Stat 355 Assignment3

Zimeng Ming V00844078

2019/10/15

Multiple regression and Dummy Variables, Section 11.2

Instructions:

1. Complete your assignment in R Markdown using this file as a template. Insert R code in the R chunks, and type in your response after the corresponding R chunk leaving one blank line between the R chunk and your comments.
2. Execute each line of code separately to ensure that it works properly.
3. Either [knit the entire document to pdf] or [knit to HTML or Word and print to pdf].
4. Submit the pdf file to CourseSpaces in the Assignment 3 activity.

Data Description:

Question 11.Review.15(new book) 11.Review.17(old book)

In a study, subjects were preterm infant with low birth weights born in three different hospitals. The variables are:

```
WEIGHT:  weight in kg  (Y variable)
WEEKS:  gestation age in weeks
HOSP:  Hospital of birth, A, B or C
```

0. Read the data into R using the read.csv function.

```
knitr::opts_chunk$set(fig.width=8, fig.height=6) #set size of graphs
HOSPB<-read.csv('REV_C11_17.csv')
dim(HOSPB)
```

```
## [1] 40  3
```

```
HOSPB$HOSPn <- as.numeric(HOSPB$HOSP)  #create a numeric HOSP for plotting
```

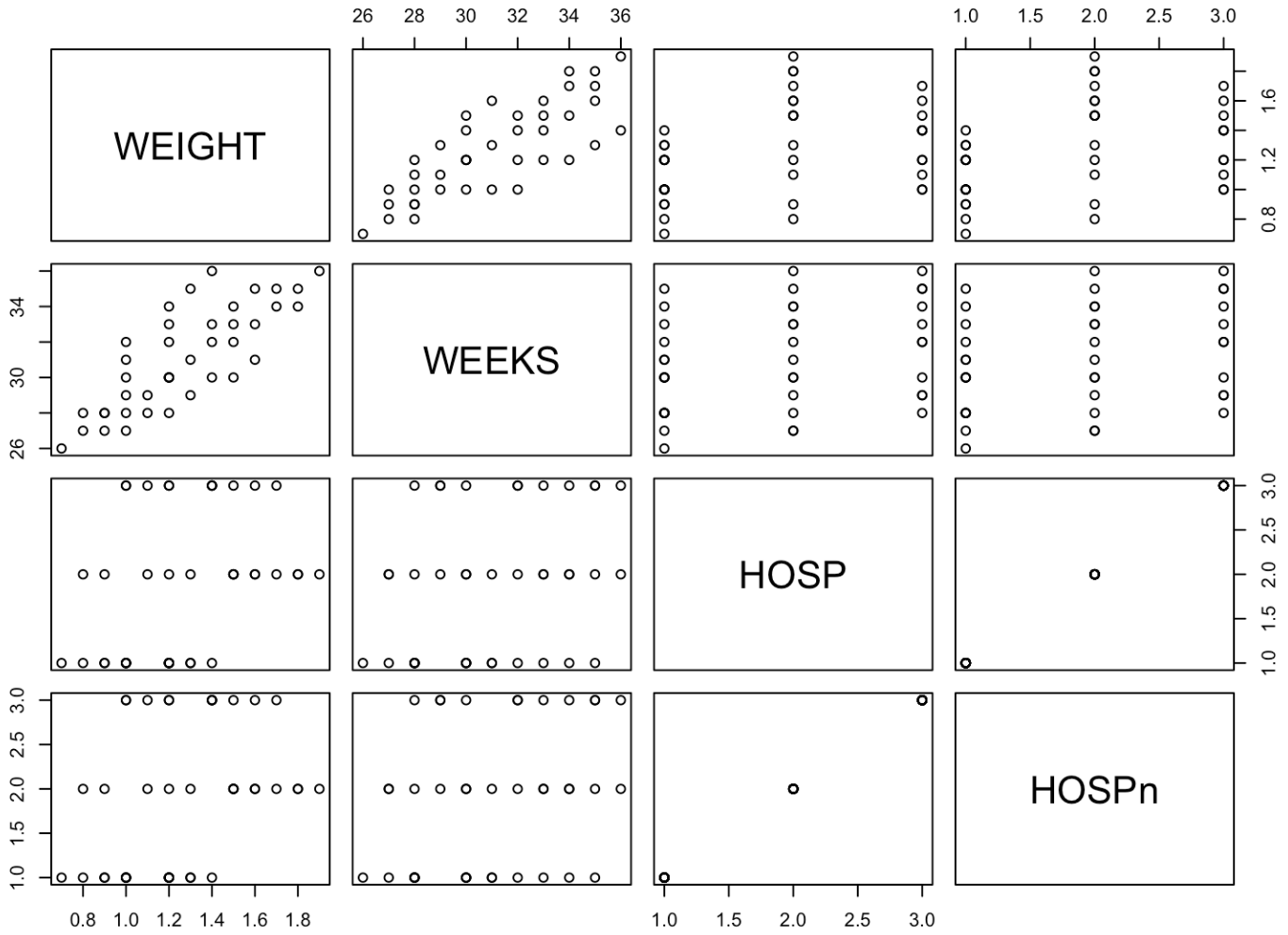
Questions:

1. Provide descriptive statistics and a pairs plot for the data.
   Comment on your results and especially on any unusual features in the data. (3 marks)

```
summary(HOSPB)
```

```
##      WEIGHT          WEEKS       HOSP       HOSPn
##  Min.   :0.70   Min.   :26.00   A:15   Min.   :1.0
##  1st Qu.:1.00   1st Qu.:28.75   B:14   1st Qu.:1.0
##  Median :1.20   Median :31.00   C:11   Median :2.0
##  Mean   :1.27   Mean   :31.07          Mean   :1.9
##  3rd Qu.:1.50   3rd Qu.:33.25          3rd Qu.:3.0
##  Max.   :1.90   Max.   :36.00          Max.   :3.0
```

```
plot(HOSPB)
```



Comment:
From the data summary, We can know the min value for weight is 0.7 and the mean is higher then median,so it is a little bit skew to left,the max value is 1.9.

The min value for Week is 26 and the mean is a little higher then median but not too much ,so it is seems normal distribution,the max value is 36.

The min value for HOSPn is 1 and the max value is 3, seems normal distribution since the median and mean and the 1st qu are looks 1/3 for the 3rd qu.

Also from the pairs plot, it is very clearly that only weight and weeks are looks has linear reationships, the others are no relationship with other variables.

---

2. Provide summary statistics of the dataset by HOSP and comment. (3 marks)

```
HOSPB$HOSP<-as.factor(HOSPB$HOSP)
by(HOSPB,HOSPB$HOSP,summary)
```

```
## HOSPB$HOSP: A
##      WEIGHT          WEEKS       HOSP       HOSPn
##  Min.   :0.700   Min.   :26.00   A:15   Min.   :1
##  1st Qu.:0.950   1st Qu.:28.00   B: 0   1st Qu.:1
##  Median :1.000   Median :30.00   C: 0   Median :1
##  Mean   :1.073   Mean   :30.07          Mean   :1
##  3rd Qu.:1.200   3rd Qu.:31.50          3rd Qu.:1
##  Max.   :1.400   Max.   :35.00          Max.   :1
## ------------------------------------------------------
## HOSPB$HOSP: B
##      WEIGHT          WEEKS       HOSP       HOSPn
##  Min.   :0.800   Min.   :27.00   A: 0   Min.   :2
##  1st Qu.:1.225   1st Qu.:29.25   B:14   1st Qu.:2
##  Median :1.500   Median :31.50   C: 0   Median :2
##  Mean   :1.443   Mean   :31.36          Mean   :2
##  3rd Qu.:1.675   3rd Qu.:33.75          3rd Qu.:2
##  Max.   :1.900   Max.   :36.00          Max.   :2
## ------------------------------------------------------
## HOSPB$HOSP: C
##      WEIGHT          WEEKS       HOSP       HOSPn
##  Min.   :1.000   Min.   :28.00   A: 0   Min.   :3
##  1st Qu.:1.150   1st Qu.:29.50   B: 0   1st Qu.:3
##  Median :1.400   Median :32.00   C:11   Median :3
##  Mean   :1.318   Mean   :32.09          Mean   :3
##  3rd Qu.:1.450   3rd Qu.:34.50          3rd Qu.:3
##  Max.   :1.700   Max.   :36.00          Max.   :3
```

Comment: There are 15 A's , 14 B's and 11 C's in the data. As for the HOSPn, A is 1, B is 2 and C is 3.

As for A, The minimum weight is 0.7 and the median is lower then mean, the max weight is 1.4. In terms of week, the weeks range is from 26 to 35, the median and mean are close but mean is a little bit higher then median.

As for B, The minimum weight is 0.8 and the median is higher then mean, the max weight is 1.9. In terms of week, the weeks range is from 27 to 36, the median and mean are close but mean is a little bit lower then median.

As for B, The minimum weight is 1.0 and the median is higher then mean, the max weight is 1.7. In terms of week, the weeks range is from 28 to 36, the median and mean are close but mean is a little bit higher then median.

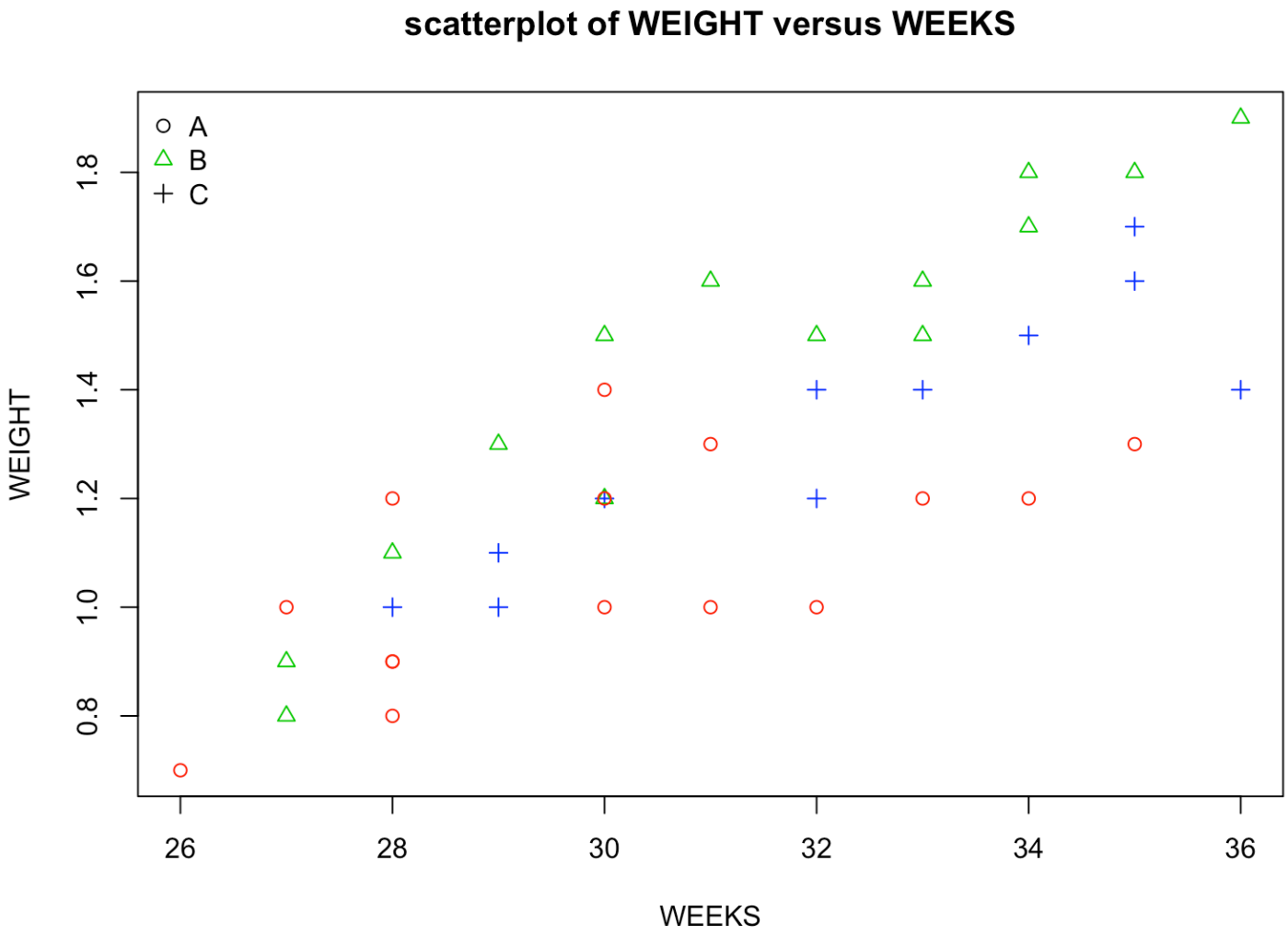
According to those three data, we can see that the Hospital C has less range of weight, but the Hosiptal B has the highest average on mean and median as well. Also Hospital C has less range of weeks and it also has the highest median and mean weeks in three hospital.

---

3. Provide a scatterplot of WEIGHT versus WEEKS, using a different plotting character and colour for each of the three hospitals. Comment on the graph. (2 marks)

```
with(HOSPB,plot(WEEKS, WEIGHT, pch=HOSPn, col=(1 + HOSPn),
     ylab='WEIGHT',xlab='WEEKS', main="scatterplot of WEIGHT versus WEEKS"))

legend('topleft', c('A', 'B','C'), pch=1:3, col=c(1,3),
       bty="n")
```



Comment:
From the graph, the red circle is represent Hosipital A, the blue cross is represent Hosipital B, the green triangle is represent Hosipital C

As we can see from the graph, the hospital A,B,C are all looks have the linear relationships for weight and weeks. The hospital B are looks very clearly has an model line, Hosipital A has very large variance and the model is not clearly. Hosiptal C are looks has linear models between weight and weeks.

4. Fit parallel lines model for WEIGHT versus WEEKS, where the intercepts may differ by HOSP but the slopes are the same. Provide a summary of the model and comment. (4 marks)

```
fit_WW<-lm(HOSPB$WEIGHT~HOSPB$WEEKS+HOSPB$HOSP)
summary(fit_WW)
```

```
##
## Call:
## lm(formula = HOSPB$WEIGHT ~ HOSPB$WEEKS + HOSPB$HOSP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29535 -0.10759  0.00870  0.08741  0.33198
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.324633   0.266333  -4.974 1.63e-05 ***
## HOSPB$WEEKS  0.079755   0.008766   9.098 7.28e-11 ***
## HOSPB$HOSPB  0.266602   0.056338   4.732 3.40e-05 ***
## HOSPB$HOSPC  0.083405   0.061567   1.355    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1485 on 36 degrees of freedom
## Multiple R-squared:  0.7821, Adjusted R-squared:  0.7639
## F-statistic: 43.07 on 3 and 36 DF,  p-value: 5.342e-12
```

Comment:
The estimate model line is WEIGHT=-1.324633 +0.079755*WEEKS +0.266602*HOSPB+ 0.083405*HOSPC+Error

The Degree of freedom is 36, the R_square is 0.7821

From the F-test: p-value is very small, so we have strong evidence to against H0, we can conclude at least one of the beta is significant.

From the sumamry, we can see that the p-vlaue for beta1 is really low so we have strong evidence against H0, so we can say the slope for Weeks(beta1) is significant.

But we can see that only the weight and weeks for hospital B and C are looks has linear models, beta(HOSPB) is very small so we have strong evidence agianst H0 and also the p-value for the hosipital C is bigger then 0.05. So we can conclude only the beta for hospital B is significant.

---

5. Fit a model for WEIGHT versus WEEKS, where the intercepts and slopes may differ by HOSP. Provide a summary of the model and comment. (4 marks)

```
fit_gww<-lm(HOSPB$WEIGHT~HOSPB$WEEKS+HOSPB$HOSP+HOSPB$WEEKS*HOSPB$HOSP)
summary(fit_gww)
```

```
##
## Call:
## lm(formula = HOSPB$WEIGHT ~ HOSPB$WEEKS + HOSPB$HOSP + HOSPB$WEEKS *
##     HOSPB$HOSP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21714 -0.07487 -0.01513  0.07181  0.32992
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.39278    0.40931  -0.960  0.34403
## HOSPB$WEEKS               0.04876    0.01357   3.595  0.00102 **
## HOSPB$HOSPB              -1.56716    0.56897  -2.754  0.00938 **
## HOSPB$HOSPC              -0.74327    0.63875  -1.164  0.25267
## HOSPB$WEEKS:HOSPB$HOSPB  0.05976    0.01848   3.233  0.00272 **
## HOSPB$WEEKS:HOSPB$HOSPC  0.02772    0.02039   1.359  0.18310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1336 on 34 degrees of freedom
## Multiple R-squared:  0.8336, Adjusted R-squared:  0.8091
## F-statistic: 34.06 on 5 and 34 DF,  p-value: 2.616e-12
```

The estimate model line is WEIGHT=-0.39278 +0.04876*WEEKS -1.56716*HOSPB-0.74327 *HOSPC+0.05976WEEKS*HOSPB+0.02772WEEKS*HOSPC+Error

The Degree of freedom is 34, the R_square is 0.8336

From the F-test: p-value is very small, so we have strong evidence to against H0, we can conclude at least one of the betas is significant.

From the sumamry, we can see that the p-vlaue for beta1(WEEKS) is really low so we have strong evidence against H0, so we can say the slope for Weeks(beta1) is significant.

As for the p-values of HOSPB and WEEKS*HOSPB is very small, so we have strong evidence against H0, so we can conclude that those two betas are significant. However, the P-value is very high when the HOSPC in the model, so beta(HOSPC) and beta(WEEKS*HOSPC) are not significant.

---

6. Provide a scatterplot of WEIGHT versus WEEKS, using a different plotting character and colour for each of the three hospitals. Overlay the fitted lines from the model in Question 5. (2 marks)

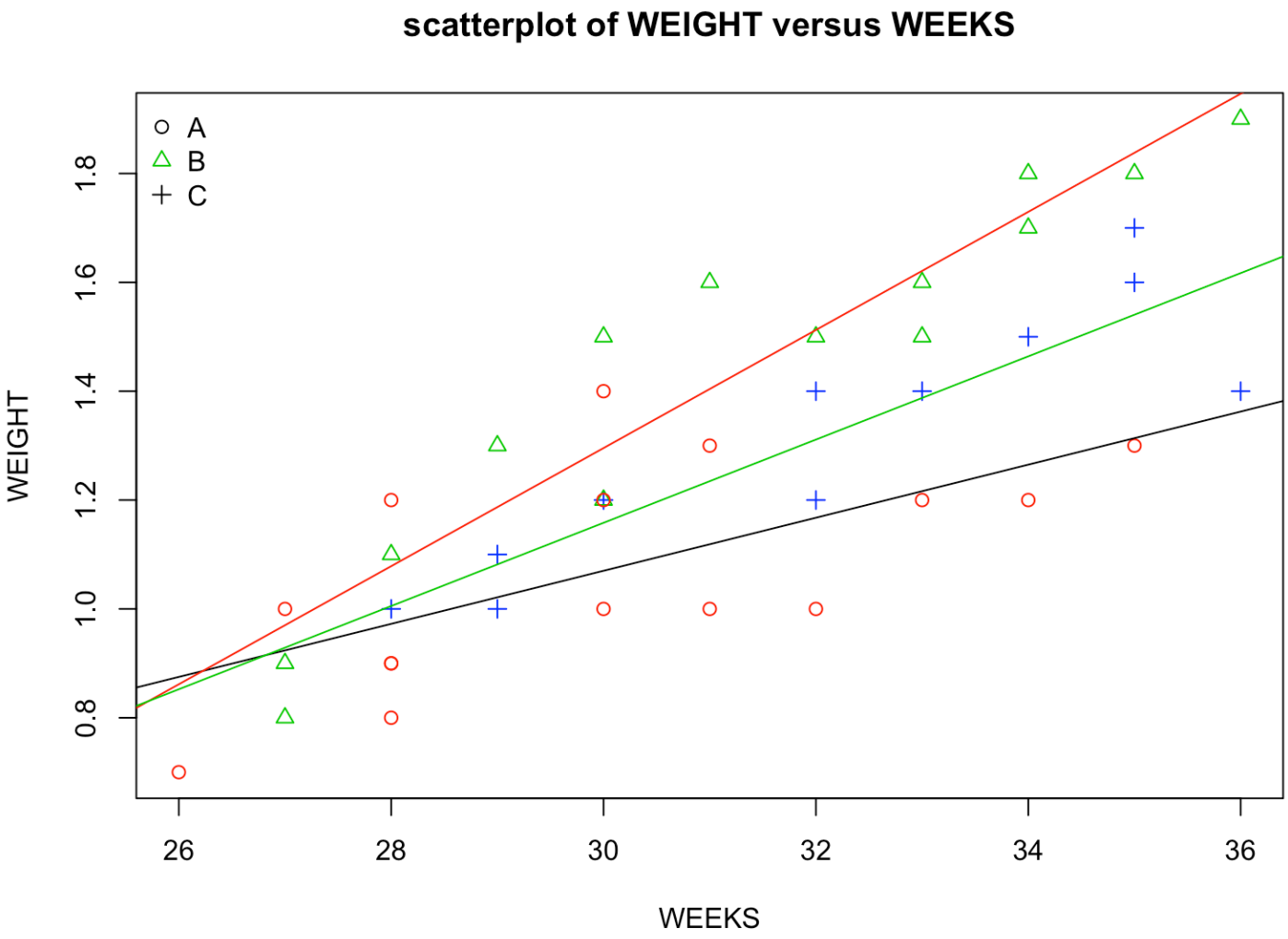
```
with(HOSPB,plot(WEEKS, WEIGHT, pch=HOSPn, col=(1 + HOSPn),
     xlab='WEEKS',ylab='WEIGHT', main="scatterplot of WEIGHT versus WEEKS"))

legend('topleft', c('A', 'B','C'), pch=1:3, col=c(1,3),
       bty="n")


ww<-coef(fit_gww)
ww
```

```
##             (Intercept)             HOSPB$WEEKS             HOSPB$HOSPB
##             -0.39277854              0.04876204             -1.56715837
##             HOSPB$HOSPC HOSPB$WEEKS:HOSPB$HOSPB HOSPB$WEEKS:HOSPB$HOSPC
##             -0.74327347              0.05975531              0.02771551
```

```
abline(ww[1], ww[2], col=1)
abline(ww[1] + ww[3], ww[2] + ww[5], col=2)
abline(ww[1] + ww[4], ww[2] + ww[6], col=3)
```



7. Compare the models in Questions 4 and 5. Which one model would you present to your boss and why? (3 marks)

```
anova(fit_WW,fit_gww)
```

```
## Analysis of Variance Table
##
## Model 1: HOSPB$WEIGHT ~ HOSPB$WEEKS + HOSPB$HOSP
## Model 2: HOSPB$WEIGHT ~ HOSPB$WEEKS + HOSPB$HOSP + HOSPB$WEEKS * HOSPB$HOSP
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     36 0.79406
## 2     34 0.60646  2    0.1876 5.2586 0.01024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment:
As the P-value is less then 0.05, we have strong evidence against H0, that we can use the model 2 which is WEIGHT~WEEKS+HOSP+WEEKS*HOSP is significant.