ST515-HW3

Nora Quick

1. Consider the gravid data set.

```
df = read.table("gravid.txt", header=TRUE);
str(df);
```

```
## 'data.frame':    95 obs. of  3 variables:
##  $ cohort: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ size  : num  4.1 4.15 4.15 4.6 4.4 4.08 4.45 3.98 4.65 4.73 ...
##  $ nembr : int  15 15 20 19 20 18 22 22 20 22 ...
```

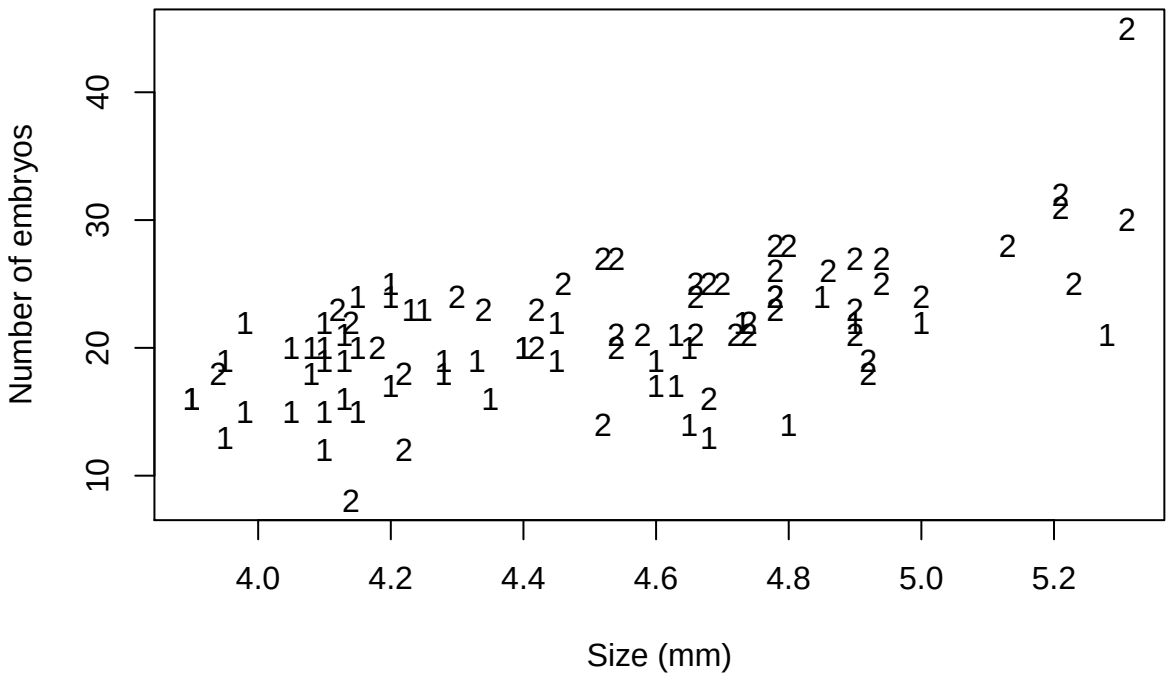
```
head(df);
```

```
##   cohort size nembr
## 1      1 4.10    15
## 2      1 4.15    15
## 3      1 4.15    20
## 4      1 4.60    19
## 5      1 4.40    20
## 6      1 4.08    18
```

```
attach(df);
```

(a) Draw a scatterplot of number of embryos (nembr) vs. carapace length (size), labeling the points with the number of the cohort (1 or 2) to which they belong.

```
## Plot the data: scatter plot
plot(size, nembr, type="n", xlab="Size (mm)", ylab="Number of embryos");
points(size[cohort==1], nembr[cohort==1], pch="1")
points(size[cohort==2], nembr[cohort==2], pch="2")
```



(b1) Create a variable, cohort1, that takes on the value one for females in cohort 1, and the value zero for females in cohort 2:

```
cohort1 <- as.numeric(cohort==1);
```

(b2) For each of the models below, show the regression equation obtained by fitting the model (e.g., nembr = 1.23 + 4.56 cohort1 + 7.89 size), and report the error (residual) sum of squares and error degrees of freedom for the model.

```
# MODEL 1: number of embryos as a function of cohort only
m1 = lm(nembr ~ cohort1)
```

```
# Output of the fit
summary(m1)
```

```
##
## Call:
## lm(formula = nembr ~ cohort1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3333  -2.6135   0.1064   2.6667  21.6667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  23.3333     0.6658   35.04  < 2e-16 ***
## cohort1      -4.4397     0.9466   -4.69 9.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.613 on 93 degrees of freedom
## Multiple R-squared:  0.1913, Adjusted R-squared:  0.1826
## F-statistic:    22 on 1 and 93 DF,  p-value: 9.353e-06
```

```
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: nembr
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## cohort1    1  468.09  468.09  21.995 9.353e-06 ***
## Residuals 93 1979.13   21.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual sum of squares is 1979.13. The residual degrees of freedom is 93.

```
# MODEL 2: number of embryos as a function of cohort and mysid size, main effects only
m2 = lm(nembr ~ cohort1 + size)
```

```
# Output of the fit
summary(m2)
```

```
##
## Call:
## lm(formula = nembr ~ cohort1 + size)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8182  -1.9960   0.0349   2.2773  17.6008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.0067     6.1620  -1.137   0.2585
## cohort1      -2.1240     0.9668  -2.197   0.0305 *
## size          6.4794     1.3098   4.947 3.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.122 on 92 degrees of freedom
## Multiple R-squared:  0.3612, Adjusted R-squared:  0.3473
## F-statistic: 26.01 on 2 and 92 DF,  p-value: 1.115e-09
```

```
anova(m2)
```

```
## Analysis of Variance Table
##
```

```
## Response: nembr
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## cohort1    1  468.09  468.09  27.547 9.792e-07 ***
## size       1  415.82  415.82  24.471 3.388e-06 ***
## Residuals 92 1563.31   16.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual sum of squares is 1563.31. The residual degrees of freedom is 92.

```
# MODEL 3: number of embryos as a function of cohort, mysid size, and an interaction between cohort and

# cohort2 <- as.numeric(cohort==2)

m3 = lm(nembr ~ cohort1 * size)
```

```
# Output of the fit
summary(m3)
```

```
##
## Call:
## lm(formula = nembr ~ cohort1 * size)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5412 -2.4557  0.2548  2.1780 14.9670
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -26.661      8.041  -3.315 0.001315 **
## cohort1        37.168     11.152   3.333 0.001245 **
## size           10.677      1.713   6.232 1.41e-08 ***
## cohort1:size   -8.738      2.472  -3.535 0.000643 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.887 on 91 degrees of freedom
## Multiple R-squared:  0.4383, Adjusted R-squared:  0.4198
## F-statistic: 23.67 on 3 and 91 DF,  p-value: 2.059e-11
```

```
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: nembr
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## cohort1       1  468.09  468.09  30.989 2.599e-07 ***
## size          1  415.82  415.82  27.529 1.003e-06 ***
## cohort1:size  1  188.75  188.75  12.496 0.0006434 ***
## Residuals    91 1374.56   15.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual sum of squares is 1374.56. The residual degrees of freedom is 91.

(c) Use the information from above to select what you consider to be the "best" model of number of embryos as a function of cohort and/or size. Explain your choice. **For this problem, do the model comparisons "by hand" using the formula on slides 20–22 of notes 3.1 "Analysis of Covariance".**

```
# Model 3 & Model 2
x = ((1563.31 - 1374.56) / (92 - 91)) / (1374.56 / 91)
pf(x, 1, 91, lower.tail = FALSE)
```

```
## [1] 0.0006435266
```

From this outcome it appears that Model 3 is better than Model 2 so we will keep Model 3 and compare it to Model 1.

```
y = ((1979.13 - 1374.56) / (93 - 91)) / (1374.56 / 91)
pf(y, 2, 91, lower.tail = FALSE)
```

```
## [1] 6.264217e-08
```

From this outcome it appears that Model 3 is still the "best" model of number of embryos as a function of cohort and/or size. I would conclude this both because of the "by hand" math this question asked for and because it has an interaction term that gives us more information on the number of embryos and size.

(d) Test the hypothesis that cohort does not influence number of embryos in any way—either as a main effect, or as a main effect plus interaction—given that mysid size is in the model. This might involve fitting a model(s) other than the three indicated above. Show your R code, and interpret your result. (For this part, you can use the R function "anova" for model comaprisons.)

```
# Model 4
m4 = lm(nembr ~ cohort1 + cohort1:size)

# Output of cohort 1 as the main effect plus interaction
summary(m4)
```

```
##
## Call:
## lm(formula = nembr ~ cohort1 + cohort1:size)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3333  -2.3800   0.4764   2.1565  21.6667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.3333     0.6664  35.013   <2e-16 ***
## cohort1      -12.8263     9.2041  -1.394    0.167
## cohort1:size   1.9390     2.1168   0.916    0.362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.617 on 92 degrees of freedom
## Multiple R-squared:  0.1986, Adjusted R-squared:  0.1812
## F-statistic:  11.4 on 2 and 92 DF,  p-value: 3.78e-05
```

```
anova(m4)
```

```
## Analysis of Variance Table
##
## Response: nembr
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## cohort1       1  468.09  468.09 21.9574 9.614e-06 ***
## cohort1:size  1   17.89   17.89  0.8391     0.362
## Residuals    92 1961.25   21.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We are testing the hypothesis that cohort does not influence the number of embryos as a main effect or a main effect plus interaction. To accept or reject this I want to look at all four models' p-values.

From Model 4 we can see that there the ANOVA test produces a very small p-value for cohort1 but the interaction is relatively large. This is confusing so I went and checked the p-values of the other models. For all cohort has a relatively small p-value indicating that there is evidence that cohort does influence the number of embryos.

(e) Use the regression output from Model 3 above to write down equations to predict number of embryos from mysid size, for cohort 1 and for cohort 2. On the scatterplot you made in (a), draw in the lines corresponding to these predictive equations, and indicate which line is for which cohort.

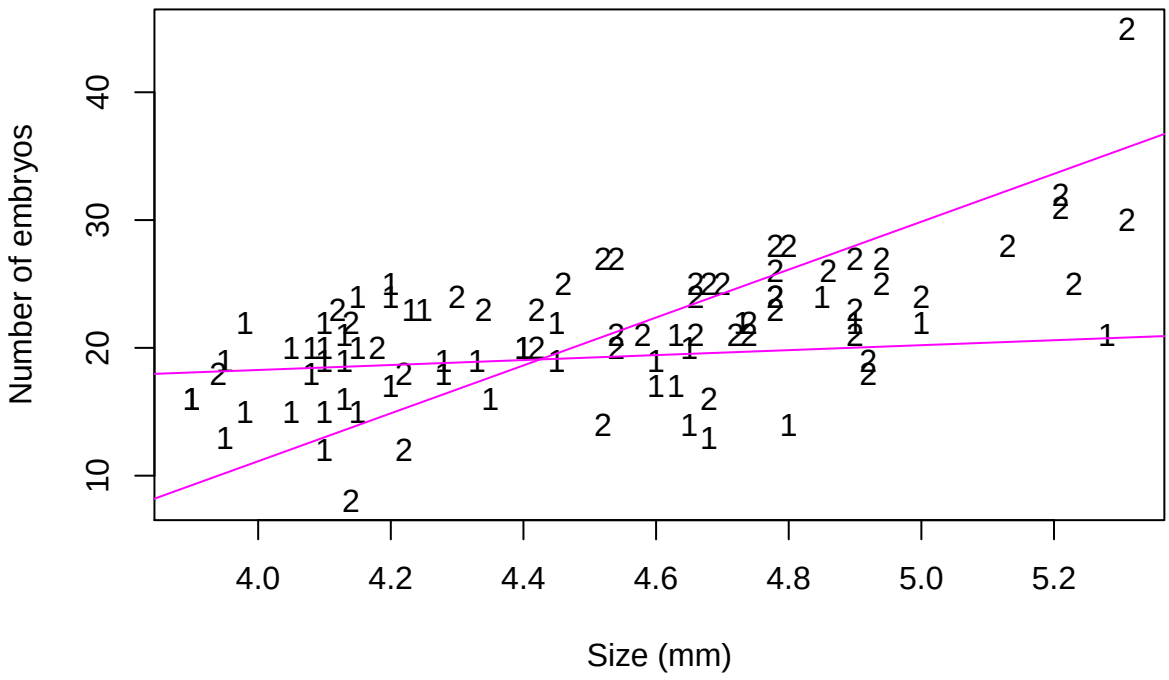
```
coef(m3)
```

```
##  (Intercept)      cohort1         size cohort1:size
##   -26.660522    37.167569    10.676744    -8.737701
```

```
## Plot the data: scatter plot
plot(size, nembr, type="n", xlab="Size (mm)", ylab="Number of embryos");
points(size[cohort==1], nembr[cohort==1], pch="1")
points(size[cohort==2], nembr[cohort==2], pch="2")

# Model 1
a1 = -26.66 + 37.17
b1 = 10.68 + (-8.74)
abline(a1, b1, col="magenta")

# Model 2
a2 = -26.66 + (-37.17)
b2 = 10 + 8.74
abline(a2, b2, col="magenta")
```



(f) After doing all this work, how would you answer the question posed at the beginning of the problem ("Does the number of embryos carried per female differ between cohorts 1 and 2?")?

After doing the plotting work in part (e) I conclude that the number of embryos carried per female depends on the cohort they are in.