Week 2 Problem Based Learning and Practical Solutions

Jeffrey O. Hanson[1]
[1]*School of Biological Sciences, The University of Queensland, Brisbane, QLD, Australia*
Correspondance should be addressed to jeffrey.hanson@uqconnect.edu.au

21 March 2016

Contents

Problem based learning workshop

Activity 1

1. *How does sample size affect your type 1 error?*
    - It doesn't.
2. *How does sample size affect your type 2 error?*
    - Increases in sample size reduce type 2 error.
3. *Discuss some key elements that will maximise the power of your experiment*
    - Effect size - the bigger the expected difference the more power you have.
    - Sample size - the more replicates you have; the more power you have.
4. *If you had limited resources (ie. small N), could you still increase the power of your experiment?*
    - Yes, select treatments that will maximimise the expected effect size.

Activity 2; 1 continuous X variable

1. *Write down the equation that describes the least square regression for the data in example 1 below.*

   - $ntc = 54.382 + (cpd \times 7.977)$

2. *Why are there two t-tests in the table above? How are the t-values calculated? Calculate them and their associated p-values.*

   - The t-tests are testing if the model coefficients are significantly different to zero. The first test is for the intercept and the second for the slope.
   - The t-values are calculated by dividing the term estimate ("Estimate") by the the uncerainty ("Std. Error) surrounding this estimate.
   - The t-values are then converted to p-values using the table in the manual. Alternatively, use the R code below, where `x` is the t-statistic and `y` is the residual degrees of freedom in the model.

```
2 * pt(abs(x), df=y, lower=FALSE)`
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	54.382	42.659	1.275	0.214
cpd	7.977	1.807	4.414	0.0001

   -

3. *Calculate the F value for the ANOVA table above and its associated p-value?*

   - For each term (row in the table), the mean sums of squares ("Mean Sq") are calculated by dividing the sum of squares ("Sum Sq") by the degrees of freedom ("Df").
   - The F-statistic ("F value") is then calculated by diving the mean sum of squares for the slope (512522) by the residual mean sum of squares for the model (26288.85).
   - The p-value is then calculated using the `R` code below, where `x` is the F-value (19.498), `df1` is the degrees of freedom consumed by the slope (1) and `df2` is the residual degrees of freedom (26).

```
pf(abs(x), df1, df2, lower=F)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cpd	1	512522	512522	19.498	0.0001
Residuals	26	683510	26288.85		

4. *What can you say about the relationship between smoking cancer according to the results above?*

Please comment on both the null hypothesis being tested and also on the fit of the model (hint: also calculate R^2)

- There seems to be a positive linear relationship between the number of tumor cells and the number of cigarettes smoked per day. The intercept was not found to be significantly different to zero. This suggests that individuals who did not smoke any cigarettes were not associated with any cancer cells. The slope of the regression was, however, found to be significantly different to zero. This suggests that individuals who smoke cigarettes had more tumor cells. This model explained a large amount of variation ($R^2 = \frac{512522}{512522+683510} = 0.95$). **However, we should also note that this model is totally inappropriate for analyzing this data since we are using count data. Stay tuned for later in the course when we learn how to deal with this.**

Activity 2; 2 nominal X variables

1. *Write down the equation that describes the least square regression for the data in example 2 on the next page.*

   - $Rate = 0.3800 + (0.33857 \times X_1) + (-0.01667 \times X_2) + (-0.03524 \times X_3)$

2. *What can you conclude about the effect of team and Experiment on the Rate response?* Looking at the boxplots, there doesn't seem to be much difference between Team A or Team B. The treatment group has slightly higher values than the control group.

3. *Is there a significant interactions between TEAM and EXPERIMENT?*

   a) *Test significance of each term*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.38	0.045	8.444	<0.001
ExperimentTreatment	0.339	0.064	5.297	<0.001
TeamB	-0.017	0.067	-0.254	0.802
ExperimentTreatment:TeamB	-0.035	0.095	-0.368	0.716

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Experiment	1	0.675	0.675	48.214	<0.001
Team	1	0.008	0.008	0.571	0.458
Experiment:Team	1	0.002	0.002	0.143	0.709
Residuals	22	0.318	0.014		

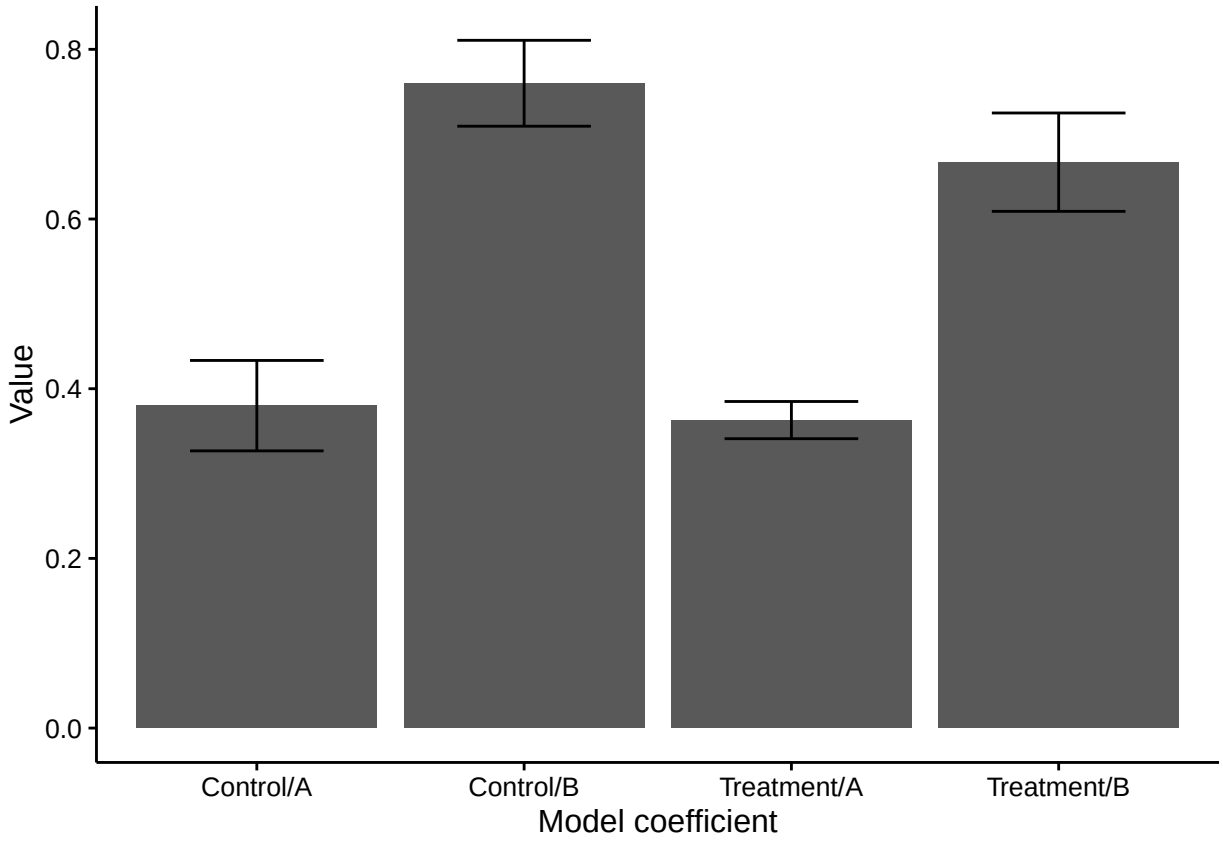
   b) *Calculate the mean for each group using the information above.*

Group	Intercept (0.38)	X_1 (0.34)	X_2 (-0.02)	X_3 (-0.04)	Prediction
Control, A	1	0	0	0	0.38
Control, B	1	1	0	0	0.76
Treatment, A	1	0	1	0	0.363
Treatment, B	1	1	1	1	0.667

The predictions are calculated by multiplying the values in each column by the coefficient for that column, and then summing all those values togeather. For example, the prediction for 'Control, A' in the first row is derived using: $(1 \times 0.38) + (0 \times 0.34) + (0 \times -0.02) + (0 \times 0.04)$

c) *Draw a single bar graph the relationship amongst the four group, and add approximate standard errors to each mean.*

```r
# init
library(ggplot2)
# create data.frame with means and sds
df1 <- data.frame(
    group=c('Control/A','Control/B', 'Treatment/A', 'Treatment/B'),
    mean=c(0.38,0.76,0.363,0.667),
    sd=c(0.141,0.124,0.058,0.142),
    n=c(7,6,7,6)
)
# calculate approx errors
df1$se <- df1$sd / sqrt(df1$n)
# make plot
ggplot(data=df1) +
    geom_bar(aes(x=group,y=mean), stat='identity') +
    geom_errorbar(aes(x=group, ymin=mean-se, ymax=mean+se), width=0.5) +
    theme_classic() +
    xlab('Model coefficient') +
    ylab('Value')
```



R practical session

P1. Entering the data

```
# Enter the data from Research Team A.
ControlA <- c(0.2, 0.4, 0.5, 0.38, 0.6, 0.2, 0.8, 0.4, 0.4, 0.2)
TreatmentA <- c(0.6, 0.8, 0.7, 0.8, 0.7, 0.6, 0.3, 0.6, 0.5, 0.9)
# Enter the data from Research Team B.
ControlB <- c(0.3, 0.6, 0.2)
TreatmentB <- c(0.99, 0.7, 0.6)
# Check the data
ControlA ; TreatmentA
```

```
##  [1] 0.20 0.40 0.50 0.38 0.60 0.20 0.80 0.40 0.40 0.20
```

```
##  [1] 0.6 0.8 0.7 0.8 0.7 0.6 0.3 0.6 0.5 0.9
```

```
ControlB ; TreatmentB
```

```
## [1] 0.3 0.6 0.2
```

```
## [1] 0.99 0.70 0.60
```

P2. Calculating summary statistics

```
NA
meanControlA <- mean(ControlA) # calculate the mean of TeamA control group
meanTreatmentA <- mean(TreatmentA) # calculate the median of TeamA control group
NA
range(ControlA)                  # calculate the range
```

```
## [1] 0.2 0.8
```

```
var(ControlA)                    # calculate the variance
```

```
## [1] 0.03664
```

```
sd(ControlA)                 # calculate the standard deviation
```

```
## [1] 0.1914158
```

```
sqrt(var(ControlA))                  # another way to calculate standard deviation
```

```
## [1] 0.1914158
```

```
sd(ControlA)/sqrt(length(ControlA))     # standard error
```

```
## [1] 0.06053098
```

```
# Repeat all of the above for Research Team B.
```

P3. Calculating summary statistics across a data frame

```
teamA <- data.frame(cbind(ControlA ,TreatmentA))      # make a data frame
teamA                                                 # have a look at it
```

```
##    ControlA TreatmentA
## 1      0.20        0.6
## 2      0.40        0.8
## 3      0.50        0.7
## 4      0.38        0.8
```

```
## 5      0.60        0.7
## 6      0.20        0.6
## 7      0.80        0.3
## 8      0.40        0.6
## 9      0.40        0.5
## 10     0.20        0.9
```

```
sapply(teamA,mean)              # calculate the mean of each column
```

```
##   ControlA TreatmentA
##      0.408      0.650
```

```
sapply(teamA,var)               # calculate the variance of each column
```

```
##   ControlA TreatmentA
## 0.03664000 0.02944444
```

```
sapply(teamA,sum)               # calculate the sum of each column
```

```
##   ControlA TreatmentA
##       4.08       6.50
```

```
sapply(teamA,min)               # calculate the minimum of each column
```

```
##   ControlA TreatmentA
##        0.2        0.3
```

```
sapply(teamA,max)               # calculate the maximum of each column
```

```
##   ControlA TreatmentA
##        0.8        0.9
```

```
sapply(teamA, quantile)             # median and quartiles of each column
```

```
##      ControlA TreatmentA
## 0%      0.200      0.300
## 25%     0.245      0.600
## 50%     0.400      0.650
## 75%     0.475      0.775
## 100%    0.800      0.900
```

```r
sapply(teamA, range)            # calculate the range of each column
```

```
##      ControlA TreatmentA
## [1,]      0.2        0.3
## [2,]      0.8        0.9
```

```r
sapply(teamA, length)           # calculate the sample size of each column
```

```
##   ControlA TreatmentA
##         10         10
```

```r
# Repeat all of the above for Research Team B.
```

P4. Hypothesis testing: the T-test

```r
attach(teamA)           # make 'teamA' the default data set
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     ControlA, TreatmentA
```

```r
t.test(ControlA, TreatmentA,var.equal=TRUE) # T-test: ControlA versus TreatmentA
```

```
##
##  Two Sample t-test
##
## data:  ControlA and TreatmentA
## t = -2.9769, df = 18, p-value = 0.008081
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.41278887 -0.07121113
## sample estimates:
## mean of x mean of y
##     0.408     0.650
```

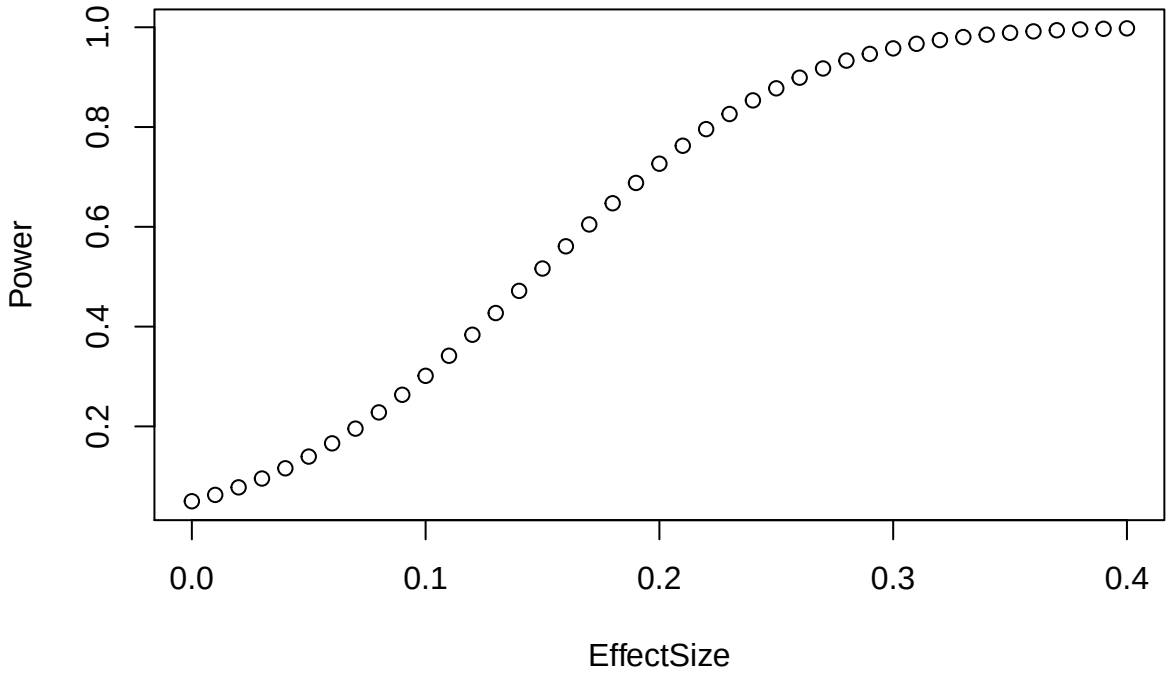
```r
power.t.test(    # calculate Power of T-test, and output the variable set "NULL"
n = 10,          # sample size of each group
delta =  mean(TreatmentA) - mean(ControlA), # difference between the two means
sd =  sd(ControlA),  # set a common standard deviation
sig.level = 0.05, # set a Type I error rate
power = NULL, # set the statistical power (1 - Type II error rate)
type = "two.sample", # request a T-test comparing means of two groups
alternative = "one.sided" # alternative to null hypothesis is Treatment > Control
) # end  "power.t.test("  command
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 0.242
##              sd = 0.1914158
##       sig.level = 0.05
##           power = 0.8584562
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

```
# Repeat all of the above for Research Team B.
```

P5. Graph of power versus effect size

```
EffectSize <- seq(0.0, 0.4, 0.01)   # set various effects sizes
Power <- vector()           # request space to hold the power estimates
length(Power) = length(EffectSize)  # allocate space to hold the power estimates
for (i in 1:length(EffectSize) ){   # loop through the effect sizes
Power[i] <- power.t.test(       # record the output at the "ith" place in "Power"
n = 10,             # sample size of each group
delta = EffectSize[i],     # "ith" effect size : difference between the two means
sd =  sd(ControlA),        # set a common standard deviation
sig.level = 0.05,          # set a Type I error rate
power = NULL,              # set the statistical power (1 - Type II error rate)
type = "two.sample",       # request a T-test comparing means of two groups
alternative = "one.sided")$power} # select the element called power then re-loop
plot(EffectSize,Power,type="b") # graph power against effect size
```




```
# Repeat all of the above for Research Team B.
```

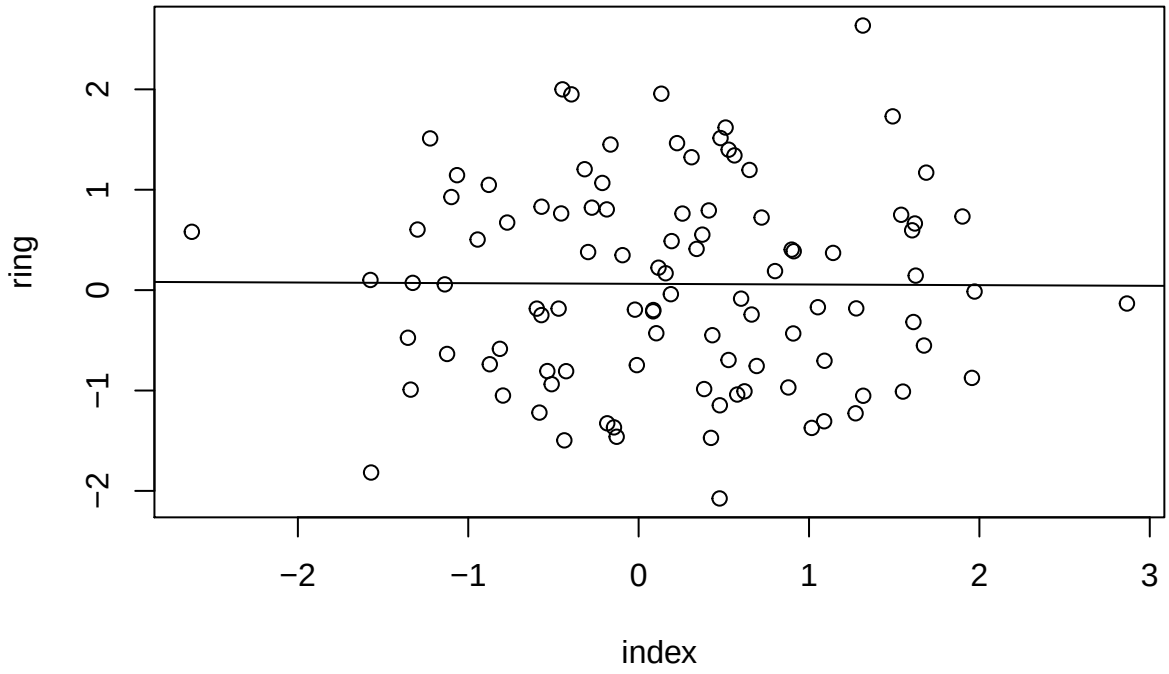
P6. Regression on single, continuous explanatory variable

```
FingerData <- data.frame(index=rnorm(100), ring=rnorm(100),
gender=sample(c('m', 'f'), 100, replace=TRUE)) # create FingerData table
attach(FingerData)      # make 'FingerData' the default data set
plot(index,ring)        # plot "index" (X) against "ring" (Y)
fit <- lm( ring ~ index )   # regress "ring" (Y) on "index" (X)
summary(fit)            # summary of regression coefficients
```

```
##
## Call:
## lm(formula = ring ~ index)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.13453 -0.81080 -0.03826  0.69896  2.58140
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.062658   0.102156   0.613    0.541
## index       -0.006559   0.103065  -0.064    0.949
##
## Residual standard error: 0.9994 on 98 degrees of freedom
## Multiple R-squared:  4.132e-05,  Adjusted R-squared:  -0.01016
## F-statistic: 0.00405 on 1 and 98 DF,  p-value: 0.9494
```

```
abline(fit)          # plot of "ring" versus "index" with regression line added
```




```
anova(fit)           # ANOVA table
```

```
## Analysis of Variance Table
##
## Response: ring
##           Df Sum Sq Mean Sq F value Pr(>F)
## index      1  0.004 0.00404   0.004 0.9494
## Residuals 98 97.874 0.99872
```

```
names(fit)           # list the elements of "fit"
```

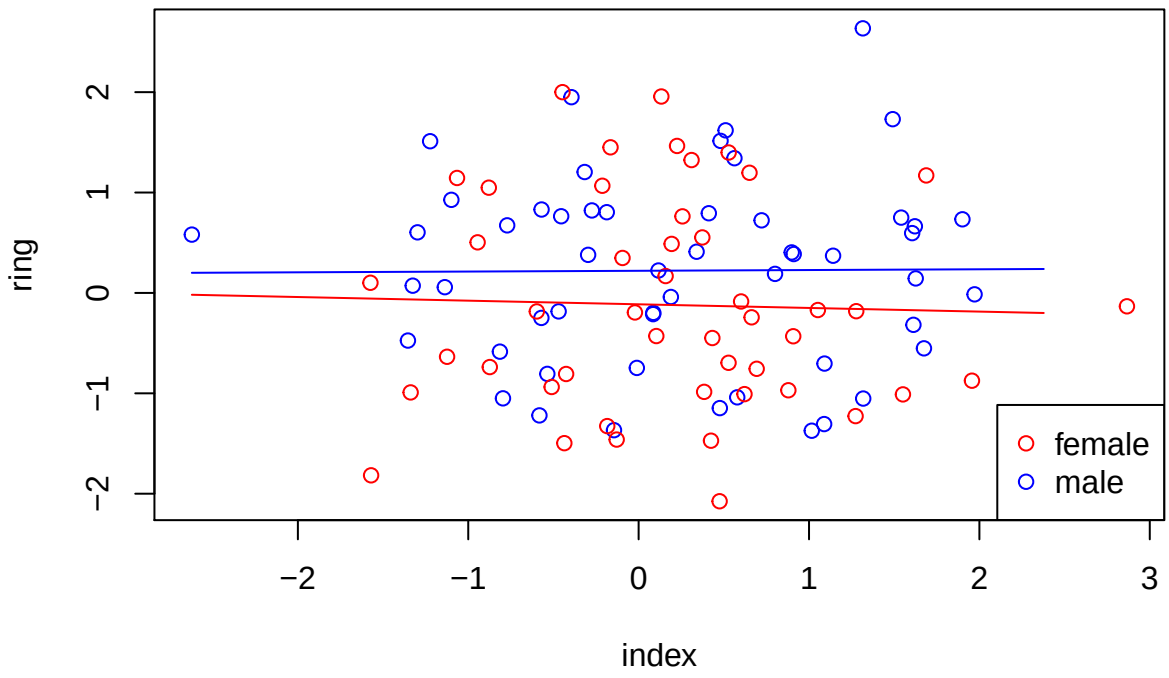
```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
names(summary(fit))     # list the elements of "summary(fit)"
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

P7. Regression on two explanatory variables: one nominal, one continuous

```
plot(index,ring,type="n")                    # set up an empty graph
points(index[gender=="m"],ring[gender=="m"],col="blue") # add points for males
points(index[gender=="f"],ring[gender=="f"],col="red")  # add points for females
# add a legend
legend("bottomright",legend=c("female","male"),pch=1,col=c("red","blue"))
fit <- lm( ring ~ index * gender )  # regress 'ring' on 'index' and 'gender'
beta <- fit$coefficients        # extract the regression coefficients from 'fit'
xvalues <- seq(min(index), max(index), 1)   # set up trial values for X
# predict Y values for females
yfemales <- beta[1] + beta[2]*xvalues
# predict Y values for males
ymales <- (beta[1] + beta[3]) + (beta[2] + beta[4])*xvalues
lines(xvalues,ymales,type="l",col="blue")   # plot the regression line for males
lines(xvalues,yfemales,type="l",col="red")  # plot the regression line for females
```



P8. A data set containing two nominal, explanatory variables

```
TreatmentA <- c(0.6, 0.8, 0.7, 0.8, 0.7, 0.6, 0.3, 0.6, 0.5, 0.9)
ControlA <- c(0.2, 0.4, 0.5, 0.38, 0.6, 0.2, 0.8, 0.4, 0.4, 0.2)
TreatmentB <- c(0.99, 0.7, 0.6)
ControlB <- c(0.3, 0.6, 0.2)
teamA <- data.frame(cbind(ControlA ,TreatmentA)) # make a data frame for 'teamA'
teamB <- data.frame(cbind(ControlB ,TreatmentB)) # make a data frame for 'teamB'
colnames(teamA) <- c("Control", "Treatment")
colnames(teamB) <- c("Control", "Treatment")
dataA <- stack(teamA)   # stack 'ControlA' and 'TreatmentA'
dataB <- stack(teamB)   # stack 'ControlB' and 'TreatmentB'
dataA$team <- 'teamA' ;  dataB$team <- 'teamB' # add indicators for 'team'
CancerStudy <- data.frame(rbind(dataA,dataB))  # assemble everything into data frame
CancerStudy          # have a look at the data set you have created
```

```
##    values       ind  team
## 1    0.20   Control teamA
## 2    0.40   Control teamA
## 3    0.50   Control teamA
## 4    0.38   Control teamA
```

```
## 5    0.60   Control teamA
## 6    0.20   Control teamA
## 7    0.80   Control teamA
## 8    0.40   Control teamA
## 9    0.40   Control teamA
## 10   0.20   Control teamA
## 11   0.60 Treatment teamA
## 12   0.80 Treatment teamA
## 13   0.70 Treatment teamA
## 14   0.80 Treatment teamA
## 15   0.70 Treatment teamA
## 16   0.60 Treatment teamA
## 17   0.30 Treatment teamA
## 18   0.60 Treatment teamA
## 19   0.50 Treatment teamA
## 20   0.90 Treatment teamA
## 21   0.30   Control teamB
## 22   0.60   Control teamB
## 23   0.20   Control teamB
## 24   0.99 Treatment teamB
## 25   0.70 Treatment teamB
## 26   0.60 Treatment teamB
```

P9. Regression with two nominal variables

```
# regress 'team' and 'treatment' (ind) onto survival
fit <- lm(values~ind*team, data=CancerStudy)
summary(fit)                    # look at the regression coefficients
```

```
##
## Call:
## lm(formula = values ~ ind * team, data = CancerStudy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3500 -0.1292 -0.0180  0.1355  0.3920
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.40800    0.05891   6.926 5.92e-07 ***
## indTreatment             0.24200    0.08331   2.905  0.00822 **
## teamteamB               -0.04133    0.12263  -0.337  0.73927
## indTreatment:teamteamB   0.15467    0.17343   0.892  0.38214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1863 on 22 degrees of freedom
## Multiple R-squared:  0.4119, Adjusted R-squared:  0.3317
## F-statistic: 5.137 on 3 and 22 DF,  p-value: 0.00763
```

```
anova(fit)                    # look at the ANOVA table
```

```
## Analysis of Variance Table
##
## Response: values
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## ind        1 0.50123 0.50123 14.4430 0.0009801 ***
## team       1 0.00598 0.00598  0.1724 0.6820460
## ind:team   1 0.02760 0.02760  0.7954 0.3821379
## Residuals 22 0.76349 0.03470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2)) # set up plotting device as 2x2 grid
plot(fit)                     # look at various diagnostic graphics
```