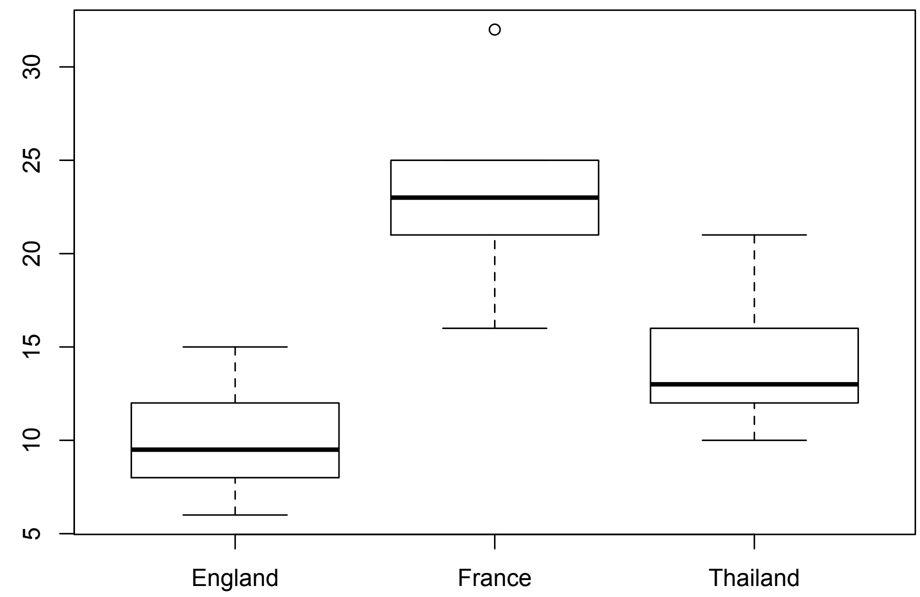
Stat 306: Finding Relationships in Data.

Lecture 11

Section 3.9 (recap) + quadratic terms + Section 3.10 - Partial correlation

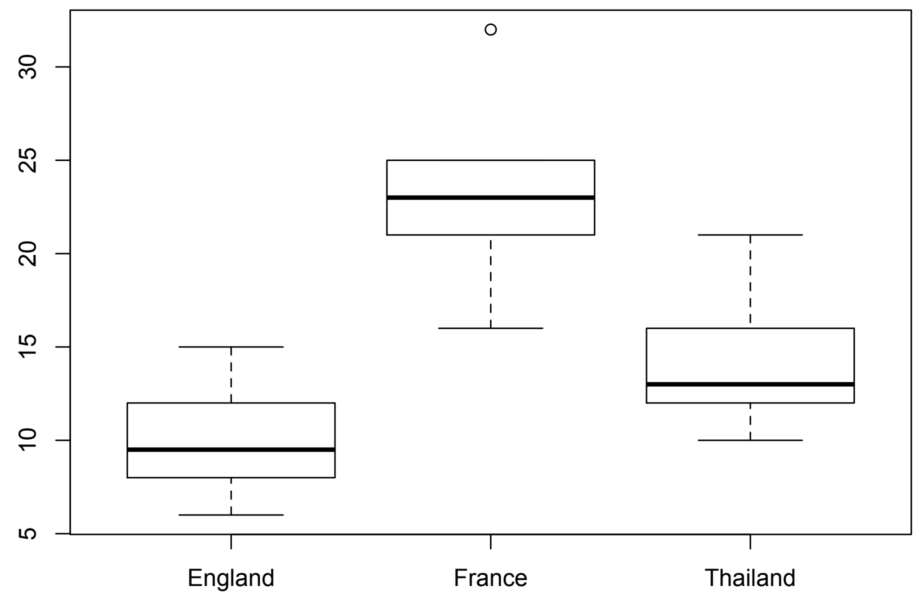
3.9 Categorical explanatory variables

```
> y <- c(23,25,21,32,16,23,15,10,8,9,6,12,13,13,12,21,16,10)
> country <- c(
+ "France", "France", "France",
+ "France", "France", "France",
+ "England", "England", "England",
+ "England", "England", "England",
+ "Thailand", "Thailand", "Thailand",
+ "Thailand", "Thailand", "Thailand")
>
> # together in a data.frame:
> mydata <- data.frame(y, country)
> mydata
    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

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> myX
        x1 x2
 [1,] 1  1  0
 [2,] 1  1  0
 [3,] 1  1  0
 [4,] 1  1  0
 [5,] 1  1  0
 [6,] 1  1  0
 [7,] 1  0  0
 [8,] 1  0  0
 [9,] 1  0  0
[10,] 1  0  0
[11,] 1  0  0
[12,] 1  0  0
[13,] 1  0  1
[14,] 1  0  1
[15,] 1  0  1
[16,] 1  0  1
[17,] 1  0  1
[18,] 1  0  1
```



3.9 Categorical explanatory variables

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> myX
        x1 x2
 [1,] 1  1  0
 [2,] 1  1  0
 [3,] 1  1  0
 [4,] 1  1  0
 [5,] 1  1  0
 [6,] 1  1  0
 [7,] 1  0  0
 [8,] 1  0  0
 [9,] 1  0  0
[10,] 1  0  0
[11,] 1  0  0
[12,] 1  0  0
[13,] 1  0  1
[14,] 1  0  1
[15,] 1  0  1
[16,] 1  0  1
[17,] 1  0  1
[18,] 1  0  1
```

Consider our model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

3.9 Categorical explanatory variables

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> myX
        x1 x2
 [1,] 1  1  0
 [2,] 1  1  0
 [3,] 1  1  0
 [4,] 1  1  0
 [5,] 1  1  0
 [6,] 1  1  0
 [7,] 1  0  0
 [8,] 1  0  0
 [9,] 1  0  0
[10,] 1  0  0
[11,] 1  0  0
[12,] 1  0  0
[13,] 1  0  1
[14,] 1  0  1
[15,] 1  0  1
[16,] 1  0  1
[17,] 1  0  1
[18,] 1  0  1
```

Consider our model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

for **France**

$$Y = \beta_0 + \beta_1(1) + \beta_2 0$$

$$= \beta_0 + \beta_1$$

3.9 Categorical explanatory variables

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> myX
        x1 x2
 [1,] 1  1  0
 [2,] 1  1  0
 [3,] 1  1  0
 [4,] 1  1  0
 [5,] 1  1  0
 [6,] 1  1  0
 [7,] 1  0  0
 [8,] 1  0  0
 [9,] 1  0  0
[10,] 1  0  0
[11,] 1  0  0
[12,] 1  0  0
[13,] 1  0  1
[14,] 1  0  1
[15,] 1  0  1
[16,] 1  0  1
[17,] 1  0  1
[18,] 1  0  1
```

Consider our model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(France) $Y = \beta_0 + \beta_1(1) + \beta_2 0$

$= \beta_0 + \beta_1$

for **England**

$$Y = \beta_0 + \beta_1 0 + \beta_2 0$$

$$= \beta_0$$

3.9 Categorical explanatory variables

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> myX
        x1 x2
 [1,] 1  1  0
 [2,] 1  1  0
 [3,] 1  1  0
 [4,] 1  1  0
 [5,] 1  1  0
 [6,] 1  1  0
 [7,] 1  0  0
 [8,] 1  0  0
 [9,] 1  0  0
[10,] 1  0  0
[11,] 1  0  0
[12,] 1  0  0
[13,] 1  0  1
[14,] 1  0  1
[15,] 1  0  1
[16,] 1  0  1
[17,] 1  0  1
[18,] 1  0  1
```

Consider our model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(France) $Y = \beta_0 + \beta_1(1) + \beta_2 0$

$= \beta_0 + \beta_1$

(England) $Y = \beta_0 + \beta_1 0 + \beta_2 0$

$= \beta_0$

for **Thailand**

$$Y = \beta_0 + \beta_1 0 + \beta_2(1)$$

$$= \beta_0 + \beta_2$$

3.9 Categorical explanatory variables

Therefore, England is the "reference category".

Consider our model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(France) $Y = \beta_0 + \beta_1(1) + \beta_2 0$

$= \beta_0 + \beta_1$

(England) $Y = \beta_0 + \beta_1 0 + \beta_2 0$

$= \beta_0$

(Thailand) $Y = \beta_0 + \beta_1 0 + \beta_2(1)$

$= \beta_0 + \beta_2$

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

H_1: $\mu_{France} \neq \mu_{England}$

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

H_1: $\mu_{France} \neq \mu_{Thailand}$

The model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(France)

$$Y = \beta_0 + \beta_1(1) + \beta_2 0$$
$$= \beta_0 + \beta_1$$

(England)

$$Y = \beta_0 + \beta_1 0 + \beta_2 0$$
$$= \beta_0$$

(Thailand)

$$Y = \beta_0 + \beta_1 0 + \beta_2(1)$$
$$= \beta_0 + \beta_2$$

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_0 + \beta_1 = \beta_0$

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_0 + \beta_1 = \beta_0 + \beta_2$

The model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(France) $Y = \beta_0 + \beta_1(1) + \beta_2 0$

$= \beta_0 + \beta_1$

(England) $Y = \beta_0 + \beta_1 0 + \beta_2 0$

$= \beta_0$

(Thailand) $Y = \beta_0 + \beta_1 0 + \beta_2(1)$

$= \beta_0 + \beta_2$

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

The model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(France) $Y = \beta_0 + \beta_1(1) + \beta_2 0$

$= \beta_0 + \beta_1$

(England) $Y = \beta_0 + \beta_1 0 + \beta_2 0$

$= \beta_0$

(Thailand) $Y = \beta_0 + \beta_1 0 + \beta_2(1)$

$= \beta_0 + \beta_2$

3.9 Categorical explanatory variables

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                      x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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
```

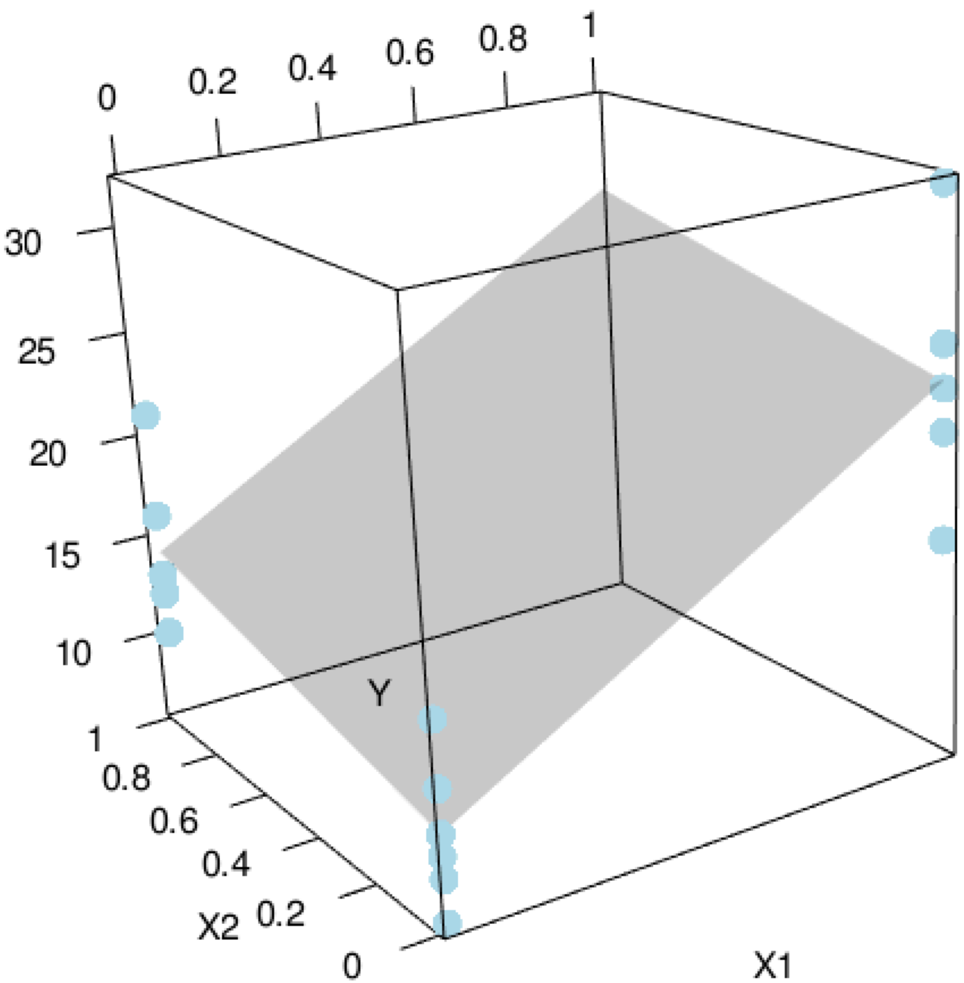
So we have that:

b_0 = 10.00 , Var(b_0) = 2.91 $\Rightarrow SE(b_0) = \sqrt{2.91} = 1.71$

b_1 = 13.33 , Var(b_1) = 5.82 $\Rightarrow SE(b_1) = \sqrt{5.82} = 2.41$

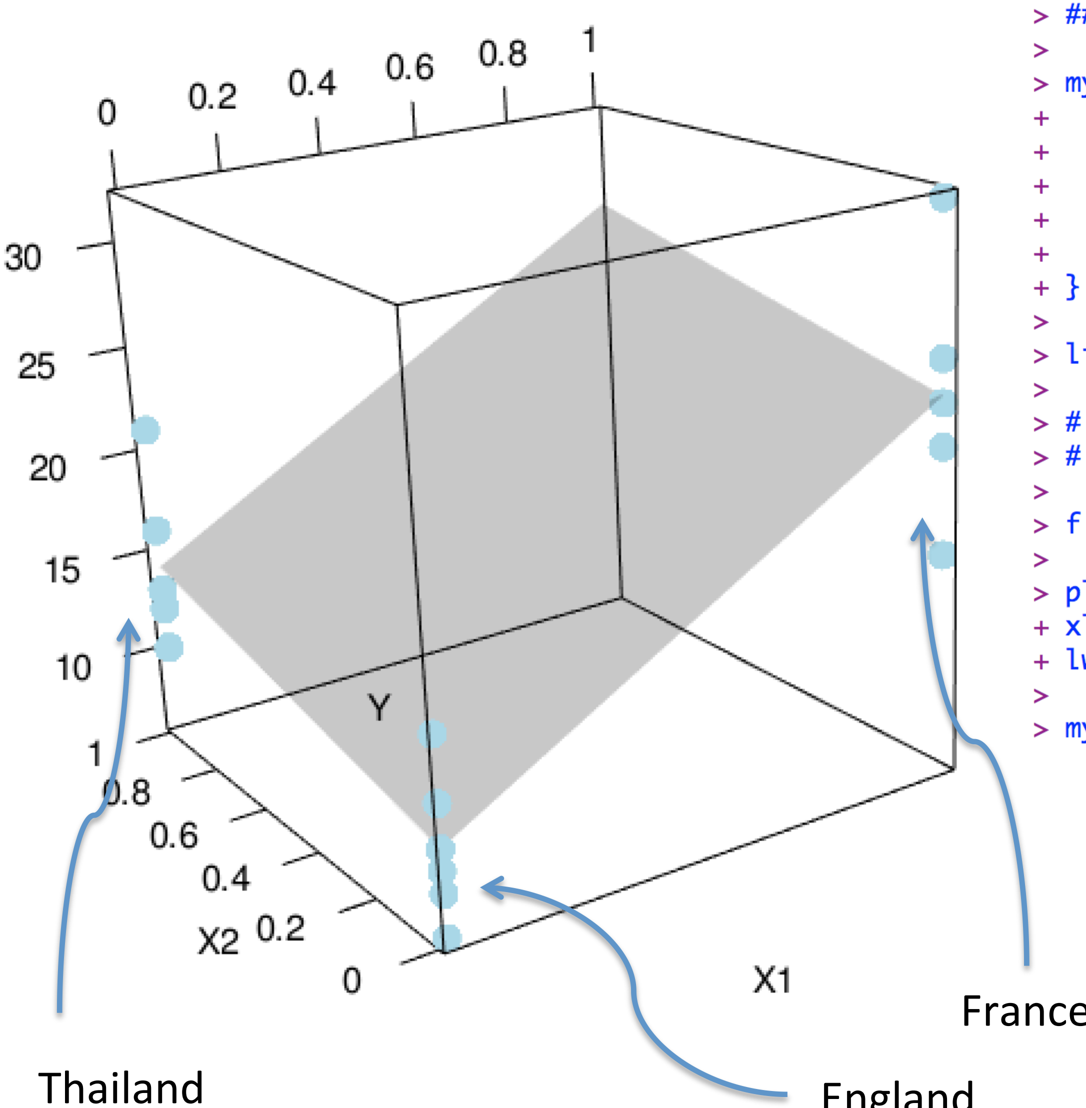
b_2 = 4.17 , $V\hat{ar}(b_2) = 5.82 \Rightarrow SE(b_2) = \sqrt{5.82} = 2.41$

3.9 Categorical explanatory variables



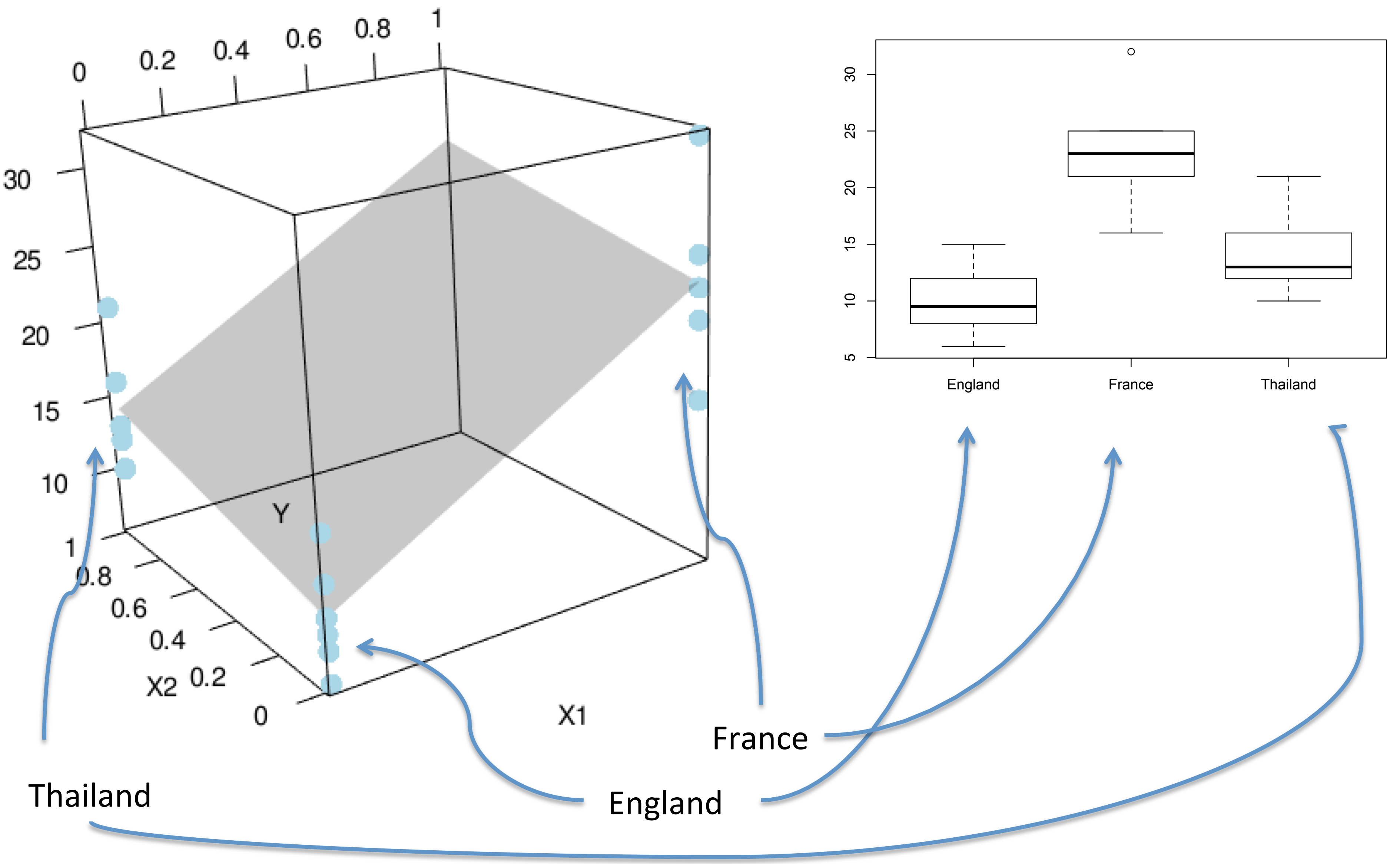

```
> ## Let's take a closer look with a 3d plot:
>
> my_surface <- function(f, n=10, ...) {
+   ranges <- rgl:::.getRanges()
+   x <- seq(ranges$xlim[1], ranges$xlim[2], length=n)
+   y <- seq(ranges$ylim[1], ranges$ylim[2], length=n)
+   z <- outer(x,y,f)
+   surface3d(x, y, z, ...)
+ }
>
> library(rgl)
>
> # This "f" is the fitted hyperplane using the beta
> # estimates from the linear regression model above
>
> f <- function(x1, x2){  10 + 13.333*x1 + 4.167*x2}
>
> plot3d(x1,x2,y, type="p", col="lightblue",
+ xlab="X1", ylab="X2", zlab="Y", site=5,
+ lwd=15, size=12)
>
> my_surface(f, alpha=.2 )
```

3.9 Categorical explanatory variables




```
> ## Let's take a closer look with a 3d plot:
>
> my_surface <- function(f, n=10, ...) {
+   ranges <- rgl:::.getRanges()
+   x <- seq(ranges$xlim[1], ranges$xlim[2], length=n)
+   y <- seq(ranges$ylim[1], ranges$ylim[2], length=n)
+   z <- outer(x,y,f)
+   surface3d(x, y, z, ...)
+ }
>
> library(rgl)
>
> # This "f" is the fitted hyperplane using the beta
> # estimates from the linear regression model above
>
> f <- function(x1, x2){  10 + 13.333*x1 + 4.167*x2}
>
> plot3d(x1,x2,y, type="p", col="lightblue",
+ xlab="X1", ylab="X2", zlab="Y", site=5,
+ lwd=15, size=12)
>
> my_surface(f, alpha=.2 )
```

3.9 Categorical explanatory variables



3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1 - b_2)/SE(b_1 + b_2)$

So we have that:

$b_0 = 10.00$, $Var(b_0) = 2.91$ $\Rightarrow SE(b_0) = \sqrt{2.91} = 1.71$

$b_1 = 13.33$, $Var(b_1) = 5.82$ $\Rightarrow SE(b_1) = \sqrt{5.82} = 2.41$

$b_2 = 4.17$, $V\hat{ar}(b_2) = 5.82 \Rightarrow SE(b_2) = \sqrt{5.82} = 2.41$

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 13.33/2.41

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1 - b_2)/SE(b_1 - b_2)$

= ???

So we have that:

$b_0 = 10.00$, $Var(b_0) = 2.91$ $\Rightarrow SE(b_0) = \sqrt{2.91} = 1.71$

$b_1 = 13.33$, $Var(b_1) = 5.82$ $\Rightarrow SE(b_1) = \sqrt{5.82} = 2.41$

$b_2 = 4.17$, $V\hat{a}r(b_2) = 5.82 \Rightarrow SE(b_2) = \sqrt{5.82} = 2.41$

3.9 Categorical explanatory variables

You can always "re-level" (i.e. change the reference level to get a comparison that you want). But this is not necessary. The fitted model will not change. And for Standard Errors and Intervals, we should remember that…

$$Var(A - B) = Var(A) + Var(B) - 2Cov(A,B)$$

Or more generally:

$$Var(aA - bB) = a^2Var(A) + b^2Var(B) - 2abCov(A,B)$$

3.9 Categorical explanatory variables

we should
remember that...

$$Var(A – B) = Var(A) + Var(B) – 2Cov(A,B)$$

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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

$$\mathrm{Var}(A - B) = \mathrm{Var}(A) + \mathrm{Var}(B) - 2\mathrm{Cov}(A,B)$$

$$\mathrm{Var}(B_1 - B_2) = \mathrm{Var}(B_1) + \mathrm{Var}(B_2) - 2\mathrm{Cov}(B_1, B_2)$$

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                     x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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

$$Var(A - B) = Var(A) + Var(B) - 2Cov(A,B)$$

$$Var(B_1 - B_2) = Var(B_1) + Var(B_2) - 2Cov(B_1, B_2)$$

This is estimated by:

$$= 5.83 \quad + \quad 5.83 \qquad - 2(2.91)$$

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                   x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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

$$Var(A - B) = Var(A) + Var(B) - 2Cov(A,B)$$

$$Var(B_1 - B_2) = Var(B_1) + Var(B_2) - 2Cov(B_1, B_2)$$

This is estimated by:

$$= 5.83 \quad + \quad 5.83 \qquad - 2(2.91)$$

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                      x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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

$$\text{Var}(A - B) = \text{Var}(A) + \text{Var}(B) - 2\text{Cov}(A,B)$$

$$\text{Var}(B_1 - B_2) = \text{Var}(B_1) + \text{Var}(B_2) - 2\text{Cov}(B_1, B_2)$$

This is estimated by:

$$=5.84$$

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                  x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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

$$\mathrm{Var}(\hat{B}_1 - \hat{B}_2) = \mathrm{Var}(\hat{B}_1) + \mathrm{Var}(\hat{B}_2) - 2\mathrm{Cov}(\hat{B}_1, \hat{B}_2)$$

$$= 5.84$$

$$\mathrm{SE}(b_1 - b_2) = \sqrt{5.84} = 2.42$$

```
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   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

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 13.33/2.41

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1-b_2)/SE(b_1 - b_2)$

= ???

So we have that:

$b_0 = 10.00$, $Var(b_0) = 2.91$ $\Rightarrow SE(b_0) = \sqrt{2.91} = 1.71$

$b_1 = 13.33$, $Var(b_1) = 5.82$ $\Rightarrow SE(b_1) = \sqrt{5.82} = 2.41$

$b_2 = 4.17$, $V\hat{ar}(b_2) = 5.82 \Rightarrow SE(b_2) = \sqrt{5.82} = 2.41$

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 13.33/2.41

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1-b_2)/SE(b_1- b_2)$

= -9.12/2.42

So we have that:

$b_0 = 10.00$, $Var(b_0) = 2.91$ $\Rightarrow SE(b_0) = \sqrt{2.91} = 1.71$

$b_1 = 13.33$, $Var(b_1) = 5.82$ $\Rightarrow SE(b_1) = \sqrt{5.82} = 2.41$

$b_2 = 4.17$, $\hat{Var}(b_2) = 5.82 \Rightarrow SE(b_2) = \sqrt{5.82} = 2.41$

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 5.53

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1 - b_2)/SE(b_1 - b_2)$

= -3.798

So we have that:

b_0 = 10.00 , $Var(b_0)$ = 2.91 $\Rightarrow SE(b_0) = \sqrt{2.91} = 1.71$

b_1 = 13.33 , $Var(b_1)$ = 5.82 $\Rightarrow SE(b_1) = \sqrt{5.82} = 2.41$

b_2 = 4.17 , $V\hat{ar}(b_2) = 5.82 \Rightarrow SE(b_2) = \sqrt{5.82} = 2.41$

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 5.53

Therefore:

p-value = 2*(1-pt(abs(5.53), n-k))

< 0.0001

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1-b_2)/SE(b_1- b_2)$

= -3.798

Therefore:

p-value = 2*(1-pt(abs(-3.798), n-k))

= 0.002

3.9 Categorical explanatory variables

Consider two hypothesis tests:

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 5.53

p-value = 2*(1-pt(abs(5.53), n-k))

< 0.0001

We reject the null hypothesis: the average for France is different than the average for England.

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1-b_2)/SE(b_1- b_2)$

= -3.798

p-value = 2*(1-pt(abs(-3.798), n-k))

= 0.002

We reject the null hypothesis: the average for France is different than the average for Thailand.

3.9 Categorical explanatory variables

What if France was selected as the reference category?

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 5.53

p-value = 2*(1-pt(abs(5.53), n-k))

< 0.0001

We reject the null hypothesis: the average For France is different than the average for England

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1 - b_2)/SE(b_1 - b_2)$

= -3.798

p-value = 2*(1-pt(abs(-3.798), n-k))

= 0.002

We reject the null hypothesis: the average for France is different than the average for Thailand.

3.9 Categorical explanatory variables

What if France was selected as the reference category?

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="England")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    23.333      1.707 13.671        19.696        26.971  0.000
x1 -13.333      2.414 -5.524       -18.478        -8.189  0.000
x2  -9.167      2.414 -3.798       -14.311        -4.022  0.002

$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

What if France was selected as the reference category?

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="England")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    23.333      1.707 13.671        19.696        26.971  0.000
x1 -13.333      2.414 -5.524       -18.478        -8.189  0.000
x2  -9.167      2.414 -3.798       -14.311        -4.022  0.002

$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
```

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1-b_2)/SE(b_1- b_2)$

= -3.798

p-value = 2*(1-pt(abs(-3.798), n-k))

= 0.002

We reject the null hypothesis: the average for France is different than the average for Thailand.

3.9 Categorical explanatory variables

What if France was selected as the reference category?

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1 - b_2)/SE(b_1 - b_2)$

= -3.798

p-value = 2*(1-pt(abs(-3.798), n-k))

= 0.002

We reject the null hypothesis: the average for France is different than the average for Thailand.

```
> # We re-code country...
> # with binary "dummy" variables...
> x1 <- as.numeric(country=="England")
> x2 <- as.numeric(country=="Thailand")
>
> # ...assemble the design matrix...
> myX<-cbind(1, x1, x2)
> # ...and use linear regression:
> linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1        x2
    2.912963 -2.912963 -2.912963
x1 -2.912963  5.825926  2.912963
x2 -2.912963  2.912963  5.825926
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    23.333      1.707 13.671        19.696        26.971  0.000
x1 -13.333      2.414 -5.524       -18.478        -8.189  0.000
x2  -9.167      2.414 -3.798       -14.311        -4.022  0.002

$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

What about two categorical covariates?

3.9 Categorical explanatory variables

What about two categorical covariates?

Consider country (3 categories) and gender (2 categories).

3.9 Categorical explanatory variables

What about two categorical covariates?

```
> ##  What if there is another categorical covariate, e.g. gender?
> gender<- c("M","F","M","F","M","F","F","F",
+ "M","M","M","M","M","M","M","F","F","M")
>
> # We re-code with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
> x3 <- as.numeric(gender=="M")
>
> myX<-cbind(1, x1, x2, x3)
>
> # ...and use linear regression:
> mod1<-linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1            x2            x3
    2.493971 -1.7640286 -1.520714e+00 -1.459886e+00
x1 -1.764029  3.1022571  1.520714e+00  3.649714e-01
x2 -1.520714  1.5207143  3.041429e+00 -1.620799e-16
x3 -1.459886  0.3649714 -6.483195e-16  2.189829e+00
> mod1
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    13.787      1.579  8.730        10.400        17.174  0.000
x1  12.387      1.761  7.033         8.609        16.164  0.000
x2   4.167      1.744  2.389         0.426         7.907  0.032
x3  -5.680      1.480 -3.838        -8.854        -2.506  0.002

$SStable
  SS_Total SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1    820.5 127.74  9.124        3.021 0.844 0.811     25.308          0
```

3.9 Categorical explanatory variables

What about two categorical covariates?

```
> ##  What if there is another categorical covariate, e.g. gender?
> gender<- c("M","F","M","F","M","F","F","F",
+ "M","M","M","M","M","M","M","F","F","M")
>
> # We re-code with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
> x3 <- as.numeric(gender=="M")
>
> myX<-cbind(1, x1, x2, x3)
>
> # ...and use linear regression:
> mod1<-linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1            x2            x3
    2.493971 -1.7640286 -1.520714e+00 -1.459886e+00
x1 -1.764029  3.1022571  1.520714e+00  3.649714e-01
x2 -1.520714  1.5207143  3.041429e+00 -1.620799e-16
x3 -1.459886  0.3649714 -6.483195e-16  2.189829e+00
> mod1
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    13.787      1.579  8.730        10.400        17.174  0.000
x1  12.387      1.761  7.033         8.609        16.164  0.000
x2   4.167      1.744  2.389         0.426         7.907  0.032
x3  -5.680      1.480 -3.838        -8.854        -2.506  0.002

$SStable
  SS_Total SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1    820.5 127.74  9.124        3.021 0.844 0.811     25.308          0
```

England is reference category (x1, x2)

“F” is reference category (x3)

3.9 Categorical explanatory variables

What about two categorical covariates?

```
> ##  What if there is another categorical covariate, e.g. gender?
> gender<- c("M","F","M","F","M","F","F","F",
+ "M","M","M","M","M","M","M","F","F","M")
>
> # We re-code with binary "dummy" variables...
> x1 <- as.numeric(country=="France")
> x2 <- as.numeric(country=="Thailand")
> x3 <- as.numeric(gender=="M")
>
> myX<-cbind(1, x1, x2, x3)
>
> # ...and use linear regression:
> mod1<-linear_reg(y, myX)
[1] "variance-covariance matrix for beta:"
                    x1            x2            x3
    2.493971 -1.7640286 -1.520714e+00 -1.459886e+00
x1 -1.764029  3.1022571  1.520714e+00  3.649714e-01
x2 -1.520714  1.5207143  3.041429e+00 -1.620799e-16
x3 -1.459886  0.3649714 -6.483195e-16  2.189829e+00
> mod1
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    13.787      1.579  8.730        10.400        17.174  0.000
x1  12.387      1.761  7.033         8.609        16.164  0.000
x2   4.167      1.744  2.389         0.426         7.907  0.032
x3  -5.680      1.480 -3.838        -8.854        -2.506  0.002

$SStable
  SS_Total SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1    820.5 127.74  9.124        3.021 0.844 0.811     25.308          0
```

England is reference category (x1, x2)

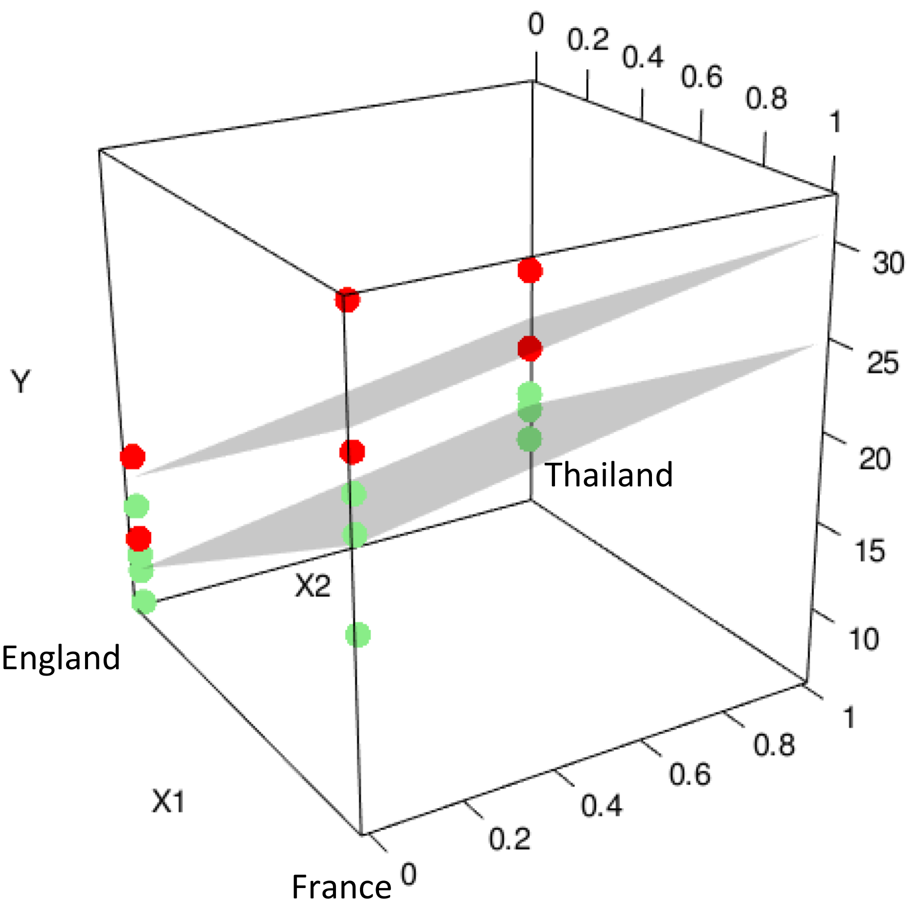
"F" is reference category (x3)

The **model**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

3.9 Categorical explanatory variables

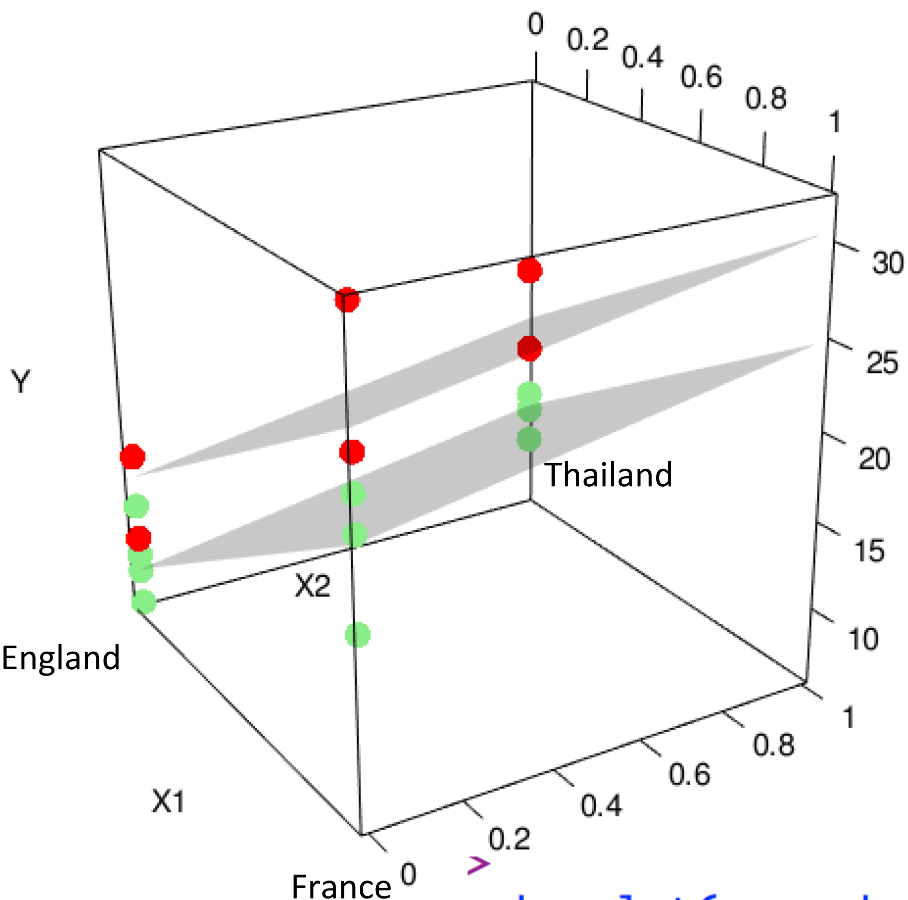
What about two categorical covariates?

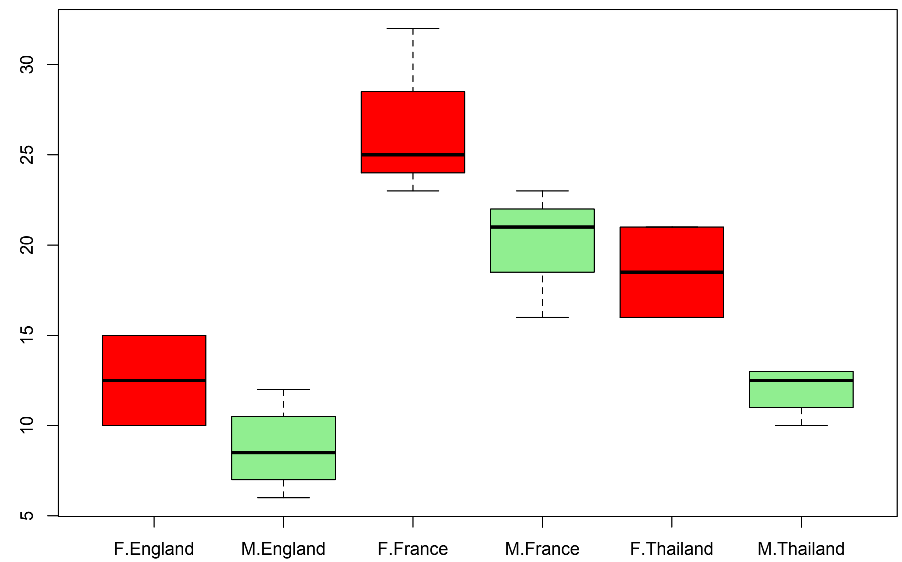



```
> col_codes<-rep(NA,n)
> for(i in 1:n){
+  if(x3[i]==1){col_codes[i]<-"lightgreen"}
+  if(x3[i]==0){col_codes[i]<-"red"}
+  }
>
> plot3d(x1,x2,y, type="p", col= col_codes,
+ xlab="X1", ylab="X2", zlab="Y", site=5,
+ lwd=15, size=12)
>
> fM <- function(x1, x2){  13.787 + 12.387*x1 + 4.167*x2 -5.680}
>
> fF <- function(x1, x2){  13.787 + 12.387*x1 + 4.167*x2}
> my_surface(fM, alpha=.2 )
>
> my_surface(fF, alpha=.2 )
```

3.9 Categorical explanatory variables

What hypotheses can we test?






```
>
> boxplot(y~gender+country, col=c("red", "lightgreen"))
>
```

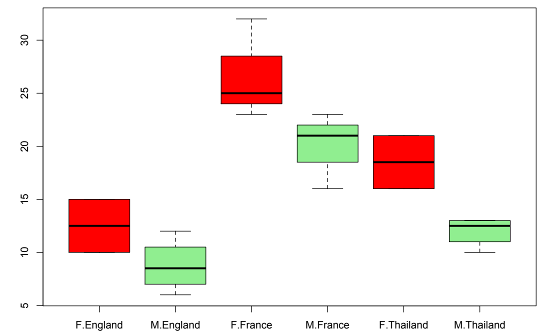
3.9 Categorical explanatory variables

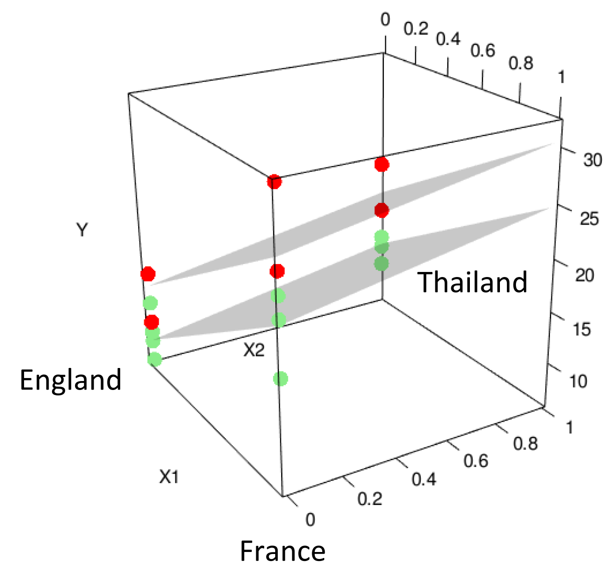
What hypotheses can we test?

Is the average for Males equal to the average for Females, when adjusted for country?

Is the average for Thailand equal to the average for France, when adjusted for gender?

Is the average for Females in England equal to the average of Males in Thailand?





3.9 Categorical explanatory variables

What hypotheses can we test?

Is the average for Males equal to the average for Females, when adjusted for country?

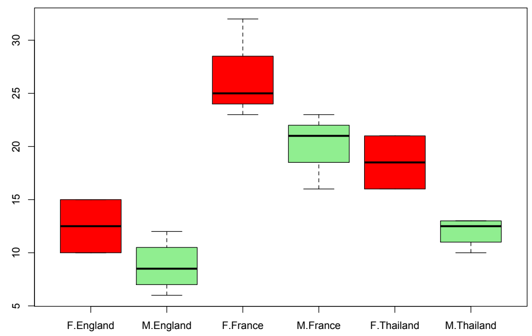
$$H_0 : \beta_3 = 0$$

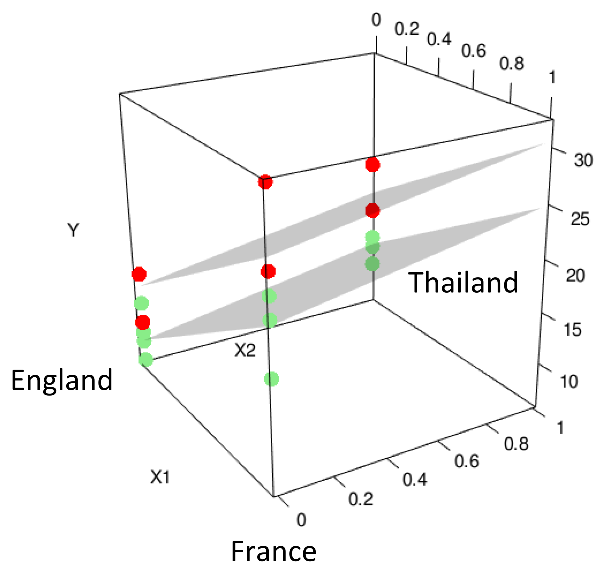
Is the average for Thailand equal to the average for France, when adjusted for gender?

$$H_0 : \beta_1 = \beta_2$$

Is the average for Females in England equal to the average of Males in Thailand?

???

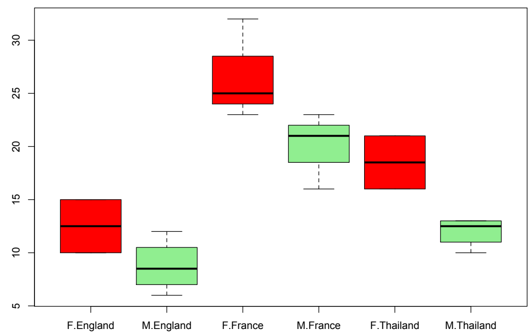




3.9 Categorical explanatory variables

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

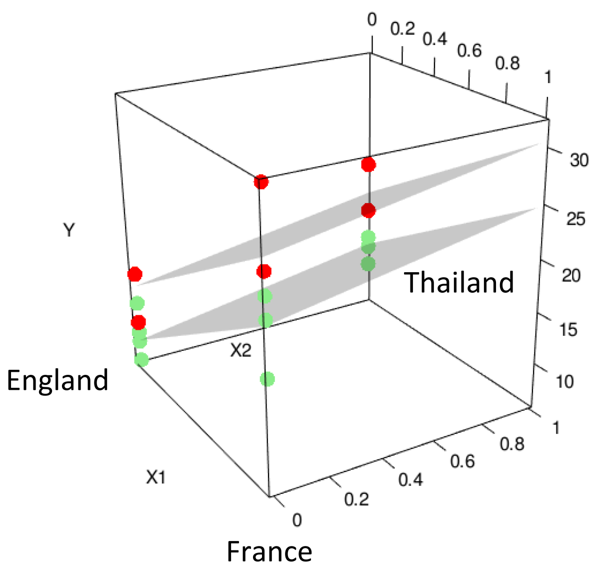


Is the average for Males equal to the average for Females, when adjusted for country?

$H_0 : \beta_3 = 0$

Is the average for Thailand equal to the average for France, when adjusted for gender?

$H_0 : \beta_1 = \beta_2$



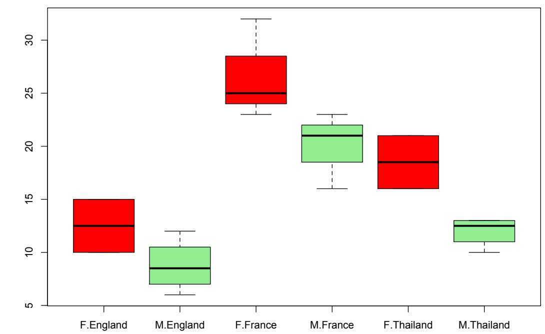
Is the average for Females in England equal to the average of Males in Thailand?

$H_0 : \beta_0 = \beta_0 + \beta_3 + \beta_2$

3.9 Categorical explanatory variables

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

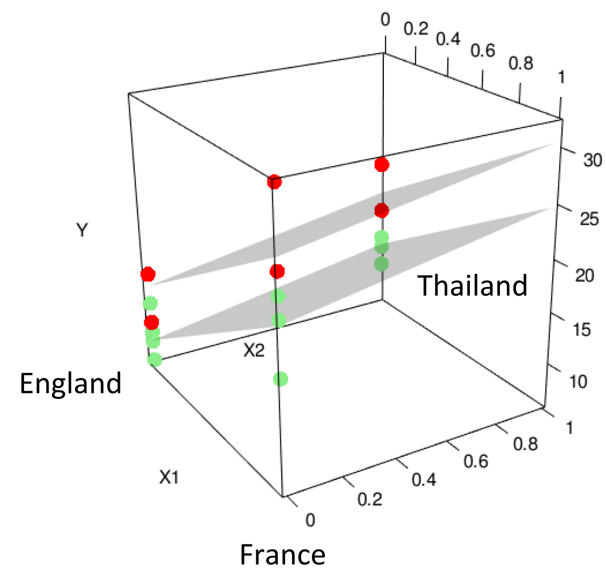


Is the average for Males equal to the average for Females, when adjusted for country?

$H_0 : \beta_3 = 0$

Is the average for Thailand equal to the average for France, when adjusted for gender?

$H_0 : \beta_1 = \beta_2$



Is the average for Females in England equal to the average of Males in Thailand?

$H_0 : \beta_0 = \beta_0 + \beta_3 + \beta_2$

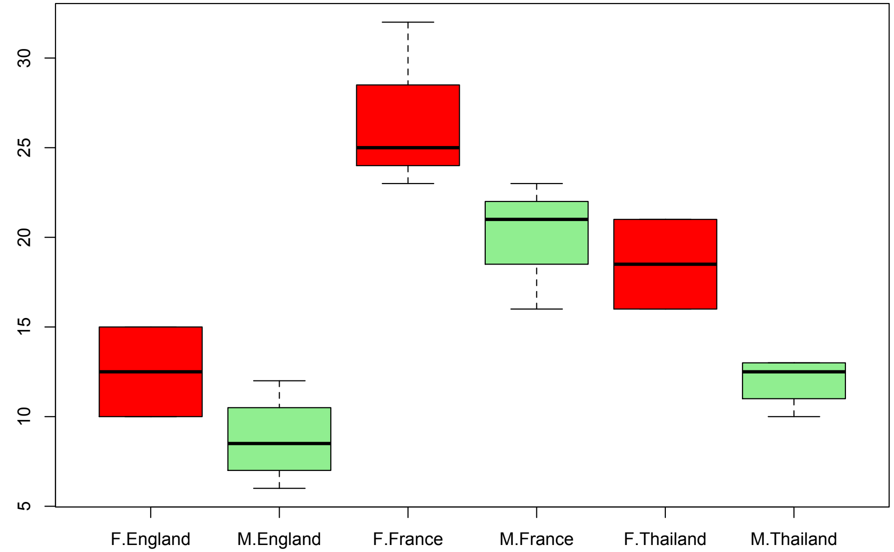
What assumption are we making with this model?

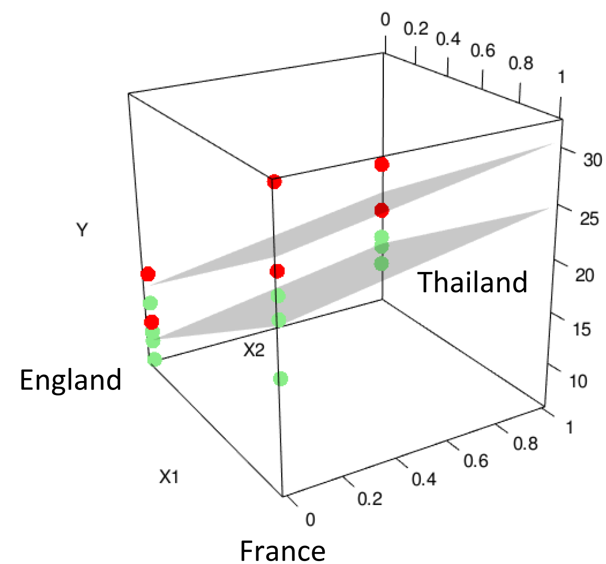
3.9 Categorical explanatory variables

The **model**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

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



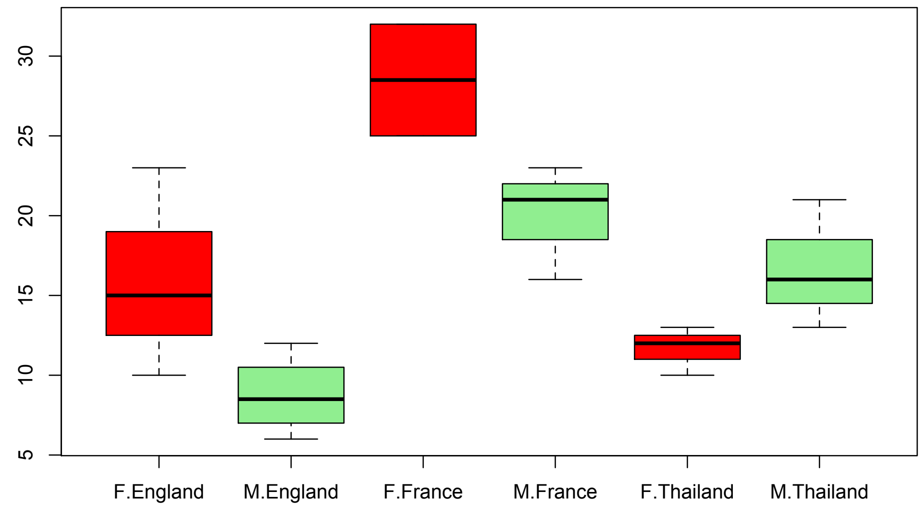


3.9 Categorical explanatory variables

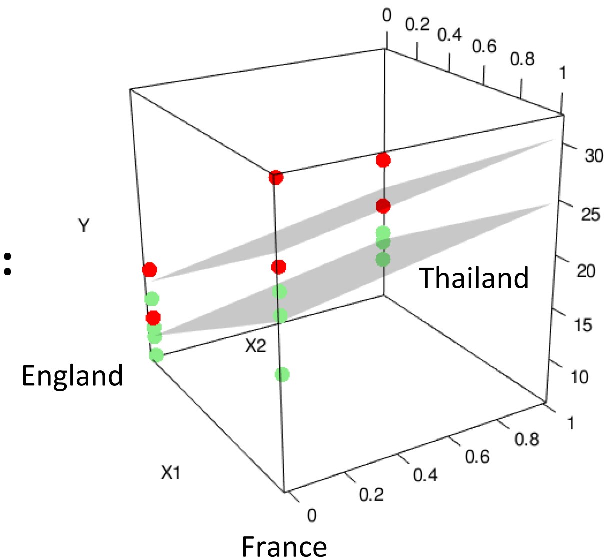
The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

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

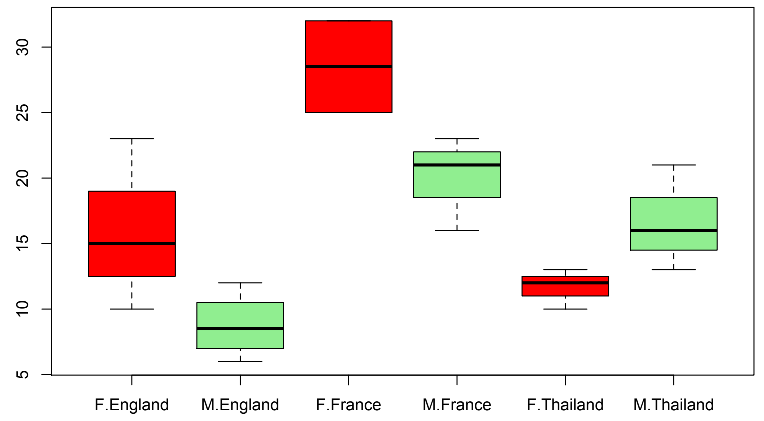


Small changes to the data:
Let's switch the labels for Thailand… and Change one "France" to "England".

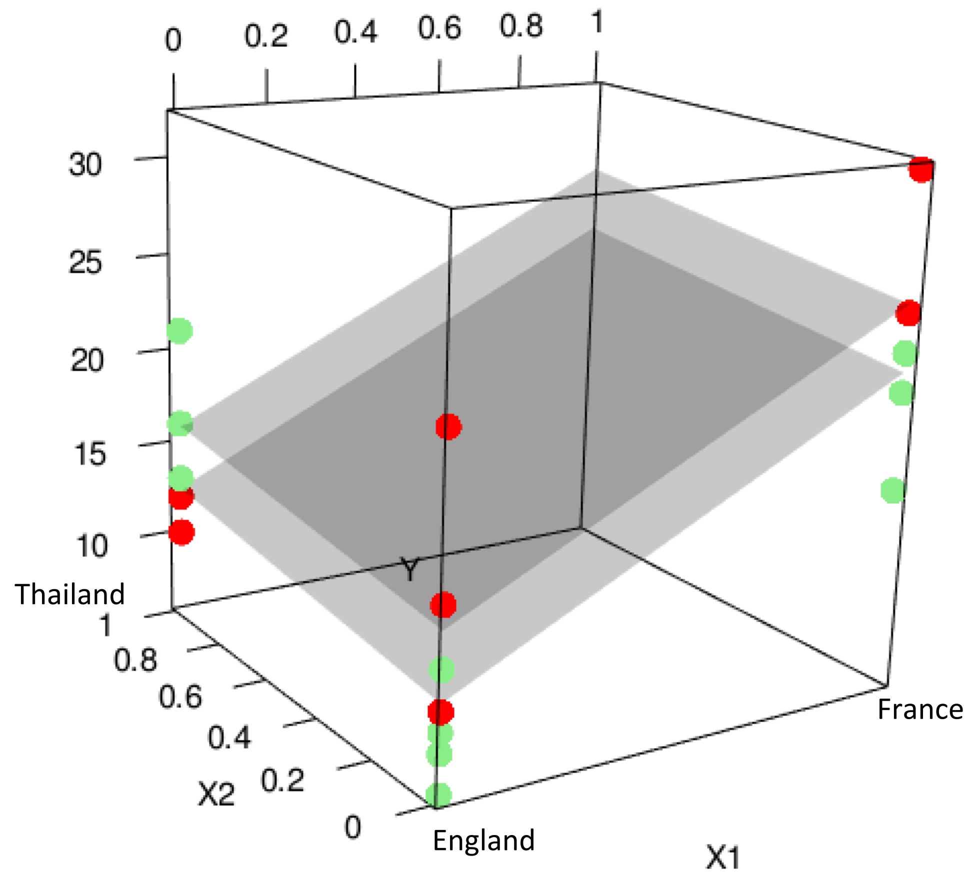


The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$




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

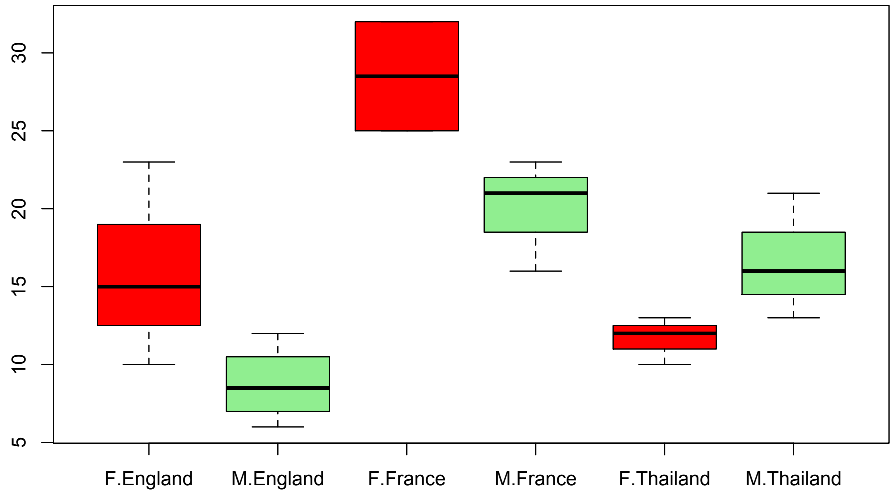


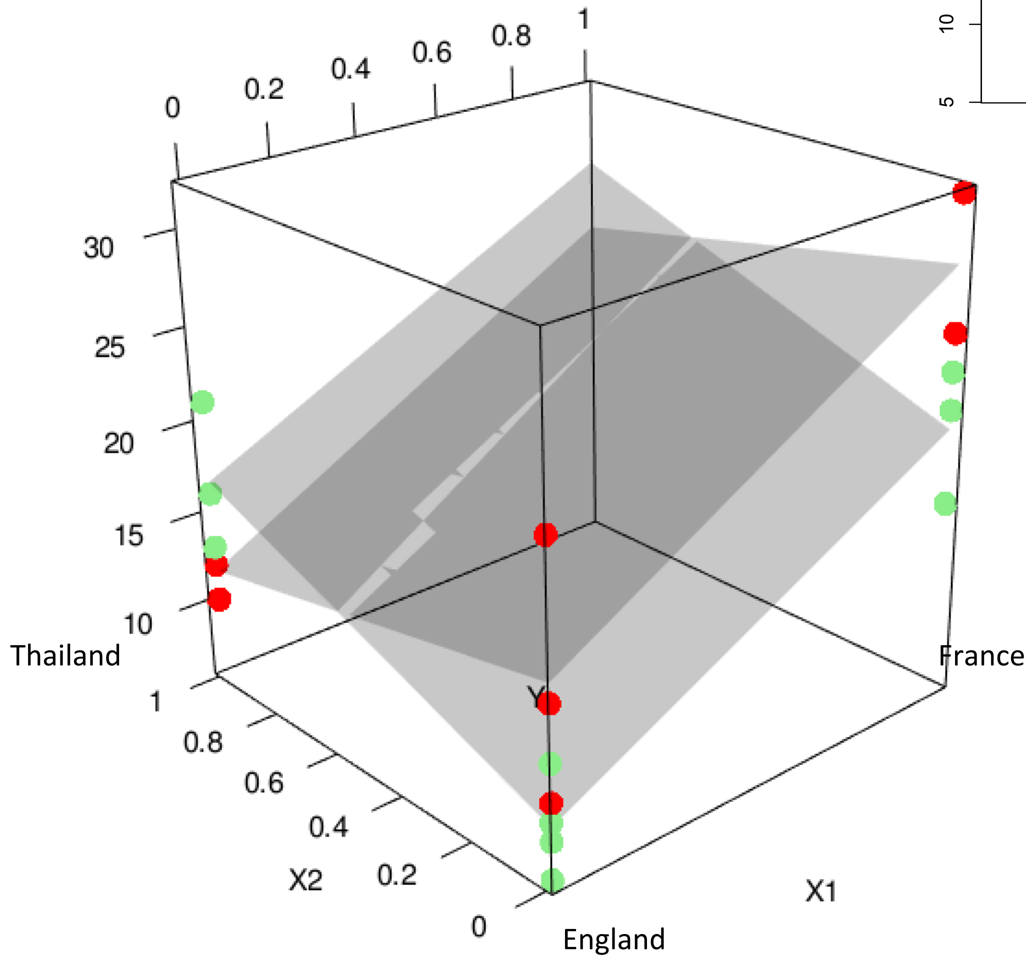
The **model**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

The **model** with an interaction effect:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_3 X_1) + \beta_5 (X_3 X_2)$$





3.9 Categorical explanatory variables

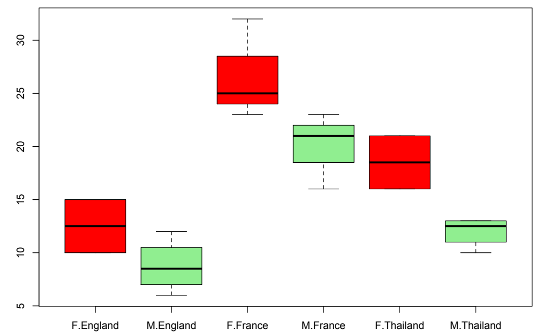
Now let's go back to the model **with the original data** and **without interaction effects**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

3.9 Categorical explanatory variables

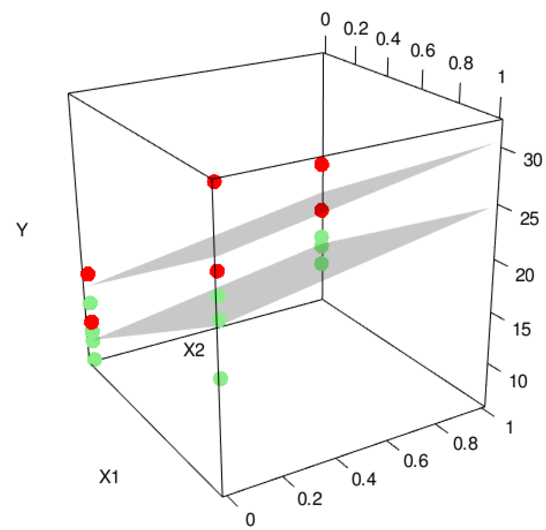
The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$



What about the more general question…

Question: Do all the categories have the same average?



3.9 Categorical explanatory variables

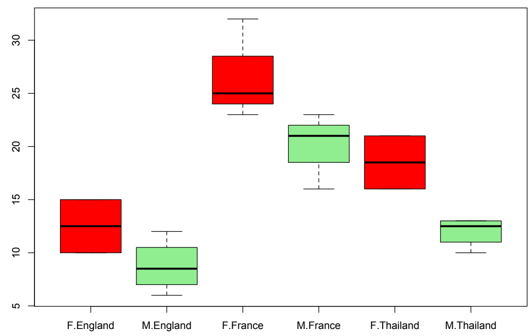
The **model**:

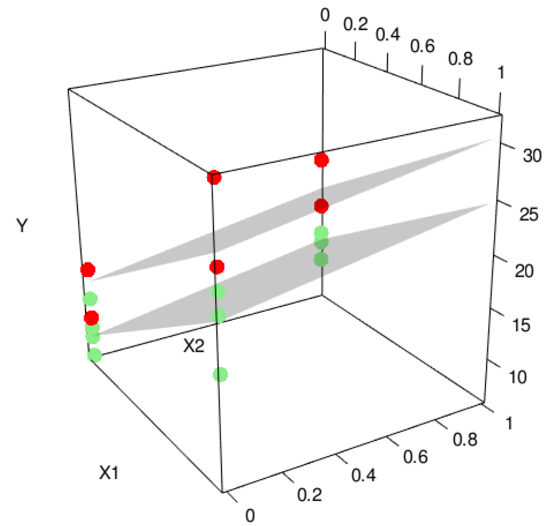
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

What about the more general question…

Question: Do all the categories have the same average?

$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \beta_3 = 0$

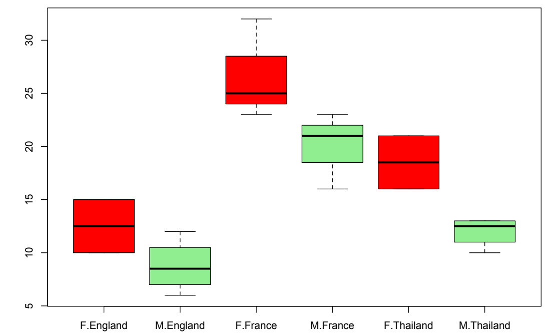




3.9 Categorical explanatory variables

The **model**:

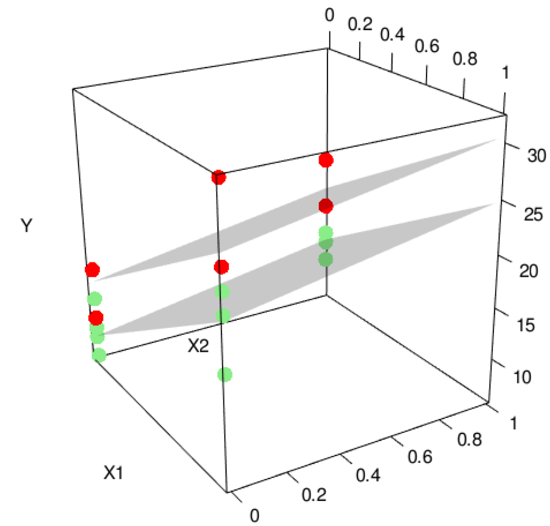
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$



What about the more general question…

Question: Do all the categories have the same average?

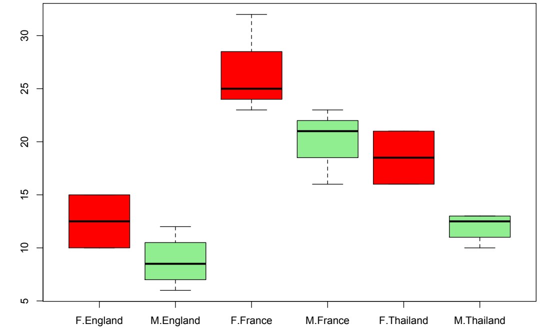
$H_0: \beta_1 = \beta_2 = \beta_3 = 0$



3.9 Categorical explanatory variables

The **model**:

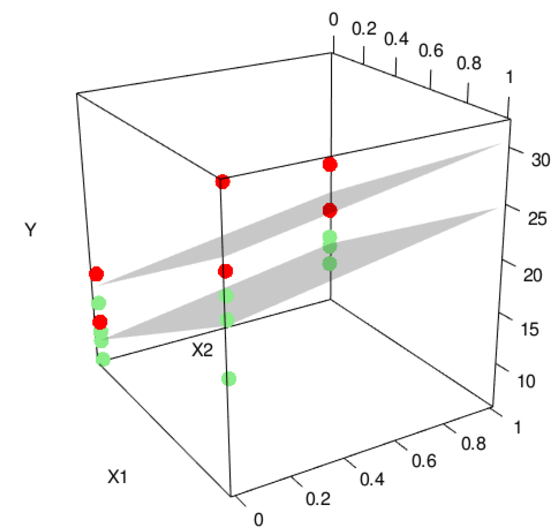
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$



What about the more general question…

Question: Do all the categories have the same average?

Answer: F-test

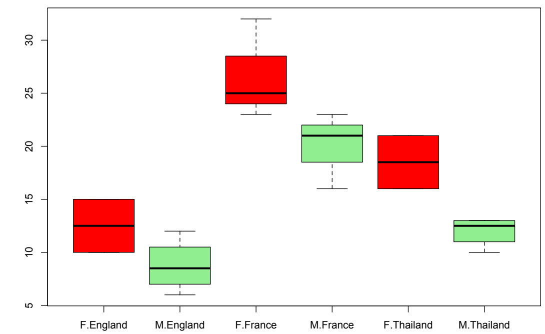



```
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    10.000      1.707  5.859         6.362        13.638  0.000
x1  13.333      2.414  5.524         8.189        18.478  0.000
x2   4.167      2.414  1.726        -0.978         9.311  0.105

$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

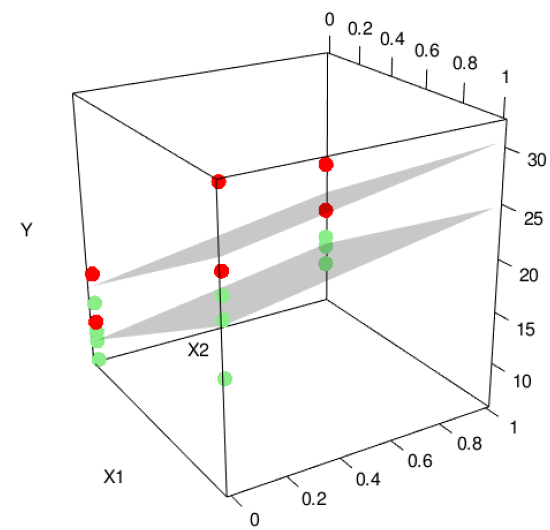
The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$



What about the more general question…

Question: Do all the categories have the same average?



Answer: F-test

	betahat	se_betahat	tratio	ci_lower_beta	ci_upper_beta	pvalue
	10.000	1.707	5.859	6.362	13.638	0.000
x1	13.333	2.414	5.524	8.189	18.478	0.000
x2	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

The **model** with **an interaction effect**:

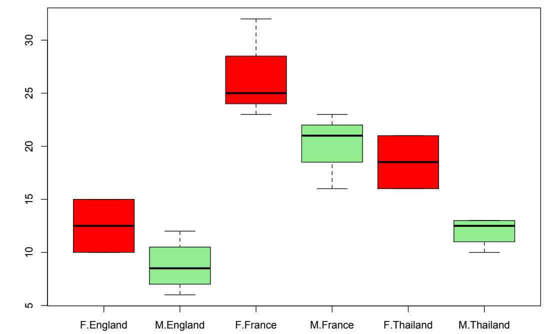
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_3 X_1) + \beta_5 (X_3 X_2)$$

What about the more general question…

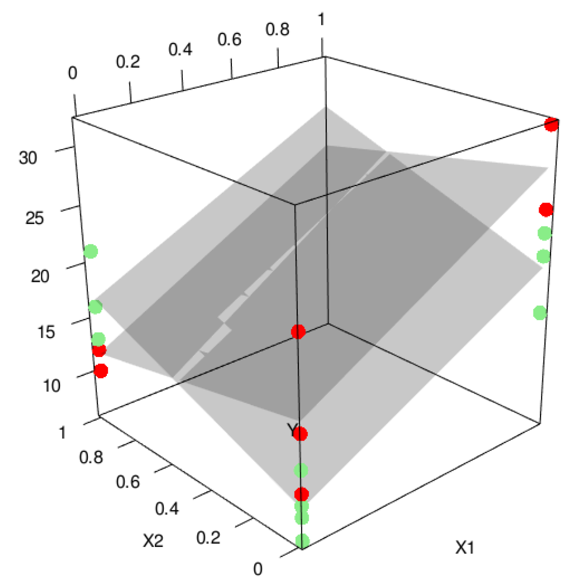
Question: Do all the categories have the same average?

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

Answer: F-test



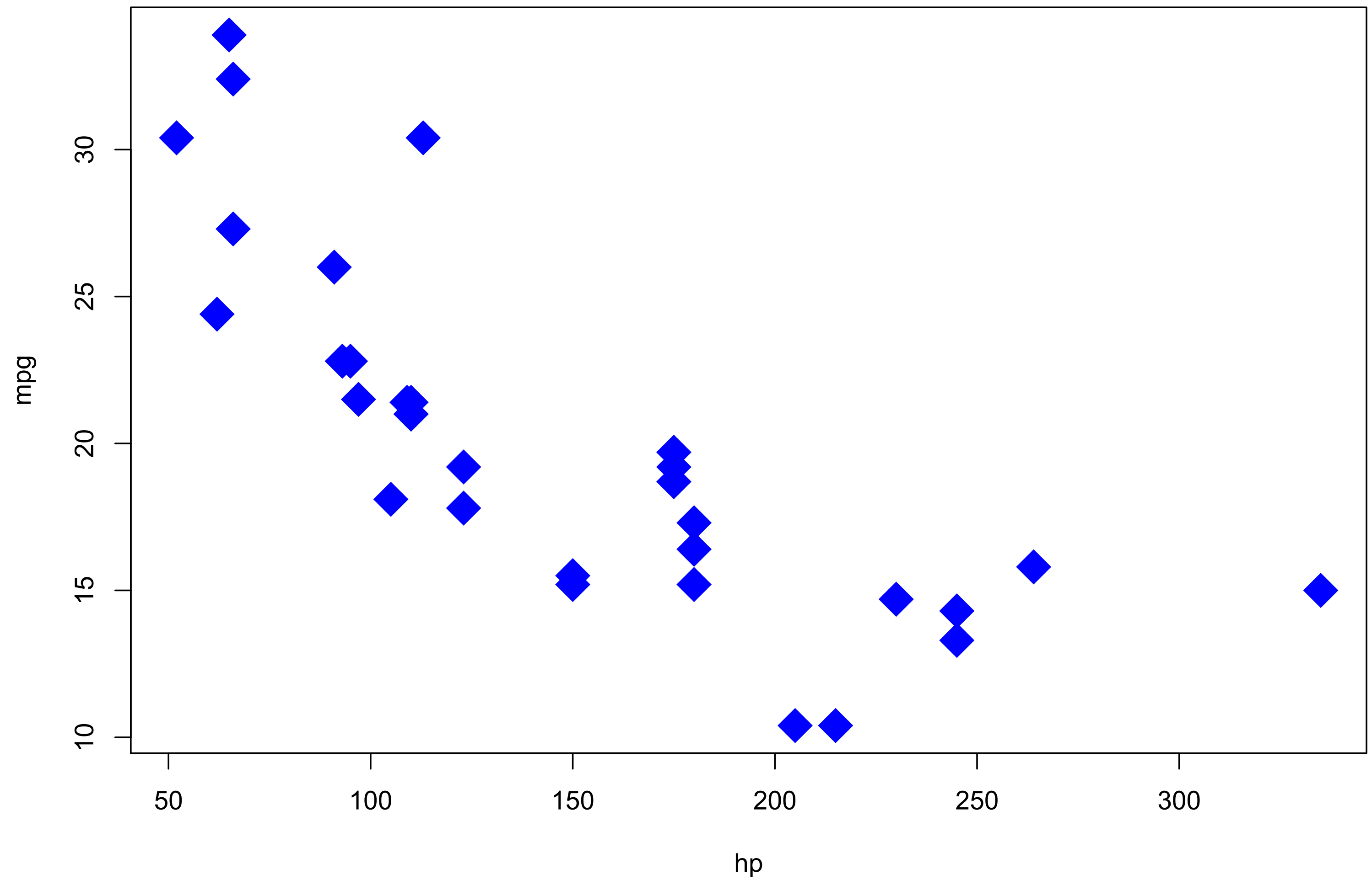

```
> mod2
$coeftable
   betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
    16.000      2.313  6.918        10.961        21.039  0.000
x1  12.500      3.657  3.418         4.532        20.468  0.005
x2  -4.333      3.271 -1.325       -11.460         2.793  0.210
x3  -7.250      3.060 -2.370       -13.916        -0.584  0.035
x4  -1.250      4.768 -0.262       -11.639         9.139  0.798
x5  12.250      4.479  2.735         2.491        22.009  0.018

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1    820.5 192.583 16.049        4.006 0.765 0.667      7.825      0.002
```



- Questions?

Quadratic Terms

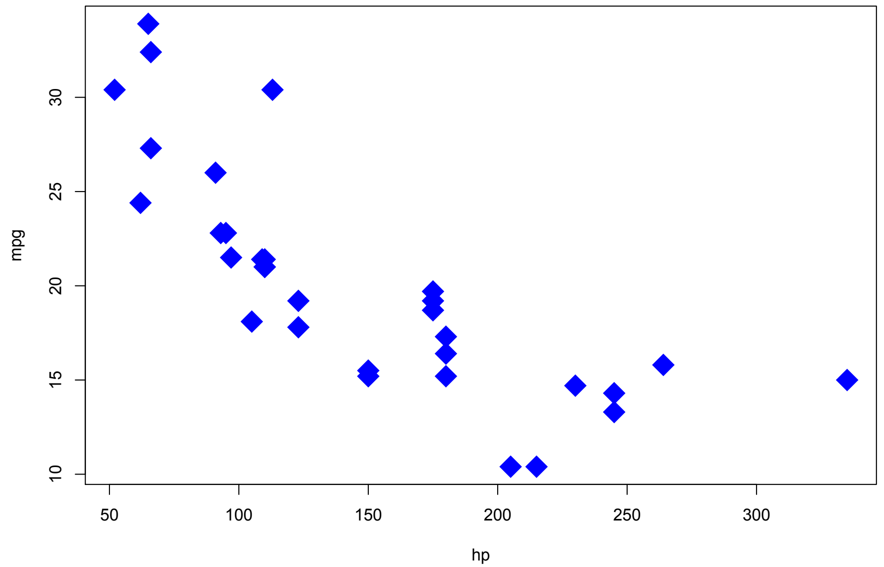
```
> data(mtcars)
> plot(mpg ~ hp, data=mtcars, cex=3, pch=18, col="blue")
```



Quadratic Terms

The **model**:

$Y = \beta_0 + \beta_1 X_1$

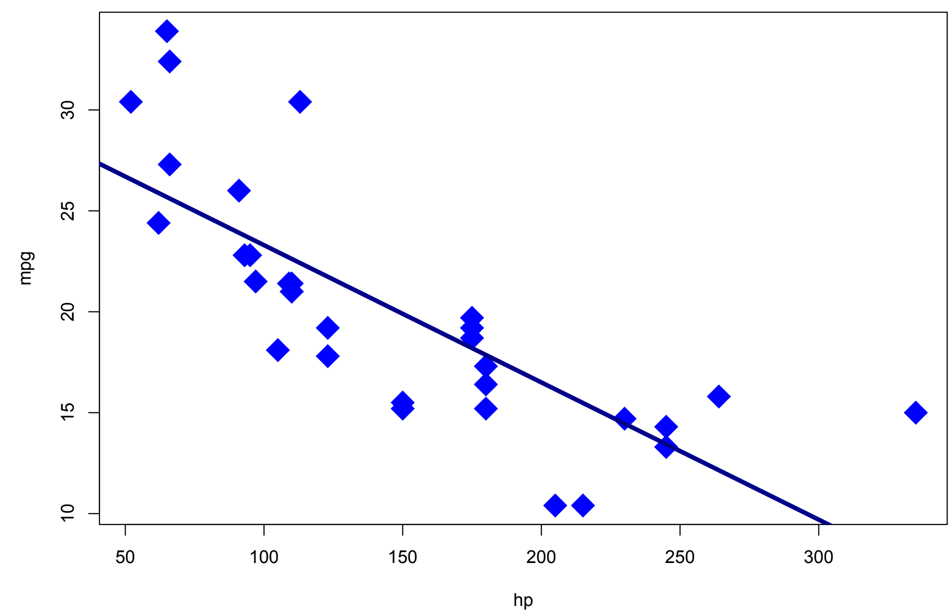



```
> # define outcome variable and design matrix:
> carsy  <- mtcars$mpg
> carsX  <- cbind(1, mtcars$hp)
>
> # ...and use linear regression:
> mod1<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
            [,1]          [,2]
[1,]  2.66969767 -0.0150208454
[2,] -0.01502085  0.0001024003
> carsmod<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
            [,1]          [,2]
[1,]  2.66969767 -0.0150208454
[2,] -0.01502085  0.0001024003
> carsmod
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  30.099      1.634 18.421        26.762        33.436      0
2  -0.068      0.010 -6.742        -0.089        -0.048      0

$SStable
  SS_Total   SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 447.674 14.922        3.863 0.602 0.589      45.46          0
```

Quadratic Terms

The **model**:

$$Y = \beta_0 + \beta_1 X_1$$

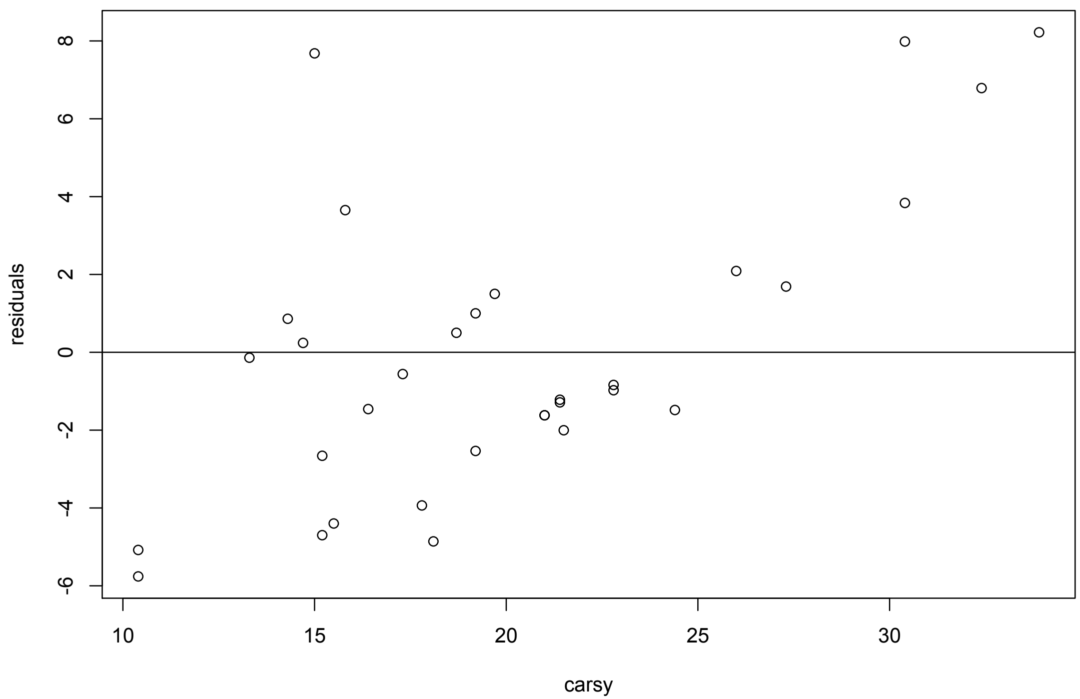



```
> # define outcome variable and design matrix:
> carsy  <- mtcars$mpg
> carsX  <- cbind(1, mtcars$hp)
>
> # ...and use linear regression:
> mod1<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
            [,1]          [,2]
[1,]  2.66969767 -0.0150208454
[2,] -0.01502085  0.0001024003
> carsmod<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
            [,1]          [,2]
[1,]  2.66969767 -0.0150208454
[2,] -0.01502085  0.0001024003
> carsmod
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  30.099      1.634 18.421        26.762        33.436      0
2  -0.068      0.010 -6.742        -0.089        -0.048      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 447.674 14.922        3.863 0.602 0.589      45.46          0
> abline(30.099,-0.068, col="darkblue", lwd=4)
```

Quadratic Terms

Plot the residuals vs. y:

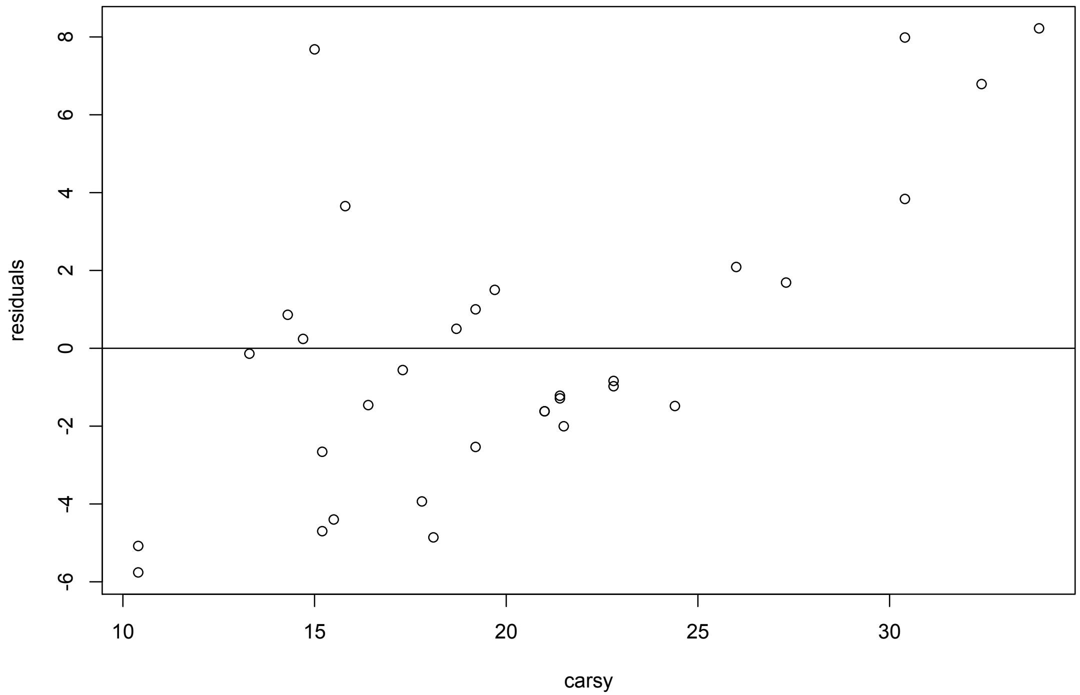
```
> yhat<-carsX%*%c(30.099, -0.068)
>
> residuals<-carsy-yhat
> plot(residuals~carsy)
> abline(0,0)
>
```



Quadratic Terms

Plot the residuals vs. y:

```
> yhat<-carsX%*%c(30.099, -0.068)
>
> residuals<-carsy-yhat
> plot(residuals~carsy)
> abline(0,0)
>
```



Not very good!

Quadratic Terms

```
> # add a "quadratic term" to the design matrix:
> carsX  <- cbind(1, mtcars$hp, mtcars$hp^2)
> carsX
     [,1] [,2]  [,3]
[1,]    1  110 12100
[2,]    1  110 12100
[3,]    1   93  8649
[4,]    1  110 12100
[5,]    1  175 30625
[6,]    1  105 11025
[7,]    1  245 60025
```

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1^2)$

Quadratic Terms

```
> # add a "quadratic term" to the design matrix:
> carsX  <- cbind(1, mtcars$hp, mtcars$hp^2)
> carsX
     [,1] [,2]  [,3]
 [1,]    1  110 12100
 [2,]    1  110 12100
 [3,]    1   93  8649
 [4,]    1  110 12100
 [5,]    1  175 30625
 [6,]    1  105 11025
```

```
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
              [,1]          [,2]          [,3]
[1,]  7.5117623243 -9.139336e-02  2.374438e-04
[2,] -0.0913933643  1.216884e-03 -3.341172e-06
[3,]  0.0002374438 -3.341172e-06  9.691325e-09
> carsmodQ
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  40.409      2.741 14.744        34.804        46.015      0
2  -0.213      0.035 -6.115        -0.285        -0.142      0
3   0.000      0.000  4.275         0.000         0.001      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0
```

Quadratic Terms

```
> # add a "quadratic term" to the design matrix:
> carsX  <- cbind(1, mtcars$hp, mtcars$hp^2)
> carsX
     [,1] [,2]  [,3]
[1,]    1  110 12100
[2,]    1  110 12100
[3,]    1   93  8649
[4,]    1  110 12100
[5,]    1  175 30625
[6,]    1  105 11025
[7,]    1  245 60025

> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
              [,1]          [,2]          [,3]
[1,]  7.5117623243 -9.139336e-02  2.374438e-04
[2,] -0.0913933643  1.216884e-03 -3.341172e-06
[3,]  0.0002374438 -3.341172e-06  9.691325e-09
> carsmodQ
$coeftable
 betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
  40.409      2.741 14.744        34.804        46.015      0
  -0.213      0.035 -6.115        -0.285        -0.142      0
   0.000      0.000  4.275         0.000         0.001      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0
```

What happened here?

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1^2)$

Quadratic Terms

```
> # scale by a factor of 100 and add a "quadratic term":
> carsX  <- cbind(1, mtcars$hp/100, (mtcars$hp/100)^2)
>
> # linear regression model
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
          [,1]      [,2]       [,3]
[1,]  7.511762 -9.139336  2.3744377
[2,] -9.139336 12.168840 -3.3411717
[3,]  2.374438 -3.341172  0.9691325
> carsmodQ
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  40.409      2.741 14.744        34.804        46.015      0
2 -21.331      3.488 -6.115       -28.465       -14.196      0
3   4.208      0.984  4.275         2.195         6.222      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0
```

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_1^2)$

with: X_1 = hp/100

Quadratic Terms

```
> # scale by a factor of 100 and add a "quadratic term":
> carsX  <- cbind(1, mtcars$hp/100, (mtcars$hp/100)^2)
>
> # linear regression model
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
          [,1]      [,2]       [,3]
[1,]  7.511762 -9.139336  2.3744377
[2,] -9.139336 12.168840 -3.3411717
[3,]  2.374438 -3.341172  0.9691325
> carsmodQ
$coeftable
  betahat  e_betahat tratio ci_lower_beta ci_upper_beta pvalue
   40.409      2.741 14.744        34.804        46.015      0
  -21.331      3.488 -6.115       -28.465       -14.196      0
    4.208      0.984  4.275         2.195         6.222      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0
```

Much better!!

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1^2)$

with: X_1 = hp/100

Quadratic Terms

```
> # linear regression model
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
          [,1]      [,2]       [,3]
[1,]  7.511762 -9.139336  2.3744377
[2,] -9.139336 12.168840 -3.3411717
[3,]  2.374438 -3.341172  0.9691325
> carsmodQ
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  40.409      2.741 14.744        34.804        46.015      0
2 -21.331      3.488 -6.115       -28.465       -14.196      0
3   4.208      0.984  4.275         2.195         6.222      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0

> plot(mpg ~ hp, data=mtcars, cex=3, pch=18, col="blue")
> yhat<-carsX%*%c(40.409,-21.331, 4.208)
> points(mtcars$hp, yhat, col="orange", pch=20, cex=3)
```

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1^2)$

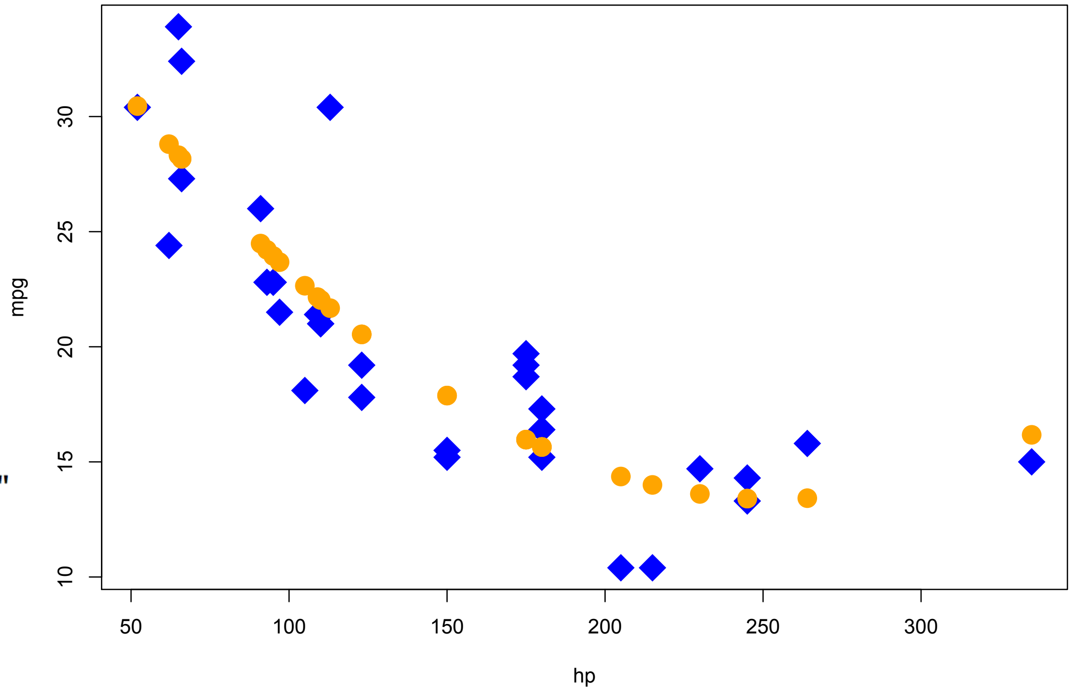
with: $X_1 = hp/100$

Quadratic Terms

The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1^2)$

with: X_1 = hp/100

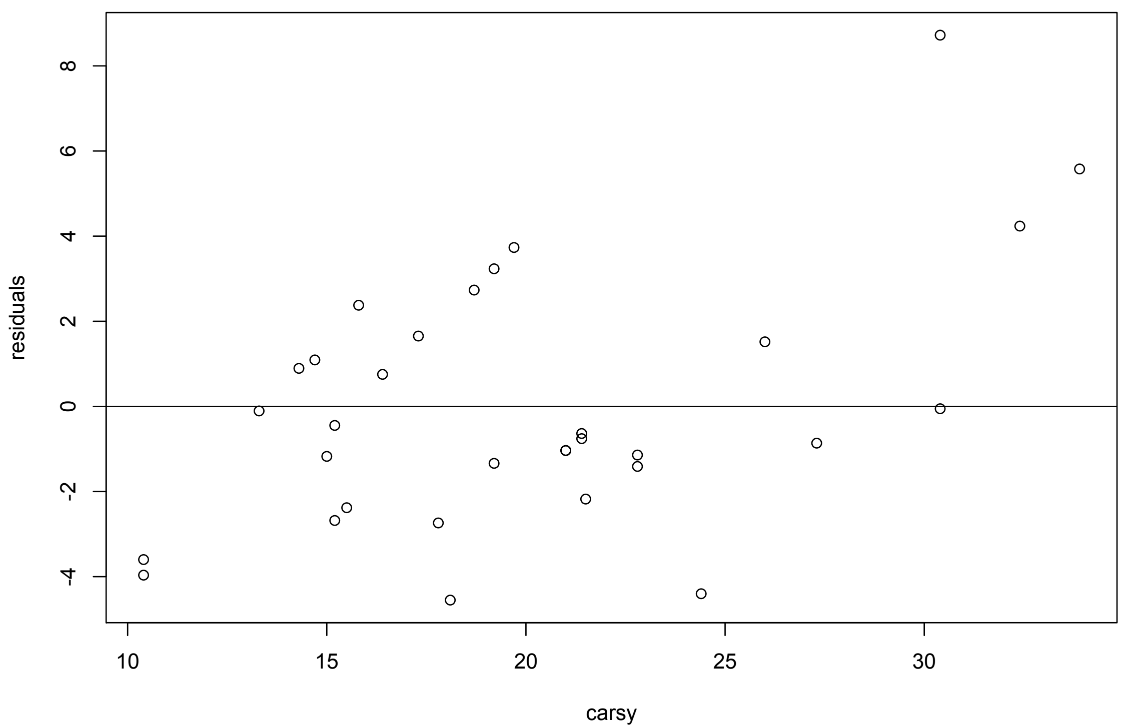



```
> # linear regression model
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
          [,1]      [,2]       [,3]
[1,]  7.511762 -9.139336  2.3744377
[2,] -9.139336 12.168840 -3.3411717
[3,]  2.374438 -3.341172  0.9691325
> carsmodQ
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  40.409      2.741 14.744        34.804        46.015      0
2 -21.331      3.488 -6.115       -28.465       -14.196      0
3   4.208      0.984  4.275         2.195         6.222      0

$SStable
  SS_Total   SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47         3.077 0.756 0.739     44.953          0

> plot(mpg ~ hp, data=mtcars, cex=3, pch=18, col="blue")
> yhat<-carsX%*%c(40.409,-21.331, 4.208)
> points(mtcars$hp, yhat, col="orange", pch=20, cex=3)
```

Quadratic Terms

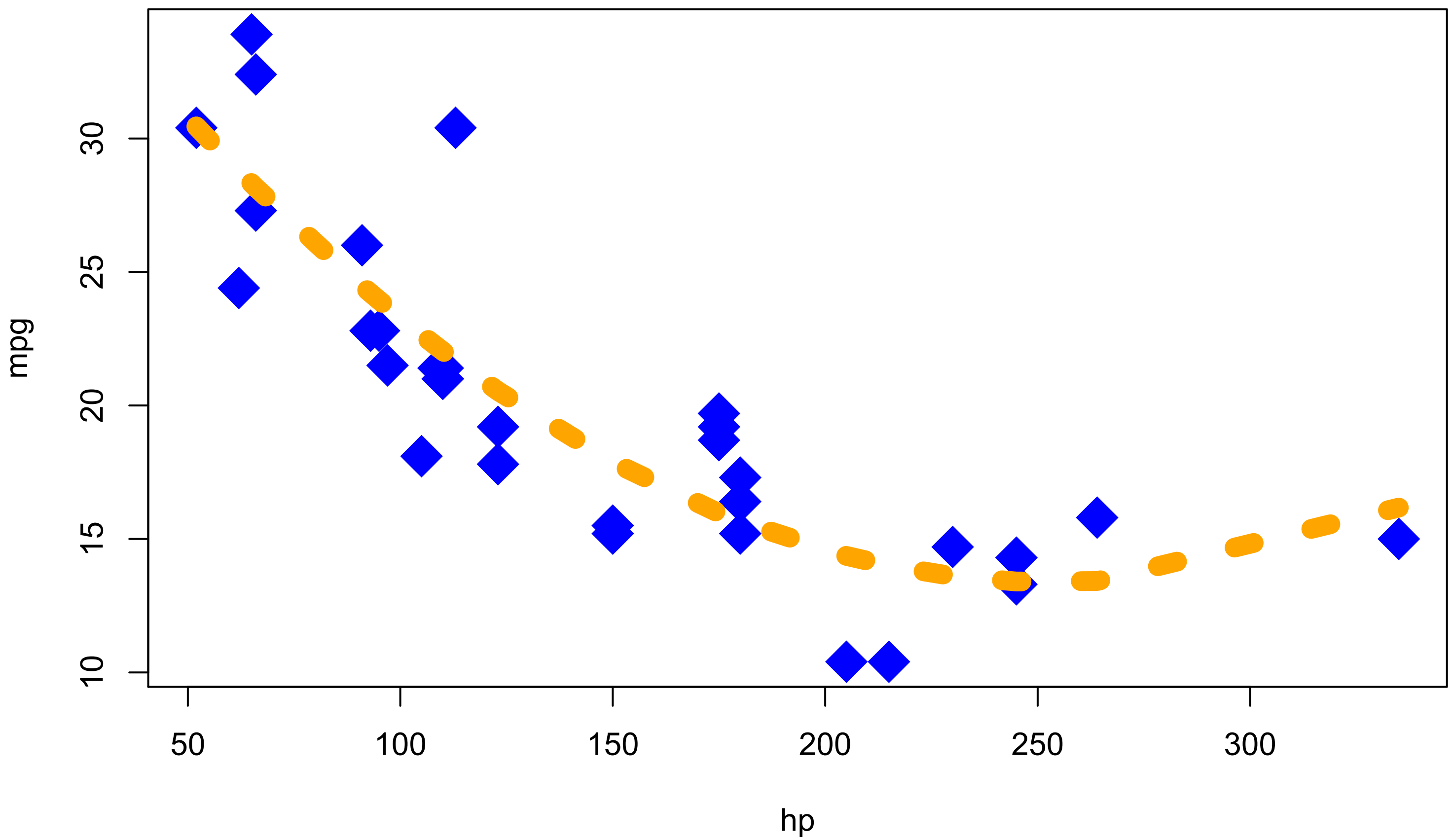
Plot the residuals vs. y:

```
> residuals<-carsy-yhat
> plot(residuals~carsy)
> abline(0,0)
```

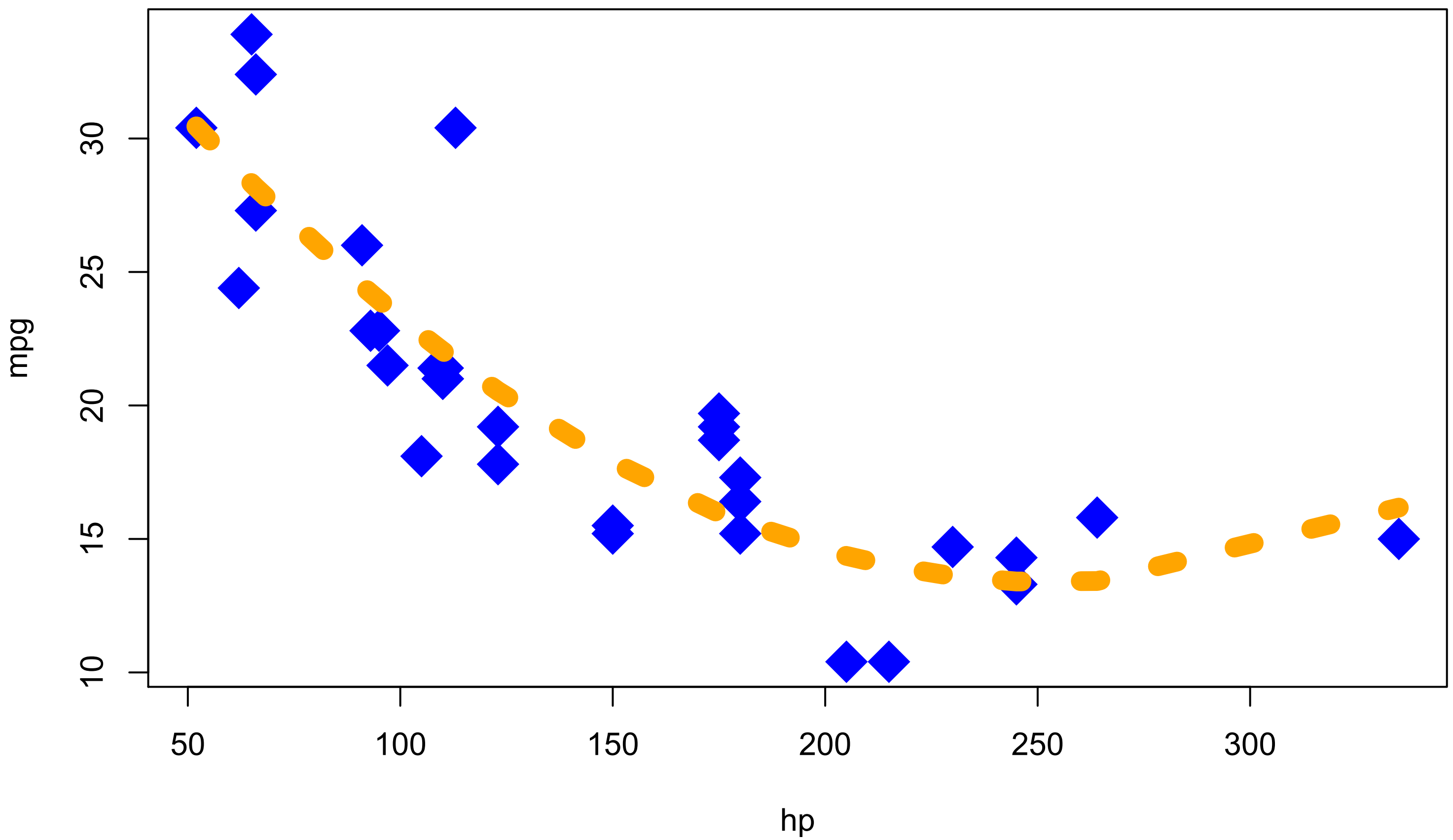


Not perfect, but better!

Quadratic Terms

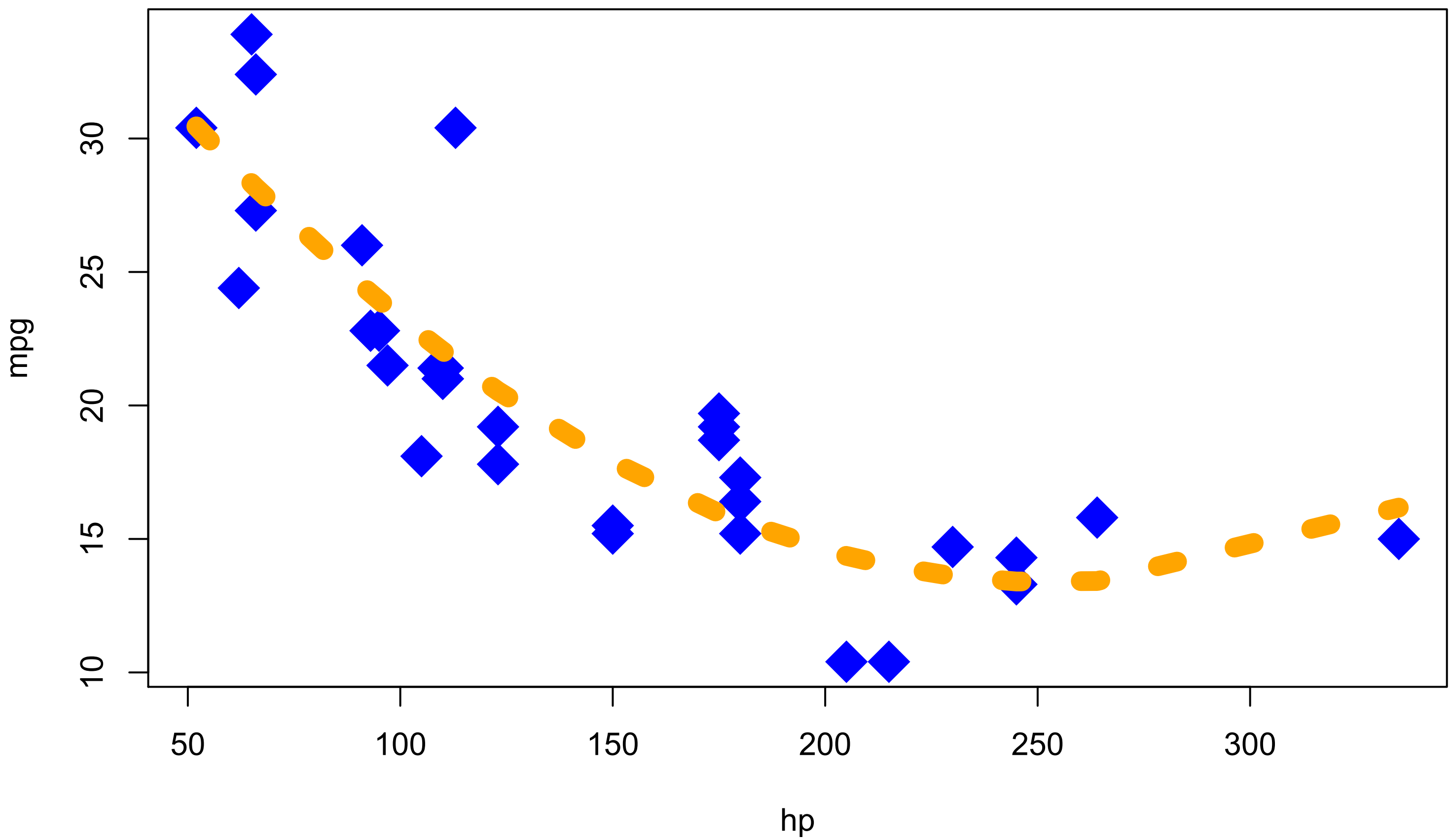


Quadratic Terms



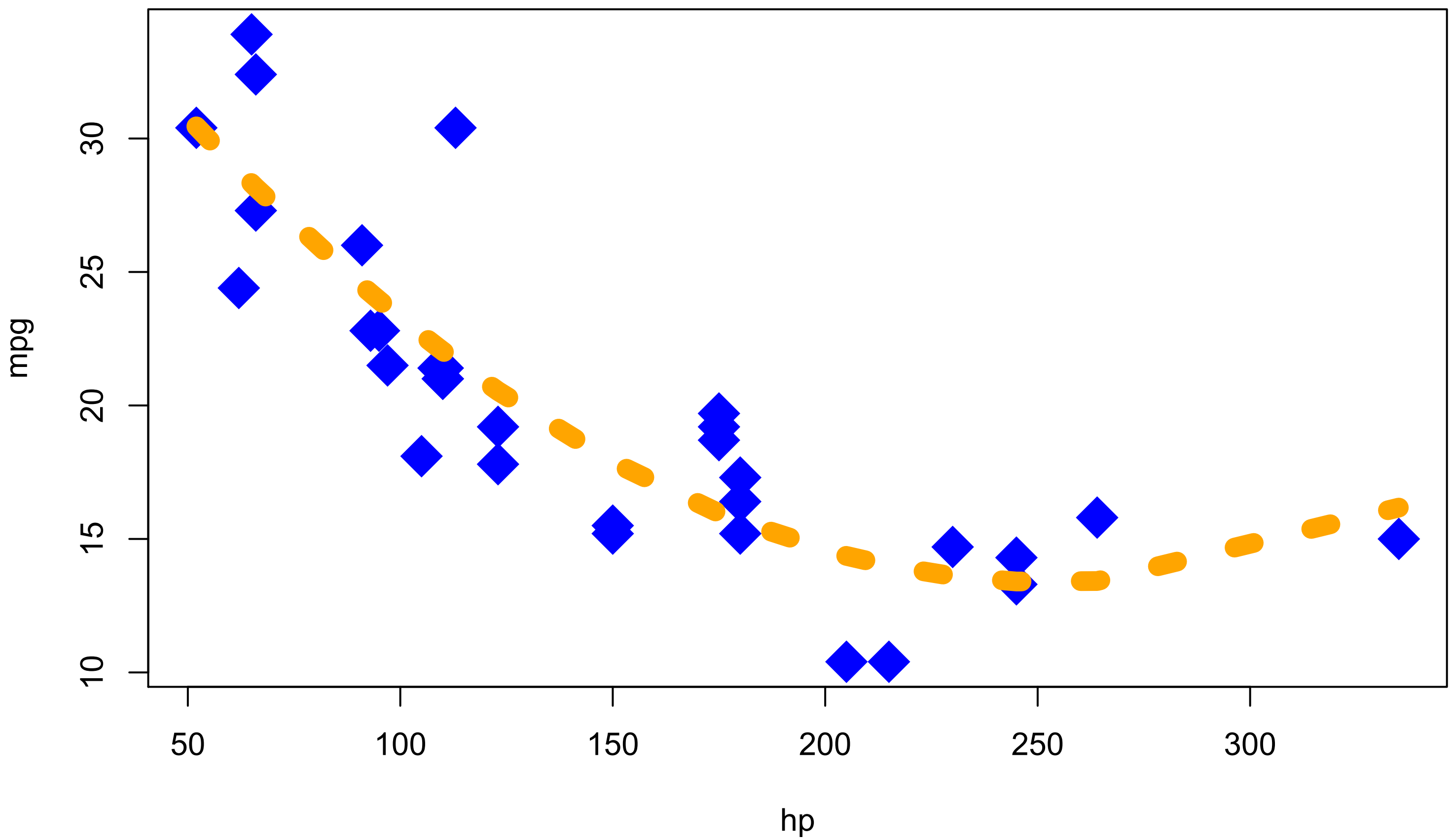
Wait a minute…

Quadratic Terms



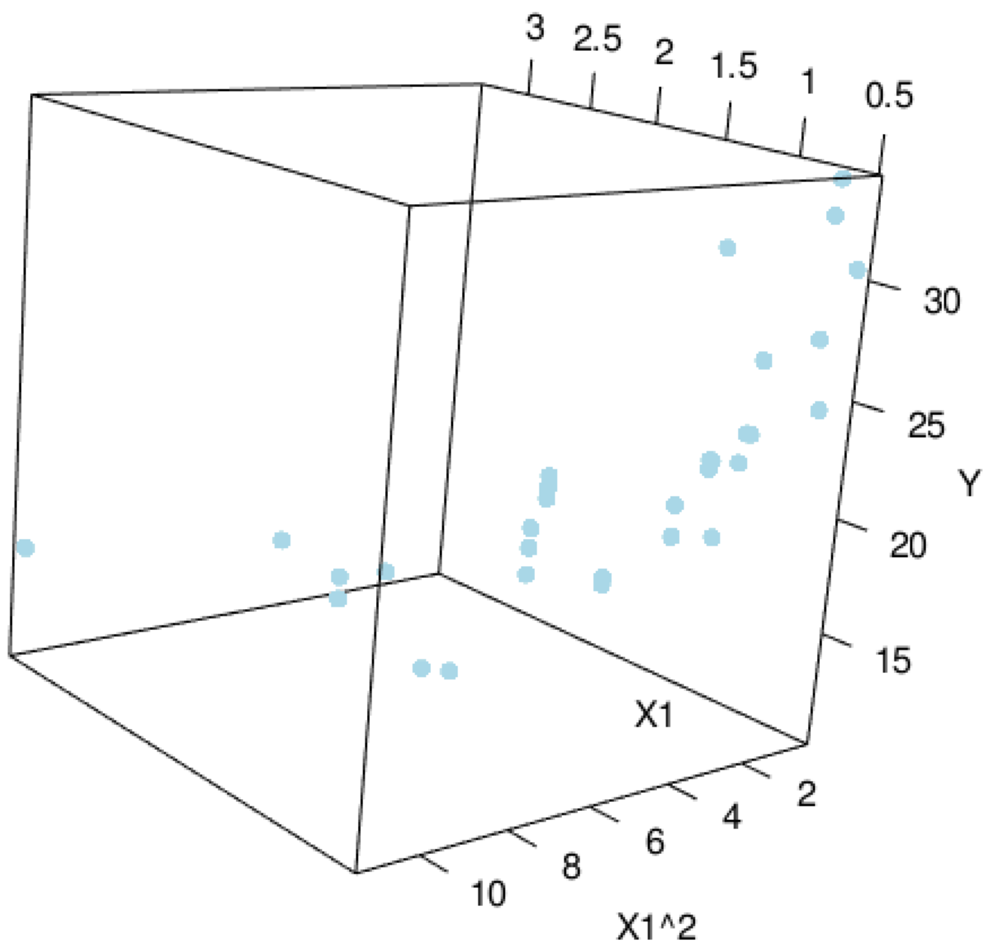
Wait a minute... isn't this supposed to be ...

Quadratic Terms



Wait a minute… isn't this supposed to be ***LINEAR*** regression?!?

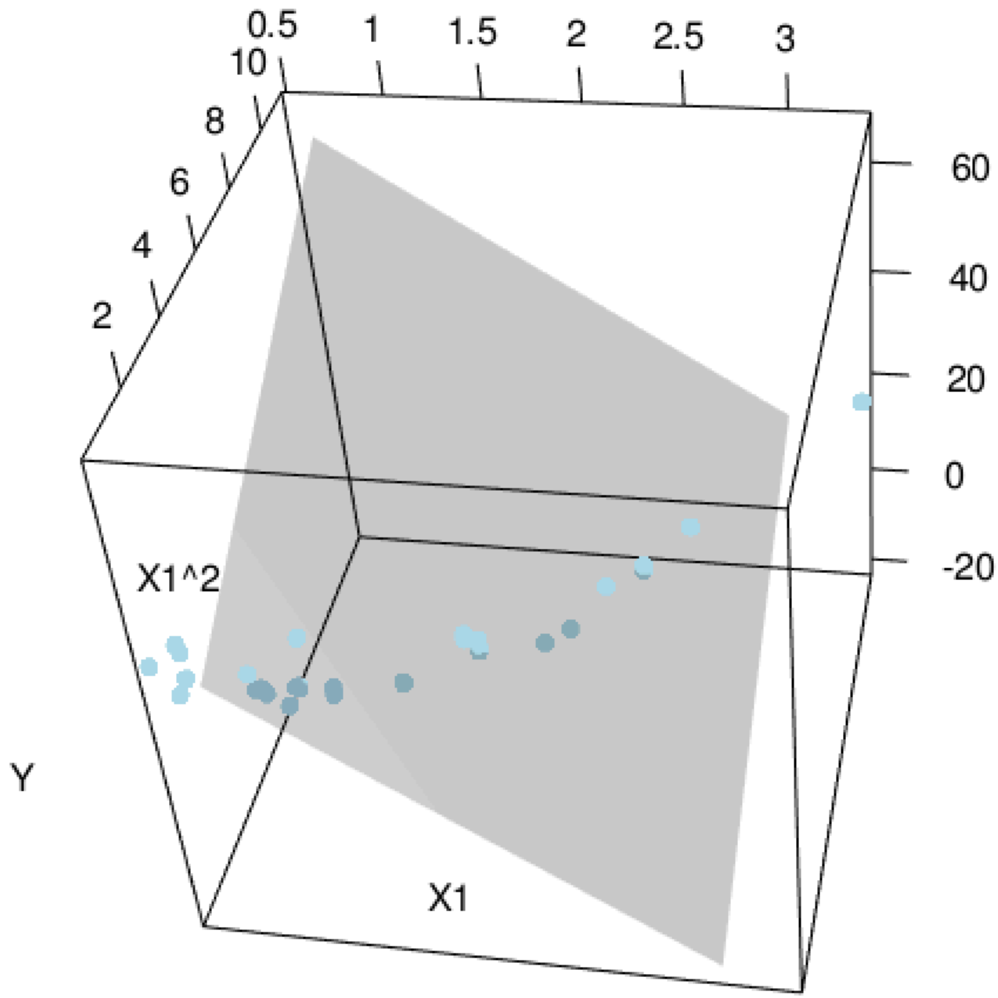
Quadratic Terms



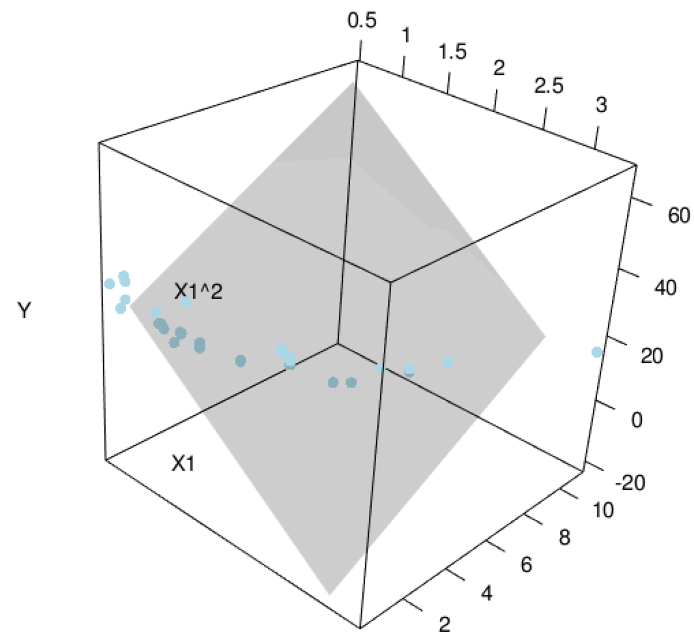
X_1 = horsepower/100

Y = mpg

Quadratic Terms

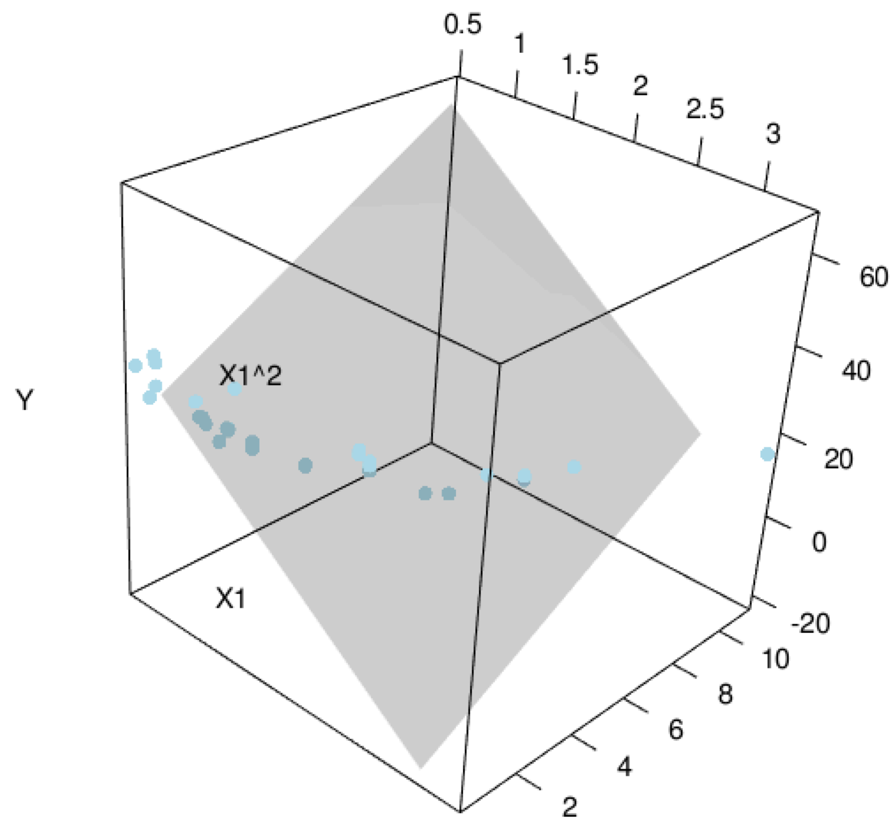


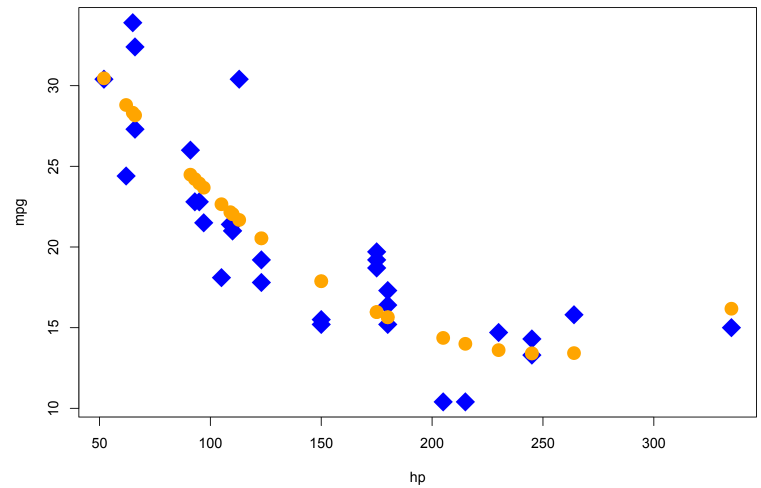
Quadratic Terms




```
> my_surface <- function(f, n=10, ...) {
+   ranges <- rgl:::.getRanges()
+   x <- seq(ranges$xlim[1], ranges$xlim[2], length=n)
+   y <- seq(ranges$ylim[1], ranges$ylim[2], length=n)
+   z <- outer(x,y,f)
+   surface3d(x, y, z, ...)
+   aspect3d(1,1,1)
+ }
>
> library(rgl)
>
> f <- function(x1, x2){40.4 -21.33*x1 + 4.21*x2}
>
> x1<-carsX[,2]
> x2<-carsX[,3]
> y<-carsy
>
> plot3d(x1,x2,y, type="p", col="lightblue", xlab="X1", ylab="X2", zlab="Y", site=5, lwd=15, size=7)
>   aspect3d(1,1,1)
> my_surface(f, alpha=.2)
>
```

Quadratic Terms



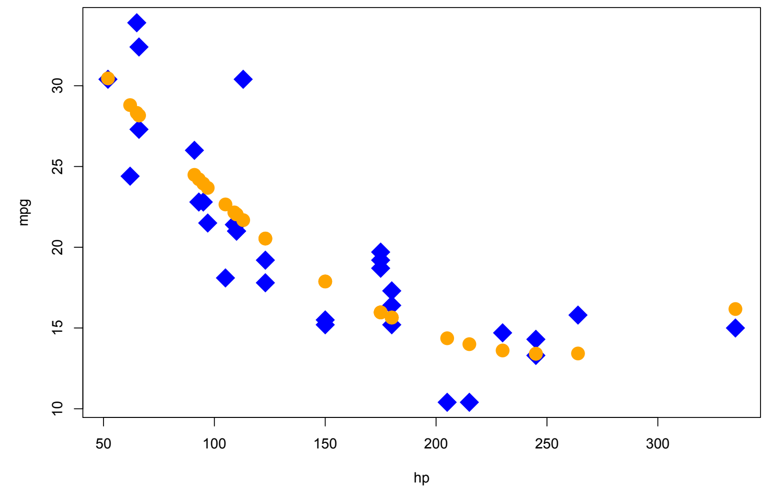


3.9 Categorical explanatory variables

What hypotheses can we test?

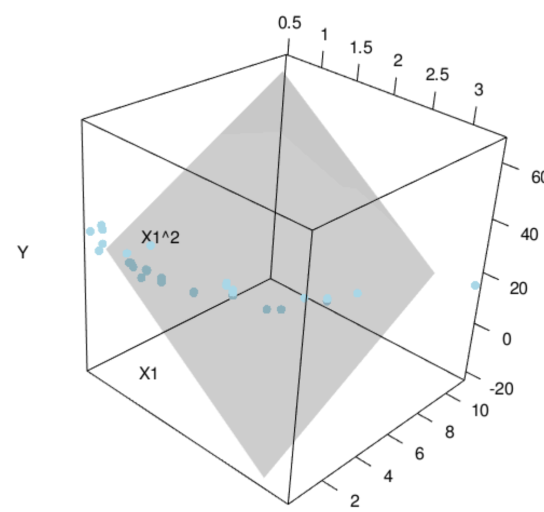
The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_1^2)$



Is the negative trend between mpg and hp, reduced as hp increases?

More general question:
Is there any association between mpg and hp?

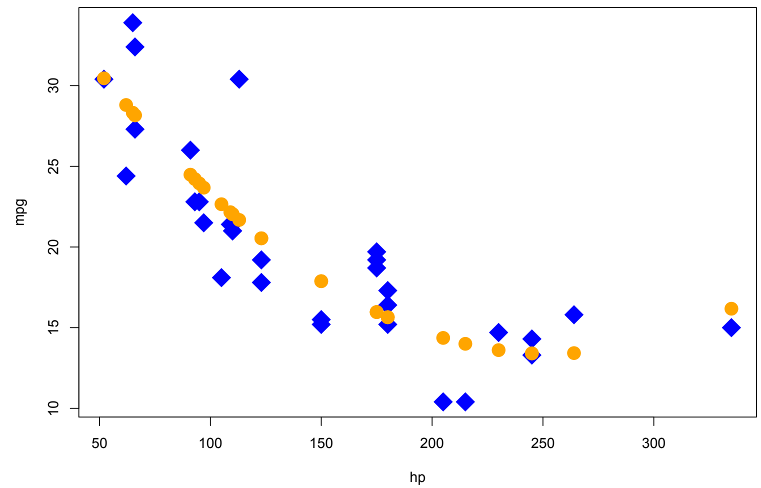


3.9 Quadratic variables

What hypotheses can we test?

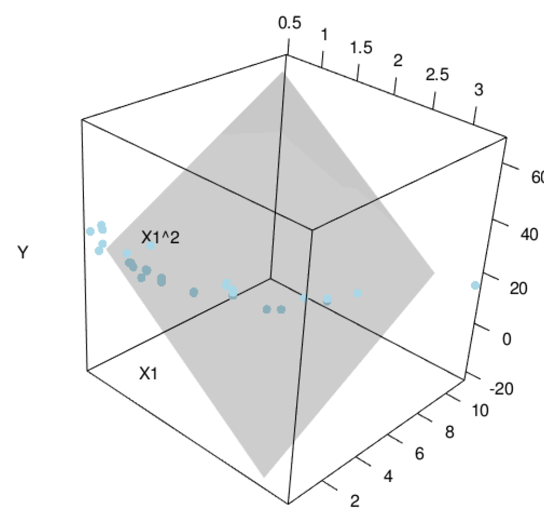
The **model**:

$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_1^2)$



Is the negative trend between mpg and hp, reduced as hp increases?

$H_0 : \beta_2 < 0$



More general question:
Is there any association between mpg and hp?

$H_0 : \beta_1 = \beta_2 = 0$

Look at the F-test p-value.

3.9 Quadratic variables

Does the negative trend between mpg and hp, decrease as hp increases?

$H_0 : \beta_2 < 0$

Is there any association between mpg and hp?

$H_0 : \beta_1 = \beta_2 = 0$

Look at the F-test p-value.

```
> # linear regression model
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
          [,1]      [,2]       [,3]
[1,]  7.511762 -9.139336  2.3744377
[2,] -9.139336 12.168840 -3.3411717
[3,]  2.374438 -3.341172  0.9691325
> carsmodQ
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  40.409      2.741 14.744        34.804        46.015      0
2 -21.331      3.488 -6.115       -28.465       -14.196      0
3   4.208      0.984  4.275         2.195         6.222      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0

> plot(mpg ~ hp, data=mtcars, cex=3, pch=18, col="blue")
> yhat<-carsX%*%c(40.409,-21.331, 4.208)
> points(mtcars$hp, yhat, col="orange", pch=20, cex=3)
```

3.9 Quadratic variables

Does the negative trend between mpg and hp, decrease as hp increases?

$H_0 : \beta_2 < 0$

Is there any association between mpg and hp?

$H_0 : \beta_1 = \beta_2 = 0$

Look at the F-test p-value.

```
> # linear regression model
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
          [,1]       [,2]       [,3]
[1,]  7.511762 -9.139336  2.3744377
[2,] -9.139336 12.168840 -3.3411717
[3,]  2.374438 -3.341172  0.9691325
> carsmodQ
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  40.409      2.741 14.744        34.804        46.015      0
2 -21.331      3.488 -6.115       -28.465       -14.196      0
3   4.208      0.984  4.275         2.195         6.222      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0

> plot(mpg ~ hp, data=mtcars, cex=3, pch=18, col="blue")
> yhat<-carsX%*%c(40.409,-21.331, 4.208)
> points(mtcars$hp, yhat, col="orange", pch=20, cex=3)
```

3.9 Quadratic variables

Does the negative trend between mpg and hp, decrease as hp increases?

$H_0 : \beta_2 < 0$

Is there any association between mpg and hp?

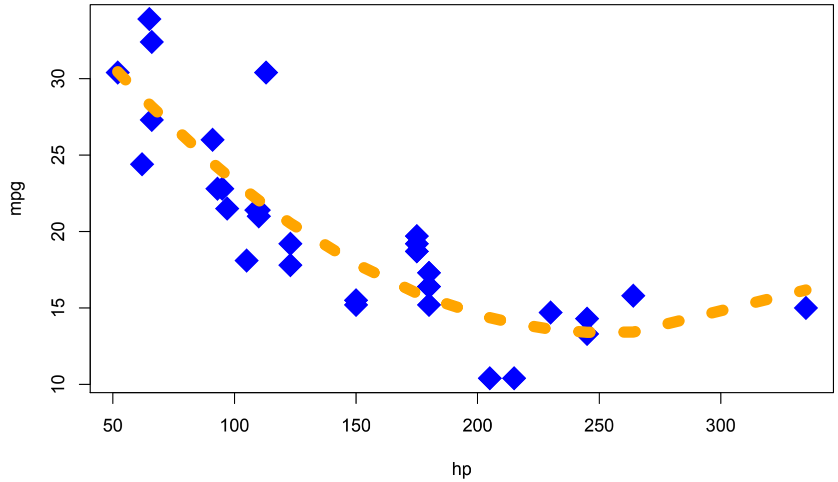
$H_0 : \beta_1 = \beta_2 = 0$

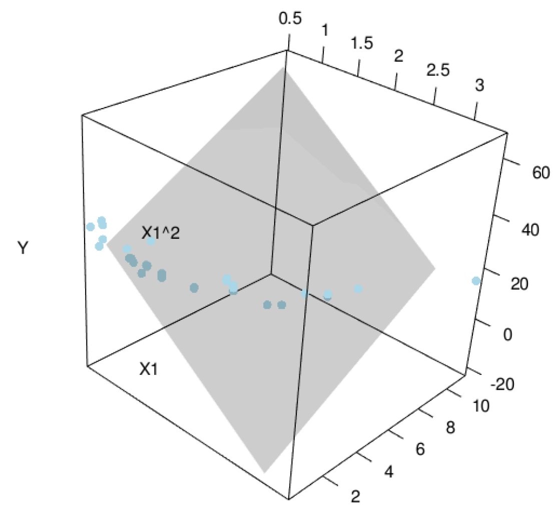
Look at the F-test p-value.

```
> # linear regression model
> carsmodQ<-linear_reg(carsy, carsX)
[1] "variance-covariance matrix for beta:"
          [,1]      [,2]       [,3]
[1,]  7.511762 -9.139336  2.3744377
[2,] -9.139336 12.168840 -3.3411717
[3,]  2.374438 -3.341172  0.9691325
> carsmodQ
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  40.409      2.741 14.744        34.804        46.015      0
2 -21.331      3.488 -6.115       -28.465       -14.196      0
3   4.208      0.984  4.275         2.195         6.222      0

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 274.632   9.47        3.077 0.756 0.739     44.953          0

> plot(mpg ~ hp, data=mtcars, cex=3, pch=18, col="blue")
> yhat<-carsX%*%c(40.409,-21.331, 4.208)
> points(mtcars$hp, yhat, col="orange", pch=20, cex=3)
```

3.9 Quadratic variables

Questions?





3.9 Quadratic + categorical variables

3.9 Quadratic + categorical variables

Consider X_2 as the type of transmission, with two categories:

“Manual”
“Automatic”

We have this coded as (0 = automatic, 1 = manual)

3.9 Quadratic + categorical variables

Consider X_2 as the type of transmission, with two categories:

"Manual"
"Automatic"

We have this coded as (0 = automatic, 1 = manual)

The **model**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1^2) + \beta_3 X_2$$

3.9 Quadratic + categorical variables

Consider X_2 as the type of transmission, with two categories: "Manual" and "Automatic".
(0 = automatic, 1 = manual)

The **model**:
$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_1^2) + \beta_3 X_2$

What hypotheses can we test?

3.9 Quadratic + categorical variables

Consider X_2 as the type of transmission, with two categories: "Manual" and "Automatic".
(0 = automatic, 1 = manual)

The **model**:
$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_1^2) + \beta_3 X_2$

What hypotheses can we test?

Is the mpg (adjusted for horsepower) different for manual and automatic cars ?

3.9 Quadratic + categorical variables

Consider X_2 as the type of transmission, with two categories: "Manual" and "Automatic".
(0 = automatic, 1 = manual)

The **model**:
$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_1^2) + \beta_3 X_2$

What hypotheses can we test?

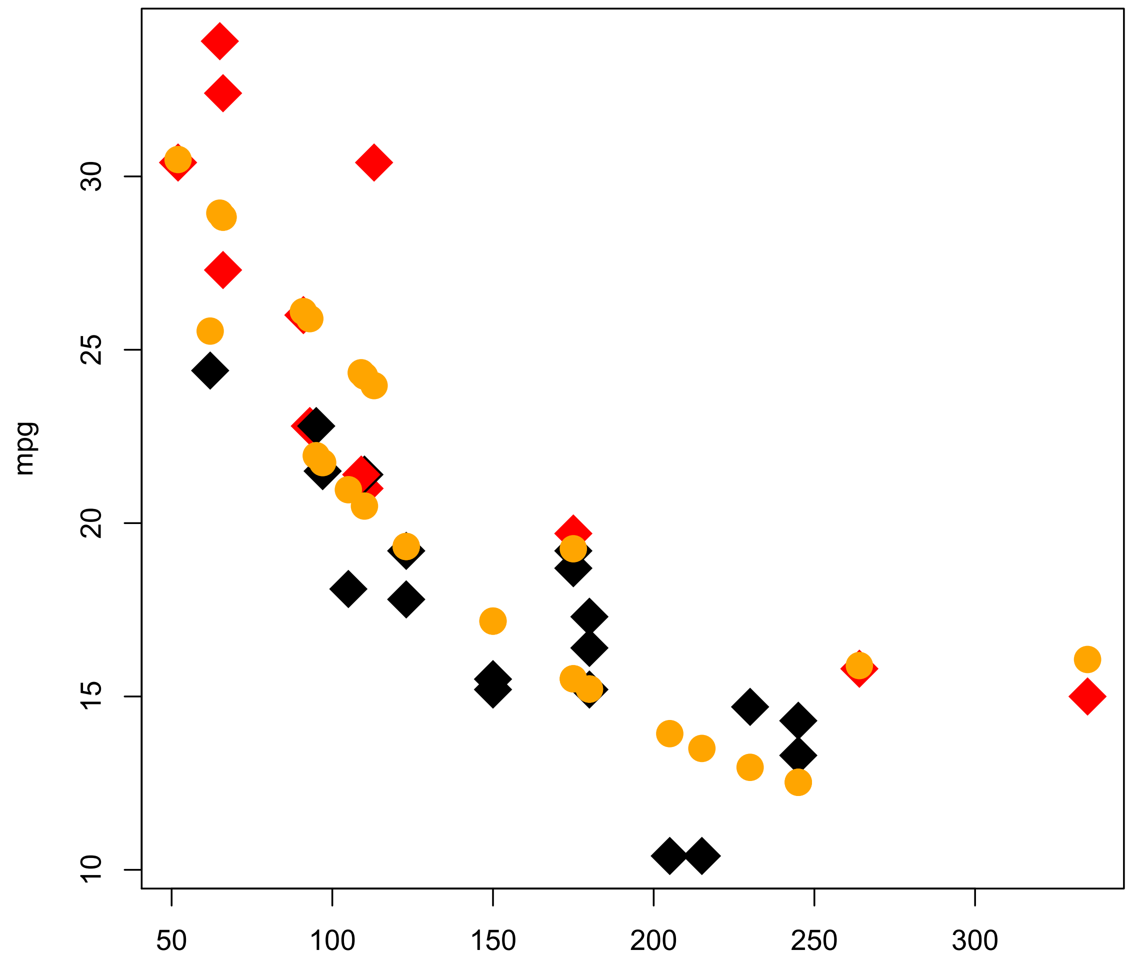
Is the mpg different for manual and automatic cars, controlling for horsepower (hp)?

$H_0 : \beta_3 = 0$

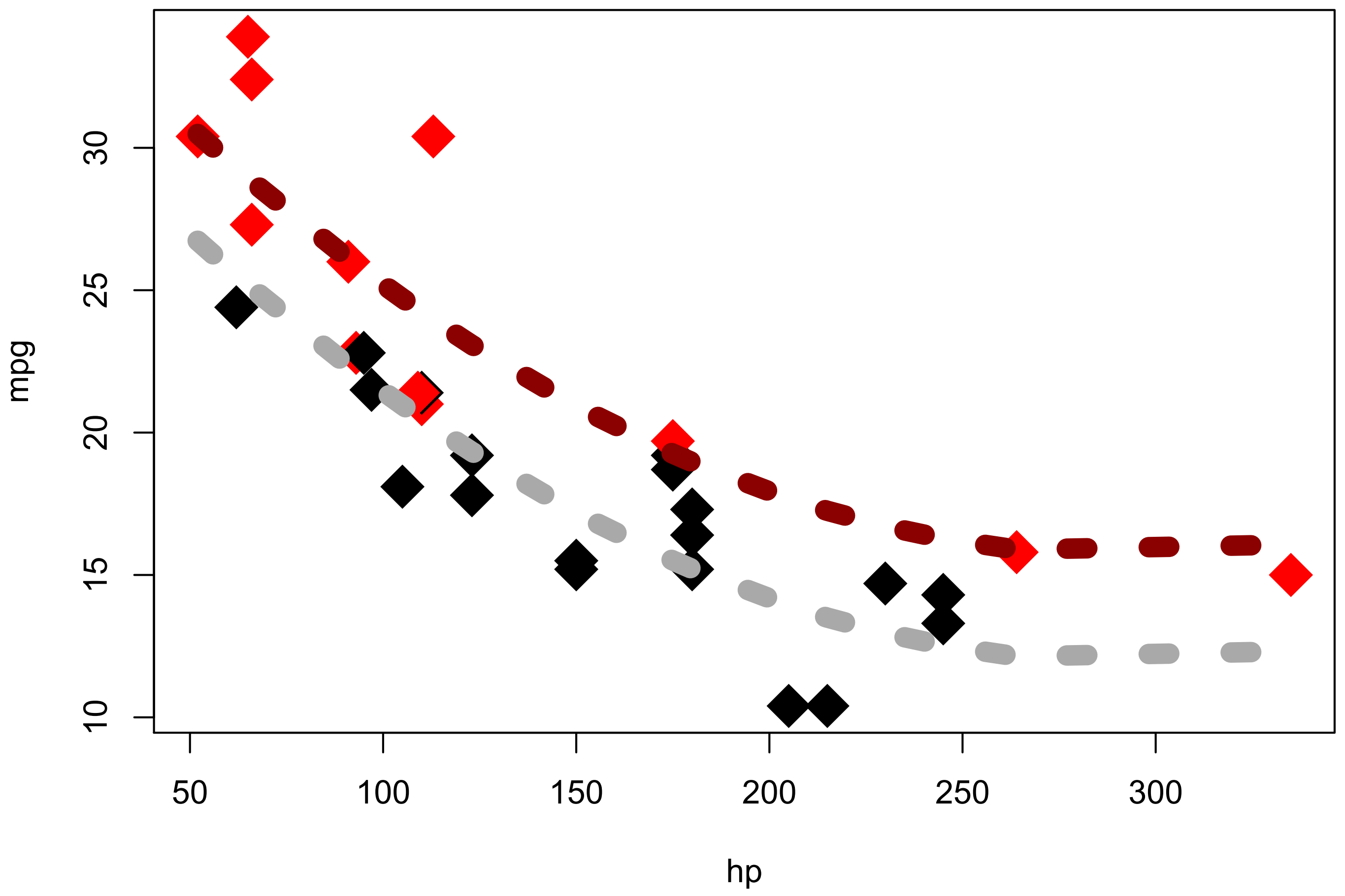
$H_0 : \beta_3 = 0$

```
> # add "am" -- Automatic(A) vs. Manual (M) covariate:
>
> carsy  <- (mtcars$mpg)
> carsX  <- cbind(1, mtcars$hp/100, (mtcars$hp/100)^2, mtcars$am)
> linear_reg(carsy, carsX)
[1] "Estimated Var-Cov matrix of beta:"
           [,1]       [,2]       [,3]       [,4]
[1,]   9.906608 -11.040062  2.8705169 -2.3937688
[2,] -11.040062  13.234580 -3.5763067  2.3395536
[3,]   2.870517  -3.576307  1.0060635 -0.6091053
[4,]  -2.393769   2.339554 -0.6091053  1.3537133
$coeftable
  betahat se_betahat tratio ci_lower_beta ci_upper_beta pvalue
1  33.776      3.147 10.731        27.329        40.223  0.000
2 -14.848      3.638 -4.081       -22.300        -7.396  0.000
3   2.520      1.003  2.513         0.466         4.575  0.018
4   3.751      1.163  3.224         1.368         6.135  0.003

$SStable
  SS_Total  SS_Res MS_Res sqrt.MS_Res.    R2 adjR2 Fstatistic Ftest_pval
1 1126.047 200.279  7.153        2.674 0.822 0.803     43.142          0


>
> plot(mpg ~ hp, data=mtcars, cex=3, pch=18, col=mtcars$am+1)
> yhat<-carsX%*%c( 33.776,-14.848, 2.520,  3.751)
>
> points(mtcars$hp, yhat, col="orange", pch=20, cex=3)
>
```

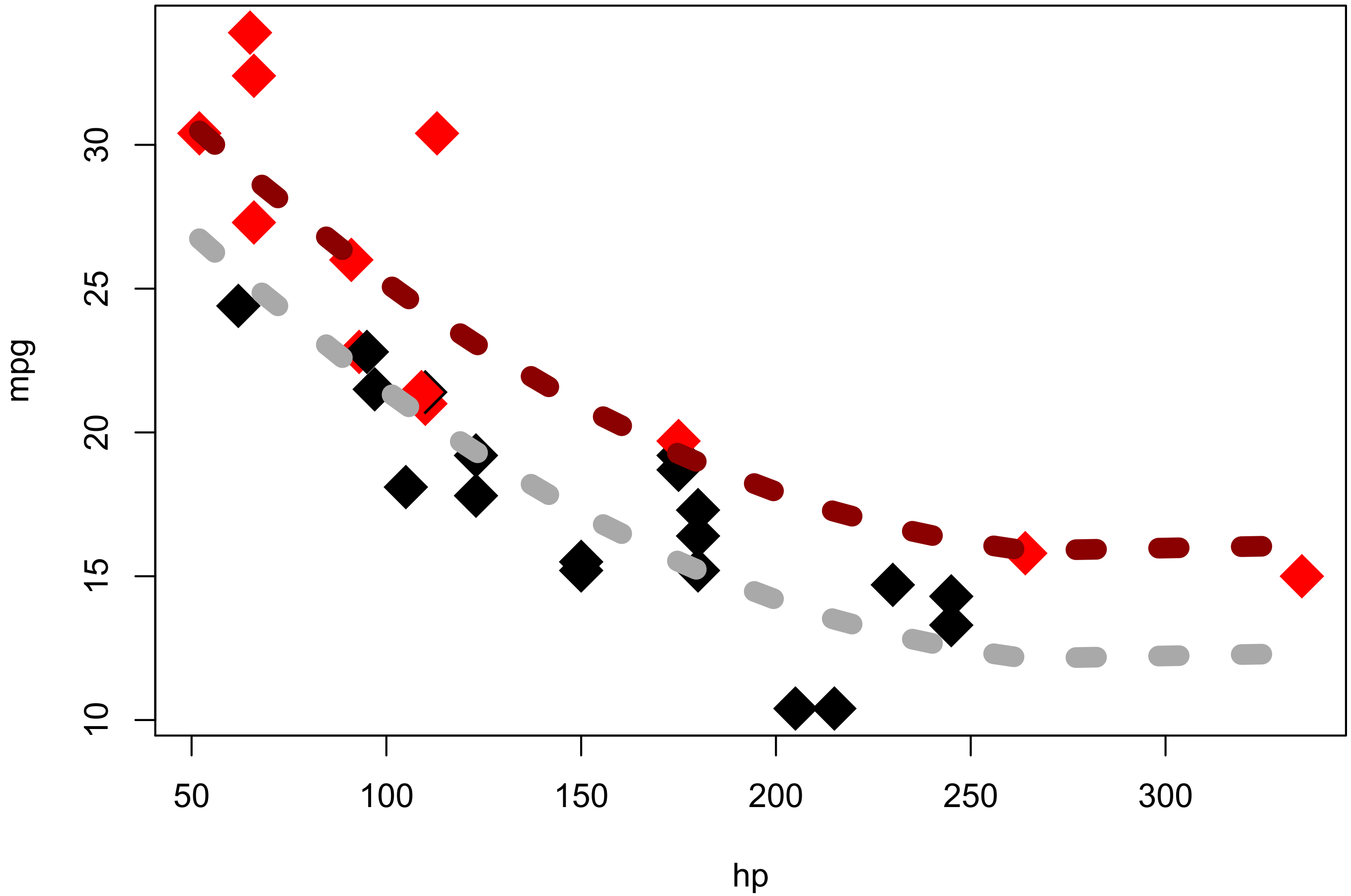
3.9 Quadratic variables + Categorical variables



3.9 Quadratic variables + Categorical variables

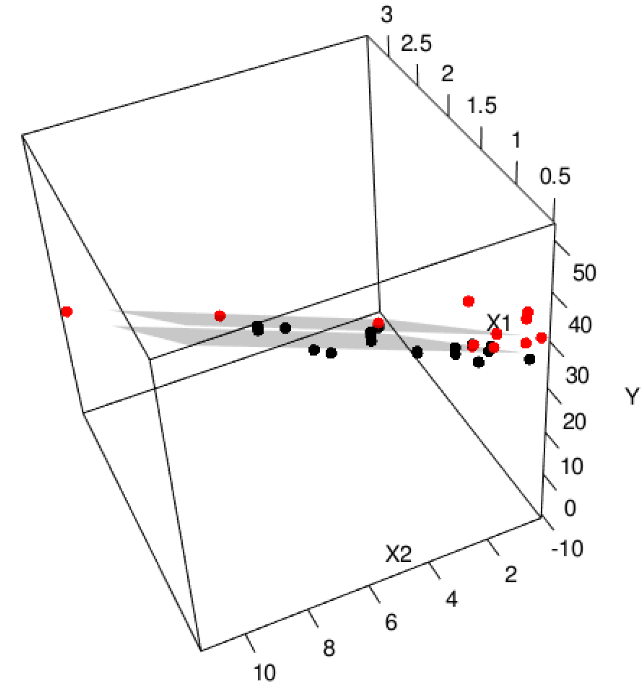


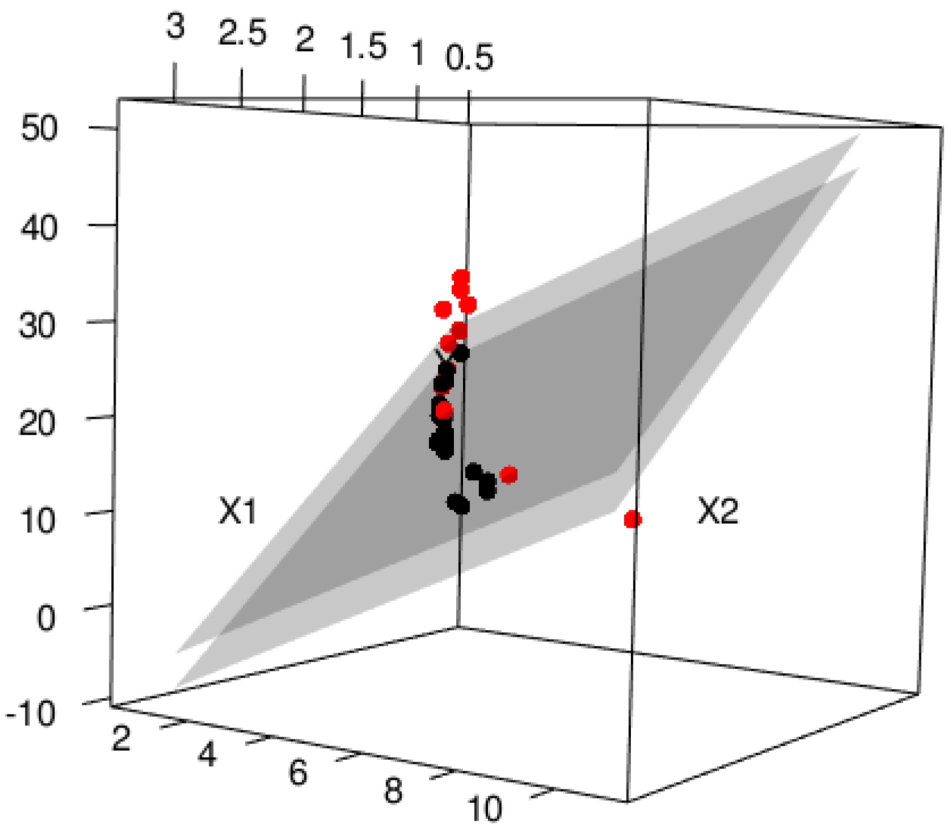
3.9 Quadratic variables + Categorical variables



Wait a minute… isn't this supposed to be ***LINEAR*** regression?!?

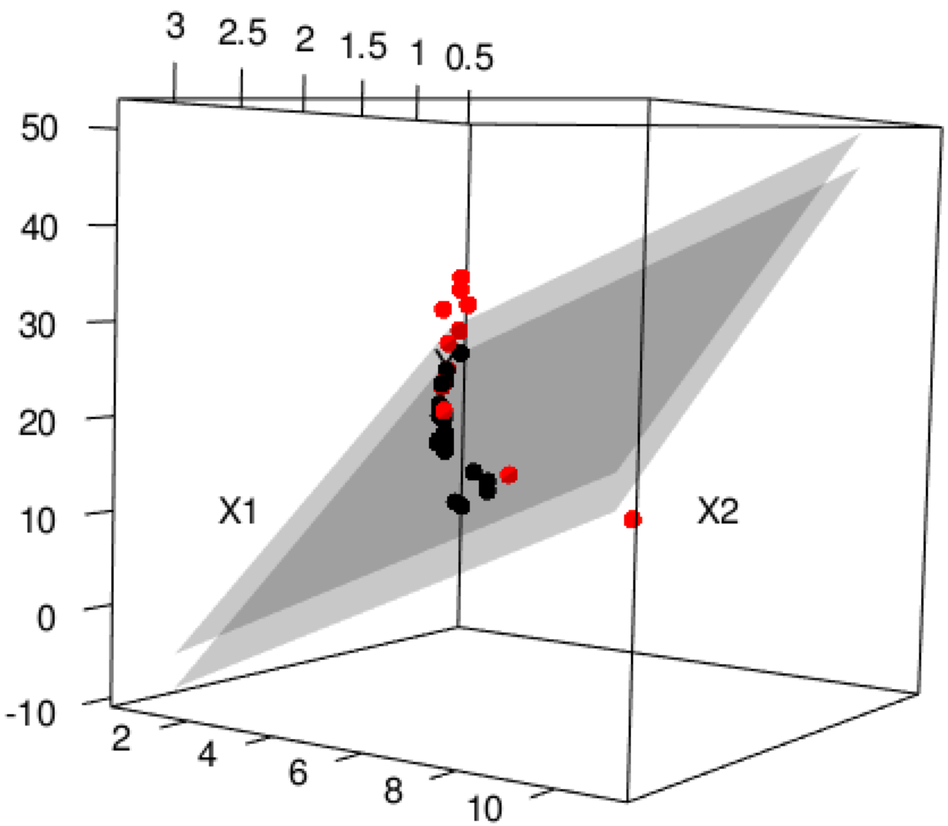
3.9 Quadratic variables variables

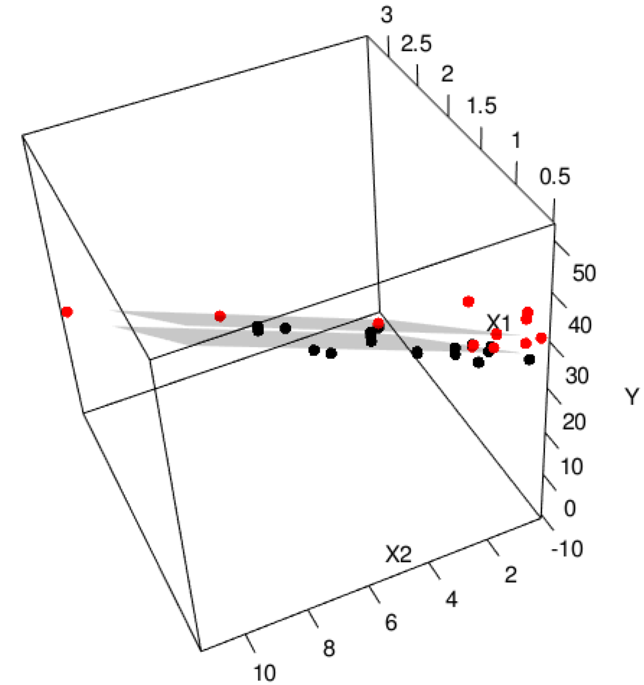




It IS linear in the higher dimension!

3.9 Quadratic variables variables



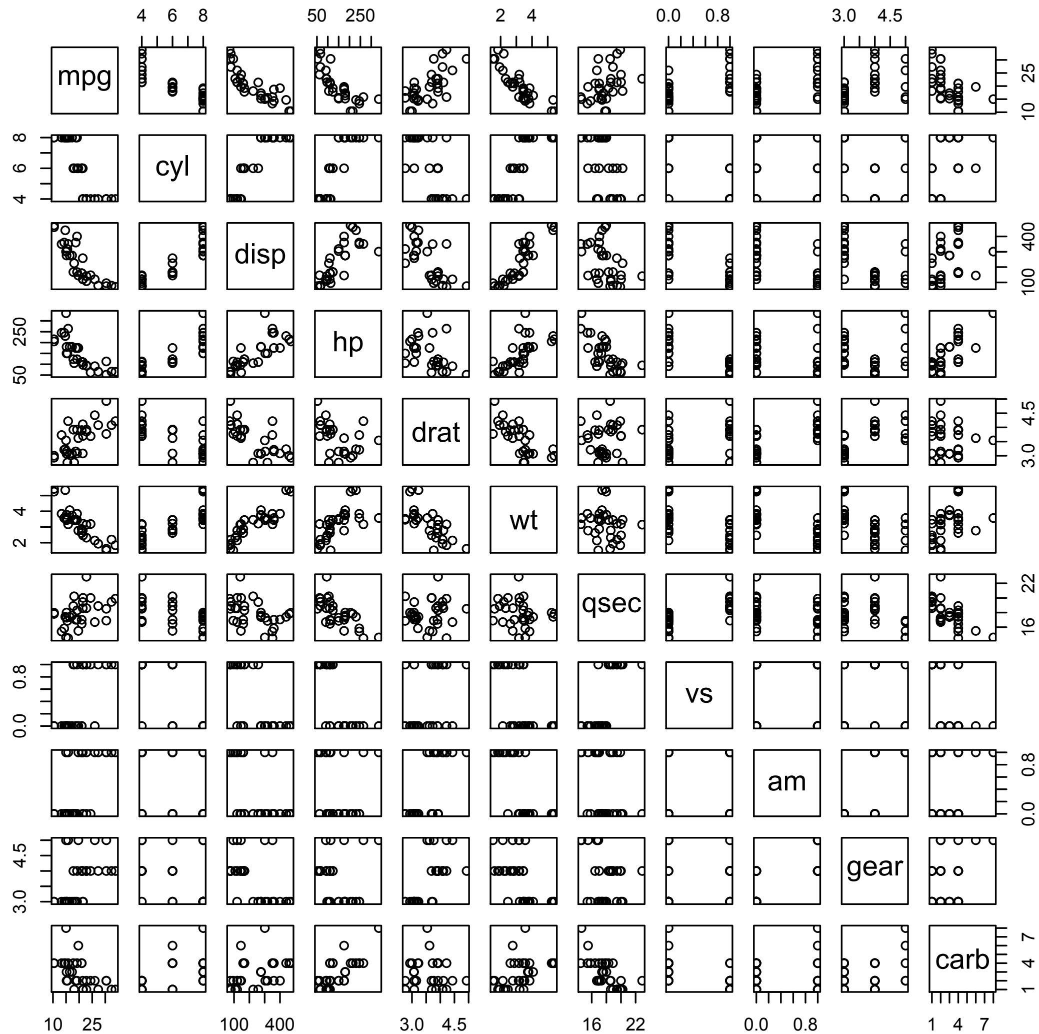


It IS linear in the higher dimension!

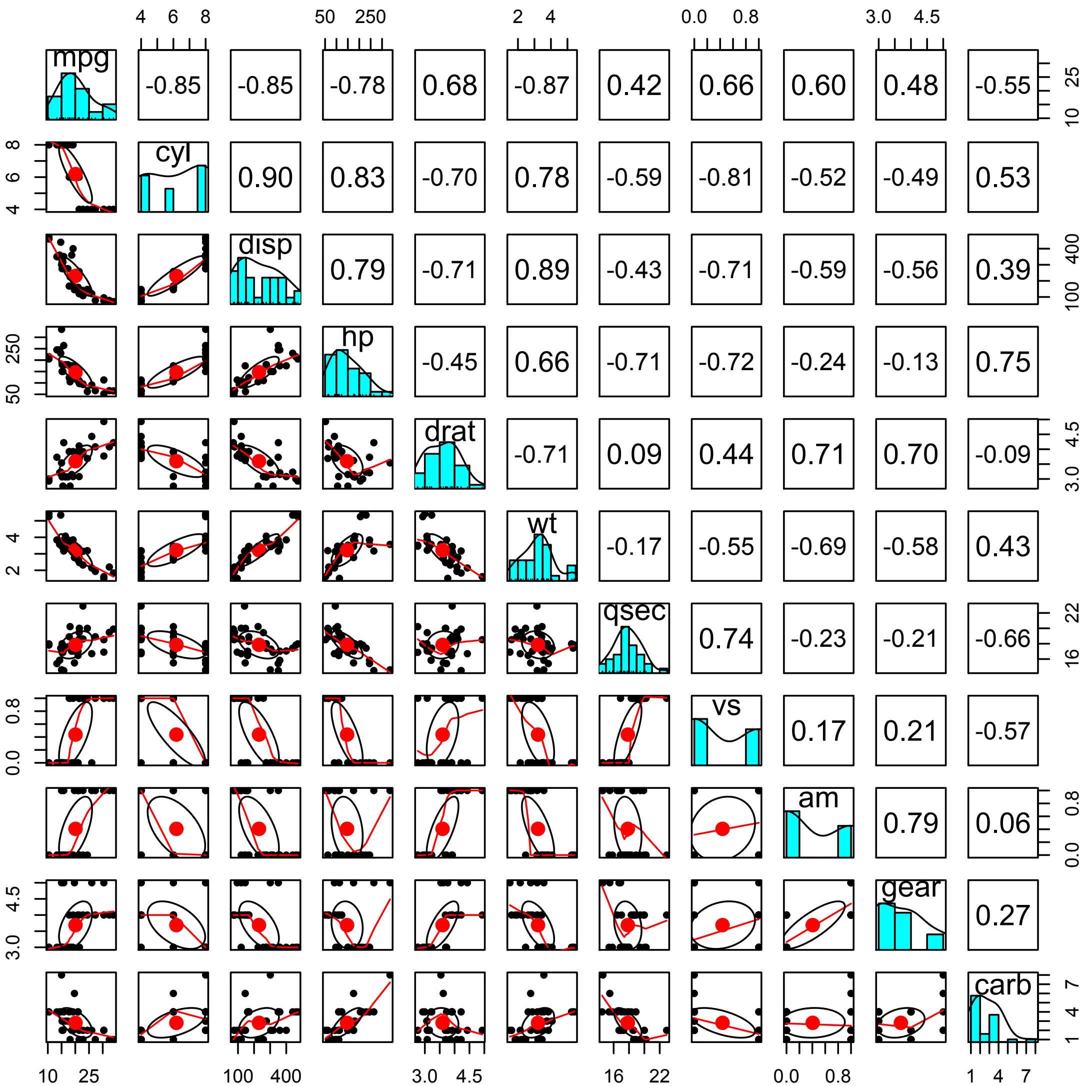
How would you include an interaction effect in the model?

What would the 3d plot look like?

```
> pairs(mtcars)
>
```



```
> library(psych)
> pairs.panels(mtcars)
```



3.9 Quadratic variables + Categorical variables

Search online for “mtcars linear regression” and you will find some great material, for example:

- https://rstudio-pubs-static.s3.amazonaws.com/193417_4e1f9d5b1c6f472885fc5b03df9d4331.html
- http://rstudio-pubs-static.s3.amazonaws.com/20516_29b941670a4b42688292b4bb892a660f.html
- https://cran.r-project.org/web/packages/olsrr/vignettes/residual_diagnostics.html

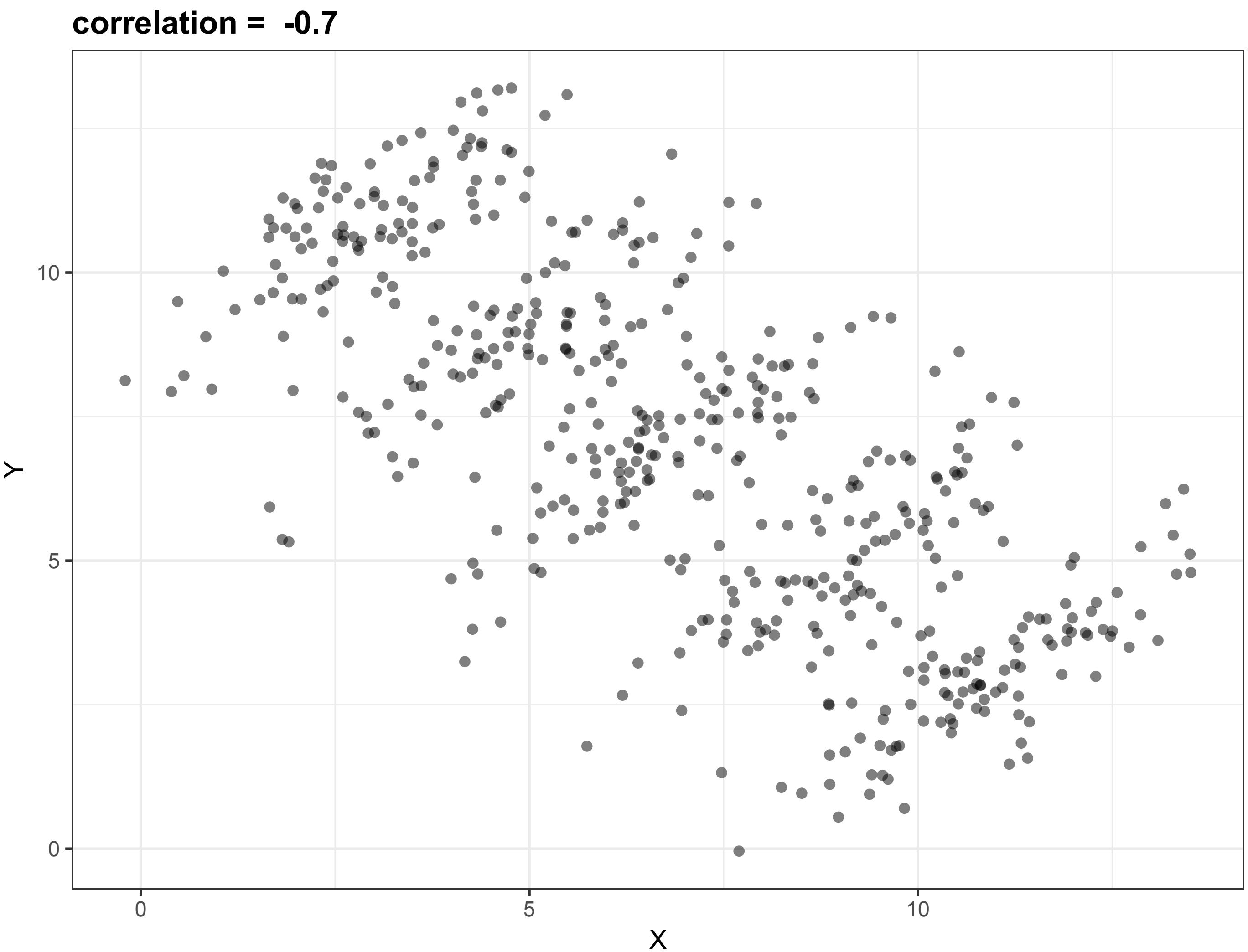
2.10 Partial correlations

Definition 3.1 Partial correlation $r_{yx;\mathbf{w}} = r_{xy;\mathbf{w}}$ of y and x given variables in $\mathbf{w}$.

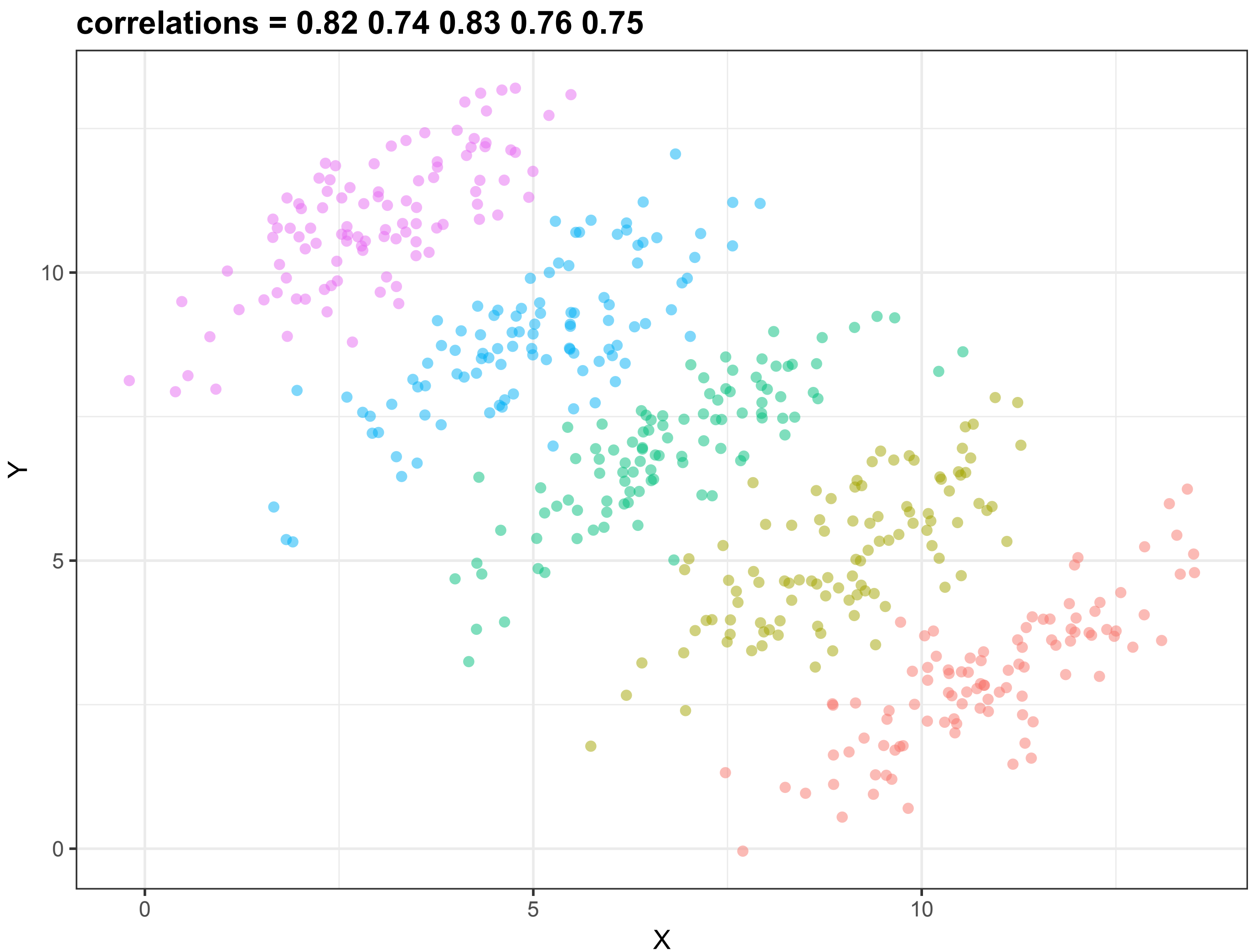
1. Regress y on $\mathbf{w}$ (say with length m) and get residuals $e_{i1} = y_i - \hat{\beta}_0 - \hat{\beta}_1 w_{i1} - \cdots - \hat{\beta}_m w_{im}$.
2. Regress x on $\mathbf{w}$ and get residuals $e_{i2} = x_i - \hat{\gamma}_0 - \hat{\gamma}_1 w_{i1} - \cdots - \hat{\gamma}_m w_{im}$.
3. The sample correlation of (e_{i1}, e_{i2}), $i = 1, \ldots, n$, is defined to be the partial correlation $r_{yx;\mathbf{w}}$.

$\square$

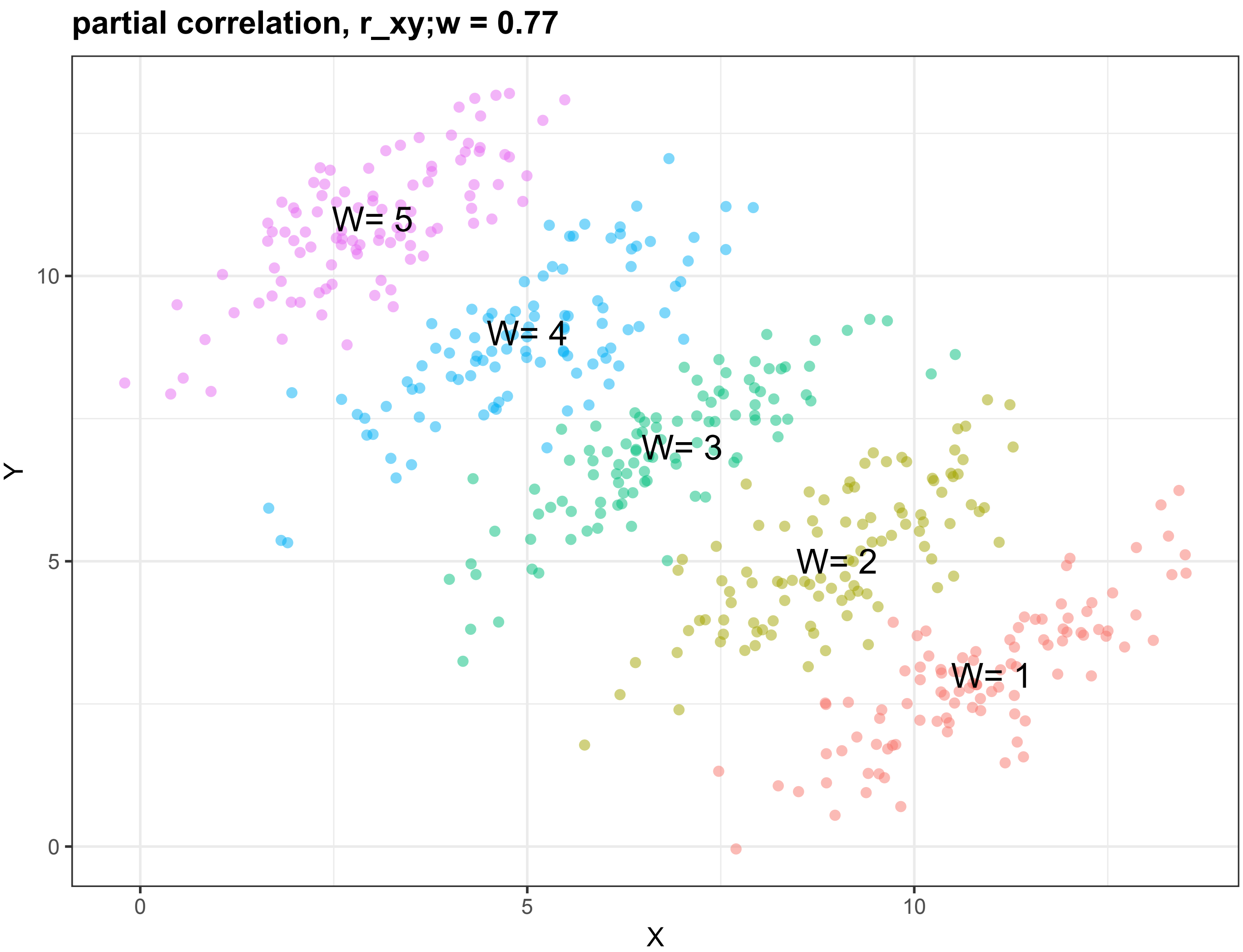
2.10 Partial correlations



2.10 Partial correlations



2.10 Partial correlations



2.10 Partial correlations

Some more results on partial correlations.
There is a recursive formula for partial correlations as more variables are conditioned on (check for conditioning on two variables by implementing the above formulas and those below, and verifying that they lead to identical results).

(3.135) $$r_{yx_2;x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_2 x_1}}{\sqrt{(1 - r^2_{yx_1})(1 - r^2_{x_2 x_1})}},$$

(3.136) $$r_{yx;z} = \frac{r_{yx} - r_{yz} r_{xz}}{\sqrt{(1 - r^2_{yz})(1 - r^2_{xz})}},$$

(3.137) $$r_{yx_3;x_1 x_2} = \frac{r_{yx_3;x_2} - r_{yx_1;x_2} r_{x_3 x_1;x_2}}{\sqrt{(1 - r^2_{yx_1;x_2})(1 - r^2_{x_3 x_1;x_2})}},$$

(3.138) $$r_{yx;z\mathbf{w}} = \frac{r_{yx;\mathbf{w}} - r_{yz;\mathbf{w}} r_{xz;\mathbf{w}}}{\sqrt{(1 - r^2_{yz;\mathbf{w}})(1 - r^2_{xz;\mathbf{w}})}},$$

where $\mathbf{w}$ can consist of more than one variable (or none).