Lecture notes: Linear model for RNA-seq analysis

Background

Linear models are a set of versatile tools for describing, investigating and statistically assess dependencies between measured variables. The linear models describe the relation between a 'dependent' variable and a set of covariates, called 'independent' variables. The dependent variable is typically an outcome of an experiment while the independent variables are inputs or parameters that are varied in the experimental setting. Linear models have a strong resemblance with other forms of commonly used statistical methods, such as ANOVA and the t-test, but offers a much higher degree of flexibility. In particular, linear models make it possible to add almost any number of independent variables that can be of a wide range of types and forms (e.g. categorical, continuous, etc).

Linear models are general statistical tools and can be applied to a large variety of problems. In these lecture notes, however, our man aim is to introduce linear models as a tool for the analysis of the expression level of genes measured by an RNA-seq experiment. In particular, we will introduce linear models as a way to model the gene expression as a function of the experimental conditions and thereby identify important patterns in the data.

General definition and notation

We will start by defining the concept of linear models and introduce basic notations. Assume that the random variable Y_j $(j = 1, \dots, m)$ is the outcome of an experiment that has been performed. We will furthermore assume that Y_j is dependent on a set of experimental conditions captured by the variables $x_{1,j}, x_{2,j}, \dots, x_{p,j}$. Note that $x_{1,j}, x_{2,j}, \dots, x_{p,j}$ can be of different types, including both categorical (e.g. 'on/off', 'exposure/no exposure', 'sick/healthy') or continuous ('concentration', 'time', etc). The linear model aims to describe the linear relationship between Y_j and $x_{1,j}, x_{2,j}, \dots, x_{p,j}$. The variable Y_j is therefore called the dependent variable while $x_{1,j}, x_{2,j}, \dots, x_{p,j}$ are called independent variables.

The linear model can be specified through the equation

$$Y_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \cdots + \beta_p x_{p,i} + \varepsilon_j$$

Here, the parameter β_0 is called the intercept and describe the baseline of Y_j, i.e. the value when all dependent variables are zero. The parameters $\beta_1, \dots, \beta_p$ describe how the independent variable influence the dependent variable Y_j. A parameter that has a non-zero value means that the independent variable influences the dependent variable. If, on the other hand, the value is zero or close to zero, the corresponding independent variable has no or little effect on the dependent variable. The parameters $\beta_0, \beta_1, \dots, \beta_p$ are typically unknown and needs to be estimated from the data.

The last term, ε_i, is an error term and describes the changes in the dependent variable that is not explained by the independent variables. The typically includes various forms of measurement errors as well as other experimental conditions that, for some reason, are not included in the model (e.g. because they were not observed in the experiment). In the linear model ε_i is assumed to be independent and to be normally distributed with expected value 0 and variance σ^2, i.e.

$$\varepsilon_j \sim \text{Normal}(0, \sigma^2).$$

The linear model describes function between the expected outcome of Y_j and the independent variables. Indeed, taking the expected value of Y_j gives us

$$\text{Exp}[Y_j] = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \cdots + \beta_p x_{p,j}.$$

An important property of the linear model is that the independent variables can be arbitrarily transformed before included in the linear model. We can, for example, define $x_{j,1}$as

$$x_{j,1} = f(z_{j,1})$$

where $z_{j,1}$ is the input to the experiment and *f* is an arbitrary, typically monotone, function (e.g. square-root or the natural logarithm). These forms of transformations can be important to properly describe the association between the dependent and independent variables. For example, in toxicology the effect of a chemical is often tested along an exponential scale, e.g. experiments are performed for e.g. 0.1, 1, 10, 100 and 1000 nM of a certain toxicant. Transforming these values using a logarithm transform is often advised since this will distribute the concentrations more uniformly.

There are several special cases of linear models.

Special case 1: If $p = 1$ and $x_{j,1}$ takes values along the x-axis, the linear model is reduced to the simple regression model, i.e.

$$Y_j = \beta_0 + \beta_1 x_{j,1} + \varepsilon$$

This model describes a line through the plane where β_0 is the point where the line crosses the y-axis and $\beta_1$is the slope.

Special case 2: If $p = 1$ and $x_{j,1}$ is a categorical variable taking on two values that indicating if sample *j* belongs to category 1 or 2 the corresponding linear model,

$$Y_j = \beta_0 + \beta_1 x_{j,1} + \varepsilon$$

is equivalent to a t-test assuming equal variance between the two categories. If instead $x_{j,1}$ takes *k* values indicating if sample *j* belongs to category 1, 2, …, *k*, the linear model will instead correspond to a one-way ANOVA. Technically, a categorical variable of more than two categories needs to be encoded into multiple binary variables when the model is fit. This is however done automatically by most statistical software.

These special cases show that linear models are flexible and captures many of the existing statistical methodologies under one single framework.

Linear models for analyzing gene expression in RNA-seq experiments

Our aim is to use linear models to analyze gene expression in RNA-seq experiments. In RNA-seq experiments, the expression of 10,000 to 100,000 genes/transcript is typically measured simultaneously. All these genes/transcripts can follow their own specific expression profile over the sample included in the experiment and thus needs to be modelled separately. This is a fundamental difference to the formulation of the linear model above, which was formulated for one single dependent variable. For our RNA-seq data we need instead 10,000 to 100,000 linear models!

We will here assume that the raw RNA-seq data has been properly pre-processed. This means that the generated reads have been quality assessed, aligned to a reference and the transcript abundances have been estimated. We will also assume that the resulting data has been normalized and transformed to be approximately follow a normal distribution. Exactly how this should be performed is dependent on the dataset and how the RNA-seq data was generated and will not be covered here. We will simply assume that these prior steps were successful.

Furthermore, assume that we have *n* genes in the RNA-seq experiment and *m* samples in total. Let $Y_{i,j}$ be the expression level of gene *i* in sample *j*. The expression for each gene *i* and sample *j* will be described using a linear model following the same structure as above, i.e.

$$Y_{i,j} = \beta_{0,i} + \beta_{1,i}x_{1,j} + \beta_{2,i}x_{2,j} + \cdots + \beta_{p,i}x_{p,j} + \varepsilon_{i,j}.$$

Please note that the indexing of the parameters and independent variables. The parameters are gene-specific, i.e. different for each gene, and are therefore indexed with *i*. Since all genes are measured in one single experiment, the independent variables are the same and does not depend on the gene. They vary, however, between the different samples as before and are therefore indexed with *j*.

The error term $\varepsilon_{i,j}$ is assumed to be independent between both genes and samples and follows a normal distribution with expected value 0 and variance σ_i^2. Thus, the variability is assumed to be different for different genes. This assumption can be supported by biological arguments. Some genes, for example actins, are known to have an often high and relatively stable expression between samples and experimental conditions (sometimes called 'housekeeping genes'). Other forms of genes can instead vary substantially. For example, the gene SRY, which is one of the few genes located on the human Y-chromosome, is only expressed in males and will therefore vary a lot between samples in experiment based on human tissue that includes both sexes.

To demonstrate how linear models can be formulated for RNA-seq data we will now take a look on a few examples.

Example: Gene expression in malign and benign cancer

In a scientific study, RNA-seq was used to characterize the expression of 45,015 transcripts ('genes') in small intestinal neuroendocrine tumors (SI-NET), a relatively uncommon form of

cancer that is often associated with poor prognosis. SI-NETs can be divided into two groups depending on its aggressiveness. In this study, the scientists wanted to analyze the expression of tumors from these two groups and identify genes that are differentially expressed. In total, 33 patients were included in the study of which 23 were from less aggressive group and 10 from more aggressive group. For each patient, a biopsy was sampled, the RNA extracted, sequenced using standard protocols, preprocessed and transformed and normalized so that the resulting data is approximately normal distributed. For each patient, the age and sex were also recorded.

The expression of each genes was described using a linear model. If we define the independent variable $x_{1,j}$ as

$$x_{1,j} = \begin{cases} 1 & \text{if patient } j \text{ has a more aggressive tumor,} \\ 0 & \text{if patient } j \text{ has a less aggresive tumor,} \end{cases}$$

the model can be formulated as

$$Y_{i,j} = \beta_{0,i} + \beta_{1,i} x_{1,j} + \varepsilon_{i,j}.$$

Here, $Y_{i,j}$ is the expression of gene *i* in patient *j*, $\beta_{0,i}$ is the base line expression of gene *i*, $\beta_{1,i}$ is the difference in expression in the more aggressive compared to the less aggressive tumor form and $\varepsilon_{i,j}$ is the error for gene *i* and patient *j*. Note that a positive value of $\beta_{1,i}$ means higher expression in the more aggressive tumors, while a negative value means a higher expression in the less aggressive tumors. If $\beta_{1,i}$ is close to zero there is no difference between the tumor types.

The scientists suspected that the age and sex may be influencing the gene expression and an extended model was therefore formulated. Let $x_{2,j}$ be the age of patient *j*. Also, let

$$x_{3,j} = \begin{cases} 1 & \text{if patient } j \text{ is a male,} \\ 0 & \text{if patient } j \text{ is a female.} \end{cases}$$

The extended model can then be formulated as

$$Y_{ij} = \beta_{0i} + \beta_{1i} x_{1j} + \beta_{2i} x_{2j} + \beta_{3i} x_{3j} + \varepsilon_{ij}.$$

The focus of this model is still $\beta_{1,i}$, i.e. the difference in gene expression between more and less aggressive tumor forms. Adding $x_{2,j}$ and $x_{3,j}$ will however explain the variability in gene expression patterns that is caused by differences in age and sex of the patients.

Example: gene expression in fish exposed to toxicants

Synthetic estrogens, such as etinylestradiol, that are used in contraceptives reach the aquatic environment through our waste water treatment plants. Even in low concentrations, etinylestradiol have been hypothesized to cause adverse effects in fish by perturbing their hormone levels and making them sterile. This hypothesis was investigated by environmental researchers who conducted an experiment where fish from the species *Onchorynchus mykiss* (rainbow trout) were exposed etinylestradiol. After two weeks of exposure, the livers were harvested from the fish and analyzed using RNA-seq to examine the gene expression. The

exposure experiment was performed in four aquaria with increasing concentrations of etinylestradiol in increasing concentrations: 10 ng/L, 100 ng/L, 1000 ng/L and 10ug/L. In addition, a reference aquarium with water with no etinylestradiol was also included.

The researchers wanted to identify genes that had increased or decreased the expression with increasing levels of etinylestradiol. This was done by formulating a regression model where the expression of gene g and sample j was assumed to depend on the etinylestradiol concentration ($x_{1,j}$), i.e.

$$Y_{i,j} = \beta_{0,i} + \beta_{1,i}x_{1,j} + \varepsilon_{i,j}.$$

Here $\beta_{0,i}$ is the intercept and $\beta_{1,i}$ the slope of the regression curve. In order to distribute the concentrations along the x-axis, the concentration variable was transformed into uniform steps with the control starting at the value of 0, the lowest concentration at 1 and so on ending with the highest concentration at the value of 4. This corresponds to a $\log_{10}$-transformation of the concentrations where the control has been added at 0 and with the other transformed concentrations starting at 1.

From previous experiences, the researchers know that fish that have lived in the same aquarium may show similar gene expression patterns. This can, for example, caused by the social structure of the fish in the aquaria which can induce a stress response in some individuals. In order to include this effect into the model, the variable $x_{2,j}$ taking values from the categories "aquarium 1", "aquarium 2", "aquarium 3" and "aquarium 4", was included into the model. The resulting model then becomes

$$Y_{i,j} = \beta_{0,i} + \beta_{1,i}x_{1,j} + \beta_{2,i}x_{2,j} + \varepsilon_{i,j}.$$

Identification of differentially expressed genes can be preformed by analyzing the parameter for the first independent variable, i.e. $\beta_{1,i}$. Note that this model combines a continuous variable ($x_{1,j}$) with a categorical variable ($x_{2,j}$) and thus a combination of a standard regression and ANOVA-analysis.

Estimation of parameters

We will now turn our attention on the estimation of the parameters in the linear model. Estimating the parameter values are done using computer algorithms but knowledge about the underlying techniques are necessary for a proper interpretation.

The parameters in linear models are estimated using the method of maximum likelihood. Thus, to find our estimates, the likelihood function needs to be formulated and then maximized. The data that will be used to estimate the parameters are observations of the dependent and independent variables and these comes in *n* tuples, one for each of the *m* samples. Thus, for a specific gene *i* the observations are $(y_{i1}, x_{11}, \dots, x_{p1}), \dots, (y_{im}, x_{1m}, \dots, x_{pm})$. Furthermore, since ε_{ij} is normally distributed it follows that

$$Y_{ij} \sim \text{Normal}(0, \sigma_i^2).$$

The density function for Y_{ij} for a gene *i* and sample *j* can then be written as

$$f(y_{ij}, x_{1,j}, \dots, x_{p,j}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\left(y_{ij}-(\beta_{0,i}+\beta_{1,i}x_{1,j}+\beta_{2,i}x_{2,j}+\cdots+\beta_{p,i}x_{p,j})\right)^2}{2\sigma_i^2}}, -\infty < y_{ij} < \infty.$$

The likelihood function L for gene i can then formulated as the product of the density over all samples, i.e.

$$L(\beta_{0i}, \beta_{1i}, \dots, \beta_{pi}, \sigma_i^2) = \prod_{j=1}^{n} f(y_{ij}, x_{1j}, \dots, x_{pj}).$$

The corresponding log likelihood function becomes

$$l(\beta_{0i}, \beta_{1i}, \dots, \beta_{pi}, \sigma_i^2) = \log L(\beta_{0i}, \beta_{1i}, \dots, \beta_{pi}, \sigma_i^2) = \sum_{j=1}^{n} \log f(y_{ij}, x_{1j}, \dots, x_{pj})$$

Using some algebra, the log likelihood function can be shown to be

$$l(\beta_{0i}, \beta_{1i}, \dots, \beta_{pi}, \sigma_i^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma_i^2 - \frac{1}{2\sigma_i^2}\sum_{j=1}^{n}\left(y_{ij} - (\beta_{0i} + \beta_{1i}x_{1j} + \beta_{2i}x_{2j} + \cdots + \beta_{pi}x_{pj})\right)^2$$

This function can be maximized to find point estimates of the parameters. For $\beta_{0,i}, \dots, \beta_{p,i}$ this is equivalent to minimizing

$$\sum_{j=1}^{n}\left(y_{ij} - (\beta_{0,i} + \beta_{1,i}x_{1,j} + \beta_{2,i}x_{2,j} + \cdots + \beta_{p,i}x_{p,j})\right)^2,$$

that is, solving a least squares problem (you probably recognize this from e.g. simple regression and ANOVA). Taking the partial derivative of the expression above generates a system of equations,

$$\begin{cases} \sum y_{ij} = n\beta_{0i} + \beta_{1i}\sum x_{1j} + \beta_{2i}\sum x_{2j} + \cdots + \beta_{pi}\sum x_{pj} \\ \sum y_{ij}x_{1j} = \beta_{0i}\sum x_{1j} + \beta_{1i}\sum x_{1j}^2 + \beta_{2i}\sum x_{1j}x_{2j} + \cdots + \beta_{pi}\sum x_{1j}x_{pj} \\ \sum y_{ij}x_{2j} = \beta_{0i}\sum x_{2j} + \beta_{1i}\sum x_{1j}x_{2j} + \beta_{2i}\sum x_{2j}^2 + \cdots + \beta_{pi}\sum x_{2j}x_{pj} \\ \vdots \\ \sum y_{ij}x_{pj} = \beta_{0i}\sum x_{pj} + \beta_{1i}\sum x_{1j}x_{pj} + \beta_{2i}\sum x_{2j}x_{pj} + \cdots + \beta_{pi}\sum x_{pj}^2 \end{cases}$$

that can be solved numerically. This results in point estimates $\hat{\beta}_{0,i}, \hat{\beta}_{1,i}, \dots, \hat{\beta}_{p,i}$ of the true parameters $\beta_{0,i}, \beta_{1,i}, \dots, \beta_{p,i}$.

The variance parameter, σ_i^2, can also be estimated by maximum likelihood and shown to be

$$\hat{\sigma}_i^2 = \frac{1}{n-(p+1)}\sum_{j=1}^{n}\left(y_{ij}-\hat{y}_{ij}\right)^2,$$

where

$$\hat{y}_{ij}=\hat{\beta}_{0i}+\hat{\beta}_{1i}x_{1j}+\hat{\beta}_{2i}x_{2j}+\cdots+\hat{\beta}_{pi}x_{pj}$$

and p is, as before, the number of independent variables.

Inference of parameters

A usual form of inference performed in a linear model is to analyze if a specific independent variable is associated with the dependent variable. This is typically done using a statistical test. If, for a specific gene i, we are interested in if x_{kj} is associated with Y_{ij} over the samples $j = 1, \ldots, m$, we can test whether its corresponding parameter β_{ki} is zero or, alternatively, has a non-zero value. Statistically, this can be formulated by the hypotheses

$$H_0: \beta_{ki} = 0,$$
$$H_A: \beta_{ki} \neq 0.$$

These hypotheses can be tested using the fact that the point estimate $\hat{\beta}_{ki}$ follows a normal distribution, i.e.

$$\hat{\beta}_{ki} \sim \text{Normal}\left(\beta_{ki}, \sigma^2_{\hat{\beta}_{ki}}\right)$$

where $\sigma^2_{\hat{\beta}_{ki}}$, which is the variance of $\hat{\beta}_{ki}$, has a known but rather complicated expression (and will not be specifically stated here). Thus, the hypotheses can be tested by using a test statistic

$$T = \frac{\hat{\beta}_{ki}}{S_{\hat{\beta}_{ki}}}$$

where $S_{\hat{\beta}_{ki}}$ is the sample standard deviation $\hat{\beta}_{ki}$, i.e. the estimated value of $\sigma_{\hat{\beta}_{ki}}$. The distribution of T can be shown to follow a t-distribution with parameter n-(k+1) (i.e. n-(k+1) degrees of freedom). This makes it possible to calculate p-values based on an observation t of T according to

$$p = \text{Prob}(\text{T} > |\text{t}|).$$

Note that the values of $\hat{\beta}_{ki}$ and $S_{\hat{\beta}_{ki}}$are typically calculated by statistical software packages and their explicit forms are not within the scope of this course.

Interpreting model fits

It is often important to evaluate how well a linear model fits the data. A common way to do that is to use the so called R^2-value, which estimates how large proportion of the total variability of the dependent variable that is explained by the linear model. The R^2-value can be calculated

by comparing the variability residuals, i.e. the errors of the model to the overall variability of the dependent variability. First, let SST denote the sum of squares of the total variability, i.e.

$$\text{SST} = \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_i\right)^2,$$

and let SSE denote the residual sum of squares ('sum of squares errors')

$$\text{SSE} = \sum_{j=1}^{n} \left(y_{ij} - \hat{y}_{ij}\right)^2.$$

In these equations, $\bar{y}_i$ denotes the sample mean of y_{ij}, i.e.

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \quad ,$$

and $\hat{y}_{ij}$, as above,

$$\hat{y}_{ij} = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_{1j} + \hat{\beta}_{2i}x_{2j} + \cdots + \hat{\beta}_{pi}x_{pj}.$$

The R^2-value is then calculated by comparing SST to SSE according to

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

This equation shows that if SSE is close to zero, i.e. the error of the model is small so that the dependent variable is almost fully described by the independent variables, the value of R^2 is close to 1. If, on the other hand, SSE is close to the total variability SST, which means that the model does a poor job of describing the dependent variable, the value of R^2 will be smaller or even close to zero.

There is, however, a major issue of the R^2-value – it always increases when adding more independent variables. Thus, based on the interpretation of R^2 a model can always be improved by adding more independent variables even if they are only vaguely associated with the dependent variable. Instead, an adjusted variant of R^2 is often preferred. The adjusted R^2-value penalize for the number of independent variables that are used and are therefore typically more informative that then standard R^2-value. The adjusted R^2-value is defined as

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-(k+1)}\frac{\text{SSE}}{\text{SST}}.$$

The interpretation of R^2_{adj} should be done in the same way as for R^2, i.e. a value close to 1 means a good fit while a value close to zero means a poor fit.

Linear models in R

Linear models can be easily used in the statistical programming language R through the **lm** function. In R, the linear model is specified using a so call 'formula' that defines the model and

specifies how the dependent variable is connected to the independent variable. Once the model is specified, **lm** calculates all point estimates and p-values for all included parameters.

The two main arguments to **lm** is the formula and a data matrix in the form of a **data.frame**. The terms used in the formula should correspond to the names of the columns in the data matrix.

Example: A first example – cats revisited
We will start a simple example on how to use **lm** where we will take a look at the **cats** dataset that you have previously analyzed in computer exercise 1. The **cats** dataset contains 144 observations of cats, their sex (male/female), their body weight ('Bwt', in kilograms) and heart weight ('Hwt', in grams).

```
> library(MASS)
> data(cats)
> dim(cats)
[1] 144   3

> head(cats)
  Sex Bwt Hwt
1   F 2.0 7.0
2   F 2.0 7.4
3   F 2.0 9.5
4   F 2.1 7.2
5   F 2.1 7.3
6   F 2.1 7.6
```

We will use a linear model to see if we describe the heart weight of a cat based on its sex and body weight. Let Y_j be the heart weight of cat j (j=1,…, 144). Let x_{1j} define the sex of cat j, i.e.

$$x_{1j} = \begin{cases} 1 & \text{if cat } j \text{ is male,} \\ 0 & \text{if cat } j \text{ is female.} \end{cases}$$

The corresponding linear model that we will be using is

$$Y_j = \beta_0 + \beta_1 x_{1j} + \varepsilon_j.$$

This model can be specified into R using the **lm** command as

```
> lm(Hwt~Sex, data=cats).

Call:
lm(formula = Hwt ~ Sex, data = cats)

Coefficients:
(Intercept)         SexM
      9.202        2.121
```

Note that the intercept (β_0) is included by default and does not need to be specified explicitly. The result from R shows us that the intercept has been estimated to 9.2 g and that the effect of

being a male increase the heart weight 2.1 g. More information about the results can be obtained using the **summary** function, i.e.,

```
> summary(lm(Hwt~Sex, data=cats))

Call:
lm(formula = Hwt ~ Sex, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8227 -1.7227  0.0273  1.2273  9.1773

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2021     0.3251  28.308  < 2e-16 ***
SexM          2.1206     0.3961   5.354 3.38e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.229 on 142 degrees of freedom
Multiple R-squared:  0.168,     Adjusted R-squared:  0.1621
F-statistic: 28.66 on 1 and 142 DF,  p-value: 3.38e-07
```

This output specifies a wide range of information about the fitted linear model. Under the title ‘Coefficients’, we can see the result for each parameter, including its point estimate (‘Estimate’), estimated standard error (‘Std. Error’), t statistic (‘t value’) and its corresponding p-value when testing the null hypothesis that the parameter should be zero (‘Pr(>|t|)’). Thus, we can see from this analysis that the sex of a cat is highly significant when describing the weight of its heart. Note also that the adjusted R-square is 0.1621, which is quite low. Thus, the sex of a cat has a limited ability to predict the weight of the heart.

We will now further refined the model and also introduce the body weight of the cat. If we let x_{2j} denote the body weight of cat j (which was, in our dataset, given in kg), the refined model can be stated as

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \varepsilon_j.$$

This model can be fitted to the data in R using

```
> lm(Hwt~Sex+Bwt, data=cats)

Call:
lm(formula = Hwt ~ Sex + Bwt, data = cats)

Coefficients:
(Intercept)         SexM          Bwt
    -0.4150      -0.0821       4.0758
```

The parameter for the body weight was estimated to more than 4, which means that the heart weight increase 4 grams for every kilogram of cat. Note that the point estimates of the intercept

and the sex has changed significantly. In fact, the parameter for sex is now very close to zero and has, in this extended model, much smaller impact on the body weight. Using the **summary** function we get

```
Call:
lm(formula = Hwt ~ Sex + Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5833 -0.9700 -0.0948  1.0432  5.1016

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4149     0.7273  -0.571    0.569
SexM         -0.0821     0.3040  -0.270    0.788
Bwt           4.0758     0.2948  13.826   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.457 on 141 degrees of freedom
Multiple R-squared:  0.6468,    Adjusted R-squared:  0.6418
F-statistic: 129.1 on 2 and 141 DF,  p-value: < 2.2e-16
```

which shows that the sex of the cat no longer is significant. A likely explanation for this phenomenon is that the body weight is highly correlated to the sex of the cat. When the body weight is added to the model, the sex of the cat does not provide any additional information for describing the heart weight. Note also that the adjusted R^2 has increased considerably, from 0.1620 to 0.6418, suggesting that this model describes the variability of the data much better.

Example: Gene expression in malign and benign cancer (cont.)
We will now continue our analysis of the gene expression data of the small intestinal neuroendocrine tumors (SI-NETs). We will start by analyzing the data with the smaller model,

$$Y_{i,j} = \beta_{0,i} + \beta_{1,i} x_{1,j} + \varepsilon_{i,j}.$$

where

$$x_{1,j} = \begin{cases} 1 & \text{if patient } j \text{ has a more aggressive tumor,} \\ 0 & \text{if patient } j \text{ has a less aggresive tumor.} \end{cases}$$

$Y_{i,j}$ is the expression of gene *i* in patient *j*, $\beta_{0,i}$ is the base line expression of gene *i*, $\beta_{1,i}$ is the difference in expression in the more aggressive compared to the less aggressive tumor form and $\varepsilon_{i,j}$ is the error for gene *i* and patient *j*. In order to analyze the entire dataset, we need to use the model for each of the 45,015 transcripts. This can easily be done in R using a for-loop. The input data, here called **data.exp**, is organized in a data.frame with a row for each gene and a column for each individual of the 33 patients.

```
> head(data.exp[,1:5])
     251485046843_1 251485046843_3 251485053884_4 251485046844_2 251485046889_1
<NA>      16.273216      16.099402      14.764864      15.183461      15.345287
<NA>       2.691950       2.661967       1.437180       2.037386       1.885127
<NA>       3.258086       5.123285       1.437180       1.766553       1.955436
```

```
<NA>       4.159501       5.123285       1.853821       1.617931       1.835417
<NA>       2.008186       3.184030       1.311090       1.617931       1.835417
<NA>       4.893363       4.958962       1.437180       1.572721       2.067911
```

Another data structure, here called **annotation**, contains information of the patients in the form of a data.frame with one row for aggressiveness of the tumor, the sex and the age of the patient.

```
> head(annotation)
          Sample Group Sex Age
1 251485046843_1    G2   Male  49
3 251485046843_3    G1 Female  81
4 251485053884_4    G1 Female  64
5 251485046844_2    G1 Female  42
6 251485046889_1    G1 Female  74
7 251485053885_3    G1 Female  58
```

The order of the column of **data.exp** and the rows of **annotation** is organized so that they correspond to the same patient. The R-code for analyzing this data can be seen below. Note that we need to use the function **summary** to make R to calculate a p-value from the results calculated by **lm**.

```
# R code for fitting a linear model to each transcript

ngenes=nrow(data.exp)      # Number of rows (transcripts) in the expression data
parameter=rep(NA, ngenes) # Allocate a vector for parameter values
pvalue=rep(NA, ngenes)    # Allocate a vector for p-values

for(i in 1:ngenes){       # Loop over the number of transcripts

  # Extract the data for transcript i by indexing 'data.exp'and 'annotation'. Note
  # that this new data needs to be a data.frame in order to feed it into lm
  data.cur=data.frame(data.exp[i,], annotation[,2], annotation[,3], annotation[,4])

  # Add column names to 'data.cur'
  colnames(data.cur)=c("Expression", "Group", "Sex", "Age")

  # Fit the linear model to the data
  res.lm=lm(Expression~Group, data=data.cur)

  # Save the parameter 1 ("Group") from the result. Note that res.lm$coefficient[1]
  # is the intercept.
  parameter1[i]=res.lm$coefficient[2]

  # Calculate p-values (and other stuffs for the model)
  res.summary.lm=summary(res.lm)

  # Extact the p-value. Note that $coefficient after running summary
  # is a matrix with a row for each parameter.
  pvalue1[i]=res.summary.lm$coefficient[2,4]
}
```

The model can be extended by adding more independent variables in the formula to **lm**. For example,

```
res.lm=lm(Expression~Group+Age+Geneder, data=data.cur)
```

corresponds to the full model described on page 4.

Interpretation of the result, which consist 45,015 linear models, is not trivial. This is however beyond the scope of this text. Refer instead to the notes from the lecture addressing analysis of RNA-seq data.