NZSSN Courses: Introduction to R

Session 8 – Advanced analysis

Statistical Consulting Centre

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

2 March, 2017

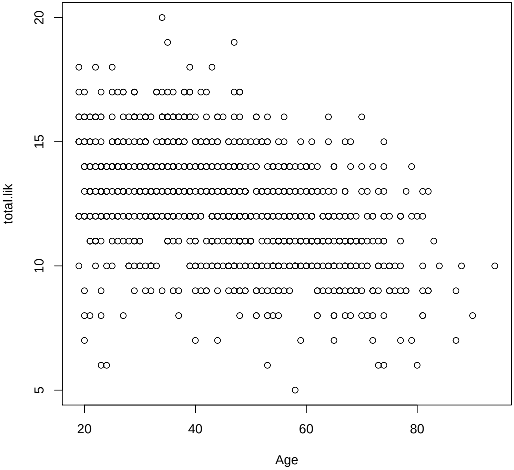
Linear regression

`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.



Simple linear regression

- The relationship between age and total score appears weakly negative, i.e. total score decreases with age.
- Let's carry out the linear regression of Age on total score, i.e.

```
try.lm <- with(issp.df, lm(total.lik~Age))
```

Simple linear regression

```
summary(try.lm)



Call:
lm(formula = total.lik ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7897 -1.2829  0.0273  1.3692  6.8813

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.192553   0.200219   75.88   <2e-16 ***
Age         -0.060995   0.004173  -14.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.082 on 952 degrees of freedom
  (93 observations deleted due to missingness)
Multiple R-squared:  0.1833,Adjusted R-squared:  0.1824
F-statistic: 213.6 on 1 and 952 DF,  p-value: < 2.2e-16
```

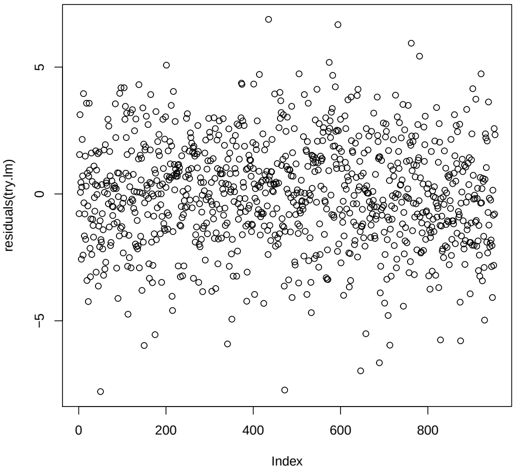
Simple linear regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.19255299	0.200218824	75.87974	0.000000e+00
Age	-0.06099486	0.004173331	-14.61539	8.528271e-44

- The estimated intercept is 15.19. There is very strong evidence that this is not zero (p-value < 0.0001).
- The estimated slope is -0.06. There is very strong evidence that this is not zero (p-value < 0.0001).
- The fitted line is `total.lik` = -0.06 × `Age` + 15.19
- For every one year increase in age, the mean total score decreases by 0.06 units on the likert scale.

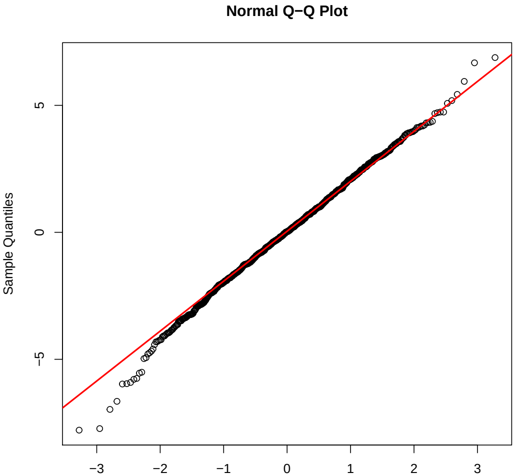
Check the fit: Residual plots

```
plot(residuals(try.lm))
```



Are the residuals approximately normal?

```
qqnorm(residuals(try.lm))
qqline(residuals(try.lm), col = 2, lwd = 2)
```

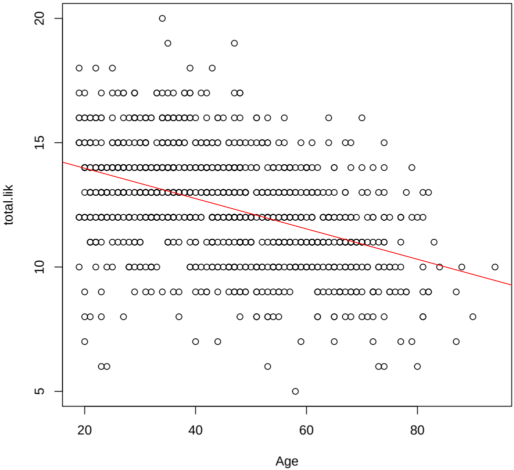


Conclusion

- The linear relationship between age and total score is statistically significant.
- Total score is negatively related to age.

Add the fitted line

```
with(issp.df, plot(Age, total.lik))
abline(try.lm, col = 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 `issp.df`.
  - Consider the variable `Q5` with values '`be obedient`' and '`think themselves`'.

Logistic regression

Question

Is `Age` a useful indicator of choosing '`being obedient`' as important in preparing for children for life?

How do we answer this?

By relating the *probability* of `being obedient` to `Age`.

- Linear regression is *not* suitable here because:
  - It assumes the response variable (`Q5`) takes values from $-\infty$ to $+\infty$.
  - But `Q5` takes only two values, namely `being obedient` or `think themselves`!

Relating a probability to an explanatory variable

Let:

- $p = \Pr(\texttt{Q5} = \texttt{being obedient})$
- $1 - p = \Pr(\texttt{Q5} = \texttt{think themselves})$

Definition: The *odds* that a respondent of `Q5` chooses `being obedient` is

$$\text{odds} = \frac{p}{1-p}.$$

- The *odds* of an event (i.e. $\texttt{Q5} = \texttt{being obedient}$) tells us how likely that event is to occur relative to it not occurring.
- To relate p to an explanatory variable, we need the **log-odds**, i.e.

$$\log\left(\frac{p}{1-p}\right) = \texttt{Intercept} + \texttt{Slope} \times \texttt{Age}.$$

- $\log\left(\frac{p}{1-p}\right)$ is known as the *logit* transformation

GLMs in R: `glm()`

```
glm(formula, family, ...)
```

- `formula`: Similar format as `lm()`; response variable and explanatory variable(s) separated by `~`.
- `family`: Use `family = binomial` for logistic regression.
- `...` See the help file of `glm` (`?glm`) for other arguments.

Logistic regression: Example

Suppose we want to find out whether older people are more likely to consider *being obedient* as more important in preparing children for life than is *thinking for themselves*.

Statisically speaking, we want to test whether the probability of choosing "`be obedient`" in `Q5` increases/decreases/does not change with `Age`.

Logistic regression: Example

- Declare the response variable `Q5` as a integer/numeric.
- `be obedient` is assigned the numeric value 1 and `think themselves` is assigned numeric value 0.
- It follows, therefore, that:
  1. $p = \Pr(\texttt{be obedient})$
  2. $p/(1-p)$ is the odds of participants selecting "being obedient" relative to selecting "thinking for themselves" as being important in preparing children for life.

Note: Here, the explanatory variable Age is integer/numeric.

GLMs in R: glm()

```r
## class of Q5?
class(issp.df$Q5)

[1] "character"

## Convert Q5 to a variable of type 'numeric'
issp.df$Q5 <- ifelse(issp.df$Q5 == "be obedient", 1, 0)

## Numeric values of Q5?
class(issp.df$Q5)

[1] "numeric"
```

Logistic regression: Example

Fit the model with `glm()`

```
try.glm <- with(issp.df, glm(Q5~Age,
                             family = binomial))
```

- `family = binomial`, logistic regression.

```
summary(try.glm)

Call:
glm(formula = Q5 ~ Age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2024  -0.7108  -0.5462  -0.4299   2.2182

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.132552   0.278107 -11.264  < 2e-16 ***
Age          0.036263   0.005163   7.024 2.16e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Logistic regression: Example

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.13255166	0.278106937	-11.263839	1.979445e-29
Age	0.03626334	0.005162854	7.023894	2.157690e-12

$$
\begin{aligned}
\text{logit(be obedient)} &= -3.13 + 0.04 \times \text{Age}. \\
\text{Odds(be obedient)} &= e^{-3.13+0.04\times\text{Age}} \\
\text{Probability(be obedient)} &= \frac{e^{-3.13+0.04\times\text{Age}}}{1 + e^{-3.13+0.04\times\text{Age}}}
\end{aligned}
$$

Prediction from the model

```
# Logit scale, usually referred to as the
# 'linear predictor' scale
lp <- predict(try.glm, data.frame(Age = 50))
lp

        1
-1.319385

# Calculate the odds
exp(lp)

        1
0.2672997
```

Interpretation: A 50-year old is **0.3 times** likely to consider *being obedient* important preparation for life than *thinking for oneself*. Or, a 50-year old is **3.7 times more** likely to *thinking for oneself* than *being obedient*.

Prediction from the model

```
#Probability scale
predict(try.glm, data.frame(Age = 50), type = "response")

        1
0.2109207
```

Interpretation: The probability that a 50-year old considers *being obedient* important preparation for life is 0.2109207.

Putting prediction into context

```
#Probability with standard error
predict(try.glm, data.frame(Age = 50),
type = "response", se.fit = TRUE)

$fit
        1
0.2109207

$se.fit
         1
0.01409851

$residual.scale
[1] 1
```

Categorical explanatory variables

```
try.glm2 <- with(issp.df, glm(Q5~age.group, family = binomial))
anova(try.glm2, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: Q5

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        920     929.45
age.group  2   46.725       918     882.73 7.143e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis of deviance table for a `glm()` object (generated using `anova()`) is analogous to ANOVA table for an `lm()` object.

Categorical explanatory variables

```
anova(try.glm2, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: Q5

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        920     929.45
age.group  2   46.725       918     882.73 7.143e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have extremely strong evidence that at least one age group has different log-odds of choosing "be obedient" from the other age groups.

Categorical explanatory variables

```
summary(try.glm2)



Call:
glm(formula = Q5 ~ age.group, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9790  -0.6169  -0.6169  -0.5233   2.0279

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.9192     0.1737 -11.048  < 2e-16 ***
age.group36 to 60   0.3568     0.2157   1.654    0.098 .
age.groupOver 61    1.4327     0.2274   6.301 2.96e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 929.45  on 920  degrees of freedom
Residual deviance: 882.73  on 918  degrees of freedom
  (126 observations deleted due to missingness)
AIC: 888.73
```

Categorical explanatory variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9192419	0.1737143	-11.048269	2.234812e-28
age.group36 to 60	0.3568389	0.2156919	1.654392	9.804798e-02
age.groupOver 61	1.4327090	0.2273911	6.300639	2.964214e-10

- `(Intercept)` corresponds to the reference age group, namely "Under 35" which is the one that is *not* listed!
- So, all subsequent rows of this table are hypothesis tests of the log-odds of the named age group relative to the reference group being zero, i.e.
  1. There is extremely strong evidence (p-value $<< 0.0001$) that the log-odds of choosing *being obedient* for the "Over 61" age group is higher than "Under 35" (p-value < 0.0001).
  2. There is no evidence (p-value $= 0.098$) that the log-odds of choosing *being obedient* for "36 to 60" is different from "Under 35" .

Compare "Over 61" with "36 to 60"

- Create another factor for age group with different reference level.

```
age.refac <- factor(as.character(issp.df$age.group),
   levels = c("Over 61", "36 to 60", "Under 35"))
```

- Re-fit the model.

```
try.glm3 <- glm(Q5~age.refac, family = binomial,
                data=issp.df)
```

Compare "Over 61" with "36 to 60"

```r
summary(try.glm3)
```

```
                    Estimate Std. Error   z value     Pr(>|z|)
(Intercept)       -0.4865329  0.1467312 -3.315810 9.137782e-04
age.refac36 to 60 -1.0758700  0.1946187 -5.528093 3.237314e-08
age.refacUnder 35 -1.4327090  0.2273911 -6.300639 2.964214e-10
```

There is extremely strong evidence that the log-odds of choosing *being obedient* for the "36 to 60" age group is *lower* than the "Over 61" age group.

Summary

- Linear regression
- Logistic Regression