Biometry. Lecture 18

Alexey Shipunov

Minot State University

April 8, 2015

> setwd("<working folder>")
>   or
>   "Change dir"
>   in menu!

On Mac, be sure that startup option is working: getwd()
(getwd() checks if R is in working folder, dir() checks the folder
content)

M

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)
```

For help, use `?summary.lm` because `summary()` is generic.

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)

M

Diagnostic plots for regression

```
> plot(lm.women) # Hit "Enter" to change plot
```

- "Residuals vs. Fitted": checks **linearity**, 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, should be random
- "Residuals vs. Leverage & Cook's distance": checks the most influential observations
- Use ?plot.lm for help.

M

Blood 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")
```

Regression diagnostics

```
> plot(thuesen.lm)
```

Correlation training

See today's script

Regression training

See today's script

Homework!

For *Cochlearia* data, tell:

- Which two measurement characters are most correlated?
- Is this correlation reliable?
- Is the linear model based on these characters reliable?

Finishing...

Save your commands!

(savehistory(<todaysdate>.r) or File -> Save as... on Mac)

Summary: most important commands

- cor()—calculates correlation coefficients
- cor.test()—runs correlation tests
- lm()—estimates the linear regression, summary(lm()) and plot(lm()) are related *methods*.

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!
Ongoing translation from Russian.