NZSSN Courses: Introduction to R

Session 8 – Advance analysis

Statistical Consulting Centre

consulting@stat.auckland.ac.nz
The Department of Statistics
The University of Auckland

20 July, 2017

SCIENCE
DEPARTMENT OF STATISTICS

Linear regression

`lm(y~x)` is used for linear regression.

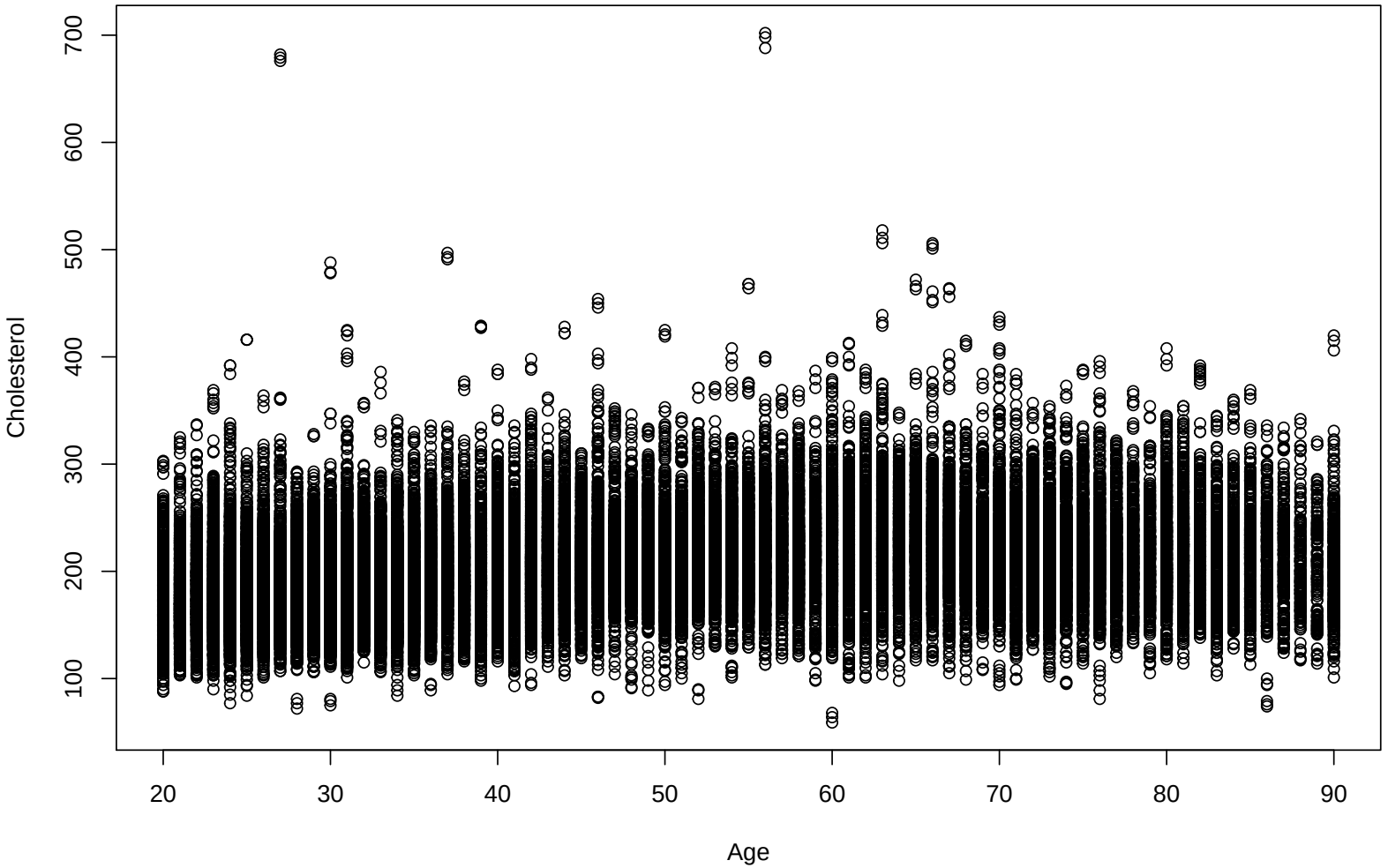
- `y`, the response variable.
- `x`, the explanatory variable.
- There can be more than one explanatory variable, called *multiple* linear regression.
- Both response variable and explanatory variable(s) should be numeric, it is *generalised* linear regression.

Simple linear regression

When there is only one predictor variable (e.g. Age) in our linear regression, we refer to this as *simple* linear regression.

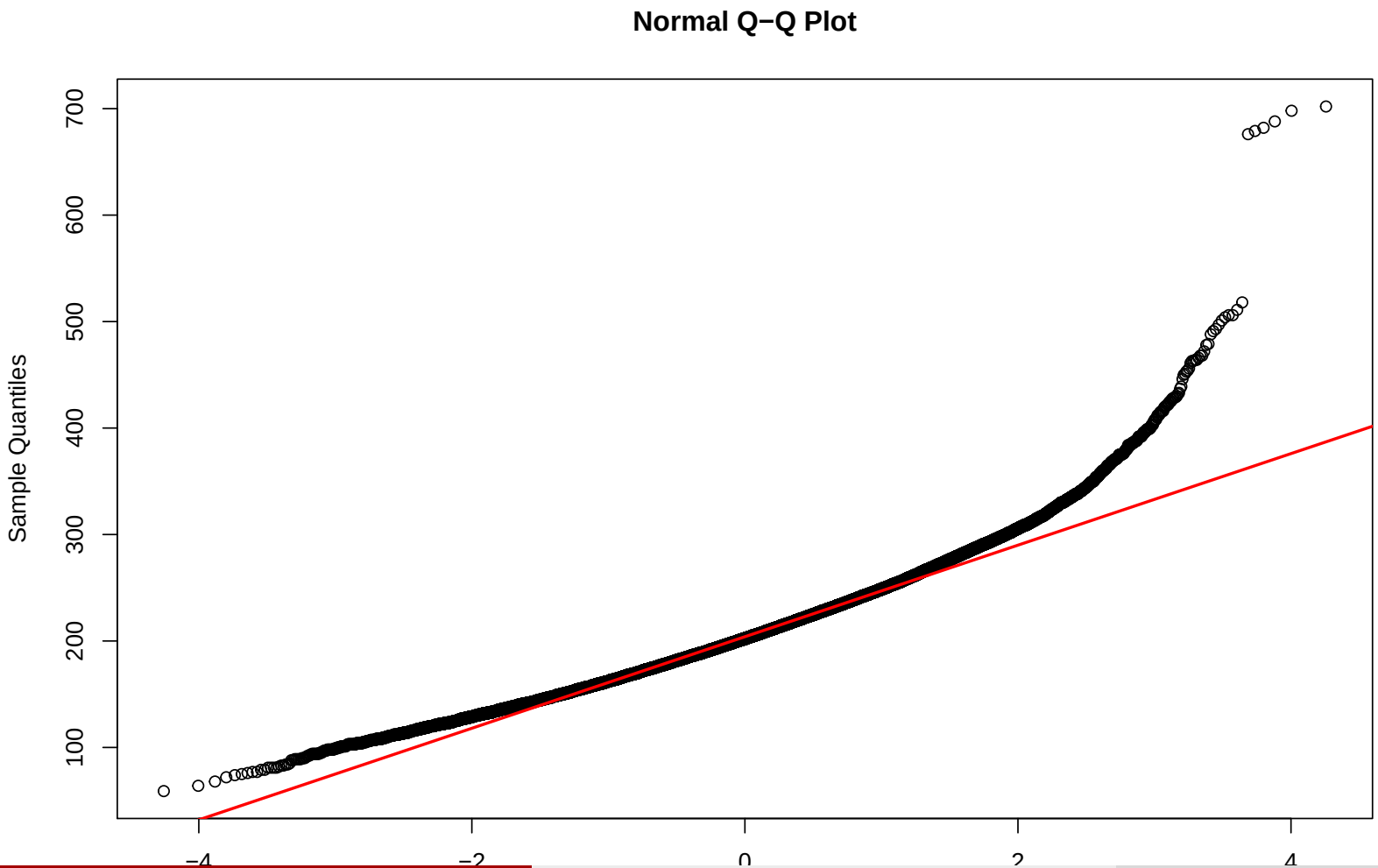
```
with(combined.long.df, plot(Age, Cholesterol))
```

Simple linear regression



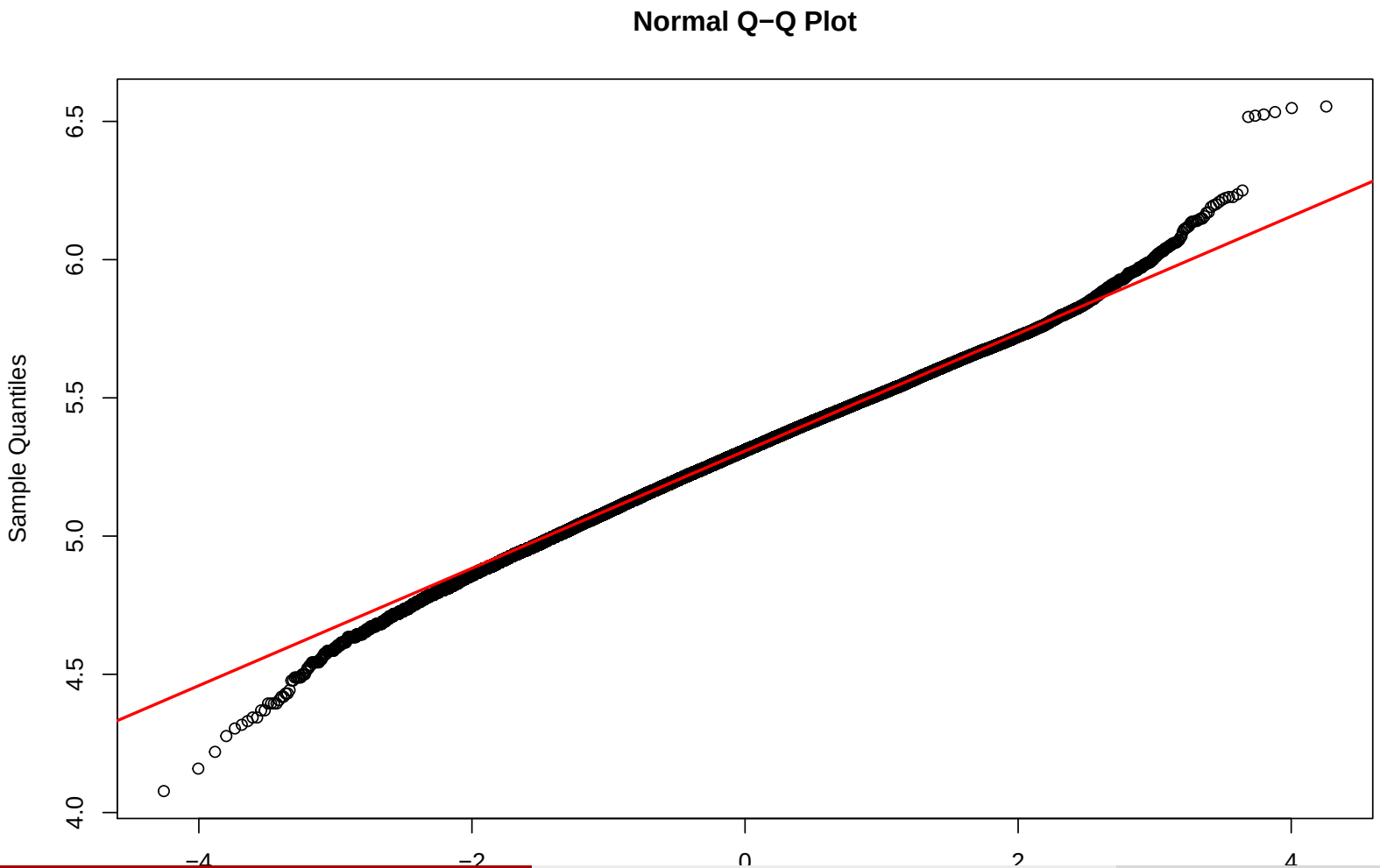
Normality check

```
qqnorm(combined.long.df$Cholesterol)
qqline(combined.long.df$Cholesterol, col = 2, lwd = 2)
```



Normality check

```
qqnorm(log(combined.long.df$Cholesterol))
qqline(log(combined.long.df$Cholesterol), col = 2, lwd = 2)
```



Simple linear regression

Let's carry out the linear regession of Age on Cholesterol level, i.e.

```
trylm <- with(combined.long.df, lm(log(Cholesterol)~Age))
summary(trylm)
```

Simple linear regression

```
##
## Call:
## lm(formula = log(Cholesterol) ~ Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26661 -0.12958  0.00596  0.13394  1.29653
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.134e+00  2.518e-03 2038.65   <2e-16 ***
## Age         3.504e-03  4.798e-05   73.04   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2065 on 48184 degrees of freedom
```

Simple linear regression

```
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)      5.1339     0.0025 2038.6540        0
## Age              0.0035     0.0000   73.0431        0
```

- The estimated intercept is 5.1339. There is a very strong evidecne that this is not zero (p-value < 0.0001).
- The estimated slope is 0.0035. There is a vereys strong evidence that this is not zero (p-value < 0.0001).
- The fitted line is:

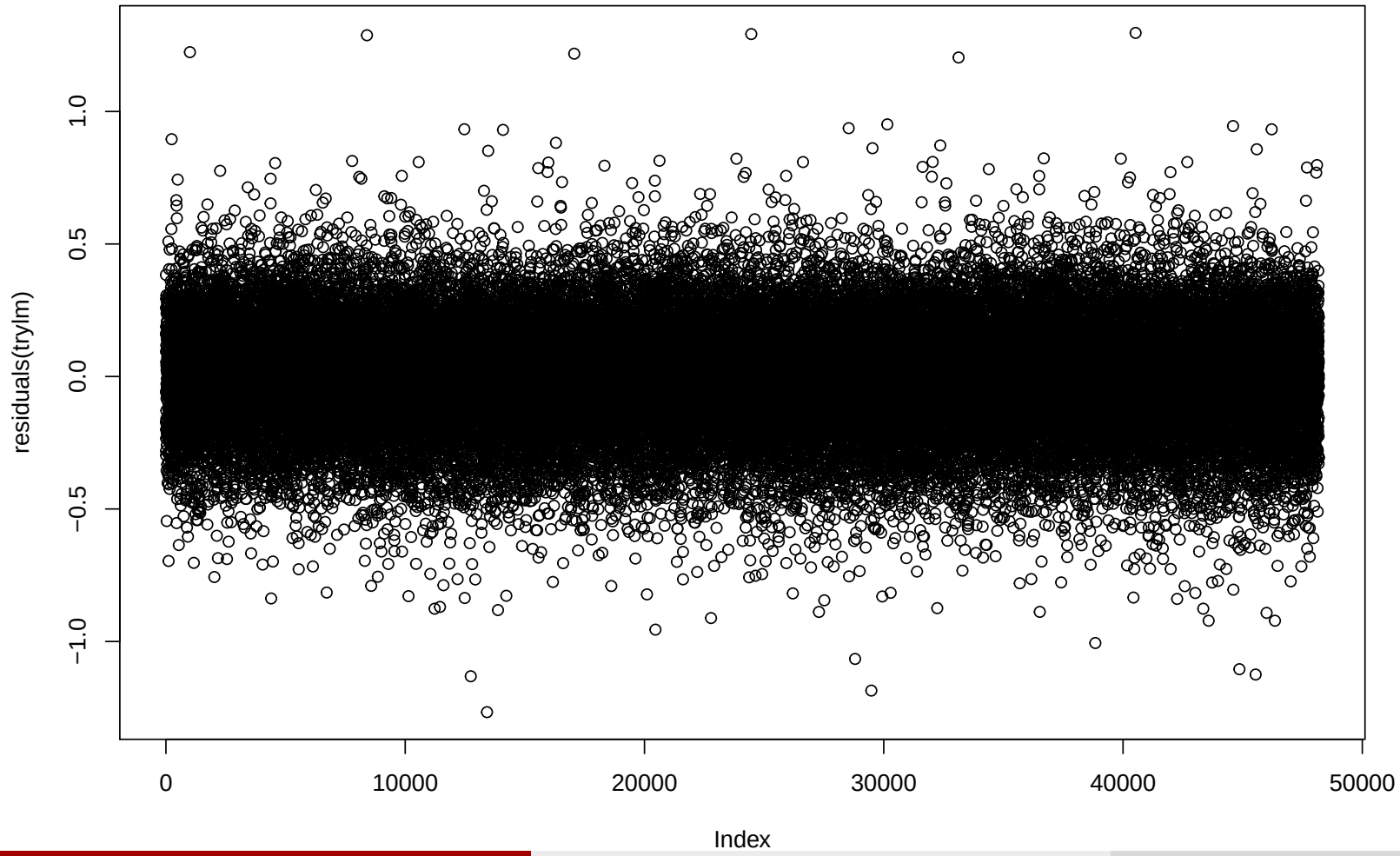
$$log(Cholesterol) = 5.1339 + 0.0035 \times Age$$

$$Cholesterol = e^{5.1339 + 0.0035 \times Age}$$

- For every one year increase in age, the Cholesterol level increase by 1.003506 times.

Check

```
plot(residuals(trylm))
```

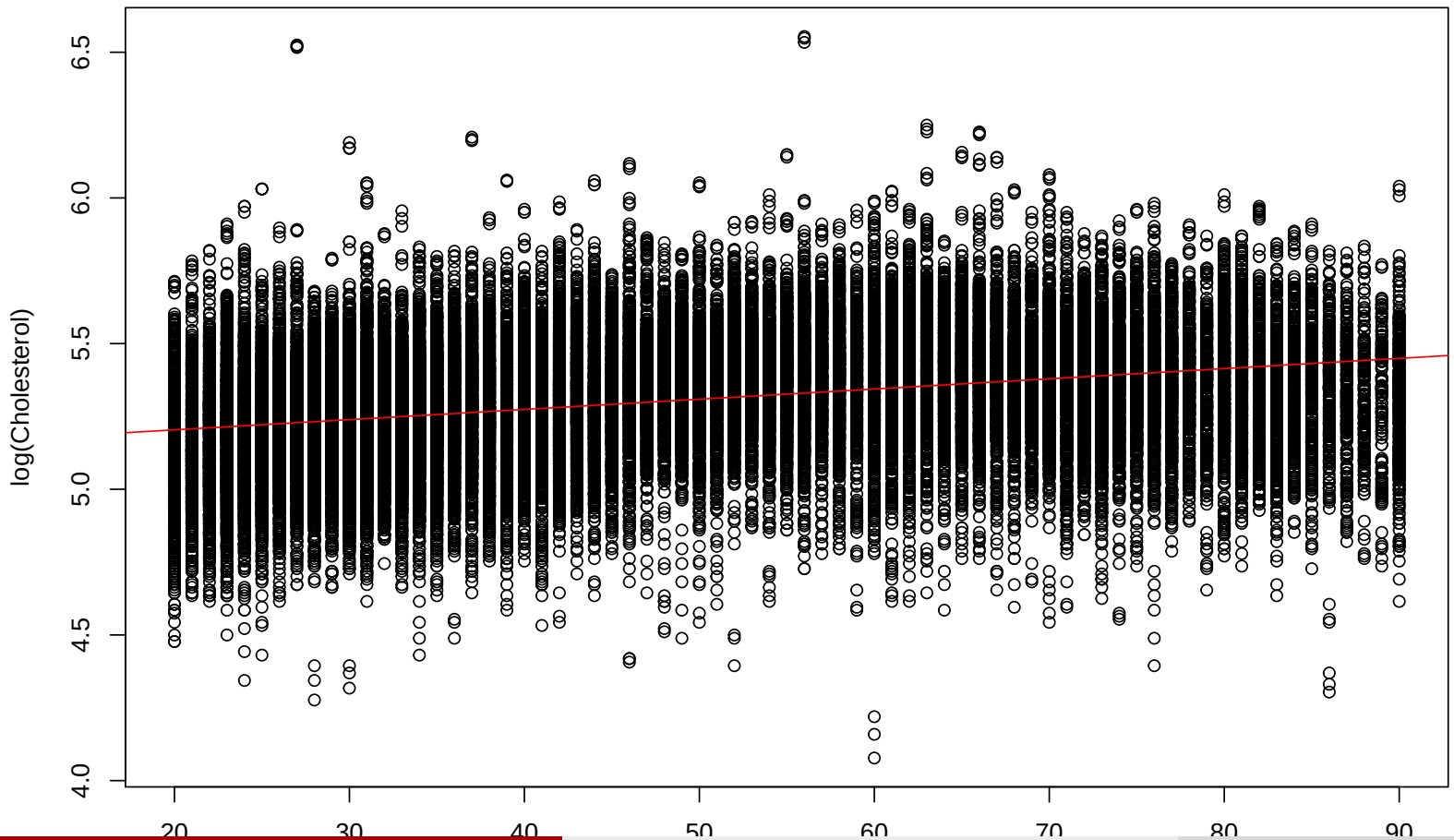


Conclusion

- The linear relationship between age and Cholesterol level is statistically significant.
- Cholesterol level is positive related to age.

Add the fitted line

```
with(combined.long.df, plot(Age, log(Cholesterol)))
abline(trylm, col = 2)
```



Quadratic term

```
tryquad <- with(combined.long.df,
        lm(log(Cholesterol) ~ Age + I(Age^2)))
round(summary(tryquad)$coef, 4)
```

```
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     4.8649     0.0067 728.3309        0
## Age             0.0156     0.0003  55.0753        0
## I(Age^2)       -0.0001     0.0000 -43.3394        0
```

`I(Age^2)` tells R to treat `^` as arithmetical operator, rather than formula operator. Our fitted curve is:

$$log(Cholesterol) = 4.8649 + 0.0035 \times Age + -0.0001 Age^2$$

What if the response variable is *not* continuous?

So far, we have considered methods for analysing response variables measured on a continuous scale. Often, measurements are:

- *Counts* per unit time, e.g. number of hours worked in a working week.
- *Binary* responses, e.g. Gender.
- *Generalised* linear models: *Poisson* (counts) and *Logistic* (binary) regression
- *Today* logistic regression *only*

Logistic regression

- Relates a *binary response variable* to a continuous and/or categorical variable.
- Let's illustrate by example using `combine.df`.
- Does smoking habit depend on cholesterol level, height or weight?
- Does the probability of smoking increase with age?
- What characteristics do smoking patients have, i.e. how do smoking patients different from non smoking patients.

Speaking in statistical language, we have a binary response. A patient is either smoking or non-smoking, there is only two possible outcomes.

Logistic regression is typically used in case control studies. In this case, we can treat smoking as case and non-smoking as control.

Summary

- Linear regression
- Logistic Regression