Section 16

Linear constraints in multiple linear regression. Analysis of variance.

Multiple linear regression with general linear constraints. Let us consider a multiple linear regression $Y = X\beta + \varepsilon$ and suppose that we want to test a hypothesis given by a set of s linear equations. In a matrix form:

$$H_0 : A\beta = c,$$

where A is a $s \times p$ matrix and c is a $s \times 1$ vector. We will assume that $s \leq p$ and the matrix A has rank s. This generalizes two types of hypotheses from previous lecture, when we considered only one linear combination of parameters ($s = 1$ case) or tested hypothesis about all parameters simultaneously ($s = p$ case).

To test this general hypothesis, a natural idea is to compare how far $A\hat{\beta}$ is from c and to do this we need to find the distribution of $A\hat{\beta}$. Clearly, this distribution is normal with mean $A\beta$ and covariance

$$\mathbb{E}A(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T A^T = A\mathrm{Cov}(\hat{\beta})A^T = \sigma^2 A(X^T X)^{-1}A^T = \sigma^2 D$$

where we introduced a notation

$$D := A(X^T X)^{-1}A^T.$$

A matrix D is a symmetric positive definite invertible $s \times s$ matrix and, therefore, we can take its square root $D^{1/2}$. It is easy to check that the covariance of $D^{-1/2}A(\hat{\beta} - \beta)$ is $\sigma^2 I$. This implies that

$$\frac{1}{\sigma^2}|D^{-1/2}A(\hat{\beta} - \beta)|^2 = \frac{1}{\sigma^2}(A(\hat{\beta} - \beta))^T D^{-1}A(\hat{\beta} - \beta) \sim \chi^2_s.$$

Under null hypothesis, $A\beta = c$, we get

$$\frac{1}{\sigma^2}(A\hat{\beta} - c)^T D^{-1}(A\hat{\beta} - c) \sim \chi^2_s. \tag{16.0.1}$$

Since $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ is independent of $\hat{\beta}$, we get

$$\frac{1}{s\sigma^2}(A\hat{\beta}-c)^T D^{-1}(A\hat{\beta}-c)\Big/\frac{n\hat{\sigma}^2}{(n-p)\sigma^2}$$
$$= \frac{n-p}{ns\hat{\sigma}^2}(A\hat{\beta}-c)^T D^{-1}(A\hat{\beta}-c) \sim F_{s,n-p}. \tag{16.0.2}$$

This is enough to test hypothesis H_0. However, in a variety of applications a different equivalent representation of (16.0.1) is more useful. It is given in terms of MLE $\hat{\beta}_A$ of β that satisfies the constraint in H_0. In other words, $\hat{\beta}_A$ is obtained by solving:

$$|Y - X\beta|^2 \to \min_{\beta} \quad \text{subject to the constraint} \quad A\beta = c. \tag{16.0.3}$$

Lemma. *If $\hat{\beta}_A$ is solution of (16.0.3) then the left hand side of (16.0.1) is equal to*

$$\frac{1}{\sigma^2}|X(\hat{\beta}_A - \hat{\beta})|^2. \tag{16.0.4}$$

Proof. First, let us find the constrained MLE $\hat{\beta}_A$ explicitly. By method of Lagrange multipliers we need to solve a system of two equations:

$$A\beta = c, \;\; \frac{\partial}{\partial\beta}\Big(|Y - X\beta|^2 + (A\beta - c)^T\lambda\Big) = 0,$$

where λ is a $s \times 1$ vector. The second equation is

$$-2X^TY + 2X^TX\beta + A^T\lambda = 0.$$

Solving this for β gives

$$\hat{\beta}_A = (X^TX)^{-1}X^TY - \frac{1}{2}(X^TX)^{-1}A^T\lambda = \hat{\beta} - \frac{1}{2}(X^TX)^{-1}A^T\lambda.$$

Since $\hat{\beta}_A$ must satisfy the linear constraint, we get

$$c = A\hat{\beta}_A = A\hat{\beta} - \frac{1}{2}A(X^TX)^{-1}A^T\lambda = A\hat{\beta} - \frac{1}{2}D\lambda.$$

Solving this for λ, $\lambda = 2D^{-1}(A\hat{\beta} - c)$, we get

$$\hat{\beta}_A = \hat{\beta} - (X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c).$$

and, therefore,

$$X(\hat{\beta}_A - \hat{\beta}) = -X(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c).$$

We can use this formula to compute

$$|X(\hat{\beta}_A - \hat{\beta})|^2 = (X(A\hat{\beta} - \hat{\beta}))^T X(\hat{\beta}_A - \hat{\beta})$$
$$= (A\hat{\beta} - c)^T(X(X^TX)^{-1}A^TD^{-1})^T X(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c)$$
$$= (A\hat{\beta} - c)^T D^{-1}A(X^TX)^{-1}X^TX(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c).$$
$$= (A\hat{\beta} - c)^T D^{-1}A(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c)$$
$$= (A\hat{\beta} - c)^T D^{-1}DD^{-1}(A\hat{\beta} - c)$$
$$= (A\hat{\beta} - c)^T D^{-1}(A\hat{\beta} - c).$$

Comparing with (16.0.1) proves Lemma.

□

Using (16.0.2) and Lemma, we get that under null hypothesis H_0:

$$\frac{n-p}{ns\hat{\sigma}^2}|X(\hat{\beta}_A - \hat{\beta})|^2 \sim F_{s,n-p}. \tag{16.0.5}$$

There are many different models that are special cases of a multiple linear regression and many hypotheses about these model can be written as a general linear constraints. We will describe one such model in detail - one-way layout in analysis of variance. Then we will describe a couple of other models without going into details since the idea will become clear.

Analysis of variance: one-way layout. Suppose that we are given p independent samples

$$\begin{gathered} Y_{11}, \ldots, Y_{1n_1} \sim N(\mu_1, \sigma^2) \\ \vdots \\ Y_{p1}, \ldots, Y_{pn_p} \sim N(\mu_p, \sigma^2) \end{gathered}$$

of sizes $n_1, \ldots, n_p$ correspondingly. We assume that the variance of all distributions are equal. We would like to test the hypothesis that the means of all distributions are equal,

$$H_0 : \mu_1 = \ldots = \mu_p.$$

This problem is in fact a special case of a multiple linear regression and testing hypothesis given by linear equations. We can write

$$Y_{ki} = \mu_k + \varepsilon_{ki}, \quad \text{where} \quad g_{ki} \sim N(0, \sigma^2), \quad \text{for} \quad k = 1, \ldots, p, \;\; i = 1 \ldots, n_i.$$

Let us consider $n \times 1$ vector, where $n = n_1 + \ldots + n_p$,

$$Y = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{p1}, \ldots, Y_{pn_p})^T$$

and $p \times 1$ parameter vector

$$\mu = (\mu_1, \ldots, \mu_p)^T.$$

Then we can write all the equations in a matrix form

$$Y = X\mu + \varepsilon,$$

where X is the following $n \times p$ matrix:

$$X = \left(\begin{array}{cccc} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 \\ \hline 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{array}\right).$$

The blocks have $n_1, \ldots, n_p$ rows. Basically, the predictor matrix X consists of indicators to which group the observation belongs to. The hypothesis H_0 can be written in a matrix form as $A\mu = 0$ for $(p-1) \times p$ matrix

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 & -1 \\ 0 & 1 & \ldots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & -1 \end{pmatrix}.$$

We need to compute the statistic in (16.0.5) that will have distribution $F_{p-1,n-p}$. First of all,

$$X^T X = \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & n_r \end{pmatrix}.$$

Since $\hat{\mu} = (X^T X)^{-1} X^T Y$ it is easy to see that for each $i \leq p$,

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik} = \bar{Y}_i \text{ - the average of } i\text{th sample.}$$

We also get

$$\hat{\sigma}^2 = \frac{1}{n} |Y - X\mu|^2 = \frac{1}{n} \sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_i)^2.$$

To find the MLE $\hat{\mu}_A$ under the linear constraints $A\mu = 0$ we simply need to minimize $|Y - X\mu|^2$ over vectors $\mu = (\mu_1, \ldots, \mu_1)^T$ with all equal coordinates. But, obviously, $X\mu$ is a vector $(\mu_1, \ldots, \mu_1)^T$ of size $n \times 1$, so we need to minimize

$$\sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - \mu_1)^2 \min_{\mu_1}$$

and we get

$$\mu_1 = \frac{1}{n} \sum_{i=1}^{p} \sum_{k=1}^{n_i} Y_{ik} = \bar{Y} \text{ - overall average of all samples.}$$

Therefore,

$$\hat{\mu}_A - \hat{\mu} = (\bar{Y} - \bar{Y}_1, \ldots, \bar{Y} - \bar{Y}_p)^T$$

and

$$|X(\hat{\mu}_A - \hat{\mu})|^2 = \sum_{i=1}^{p} \sum_{k=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{p} n_i (\bar{Y}_i - \bar{Y})^2.$$

By (16.0.5), under the null hypothesis H_0,

$$F := \frac{n-p}{p-1} \frac{\sum_{i=1}^{p} n_i (\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_i)^2} \sim F_{p-1,n-p}. \tag{16.0.6}$$

In order to test H_0, we define a decision rule

$$\delta = \begin{cases} H_0, & F \le c_\alpha \\ H_1, & F > c_\alpha \end{cases}$$

where $F_{p-1,n-p}(c_\alpha, +\infty) = \alpha$. The sum in the numerator in (16.0.6) represents the total variation of the sample means $\bar{Y}_i$ of each population around the overall mean $\bar{Y}$. The sum in the numerator represent the total variation of the observations Y_{ik} around their particular sample means $\hat{Y}_i$. This interpretation of the test statistic explains the name - analysis of variance, or anova.

□

Example. Let us again consider normal body temperature dataset and perform anova test to compare the mean body temperature for men and women. Previously we have tested this using t-tests and KS test for two samples. We use Matlab function

```
[p,tbl,stats]=anova1([men, women])
```

where 'men' and 'women' are 65×1 vectors. For unequal groups 'anova1' requires a second argument with group labels. The output produces a table 'tbl':

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
'Columns'	[2.7188]	[1]	[2.7188]	[5.2232]	[0.0239]
'Error'	[66.6262]	[128]	[0.5205]		
'Total'	[69.3449]	[129]			

'SS' gives the sum of squares in the numerator of (16.0.6) ('Columns'), denominator ('Error'), and their total sum. Degrees of freedom 'df' represent degrees of freedom $p-1$ and $n-p$. 'MS' represents the normalized sums of squares by corresponding degrees of freedom. 'F' is a statistic in (16.0.6) and 'Prob>F' is a p-value corresponding to this F-statistic. This means that at the level of significance $\alpha = 0.05$ we reject the null hypothesis that the means are equal.

□

Analysis of variance: two-way layout. Suppose that we again have samples from different groups only now the groups will have two categories defined by two factors. For example, if we want to compare SAT scores in different states but also separate public and private schools then we will have groups defined by two factors - state and school type. We consider the following model of the data:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

for $i = 1, \dots, a, j = 1, \dots, b$ and $k = 1, \dots, n_{ij}$, i.e. we have a categories of the first factor, b categories of the second factor and n_{ij} observations in group (i, j). This model is not any different from one-way anova, simply the groups are indexed by two parameters/factors, but the estimation of parameters can be carried out as in the one-way anova. However, to test various hypotheses about the effects of these two factors it is more convenient to write the model in an equivalent way as follow:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where we assume that

$$\sum_{i=1}^{a} \alpha_i = 0, \ \sum_{j=1}^{b} \beta_j = 0, \ \sum_{i=1}^{b} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0.$$

These constraints define all parameters uniquely from original parameters μ_{ij}. Parameter μ is called the *overall mean.* The reason we separate *additive effects* α_i and β_j of two factors from the most general *interaction effect* γ_{ij} is because it is easier to formulate various hypotheses in terms of these parameters. For example:

- $H_0 : \alpha_1 = \ldots = \alpha_a = 0$ - the *additive* effect of the first factor is insignificant;
- $H_0 : \beta_1 = \ldots = \beta_b = 0$ - the *additive* effect of the second factor is insignificant;
- $H_0 :$ all $\gamma_{ij} = 0$ - the effect of the *interaction* of both factors is insignificant, i.e. the effect of factors is additive.

Matlab function 'anova2' performs two-way layout of anova if the sizes of all groups n_{ij} are equal, i.e. the data is *balanced*. If the sizes of groups are different one should use 'anovan' - a general N-way anova.

□

Analysis of covariance. This is another special case of multiple regression when all groups of data have a continuous predictor variable. The model is:

$$Y_{ik} = \alpha + \alpha_i + (\beta + \beta_i)X_{ik} + \varepsilon_{ik}$$

for $i = 1, \ldots, a$ and $k = 1, \ldots, n_i$. We have a groups and n_i observations in ith group. To determine the parameters uniquely we assume that

$$\sum_{i=1}^{a} \alpha_i = 0, \ \sum_{i=1}^{a} \beta_i = 0.$$

Example. (*Fruitfly dataset*) We consider a dataset from [1] (available on the journal's website) and [2]. The experiment consisted of five groups of male fruitflies, 25 male fruitflies in each group. The males in each group were supplied with different number of either receptive or non receptive females each day.

Group 1: 8 newly inseminated non-receptive females per day;

Group 2: no females;

Group 3: 1 newly inseminated non-receptive female per day;

Group 4: 1 receptive female per day;

Group 5: 8 receptive females per day.

The experiment was designed to test if the increased reproduction results in decreased longevity, so the lifespan of each male fruitfly was the response variable Y.

One-way anova. Let us start with a one-way anova, i.e. we consider a model

$$Y_{ij} = \mu_i + \varepsilon_{ik}, \quad \text{where } i = 1, \ldots, 5, \ \ k = 1, \ldots, 25$$

and test the hypothesis $H_0 : \mu_1 = \ldots = \mu_5$. Suppose that 'lifespan1' is a 25×5 matrix such that each column contains observations from one group. Then running

```
[p,tbl,stats]=anova1(lifespan1);
```

produces the boxplot in figure 16.1 and a table 'tbl':

```
'Source'     'SS'            'df'   'MS'            'F'        'Prob>F'
'Columns'    [1.1939e+004]   [  4]  [2.9848e+003]  [13.6120]  [3.5156e-009]
'Error'      [2.6314e+004]   [120]  [   219.2793]
'Total'      [3.8253e+004]   [124]
```

p-value suggests how unlikely hypothesis H_0 is. The boxplot suggests that the last group's

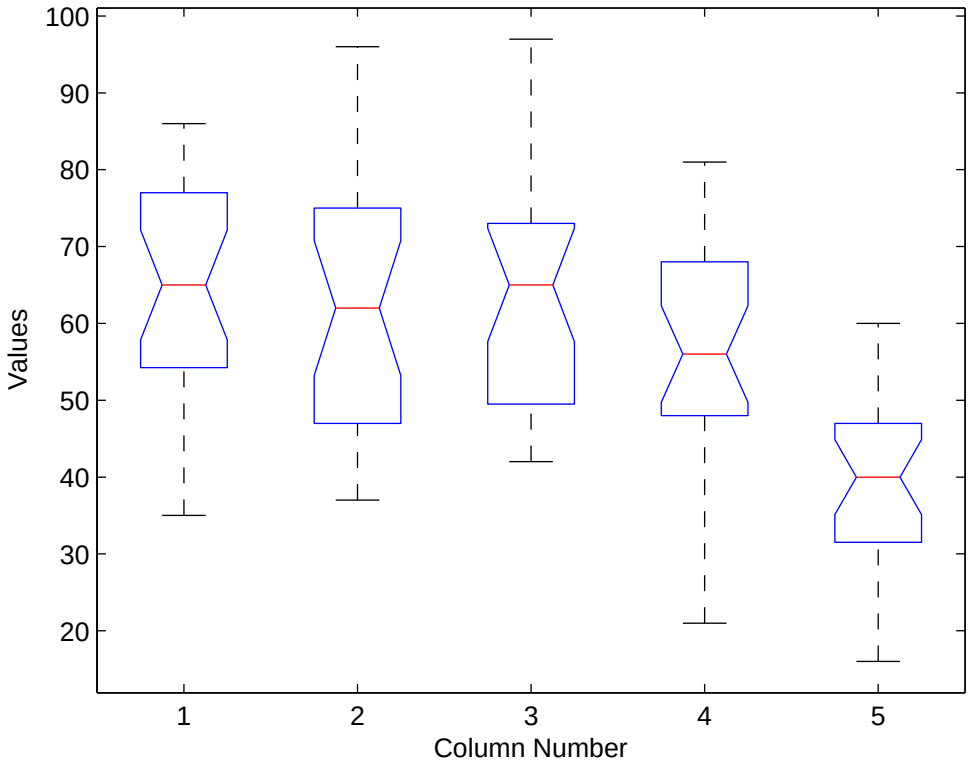


Figure 16.1: Boxplot for one-way ANOVA.

lifespan is most different from the other four groups. As a result, we might want to test the hypothesis $H_0 : \mu_1 = \ldots = \mu_4$ that the means of the first four groups are equal,

```
[p,tbl,stats]=anova1(lifespan1(:,1:4));
```

we get the following table

```
'Source'   'SS'           'df' 'MS'        'F'       'Prob>F'
'Columns'  [   988.0800] [ 3] [329.3600] [1.3869] [0.2515]
'Error'    [2.2798e+004] [96] [237.4842]
'Total'    [2.3787e+004] [99]
```

and we see that the p-value is 0.2515, so we accept H_0 if the level of significance $\alpha \leq p$-value.

□

Two-way anova. Let us now consider four groups without the second group (no females) and test the effects of two factors:

- Factor A: 'receptive' or 'non-receptive';
- Factor B: '1' or '8'.

This means that we consider a model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

for $i = 1, \ldots, 2, j = 1, \ldots, 2$ and $k = 1, \ldots, 25$. To use Matlab function 'anova2' we arrange the data into a 50×2 matrix 'lifespan2' such that two columns represent two categories of Factor A, the first 25 rows represent group '1' in Factor B and rows 26 through 50 represent group '8' in Factor B. Then

```
[p,tbl,stats]=anova2(lifespan2,25)
```

produces (here 25 indicates the number of replicas in one cell) the table

```
'Source'      'SS'           'df' 'MS'           'F'       'Prob>F'
'Columns'     [6.6749e+003] [ 1] [6.6749e+003] [32.3348] [1.3970e-007]
'Rows'        [1.7223e+003] [ 1] [1.7223e+003] [ 8.3430] [     0.0048]
'Interaction' [2.3717e+003] [ 1] [2.3717e+003] [11.4890] [     0.0010]
'Error'       [1.9817e+004] [96] [   206.4308]
'Total'       [3.0586e+004] [99]
```

p-values in the last column correspond to three hypotheses:

- $H_0 : \alpha_1 = \alpha_2 = 0$, i.e. the effect of Factor A is insignificant;
- $H_0 : \beta_1 = \beta_2 = 0$, i.e. the effect of Factor B is insignificant;
- $H_0 : \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$, i.e. the effect of the 'interaction' between Factors A and B is insignificant.

Small p-values suggest that all these hypotheses should be rejected.

□

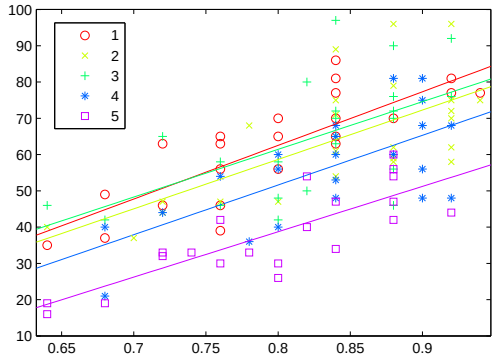
Analysis of covariance. Besides reproduction factors A and B, another continuous explanatory variable for longevity was used - the length of thorax (a division of a body between the head and the abdomen - chest). We are now in the setting of ancova:

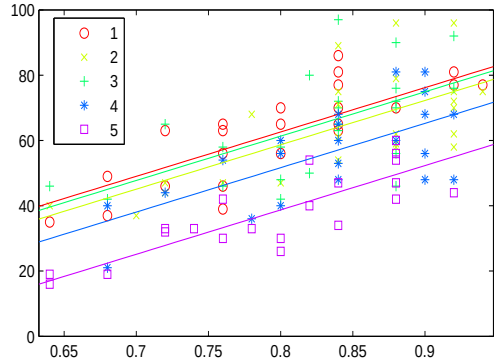
$$Y_{ik} = \alpha + \alpha_i + (\beta + \beta_i)X_{ik} + \varepsilon_{ik}$$

for $i = 1, \ldots, 5$ and $k = 1, \ldots, 25$. Analysis of covariance tool in Matlab

```
aoctool(thorax,lifespan,groups);
```

produces the following output, figure 16.2:





Coefficient Estimates

Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	-53.788	10.5457	-5.1	0
1	-1.912	19.7963	-0.1	0.9232
2	3.546	19.8575	0.18	0.8586
3	10.063	22.6154	0.44	0.6572
4	-4.204	22.703	-0.19	0.8534
5	-7.492	20.2771	-0.37	0.7124
Slope	135.473	12.7783	10.6	0
1	12.316	24.3219	0.51	0.6136
2	0.653	23.7582	0.03	0.9781
3	-4.024	27.3217	-0.15	0.8832
4	1.528	27.1208	0.06	0.9552
5	-10.473	25.0536	-0.42	0.6767

Coefficient Estimates

Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	-54.062	10.2551	-5.27	0
1	8.006	1.8898	4.24	0
2	4.077	1.8894	2.16	0.0329
3	6.73	1.881	3.58	0.0005
4	-2.94	1.8914	-1.55	0.1227
5	-15.873	1.8981	-8.36	0
Slope	135.819	12.439	10.92	0

ANOVA Table

Source	d.f.	Sum Sq	Mean Sq	F	Prob>F
groups	4	9611.5	2402.9	21.09	0
thorax	1	13168.9	13168.9	115.59	0
groups*thorax	4	42.5	10.6	0.09	0.9844
Error	115	13102.1	113.9		

ANOVA Table

Source	d.f.	Sum Sq	Mean Sq	F	Prob>F
groups	4	9611.5	2402.9	21.75	0
thorax	1	13168.9	13168.9	119.22	0
Error	119	13144.7	110.5		

Figure 16.2: Left column top to bottom: graph of fitted line for each group, estimates of coefficients, ancova test table. Right column: same under assumption that all slopes are equal.

We see that the p-value of 'groups*thorax' interaction, corresponding to the hypothesis that all $\beta_i = 0$, is equal to 0.9844, which means that we can accept this hypothesis. As a result, we fit the model with equal slopes for all groups, figure 16.2, right column. The p-values for 'groups' and 'thorax', corresponding to the hypotheses all $\alpha_i = 0$ and $\beta = 0$, are almost 0 and we should reject these hypotheses.

□

References.

[1] Hanley, J. A., and Shapiro, S. H. (1994), "Sexual Activity and the Lifespan of Male Fruitflies: A Dataset That Gets Attention," *Journal of Statistics Education*, Volume 2, Number 1.

[2] Linda Partridge and Marion Farquhar (1981), "Sexual Activity and the Lifespan of Male Fruitflies," *Nature*, 294, 580-581.