Stat 306:
Finding Relationships in Data.
Lecture 10
Section 3.9 – Categorical explanatory variables

Age vs. Money

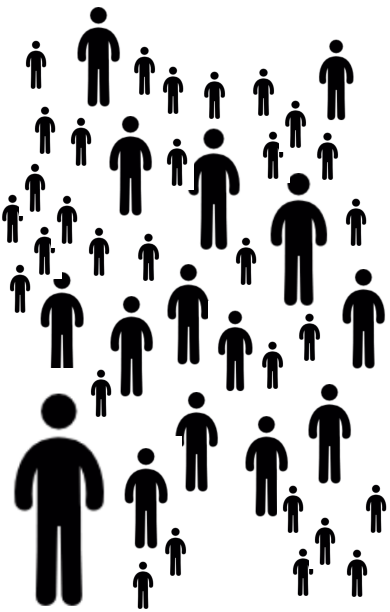
Dependent variable

χ → old (0)
χ → young (1)

Independent variable

γ → Cash ($)
On hand

Population



Population parameters

μ_0 , μ_1 , σ^2

Hypothesis Test

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Sample statistics

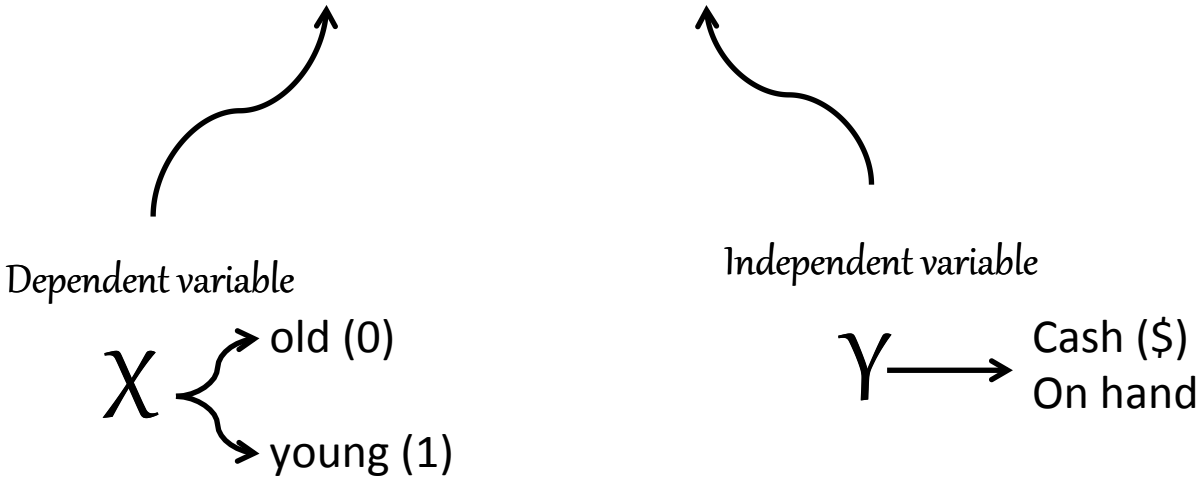
$\bar{y}_0 = 56$
$\bar{y}_1 = 27$
$\bar{y}_0 - \bar{y}_1 = 29$
$s_p = 10.81$

t = 2.68, df = 7
p-value = 0.03
95% C.I. = [3.4, 54.6]

Sample, n=9

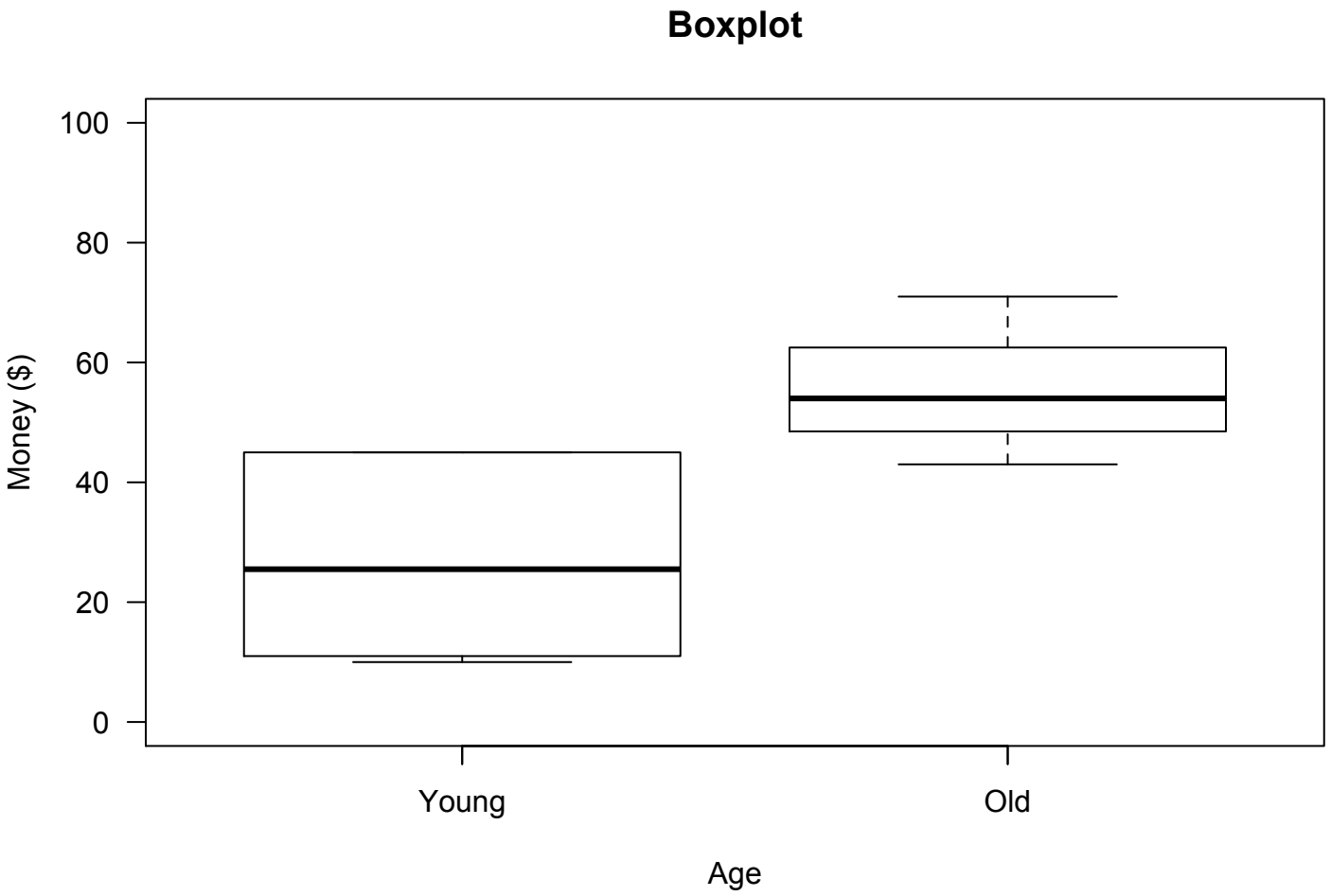
χ	y
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

Age vs. Money

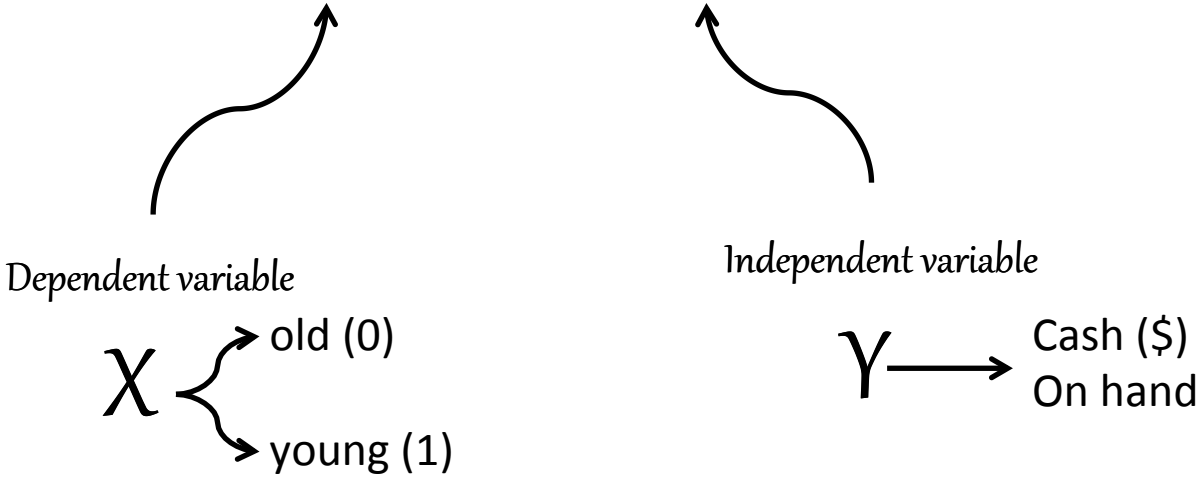



```
> old=c(71,54,43); young=c(45,21,11,30,45,10)
> t.test(x=x2, y=x1, var.equal=TRUE)

    Two Sample t-test

data:  x2 and x1
t = 2.6827, df = 7, p-value = 0.03142
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.438298 54.561702
sample estimates:
mean of x mean of y
       56        27
```

Age vs. Money



Age vs. Money

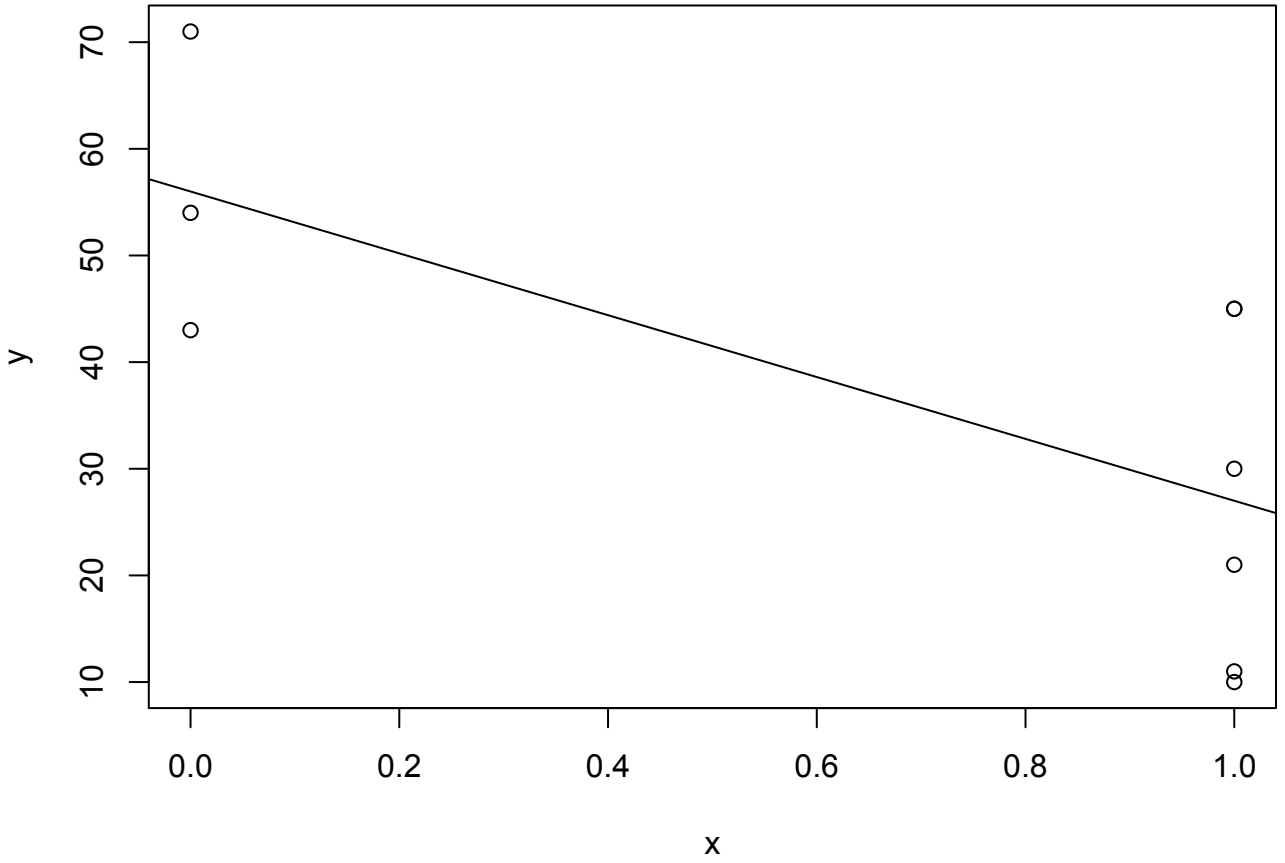



```
> x <- c(0, 0, 0, 1, 1, 1, 1, 1, 1)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> summary(lm(y~x))
```

Age vs. Money

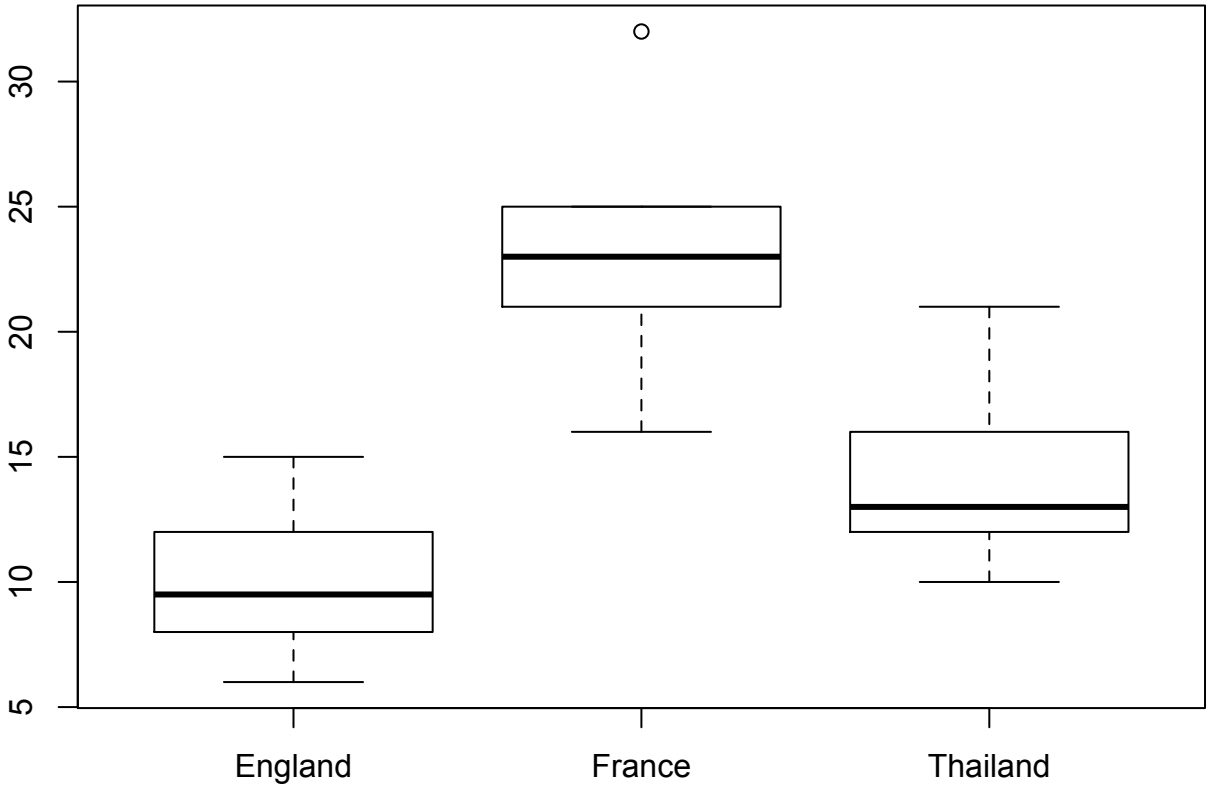
```
> x <- c(0, 0, 0, 1, 1, 1, 1, 1, 1)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
   -17    -13     -2     15     18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.000      8.826   6.345 0.000387 ***
x            -29.000     10.810  -2.683 0.031417 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.29 on 7 degrees of freedom
Multiple R-squared:  0.5069,Adjusted R-squared:  0.4365
F-statistic: 7.197 on 1 and 7 DF,  p-value: 0.03142
```

Age vs. Money



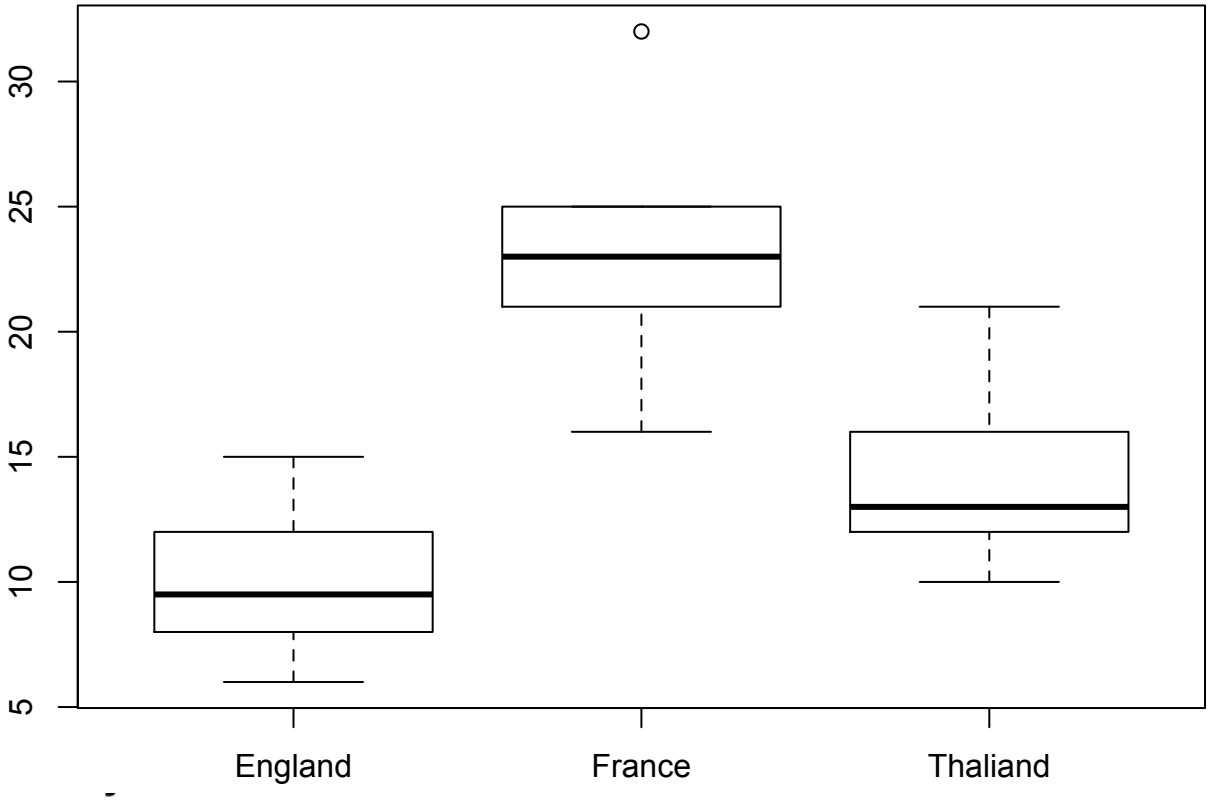

```
> plot(y~x)
> abline(lm(y~x))
```

3.9 Categorical explanatory variables




```
> country<-c(rep("France",6),rep("England",6),rep("Thailand",6))
> y<-c(23, 25, 21, 32, 16, 23, 15, 10,  8,  9,  6, 12, 13, 13, 12, 21, 16, 10)
> boxplot(y~country)
> 
```

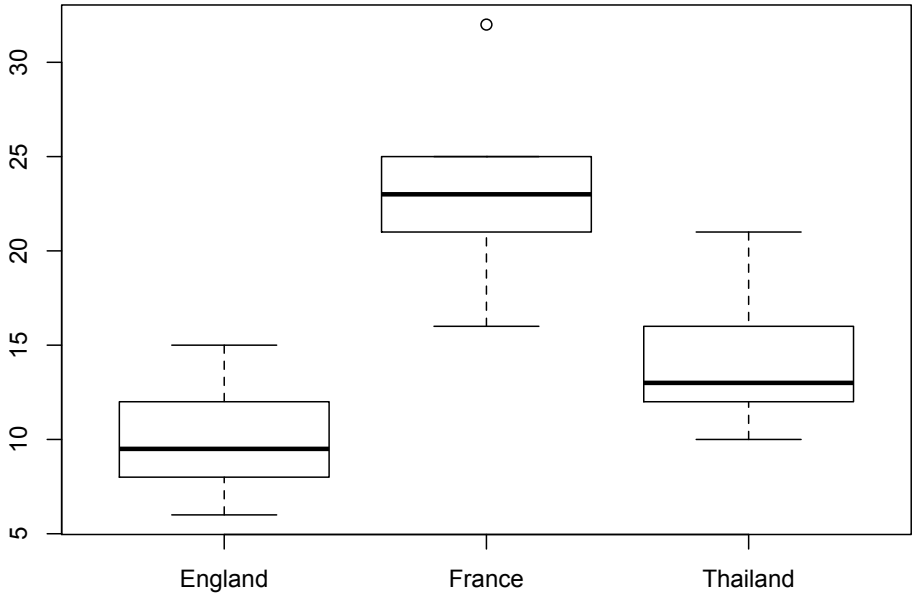
3.9 Categorical explanatory variables




```
> summary(aov(y ~ country, data = mydata))
            Df Sum Sq Mean Sq F value   Pr(>F)
country      2  558.3  279.17   15.97 0.000192 ***
Residuals   15  262.2   17.48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

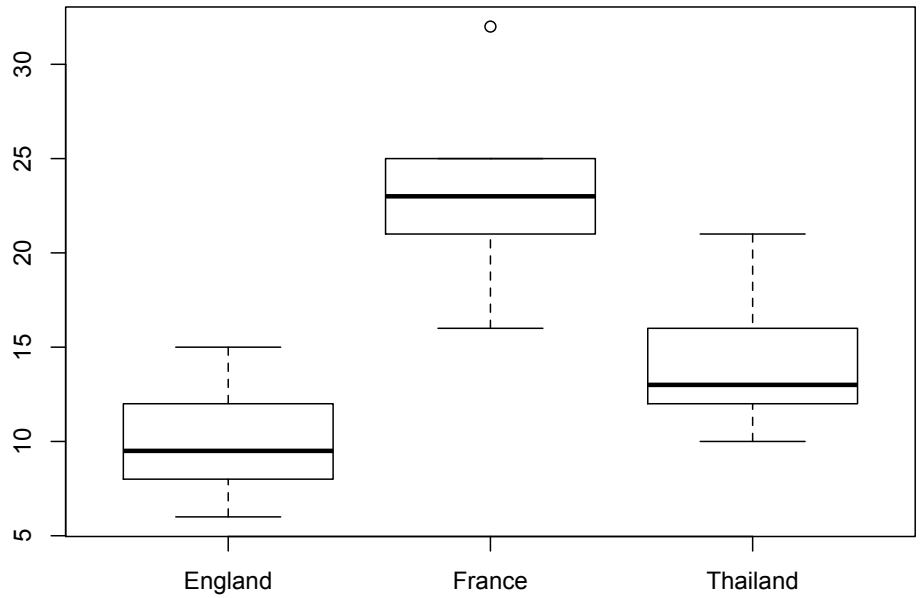
3.9 Categorical explanatory variables

```
> data.frame(y, country)
    y  country
1  23   France
2  25   France
3  21   France
4  32   France
5  16   France
6  23   France
7  15  England
8  10  England
9   8  England
10  9  England
11  6  England
12 12  England
13 13 Thailand
14 13 Thailand
15 12 Thailand
16 21 Thailand
17 16 Thailand
18 10 Thailand
```



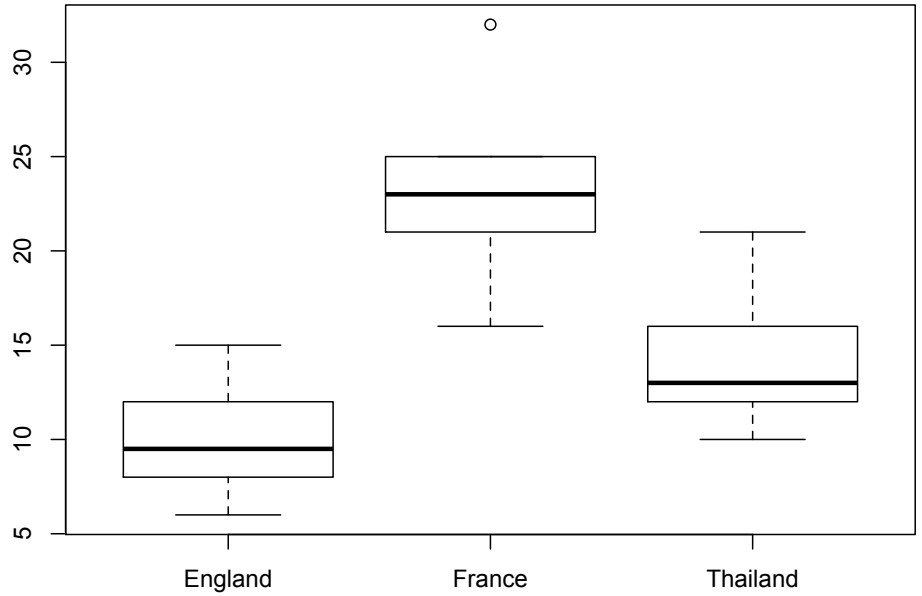
3.9 Categorical explanatory variables

```
> data.frame(y, country, x=as.numeric(as.factor(mydata$country))-1)
    y  country x
1  23   France 1
2  25   France 1
3  21   France 1
4  32   France 1
5  16   France 1
6  23   France 1
7  15  England 0
8  10  England 0
9   8  England 0
10  9  England 0
11  6  England 0
12 12  England 0
13 13 Thailand 2
14 13 Thailand 2
15 12 Thailand 2
16 21 Thailand 2
17 16 Thailand 2
18 10 Thailand 2
```



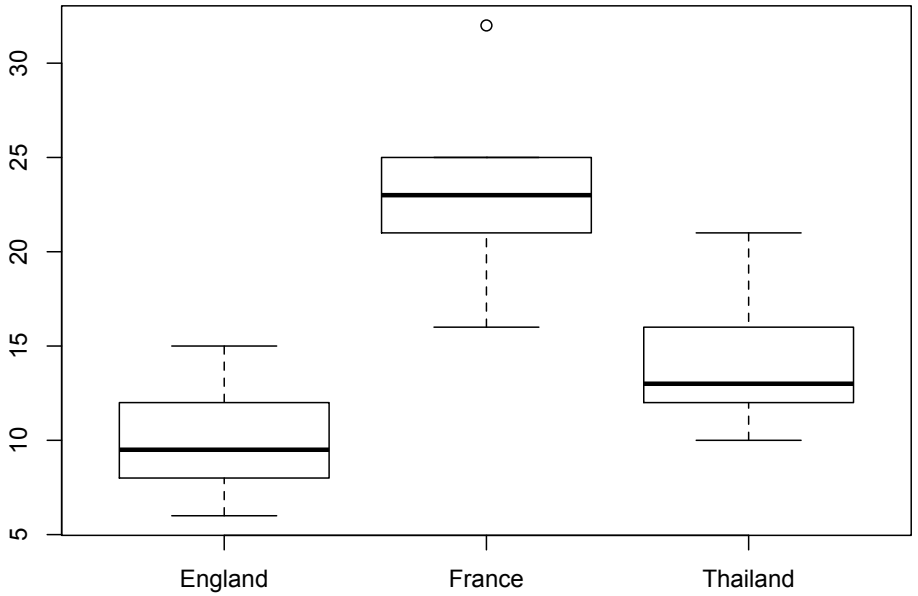
3.9 Categorical explanatory variables

```
> data.frame(y, country, x=as.numeric(as.factor(mydata$country))-1)
    y  country x
1  23   France 1
2  25   France 1
3  21   France 1
4  32   France 1
5  16   France 1
6  23   France 1
7  15   [illegible]
8  10   [illegible]
9   8   [illegible]
10  9   [illegible]
11  6   [illegible]
12 12   [illegible]
13 15   [illegible]
14 13 Thailand 2
15 12 Thailand 2
16 21 Thailand 2
17 16 Thailand 2
18 10 Thailand 2
```



3.9 Categorical explanatory variables

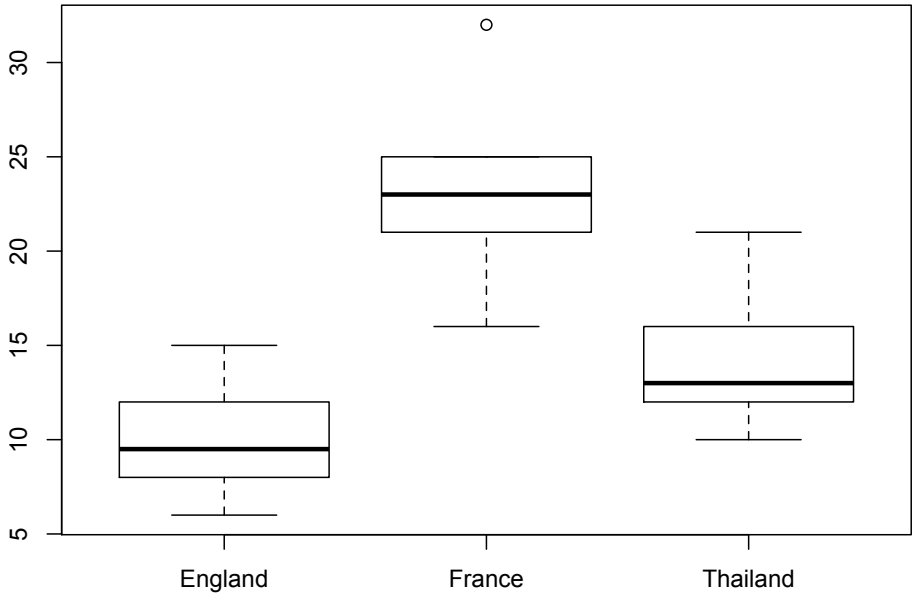
```
> data.frame(y, country, x1, x2)
    y  country x1 x2
1  23   France  1  0
2  25   France  1  0
3  21   France  1  0
4  32   France  1  0
5  16   France  1  0
6  23   France  1  0
7  15  England  0  0
8  10  England  0  0
9   8  England  0  0
10  9  England  0  0
11  6  England  0  0
12 12  England  0  0
13 13 Thailand  0  1
14 13 Thailand  0  1
15 12 Thailand  0  1
16 21 Thailand  0  1
17 16 Thailand  0  1
18 10 Thailand  0  1
```



3.9 Categorical explanatory variables

```
> mydata<-data.frame(y, country, x1, x2)
> summary(lm(y ~ x1 + x2, data=mydata))

Call:
lm(formula = y ~ x1 + x2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3333 -2.1250 -0.6667  1.7917  8.6667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.000      1.707   5.859 3.14e-05 ***
x1            13.333      2.414   5.524 5.84e-05 ***
x2             4.167      2.414   1.726    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.181 on 15 degrees of freedom
Multiple R-squared:  0.6805,   Adjusted R-squared:  0.6379
F-statistic: 15.97 on 2 and 15 DF,  p-value: 0.0001922
```

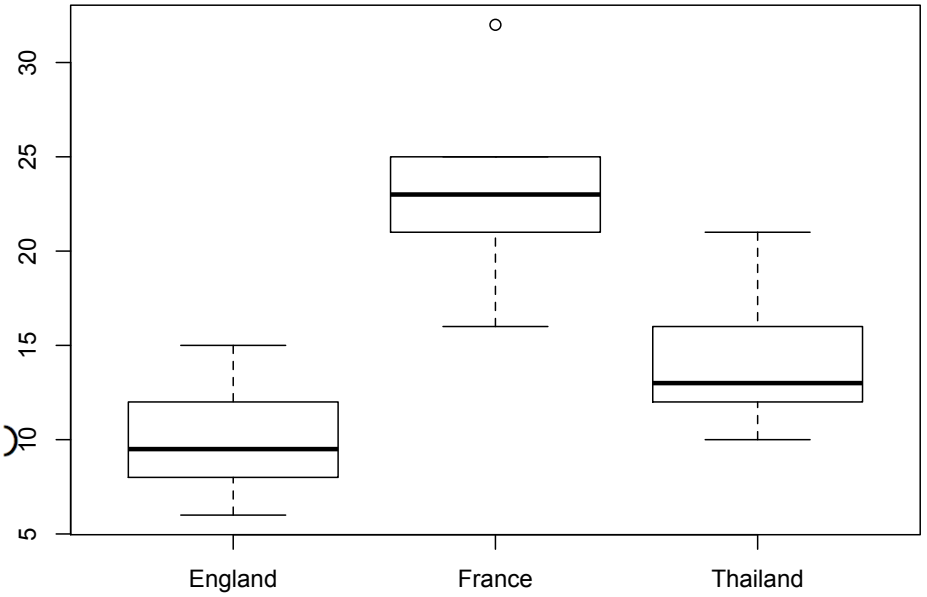
3.9 Categorical explanatory variables




```
> mydata<-data.frame(y, country, x1, x2)
> summary(lm(y ~ x1 + x2, data=mydata))

Call:
lm(formula = y ~ x1 + x2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3333 -2.1250 -0.6667  1.7917  8.6667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.000      1.707   5.859 3.14e-05 ***
x1            13.333      2.414   5.524 5.84e-05 ***
x2             4.167      2.414   1.726    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.181 on 15 degrees of freedom
Multiple R-squared:  0.6805,   Adjusted R-squared:  0.6379
F-statistic: 15.97 on 2 and 15 DF,  p-value: 0.0001922
```

3.9 Categorical explanatory variables

```
> summary(lm(y ~ as.factor(country), data=mydata))

Call:
lm(formula = y ~ as.factor(country), data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3333 -2.1250 -0.6667  1.7917  8.6667

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 10.000      1.707   5.859 3.14e-05 ***
as.factor(country)France    13.333      2.414   5.524 5.84e-05 ***
as.factor(country)Thailand   4.167      2.414   1.726    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.181 on 15 degrees of freedom
Multiple R-squared:  0.6805,   Adjusted R-squared:  0.6379
F-statistic: 15.97 on 2 and 15 DF,  p-value: 0.0001922
```



3.9 Categorical explanatory variables

How do we interpret this model?

```
> x3<-c(34, 39, 32, 44, 22, 39, 41, 33, 37, 37, 27, 36, 67, 65, 56, 68, 60, 59)
> x3
 [1] 34 39 32 44 22 39 41 33 37 37 27 36 67 65 56 68 60 59
> mydata<-data.frame(y, country, x1, x2, x3)
> summary(lm(y ~ x1 + x2 +x3, data=mydata))

Call:
lm(formula = y ~ x1 + x2 + x3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7297 -2.1410  0.1536  1.5329  3.7008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.0295     4.0969  -2.448 0.028149 *
x1           13.4283     1.4859   9.037 3.23e-07 ***
x2          -11.4013     3.4177  -3.336 0.004899 **
x3            0.5696     0.1126   5.058 0.000175 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.574 on 14 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8628
F-statistic: 36.63 on 3 and 14 DF,  p-value: 7.02e-07
```

3.9 Categorical explanatory variables

How do we make predictions from this model?

```
> x3<-c(34, 39, 32, 44, 22, 39, 41, 33, 37, 37, 27, 36, 67, 65, 56, 68, 60, 59)
> x3
 [1] 34 39 32 44 22 39 41 33 37 37 27 36 67 65 56 68 60 59
> mydata<-data.frame(y, country, x1, x2, x3)
> summary(lm(y ~ x1 + x2 +x3, data=mydata))

Call:
lm(formula = y ~ x1 + x2 + x3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7297 -2.1410  0.1536  1.5329  3.7008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.0295     4.0969  -2.448 0.028149 *
x1           13.4283     1.4859   9.037 3.23e-07 ***
x2          -11.4013     3.4177  -3.336 0.004899 **
x3            0.5696     0.1126   5.058 0.000175 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.574 on 14 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8628
F-statistic: 36.63 on 3 and 14 DF,  p-value: 7.02e-07
```

3.9 Categorical explanatory variables

3.9 Categorical explanatory variables

In this section, we indicate how to code categorical explanatory variables into binary dummy variables to use for multiple regression. A categorical variable with m categories ($m \geq 2$) is converted into $m-1$ binary variables, where one category is chosen as a baseline category and the $m-1$ binary variables represent differences of the other categories relative to the baseline. The selection of the baseline category is not unique.

To show the main ideas, first consider a multiple regression with one continuous and one categorical explanatory variable. If the categorical variable has two categories (e.g., female and male), define

$$z_i = \begin{cases} 1 & \text{if category 2 for } i\text{th case,} \\ 0 & \text{if category 1 for ith case.} \end{cases} \tag{3.109}$$

3.9 Categorical explanatory variables

3.9 Categorical explanatory variables

To show the main ideas, first consider a multiple regression with one continuous and one categorical explanatory variable. If the categorical variable has two categories (e.g., female and male), define

$$z_i = \begin{cases} 1 & \text{if category 2 for } i\text{th case,} \\ 0 & \text{if category 1 for ith case.} \end{cases} \tag{3.109}$$

```
> # Consider a multiple regression with
> # one continuous and one categorical
> # explanatory variable.
>
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> n <- length(y)
>
> gender<- c("Male", "Female", "Female", "Male", "Female", "Female",
"Male", "Male", "Female")
> data.frame(y, x1, gender)
   y x1 gender
1 71 82   Male
2 54 45 Female
3 43 71 Female
4 45 22   Male
5 21 29 Female
6 11  9 Female
7 30 12   Male
8 45 18   Male
9 10 24 Female
```

3.9 Categorical explanatory variables

3.9 Categorical explanatory variables

To show the main ideas, first consider a multiple regression with one continuous and one categorical explanatory variable. If the categorical variable has two categories (e.g., female and male), define

(3.109)
$$z_i = \begin{cases} 1 & \text{if category 2 for } i\text{th case,} \\ 0 & \text{if category 1 for ith case.} \end{cases}$$

```
> # If the categorical variable has two categories (e.g., female and
male), define:
>
> z <- rep(NA,n)
> for(i in 1:n){
+  if(gender[i]=="Male") {z[i] <- 1}
+  if(gender[i]=="Female") {z[i] <- 0}
+ }
>
> z
[1] 1 0 0 1 0 0 1 1 0
>
```

3.9 Categorical explanatory variables

3.9 Categorical explanatory variables

In this case, category 1 is considered as the baseline category. The data are converted (y_i, x_i, z_i), $i = 1, \ldots, n$. The regression equation becomes $Y_i = \mu_Y(x_i, z_i) + \epsilon_i$, where

$$\mu_Y(x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i = \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ \beta_0 + \beta_1 x_i + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_i & \text{if category 2.} \end{cases} \tag{3.110}$$

This implies a model where the relation of y with x is linear for both categories and there is a common slope. So on a scatterplot, the data for the two categories should lie roughly on parallel lines. β_2 is interpreted as the separation distance of the two lines.

```
> # Female is considered as the baseline category
> linear_reg(y, cbind(1, x1, z))
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
     7.639      7.102  1.076        -9.739        25.018  0.323
x1   0.566      0.146  3.866         0.208         0.925  0.008
z   21.139      7.238  2.921         3.429        38.849  0.027

$SStable
  SS_Total  SS_Res  MS_Res sqrt.MS_Res.   R2 adjR2 Fstatistic
1     3318 697.196 116.199        10.78 0.79  0.72     11.277
  Ftest_pval
1      0.009

>
```

3.9 Categorical explanatory variables

3.9 Categorical explanatory variables

In this case, category 1 is considered as the baseline category. The data are converted (y_i, x_i, z_i), $i = 1, \ldots, n$. The regression equation becomes $Y_i = \mu_Y(x_i, z_i) + \epsilon_i$, where

$$\mu_Y(x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i = \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ \beta_0 + \beta_1 x_i + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_i & \text{if category 2.} \end{cases} \tag{3.110}$$

This implies a model where the relation of y with x is linear for both categories and there is a common slope. So on a scatterplot, the data for the two categories should lie roughly on parallel lines. β_2 is interpreted as the separation distance of the two lines.

```
> # Female is considered as the baseline category
> linear_reg(y, cbind(1, x1, z))
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
     7.639      7.102  1.076        -9.739        25.018  0.323
x1   0.566      0.146  3.866         0.208         0.925  0.008
z   21.139      7.238  2.921         3.429        38.849  0.027

$SStable
  SS_Total  SS_Res  MS_Res sqrt.MS_Res.   R2 adjR2 Fstatistic
1     3318 697.196 116.199        10.78 0.79  0.72     11.277
  Ftest_pval
1      0.009
```

3.9 Categorical explanatory variables

3.9 Categorical explanatory variables

In this case, category 1 is considered as the baseline category. The data are converted (y_i, x_i, z_i), $i = 1, \ldots, n$. The regression equation becomes $Y_i = \mu_Y(x_i, z_i) + \epsilon_i$, where

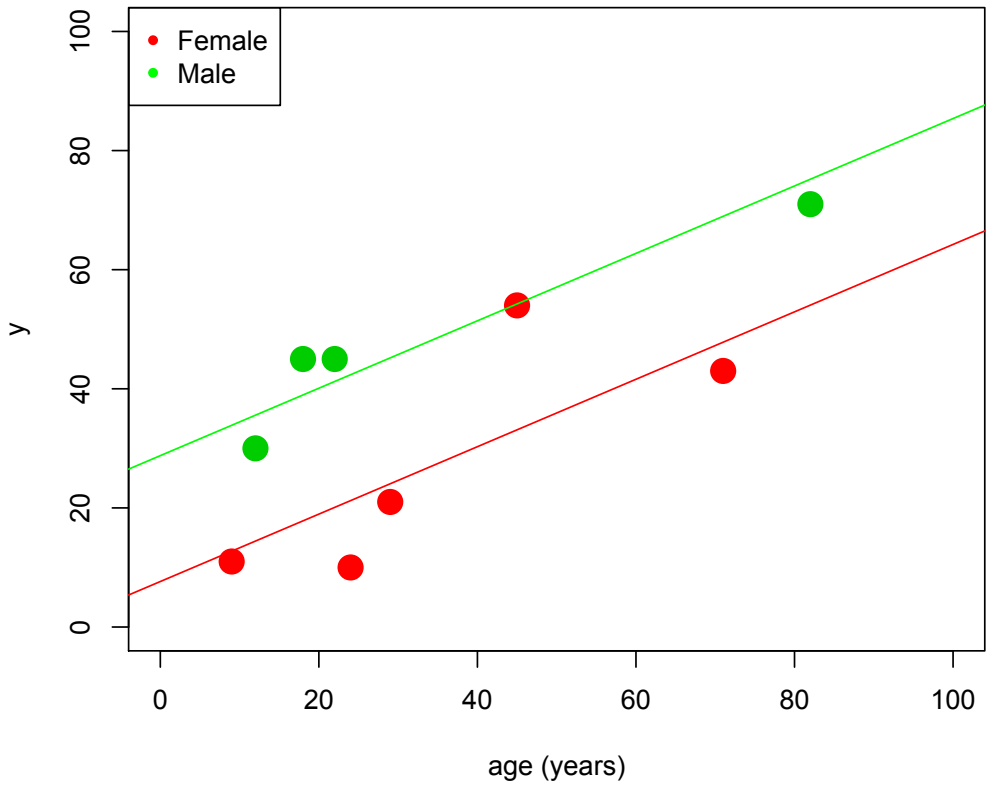
$$\mu_Y(x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i = \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ \beta_0 + \beta_1 x_i + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_i & \text{if category 2.} \end{cases} \tag{3.110}$$

This implies a model where the relation of y with x is linear for both categories and there is a common slope. So on a scatterplot, the data for the two categories should lie roughly on parallel lines. β_2 is interpreted as the separation distance of the two lines.

```
> plot(y~x1, ylim=c(0,100), xlim=c(0,100), xlab="age (years)", col=z+2, pch=20, cex=3)
> legend("topleft",c("Female", "Male"), pch=20, col=c("red", "green"))
> abline(7.639, 0.566, col="red")
> abline(7.639+21.139, 0.566, col="green")
>
```

3.9 Categorical explanatory variables

```
> plot(y~x1, ylim=c(0,100), xlim=c(0,100), xlab="age (years)", col=z+2, pch=20, cex=3)
> legend("topleft",c("Female", "Male"), pch=20, col=c("red", "green"))
> abline(7.639, 0.566, col="red")
> abline(7.639+21.139, 0.566, col="green")
>
```



3.9 Categorical explanatory variables

If the scatterplot shows linear relationships with different slopes for the two categories, then for multiple regression, use converted $(y_i, x_i, z_i, x_i z_i)$, $i = 1, \ldots, n$. The regression equation becomes $Y_i = \mu_Y^*(x_i, z_i) + \epsilon_i$, where

$$\mu_Y^*(x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i \tag{3.111}$$

$$= \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i & \text{if category 2.} \end{cases} \tag{3.112}$$

3.9 Categorical explanatory variables

If the scatterplot shows linear relationships with different slopes for the two categories, then for multiple regression, use converted $(y_i, x_i, z_i, x_i z_i)$, $i = 1, \ldots, n$. The regression equation becomes $Y_i = \mu_Y^*(x_i, z_i) + \epsilon_i$, where

$$\mu_Y^*(x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i \quad (3.111)$$

$$= \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i & \text{if category 2.} \end{cases} \quad (3.112)$$

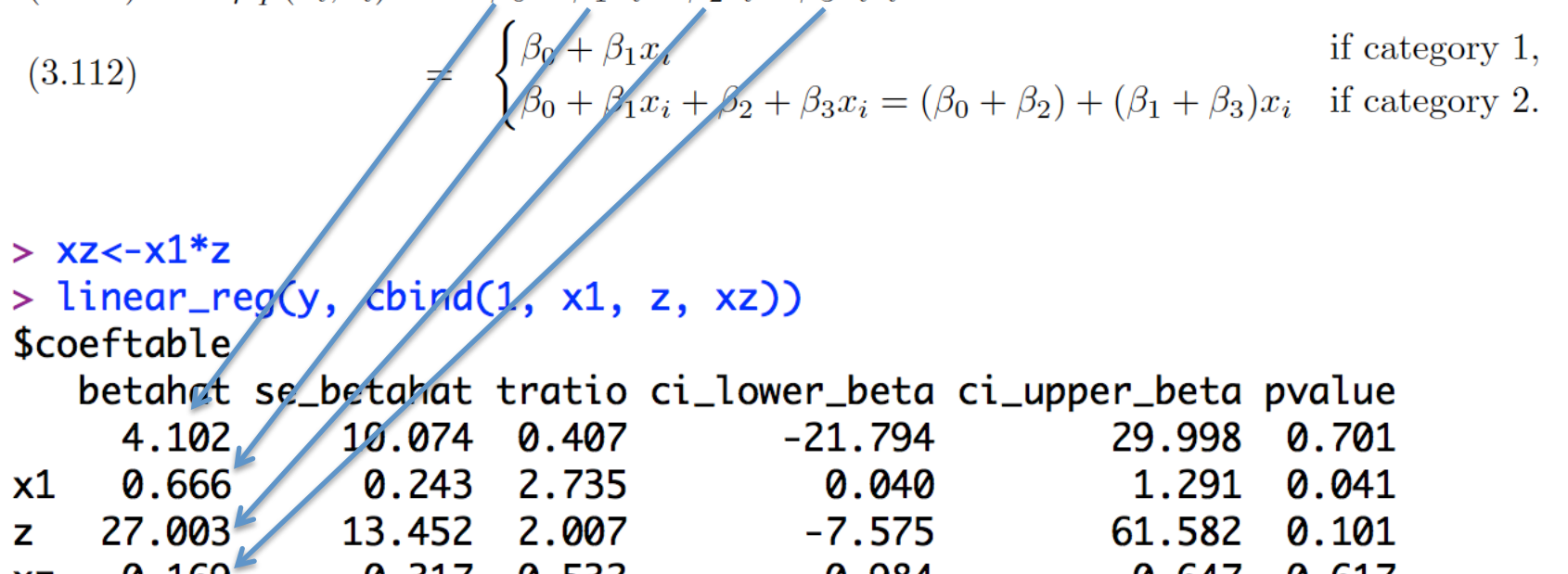
```
> xz<-x1*z
> linear_reg(y, cbind(1, x1, z, xz))
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
     4.102     10.074  0.407       -21.794        29.998  0.701
x1   0.666      0.243  2.735         0.040         1.291  0.041
z   27.003     13.452  2.007        -7.575        61.582  0.101
xz  -0.169      0.317 -0.532        -0.984         0.647  0.617

$SStable
  SS_Total  SS_Res  MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1     3318 659.833 131.967       11.488 0.801 0.682      6.714      0.033
```

3.9 Categorical explanatory variables

If the scatterplot shows linear relationships with different slopes for the two categories, then for multiple regression, use converted $(y_i, x_i, z_i, x_i z_i)$, $i = 1, \ldots, n$. The regression equation becomes $Y_i = \mu_Y^*(x_i, z_i) + \epsilon_i$, where

$$\mu_Y^*(x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i \tag{3.111}$$

$$= \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i & \text{if category 2.} \end{cases} \tag{3.112}$$



```
> xz<-x1*z
> linear_reg(y, cbind(1, x1, z, xz))
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
     4.102     10.074  0.407       -21.794        29.998  0.701
x1   0.666      0.243  2.735         0.040         1.291  0.041
z   27.003     13.452  2.007        -7.575        61.582  0.101
xz  -0.169      0.317 -0.532        -0.984         0.647  0.617

$SStable
  SS_Total  SS_Res  MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1     3318 659.833 131.967       11.488 0.801 0.682      6.714      0.033
```

3.9 Categorical explanatory variables

Hence β_3 is interpreted as the difference in slope for category 2 versus category 1. Is there a simple interpretation for β_2 in this case? The product $x_i z_i$ is an example of what is called an *interaction term* in multiple regression. Interaction terms involving products of other explanatory variables indicate that the two variables do not influence the mean value of the response in an additive manner.

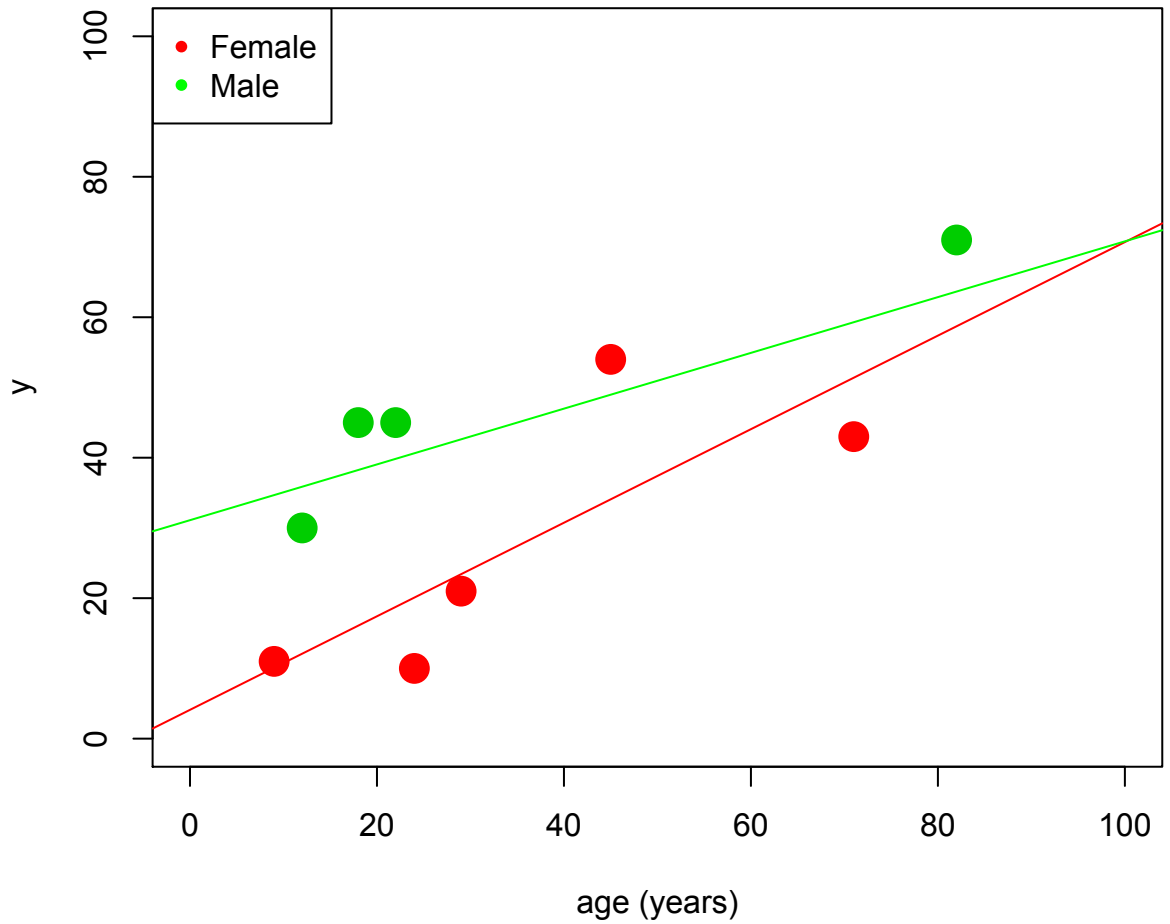
$$\mu_Y^*(x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i \tag{3.111}$$

$$= \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i & \text{if category 2.} \end{cases} \tag{3.112}$$

```
> xz<-x1*z
> linear_reg(y, cbind(1, x1, z, xz))
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
     4.102     10.074  0.407       -21.794        29.998  0.701
x1   0.666      0.243  2.735         0.040         1.291  0.041
z   27.003     13.452  2.007        -7.575        61.582  0.101
xz  -0.169      0.317 -0.532        -0.984         0.647  0.617

$SStable
  SS_Total  SS_Res  MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1     3318 659.833 131.967       11.488 0.801 0.682      6.714      0.033
```

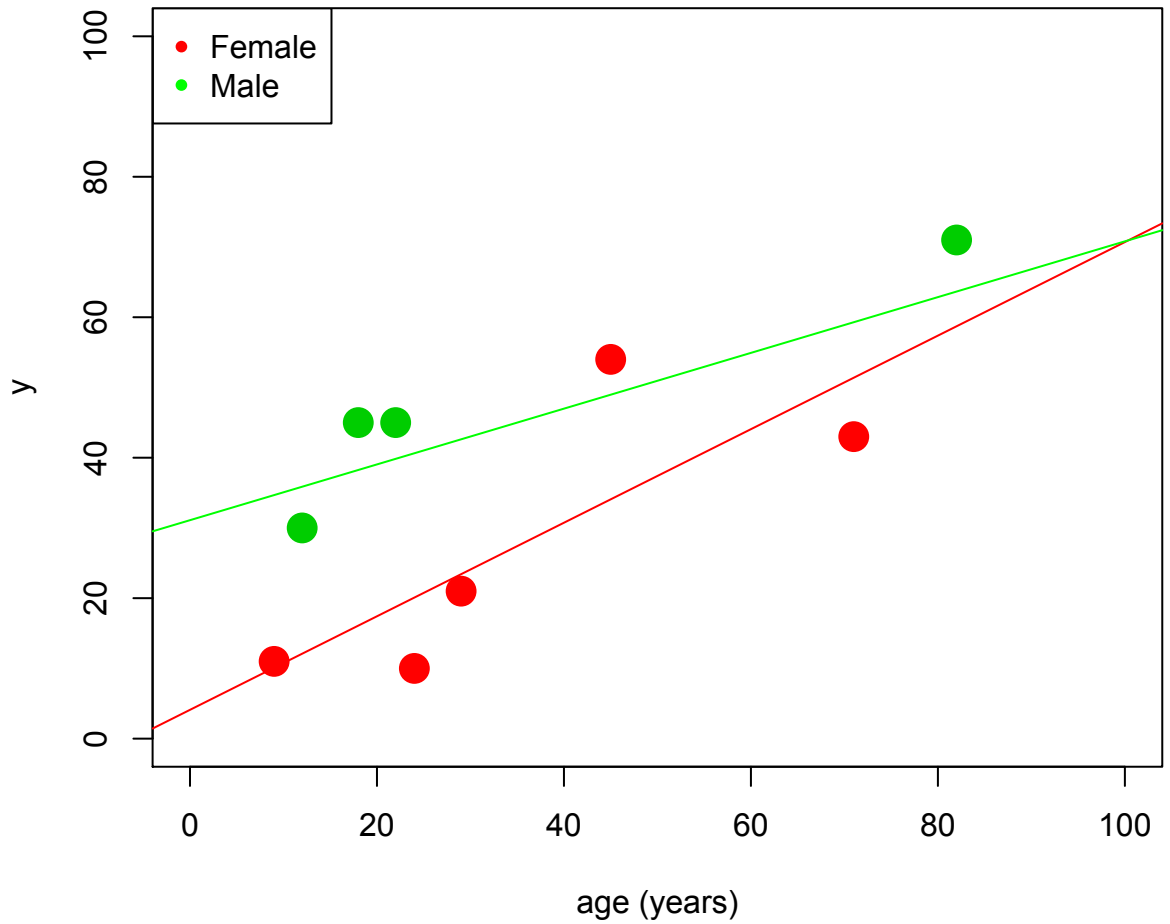
3.9 Categorical explanatory variables

```
> plot(y~x1, ylim=c(0,100), xlim=c(0,100), xlab="age (years)", col=z+2, pch=20, cex=3)
> legend("topleft",c("Female", "Male"), pch=20, col=c("red", "green"))
> abline(4.102, 0.666, col="red")
> abline(4.102 + 27.003, 0.566 + -0.169, col="green")
>
```



3.9 Categorical explanatory variables

```
> plot(y~x1, ylim=c(0,100), xlim=c(0,100), xlab="age (years)", col=z+2, pch=20, cex=3)
> legend("topleft",c("Female", "Male"), pch=20, col=c("red", "green"))
> abline(4.102, 0.666, col="red")
> abline(4.102 + 27.003, 0.566 + -0.169, col="green")
>
```



3.9 Categorical explanatory variables

For categorical variable with m categories, create $m-1$ binary dummy variables $z_{i2}, \ldots, z_{im}$ where

$$z_{ij} = \begin{cases} 1 & \text{if category } j \text{ for } i\text{th case,} \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } j = 2, \ldots, m. \tag{3.113}$$

Here, category 1 is considered as the baseline category. If the ith observation is in category 1, then $(z_{i2}, \ldots, z_{im}) = (0, \ldots, 0)$. If the ith observation is in category 2, then $(z_{i2}, \ldots, z_{im}) = (1, 0, \ldots, 0)$. If the ith observation is in category 3, then $(z_{i2}, \ldots, z_{im}) = (0, 1, 0, \ldots, 0)$, etc. If the ith observation is in category m, then $(z_{i2}, \ldots, z_{im}) = (0, \ldots, 0, 1)$. The regression equation becomes $Y_i = \mu_Y(x_i, z_{i2}, \ldots, z_{im}) + \epsilon_i$,

3.9 Categorical explanatory variables

For categorical variable with m categories, create $m-1$ binary dummy variables $z_{i2}, \ldots, z_{im}$ where

$$z_{ij} = \begin{cases} 1 & \text{if category } j \text{ for } i\text{th case,} \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } j = 2, \ldots, m. \tag{3.113}$$

s in category 1, then
$_{im}) = (1, 0, \ldots, 0)$. If
h observation is in cat-
$_Y(x_i, z_{i2}, \ldots, z_{im}) + \epsilon_i,$

```
> # Alternatively, one could re-code country
> # with binary "dummy" variables...
> z2 <- as.numeric(country=="France")
> z3 <- as.numeric(country=="Thailand")
>
> # ...and use linear regression:
> linear_reg(y, cbind(1, z2, z3))
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
z2  13.333      2.414  5.524         8.189        18.478  0.000
z3   4.167      2.414  1.726        -0.978         9.311  0.105

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.   R2 adjR2 Fstatistic Ftest_pval
1    820.5 262.167 17.478        4.181 0.68 0.638     15.973          0
```

3.9 Categorical explanatory variables

$$\mu_Y(x_i, z_{i2}, \ldots, z_{im}) = \beta_0 + \beta_1 x_i + \beta_2 z_{i2} + \cdots + \beta_m z_{im} \tag{3.114}$$

$$= \begin{cases} \beta_0 + \beta_1 x_i & \text{if category 1,} \\ (\beta_0 + \beta_2) + \beta_1 x_i & \text{if category 2,} \\ (\beta_0 + \beta_3) + \beta_1 x_i & \text{if category 3,} \\ \cdots & \text{etc.} \end{cases} \tag{3.115}$$

This implies a model where the relation of y with x is linear for all categories and there is a common slope β_1. $\beta_2, \ldots, \beta_m$ are the distances of the parallel lines relative to that for category 1.