Statistics with Spa  ows

Lecture 11-a

Julia Schroeder

Julia.schroeder@imperial.ac.uk

Outline

- Linear models – going big
- Categorical predictors

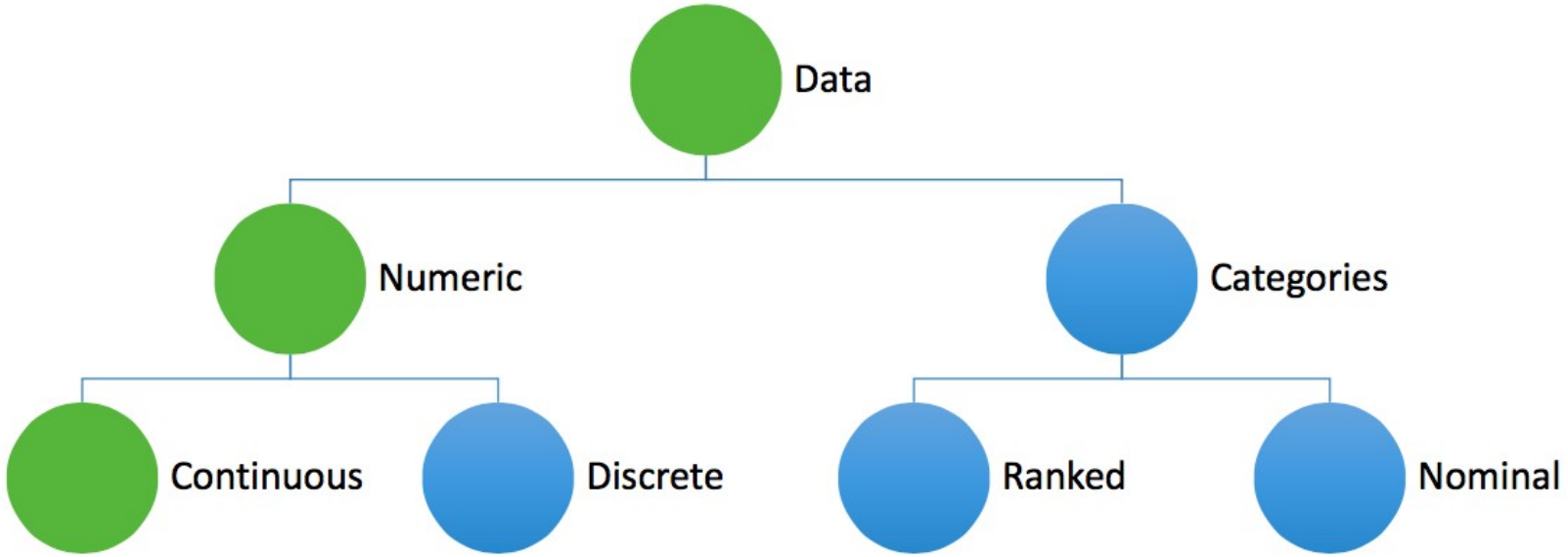
Linear models – different predictors

```
lm(response~explanatory)
```

Linear models – different predictors

```
lm(response~explanatory)
```

Data types



Categorical predictor

`lm(response~explanatory)`

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Male
Female
Female
Female
Male
Male

Categorical predictor

`lm(response~explanatory)`

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

1
0
0
0
1
1

```
> a<-read.table("SparrowSize.txt", header=T)
> str(a)
'data.frame':   1770 obs. of  8 variables:
 $ BirdID: int  1 2 2 2 2 2 2 2 2 ...
 $ Year  : int  2002 2001 2002 2003 2004 2004 2004 2004 2004 2005 ...
 $ Tarsus: num  16.9 16.8 17.2 17.5 17.8 ...
 $ Bill  : num  NA NA NA 13.5 13.4 ...
 $ Wing  : num  76 76 76 76 77 78 77 77 77 77 ...
 $ Mass  : num  23.6 27.5 28.1 27.8 26.5 ...
 $ Sex   : int  0 1 1 1 1 1 1 1 1 ...
 $ Sex.1 : chr  "female" "male" "male" "male" ...
> CatMod1<-lm(Mass~Sex, data=a)
> summary(CatMod1)

Call:
lm(formula = Mass ~ Sex, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4356 -1.4685 -0.0685  1.3644  8.7315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.46852    0.07259 378.414  < 2e-16 ***
Sex          0.56706    0.10176   5.572 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1 on 1702 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.01792,   Adjusted R-squared:  0.01734
F-statistic: 31.05 on 1 and 1702 DF,  p-value: 2.917e-08
```

Hypothesis:
Body mass is sexually selected
-> Males predicted to be heavier

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

23.6
27.5
28.1
27.8
26.5
…
…

0
1
1
1
1
0
…
…

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

$x_i = 0$

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

$x_i = 1$

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

23.6	0
27.5	1
28.1	1
27.8	1
26.5	0
…	…
…	…

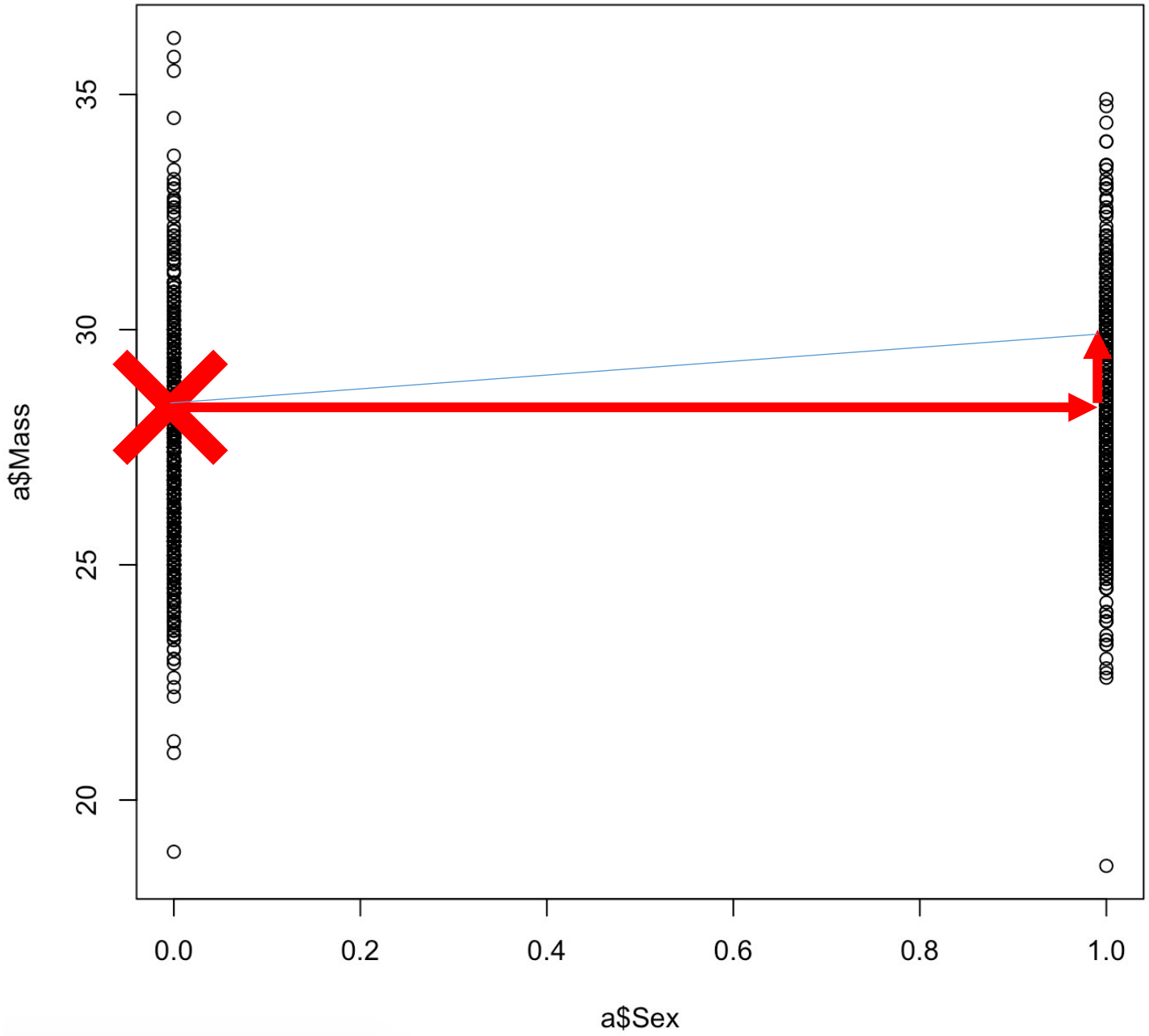
$$y_i = b_0 + \varepsilon_i$$

$x_i = 0$ Intercept

$$y_i = b_0 + b_1 + \varepsilon_i$$

$x_i = 1$ Intercept + sex estimate

```
> summary(CatMod1)

Call:
lm(formula = Mass ~ Sex, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4356 -1.4685 -0.0685  1.3644  8.7315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.46852    0.07259 378.414  < 2e-16 ***
Sex          0.56706    0.10176   5.572 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1 on 1702 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.01792,   Adjusted R-squared:  0.01734
F-statistic: 31.05 on 1 and 1702 DF,  p-value: 2.917e-08
```

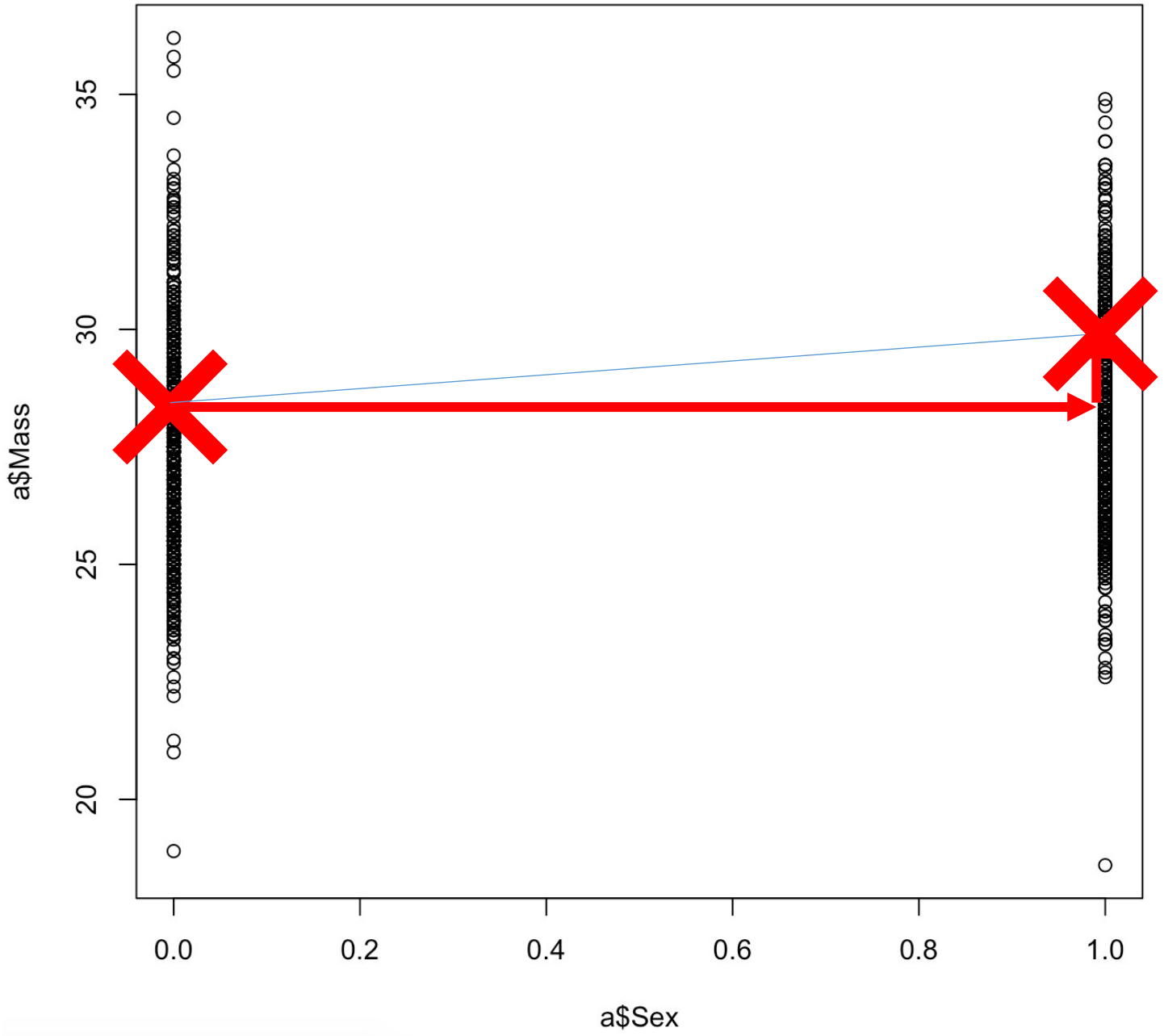


```
> summary(CatMod1)

Call:
lm(formula = Mass ~ Sex, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4356 -1.4685 -0.0685  1.3644  8.7315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.46852    0.07259 378.414  < 2e-16 ***
Sex          0.56706    0.10176   5.572 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1 on 1702 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.01792,   Adjusted R-squared:  0.01734
F-statistic: 31.05 on 1 and 1702 DF,  p-value: 2.917e-08
```

Means of each group:

Intercept = mean of x = 0

Intercept + b_1 = mean of x =1

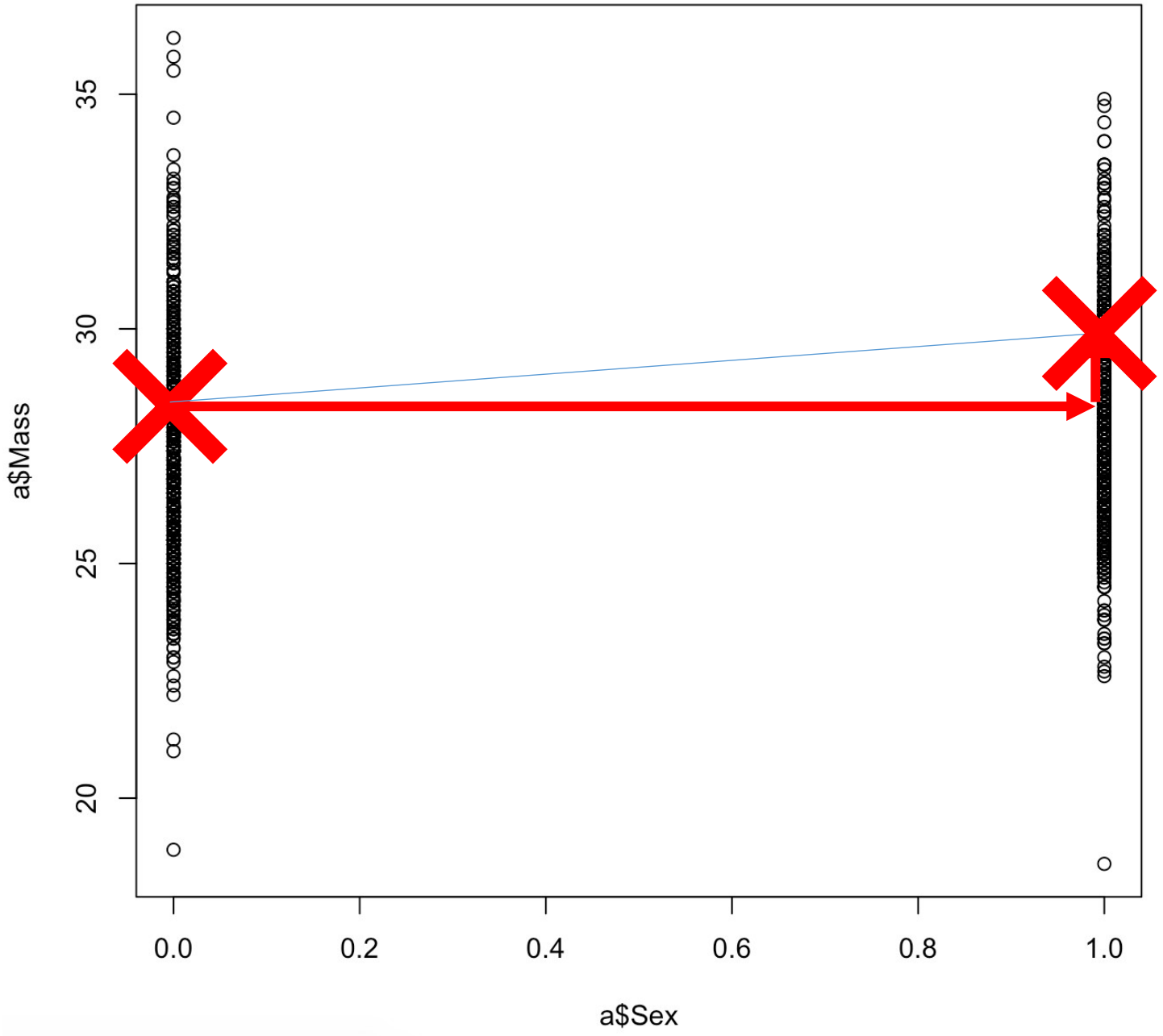


```
> summary(CatMod1)

Call:
lm(formula = Mass ~ Sex, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4356 -1.4685 -0.0685  1.3644  8.7315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.46852    0.07259 378.414  < 2e-16 ***
Sex          0.56706    0.10176   5.572 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1 on 1702 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.01792,   Adjusted R-squared:  0.01734
F-statistic: 31.05 on 1 and 1702 DF,  p-value: 2.917e-08
```

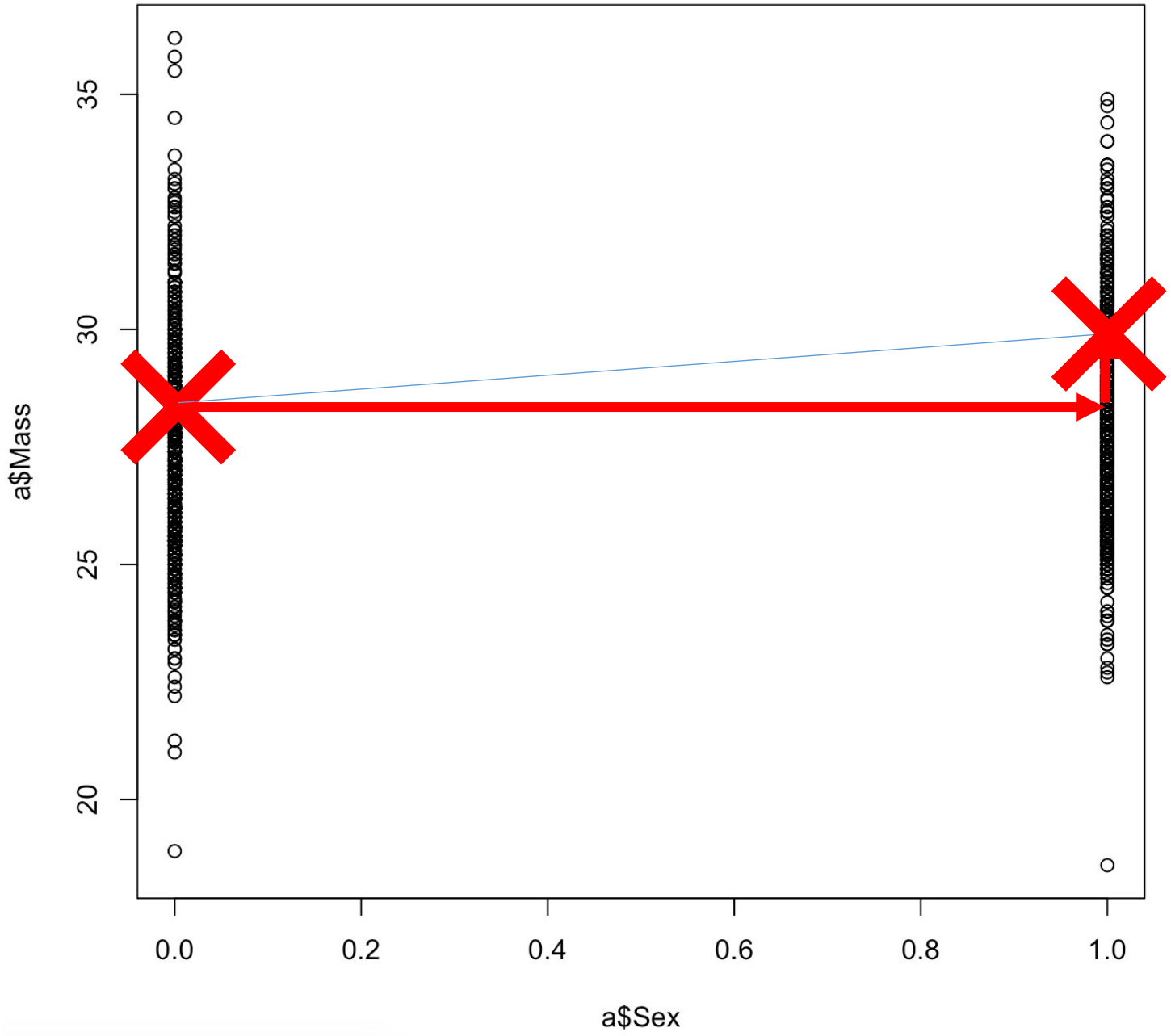
Means of each group:

Intercept = mean of x0 = 27.47

Intercept + b_1 = mean of x1 = 28.02



```
> summary(CatMod1)

Call:
lm(formula = Mass ~ Sex, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4356 -1.4685 -0.0685  1.3644  8.7315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.46852    0.07259 378.414  < 2e-16 ***
Sex          0.56706    0.10176   5.572 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1 on 1702 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.01792,   Adjusted R-squared:  0.01734
F-statistic: 31.05 on 1 and 1702 DF,  p-value: 2.917e-08
```



Means of each group:

Intercept = mean of x0 = 27.47

Intercept + b_1 = mean of x1 = 28.02

b_1 = difference between both groups

```
> summary(CatMod1)

Call:
lm(formula = Mass ~ Sex, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4356 -1.4685 -0.0685  1.3644  8.7315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.46852    0.07259 378.414  < 2e-16 ***
Sex          0.56706    0.10176   5.572 2.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.1 on 1702 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.01792,   Adjusted R-squared:  0.01734
F-statistic: 31.05 on 1 and 1702 DF,  p-value: 2.917e-08
```

Categorical predictors:

- Intercept is mean of reference category
- Categorical predictor is difference to reference category

Means of each group:

Intercept = mean of x0 = 27.47

Intercept + b_1 = mean of x1 = 28.02

b_1 = difference between both groups

```
> a<-read.table("SparrowSize.txt", header=T)
> str(a)
'data.frame':   1770 obs. of  8 variables:
 $ BirdID: int  1 2 2 2 2 2 2 2 2 ...
 $ Year  : int  2002 2001 2002 2003 2004 2004 2004 2004 2004 2005 ...
 $ Tarsus: num  16.9 16.8 17.2 17.5 17.8 ...
 $ Bill  : num  NA NA NA 13.5 13.4 ...
 $ Wing  : num  76 76 76 76 77 78 77 77 77 77 ...
 $ Mass  : num  23.6 27.5 28.1 27.8 26.5 ...
 $ Sex   : int  0 1 1 1 1 1 1 1 1 1 ...
 $ Sex.1 : chr  "female" "male" "male" "male" ...
```

```
> CatMod2<-lm(Mass~Sex.1, data=a)
> summary(CatMod2)

Call:
lm(formula = Mass ~ Sex.1, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4356 -1.4685 -0.0685  1.3644  8.7315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.46852    0.07259 378.414  < 2e-16 ***
Sex.1male    0.56706    0.10176   5.572 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1 on 1702 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.01792,   Adjusted R-squared:  0.01734
F-statistic: 31.05 on 1 and 1702 DF,  p-value: 2.917e-08
```

R choose the reference category alphanumerically and uses internal dummy codes:

Female = 0

→ Male = 1

Categorical predictors

- But what about categorical predictors with more than two levels?

```
> a<-read.table("SparrowSize.txt", header=T)
> str(a)
'data.frame':   1770 obs. of  8 variables:
 $ BirdID: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Year  : int  2002 2001 2002 2003 2004 2004 2004 2004 2004 2005 ...
 $ Tarsus: num  16.9 16.8 17.2 17.5 17.8 ...
 $ Bill  : num  NA NA NA 13.5 13.4 ...
 $ Wing  : num  76 76 76 76 77 78 77 77 77 77 ...
 $ Mass  : num  23.6 27.5 28.1 27.8 26.5 ...
 $ Sex   : int  0 1 1 1 1 1 1 1 1 1 ...
 $ Sex.1 : chr  "female" "male" "male" "male" ...
```

Year as categorical predictor

Hypothesis:
Different years have different food supply

Prediction:
Mass differs between years

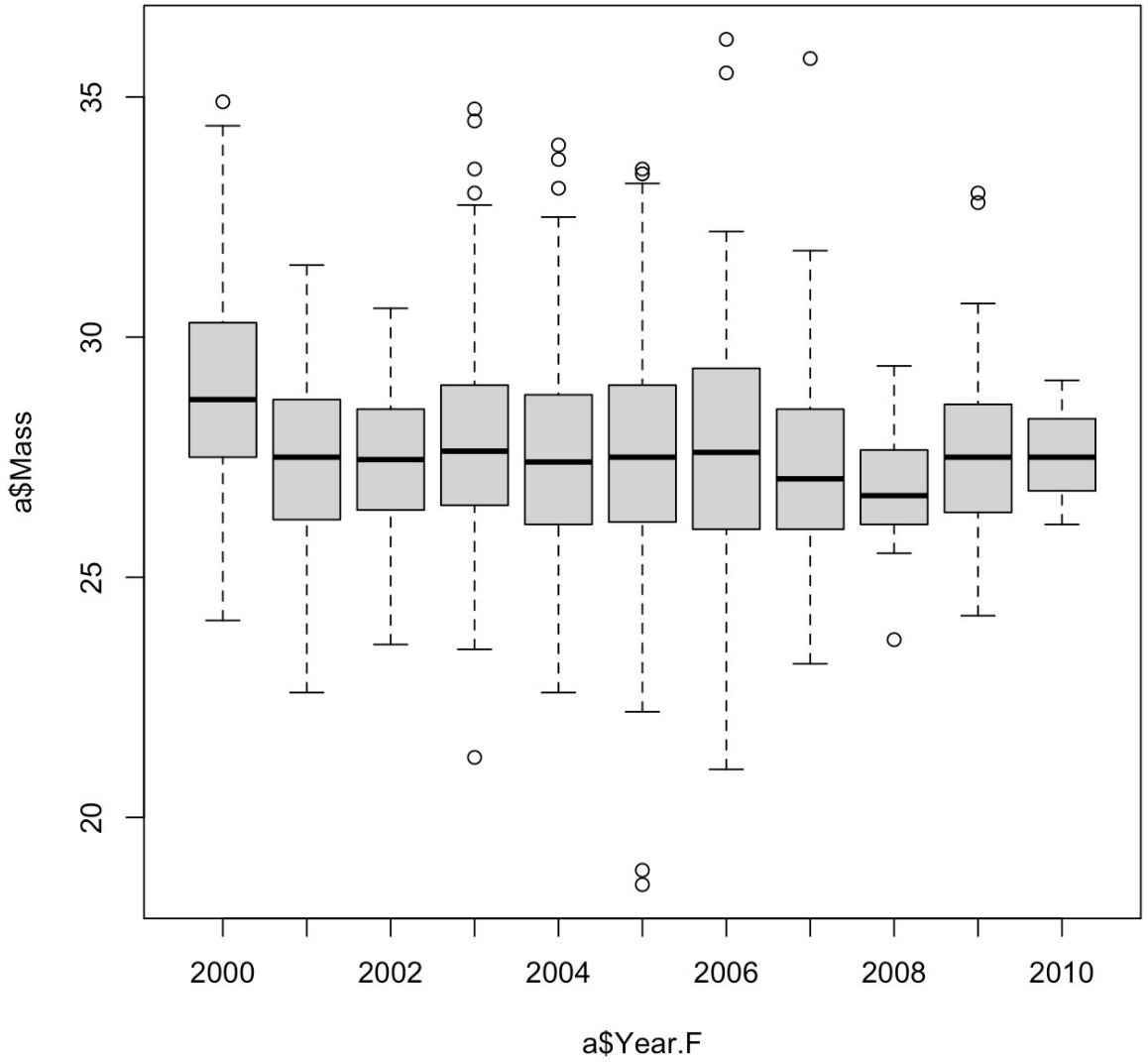
```
> a$Year.F<-as.factor(a$Year)
```

```
> a$Year.F<-as.factor(a$Year)
> str(a$Year.F)
 Factor w/ 11 levels "2000","2001",..: 3 2 3 4 5 5 5 5 5 6 ...
```

```
> a$Year.F<-as.factor(a$Year)
> str(a$Year.F)
 Factor w/ 11 levels "2000","2001",..: 3 2 3 4 5 5 5 5 5 6 ...
> CatMod3<-lm(Mass~Year.F, data=a)
```

```
> a$Year.F<-as.factor(a$Year)
> str(a$Year.F)
 Factor w/ 11 levels "2000","2001",..: 3 2 3 4 5 5 5 5 5 6 ...
> CatMod3<-lm(Mass~Year.F, data=a)
> summary(CatMod3)

Call:
lm(formula = Mass ~ Year.F, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0051 -1.4051 -0.1089  1.2645  8.4874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8537     0.1375 209.772  < 2e-16 ***
Year.F2001   -1.3620     0.2518  -5.410 7.22e-08 ***
Year.F2002   -1.3871     0.2903  -4.778 1.92e-06 ***
Year.F2003   -1.0596     0.2105  -5.035 5.30e-07 ***
Year.F2004   -1.3182     0.1686  -7.816 9.49e-15 ***
Year.F2005   -1.2486     0.1695  -7.367 2.71e-13 ***
Year.F2006   -1.1411     0.2349  -4.858 1.29e-06 ***
Year.F2007   -1.4176     0.2546  -5.568 2.99e-08 ***
Year.F2008   -2.0810     0.6411  -3.246  0.00119 **
Year.F2009   -0.9842     0.4544  -2.166  0.03046 *
Year.F2010   -1.2871     1.2070  -1.066  0.28642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.077 on 1693 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.04451,   Adjusted R-squared:  0.03887
F-statistic: 7.887 on 10 and 1693 DF,  p-value: 1.721e-12
```

```
> plot(a$Mass~a$Year.F)
>
```



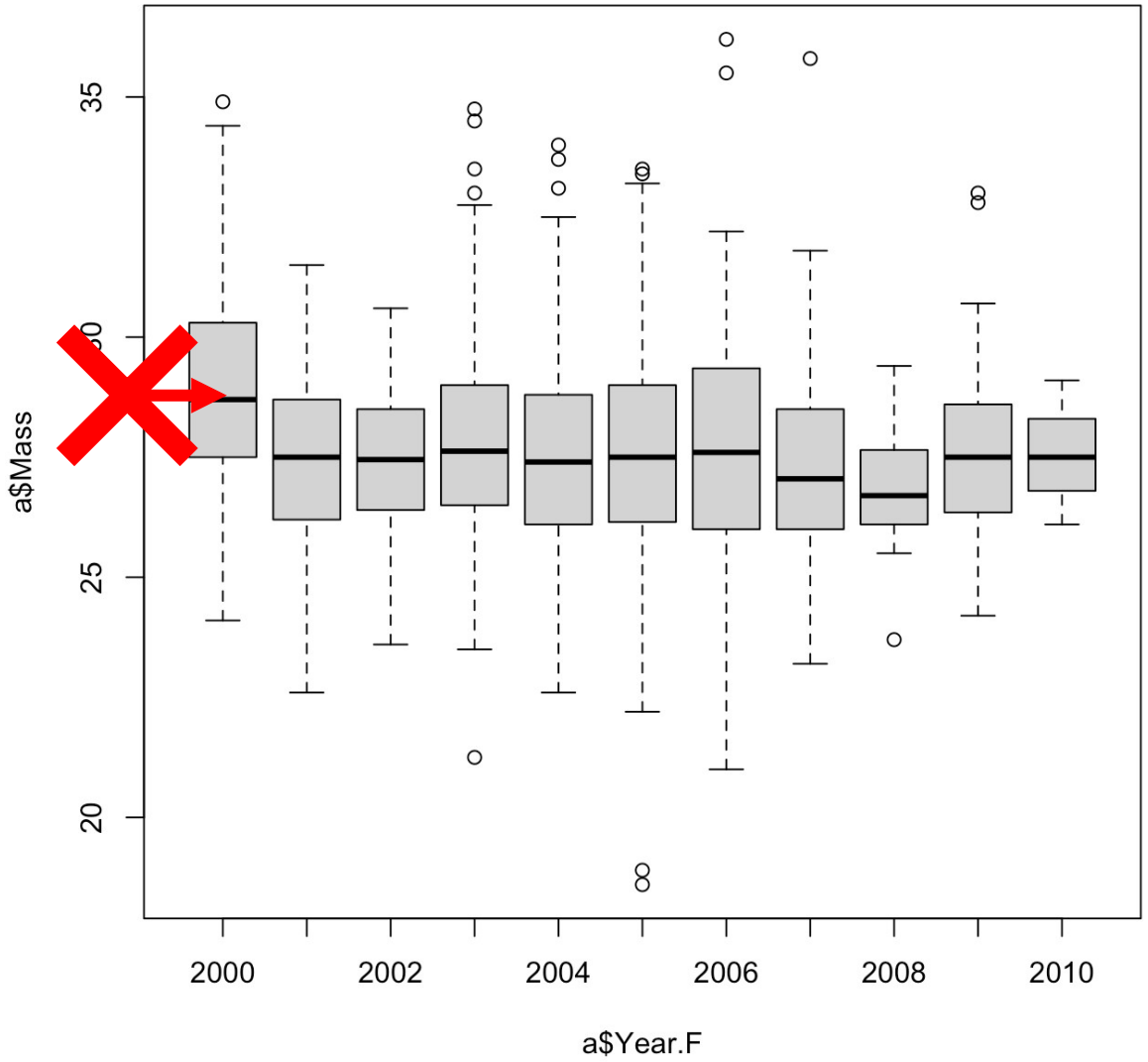
```
> a$Year.F<-as.factor(a$Year)
> str(a$Year.F)
 Factor w/ 11 levels "2000","2001",..: 3 2 3 4 5 5 5 5 5 6 ...
> CatMod3<-lm(Mass~Year.F, data=a)
> summary(CatMod3)

Call:
lm(formula = Mass ~ Year.F, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0051 -1.4051 -0.1089  1.2645  8.4874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8537     0.1375 209.772  < 2e-16 ***
Year.F2001   -1.3620     0.2518  -5.410 7.22e-08 ***
Year.F2002   -1.3871     0.2903  -4.778 1.92e-06 ***
Year.F2003   -1.0596     0.2105  -5.035 5.30e-07 ***
Year.F2004   -1.3182     0.1686  -7.816 9.49e-15 ***
Year.F2005   -1.2486     0.1695  -7.367 2.71e-13 ***
Year.F2006   -1.1411     0.2349  -4.858 1.29e-06 ***
Year.F2007   -1.4176     0.2546  -5.568 2.99e-08 ***
Year.F2008   -2.0810     0.6411  -3.246  0.00119 **
Year.F2009   -0.9842     0.4544  -2.166  0.03046 *
Year.F2010   -1.2871     1.2070  -1.066  0.28642
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.077 on 1693 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.04451,   Adjusted R-squared:  0.03887
F-statistic: 7.887 on 10 and 1693 DF,  p-value: 1.721e-12
```

```
> plot(a$Mass~a$Year.F)
>
```

2000 is the reference level

-> 2000 is the intercept – 28.85

-> 2001 is the intercept + 1.36



```
> a$Year.F<-as.factor(a$Year)
> str(a$Year.F)
 Factor w/ 11 levels "2000","2001",..: 3 2 3 4 5 5 5 5 5 6 ...
> CatMod3<-lm(Mass~Year.F, data=a)
> summary(CatMod3)

Call:
lm(formula = Mass ~ Year.F, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0051 -1.4051 -0.1089  1.2645  8.4874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8537     0.1375 209.772  < 2e-16 ***
Year.F2001   -1.3620     0.2518  -5.410 7.22e-08 ***
Year.F2002   -1.3871     0.2903  -4.778 1.92e-06 ***
Year.F2003   -1.0596     0.2105  -5.035 5.30e-07 ***
Year.F2004   -1.3182     0.1686  -7.816 9.49e-15 ***
Year.F2005   -1.2486     0.1695  -7.367 2.71e-13 ***
Year.F2006   -1.1411     0.2349  -4.858 1.29e-06 ***
Year.F2007   -1.4176     0.2546  -5.568 2.99e-08 ***
Year.F2008   -2.0810     0.6411  -3.246  0.00119 **
Year.F2009   -0.9842     0.4544  -2.166  0.03046 *
Year.F2010   -1.2871     1.2070  -1.066  0.28642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.077 on 1693 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.04451,   Adjusted R-squared:  0.03887
F-statistic: 7.887 on 10 and 1693 DF,  p-value: 1.721e-12
```

```
> plot(a$Mass~a$Year.F)
>
```

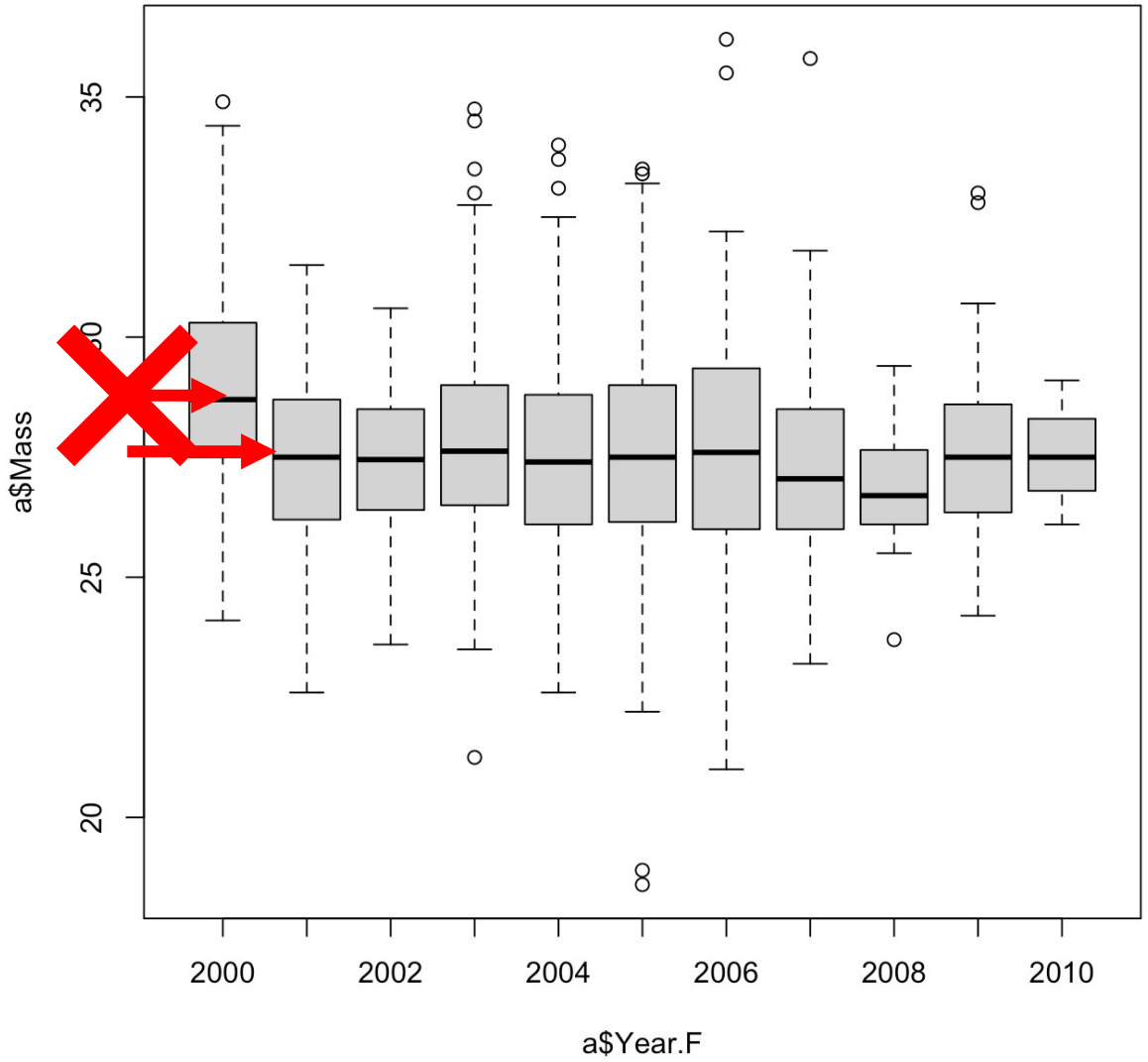
2000 is the reference level

-> 2000 is the intercept – 28.85

-> 2001 is the intercept - 1.36

-> 2002 is the intercept - 1.39

-> 2003 is the intercept – 1.06



```
> a$Year.F<-as.factor(a$Year)
> str(a$Year.F)
 Factor w/ 11 levels "2000","2001",..: 3 2 3 4 5 5 5 5 5 6 ...
> CatMod3<-lm(Mass~Year.F, data=a)
> summary(CatMod3)

Call:
lm(formula = Mass ~ Year.F, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0051 -1.4051 -0.1089  1.2645  8.4874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8537     0.1375 209.772  < 2e-16 ***
Year.F2001   -1.3620     0.2518  -5.410 7.22e-08 ***
Year.F2002   -1.3871     0.2903  -4.778 1.92e-06 ***
Year.F2003   -1.0596     0.2105  -5.035 5.30e-07 ***
Year.F2004   -1.3182     0.1686  -7.816 9.49e-15 ***
Year.F2005   -1.2486     0.1695  -7.367 2.71e-13 ***
Year.F2006   -1.1411     0.2349  -4.858 1.29e-06 ***
Year.F2007   -1.4176     0.2546  -5.568 2.99e-08 ***
Year.F2008   -2.0810     0.6411  -3.246  0.00119 **
Year.F2009   -0.9842     0.4544  -2.166  0.03046 *
Year.F2010   -1.2871     1.2070  -1.066  0.28642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.077 on 1693 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared:  0.04451,   Adjusted R-squared:  0.03887
F-statistic: 7.887 on 10 and 1693 DF,  p-value: 1.721e-12
```

```
> plot(a$Mass~a$Year.F)
>
```

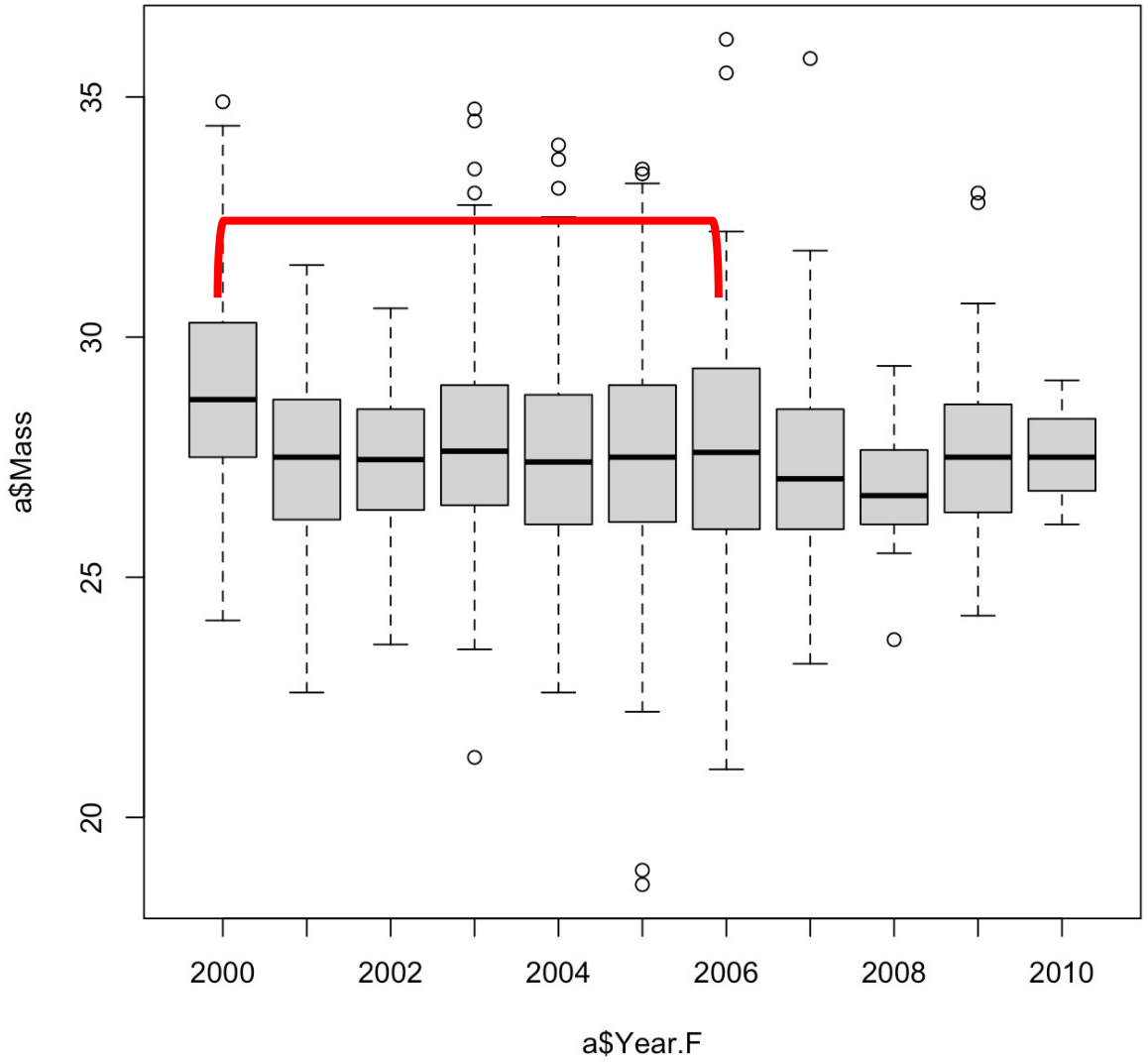
2000 is the reference level

-> 2000 is the intercept – 28.85

-> 2001 is the intercept - 1.36

-> 2002 is the intercept - 1.39

-> 2003 is the intercept – 1.06



Take home: categorical predictors:

- R chooses reference level alpha-numerical
- Intercept = mean at reference level
- Estimates: = difference to reference level
- T-test: difference to reference level
- Using categorical variables with many levels – not so good because loosing lots of d.f.s