STA 303H1S / 1002 HS -Winter 2018 Assignment # 2 Solution Guide
What factors affect baby birth weight?
### *Shivon Sue-Chee*
### February 16, 2018

Grading Scheme Notes

- 10 marks: Presentation (Use of sentences. Use of plain English. Relevant plots and numbers in main section. Inclusion of the bulk of R codes and output in Appendix. Personalized R codes. All required plots are given a title that includes the last 4 digits of the student's UT number.)

- 90 marks: Parts 1-9.

  1. (15 marks):
     - (10) Boxplots with horizontal and vertical labels and ticks,
     - (5) Interpretations of box plots
  2. (10 marks):
     - (5) Use of `t.test`. Refer to R code in Appendix.
     - (5) Explanation of results supported by relevant analysis (test statistic, distribution of test statistic, p-value or confidence interval, interpretation of p-value or confidence interval).
  3. (15 marks)
     - (5) Variable used as a factor. Refer to R code in Appendix.
     - (5) Explanation of results supported by relevant analysis (test statistic, distribution of test statistic, p-value or confidence interval, interpretation of p-value or confidence interval).
     - (5) Pairwise comparisons by Tukey's or Bonferroni. Explanation of results.
  4. (15 marks)
     - (5) Variable used as a factor. Refer to R code in Appendix.
     - (5) Explanation of results supported by relevant analysis (test statistic, distribution of test statistic, p-value or confidence interval, interpretation of p-value or confidence interval).
     - (5) Pairwise comparisons by Tukey's or Bonferroni. Explanation of results.
  5. (10 marks)
     - (5) Normality checks and interpretation
     - (5) Check on homoscedasticity and interpretation. Discuss independence assumption.
  6. (10 marks) No two-way analysis/results. Interpretation of interaction or additivity from part 4.
  7. (5 marks) See solution
  8. (5 marks) See solution.
  9. (5 marks) Categorical (Factor) variables and not numeric ones.

1. (15 marks) The boxplots below show baby birth weight for babies whose mother smoked during pregnancy and those whose mothers did smoke during pregnancy, for babies classified by maturity and babies classified by both maternal smoking status and maturity.

There seems to be no particular outlier in the data. The red dots in the boxplots give the group means.

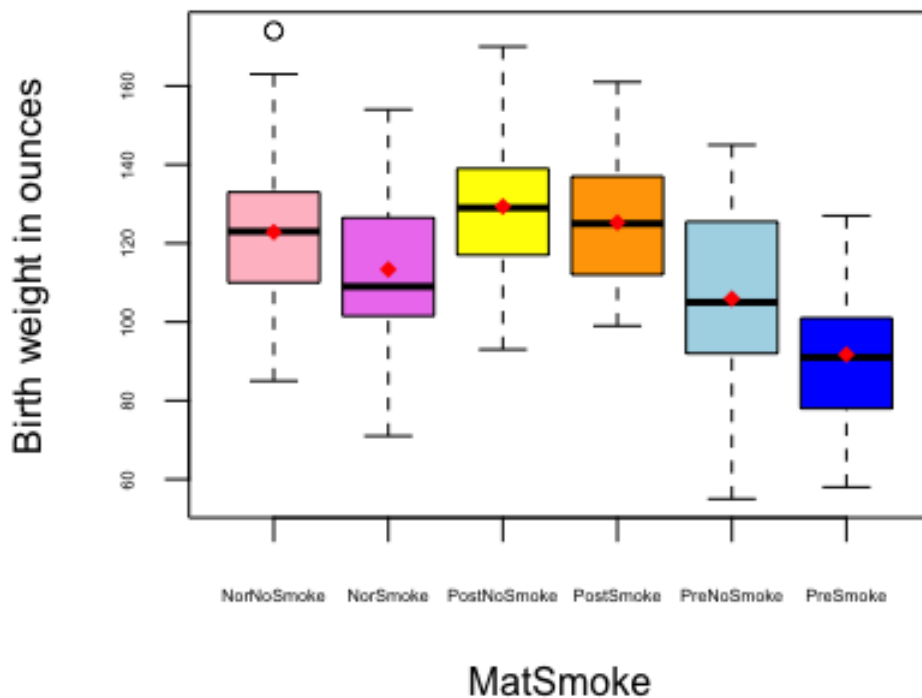**Birth weight by maternal smoking status (UT#)**
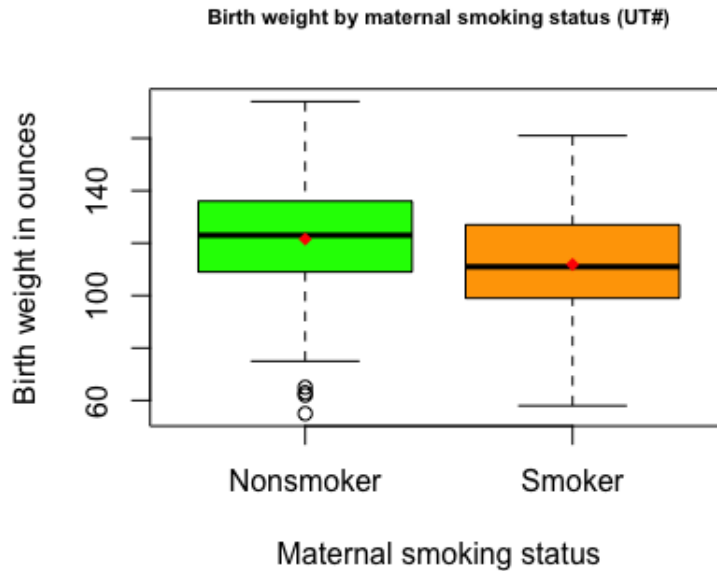
The distribution of birth weight among babies whose mother smoked and those whose mother did not smoke seems similar in means and variances, however, babies born to non-smokers were slightly heavier on average.

Birth weight by maturity (UT#)
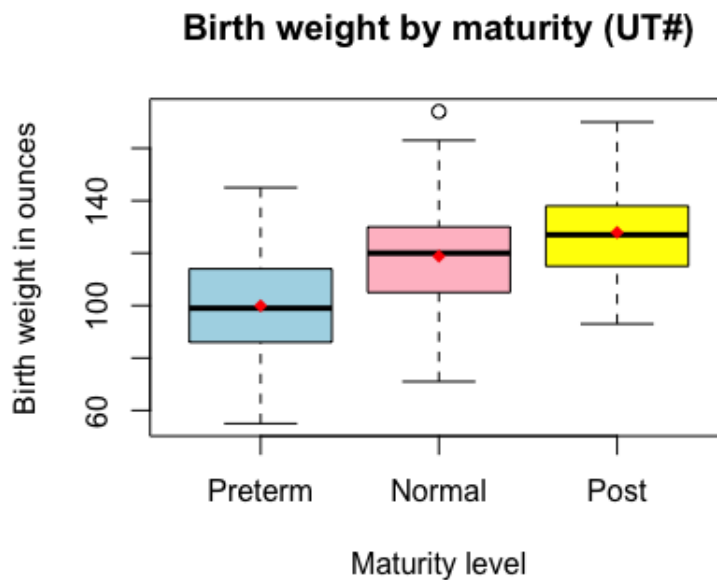
As maturity level increased from level 1 to 2 to 3, birth weight also increased.

Birth weight by combinations (UT#)

Birth weight in ounces

MatSmoke

When babies are classified by both maturity level and maternal smoking status, the most noticeable aspect is that across maturity levels, the difference between mean birth weight of babies whose mother smoked to mean birth weight of babies whose mother did not smoke decreased. For instance, for post mature babies, baby birth weight of smokers and non-smokers are similar while for premature babies, baby birth weight of smokers and non-smokers are different.

2. (10 marks) The tables below gives the mean birth weight for babies whose mothers were smokers and non-smokers, along with the test statistics and p-values for the Pooled and Satterthwaite approaches.

Smoking status	Mean birth weight	Two-sample t test	Test statistic (df)	p-value
Smoker	111.86	Pooled	4.69 (407.0)	< 0.0001
Non-smoker	121.55	Satterthwaite	4.64 (333.5)	< 0.0001

From the F-test for equality of variances (p=0.43), we have strong evidence to assume that the variances are equal and subsequently use the results from the pooled two-sample t-test.

On average, babies whose mothers smoked during pregnancy have lower birth weight than babies whose mothers were non-smokers during pregnancy; the mean birth weight for babies of smokers was 111.86 ounces while babies of non-smokers had a mean birth weight of 121.55 ounces. There is strong evidence ($p < 0.0001$) that the mean birth weight differs for babies whose mothers smoked to those whose mother did not smoke during pregnancy.

3

3. (15 marks) The table below gives the mean birth weight according to the three levels of maturity and the test statistic and p-value of the analysis of variance F-test of equal means.

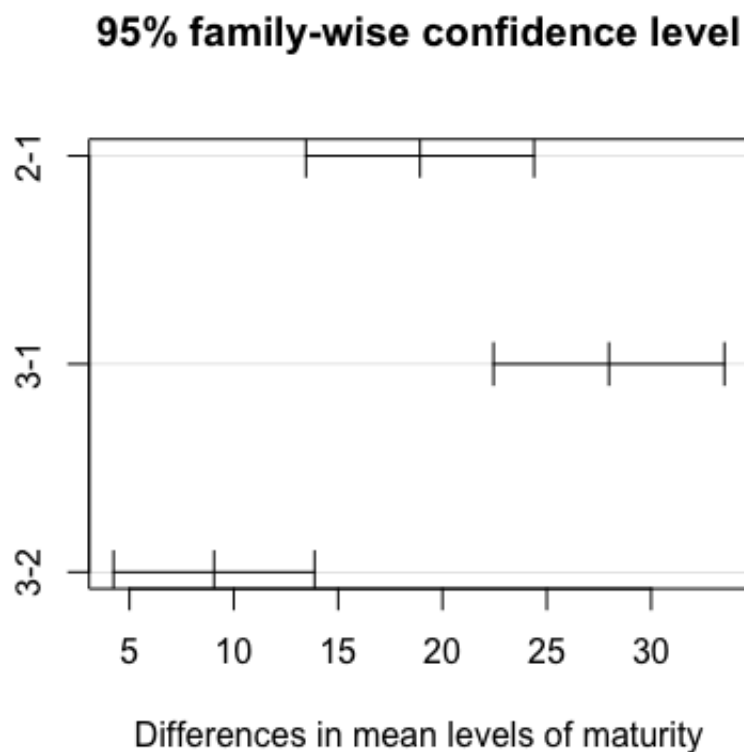| Mean birth weight (ounces)				
Premature	Normal	Postmature	Test statistic (df)	p-value
99.91	118.83	127.89	71.28 (2,406)	< 0.0001

There is strong evidence of differences in the mean birth weight among the three levels of maturity.

A post-hoc pairwise comparison of the means using Tukey's procedure, adjusted for unequal sample sizes and controlling the overall Type I error rate, indicates that the three levels are significantly different from each other. On average, postmature babies are heavier than normally matured babies ($p < 0.0001$) who in turn are heavier than premature babies ($p < 0.0001$).

Confidence intervals and p-values for all pairwise comparisons with Tukey's adjustment are given in the table below:

Levels	Mean difference	Confidence intervals	p-value
(2-1) Normal - Premature	18.93	(13.46, 24.39)	< 0.0001
(3-1) Postmature - Premature	27.98	(22.45, 33.51)	< 0.0001
(3-2) Postmature - Normal	9.06	(4.24, 13.87)	< 0.0001

The plot below shows the confidence intervals found by Tukey's approach. We notice that the confidence intervals do not include 0.

95% family-wise confidence level

Differences in mean levels of maturity

4. (15 marks) The table below gives the mean birth weight for babies classified by both maternal smoking status and maturity level, and the test statistic and p-value of the analysis of variance F-test of equal means..

Mean birth weight (ounces)							
Premature		Normal		Postmature		Test statistic (df)	p-value
Non-smoker	Smoker	Non-smoker	Smoker	Non-smoker	Smoker		
105.89	91.73	122.89	113.35	129.35	125.26	36.09 (5,403)	< 0.0001

There is strong evidence that not all classifications have the same mean birth weight ($p < 0.0001$).

A post-hoc pairwise comparison of the means using Tukeys procedure, adjusted for unequal sample sizes and controlling the overall Type I error rate, indicates that there is evidence that post mature babies born to a mother who smoked were not different in mean weight to post mature babies born to mothers who did not smoke ($p = 0.7422$). On the other hand, for the other levels of maturity, babies born to non-smokers were significantly different in mean weight to babies born to smokers (Normal, $p = 0.0099$; Pre, $p = 0.0014$).

The table below gives confidence intervals and p-values for all pairwise comparisons by the adjusted Tukeys approach.

| Levels	Mean difference	Confidence intervals	p-value
NorSmoke-NorNoSmoke	-9.49	(-17.50, -1.48)	0.0099
PostNoSmoke-NorNoSmoke	6.51	(-0.77 , 13.80)	0.1098
PostSmoke-NorNoSmoke	2.42	(-6.17, 11.01)	0.9661
PreNoSmoke-NorNoSmoke	-16.95	(-25.44, -8.46)	< 0.0001
PreSmoke-NorNoSmoke	-31.11	(-40.52, -21.70)	< 0.0001
PostNoSmoke-NorSmoke	16.00	(8.06, 23.94)	< 0.0001
PostSmoke-NorSmoke	11.91	(2.76, 21.06)	0.0030
PreNoSmoke-NorSmoke	-7.46	(-16.52, 1.60)	0.1736
PreSmoke-NorSmoke	-21.62	(-31.55, -11.70)	< 0.0001
PostSmoke-PostNoSmoke	-4.09	(-12.62, 4.43)	0.7422
PreNoSmoke-PostNoSmoke	-23.46	(-31.88, -15.03)	< 0.0001
PreSmoke-PostNoSmoke	-37.62	(-46.97, -28.27)	< 0.0001
PreNoSmoke-PostSmoke	-19.37	(-28.94, -9.80)	< 0.0001
PreSmoke-PostSmoke	-33.53	(-43.92, -23.13)	< 0.0001
PreSmoke-PreNoSmoke	-14.16	(-24.48, -3.85)	0.0014

The plot below shows the confidence intervals found by Tukey's approach. We notice that the effect of smoking on birth weight is weaker as maturity level increases. For instance, the $10th$ confidence interval from the top, at postmature maturity level, does include zero.

95% family-wise confidence level
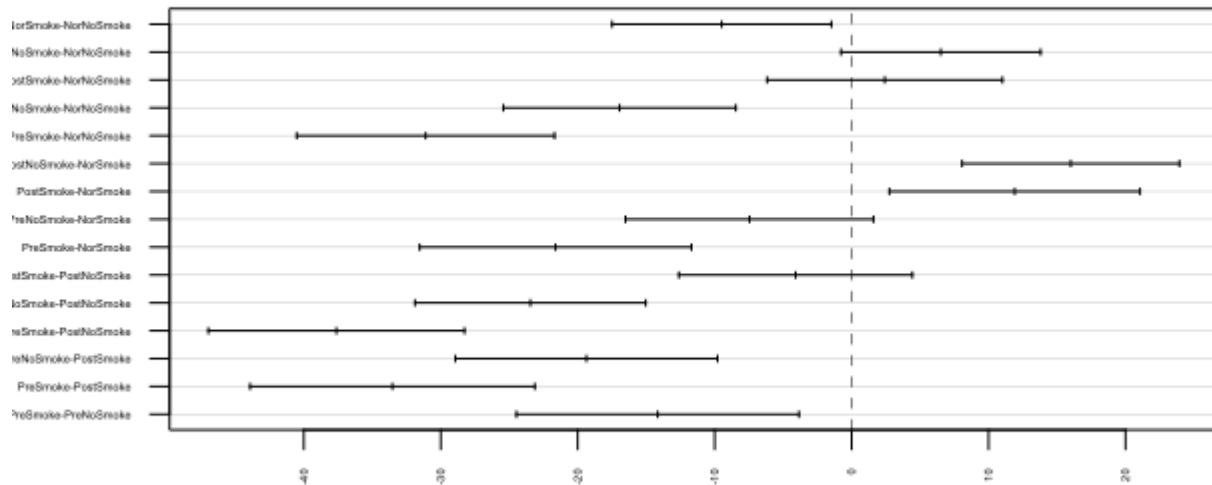
Differences in mean levels of MatSmoke

5. (10 marks) The diagnostic panel below give plots of the residuals of the analyses.
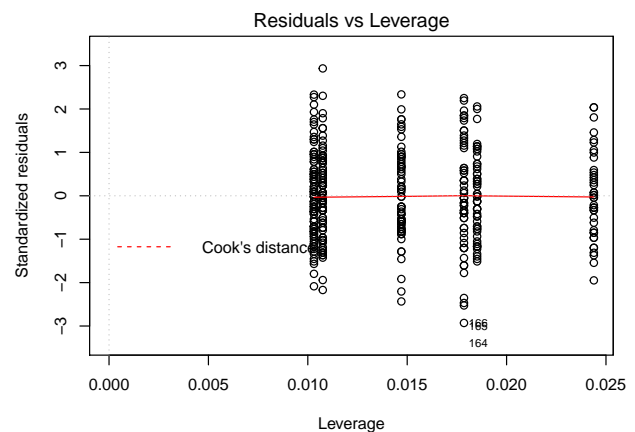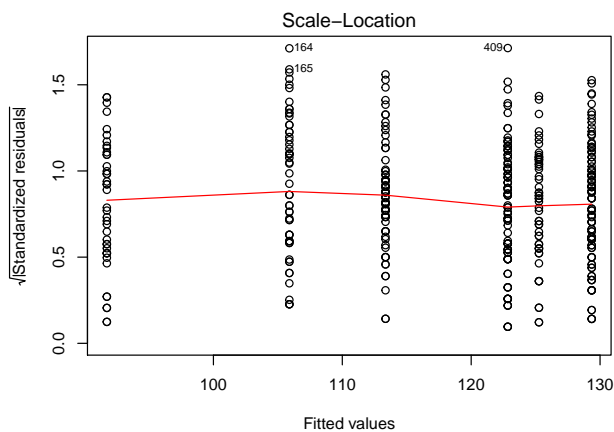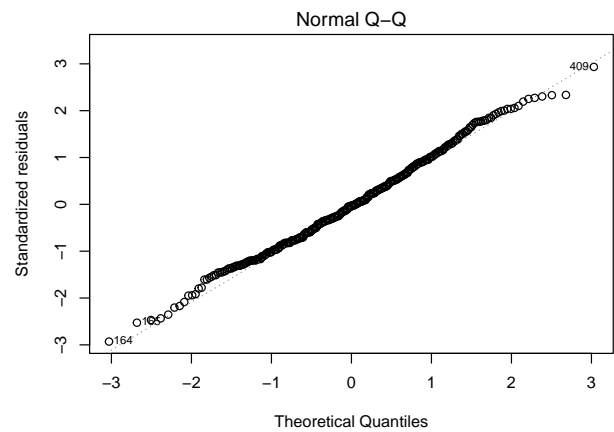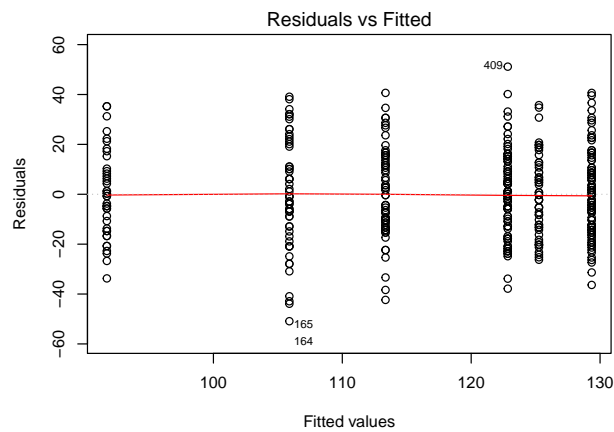
Based on the residual plots:

- Residuals versus Fitted values: There is no distinguishable pattern in the variation of the residuals, hence the assumption of constant variance seems reasonable. Moreover, the standard deviations for the observations from the 6 categories of babies range from 16.02 to 22.16, which is well within the 2:1 ratio suggested by the rule of thumb for equal variances.

- Root of Standardized absolute residuals versus Fitted values: There are no large outliers of concern. A few observations have these residuals greater than $\sqrt{2} = 1.41$ but a few in this range are expected in a data set this large.

- Normal quantile plot: Most of the residuals point fall on the straight line. Hence, the assumption of normality holds.

There were no extreme outliers in the data else sensitivity analysis would have been necessary.

We assume that babies include in this study were not twins or were not related in any way. It is likely that with a sample this large that some babies are related but no additional information was given to verify relatedness.

Since we have no serious concerns about violations of the model assumptions, we can trust the statistical tests carried out using all the data.

6. (10 marks)

   (a) The model fit in question 4 has 5 predictor variables (which are all indicator variables) since there are six categories of babies. The predictor variables in the two-way model would be 1 indicator variable for maternal smoking status (since it has two levels), 2 indicator variables for maturity level (since it has three levels) and then $2(= 1 \times 2)$ interaction terms, for a total of 5. Therefore, both models have the same number of predictor variables.

   (b) In question 4, our results show that for postmature babies, there is no significant difference in the mean birth weight between babies of smokers and those of non-smoker mothers ($p = 0.7422$). While for the other levels of maturity, there is a significant difference between babies of smoker and non-smoker mothers. Thus, there is an interaction between maturity level and maternal smoking status of the babies and an F-test for the interaction in the two-way model will likely be statistically significant.

7. (5 marks) For the combined samples by maturity level, we expect there to be lower variation for groups comprised of more babies. This violates the assumption that the error term in the model has the same variance for all observations. So we should be concerned about this.

   However, using the rule of thumb to compare variances, we have that the largest variance divided by the smallest variance, $491.2/256.6 < 4$, hence we have reason to believe that the assumption of equal variance across observations is not violated for the analysis of this data.

   Further, scientific studies are less powerful if the design is unbalanced, as is the case here.

8. (5 marks) Often in medical studies, it is more convenient to discuss the effect of numerical predictor variable, such as gestational age, in terms of suitable classifications such as premature, normal or postmature. However, from a statistical point of view, details are lost when a grouping variable is created from a quantitative variable. For instance, with our example, a two-way model cannot describe the change in baby birth weight for each unit change in gestational age. We are limited to studying the effect of gestational age in terms of maturity levels. Mathematically, the difference in models is described as:

(1) Two-way ANOVA (with maturity):

$$Weight_i|(Maturity, Smoke) = \beta_0 + \beta_1 I_{Pre,i} + \beta_2 I_{Nor,i} + \beta_3 I_{smoke,i} + \epsilon_i, \ i = 1, \ldots, 409$$

where, $I_{Pre,i} = 1$ if the ith baby was premature and is 0 otherwise, $I_{Nor,i} = 1$ if the ith baby attained normal maturity and is 0 otherwise, $I_{smoke,i} = 1$ if the ith baby's mother was a smoker and is 0 otherwise, and ϵ_i's are uncorrelated errors with constant variance, σ_1^2.

(2) ANCOVA (with gestational age):

$$Weight_i|(Gestation, Smoke) = \alpha_0 + \alpha_1 I_{Age_i} + \alpha_2 I_{smoke,i} + e_i, \ i = 1, \ldots, 409$$

where, Age, i is the gestational age in days of the ith baby, $I_{smoke,i} = 1$ if the ith baby's mother was a smoker and is 0 otherwise, and e_i's are uncorrelated errors with constant variance, σ_2^2.

The latter model is referred to as an ANCOVA- analysis of covariance model, which you may have studied in STA302 or equivalent.

9. (5 marks) Name two additional potential factors of baby birth weight and briefly describe their levels.

Answers will vary. Note that a factor is a categorical predictor variable defined by at least 2 levels. Some possible factors may include sex of baby, paternal smoking status, nutritional level of mother, education level attained by mother and employment status of mother.

Appendix of R code and Partial Output

```
> bbw<-read.csv("bbw.csv", header=T)
> head(bbw)
  bwt gestation smoke
1  68       223     1
2  92       224     1
3  91       225     1
4  69       232     1
5  65       232     1
6  71       234     1

> attach(bbw)
> dim(bbw)
[1] 409   3
```

```
> str(bbw)
'data.frame': 409 obs. of  3 variables:
 $ bwt      : int  68 92 91 69 65 71 77 77 75 96 ...
 $ gestation: int  223 224 225 232 232 234 238 238 239 241 ...
 $ smoke    : int  1 1 1 1 1 1 1 1 1 1 ...

> maturity=array(0,409)
> MatSmoke=array(0,409)

> for (i in 1:409)
+ {
+ if (gestation[i]<259)
+    {maturity[i]=1}
+ else if (gestation[i]>293)
+    {maturity[i]=3}
+ else {maturity[i]=2}
+ }

> for (i in 1:409)
+ {
+    if (maturity[i]==1 & smoke[i]==1)
+    {MatSmoke[i]="PreSmoke"}
+    else if (maturity[i]==1 & smoke[i]==0)
+    {MatSmoke[i]= .... [TRUNCATED]

> smoke=as.factor(smoke)
> maturity=as.factor(maturity)
> MatSmoke=as.factor(MatSmoke)

> ##Part 1
> #Boxplot by smoking status
> boxplot(bwt~smoke, main="Birth weight by maternal smoking status (UT#)",
+         ylab="Birth weight in ou ..." ... [TRUNCATED]
> (smeans <- tapply(bbw$bwt,bbw$smoke,mean))
       0        1
121.5488 111.8589
> points(smeans,col="red",pch=18)

> #Boxplot by maturity
> boxplot(bwt~maturity, main="Birth weight by maturity (UT#)",
+         ylab="Birth weight in ounces", xlab="Maturity level", .... [TRUNCATED]
> (mmeans <- tapply(bbw$bwt,maturity,mean))
        1         2         3
 99.90722 118.83230 127.88742
> points(mmeans,col="red",pch=18)

> #Boxplot by MatSmoke
```

```
> boxplot(bwt~MatSmoke, main="Birth weight by combinations (UT#)",
+         ylab="Birth weight in ounces",xlab="MatSmoke", ce .... [TRUNCATED]
> (msmeans <- tapply(bbw$bwt,MatSmoke,mean))
 NorNoSmoke    NorSmoke PostNoSmoke   PostSmoke  PreNoSmoke    PreSmoke
  122.83871   113.35294   129.35052   125.25926   105.89286    91.73171
> points(msmeans,col="red",pch=18)

> ##Part 2
> tapply(bbw$bwt,bbw$smoke,mean)
        0         1
121.5488 111.8589
> tapply(bbw$bwt,bbw$smoke,var)
        0         1
399.1221 446.1466
> tapply(bbw$bwt,bbw$smoke,length)
   0    1
246  163

> t.test(bwt~smoke) # Welch's

Welch Two Sample t-test

data:  bwt by smoke
t = 4.6409, df = 333.49, p-value = 4.994e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.582669 13.797101
sample estimates:
mean in group 0 mean in group 1
       121.5488        111.8589

> t.test(bwt~smoke, var.equal=TRUE) # Pooled

Two Sample t-test

data:  bwt by smoke
t = 4.6937, df = 407, p-value = 3.672e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.631563 13.748207
sample estimates:
mean in group 0 mean in group 1
       121.5488        111.8589

> var.test(bwt~smoke)

F test to compare two variances
```

```
data:  bwt by smoke
F = 0.8946, num df = 245, denom df = 162, p-value = 0.43
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6723307 1.1806435
sample estimates:
ratio of variances
         0.8945985


> ##Part 3
> tapply(bbw$bwt,maturity,mean)
         1         2         3
 99.90722 118.83230 127.88742

> tapply(bbw$bwt,maturity,var)
        1        2        3
452.0226 306.4030 268.4072

> tapply(bbw$bwt,maturity,length)
  1   2   3
 97 161 151

> fitm=lm(bwt~maturity)

> anova(fitm)
Analysis of Variance Table

Response: bwt
           Df Sum Sq Mean Sq F value    Pr(>F)
maturity    2  46586 23293.1  71.277 < 2.2e-16 ***
Residuals 406 132680   326.8
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> TukeyHSD(aov(bwt~maturity))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bwt ~ maturity)

> plot(TukeyHSD(aov(bwt~maturity)))

> ##Part 4
> tapply(bbw$bwt,MatSmoke,mean)
 NorNoSmoke    NorSmoke PostNoSmoke   PostSmoke  PreNoSmoke    PreSmoke
```

```
    122.83871   113.35294   129.35052   125.25926   105.89286    91.73171

> tapply(bbw$bwt,MatSmoke,var)
 NorNoSmoke    NorSmoke PostNoSmoke   PostSmoke  PreNoSmoke    PreSmoke
   270.6802    307.2766    271.6884    256.5730    491.1883    290.8012

> sqrt(tapply(bbw$bwt,MatSmoke,var))
 NorNoSmoke    NorSmoke PostNoSmoke   PostSmoke  PreNoSmoke    PreSmoke
   16.45236    17.52931    16.48297    16.01790    22.16277    17.05289

> tapply(bbw$bwt,MatSmoke,length)
 NorNoSmoke    NorSmoke PostNoSmoke   PostSmoke  PreNoSmoke    PreSmoke
         93          68          97          54          56          41

> fitms=lm(bwt~MatSmoke)

> anova(fitms)
Analysis of Variance Table

Response: bwt
           Df Sum Sq Mean Sq F value    Pr(>F)
MatSmoke    5  55448 11089.6  36.094 < 2.2e-16 ***
Residuals 403 123818   307.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> TukeyHSD(aov(bwt~MatSmoke))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bwt ~ MatSmoke)


> plot(TukeyHSD(aov(bwt~MatSmoke)), las=2, cex.axis=0.35)

> ##Part 5
> par(mfrow=c(2,2))
> plot(fitms)
```