1             0             0             1
## 23            1             0             0             1
## 24            1             0             0             1
## attr("assign")
## [1] 0 1 1 1
## attr("contrasts")
## attr("contrasts")$`coag_data$diet`
## [1] "contr.treatment"
```

The coefficients can be interpreted as follows: mean coagulation for treatment A is 61 (intercept) and that of treatment B is 5 more or 66, C is 7 more or 68, and the mean of D is essentially equivalent to that of A (61).

c.)

```
lin_mod_2 = lm(coag_data$coag~coag_data$diet -1)
summary(lin_mod_2)
```

```
##
## Call:
## lm(formula = coag_data$coag ~ coag_data$diet - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## coag_data$dietA  61.0000     1.1832   51.55  <2e-16 ***
## coag_data$dietB  66.0000     0.9661   68.32  <2e-16 ***
```

```
## coag_data$dietC 68.0000    0.9661    70.39    <2e-16 ***
## coag_data$dietD 61.0000    0.8367    72.91    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9986
## F-statistic: 4399 on 4 and 20 DF,  p-value: < 2.2e-16
```

These coefficients directly represent the mean coagulation value for a given diet A through D.

d.)

```
lin_mod_0 = lm(coag_data$coag~1)
anova(lin_mod_0, lin_mod_1)
```

```
## Analysis of Variance Table
##
## Model 1: coag_data$coag ~ 1
## Model 2: coag_data$coag ~ coag_data$diet
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1      23 340
## 2      20 112  3      228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value < 0.05, we reject the null hypothesis that the means of the diet levels are equal.

e.)

```
options(contrasts = c("contr.sum", "contr.poly"))
lin_mod_2 = lm(coag~diet, coag_data)
summary(lin_mod_2)
```

```
##
## Call:
## lm(formula = coag ~ diet, data = coag_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0000    0.4979 128.537  < 2e-16 ***
## diet1        -3.0000    0.9736  -3.081 0.005889 **
## diet2         2.0000    0.8453   2.366 0.028195 *
## diet3         4.0000    0.8453   4.732 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

```
model.matrix(~diet, data = coag_data, contrasts.arg = list(diet = "contr.sum"))
```

```
##      (Intercept) diet1 diet2 diet3
```

```
## 1      1      1      0      0
## 2      1      1      0      0
## 3      1      1      0      0
## 4      1      1      0      0
## 5      1      0      1      0
## 6      1      0      1      0
## 7      1      0      1      0
## 8      1      0      1      0
## 9      1      0      1      0
## 10     1      0      1      0
## 11     1      0      0      1
## 12     1      0      0      1
## 13     1      0      0      1
## 14     1      0      0      1
## 15     1      0      0      1
## 16     1      0      0      1
## 17     1     -1     -1     -1
## 18     1     -1     -1     -1
## 19     1     -1     -1     -1
## 20     1     -1     -1     -1
## 21     1     -1     -1     -1
## 22     1     -1     -1     -1
## 23     1     -1     -1     -1
## 24     1     -1     -1     -1
## attr("assign")
## [1] 0 1 1 1
## attr("contrasts")
## attr("contrasts")$diet
## [1] "contr.sum"
```

Here the coefficients are interpreted as follows: the intercept represents the mean coagulation value and each slope term corresponding to the diet is the amount of deviation from the intercept (i.e. mean for diet D is 61).

f.)

```
pairwise.t.test(coag_data$coag, coag_data$diet, p.adjust.method = "none")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: coag_data$coag and coag_data$diet
##
##      A      B      C
## B 0.00380 -      -
## C 0.00018 0.15878 -
## D 1.00000 0.00086 2.3e-05
##
## P value adjustment method: none
```

Thus, the pairwise t-test indicates that the difference in means between treatment A and D is not statistically significant as p-value = 1.00.

g.)

```
pairwise.t.test(coag_data$coag, coag_data$diet, p.adjust.method = "bonferroni")
```

```
##
```

```
## Pairwise comparisons using t tests with pooled SD
##
## data: coag_data$coag and coag_data$diet
##
##      A      B      C
## B 0.02282 -      -
## C 0.00108 0.95266 -
## D 1.00000 0.00518 0.00014
##
## P value adjustment method: bonferroni
```

h.)

```
TukeyHSD(aov(coag ~ diet, coag_data))
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = coag ~ diet, data = coag_data)
##
## $diet
##      diff      lwr      upr      p adj
## B-A      5  0.7245544  9.275446 0.0183283
## C-A      7  2.7245544 11.275446 0.0009577
## D-A      0 -4.0560438  4.056044 1.0000000
## C-B      2 -1.8240748  5.824075 0.4766005
## D-B     -5 -8.5770944 -1.422906 0.0044114
## D-C     -7 -10.5770944 -3.422906 0.0001268
```

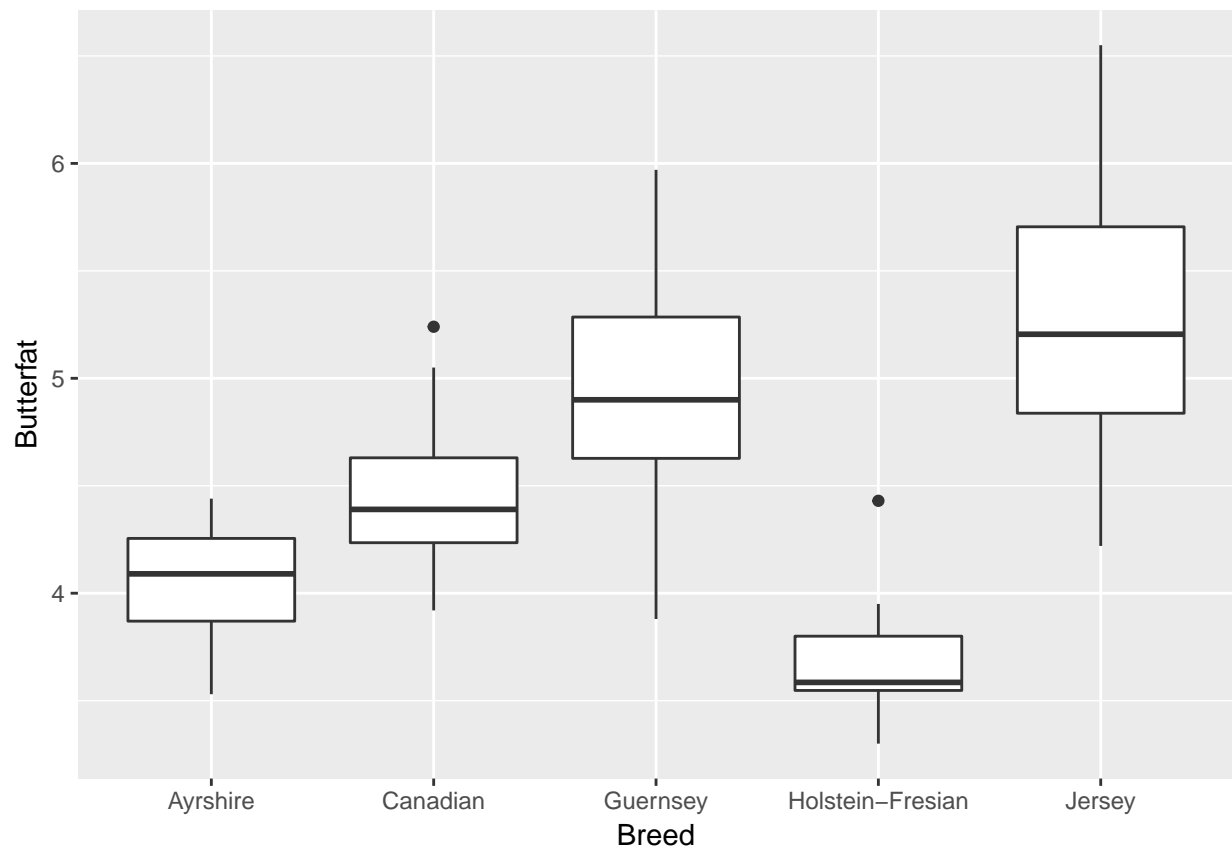
Although the exact p-adjusted values are different between the TukeyHSD and Bonferroni correction (bar A-D), at the 5% significance level, the pairwise tests yield equivalent results.

2.)

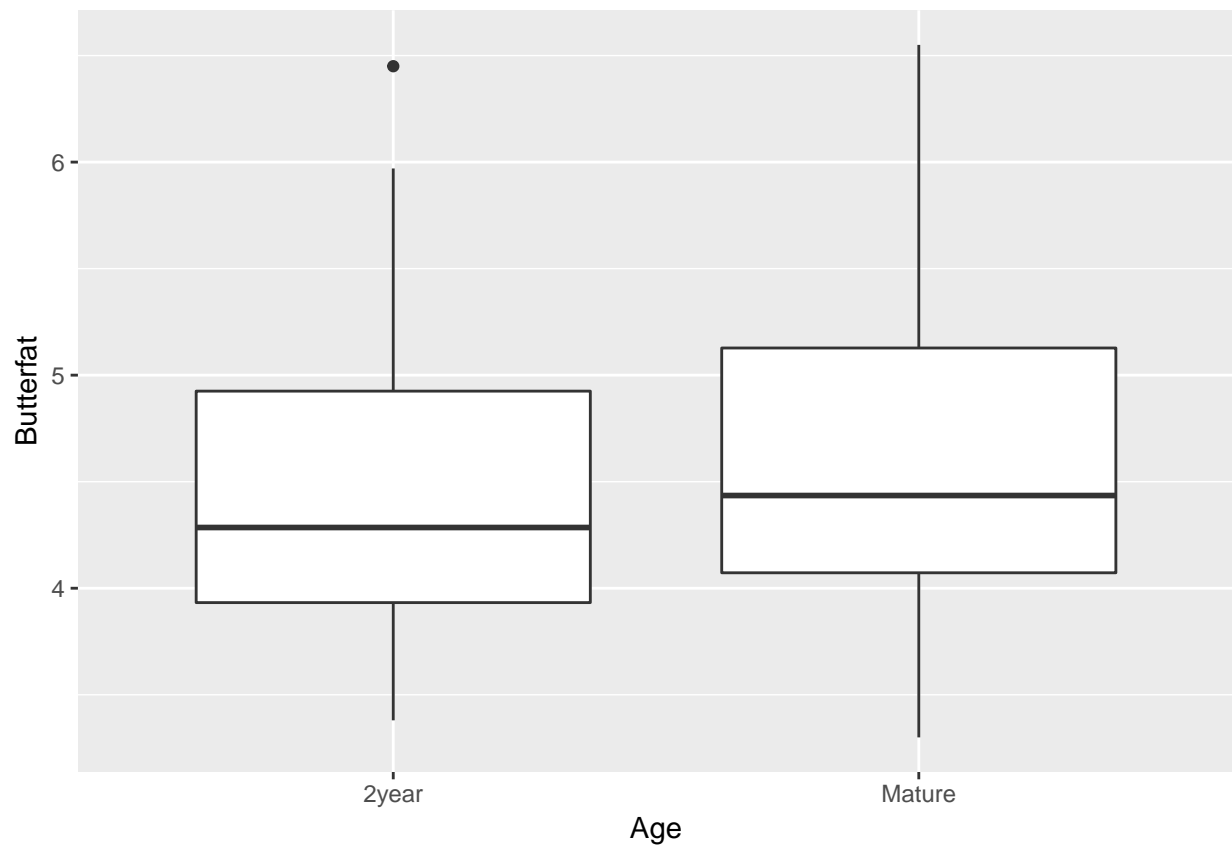
We start by plotting the dependent variable with each factor separately and then with the interaction term.

```
butter_data = faraway::butterfat
interact_term = butter_data$Breed:butter_data$Age
butter_data$interaction = interact_term

ggplot(butter_data, aes(x = Breed, y = Butterfat)) + geom_boxplot()
```

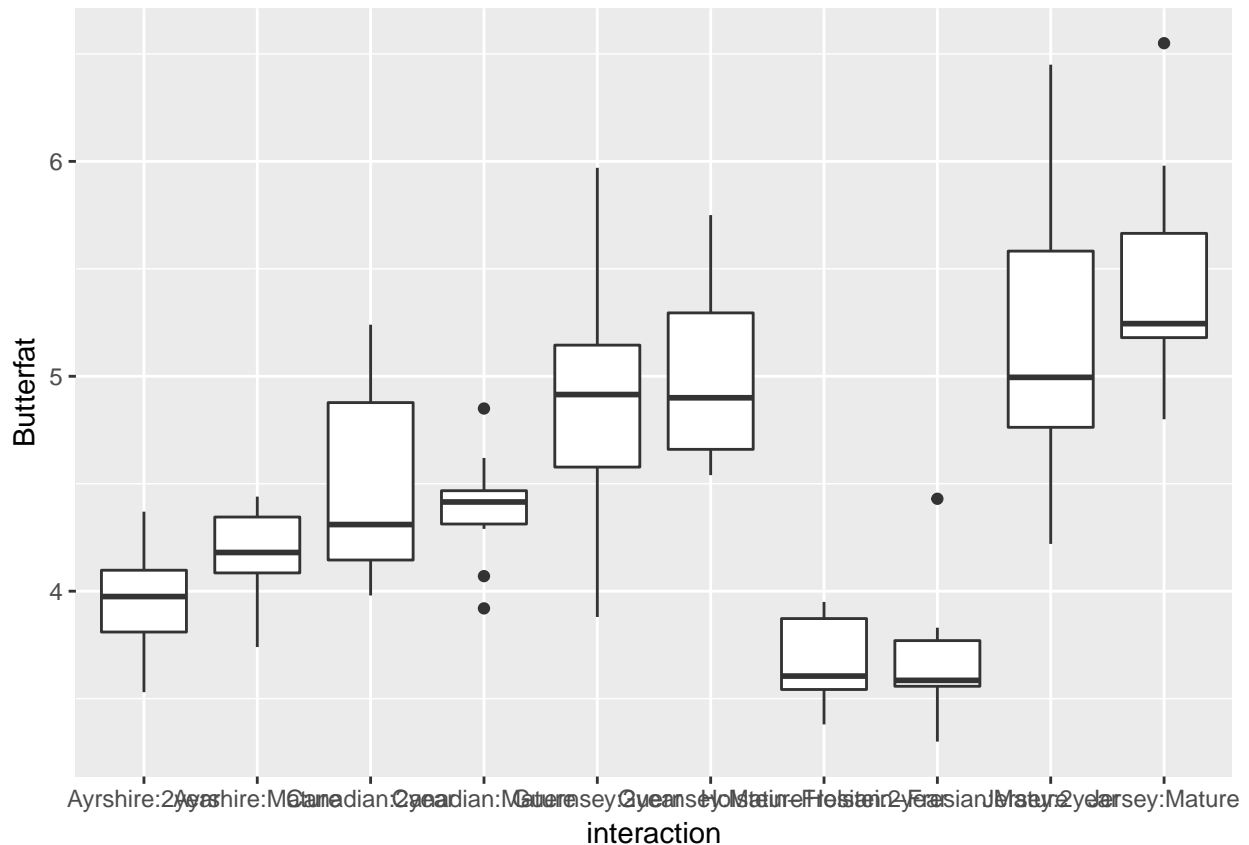


```
ggplot(butter_data, aes(x = Age, y = Butterfat)) + geom_boxplot()
```



```
ggplot(butter_data, aes(x = interaction, y = Butterfat)) + geom_boxplot()
```





```
anova(lm(Butterfat ~ Breed*Age, butterfat))
```

```
## Analysis of Variance Table
##
## Response: Butterfat
##          Df Sum Sq Mean Sq F value Pr(>F)
## Breed      4 34.321  8.5803 49.5651 <2e-16 ***
## Age        1  0.274  0.2735  1.5801 0.2120
## Breed:Age   4  0.514  0.1285  0.7421 0.5658
## Residuals 90 15.580  0.1731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the Age factor and interaction term are not statistically significant, we complete the analysis by performing pairwise tests on Breed:

```
pairwise.t.test(butterfat$Butterfat, butterfat$Breed, p.adjust.method = 'none')
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: butterfat$Butterfat and butterfat$Breed
##
##           Ayrshire Canadian Guernsey Holstein-Fresian
## Canadian    0.00486      -      -      -
## Guernsey     1.0e-09 0.00018      -      -
## Holstein-Fresian 0.00371 6.7e-08 5.5e-16      -
## Jersey       3.3e-15 3.6e-09 0.01054 < 2e-16
```

```
##  
## P value adjustment method: none
```