Biometry. Lecture 16

Alexey Shipunov

Minot State University

April 2, 2014

Outline

1. Questions and answers

2. Two-dimensional statistics
   - Regression

Outline

1. Questions and answers

2. Two-dimensional statistics
   - Regression

Starting...

```
> setwd("<working folder>")
```

or

“`Change dir`”

in menu!

Two-dimensional statistics

Regression

The idea of regression

- The basic regression analysis will apply the linear model for data
- It will study not only if variables are associates and not only the strength of association but also the form of association (law of association)

Regression formula

- The simplest is a linear regression,

  $$m = b_0 + b_1 \times x,$$

  where m is a **predicted value**, x is an **independent variable** and b_0 and b_1 are coefficients (so-called **intersect** and **slope**).
- In other terms, linear regression is

  ```
  response = intersect + slope * influence
  ```
- In R model formula language, it is simply

  ```
  response ~ influence
  ```

Analysis of regression model

- If y is a real response, then error of model

$$E = y - m$$

- If σ^2 are dispersions of m and y, then

$$R^2 = 1 - \sigma_m^2/\sigma_y^2,$$

- In a background, R^2 is similar to coefficient of determination

Test of regression

- To test if regression model is correct, the Fisher test is normally applied
- Null hypothesis for Fisher test is that a model is not reliable

Regression example: women data

```
> lm.women <- lm(weight ~ height, data = women)
> plot(weight ~ height, data = women, main="",
+ xlab="Height (feet)", ylab="Weight (pounds)")
> grid()
> abline(lm.women, col="red")
```

Analysis of regression

```
> summary(lm.women)
```

Analysis of analysis

- Resulted model: `weight = -87.51667 + 3.45 * height`
- Maximum deviations from model are -1.7333 and 3.1167 pounds
- Almost half of residuals are between first and third quartiles
- All coefficients are significant
- Adjusted R-squared is close to 1 (very high!)
- The overall p-value is much less than 0.05 therefore the model is reliable
- There are 1 and 13 degrees of freedom (for columns and for rows)

Thuesen data example

- 24 rows and 2 columns data for observations of ventricular velocity with different levels of blood glucose
- Data was taken from patients with diabetes type I.

Running the example and explaining results

```
> install.packages("ISwR")
> library(ISwR)
> str(thuesen); head(thuesen)
> thuesen <- na.omit(thuesen)
> thuesen.lm <- lm(short.velocity ~ blood.glucose,
+ data=thuesen)
> thuesen.lm
> summary(thuesen.lm)
```

Scatterplot with regression line

```
> plot(short.velocity ~ blood.glucose, data=thuesen)
> abline(thuesen.lm)
```

Visualizing residuals

```
> with(thuesen, segments(blood.glucose,
+ fitted(thuesen.lm), blood.glucose, short.velocity))
```

Confidence intervals for regression

```
> pred.frame <- data.frame(blood.glucose=4:20)
> pc <- predict(thuesen.lm, int="c", newdata=pred.frame)
> plot(short.velocity ~ blood.glucose, data=thuesen)
> pred.gluc <- pred.frame$blood.glucose
> matlines(pred.gluc, pc, lty=c(1,2,2), col="black")
```

Diagnostic plots for regression

- “Residuals vs. Fitted”: checks outliers, the best is flat line
- “Normal Q-Q”: checks residuals for normal distribution, if they are not normal then our regression is not linear
- “Scale-Location”: checks the trend in dispersion
- “Residuals vs. Leverage & Cook’s distance”: checks the most influential observations

Regression diagnostics

```
> plot(thuesen.lm)
> plot(lm(height ~ weight, data=women))
```

Finishing...

```
> savehistory("20140402.r")
```

Final question (3 points)

Final question (3 points)

In the embedded data `USArrests`,
there are numbers of murders and rapes per 100,000 for every state.
Are murders and rapes correlated? Is that correlation significant?

Summary: most important commands

- `lm()`—estimate the linear regression
- `predict()`—predict values with model

For Further Reading

A. Shipunov.
Biometry [Electronic resource].
2012—onwards.
Mode of access:
`http://ashipunov.info/shipunov/school/biol_240`

A. Shipunov, and others.
Visual statistics. Use R!
DMK Press, 2012. [Under translation from Russian.]