Mayo-Illinois Computational Genomics Course: Statistics and R

Dave Zhao

Department of Statistics
University of Illinois at Urbana-Champaign (UIUC)

Thanks for Casey Hanson for parts of the lab component of these slides.
Outline

General concepts
  Data
  Inference
  Causality

Hypothesis testing

Statistical modeling
  Overview
  Ex. #1: Is USP37 associated with high risk MM?
  Ex. #2: Is USP37 associated with high risk MM, controlling for gender?
  Ex. #3: Does USP37 affect MM risk differently in men vs. women?
  Ex. #4: Are USP37, gender, and age predictive of MM risk?

Multiple testing
General concepts: Data: Data table

Fundamental quantity of statistical analysis:
Feature 1 . . . Feature j . . . Feature p

Subject 1
X 11 . . . X 1j . . . X 1p

Subject 2
X 21 . . . X 2j . . . X 2p

... ... ... ... ...

Subject i
X i1 . . . X ij . . . X ip

... ... ... ... ...

Subject n
X n1 . . . X nj . . . X np

▶ One row per subject, subjects typically independent
▶ Dependent/independent variables are chosen from features
Fundamental quantity of statistical analysis:

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>...</th>
<th>Feature $j$</th>
<th>...</th>
<th>Feature $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>$X_{11}$</td>
<td>...</td>
<td>$X_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>Subject 2</td>
<td>$X_{21}$</td>
<td>...</td>
<td>$X_{2j}$</td>
<td>...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Subject $i$</td>
<td>$X_{i1}$</td>
<td>...</td>
<td>$X_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Subject $n$</td>
<td>$X_{n1}$</td>
<td>...</td>
<td>$X_{nj}$</td>
<td>...</td>
</tr>
</tbody>
</table>
General concepts: Data: Data table

Fundamental quantity of statistical analysis:

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>...</th>
<th>Feature j</th>
<th>...</th>
<th>Feature p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>(X_{11})</td>
<td>...</td>
<td>(X_{1j})</td>
<td>...</td>
</tr>
<tr>
<td>Subject 2</td>
<td>(X_{21})</td>
<td>...</td>
<td>(X_{2j})</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Subject (i)</td>
<td>(X_{i1})</td>
<td>...</td>
<td>(X_{ij})</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Subject (n)</td>
<td>(X_{n1})</td>
<td>...</td>
<td>(X_{nj})</td>
<td>...</td>
</tr>
</tbody>
</table>

- One row per subject, subjects typically independent
Fundamental quantity of statistical analysis:

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>...</th>
<th>Feature j</th>
<th>...</th>
<th>Feature p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>$X_{11}$</td>
<td>...</td>
<td>$X_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>Subject 2</td>
<td>$X_{21}$</td>
<td>...</td>
<td>$X_{2j}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Subject i</td>
<td>$X_{i1}$</td>
<td>...</td>
<td>$X_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Subject n</td>
<td>$X_{n1}$</td>
<td>...</td>
<td>$X_{nj}$</td>
<td>...</td>
</tr>
</tbody>
</table>

- One row per subject, subjects typically independent
- Dependent/independent variables are chosen from features
In bioinformatics you may be given the table as:

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Subject 1</th>
<th>…</th>
<th>Subject $i$</th>
<th>…</th>
<th>Subject $n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>$X_{11}$</td>
<td>…</td>
<td>$X_{i1}$</td>
<td>…</td>
<td>$X_{n1}$</td>
</tr>
<tr>
<td>Feature 2</td>
<td>$X_{12}$</td>
<td>…</td>
<td>$X_{i2}$</td>
<td>…</td>
<td>$X_{n2}$</td>
</tr>
<tr>
<td>Feature $j$</td>
<td>$X_{1j}$</td>
<td>…</td>
<td>$X_{ij}$</td>
<td>…</td>
<td>$X_{nj}$</td>
</tr>
<tr>
<td>Feature $p$</td>
<td>$X_{1p}$</td>
<td>…</td>
<td>$X_{ip}$</td>
<td>…</td>
<td>$X_{np}$</td>
</tr>
</tbody>
</table>

Make sure you’re giving the analysis software the orientation that it expects.
General concepts: Data: Data types

Two main types of features:

- Continuous:
  - Age
  - Blood albumin level
  - Microarray gene expression levels

- Discrete:
  - Categorical:
    - Gender
    - Experimental group
  - Count
  - RNA-seq gene expression levels

Data types are important for choosing the right statistical analysis procedure.
General concepts : Data : Data types

Two main type of features:

1. Continuous
   - Age
   - Blood albumin level
   - Microarray gene expression levels

Data types are important for choosing the right statistical analysis procedure.
Two main types of features:

1. Continuous
   - Age
   - Blood albumin level
   - Microarray gene expression levels

2. Discrete
General concepts: Data: Data types

Two main types of features:

1. Continuous
   - Age
   - Blood albumin level
   - Microarray gene expression levels

2. Discrete
   - Categorical
     - Gender
     - Experimental group
General concepts: Data: Data types

Two main types of features:

1. Continuous
   - Age
   - Blood albumin level
   - Microarray gene expression levels

2. Discrete
   - Categorical
     - Gender
     - Experimental group
   - Count
     - RNA-seq gene expression levels
General concepts : Data : Data types

Two main types of features:

1. Continuous
   - Age
   - Blood albumin level
   - Microarray gene expression levels

2. Discrete
   - Categorical
     - Gender
     - Experimental group
   - Count
     - RNA-seq gene expression levels

Data types are important for choosing the right statistical analysis procedure
General concepts: Data: Lab: RStudio

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale.

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.
Creating a single number

```r
> x = 3
> x
[1] 3
```
Creating a single number

> x = 3
> x
[1] 3

- The first line does two things:
  1. Creates a variable called `x` in the current R environment
  2. Sets the value of `x` to 3
Creating a single number

> x = 3
> x
[1] 3

- The first line does two things:
  1. Creates a variable called x in the current R environment
  2. Sets the value of x to 3
- The second line prints the value of x to the R console
Creating a single number

> x = 3
> x
[1] 3

- The first line does two things:
  1. Creates a variable called x in the current R environment
  2. It sets the value of x to 3
- The second line prints the value of x to the R console
- The third line indicates that x contains one element, equal to 3.
Creating a single number

> x = 3
> x
[1] 3

The first line does two things:
1. Creates a variable called x in the current R environment
2. It sets the value of x to 3

The second line prints the value of x to the R console

The third line indicates that x contains one element, equal to 3.

Variable names are case-sensitive, should not contain spaces, cannot start with numbers or certain special characters
The R Environment is the set of variables that the user creates while inputting commands into R.
The R Environment is the set of variables that the user creates while inputting commands into R.

Unknown variable

> y
Error: object ’y’ not found
Creating a vector of numbers

```r
> z = c(1,2,3)
> z
[1] 1 2 3
```
There are many possible data structures in addition to single numbers and vectors:

http://www.statmethods.net/input/datatypes.html
There are many possible data structures in addition to single numbers and vectors:
http://www.statmethods.net/input/datatypes.html

These data structures can contain numbers, characters, or mixtures of both
Mathematical operations

> x = 3
> y = 5
> z = c(1,2,3)
> 3 + 5
[1] 8
> x + 5
[1] 8
> x + y
[1] 8
Vectorized operations

> z
[1] 1 2 3
> z + 3
[1] 4 5 6
> z * 3
[1] 3 6 9
> z / 4
[1] 0.25 0.50 0.75
> z - 5
[1] -4 -3 -2
General concepts : Data : Lab : Manipulating data

Functions

```r
> z
[1] 1 2 3
> mean(z)
[1] 2
> mean(z / 4)
[1] 0.5
> z = mean(z / 4)
> z
[1] 0.5
```
General concepts: Data: Lab: Manipulating data

Functions

```r
> z
[1] 1 2 3
> mean(z)
[1] 2
> mean(z / 4)
[1] 0.5
> z = mean(z / 4)
> z
[1] 0.5
```

Notice that we reassigned the variable `z` at the end. Look at the R Environment.
In practice we will usually load data from existing files.

This lab will use a subset of the data obtained from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24080

I’ve prepared the data in a R-specific file format. The file is mm_tt2.RData
Step 1: clear the current environment

```r
> rm(list=ls())
> ls()
character(0)
```
Step 2: set working directory

```r
> setwd("~/Dropbox/teaching/mayo_genomics_2017/")
> getwd()
```

- Set the working directory to the directory that contains the file `mm_tt2.RData`.
- This can also be done graphically in RStudio.
General concepts: Data: Lab: Loading data

Step 3: load dataset

```R
> load("mm_tt2.RData")
```
Step 3: load dataset

> load("mm_tt2.RData")

Notice the change in the R environment
Examine the data

```r
> dim(mm_tt2)
[1] 336 54678
> head(colnames(mm_tt2))
[1] "deceased24" "X1007_s_at" "X1053_at" "X117_at"
[5] "X121_at" "X1255_g_at"
> tail(colnames(mm_tt2))
[1] "XAFFX-ThrX-M_at" "XAFFX-TrpnX-3_at" "XAFFX-TrpnX-5_at"
[4] "XAFFX-TrpnX-M_at" "age" "sex"
```

In RStudio you can click the `mm_tt2` object in the R environment pane to view the data.
**Subsetting the dataframe**

```r
> mm_tt2[1:5,1:3]

deceased24  X1007_s_at  X1053_at  
GSM592395   0   10.1735    9.1585  
GSM592396   0   9.7107     8.7031  
GSM592397   0   10.1822    8.5578  
GSM592398   0   9.4696     8.5545  
GSM592399   0   8.7541     8.7543  
```

▶ `mm_tt2[,1:3]` will print all rows of the first 3 columns.

▶ What data type is `deceased24`?
- Discrete categorical

▶ What data type is `x1007_s_at`?
- Continuous
Subsetting the dataframe

```r
> mm_tt2[1:5,1:3]

deceased24   X1007_s_at   X1053_at
GSM592395     0   10.1735    9.1585
GSM592396     0   9.7107     8.7031
GSM592397     0   10.1822    8.5578
GSM592398     0   9.4696     8.5545
GSM592399     0   8.7541     8.7543
```

▸ `mm_tt2[,1:3]` will print all rows of the first 3 columns.
Subsetting the dataframe

\[
> \text{mm\_tt2}[1:5,1:3]
\]

| GSM592395 | 0 | 10.1735 | 9.1585 |
| GSM592396 | 0 | 9.7107  | 8.7031 |
| GSM592397 | 0 | 10.1822 | 8.5578 |
| GSM592398 | 0 | 9.4696  | 8.5545 |
| GSM592399 | 0 | 8.7541  | 8.7543 |

- mm\_tt2[,1:3] will print all rows of the first 3 columns.
- What data type is deceased24?
Subsetting the dataframe

```r
> mm_tt2[1:5,1:3]

deceased24   X1007_s_at   X1053_at
GSM592395    0     10.1735     9.1585
GSM592396    0     9.7107     8.7031
GSM592397    0     10.1822     8.5578
GSM592398    0     9.4696     8.5545
GSM592399    0     8.7541     8.7543
```

- `mm_tt2[,1:3]` will print all rows of the first 3 columns.
- What data type is deceased24? Discrete categorical.
Subsetting the dataframe

> mm_tt2[1:5,1:3]

<table>
<thead>
<tr>
<th>deceased24</th>
<th>X1007_s_at</th>
<th>X1053_at</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM592395</td>
<td>0</td>
<td>10.1735</td>
</tr>
<tr>
<td>GSM592396</td>
<td>0</td>
<td>9.7107</td>
</tr>
<tr>
<td>GSM592397</td>
<td>0</td>
<td>10.1822</td>
</tr>
<tr>
<td>GSM592398</td>
<td>0</td>
<td>9.4696</td>
</tr>
<tr>
<td>GSM592399</td>
<td>0</td>
<td>8.7541</td>
</tr>
</tbody>
</table>

- `mm_tt2[,1:3]` will print all rows of the first 3 columns.
- What data type is `deceased24`? Discrete categorical
- What data type is `X1007_s_at`?
Subsetting the dataframe

```r
> mm_tt2[1:5,1:3]
deceased24    X1007_s_at   X1053_at
GSM592395      0      10.1735    9.1585
GSM592396      0      9.7107     8.7031
GSM592397      0      10.1822    8.5578
GSM592398      0      9.4696     8.5545
GSM592399      0      8.7541     8.7543
```

- `mm_tt2[,1:3]` will print all rows of the first 3 columns.
- What data type is `deceased24`? Discrete categorical
- What data type is `X1007_s_at`? Continuous
Extracting specific features

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>285</td>
<td>51</td>
</tr>
</tbody>
</table>

> `table(mm_tt2$deceased24)`

> `mm_tt2$deceased24[5:20] + 2`

[1] 2 2 2 3 2 2 3 2 3 2 2 2 2 2 2 2

The second command shows that the extracted feature behaves just like a (numeric) vector.
General concepts : Inference : Sample vs. population
General concepts : Inference : Sample vs. population

We are usually interested learning about characteristics of some target **population**, e.g. “all patients with multiple myeloma”.

We usually can’t collect the entire population so instead we randomly draw **sample** of $n$ observations to experiment on; these produce the data table `mm_tt2.RData`.
General concepts: Inference: Sample vs. population

We are usually interested in learning about characteristics of some target **population**, e.g. “all patients with multiple myeloma”.

We usually can’t collect the entire population so instead we randomly draw **sample** of \( n \) observations to experiment on; these produce the data table `mm_tt2.RData`.

http://ssrmc.wm.edu/case-selection/
Repeated experiments will give different observed data due to random sampling.

We usually don’t care about the sample *per se*, since the data in the sample are random. We care more about the population.
General concepts : Inference : Definition

**Statistical inference** is the process of learning characteristics of the population by using a sample.

The role of statistics:
- Propose methods of inference that are as accurate as possible
- Quantify the uncertainty of the resulting inferences
General concepts : Causality : Correlation vs. causation
An experimental study (rather than an observation study) can distinguish correlation and causation.
Population stratification is an example of correlation vs. causation.
Hypothesis testing : Population

The population we want to learn about the 12 balls in this urn.

http://www.statisticshowto.com/polya-urn/
Hypothesis testing: Research question

Suppose we know the balls are only two colors: red and green. Research question: what is the distribution of the colors?

http://www.statisticshowto.com/polya-urn/
Hypothesis testing: Null hypothesis

We think the color distribution might look like this:

http://www.statisticshowto.com/polya-urn/
Hypothesis testing : Null hypothesis

We think the color distribution might look like this:

Null hypothesis = specification of what we think the population might look like.
Hypothesis testing : Test statistic

We draw one ball at random to see if our null is correct:

http://www.statisticshowto.com/polya-urn/
Hypothesis testing : Test statistic

We draw one ball at random to see if our null is correct:

Test statistic = sample statistic used to **falsify** null hypothesis. Here, sample = the ball you drew, test statistic = its color (green).
Hypothesis testing: $P$-value

Do you think the null hypothesis is correct?

http://www.statisticshowto.com/polya-urn/
Hypothesis testing: $P$-value

Do you think the null hypothesis is correct?

$P$-value $\approx$ probability of getting the observed test statistic if the null were true.

Here, $p$-value $\approx$ probability of drawing a green ball if the null were true.

http://www.statisticshowto.com/polya-urn/
Hypothesis testing : Significance level

If the $p$-value is very low, we conclude that the null is false. $P$-values measure evidence against the null.

http://www.statisticshowto.com/polya-urn/
Hypothesis testing: Significance level

If the \( p \)-value is very low, we conclude that the null is false. \( P \)-values measure evidence against the null.

Significance level (\( \alpha \)) = pre-determined cutoff for \( p \)-values; if \( p < \alpha \) then **reject** null, otherwise you don’t have enough evidence to reject. Usually we set \( \alpha = 0.05 \).
Hypothesis testing: Power

Probability of:

<table>
<thead>
<tr>
<th></th>
<th>Reject null</th>
<th>Don’t reject null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null false</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Power: the probability of correctly rejecting the null when it’s false. Here: power = probability of rejecting the hypothesis that the urn contains only one green ball.

Power depends on:
1. True contents of urn (effect size)
2. How many balls you draw (sample size)
## Hypothesis testing: Power

Probability of:

<table>
<thead>
<tr>
<th></th>
<th>Reject null</th>
<th>Don’t reject null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>Significance level ($\alpha$)</td>
<td>$1 - \alpha$</td>
</tr>
<tr>
<td>Null false</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Hypothesis testing: Power

### Probability of:

<table>
<thead>
<tr>
<th></th>
<th>Reject null</th>
<th>Don’t reject null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>Significance level ($\alpha$)</td>
<td>$1 - \alpha$</td>
</tr>
<tr>
<td>Null false</td>
<td><strong>Power</strong></td>
<td>1 - power</td>
</tr>
</tbody>
</table>
Hypothesis testing: Power

Probability of:

<table>
<thead>
<tr>
<th></th>
<th>Reject null</th>
<th>Don’t reject null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>Significance level ($\alpha$)</td>
<td>$1 - \alpha$</td>
</tr>
<tr>
<td>Null false</td>
<td><strong>Power</strong></td>
<td>$1 - \text{power}$</td>
</tr>
</tbody>
</table>

**Power**: the probability of correctly rejecting the null when it’s false.

Here: power = probability of rejecting the hypothesis that the urn contains only one green ball.
Hypothesis testing: Power

Probability of:

<table>
<thead>
<tr>
<th>Null true</th>
<th>Reject null</th>
<th>Don’t reject null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>Significance level ($\alpha$)</td>
<td>$1 - \alpha$</td>
</tr>
<tr>
<td>Null false</td>
<td><strong>Power</strong></td>
<td>$1 - \text{power}$</td>
</tr>
</tbody>
</table>

**Power**: the probability of correctly rejecting the null when it’s false.
Here: power = probability of rejecting the hypothesis that the urn contains only one green ball.

Power depends on

1. True contents of urn (**effect size**)
2. How many balls you draw (**sample size**)


Hypothesis testing: Confidence intervals

Scientific journals are now requiring more than just \( p \)-values.
Hypothesis testing: Confidence intervals

Scientific journals are now requiring more than just $p$-values.

One reason: you can usually always get a significant $p$-value by increasing the sample size (increasing power).
Hypothesis testing: Confidence intervals

Scientific journals are now requiring more than just $p$-values.

One reason: you can usually always get a significant $p$-value by increasing the sample size (increasing power).

Better approach: report confidence intervals: a range of estimates of the effect size.

Here: an example 95% confidence interval would be $[0,0.63]$, which says that we’re 95% “confident” that the true proportion of green balls is between 0 and 0.63.
We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a statistical model
4. Identify parameters of interest
5. Formulate null hypothesis
6. Perform hypothesis test

Terms:
- **Statistical model**: a precise mathematical statement about the population-level relationship between the dependent variable and the independent variables
- **Parameters**: unknown variables in the statistical model whose values we want to infer
We will explore various research questions using the myeloma data following these steps:
We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
Statistical modeling: Overview: Basic analysis workflow

We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables

Terms:
- **Statistical model**: a precise mathematical statement about the population-level relationship between the dependent variable and the independent variables
- **Parameters**: unknown variables in the statistical model whose values we want to infer
Statistical modeling: Overview: Basic analysis workflow

We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a statistical model

Terms:
- Statistical model: a precise mathematical statement about the population-level relationship between the dependent variable and the independent variables
- Parameters: unknown variables in the statistical model whose values we want to infer
Statistical modeling: Overview: Basic analysis workflow

We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a statistical model
4. Identify parameters of interest
We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a statistical model
4. Identify parameters of interest
5. Formulate null hypothesis
Statistical modeling: Overview: Basic analysis workflow

We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a **statistical model**
4. Identify **parameters** of interest
5. Formulate null hypothesis
6. Perform hypothesis test
We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a **statistical model**
4. Identify **parameters** of interest
5. Formulate null hypothesis
6. Perform hypothesis test

**Terms:**

- **Statistical model**: a precise mathematical statement about the population-level relationship between the dependent variable and the independent variables
- **Parameters**: unknown variables in the statistical model whose values we want to infer
Statistical modeling: Overview: Basic analysis workflow

We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a **statistical model**
4. Identify **parameters** of interest
5. Formulate null hypothesis
6. Perform hypothesis test

Terms:

- **Statistical model**: a precise mathematical statement about the *population-level* relationship between the dependent variable and the independent variables
Statistical modeling: Overview: Basic analysis workflow

We will explore various research questions using the myeloma data following these steps:

1. Formulate biological question
2. Identify independent and dependent variables
3. Write down a **statistical model**
4. Identify **parameters** of interest
5. Formulate null hypothesis
6. Perform hypothesis test

Terms:

- **Statistical model**: a precise mathematical statement about the *population-level* relationship between the dependent variable and the independent variables
- **Parameters**: unknown variables in the statistical model whose values we want to infer
Statistical modeling : Example #1

1. Biological question: Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma?

2. Variables:
   ▶ (Dependent) expression_i = expression (usually log-expression) of USP37 in subject i
   ▶ (Independent) deceased_i = 1 if subject i was deceased by 24 months, 0 otherwise

3. Regression model:
   expression_i ∼ β_0 + β_1 deceased_i

This is a one-way ANOVA model.
1. Biological question:

*Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma?*
Statistical modeling : Example #1

1. Biological question:

   *Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma?*

2. Variables:

   ▶ (Dependent)
   
   expression \(i\) = expression (usually log-expression) of USP37 in subject \(i\)

   ▶ (Independent)
   
   deceased \(i\) = 1 if subject \(i\) was deceased by 24 months, 0 otherwise

This a one-way ANOVA model.
Statistical modeling: Example #1

1. Biological question:

   Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma?

2. Variables:

   ▶ (Dependent) expression\(_i\) = expression (usually log-expression) of USP37 in subject \(i\)

   ▶ (Independent) deceased\(_i\) = 1 if subject \(i\) was deceased by 24 months, 0 otherwise
Statistical modeling : Example #1

1. Biological question:

   Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma?

2. Variables:

   ▶ (Dependent) $expression_i = \text{expression (usually log-expression)}$ of USP37 in subject $i$

   ▶ (Independent) $deceased_i = 1$ if subject $i$ was deceased by 24 months, 0 otherwise

3. Regression model:
1. Biological question:

   *Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma?*

2. Variables:

   - (Dependent) $expression_i = \text{expression (usually log-expression)}$ of USP37 in subject $i$
   - (Independent) $deceased_i = 1$ if subject $i$ was deceased by 24 months, 0 otherwise

3. Regression model:

   $$expression_i \sim \beta_0 + \beta_1 deceased_i$$

   This a **one-way ANOVA** model.
4. Parameters of interest:

The regression model implies $\beta_0 + \beta_1$ when $i$ deceased

When $i$ = 0 is $\beta_0$

This means that $\beta_1$ quantifies whether average USP37 expression changes between high- and low-risk patients.

Parameters of interest: $\beta_1$ (usually in units of log-fold change)

5. Null hypothesis:

Recall: null = what we think the population looks like.

We usually choose the null by asking: what would the population look like if "nothing interesting were going on"?

Here we set the null as $H_0: \beta_1 = 0$. If we reject $H_0$, we conclude that USP37 is differentially expressed.
4. Parameters of interest:
The regression model implies

- expression; when deceased; = 1 is $\beta_0 + \beta_1$

This means that $\beta_1$ quantifies whether average USP37 expression changes between high- and low-risk patients.
4. Parameters of interest:
The regression model implies

- \textit{expression}_i \text{ when } \text{deceased}_i = 1 \text{ is } \beta_0 + \beta_1
- \textit{expression}_i \text{ when } \text{deceased}_i = 0 \text{ is } \beta_0
4. Parameters of interest:
The regression model implies

- $expression_i$ when $deceased_i = 1$ is $\beta_0 + \beta_1$
- $expression_i$ when $deceased_i = 0$ is $\beta_0$
- This means that $\beta_1$ quantifies whether average USP37 expression changes between high- and low-risk patients
4. Parameters of interest:
The regression model implies

- $expression_i$ when $deceased_i = 1$ is $\beta_0 + \beta_1$
- $expression_i$ when $deceased_i = 0$ is $\beta_0$
- This means that $\beta_1$ quantifies whether average USP37 expression changes between high- and low-risk patients

Parameters of interest: $\beta_1$ (usually in units of log-fold change)
4. Parameters of interest:
The regression model implies

- \( expression_i \) when \( deceased_i = 1 \) is \( \beta_0 + \beta_1 \)
- \( expression_i \) when \( deceased_i = 0 \) is \( \beta_0 \)
- This means that \( \beta_1 \) quantifies whether average USP37 expression changes between high- and low-risk patients

Parameters of interest: \( \beta_1 \) (usually in units of log-fold change)

5. Null hypothesis:
4. Parameters of interest:
   The regression model implies
   - $expression_i$ when $deceased_i = 1$ is $\beta_0 + \beta_1$
   - $expression_i$ when $deceased_i = 0$ is $\beta_0$
   - This means that $\beta_1$ quantifies whether average USP37 expression changes between high- and low-risk patients

Parameters of interest: $\beta_1$ (usually in units of log-fold change)

5. Null hypothesis:
   - Recall: null = what we think the population looks like.
Statistical modeling: Example #1

4. Parameters of interest:
The regression model implies

- $expression_i$ when $deceased_i = 1$ is $\beta_0 + \beta_1$
- $expression_i$ when $deceased_i = 0$ is $\beta_0$
- This means that $\beta_1$ quantifies whether average USP37 expression changes between high- and low-risk patients

Parameters of interest: $\beta_1$ (usually in units of log-fold change)

5. Null hypothesis:

- Recall: null = what we think the population looks like.
- We usually choose the null by asking: what would the population look like if “nothing interesting were going on”?
4. Parameters of interest:
The regression model implies
- expression$_i$ when deceased$_i = 1$ is $\beta_0 + \beta_1$
- expression$_i$ when deceased$_i = 0$ is $\beta_0$
- This means that $\beta_1$ quantifies whether average USP37 expression changes between high- and low-risk patients
Parameters of interest: $\beta_1$ (usually in units of log-fold change)

5. Null hypothesis:
- Recall: null = what we think the population looks like.
- We usually choose the null by asking: what would the population look like if "nothing interesting were going on"?
- Here we set the null as $H_0 : \beta_1 = 0$. If we reject $H_0$, we conclude that USP37 is differentially expressed.
One-way ANOVA

```r
> fit = glm(X232033_at ~ deceased24, data = mm_tt2, family = "gaussian")
> fit

Call: glm(formula = X232033_at ~ deceased24, family = "gaussian", data = mm_tt2)

Coefficients:
(Intercept) deceased24
       7.1490      -0.2201

Degrees of Freedom: 335 Total (i.e. Null); 334 Residual
Null Deviance:       280.1
Residual Deviance:   278  AIC: 895.8
```
X232033_at ~ deceased24 specifies the regression model; the intercept is included by default.

data = mm_tt2 tells R what dataset contains the variables ‘232033_at‘ and deceased24.

family = "gaussian" tells R that the expression outcome is a continuous data type (an normally distributed)
One-way ANOVA, more information about results

> summary(fit)

Call:
glm(formula = X232033_at ~ deceased24, family = "gaussian", data = mm_tt2)

Deviance Residuals:

    Min      1Q  Median      3Q     Max
-4.4911  -0.4224   0.1507   0.5808   1.8405

Coefficients:

                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               7.14905    0.05404  132.291  <2e-16 ***
deceased24               -0.22008    0.13871   -1.587   0.114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
Is USP37 differentially expressed at the 0.05 significance level?

| Coefficients | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | 7.14905  | 0.05404    | 132.291 | <2e-16 *** |
| deceased24   | -0.22008 | 0.13871    | -1.587  | 0.114    |
Is USP37 differentially expressed at the 0.05 significance level?

### Coefficients:

| Term               | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------|----------|------------|---------|----------|
| (Intercept)        | 7.14905  | 0.05404    | 132.291 | <2e-16   | ***    |
| deceased24         | -0.22008 | 0.13871    | -1.587  | 0.114    |        |

No.
Statistical modeling : Example #2

1. Biological question:
   Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma, controlling for gender?

2. Variables:
   ■ (Dependent) $expression_i = \text{expression (usually log-expression)} \text{ of USP37 in subject } i$
   ■ (Independent) $deceased_i = 1 \text{ if subject } i \text{ was deceased by 24 months, 0 otherwise}$
   ■ (Independent) $sex_i = 1 \text{ if subject } i \text{ is female, 0 otherwise}$
1. Biological question:

*Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma, controlling for gender?*
1. Biological question:

*Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma, controlling for gender?*

2. Variables:
1. Biological question:

Is the expression of gene 232033_at (USP37) associated with high risk multiple myeloma, controlling for gender?

2. Variables:

- (Dependent) $\text{expression}_i = \text{expression (usually log-expression)}$ of USP37 in subject $i$
- (Independent) $\text{deceased}_i = 1$ if subject $i$ was deceased by 24 months, 0 otherwise
- (Independent) $\text{sex}_i = 1$ if subject $i$ is female, 0 otherwise
Statistical modeling: Example #2: Simpson’s paradox

What does it mean to “control for gender”?
Statistical modeling: Example #2: Simpson’s paradox

What does it mean to “control for gender”?

(Figure from Wikipedia)

Blue = male, red = female, analyses with and without accounting for gender give opposite results
Statistical modeling : Example #2

3. Regression model:

\[ i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i \]

This is a two-way ANOVA model.

4. Parameters of interest: The regression model implies that

\[ \text{expression}_i \text{ when deceased}_i = 1 \text{ and } \text{sex}_i = s \] is

\[ \beta_0 + \beta_1 + \beta_2 s \]

\[ \text{expression}_i \text{ when deceased}_i = 0 \text{ and } \text{sex}_i = s \] is

\[ \beta_0 + \beta_2 s \]

Parameters of interest: \( \beta_1 \)

5. Null hypothesis: \( H_0: \beta_1 = 0 \)

Is \( \beta_1 \) in Example #2 equal to \( \beta_1 \) in Example #1? Usually not. This is an example of Simpson's paradox.
3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 deceased_i + \beta_2 sex_i \]

This a two-way ANOVA model.
Statistical modeling: Example #2

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 deceased_i + \beta_2 sex_i \]

This is a **two-way ANOVA** model.

4. Parameters of interest:
3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i \]

This is a **two-way ANOVA** model.

4. Parameters of interest:
The regression model implies that

- *expression* \(_i\) when deceased \(_i\) = 1 and sex \(_i\) = s is \(\beta_0 + \beta_1 + \beta_2 s\)
Statistical modeling: Example #2

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i \]

This a two-way ANOVA model.

4. Parameters of interest:
The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = s \) is \( \beta_0 + \beta_1 + \beta_2 s \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = s \) is \( \beta_0 + \beta_2 s \)
Statistical modeling : Example #2

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i \]

This a two-way ANOVA model.

4. Parameters of interest:

The regression model implies that
- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = s \) is \( \beta_0 + \beta_1 + \beta_2 s \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = s \) is \( \beta_0 + \beta_2 s \)

Parameters of interest: still \( \beta_1 \)
3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 deceased_i + \beta_2 sex_i \]

This a **two-way ANOVA** model.

4. Parameters of interest:

The regression model implies that

- \( expression_i \) when \( deceased_i = 1 \) and \( sex_i = s \) is \( \beta_0 + \beta_1 + \beta_2 s \)
- \( expression_i \) when \( deceased_i = 0 \) and \( sex_i = s \) is \( \beta_0 + \beta_2 s \)

Parameters of interest: still \( \beta_1 \)

5. Null hypothesis: \( H_0 : \beta_1 = 0 \)
Statistical modeling : Example #2

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 deceased_i + \beta_2 sex_i \]

This a two-way ANOVA model.

4. Parameters of interest:
The regression model implies that

- \( expression_i \) when \( deceased_i = 1 \) and \( sex_i = s \) is \( \beta_0 + \beta_1 + \beta_2 s \)
- \( expression_i \) when \( deceased_i = 0 \) and \( sex_i = s \) is \( \beta_0 + \beta_2 s \)

Parameters of interest: still \( \beta_1 \)

5. Null hypothesis: \( H_0 : \beta_1 = 0 \)

Is \( \beta_1 \) in Example #2 equal to \( \beta_1 \) in Example #1?
Statistical modeling : Example #2

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i \]

This a two-way ANOVA model.

4. Parameters of interest:
The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = s \) is \( \beta_0 + \beta_1 + \beta_2 s \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = s \) is \( \beta_0 + \beta_2 s \)

Parameters of interest: still \( \beta_1 \)

5. Null hypothesis: \( H_0 : \beta_1 = 0 \)

Is \( \beta_1 \) in Example #2 equal to \( \beta_1 \) in Example #1?
Usually not. This is an example of Simpson's paradox.
Two-way ANOVA

> fit = glm(X232033_at ~ deceased24 + sex, data = mm_tt2, family = "gaussian")
> summary(fit)

Call:
glm(formula = X232033_at ~ deceased24 + sex, family = "gaussian",
    data = mm_tt2)

Deviance Residuals:
          Min       1Q   Median       3Q      Max
-4.4662  -0.4207   0.1483   0.5835  1.8075

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)   7.12441   0.06916 103.0100  <2e-16 ***
deceased24  -0.22027   0.13885  -1.5858    0.1134
 sex        0.05756   0.10067   0.5722    0.5681

44 / 58
Is USP37 differentially expressed at the 0.05 significance level, controlling for gender?
Is USP37 differentially expressed at the 0.05 significance level, controlling for gender?

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 7.12441  | 0.06916    | 103.010 | <2e-16   *** |
| deceased24     | -0.22027 | 0.13885    | -1.586  | 0.114    |
| sex            | 0.05756  | 0.10067    | 0.572   | 0.568    |

No.
Statistical modeling: Example #3

1. Biological question: Does gene 232033_at (USP37) affect multiple myeloma risk differently in men vs. women?

2. Variables:
   - (Dependent) expression_i = expression (usually log-expression) of USP37 in subject i
   - (Independent) deceased_i = 1 if subject i was deceased by 24 months, 0 otherwise
   - (Independent) sex_i = 1 if subject i is female, 0 otherwise
1. Biological question:

Does gene 232033_at (USP37) affect multiple myeloma risk differently in men vs. women?
1. Biological question:

Does gene 232033_at (USP37) affect multiple myeloma risk differently in men vs. women?

2. Variables:
1. Biological question:

*Does gene 232033_at (USP37) affect multiple myeloma risk differently in men vs. women?*

2. Variables:

- (Dependent) \( expression_i \) = expression (usually log-expression) of USP37 in subject \( i \)
- (Independent) \( deceased_i \) = 1 if subject \( i \) was deceased by 24 months, 0 otherwise
- (Independent) \( sex_i \) = 1 if subject \( i \) is female, 0 otherwise
3. Regression model:

\[ i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \times \text{sex}_i \]

This is a saturated two-way ANOVA model.

4. Parameters of interest:

The regression model implies that when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 1 \) is

\[ \beta_0 + \beta_1 + \beta_2 + \beta_3 \]

when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 1 \) is

\[ \beta_0 + \beta_2 \]

when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 0 \) is

\[ \beta_0 + \beta_1 + \beta_3 \]

when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 0 \) is

\[ \beta_0 + \beta_3 \]

5. Null hypothesis:

\[ H_0: \beta_3 = 0 \]
3. Regression model:

\[ \text{expression}_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \times \text{sex}_i \]

This is a saturated two-way ANOVA model.
Statistical modeling : Example #3

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 deceased_i + \beta_2 sex_i + \beta_3 deceased_i \times sex_i \]

This a saturated two-way ANOVA model.

4. Parameters of interest:
3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \times \text{sex}_i \]

This is a saturated two-way ANOVA model.

4. Parameters of interest:
The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_1 + \beta_2 + \beta_3 \)

Effect of USP37 is \( \beta_1 + \beta_3 \) in females and \( \beta_1 \) in males.
Statistical modeling: Example #3

3. Regression model:

$$expression_i \sim \beta_0 + \beta_1 deceased_i + \beta_2 sex_i + \beta_3 deceased_i \times sex_i$$

This is a saturated two-way ANOVA model.

4. Parameters of interest:

The regression model implies that

- $expression_i$ when $deceased_i = 1$ and $sex_i = 1$ is $\beta_0 + \beta_1 + \beta_2 + \beta_3$
- $expression_i$ when $deceased_i = 0$ and $sex_i = 1$ is $\beta_0 + \beta_2$
Statistical modeling: Example #3

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \times \text{sex}_i \]

This is a saturated two-way ANOVA model.

4. Parameters of interest:

The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_1 + \beta_2 + \beta_3 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_2 \)
- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 0 \) is \( \beta_0 + \beta_1 \)
3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \times \text{sex}_i \]

This is a saturated two-way ANOVA model.

4. Parameters of interest:

The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_1 + \beta_2 + \beta_3 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_2 \)
- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 0 \) is \( \beta_0 + \beta_1 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 0 \) is \( \beta_0 \)
Statistical modeling: Example #3

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \ast \text{sex}_i \]

This is a **saturated two-way ANOVA** model.

4. Parameters of interest:
The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_1 + \beta_2 + \beta_3 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_2 \)
- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 0 \) is \( \beta_0 + \beta_1 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 0 \) is \( \beta_0 \)
- Effect of USP37 is \( \beta_1 + \beta_3 \) in females and \( \beta_1 \) in males
Statistical modeling: Example #3

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \ast \text{sex}_i \]

This is a saturated two-way ANOVA model.

4. Parameters of interest:

The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_1 + \beta_2 + \beta_3 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_2 \)
- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 0 \) is \( \beta_0 + \beta_1 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 0 \) is \( \beta_0 \)
- Effect of USP37 is \( \beta_1 + \beta_3 \) in females and \( \beta_1 \) in males

Parameters of interest: \( \beta_3 \) (interaction term)
Statistical modeling: Example #3

3. Regression model:

\[ expression_i \sim \beta_0 + \beta_1 \text{deceased}_i + \beta_2 \text{sex}_i + \beta_3 \text{deceased}_i \times \text{sex}_i \]

This is a saturated two-way ANOVA model.

4. Parameters of interest:
The regression model implies that

- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_1 + \beta_2 + \beta_3 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 1 \) is \( \beta_0 + \beta_2 \)
- \( expression_i \) when \( \text{deceased}_i = 1 \) and \( \text{sex}_i = 0 \) is \( \beta_0 + \beta_1 \)
- \( expression_i \) when \( \text{deceased}_i = 0 \) and \( \text{sex}_i = 0 \) is \( \beta_0 \)
- Effect of USP37 is \( \beta_1 + \beta_3 \) in females and \( \beta_1 \) in males

Parameters of interest: \( \beta_3 \) (interaction term)

5. Null hypothesis: \( H_0 : \beta_3 = 0 \)
Saturated two-way ANOVA

> fit = glm(X232033_at ~ deceased24 + sex + deceased24:sex, data = mm_tt2, family = "gaussian")
> summary(fit)

Call:
glm(formula = X232033_at ~ deceased24 + sex + deceased24:sex, family = "gaussian", data = mm_tt2)

Deviance Residuals:

Min 1Q Median 3Q Max
-4.0215 -0.4360 0.1417 0.5797 1.9133

Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.20353 0.06958 103.521 < 2e-16 ***
deceased24 -0.74414 0.17905 -4.156 4.12e-05 ***
sex -0.12729 0.10635 -1.197 0.232
deceased24:sex 1.21583 0.27277 4.457 1.14e-05 ***
Statistical modeling : Example #3 : Lab

Does USP37’s association with multiple myeloma different in men vs. women at the 0.05 significance level?
Does USP37’s association with multiple myeloma different in men vs. women at the 0.05 significance level?

Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|)   |
|---------------------|----------|------------|---------|------------|
| (Intercept)         | 7.20353  | 0.06958    | 103.521 | < 2e-16 ***|
| deceased24          | -0.74414 | 0.17905    | -4.156  | 4.12e-05 ***|
| sex                 | -0.12729 | 0.10635    | -1.197  | 0.232      |
| deceased24:sex      | 1.21583  | 0.27277    | 4.457   | 1.14e-05 ***|

Yes.
Does USP37’s association with multiple myeloma different in men vs. women at the 0.05 significance level?

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | 7.20353  | 0.06958    | 103.521 | < 2e-16  *** |
| deceased24           | -0.74414 | 0.17905    | -4.156  | 4.12e-05 *** |
| sex                  | -0.12729 | 0.10635    | -1.197  | 0.232    |
| deceased24:sex       | 1.21583  | 0.27277    | 4.457   | 1.14e-05 *** |

Yes.

Why are deceased24 and sex significantly non-zero now, when they weren’t before?
Does USP37’s association with multiple myeloma different in men vs. women at the 0.05 significance level?

Coefficients:

| Term                | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 7.20353  | 0.06958    | 103.521 | < 2e-16  *** |
| deceased24          | -0.74414 | 0.17905    | -4.156  | 4.12e-05 *** |
| sex                 | -0.12729 | 0.10635    | -1.197  | 0.232    |
| deceased24:sex      | 1.21583  | 0.27277    | 4.457   | 1.14e-05 *** |

Yes.

Why are deceased24 and sex significantly non-zero now, when they weren’t before? $\beta_1$ and $\beta_2$ have different interpretations in this model, as compared to the previous models.
1. Biological question: Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?

2. Variables:
   - **(Independent)** deceased = 1 if subject was deceased by 24 months, 0 otherwise
   - **(Dependent)** expression = expression (usually log-expression) of USP37 in subject
   - **(Dependent)** sex = 1 if subject is female, 0 otherwise
   - **(Dependent)** age = age of subject in years

3. Regression model: deceased \(i\) \sim \beta_0 + \beta_1 \text{expression}_i + \beta_2 \text{sex}_i + \beta_3 \text{age}_i

   This a multiple logistic regression model.
Statistical modeling : Example #4

1. Biological question:

Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?
1. Biological question:

   Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?

2. Variables:
1. Biological question:

   *Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?*

2. Variables:

   ▶ (Independent) $deceased_i = 1$ if subject $i$ was deceased by 24 months, 0 otherwise

   ▶ (Dependent) $expression_i = $ expression (usually log-expression) of USP37 in subject $i$

   ▶ (Dependent) $sex_i = 1$ if subject $i$ is female, 0 otherwise

   ▶ (Dependent) $age_i = $ age of subject $i$ in years
1. Biological question:

*Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?*

2. Variables:

- (Independent) $deceased_i = 1$ if subject $i$ was deceased by 24 months, 0 otherwise
- (Dependent) $expression_i =$ expression (usually log-expression) of USP37 in subject $i$
- (Dependent) $sex_i = 1$ if subject $i$ is female, 0 otherwise
- (Dependent) $age_i =$ age of subject $i$ in years

3. Regression model:

$$deceased_i \sim \beta_0 + \beta_1 expression_i + \beta_2 sex_i + \beta_3 age_i$$

This is a multiple logistic regression model.
Statistical modeling : Example #4

1. Biological question:

   Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?

2. Variables:
   ▶ (Independent) $deceased_i = 1$ if subject $i$ was deceased by 24 months, 0 otherwise
   ▶ (Dependent) $expression_i = \text{expression (usually log-expression)}$ of USP37 in subject $i$
   ▶ (Dependent) $sex_i = 1$ if subject $i$ is female, 0 otherwise
   ▶ (Dependent) $age_i = \text{age of subject } i \text{ in years}$

3. Regression model:

   $$deceased_i \sim \beta_0 + \beta_1 expression_i + \beta_2 sex_i + \beta_3 age_i$$

   This a multiple logistic regression model.
4. Parameters of interest:
Statistical modeling: Example #4

4. Parameters of interest:
   1. $\beta_1$ quantifies association between expression and MM risk
   2. $\beta_2$ quantifies association between sex and MM risk
   3. $\beta_3$ quantifies association between age and MM risk
4. Parameters of interest:

1. $\beta_1$ quantifies association between expression and MM risk
2. $\beta_2$ quantifies association between sex and MM risk
3. $\beta_3$ quantifies association between age and MM risk

5. Null hypotheses:
4. Parameters of interest:
   1. $\beta_1$ quantifies association between expression and MM risk
   2. $\beta_2$ quantifies association between sex and MM risk
   3. $\beta_3$ quantifies association between age and MM risk

5. Null hypotheses:
   1. $H_0: \beta_1 = 0$
   2. $H_0: \beta_2 = 0$
   3. $H_0: \beta_3 = 0$
Multiple logistic regression

\[ \texttt{fit} = \texttt{glm(deceased24} \sim \texttt{X232033\_at + sex + age, data = mm\_tt2, family = \texttt{\"binomial\")} \]

Call: \texttt{glm(formula = deceased24} \sim \texttt{X232033\_at + sex + age, family = \texttt{\"binomial\", data = mm\_tt2)}

Coefficients:
\[
\begin{array}{cccc}
(\text{Intercept}) & \text{X232033\_at} & \text{sex} & \text{age} \\
-1.23742 & -0.27695 & 0.01491 & 0.02561 \\
\end{array}
\]

Degrees of Freedom: 335 Total (i.e. Null); 332 Residual
Null Deviance: 286.1
Residual Deviance: 281.3 AIC: 289.3
Multiple logistic regression

```r
> fit = glm(deceased24 ~ X232033_at + sex + age, data = mm_tt2, family = "binomial")
> fit

Call: glm(formula = deceased24 ~ X232033_at + sex + age, family = "binomial",
  data = mm_tt2)

Coefficients:
(Intercept)    X232033_at       sex       age
  -1.23742    -0.27695       0.01491   0.02561

Degrees of Freedom:  335 Total (i.e. Null); 332 Residual
Null Deviance: 286.1
Residual Deviance: 281.3  AIC: 289.3
```

- `family = "binomial"` tells R that the expression outcome is a discrete categorical data type with two categories
Statistical modeling : Example #4 : Lab

Multiple logistic regression

\[
\text{> summary(fit)}
\]

Call:
\[
\text{glm(formula = deceased24 ~ X232033_at + sex + age, family = "binomial", data = mm_tt2)}
\]

Deviance Residuals:
\[
\begin{array}{ccccc}
\text{Min} & \text{1Q} & \text{Median} & \text{3Q} & \text{Max} \\
-0.8323 & -0.6094 & -0.5387 & -0.4715 & 2.2929 \\
\end{array}
\]

Coefficients:
\[
\begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & \text{z value} & \text{Pr(>|z|)} \\
\text{(Intercept)} & -1.23742 & 1.33936 & -0.924 & 0.3555 \\
X232033_at & -0.27695 & 0.15652 & -1.769 & 0.0768 . \\
sex & 0.01491 & 0.30965 & 0.048 & 0.9616 \\
age & 0.02561 & 0.01653 & 1.549 & 0.1214 \\
\end{array}
\]
Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?
Are gender, age, and the expression of gene 232033_at (USP37) predictive of multiple myeloma risk?

Coefficients:

| Estimate  | Std. Error | z value | Pr(>|z|) |
|-----------|------------|---------|----------|
| (Intercept) | -1.23742   | 1.33936 | -0.924   | 0.3555  |
| X232033_at | -0.27695   | 0.15652 | -1.769   | 0.0768  |
| sex       | 0.01491    | 0.30965 | 0.048    | 0.9616  |
| age       | 0.02561    | 0.01653 | 1.549    | 0.1214  |

USP37 is moderately predictive, but not at the 0.05 significance level. Gender and age are not predictive.
Multiple testing: Problem

The more hypothesis tests performed, the more likely to see a significant p-value just by chance, even if the null is true.

```r
set.seed(1)
summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients

Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.036360  0.057031  0.63754  0.52420
mm_tt2$deceased24 0.076241  0.146385  0.52082  0.60283

set.seed(1)
summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients

Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.111496  0.063029 -1.7690  0.07781
mm_tt2$deceased24 0.331388  0.161780  2.0484  0.04130

set.seed(1)
summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients

Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.014264  0.063623 -0.2242  0.82274
mm_tt2$deceased24 -0.068756  0.163306 -0.4210  0.67401

set.seed(1)
summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients

Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.027547  0.060285 -0.4569  0.64800
mm_tt2$deceased24  0.214504  0.154736  1.3863  0.16659

set.seed(1)
summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients

Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.046413  0.059125 -0.7850  0.43301
mm_tt2$deceased24  0.111511  0.151760  0.7348  0.46299
```
Multiple testing : Problem

The more hypothesis tests performed, the more likely to see a significant \( p \)-value just by chance, even if the null is true.

### Five hypothesis tests

```r
> set.seed(1)
> summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients
   Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03635970  0.05703107 0.6375419  0.5242087
mm_tt2$deceased24 0.07624081  0.14638483 0.5208245  0.6028341
> summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients
   Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1114958  0.06302866 -1.768970  0.07781151
mm_tt2$deceased24 0.3313876  0.16177917  2.048394  0.04130271
> summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients
   Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01426440  0.06362332 -0.2242008  0.8227382
mm_tt2$deceased24 -0.06875554  0.16330552 -0.4210240  0.6740083
> summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients
   Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0275471  0.06028464 -0.4569505  0.6480036
mm_tt2$deceased24  0.2145044  0.15473595  1.3862607  0.1665920
> summary(glm(rnorm(336) ~ mm_tt2$deceased24, family = "gaussian"))$coefficients
   Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04641295  0.05912536 -0.7849922  0.4330146
mm_tt2$deceased24  0.11151148  0.15176037  0.7347866  0.4629850
```
Multiple testing: Corrections

We must account for the number of total number of hypothesis tests we perform.

Two ways:

1. Bonferroni correction: new significant level $= \alpha / \text{total \# of hypothesis tests}$
2. False discovery rate correction: `p.adjust()` function R
Multiple testing: Lab

Which genes are significantly associated with myeloma at the 0.05 significance level, after Bonferroni correction?

Testing all genes

```r
ps = rep(NA, 54675)
for(i in 1:54675){
    fit = glm(mm_tt2[,i+1] ~ mm_tt2$deceased24, family = "gaussian")
    ps[i] = summary(fit)$coefficients[2,4]
}
```
Multiple testing: Lab

Which genes are significantly associated with myeloma at the 0.05 significance level, after Bonferroni correction?

Testing all genes

```r
ps = rep(NA, 54675)
for(i in 1:54675){
    fit = glm(mm_tt2[,i+1] ~ mm_tt2$deceased24, family = "gaussian")
    ps[i] = summary(fit)$coefficients[2,4]
}

> names(ps) = colnames(mm_tt2)[2:54676]
> ps[which(ps <= 0.05/54675)]
X201590_x_at  X210427_x_at
3.769075e-07  2.665137e-07

Both of these microarray probes correspond to ANXA2 (annexin A2).
Questions?
sdzhao@illinois.edu