6. Comparing Regression Lines

Dummy Variables in Regression & Comparing Two Regression Lines

Baby Birth-Weight Dataset:

Data on the birth weight (g) and estimated gestational age (weeks) of 12 male and 12 female babies were recorded. This data is taken from Dobson, A.J. (1990), "*An Introduction to Generalized Linear Models*", Chapman and Hall, p.17.

Female babies		**Male babies**	
Birth Weight (g) (Y)	**Gestational Age (weeks)(X)**	**Birth Weight (g) (Y)**	**Gestational Age (weeks)(X)**
3317	40	2968	40
2729	36	2795	38
.	.	.	.

We wish to use Gestational Age to predict Birth Weight and compare the separate regression lines for female and male babies.

We could fit two separate regression lines, one for female babies and one for male babies:

$$Y = \beta_0^{(f)} + \beta_1^{(f)} X + \varepsilon \,,$$

$$Y = \beta_0^{(m)} + \beta_1^{(m)} X + \varepsilon \,.$$

We would like to test the following hypotheses:

$H_0 : \beta_1^{(f)} = \beta_1^{(m)}$ (the two regression lines have the same slope)

$H_0 : \beta_0^{(f)} = \beta_0^{(m)}$ (the two regression lines have the same intercept)

$H_0 : \beta_0^{(f)} = \beta_0^{(m)}, \beta_1^{(f)} = \beta_1^{(m)}$ (the two regression lines coincide)

This is also easily achieved as follows:

Fit **one** regression model to the **combined** set of data using a **dummy variable** Z :

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

where

$$Z = \begin{cases} 0, & \text{if female} \\ 1, & \text{if male} \end{cases} .$$

XZ is known as an **interaction** term.

The combined data looks like this:

	Birth Weight (Y)	**Age** (X)	Z	XZ
Female	3317	40	0	0
.	2729	36	0	0
.	.	.	.	.
Male	2968	40	1	40
.	2795	38	1	38
.	.	.	.	.

To understand the interpretation of the coefficients in this regression model, consider the forms of the model for female babies and for male babies:

$$Z = 0 \text{ (female babies): } Y = \beta_0 + \beta_1 X + \varepsilon$$

Thus $\beta_0 = \beta_0^{(f)}$ and $\beta_1 = \beta_1^{(f)}$, are the intercept and slope, respectively, of the (true) regression line for **female** babies.

$$Z = 1 \text{ (male babies): } Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \varepsilon \, .$$

Thus $\beta_0 + \beta_2 = \beta_0^{(m)}$ and $\beta_1 + \beta_3 = \beta_1^{(m)}$ are the intercept and slope, respectively, of the (true) regression line for **male** babies.

Video 6.1

Note:

This interpretation of the coefficients $\beta_0, \beta_1, \beta_2$ and β_3 in the combined model is **dependent on the choice of coding for the dummy variable** Z. If we use an **alternative coding scheme**, such as

$$Z = \begin{cases} -1, & \text{if female} \\ +1, & \text{if male} \end{cases},$$

the interpretation of the coefficients $\beta_0, \beta_1, \beta_2$ and β_3 is **different** from that given above for the coding scheme

$$Z = \begin{cases} 0, & \text{if female} \\ 1, & \text{if male} \end{cases}$$

Because of the interpretation of the coefficients $\beta_0, \beta_1, \beta_2$ and β_3 in the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

with the recommended coding scheme

$$Z = \begin{cases} 0, & \text{if female} \\ 1, & \text{if male} \end{cases}$$

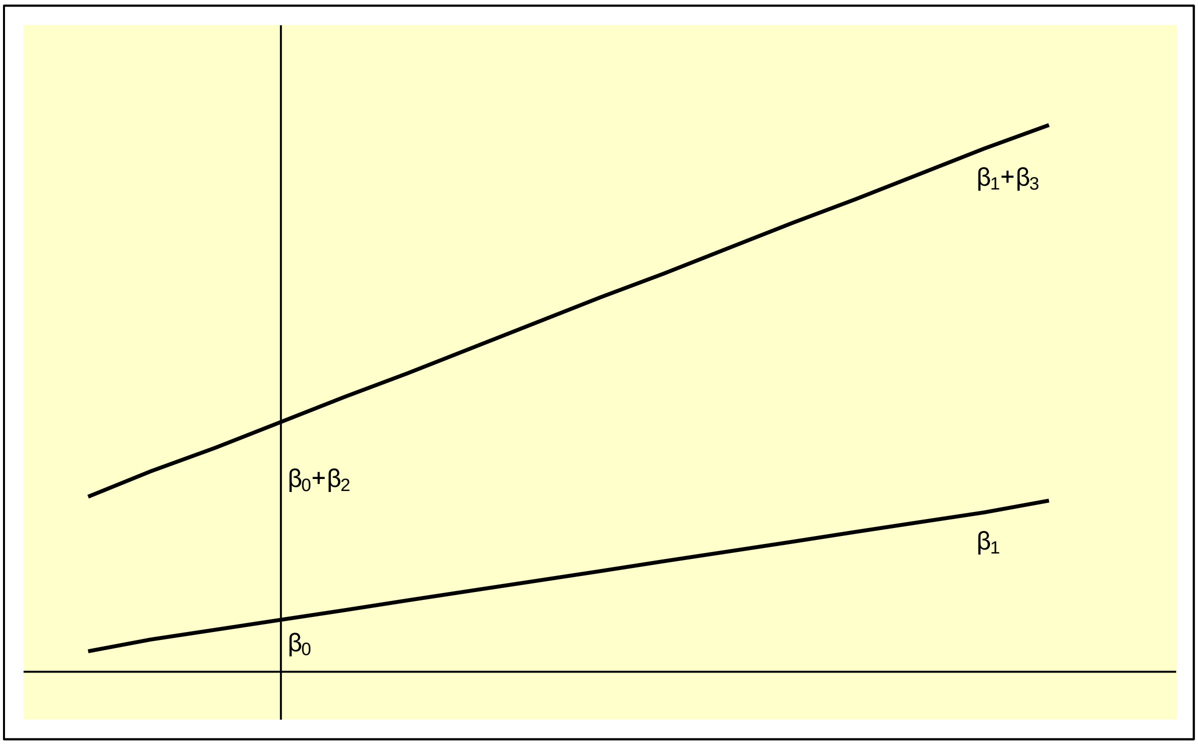
the hypotheses that we wish to test to compare the separate regression lines are easily expressed in terms of the coefficients β_2 and β_3 :

$H_0 : \beta_3 = 0$ (the two regression lines have the same slope)

$H_0 : \beta_2 = 0$ (the two regression lines have the same intercept)

$H_0 : \beta_2 = \beta_3 = 0$ (the two regression lines coincide)

These hypotheses are easily tested in this model.



Comparing more than two regression lines

The above technique of using dummy variables can be extended to the comparison of more than two regression lines.

For **two groups**, we need **one** dummy variable Z . The following coding scheme is recommended for Z :

	Z
Group 1	0
Group 2	1

and the model is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

with **one** interaction term XZ .

For **three groups**, we need **two** dummy variables Z_1 and Z_2 . The following coding scheme is recommended for Z_1 and Z_2 :

	Z_1	Z_2
Group 1	0	0
Group 2	1	0
Group 3	0	1

and the model is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 XZ_1 + \beta_5 XZ_2 + \varepsilon$$

with **two** interaction terms XZ_1 and XZ_2 .

To understand the interpretation of the coefficients in this regression model, consider the forms of the model for each group:

Group 1: $Z_1 = 0$ and $Z_2 = 0$: $Y = \beta_0 + \beta_1 X + \varepsilon$

Thus β_0 and β_1, are the intercept and slope, respectively, of the (true) regression line for **Group 1**.

Group 2: $Z_1 = 1$ and $Z_2 = 0$: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X + \varepsilon$.

Thus $\beta_0 + \beta_2$ and $\beta_1 + \beta_4$ are the intercept and slope, respectively, of the (true) regression line for **Group 2**.

Group 3: $Z_1 = 0$ and $Z_2 = 1$: $Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X + \varepsilon$.

Thus $\beta_0 + \beta_3$ and $\beta_1 + \beta_5$ are the intercept and slope, respectively, of the (true) regression line for **Group 3**.

Because of the above interpretation of the coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and β_5 in the model $Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 X Z_1 + \beta_5 X Z_2 + \varepsilon$

with the recommended coding scheme, the hypotheses that we wish to test to compare the separate regression lines are easily expressed in terms of the coefficients $\beta_2, \beta_3, \beta_4$ and β_5:

$H_0 : \beta_5 = 0$ (the regression lines for Groups 1 and 3 have same slope)
$H_0 : \beta_4 = 0$ (the regression lines for Groups 1 and 2 have same slope)

$H_0 : \beta_4 = \beta_5 = 0$ (the regression lines for Groups 1, 2 and 3 have same slope)

$H_0 : \beta_3 = 0$ (the regression lines for Groups 1 and 3 have same intercept)
$H_0 : \beta_2 = 0$ (the regression lines for Groups 1 and 2 have same intercept)

$H_0 : \beta_2 = \beta_3 = 0$ (the regression lines for Groups 1, 2, and 3 have same intercept)

$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (the regression lines for Groups 1, 2 and 3 coincide)

There are four situations to consider when comparing regression lines:

General case:	intercepts different & slopes different.
Parallel regressions:	intercepts different & slopes the same.
Concurrent regressions:	intercepts the same & slopes different.
Coincident regressions:	intercepts the same & slopes the same.

Video 6.2

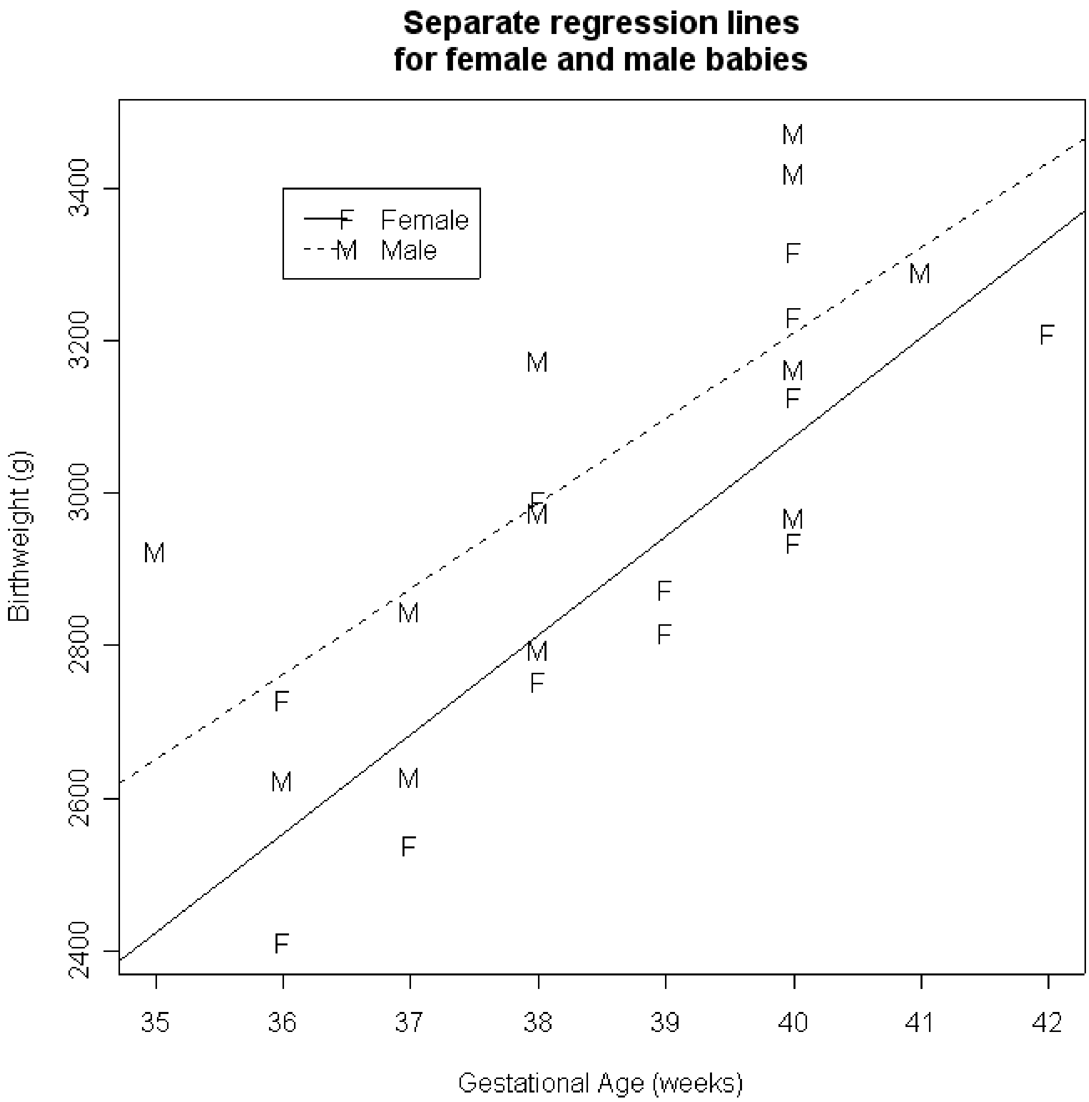
Analysis of the baby birth-weight data:

```
> babies.df <- read.table("P:\\ST2053\\babies.txt",
+ header=T)
> babies.df
   Birthwt Age Gender
1     2968  40   Male
2     2795  38   Male
3     3163  40   Male
4     2925  35   Male
5     2625  36   Male
6     2847  37   Male
7     3292  41   Male
8     3473  40   Male
9     2628  37   Male
10    3176  38   Male
11    3421  40   Male
12    2975  38   Male
13    3317  40 Female
14    2729  36 Female
15    2935  40 Female
16    2754  38 Female
17    3210  42 Female
18    2817  39 Female
19    3126  40 Female
20    2539  37 Female
21    2412  36 Female
22    2991  38 Female
23    2875  39 Female
24    3231  40 Female
```

We first draw a scatter plot for this data, using the symbols F and M to label the points for female and male babies, respectively, as shown below.

```
> attach(babies.df)
> plot(Age,Birthwt,type="n",
+ main="Separate regression lines
+ for female and male babies",
+ xlab="Gestational Age (weeks)",
+ ylab="Birthweight (g)")
> text(Age,Birthwt,c("F","M")[Gender])

> legend (36,3400,pch="FM",merge=FALSE,
+ lty=c(1,2),legend=c("Female","Male"))
```



Notes on above R code:
In the `plot` function above, the argument `type="n"` suppresses the plotting of points, so that only the title, axes and axis labels are plotted.
The `text` function plots the characters F and M in the scatter plot, depending on the `Gender` of the corresponding observation.

In the `legend` function, the first two arguments (36 and 3400) are the (*x*, *y*) coordinates of the top left corner of the box for the **legend** which describes what each of the characters F and M represents.
The `pch` argument to the `legend` function specifies which plotting characters to display. Note that this argument consists of a single character string `"FM"` containing the plotting characters to be used in the legend, not a vector of single characters.
The `legend` argument to the `legend` function is a vector of character strings `c("Female","Male")` to be associated with the plotting characters F and M.
The `lty` argument to the `legend` function specifies the types of lines (solid for `lty=1` and dotted for `lty=2`) corresponding to the categories Female and Male.
By default, the legend is contained in a box; the drawing of the box can be suppressed by the argument `bty = "n"`.

By using the `subset` argument to the `lm` function, we can fit two models to model `Birthwt` by `Age` for female and male babies separately as follows:

```
> female.babies1.lm <- lm(Birthwt~Age,
+ data = babies.df, subset= Gender=="Female")
>
> summary(female.babies1.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2141.67    1016.05  -2.108 0.061265 .
Age           130.40      26.19   4.978 0.000555 ***
```

```
> male.babies1.lm <- lm(Birthwt~Age,
+ data = babies.df, subset= Gender=="Male")
>
> summary(male.babies1.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1268.67    1239.97  -1.023  0.33035
Age           111.98      32.31   3.466  0.00606 **
```

From the above, we see that for female babies the regression line of `Birthwt` on `Age` has intercept -2141.67 and slope 130.40; the corresponding regression line for male babies has intercept -1268.67 and slope 111.98. These regression lines are plotted in above using the `abline` function as follows:

```
> abline(female.babies1.lm,lty=1)
> abline(male.babies1.lm,lty=2)
```

It is clear that the fitted regression lines have very similar slopes, but the regression line for male babies is above that for female babies.

Video 6.3

Fitting parallel regression lines

Rather than fitting separate models for female and male babies, we can fit a **series of models** to the combined group of babies. In these models, we can fit:

- the same regression line for female and male babies
- parallel regression lines for female and male babies
- separate regression lines for female and male babies

From these models, we can test the significance of the differences between the regression lines and so decide which model best represents the data.

Firstly, we use the model `babies1.lm` to fit the same regression line for female and male babies:

```
> babies1.lm <- lm(Birthwt~Age, data = babies.df)
> coef(babies1.lm)
```

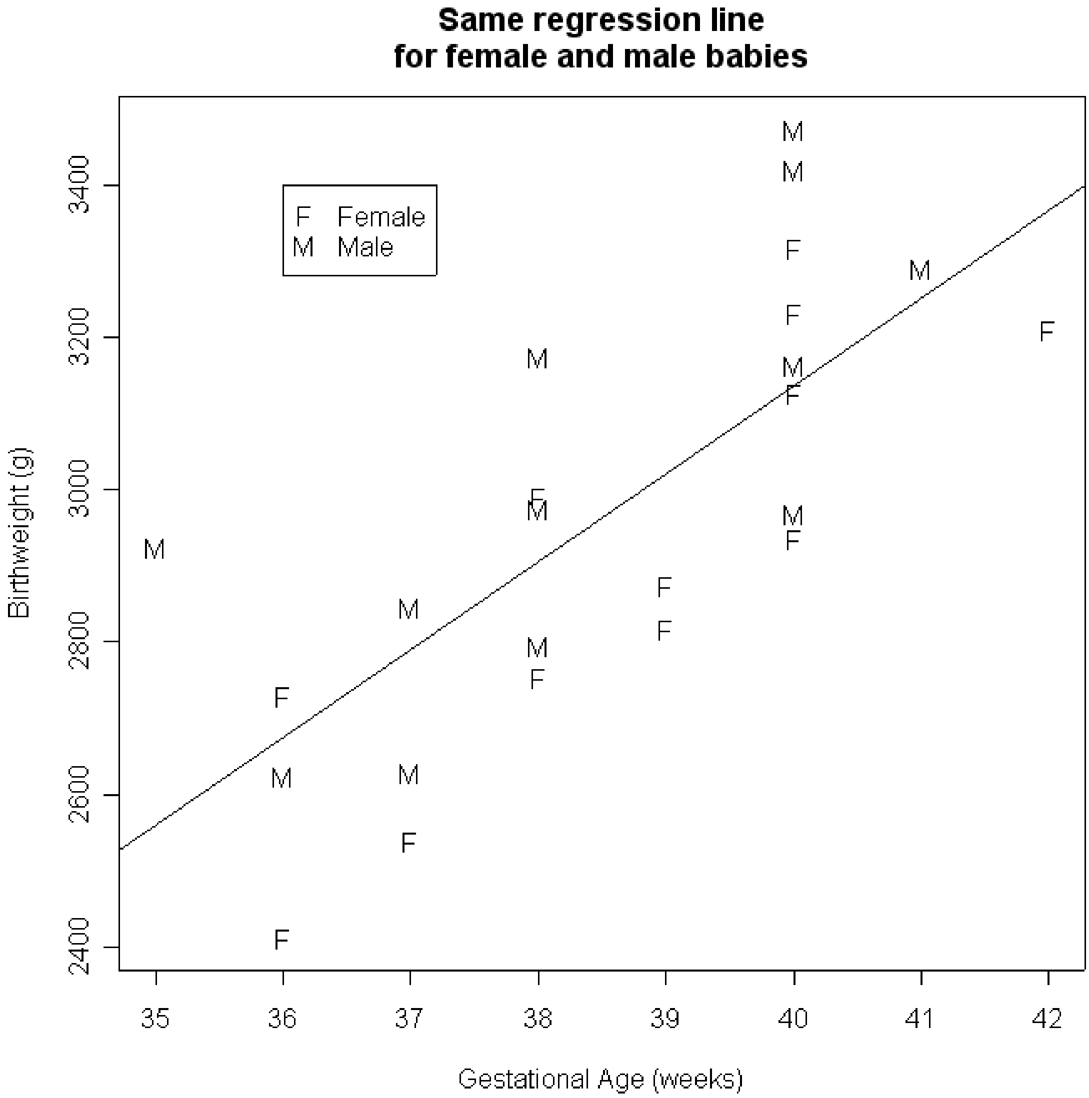
```
(Intercept)         Age
 -1484.9846    115.5283
```

From the above, we see that the fitted regression line of `Birthwt` on `Age` for all babies has intercept −1484.9846 and slope 115.5283.

We draw the scatter plot with this regression line superimposed, as shown below.

```
> plot(Age,Birthwt,type="n",
+ main="Same regression line
+ for female and male babies",
+ xlab="Gestational Age (weeks)",
+ ylab="Birthweight (g)")
> text(Age,Birthwt, c("F","M")[Gender])

> legend(36,3400,pch="FM",
+ legend=c("Female","Male"))
> abline(babies1.lm)
```



Same regression line for female and male babies

In the first scatter-plot, we saw that the separate regression lines for female and male babies had very similar slopes. This leads us to fit a model in which we assume that the regression lines for female and male babies have the same slope, but different intercepts.
In the data frame `babies.df`, the variable Gender is a factor with two levels:

```
> class(Gender)
[1] "factor"
> levels(Gender)
[1] "Female" "Male"
```

To model `Birthwt` by `Age` and `Gender`, we specify a model with the following formula:

```
Birthwt ~ Age + Gender
```

In the usual notation, the model fitted by this formula is:

$$\text{Birthwt} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Gender} + \varepsilon$$

By default, R will code the dummy variable `Gender` in the above model as follows:

$$\text{Gender} = \begin{cases} -1 \text{ for female} \\ \ \ 1 \text{ for male} \end{cases}$$

However, to make it easier to interpret the coefficients of the fitted model, the recommended coding for a dummy variable like `Gender` is the following:

$$\text{Gender} = \begin{cases} 0 \text{ for female} \\ 1 \text{ for male} \end{cases}$$

Use the `options` function as follows to force R to use the preferred 0-1 coding.

```
> options(contrasts=c(factor="contr.treatment",
+ ordered="contr.poly"))
```

Fit the model `babies2.lm`:

```
> babies2.lm <- lm(Birthwt~Age + Gender, data = babies.df)
```

Use the `model.matrix` function to view the model matrix of the fitted model:

```
> model.matrix(babies2.lm)
   (Intercept) Age GenderMale
1            1  40          1
2            1  38          1
3            1  40          1
4            1  35          1
5            1  36          1
6            1  37          1
7            1  41          1
8            1  40          1
9            1  37          1
10           1  38          1
11           1  40          1
12           1  38          1
13           1  40          0
14           1  36          0
15           1  40          0
16           1  38          0
17           1  42          0
18           1  39          0
19           1  40          0
20           1  37          0
21           1  36          0
22           1  38          0
23           1  39          0
24           1  40          0
```

View the coefficients of the fitted model as follows:

```
> coef(babies2.lm)
(Intercept)          Age  GenderMale
 -1773.3218    120.8943    163.0393
```

We can deduce the estimates of the intercepts for female and male:

$$\hat{\beta}_0 = -1773.3218 \text{ and } \hat{\beta}_0 + \hat{\beta}_2 = -1773.3218 + 163.0393 = -1610.283,$$

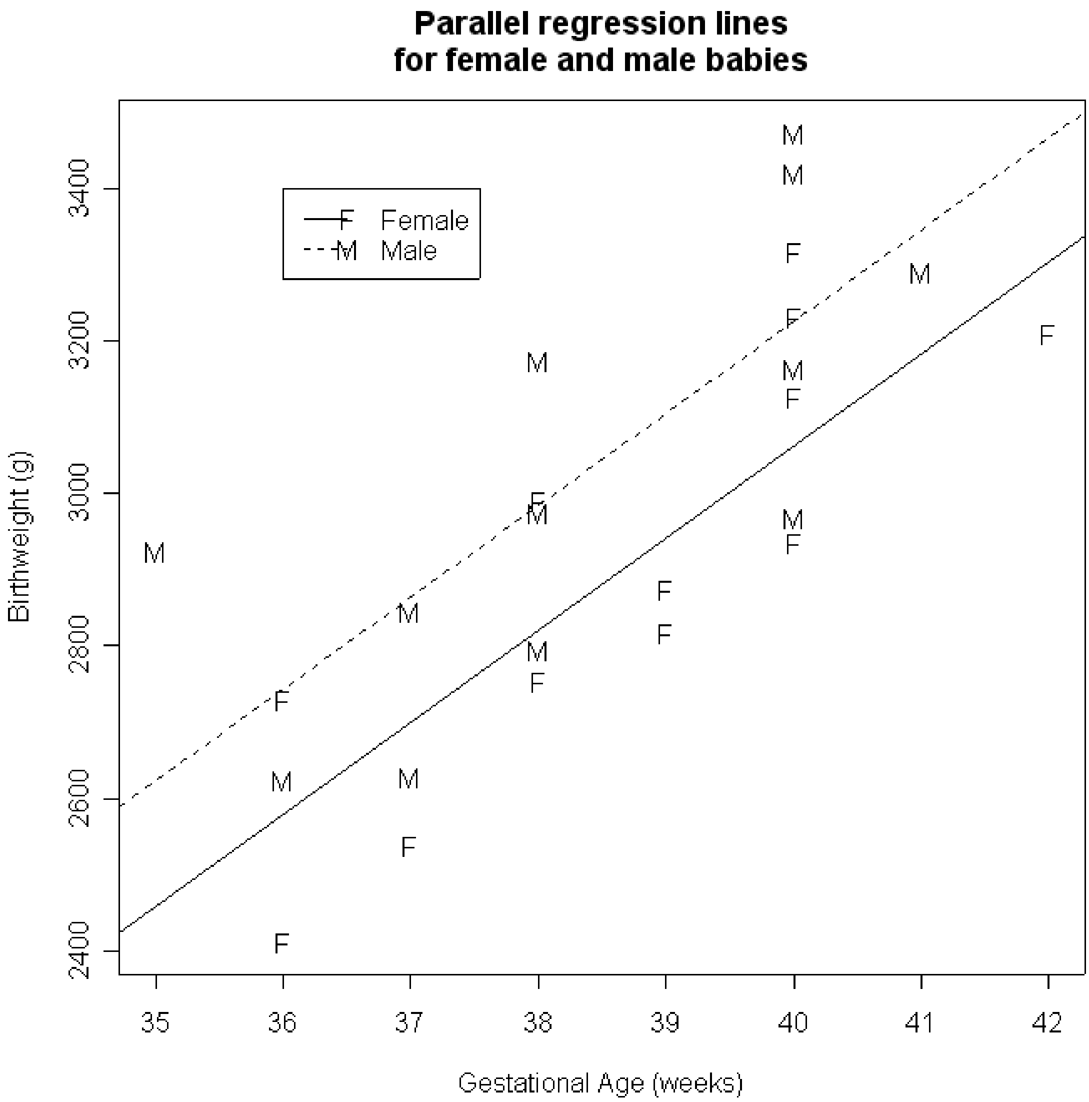
respectively, and the estimate of the common slope is

$$\hat{\beta}_1 = 120.8943$$

The parallel regression lines are plotted on the scatter plot, as shown below.

```
> plot(Age,Birthwt,type="n",
+ main="Parallel regression lines
+ for female and male babies",
+ xlab="Gestational Age (weeks)",
+ ylab="Birthweight (g)")
> text(Age,Birthwt, c("F","M")[Gender])

> legend (36,3400, c("Female","Male"),
+ pch="FM", merge=FALSE, lty=c(1,2))
> abline(-1773.3218,120.8943,lty=1)
> abline(-1773.3218+163.0393,120.8943,lty=2)
```



Parallel regression lines for female and male babies

Video 6.4

Fitting separate regression lines

In the previous section, we fitted a model with the formula `Birthwt ~ Age + Gender` to obtain the coefficients of regression lines with the same slope but different intercepts. By adding the interaction term `Age:Gender` to the model formula, we can obtain the coefficients of separate regression lines for female and male babies.

We fit a model `babies3.lm` with the formula

```
Birthwt ~ Age + Gender + Age:Gender
```

This model is fitted as follows:

```
> babies3.lm <- lm(Birthwt~Age + Gender + Age:Gender, data
= babies.df)

> model.matrix(babies3.lm)
   (Intercept) Age GenderMale Age:GenderMale
1            1  40          1             40
2            1  38          1             38
3            1  40          1             40
4            1  35          1             35
5            1  36          1             36
6            1  37          1             37
7            1  41          1             41
8            1  40          1             40
9            1  37          1             37
10           1  38          1             38
11           1  40          1             40
12           1  38          1             38
13           1  40          0              0
14           1  36          0              0
15           1  40          0              0
16           1  38          0              0
17           1  42          0              0
18           1  39          0              0
19           1  40          0              0
20           1  37          0              0
21           1  36          0              0
22           1  38          0              0
23           1  39          0              0
24           1  40          0              0
```

In the usual notation, the model we have fitted is:

$$\text{Birthwt} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Gender} + \beta_3(\text{Age})(\text{Gender}) + \varepsilon$$

Equivalently, the model fitted by this formula is

$\texttt{Birthwt} = \beta_0 + \beta_1 \texttt{Age} + \varepsilon$ if `Gender` is `Female` and

$\texttt{Birthwt} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\texttt{Age} + \varepsilon$ if `Gender` is `Male`.

We see that the column for the interaction term `Age:Gender` is the product of the column for `Age` and the column for `Gender`; thus the interaction effect of a numeric variable (`Age`) and a factor (`Gender`) is **multiplicative**. The coefficients of the fitted model are:

```
> coef(babies3.lm)
 (Intercept)          Age   GenderMale Age:GenderMale
-2141.66667    130.40000    872.99425      -18.41724
```

For female babies, the estimates of the intercept and slope are

$$\hat{\beta}_0 = -2141.66667 \quad \text{and} \quad \hat{\beta}_1 = 130.4,$$

respectively, while for male babies the corresponding estimates are

$$\hat{\beta}_0 + \hat{\beta}_2 = -2141.66667 + 872.99425 = -1268.672$$

and

$$\hat{\beta}_1 + \hat{\beta}_3 = 130.4 - 18.41724 = 111.9828$$

These regression coefficients agree with those obtained by fitting separate models for female and male babies earlier.

With the coefficients extracted from the above model, we can use the `abline` function as follows to plot separate regression lines for female and male babies as shown in the first scatter-plot.

```
> abline(-2141.66667,130.4,lty=1)
> abline(-2141.66667+872.99425,130.4-18.41724,lty=2)
```

Video 6.5

Comparing models

Recall the model formulae for the models `babies1.lm`, `babies2.lm` and `babies3.lm`:

```
> formula(babies1.lm)
Birthwt ~ Age
> formula(babies2.lm)
Birthwt ~ Age + Gender
> formula(babies3.lm)
Birthwt ~ Age + Gender + Age:Gender
```

The model for `babies3.lm` is

$$\texttt{Birthwt} = \beta_0 + \beta_1\,\texttt{Age} + \beta_2\,\texttt{Gender} + \beta_3\,(\texttt{Age})(\texttt{Gender}) + \varepsilon$$

In comparing these models, there are a number of hypotheses that we wish to test:

- H_0 : the (true) regression lines are parallel ($\beta_3 = 0$)
- H_0 : the (true) regression lines coincide ($\beta_2 = \beta_3 = 0$)

We can test the hypothesis H_0: $\beta_3 = 0$ by testing the significance of the interaction term `Age:Gender` in the model `babies3.lm`:

```
> summary(babies3.lm)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2141.67    1163.60  -1.841 0.080574 .
Age              130.40      30.00   4.347 0.000313 ***
GenderMale       872.99    1611.33   0.542 0.593952
Age:GenderMale   -18.42      41.76  -0.441 0.663893

> anova(babies3.lm)
Analysis of Variance Table

Response: Birthwt
           Df  Sum Sq Mean Sq F value    Pr(>F)
Age         1 1013799 1013799 31.0779 1.862e-05 ***
Gender      1  157304  157304  4.8221   0.04006 *
Age:Gender  1    6346    6346  0.1945   0.66389
Residuals  20  652425   32621
```

The output for the `summary` and `anova` functions indicates that the coefficient of the `Age:Gender` term in the `babies3.lm` model is not significant, so the true regression lines can be assumed to be parallel.

We test the hypothesis H_0: $\beta_2 = \beta_3 = 0$ by using the `anova` function to compare the **nested** (the terms of a smaller model are contained in a larger model) models `babies1.lm` and `babies3.lm`:

```
> anova(babies1.lm, babies3.lm)
Analysis of Variance Table

Model 1: Birthwt ~ Age
Model 2: Birthwt ~ Age + Gender + Age:Gender
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     22 816074
2     20 652425  2    163650 2.5083 0.1067
```

The Extra Sum of Squares due to the `Gender` and `Age:Gender` terms is not significant, so the true regression lines can be assumed to coincide.

The techniques described above can also be used to compare three or more regression lines.

Video 6.6

Twins Dataset

Data were collected on the IQ scores of 27 pairs of identical twins, one raised in a foster home and the other raised by natural parents. The cases are divided into three groups according to social class of the natural parents.

```
   Foster Natural Social
1      82      82 Class1
2      80      90 Class1
3      88      91 Class1
4     108     115 Class1
5     116     115 Class1
6     117     129 Class1
7     132     131 Class1
8      71      78 Class2
9      75      79 Class2
10     93      82 Class2
11     95      97 Class2
12     88     100 Class2
13    111     107 Class2
14     63      68 Class3
15     77      73 Class3
16     86      81 Class3
17     83      85 Class3
18     93      87 Class3
19     97      87 Class3
20     87      93 Class3
21     94      94 Class3
22     96      95 Class3
23    112      97 Class3
24    113      97 Class3
25    106     103 Class3
26    107     106 Class3
27     98     111 Class3
```

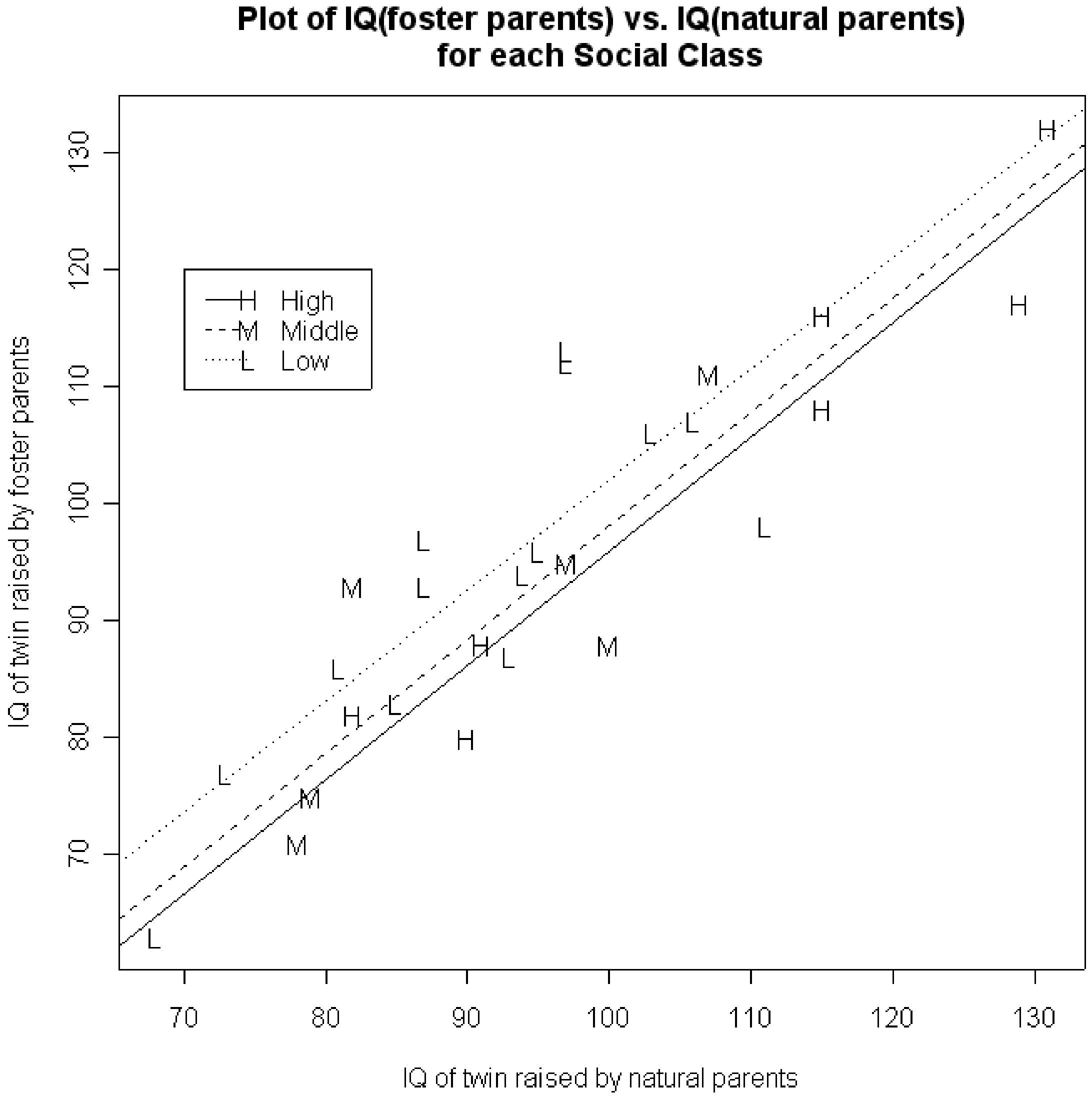
```
> # R code and output for analysis of the Twins dataset
>
> load(".rdata")

> twins.df <-
+read.table("P:\\ST2053\\twins.txt",
+header=T)
> # attach twins.df
> attach(twins.df)

> # verify that Social is a Factor with 3 levels
> class(Social)
[1] "factor"
> levels(Social)
[1] "Class1" "Class2" "Class3"


> # Plot of IQ(foster parents) vs. IQ(natural parents)
> # for each Social Class
> plot(Natural,Foster,type="n",
+ main="Plot of IQ(foster parents) vs. IQ(natural parents)
+ for each Social Class",
+ xlab="IQ of twin raised by natural parents",
+ ylab="IQ of twin raised by foster parents")
> text(Natural,Foster,c("H","M","L")[Social])

> legend(70,120,pch="HML",merge=FALSE,
+ lty=c(1,2,3),legend=c("High","Middle","Low"))
```


Plot of IQ(foster parents) vs. IQ(natural parents)
for each Social Class
IQ of twin raised by foster parents
IQ of twin raised by natural parents
High
Middle
Low
70
80
90
100
110
120
130

```
> # fit separate regression lines for each social class
>
> class1.twins1.lm <- lm(Foster ~ Natural,
+ data=twins.df,subset= Social=="Class1")
> coef(class1.twins1.lm)
(Intercept)     Natural
 -1.8720437   0.9775622
> # intercept for Class1 = -1.8720437
> # slope for Class 1 = 0.9775622
>
> class2.twins1.lm <- lm(Foster ~ Natural,
+ data=twins.df,subset= Social=="Class2")
> coef(class2.twins1.lm)
(Intercept)     Natural
  0.8160244   0.9725669
> # intercept for Class2 = 0.8160244
> # slope for Class2 = 0.9725669
>
> class3.twins1.lm <- lm(Foster ~ Natural,
+ data=twins.df,subset= Social=="Class3")
> coef(class3.twins1.lm)
(Intercept)     Natural
  7.2046099   0.9484224
> # intercept for Class3 = 7.2046099
> # slope for Class3 = 0.9484224
>
> # plot separate regression lines (as in the scatter-plot
above)
> # for each Social Class
> abline(class1.twins1.lm,lty=1)
> abline(class2.twins1.lm,lty=2)
> abline(class3.twins1.lm,lty=3)
```

Video 6.7

```
> # fit the same regression line
> # for all Social Classes
> twins4.lm <- lm(Foster ~ Natural,data=twins.df)
> coef(twins4.lm)
(Intercept)     Natural
   9.207599    0.901436

> # specify treatment contrasts;
> # this will ensure dummy variables used are the preferred
0-1 codes
> options(contrasts=c(factor="contr.treatment",
+ ordered="contr.poly"))

> # fit parallel regression lines
> twins2.lm <- lm( Foster ~ Natural + Social,
+ data=twins.df)

> # view model matrix for Model 2

> model.matrix(twins2.lm)
   (Intercept) Natural SocialClass2 SocialClass3
1            1      82            0            0
2            1      90            0            0
3            1      91            0            0
4            1     115            0            0
5            1     115            0            0
6            1     129            0            0
7            1     131            0            0
8            1      78            1            0
9            1      79            1            0
10           1      82            1            0
11           1      97            1            0
12           1     100            1            0
13           1     107            1            0
14           1      68            0            1
15           1      73            0            1
16           1      81            0            1
17           1      85            0            1
18           1      87            0            1
19           1      87            0            1
20           1      93            0            1
21           1      94            0            1
22           1      95            0            1
23           1      97            0            1
24           1      97            0            1
25           1     103            0            1
26           1     106            0            1
27           1     111            0            1
```

```
> # regression coefficients for Model 2
> summary(twins2.lm)

Call:
lm(formula = Foster ~ Natural + Social, data = twins.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8235  -5.2366  -0.1111   4.4755  13.6978 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.6076    11.8551  -0.051    0.960    
Natural        0.9658     0.1069   9.031 5.05e-09 ***
SocialClass2   2.0353     4.5908   0.443    0.662    
SocialClass3   6.2264     3.9171   1.590    0.126    
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1 

Residual standard error: 7.571 on 23 degrees of freedom
Multiple R-Squared: 0.8039,     Adjusted R-squared: 0.7784 
F-statistic: 31.44 on 3 and 23 DF,  p-value: 2.604e-08 

> # intercept = -0.6076 ; coef(Natural) = 0.9658
> # coef(Class2) =  2.0353 ; coef(Class3)= 6.2264
> # intercept for Class1 = -0.6076
> # intercept for Class2 = -0.6076 + 2.0353 = 1.4277
> # intercept for Class3 = -0.6076 + 6.2264 = 5.6188
> # common slope for all classes = 0.9658
```

```
> # fit separate regression lines for each class
> twins1.lm <- + lm( Foster ~ Natural + Social +
Natural:Social,data=twins.df)
> model.matrix(twins1.lm)
   (Intercept) Natural SocialClass2 SocialClass3 Natural:SocialClass2
1            1      82            0            0                    0
2            1      90            0            0                    0
3            1      91            0            0                    0
4            1     115            0            0                    0
5            1     115            0            0                    0
6            1     129            0            0                    0
7            1     131            0            0                    0
8            1      78            1            0                   78
9            1      79            1            0                   79
10           1      82            1            0                   82
11           1      97            1            0                   97
12           1     100            1            0                  100
13           1     107            1            0                  107
14           1      68            0            1                    0
15           1      73            0            1                    0
16           1      81            0            1                    0
17           1      85            0            1                    0
18           1      87            0            1                    0
19           1      87            0            1                    0
20           1      93            0            1                    0
21           1      94            0            1                    0
22           1      95            0            1                    0
23           1      97            0            1                    0
24           1      97            0            1                    0
25           1     103            0            1                    0
26           1     106            0            1                    0
27           1     111            0            1                    0
   Natural:SocialClass3
1                     0
2                     0
3                     0
4                     0
5                     0
6                     0
7                     0
8                     0
9                     0
10                    0
11                    0
12                    0
13                    0
14                   68
15                   73
16                   81
17                   85
18                   87
19                   87
20                   93
21                   94
22                   95
23                   97
24                   97
25                  103
26                  106
27                  111
```

```
> # regression coefficients for Model 1
> summary(twins1.lm)

Call:
lm(formula = Foster ~ Natural + Social + Natural:Social,
data = twins.df)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4795  -5.2484  -0.1550   4.5822  13.7984

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.872044  17.808264  -0.105    0.917
Natural               0.977562   0.163192   5.990 6.04e-06
***
SocialClass2          2.688068  31.604178   0.085    0.933
SocialClass3          9.076654  24.448704   0.371    0.714
Natural:SocialClass2 -0.004995   0.329525  -0.015    0.988
Natural:SocialClass3 -0.029140   0.244580  -0.119    0.906
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 7.921 on 21 degrees of freedom
Multiple R-Squared: 0.8041,     Adjusted R-squared: 0.7574
F-statistic: 17.24 on 5 and 21 DF,  p-value: 8.31e-07

> # intercept for Class1 = -1.872044
> # slope for Class 1 = 0.977562
> # intercept for Class2 = -1.872044 + 2.688068 = 0.816024
> # slope for Class2 = 0.977562 -0.004995 = 0.972567
> # intercept for Class3 = -1.872044 + 9.076654 = 7.20461
> # slope for Class3 = 0.977562 -0.029140 = 0.948422
```

Video 6.8

```
> # comparing models
>
> # testing for parallel regression lines;
> # compare Model 2 and Model 1
> anova(twins2.lm,twins1.lm)
Analysis of Variance Table

Model 1: Foster ~ Natural + Social
Model 2: Foster ~ Natural + Social + Natural:Social
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     23 1318.40
2     21 1317.47  2      0.93 0.0074 0.9926


> # testing for coincident regression lines;
> # compare Model 4 and Model 1
> anova(twins4.lm,twins1.lm)
Analysis of Variance Table

Model 1: Foster ~ Natural
Model 2: Foster ~ Natural + Social + Natural:Social
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     25 1493.53
2     21 1317.47  4    176.06 0.7016 0.5996

> # quit R
> q("yes")
```

Video 6.9

Summer 2006 Question 6

An experiment was conducted to study the relation between the depth of ruts in a pavement and the volume of traffic. Three different types of pavements (T1, T2 and T3) were considered. Several samples of each pavement type were observed over a period of time as they were exposed to increasing amounts of traffic.

A model of the following form was fitted to these data:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 X Z_1 + \beta_5 X Z_2 + e$$

where Y = Depth and X = Volume. The following set of dummy variables Z_1 and Z_2 was used to indicate the Type of pavement:

	Z_1	Z_2
T1	0	0
T2	1	0
T3	0	1

R output from this and related models is shown on the following page.
Use this output to test the following hypotheses:

(i) the true regression lines for each type are parallel

(ii) the true regression lines for each type coincide

(iii) the true regression lines for types T1 and T2 have the same intercept

(iv) the true regression lines for types T1 and T2 have the same slope

In each case, express the hypothesis in terms of the parameters of the above model. For each test, quote the value of the test statistic and the associated *p*-value.

For each type of pavement, write down the equation of the fitted regression line of Depth on Volume

(v) when parallel regression lines are fitted for each type.

(vi) when separate regression lines are fitted for each type.

R output for Question 6

```
> options(contrasts=c(factor="contr.treatment",
+ ordered="contr.poly"))

> summary(rutting2.lm)
Call:
lm(formula = Depth ~ Volume + Type, data = rutting.df)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2260     0.2960  10.900 7.29e-10 ***
Volume        1.9537     0.3411   5.728 1.32e-05 ***
TypeT2       -1.6660     0.2076  -8.025 1.11e-07 ***
TypeT3       -3.0345     0.1963 -15.461 1.38e-12 ***

> summary(rutting1.lm)
Call:
lm(formula = Depth ~ Volume + Type + Volume:Type,
data = rutting.df)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.2503     0.3457   3.617  0.00197 **
Volume          4.6795     0.4600  10.173 6.85e-09 ***
TypeT2          0.4611     0.4096   1.126  0.27502
TypeT3         -0.3785     0.4132  -0.916  0.37166
Volume:TypeT2  -2.9418     0.5469  -5.379 4.12e-05 ***
Volume:TypeT3  -3.6634     0.5492  -6.671 2.94e-06 ***

> rutting4.lm <- lm(Depth ~ Volume,data=rutting.df)

> anova(rutting2.lm,rutting1.lm)
Analysis of Variance Table

Model 1: Depth ~ Volume + Type
Model 2: Depth ~ Volume + Type + Volume:Type
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     20 2.64838
2     18 0.75173  2   1.89665 22.708 1.196e-05 ***

> anova(rutting4.lm,rutting1.lm)
Analysis of Variance Table

Model 1: Depth ~ Volume
Model 2: Depth ~ Volume + Type + Volume:Type
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     22 35.241
2     18  0.752  4    34.489 206.46 8.968e-15 ***
```

Summer 2005 Question 6

Data were collected on the Price (in dollars per hundred weight) and the Weight (in hundreds of pounds) of 29 heifers. These heifers were also graded into one of the following Grades: G1, G2 and G3.

A model of the following form was fitted to these data:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 X Z_1 + \beta_5 X Z_2 + e$$

where Y = Price and X = Weight. The following set of dummy variables Z_1 and Z_2 was used to indicate the Grade of the heifers:

	Z_1	Z_2
G1	0	0
G2	1	0
G3	0	1

S-PLUS output from this and related models is shown on the following page.
Use this output to test the following hypotheses:

(i) the true regression lines for each Grade are parallel

(ii) the true regression lines for each Grade coincide

(iii) the true regression lines for Grades G1 and G2 have the same intercept

(iv) the true regression lines for Grades G1 and G2 have the same slope

In each case, express the hypothesis in terms of the parameters of the above model. For each test, quote the value of the test statistic and the associated *p*-value.

Write down the values of the estimated regression coefficients

(v) when parallel regression lines are fitted for each Grade

(vi) when separate regression lines are fitted for each Grade

R output for Question 6

```
> options(contrasts=c(factor="contr.treatment",
+ ordered="contr.poly"))

> summary(livstock2.lm)
Call: lm(formula = Price ~ Weight + Grade,
data = livstock.df)
Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)  71.6423   5.8474    12.2520   0.0000
     Weight  -6.4960   2.1191    -3.0655   0.0052
    GradeG2 -11.7054   2.5488    -4.5924   0.0001
    GradeG3 -14.1778   2.5173    -5.6321   0.0000

> summary(livstock1.lm)
Call: lm(formula = Price ~ Weight + Grade + Weight:Grade,
data = livstock.df)

Coefficients:
                   Value Std. Error   t value  Pr(>|t|)
  (Intercept)  159.8328   23.6608      6.7552    0.0000
       Weight  -39.9156    8.9509     -4.4594    0.0002
      GradeG2  -99.1754   24.8581     -3.9897    0.0006
      GradeG3 -110.9903   24.7968     -4.4760    0.0002
WeightGradeG2   33.1940    9.2560      3.5862    0.0016
WeightGradeG3   36.4892    9.3180      3.9160    0.0007

> livstock4.lm <- lm(Price ~ Weight,data=livstock.df)

> anova(livstock2.lm,livstock1.lm)
Analysis of Variance Table
Response: Price
                          Terms Resid. Df      RSS
1                Weight + Grade        25 657.2615
2 Weight + Grade + Weight:Grade        23 394.1409
           Test Df Sum of Sq  F Value       Pr(F)
1
2 +Weight:Grade  2  263.1206 7.677171 0.002792563

>  anova(livstock4.lm,livstock1.lm)
Analysis of Variance Table
Response: Price
                          Terms Resid. Df      RSS
1                        Weight        27 1590.259
2 Weight + Grade + Weight:Grade        23  394.141
                 Test Df Sum of Sq F Value         Pr(F)
1
2 +Grade+Weight:Grade  4  1196.118 17.4498 1.041425e-006
```