STA302 - Lecture 7

Cedric Beaulac

May 30, 2019

Introduction

Today's plan

- Today we conclude the introduction to Multiple Linear Regression:
  - Interactions
  - Polynomial fit
  - Diagnostic and New Problems (collinearity)
- It is the last lecture before test#2 and the last lecture of the core content of STA302.
- Next week we will make a leap forward in time and discuss modern set up and problems!

Interactions

Interactions

- Interactions only make sense when we have multiples predictors.
- Interactions are hard to interpret.
- They represent the interaction between the EFFECT of predictors on the response.

Interactions

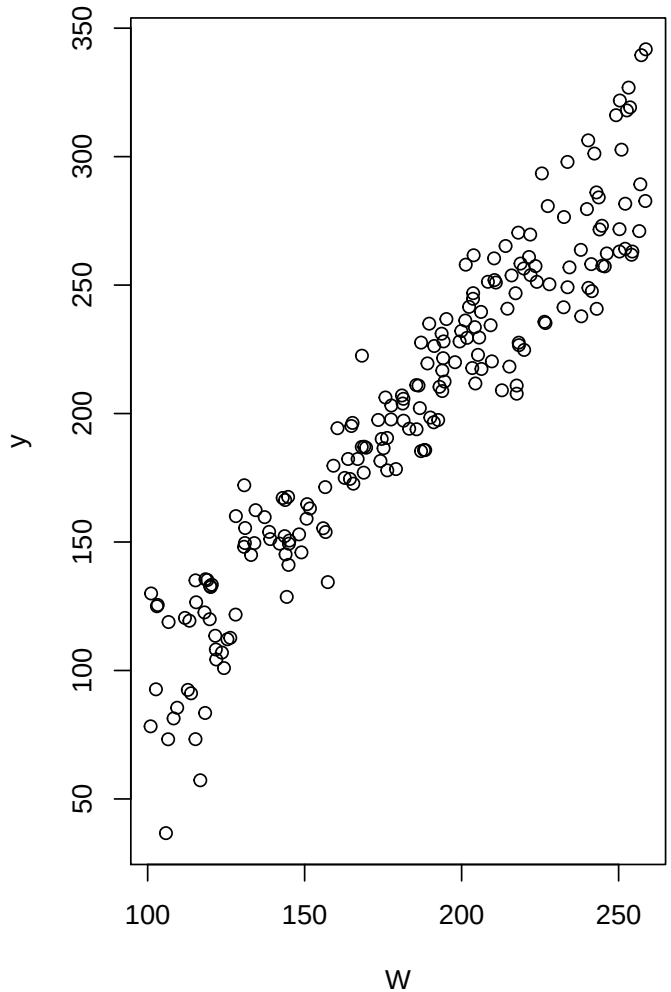
- Interaction IS NOT the effect of predictor x_1 on predictor x_2.
- Interaction IS the effect of predictor x_1 on the effect of predictor x_2 on y.
- As x_1 varies, the effect of x_2 on y is different.
- As x_1 varies, the relationship between x_2 and y is different.
- Also, as x_2 varies, the relationship between x_1 and y is different.

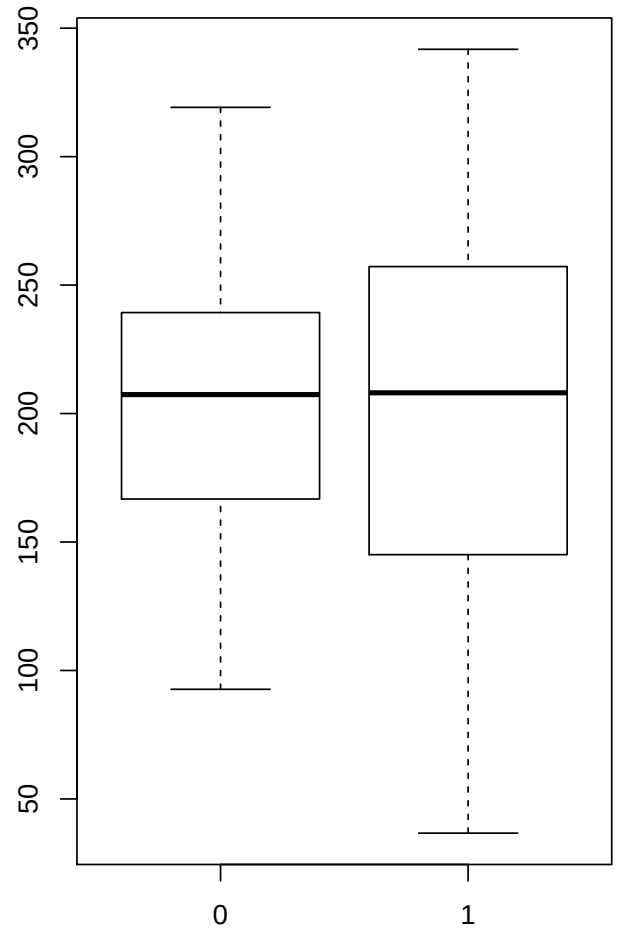
Interactions

- Interaction between two categorical predictors means that every combinaison are actually categories with respective effects.
- Interaction between a categorical predictor and a numerical predictor implies different intercepts and different slopes.
- Interaction between two numerical predictors are... much more complicated. I'm not going to try to simply explain it.

Interactions

- Let's reintroduce last lecture problem for a simple demonstration of interactions.

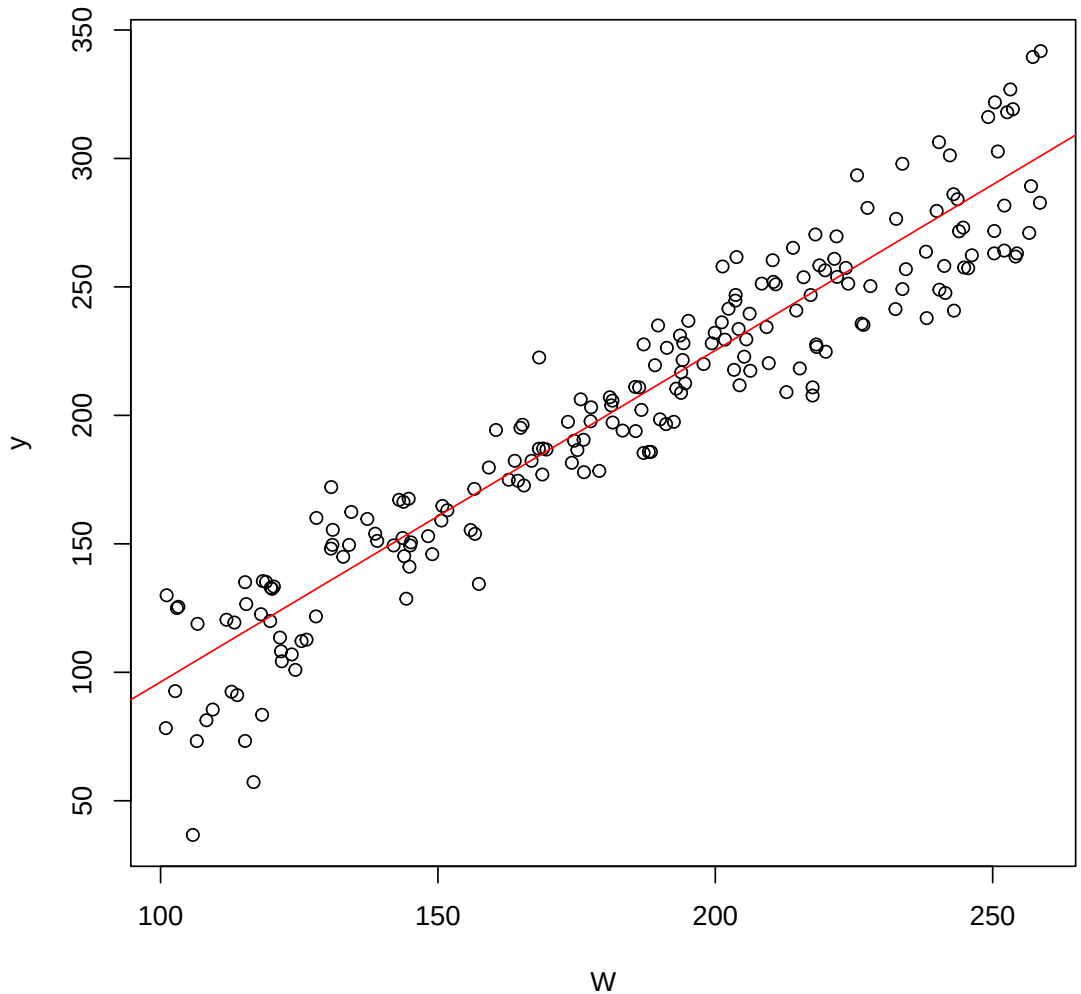




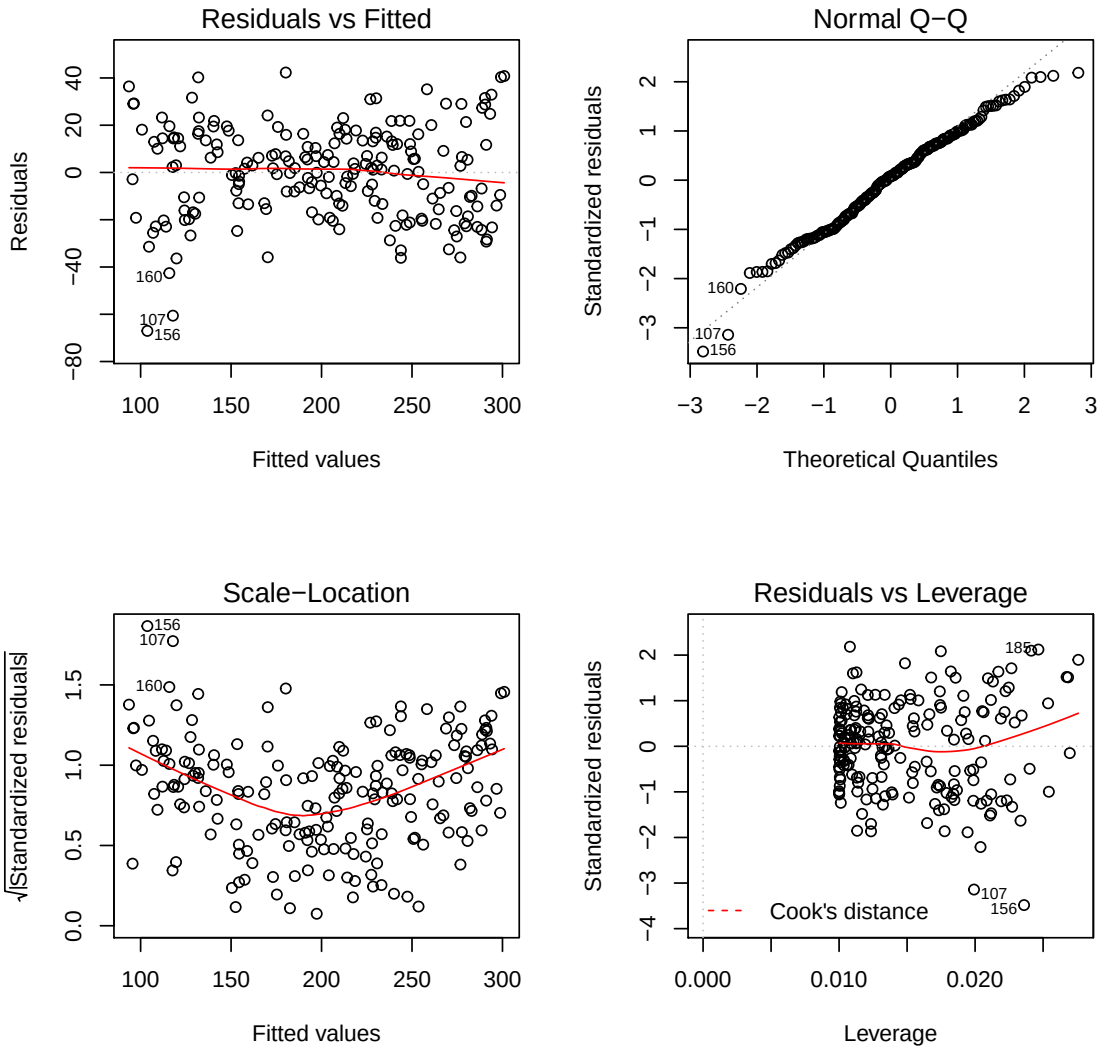
Interactions

```
##
## Call:
## lm(formula = y ~ W + G)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.034 -14.143   1.545  14.315  42.286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.90458    5.97129  -6.180 3.61e-09 ***
## W             1.29058    0.03029  42.612  < 2e-16 ***
## G             4.06778    2.75933   1.474    0.142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.47 on 197 degrees of freedom
## Multiple R-squared:  0.9022, Adjusted R-squared:  0.9012
## F-statistic: 908.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Interactions



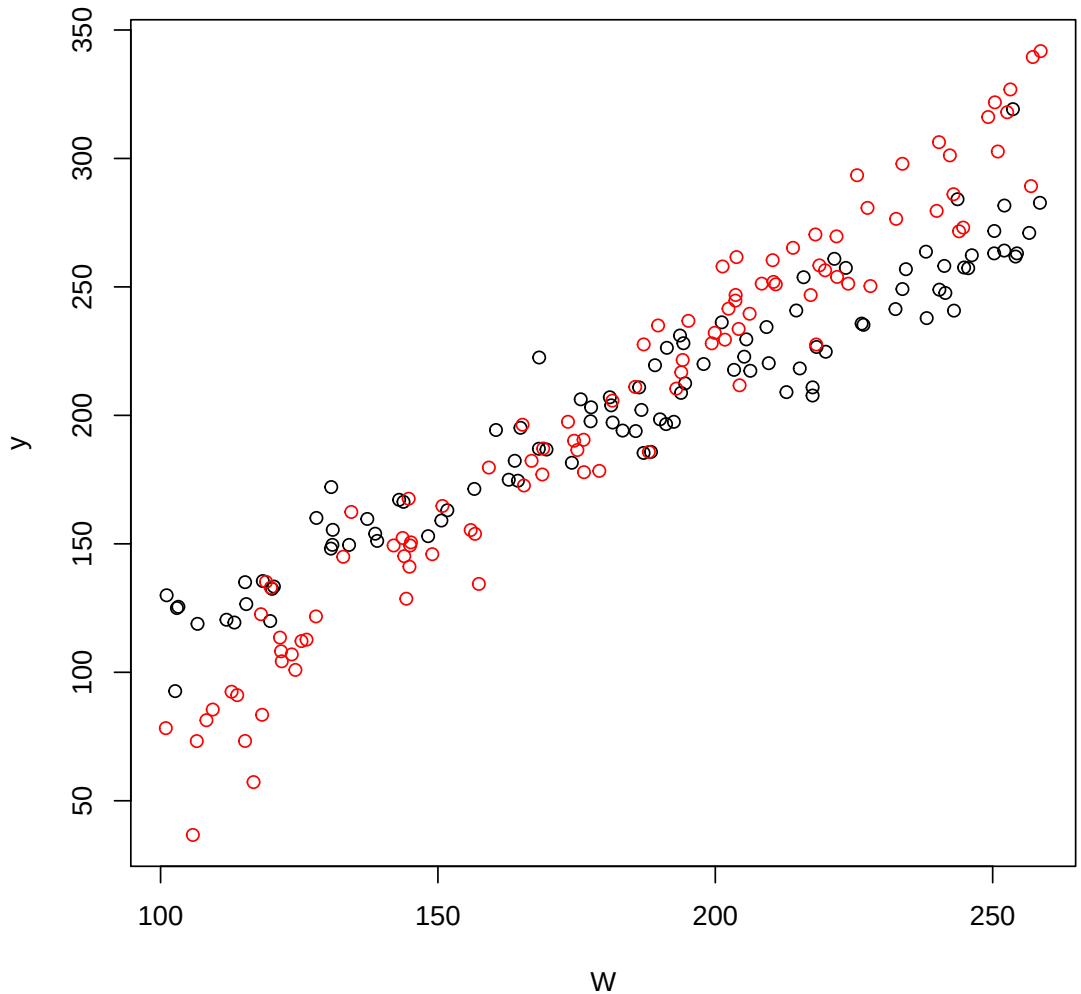
Interactions



Interactions

- It's a perfectly valid model. It's a great fit overall.
- The group means seem relatively similar.
- And it seems like we could fit a single slope for both groups.
- Overall from the look of it we could conclude gender is not even usefull here.
- Actually the parameter attached to gender is not significant (according to our t-test).
- So why should we consider interaction.
- Let's add colors to the plot!

Interactions



Interactions

- *Wait a minute!*
- It seems like the two groups have both different slopes AND different intercept.
- Why did the our output told us the groups were the same ?
- Well the red dots have lower intercept and bigger slope an both effect counters one another.
- We did not allow for different slopes when we fit the model!

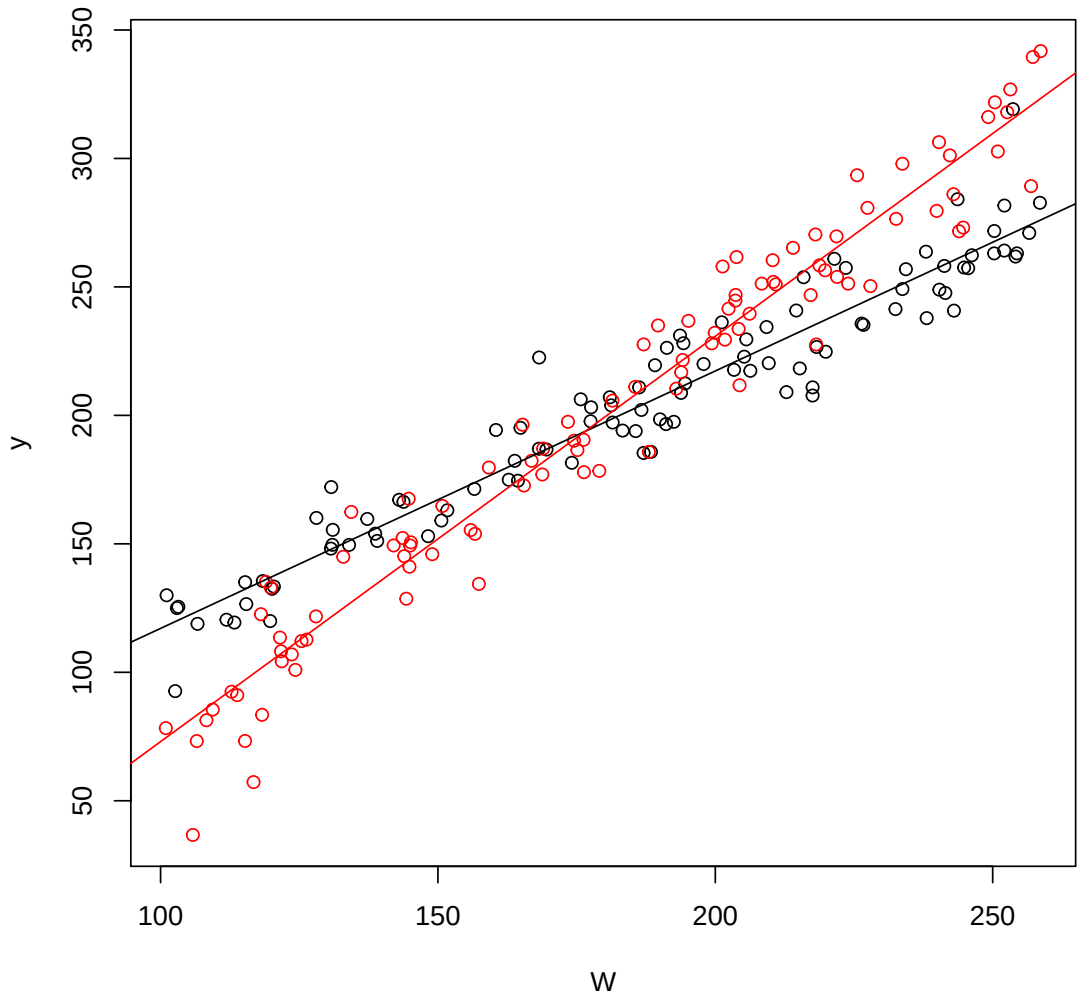
Interactions

- An interaction is actually a product of effect :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 + e$
- For a fix weight x, if the observation is a male, then
  $\mathbf{E}(y) = \beta_0 + \beta_1 x$ and if the observation is a female then
  $\mathbf{E}(y) = \beta_0 + \beta_1 x + \beta_2 + \beta_{1,2} x = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,2})x$.
- It is literrally computed as x_1 times x_2 and for a continuous and a categorical predictor it leads to different intercept and different slopes.

Interactions

```
##
## Call:
## lm(formula = y ~ W * G)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.489  -8.476  -0.850   8.182  48.144
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.88110    6.06629   2.783  0.00592 **
## W              1.00200    0.03162  31.686  < 2e-16 ***
## G           -101.66786    8.45048 -12.031  < 2e-16 ***
## W:G            0.57591    0.04467  12.892  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.36 on 196 degrees of freedom
## Multiple R-squared:  0.9471, Adjusted R-squared:  0.9463
## F-statistic:  1169 on 3 and 196 DF,  p-value: < 2.2e-16
```

Interactions



Interactions

- You can see how important interaction terms are!
- They allow the EFFECT of weight to be different given the biological gender.
- They are complicated to fully understand but they are extremely important to better fit the data.
- Simple plots are not enough to detect if we need them (as shown above).
- The model is completely *wrong* without the interaction term.

Interactions

- It is kind of scary. Should we always add interaction terms ?
- Well, it makes the number of parameter quickly explode!
- If we have p predictors, we have $\binom{p}{2}$ intercation terms. (Technically there is a lot more but we won't talk about it yet)
- So it may not be the right decision to always add all the interactions terms!
- Before we introduce graphical tools to detect interactions, let's introduce Categorical vs Catergorical interactions and finally Numerical vs Numerical!

Interactions

- Interactions between two categorical predictors.
- Let's stick with the simplest case of two binary categorical variables. So predictor 1, has Group A1 and B1 and predictor 2 has Group A2 and B2.
- We identify them with dummy variables x_1 and x_2. We have :

Predictor 1	x_1	Predictor 2	x_2
A1	0	A2	0
B1	1	B2	1

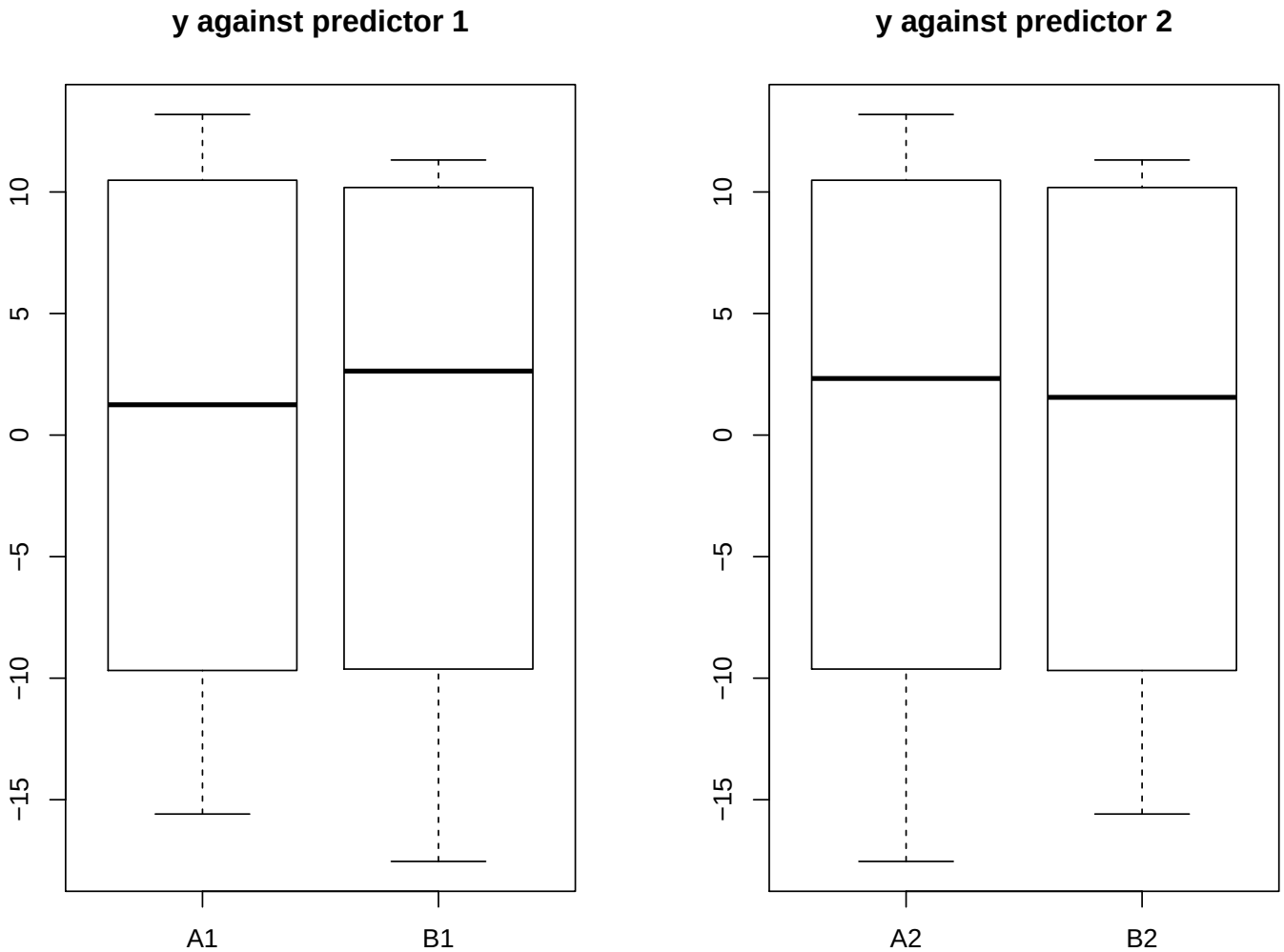
Interactions

- If we fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$ we consider the difference between Group A1 and B1 and the difference between Group A2 and B2.

E(y|A1,A2)	β_0
E(y|B1,A2)	$\beta_0 + \beta_1$
E(y|A1,B2)	$\beta_0 + \beta_2$
E(y|B1,B2)	$\beta_0 + \beta_1 + \beta_2$

- Sometimes it won't be enough. Since here changing from A1 to B1 always has the same effect. The effect is *additive*.

Interactions



Interactions

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.457  -9.820   2.236  10.226  12.925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.26390    1.82881   0.144    0.886
## x1          -0.34388    2.11172  -0.163    0.871
## x2           0.01734    2.11172   0.008    0.993
##
## Residual standard error: 10.56 on 97 degrees of freedom
## Multiple R-squared:  0.000274,   Adjusted R-squared:  -0.02034
## F-statistic: 0.01329 on 2 and 97 DF,  p-value: 0.9868
```

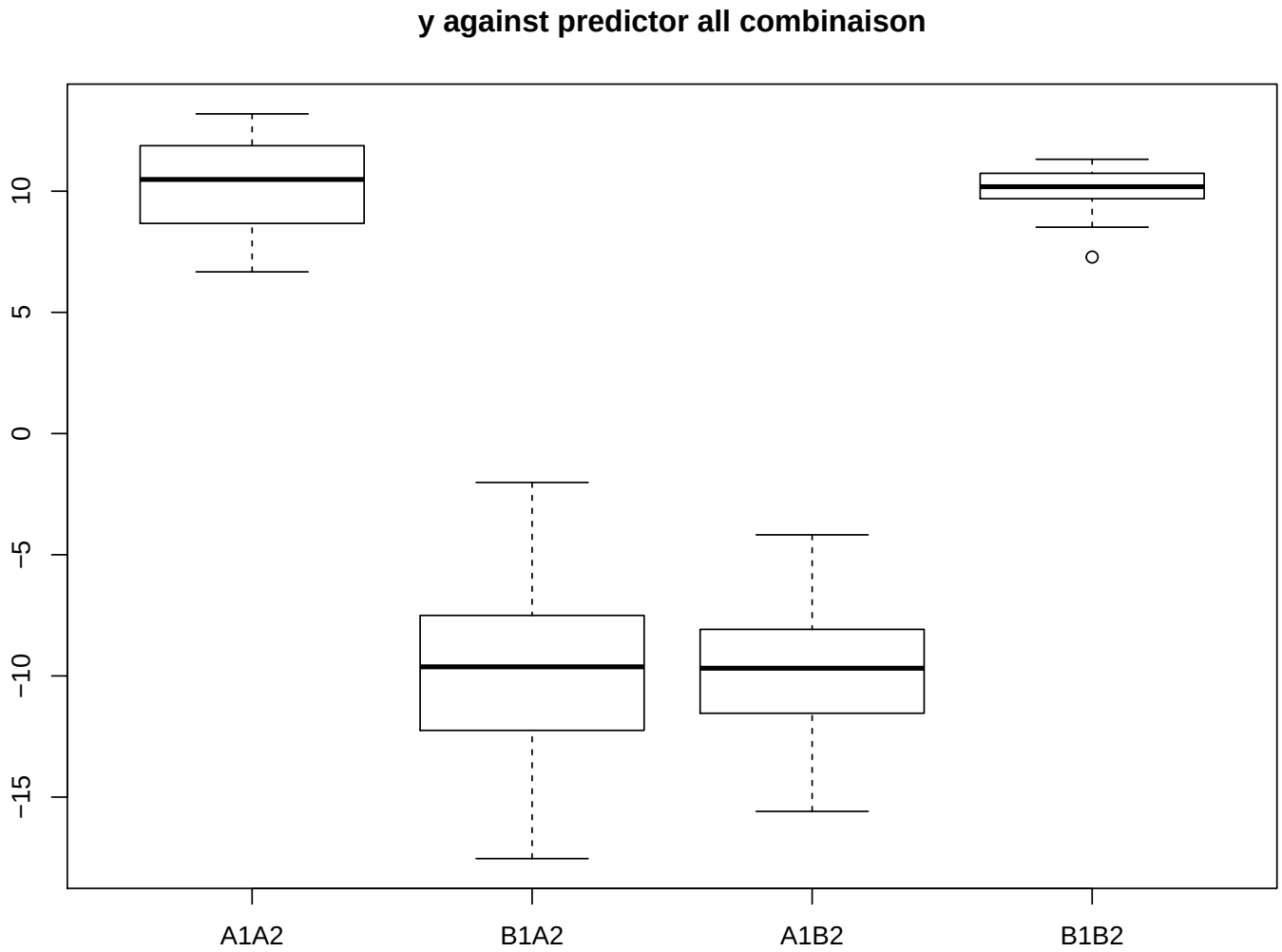
Interactions

- If we had an interaction term :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 + e$

E(y|A1,A2)	β_0
E(y|B1,A2)	$\beta_0 + \beta_1$
E(y|A1,B2)	$\beta_0 + \beta_2$
E(y|B1,B2)	$\beta_0 + \beta_1 + \beta_2 + \beta_{1,2}$

-$\beta_{1,2}$ allow for a different effect ffrom changing to A1 to B1 depending on the value of predictor 2.

Interactions



- The interaction allows the model to consider all combinaison of categorical variables.

Interactions

```
##
## Call:
## lm(formula = y ~ x1 * x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3830 -1.4124  0.1586  1.3116  8.1377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.3383     0.5262   19.65   <2e-16 ***
## x1          -20.4927     0.7442  -27.54   <2e-16 ***
## x2          -20.1315     0.7442  -27.05   <2e-16 ***
## x1:x2        40.2977     1.0524   38.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.631 on 96 degrees of freedom
## Multiple R-squared:  0.9386, Adjusted R-squared:  0.9366
## F-statistic: 488.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

Interactions

- Interactions between two continuous predictors, it is more complicated.
- The slope and intercept of x_1 is different across values x_2.
- Once again, we might miss it and it is not obvious how to detect it.
- Here are 3D plots (credit to Craig Burkett!)

Interactions

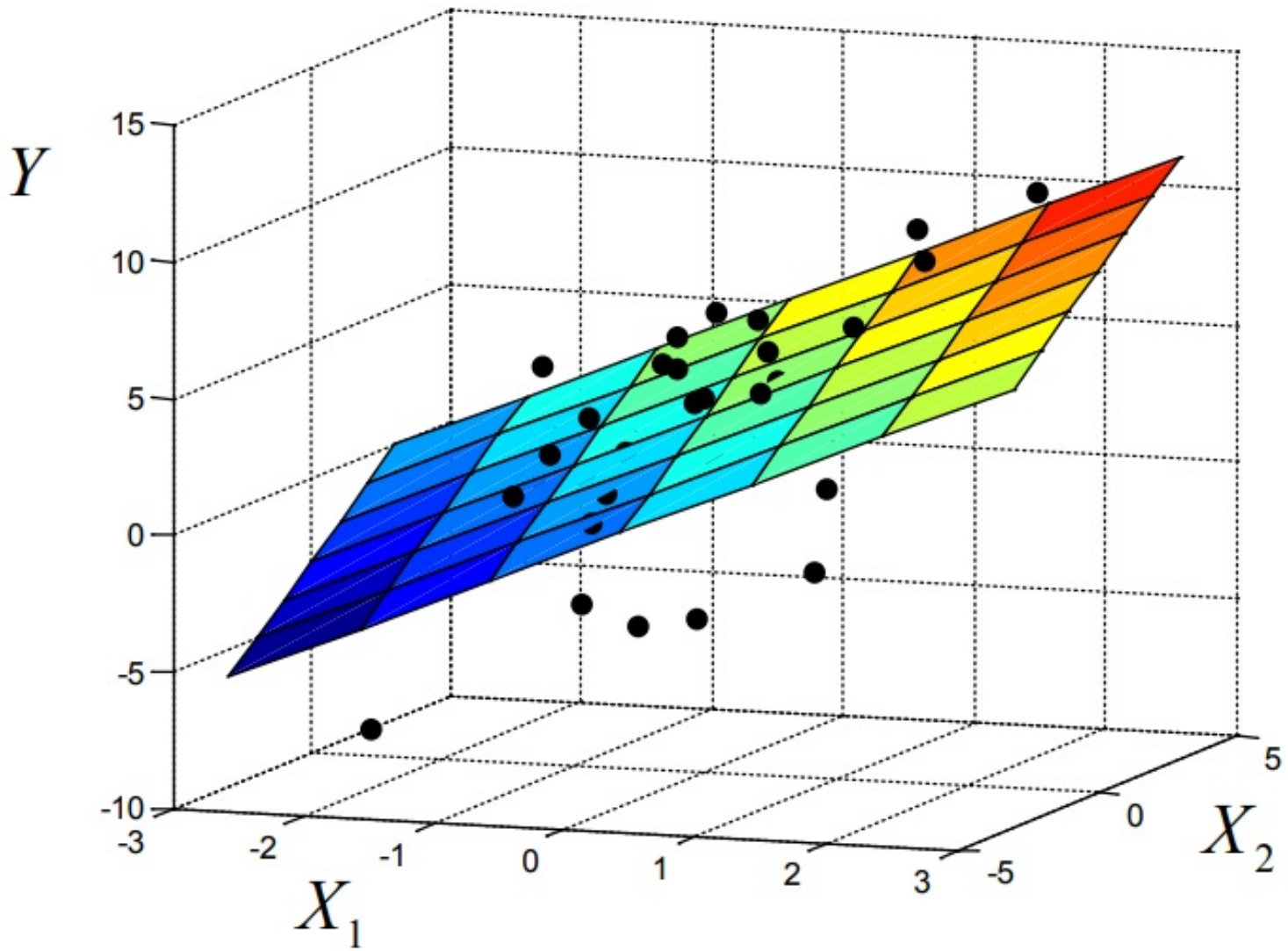


Figure 1: No interaction

Intercations

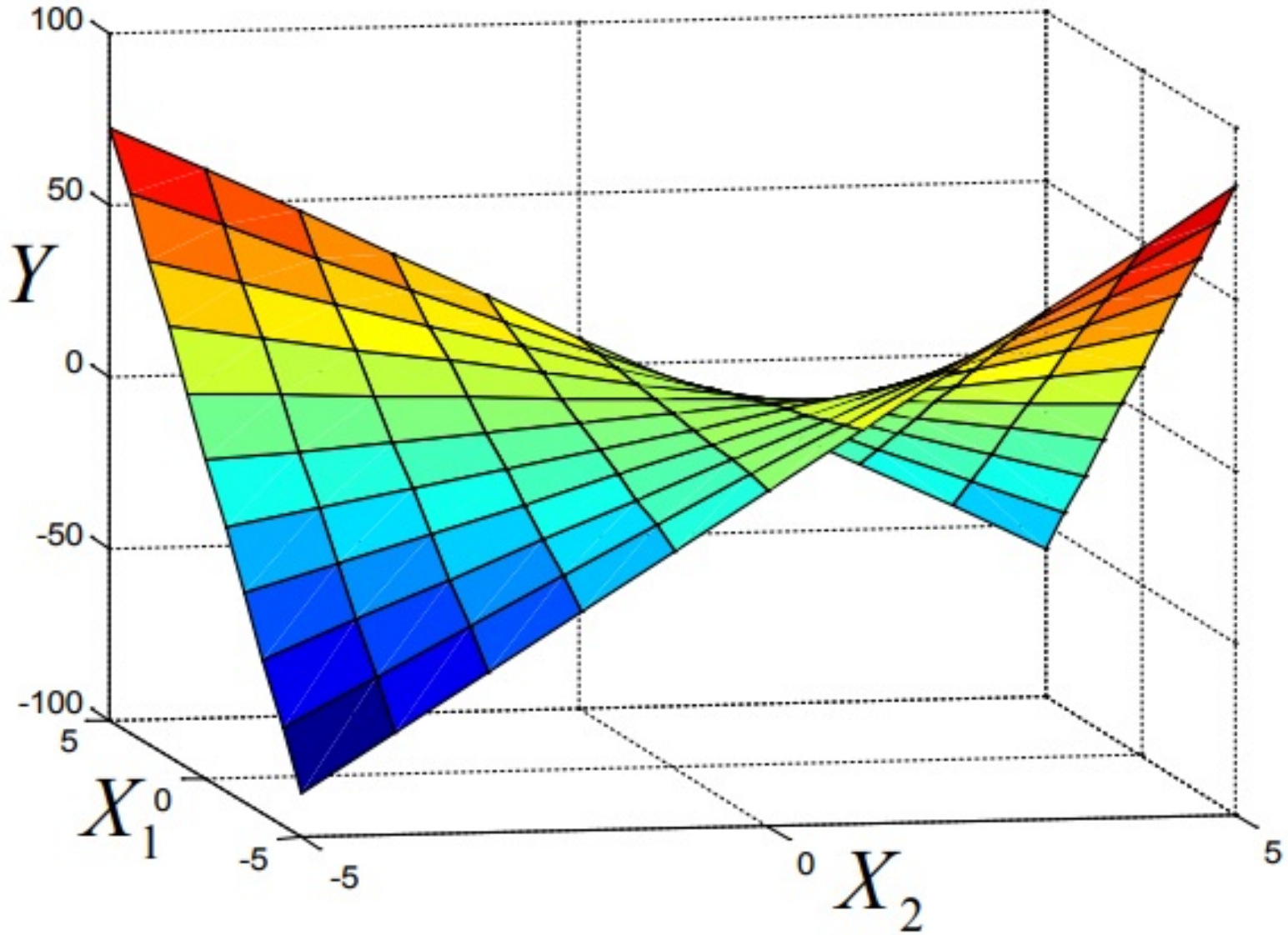


Figure 2: With interaction

Interactions

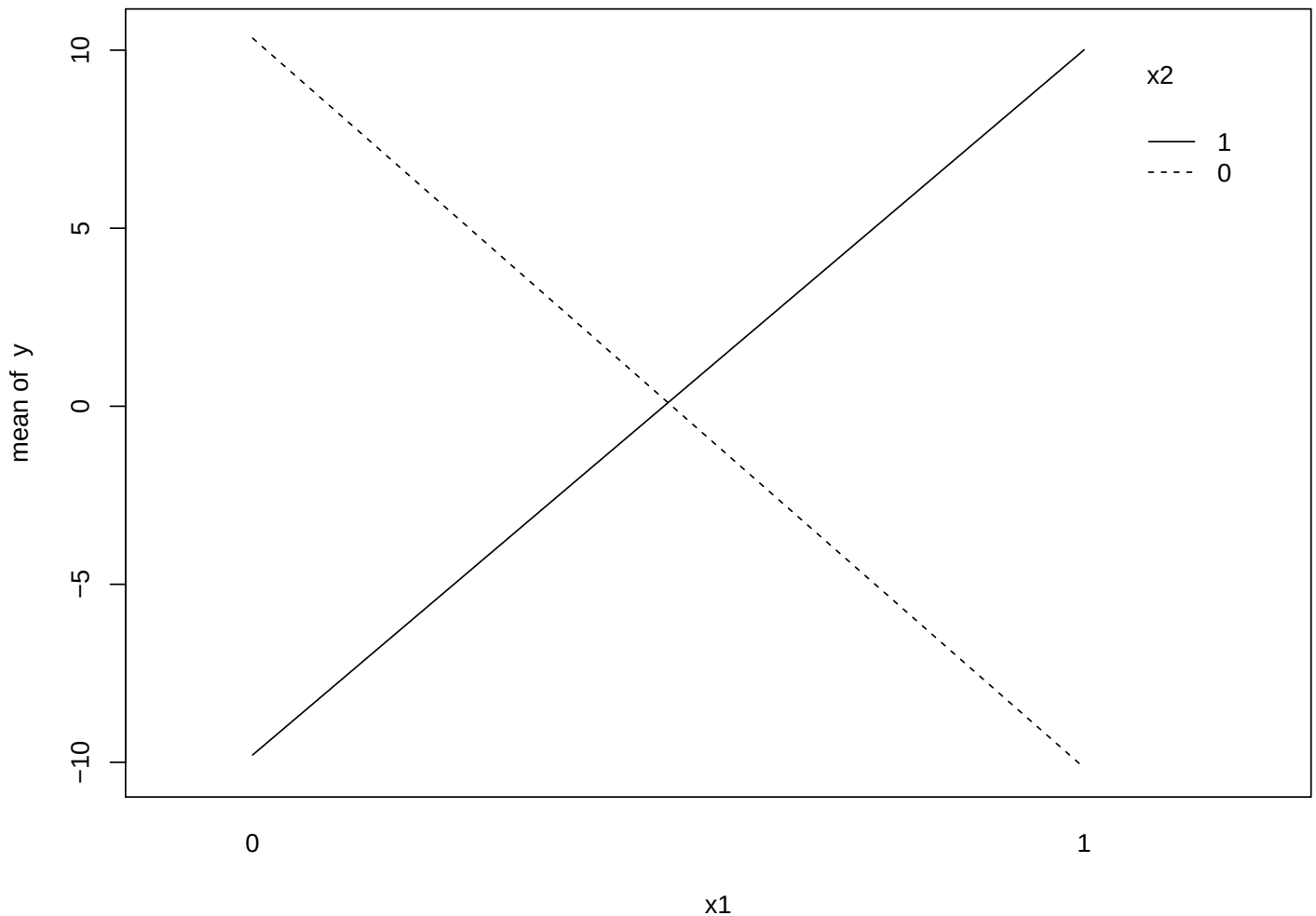
- You can see how important interactions terms are!
- They are complicated to fully understand but they are extremely important to better fit the data.
- Simple plots are not enough to decide if we need them (as shown above).
- The model is completely *wrong* without the interaction term.

Interactions

- It is kind of scary. Should we always add interaction terms ?
- Well, it makes the number of parameter quickly explode!
- If we have p predictors, we have $\binom{p}{2}$ intercation terms. (Technically there is a lot more but we won't talk about it yet)
- So it may not be the right decision to always add all the interactions terms!
- Some graphical tools can help us.

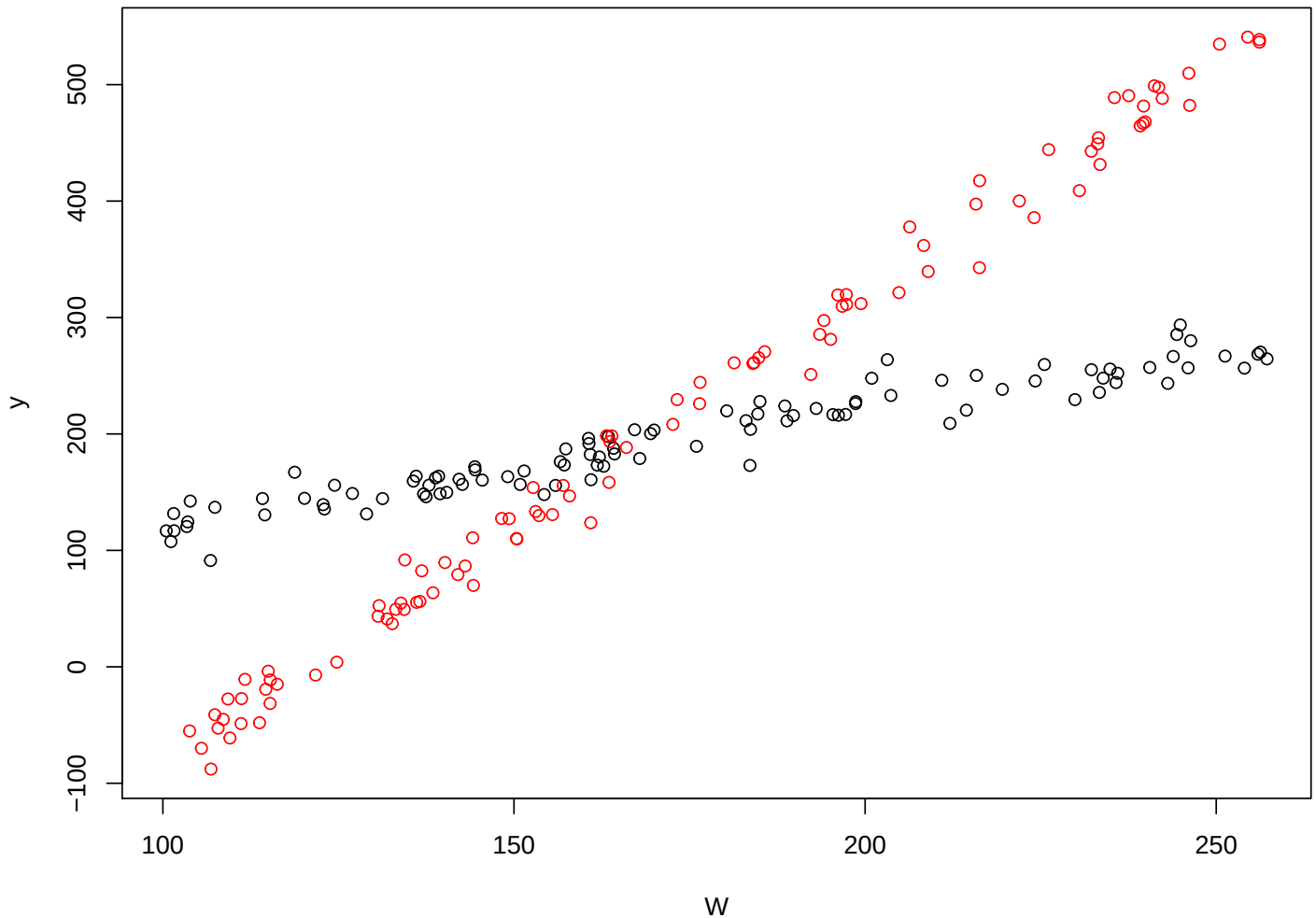
Interactions

```
interaction.plot(x1,x2,y)
```



Interactions

```
plot(y~W,col=as.factor(G))
```



Interactions

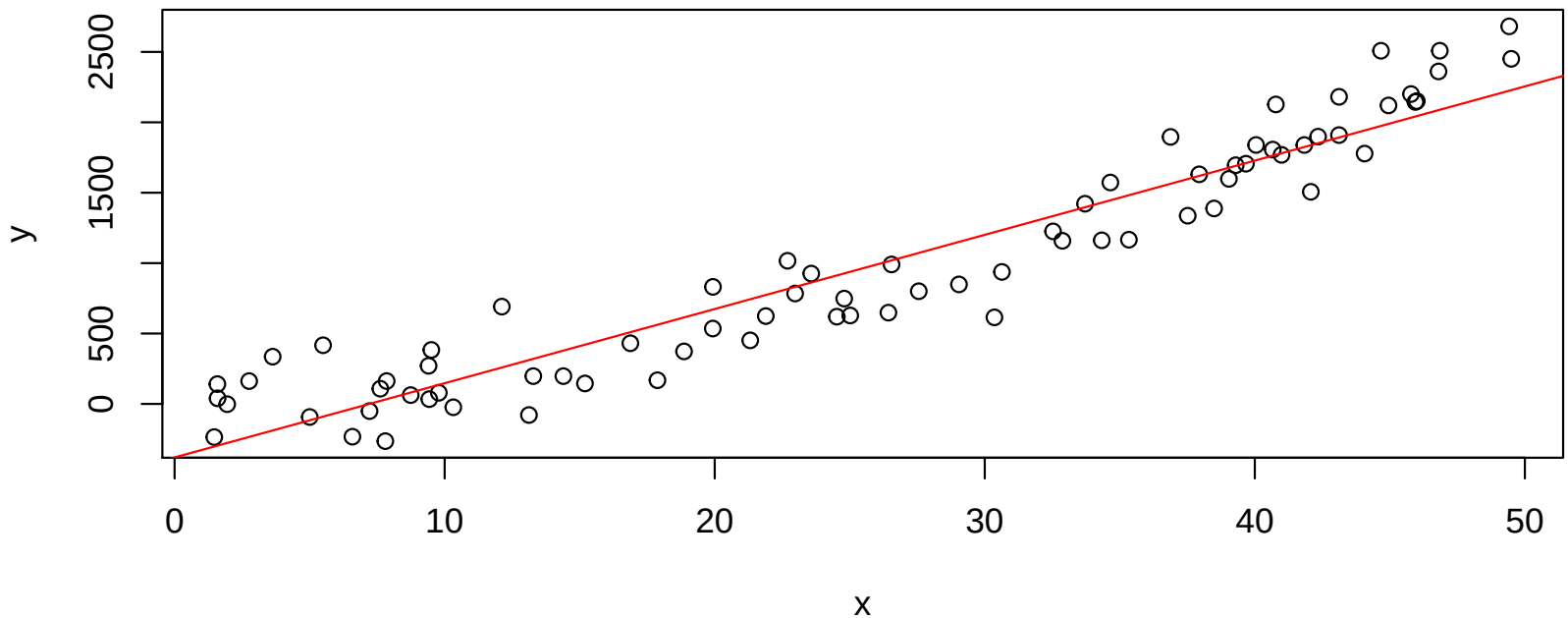
- We could also consider interactions of 3rd order :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{1,2} x_1 x_2 + ... + \beta_{1,2,3} x_1 x_2 x_3$.
- It leads to SO MANY parameters.
- Selecting the variables to keep in a model is a complicated question.
- It is an extremly modern problem now that we have a huge amount of parameters already.
- Model selection will be discussed after test#2.

Polynomial fit

Polynomial fit

- Interaction : As x_1 varies, the effect of x_2 on y is different.
- What if : As x_1 varies, the effect of x_1 on y is different.
- The model with interaction would be $y = \beta_0 + \beta_1 x_1 + \beta_{1,1} x_1 x_1 = \beta_0 + \beta_1 x_1 + \beta_{1,1} x_1^2$.
- A polynomial fit is a model where we predict y using different powers of x.
- It can be understood as a special case of interactions.
- It can also solves the issue of observable pattern in the residuals!

Polynomial fit

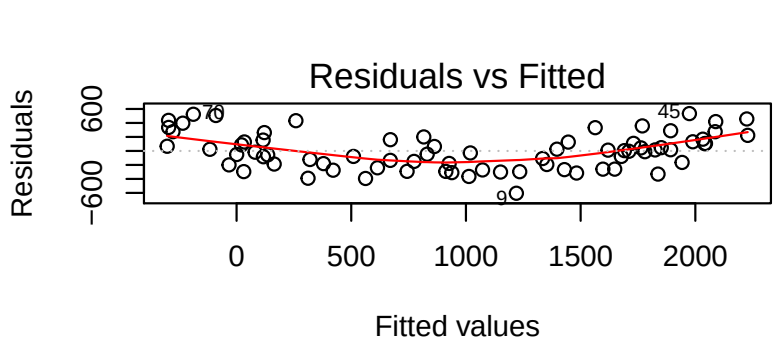


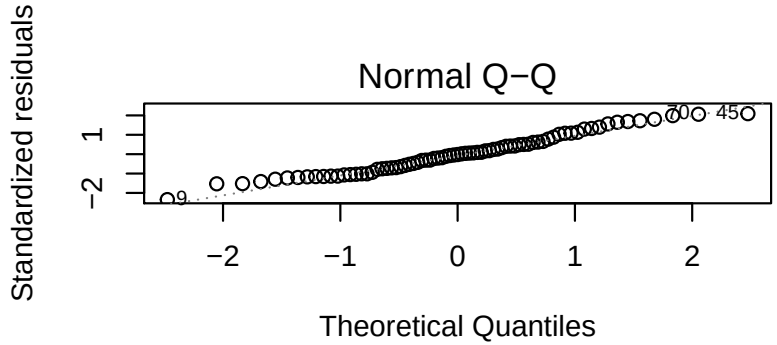
Polynomial fit

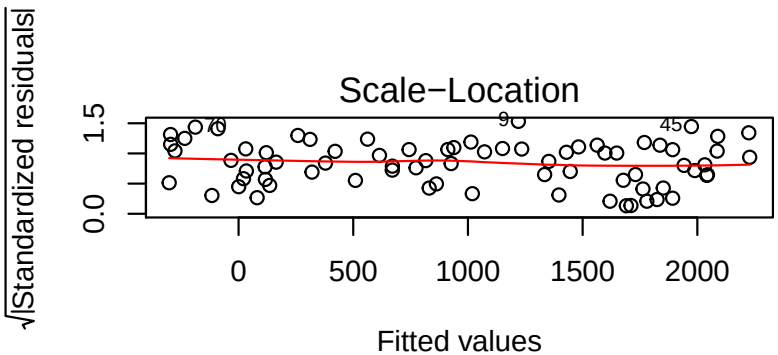
```
summary(model)
```

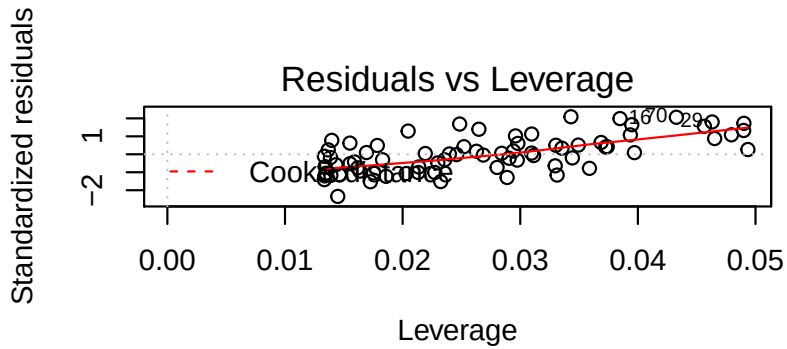
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -605.11 -196.97   -4.81  157.79  533.96
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -379.894     60.286  -6.302    2e-08 ***
## x             52.697      2.013  26.178   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 259.9 on 73 degrees of freedom
## Multiple R-squared:  0.9037, Adjusted R-squared:  0.9024
## F-statistic: 685.3 on 1 and 73 DF,  p-value: < 2.2e-16
```

Polynomial fit



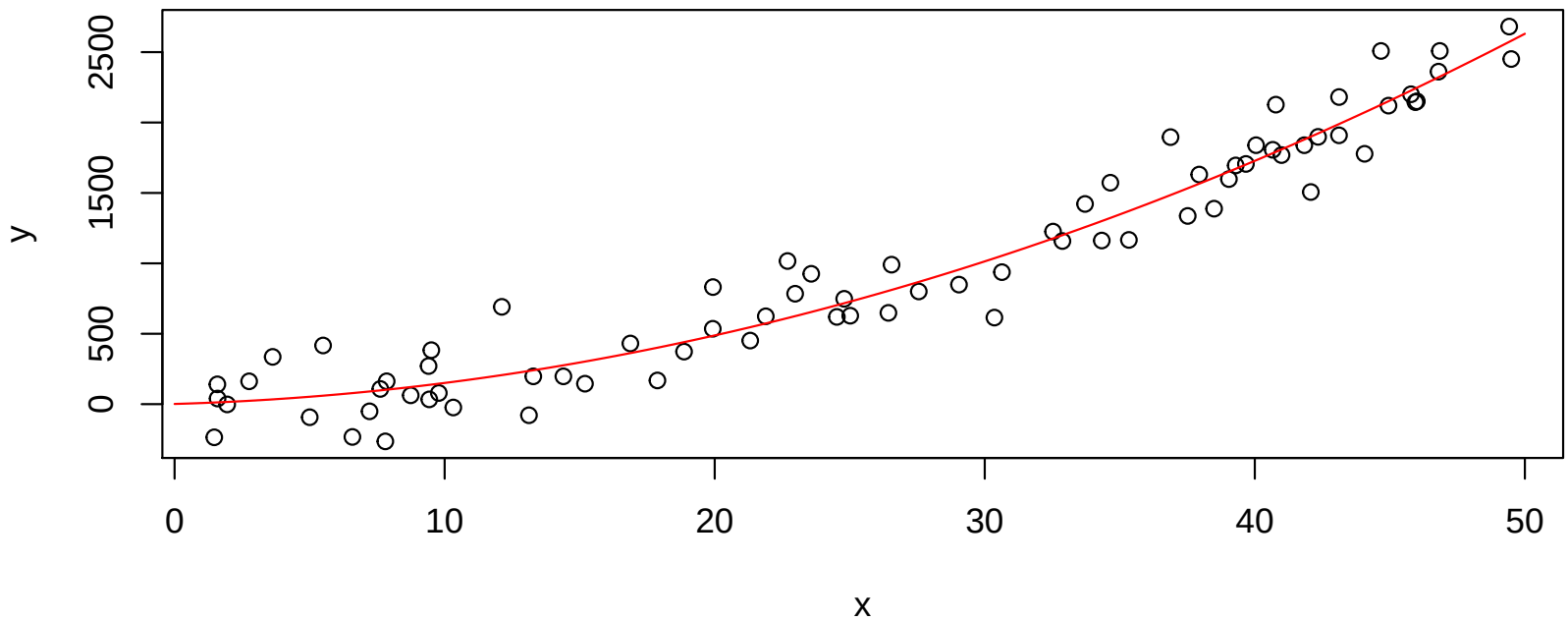






Polynomial fit

- Let's try simply including x^2 as a predictor .

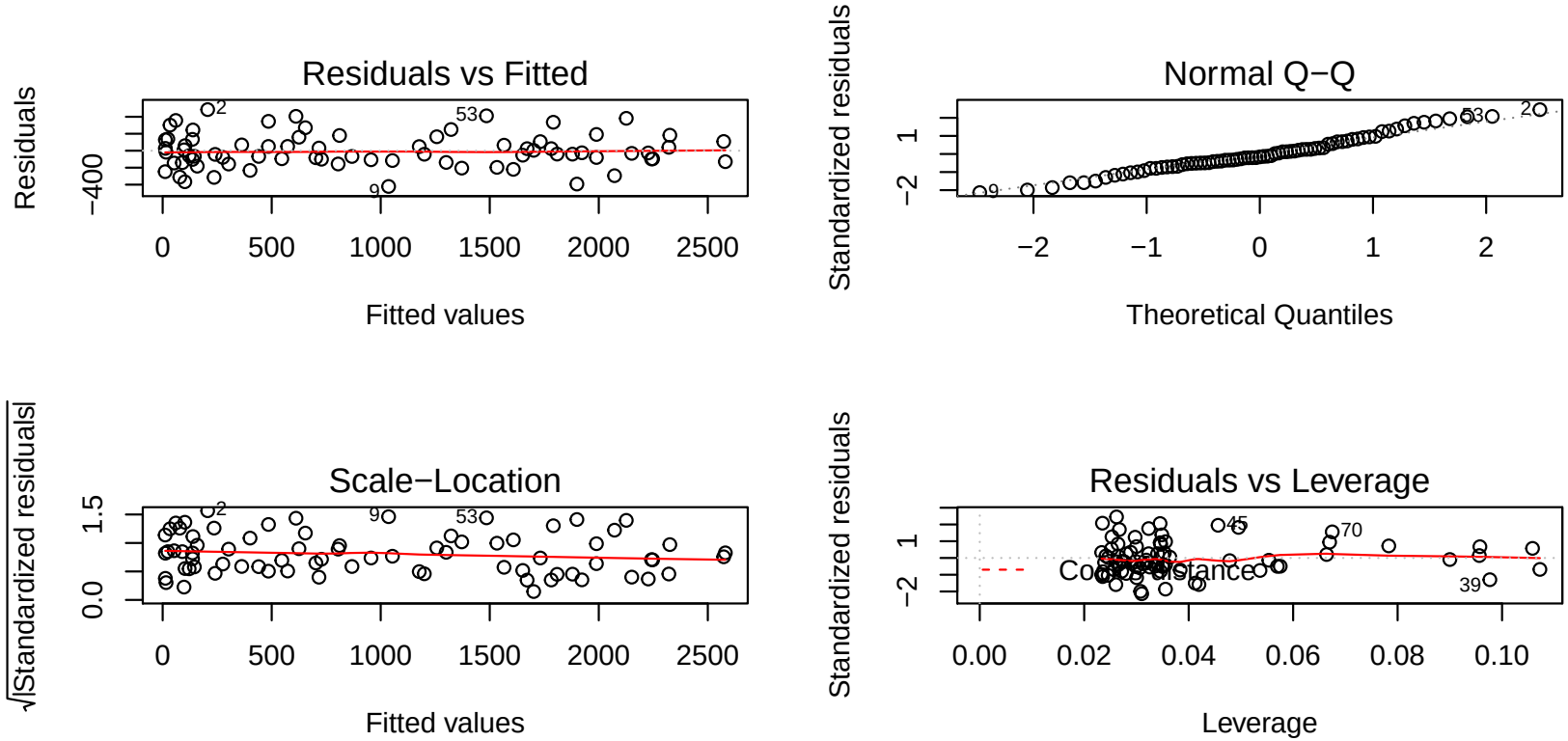


Polynomial fit

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -421.46 -111.46  -31.32  118.43  484.65
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3943    70.8490    0.02    0.984
## x             5.5065     6.7981    0.81    0.421
## x2            0.9415     0.1320    7.13 6.36e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.4 on 72 degrees of freedom
## Multiple R-squared:  0.9436, Adjusted R-squared:  0.942
## F-statistic:   602 on 2 and 72 DF,  p-value: < 2.2e-16
```

Polynomial fit



Polynomial fit

- It is still a linear relationship, y is linear with x^2.
- Once again, why should we stop at x^2 ? Why not x^3 ?
- This is once again a model selection problem.
- It will be discussed in lecture 8 and 9 and it is the last topic for STA302! (yay!)

Model checking

Model checking

- Multiple linear regression shares the same assumptions has simple linear regression and thus model checking is pretty similar.
- We fit residuals against the fitted values to check most the error assumptions. We can also fit the residuals against the predictors (we just have more plots to look at now!)
- We look for leverage points (across ALL predictors) and for outliers.
- I think lecture 5 covered these things already.
- One new problem that have to check for is **collinearity** (sometimes called multicollinearity).

Collinearity

Collinearity

- Intuitively, we add more predictor to a model to increase it's predictive power: we would like the new predictors to bring *new* information to the model.
- Mathematically speaking, we would like the new predictor to be *orthagonal*, or statistically speaking *uncorrolated*.
- Formally speaking it is not an *assumption* of the model, but informally it is.
- When we add a new variable to the model it needs to contain different information than the information contained in the other predictors.

Collinearity

- We can grasp some interpration problem right away.
- Suppose x_1 is a good predictor for y (R^2 is big, β_1 is significance ladelida)
- Suppose also that x_2, a new predictor, is highly corrolated with x_1, then x_2 is also a good predictor for y.
- Now, if both are in the model, how can the model *know* which predictor is *the right* predictor to use.

Collinearity

- Collinearity is the problem of fitting a linear model where two or more predictors are highly corrolated.
- *When two or more highly correlated predictor variables are included in a regression model, they are effectively carrying very similar information about the response variable. Thus, it is difficult for least squares to distinguish their separate effects on the response variable.*

Collinearity

- Alright let's do a bit of Math (Matrix stuff we all need to get better at):
- If we have a matrix $\mathbf{A}$. There exist a series of equivalence defining what a $\det(\mathbf{A})=0$ means.
- If a vector is a linear combinaision of other vectors of the matrix, $\det(\mathbf{A})=0$. (It is known)(The volume of the parallelipiped determined be the vectors)
- If the determinant of a matrix is 0, then the matrix is singular (has no inverse), that's due the definition of the inverse itself commonly defined as $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})$.
- A simple case is when a vector a_i is simply a scaled version of another vector a_j. For example: $a_i = ca_j$.

Collinearity

- Let's assume we have have matrix of predictors $\mathbf{X}$.
- Assume two variables are perfectly correlated $\rho(\mathbf{x_i}, \mathbf{x_j}) = 1(-1)$, it implies a perfect direct increasing (decreasing) linear relationship ($\mathbf{x_i} = a + b\mathbf{x_j}$).
- This implies the $\det(\mathbf{X}) = 0$
- Since $\det(\mathbf{A}^T)$=$\det(\mathbf{A})$, then $\det(\mathbf{X}^T)$=0.
- Finally since $\det(\mathbf{AB})$=$\det(\mathbf{A})\det(\mathbf{B})$ for square matrix of equal size we have $\det(\mathbf{X}^T\mathbf{X})$= 0.
- It implies $\mathbf{X}^T\mathbf{X}$ has no inverse.

Collinearity

- *Is that big deal ?* Yes
- Since $\mathbf{X}^t\mathbf{X}$ has no inverse, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$ does not exist.
- From a simple mathematical perspective, if the predictors are perfectly correlated, the least square estimate (and maximum likelihood estimate) does not exist.

Collinearity

- Well perfect corrolation shouldn't happen.
- You can imagine that a correlation close to 1 (-1) is problematic on its own.
- As the correlation between two predictors become closer and closer to 1 (-1), the determinant of $\mathbf{X}^T\mathbf{X}$ becomes closer and closer to 0 and the $\text{Var}[\hat{\beta}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ becomes extremely large. ($\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})$)
- Not only it is undesirable to have a large variange for our estimates but it also leads to model were all the parameters are non significant (due to the large variance).

Collinearity

- To illustrate the problems and checking procedure we will import the data set introduced in Linear Models with R (ch.5).

```
library(faraway)
data(seatpos)
```

Collinearity

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Collinearity

- All of the parameters are not significantly different from 0.
- *So the predictors are useless at predicting y ?*
- Well no, R^2 is quite large, so our model does provide a significant improvement compared to $\bar{y}$.
- So now that we are convinced that collinearity is a problem what should we do ?
- Like for model checking, we should at least check for collinearity.

Collinearity

- How to check for collinearity :
  - We can look at the correlation matrix of the predictors. Large pairwise correlation indicates a problem.
  - Build a regression model for x_i on all other predictors to asses R_i^2. When computing that value for all predictors, high R_i^2 (close to one) indicates a problem.
  - We can look at the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Small eigenvalues indicates a problem.

Collinearity

- Here is the correlation matrix :

```
library(faraway)
data <- seatpos
round(cor(seatpos[,1:(ncol(seatpos)-1)]),3)
```

```
##              Age Weight HtShoes     Ht Seated   Arm Thigh    Leg
## Age        1.000  0.081  -0.079 -0.090 -0.170 0.360 0.091 -0.042
## Weight     0.081  1.000   0.828  0.829  0.776 0.698 0.573  0.784
## HtShoes   -0.079  0.828   1.000  0.998  0.930 0.752 0.725  0.908
## Ht        -0.090  0.829   0.998  1.000  0.928 0.752 0.735  0.910
## Seated    -0.170  0.776   0.930  0.928  1.000 0.625 0.607  0.812
## Arm        0.360  0.698   0.752  0.752  0.625 1.000 0.671  0.754
## Thigh      0.091  0.573   0.725  0.735  0.607 0.671 1.000  0.650
## Leg       -0.042  0.784   0.908  0.910  0.812 0.754 0.650  1.000
```

Collinearity

- Should we include the response as well, some say yes :

```
library(faraway)
data <- seatpos
round(cor(seatpos),2)
```

```
##               Age Weight HtShoes    Ht Seated   Arm Thigh   Leg hipcenter
## Age          1.00   0.08   -0.08 -0.09  -0.17  0.36  0.09 -0.04      0.21
## Weight       0.08   1.00    0.83  0.83   0.78  0.70  0.57  0.78     -0.64
## HtShoes     -0.08   0.83    1.00  1.00   0.93  0.75  0.72  0.91     -0.80
## Ht          -0.09   0.83    1.00  1.00   0.93  0.75  0.73  0.91     -0.80
## Seated      -0.17   0.78    0.93  0.93   1.00  0.63  0.61  0.81     -0.73
## Arm          0.36   0.70    0.75  0.75   0.63  1.00  0.67  0.75     -0.59
## Thigh        0.09   0.57    0.72  0.73   0.61  0.67  1.00  0.65     -0.59
## Leg         -0.04   0.78    0.91  0.91   0.81  0.75  0.65  1.00     -0.79
## hipcenter    0.21  -0.64   -0.80 -0.80  -0.73 -0.59 -0.59 -0.79      1.00
```

Collinearity

- Now let's look at the eignen value of $\mathbf{X}^T\mathbf{X}$.
- Values close to 0 indicate a problem. (If an eigenvalue = 0, the determinant is 0 as well).

```
X <- seatpos[,1:(ncol(seatpos)-1)]
X <- as.matrix(cbind(rep(1,nrow(X)),X))
v <- eigen(t(X)%*%X)
v$values
```

```
## [1] 3.653709e+06 2.147979e+04 9.043226e+03 2.989641e+02 1.484005e+02
## [6] 8.117464e+01 5.336195e+01 7.298226e+00 5.127424e-02
```

Collinearity

- The last tool we are going to introduce is R_i^2 which is the R^2 coefficient when x_i is considered the response.

```
X <- as.matrix(seatpos[,1:(ncol(seatpos)-1)])
summary (lm(X[,1]~X[,-1]))$r.squared
```

```
## [1] 0.4994823
```

Collinearity

```
X <- as.matrix(seatpos[,1:(ncol(seatpos)-1)])
np <- ncol(X)
R <- rep(0,np)
for (i in 1:np) {
  R[i] <- summary (lm(X[,i]~X[,-i]))$r.squared
}
R
```

```
## [1] 0.4994823 0.7258043 0.9967472 0.9969982 0.8882813 0.7775983 0.63
## [8] 0.8506190
```

- Wow, these R^2 coefficients are so large. We can almost perfectly predict some of the predictors using the other one.

Collinearity

- I think the vector of R_i^2 is good enough but some statisticians like to talk about the *Variance Inflating Factors* (VIFs).
- Here's the idea:
- We can actually show that for the following model :

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + ... + \beta_p \mathbf{x_p} + \mathbf{e}$$

we have :

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)SSX_j^2}$$

- As expected, as the correlation between the predictors goes up (expressed by the R^2 coefficients) so does the variance for the estimates of a single parameter.

- $\frac{1}{1-R_j^2}$ are defined as the VIFs

Collinearity

```
round(R,4)
```

```
## [1] 0.4995 0.7258 0.9967 0.9970 0.8883 0.7776 0.6381 0.8506
```

```
VIF <- 1/(1-R)
round(VIF,4)
```

```
## [1]   1.9979   3.6470 307.4294 333.1378   8.9511   4.4964   2.7629
```

Collinearity

- The faraway package also offers direct VIFs computation tools:

```
vif(X)
```

```
##        Age     Weight    HtShoes         Ht     Seated        Arm
##   1.997931   3.647030 307.429378 333.137832   8.951054   4.496368
##      Thigh        Leg
##   2.762886   6.694291
```

Collinearity

- Now that we do understand that collinearity is a grave problem.
- We have tools (3 at least) to asses the collinearity of a data set.
- *How to fix it ?* Well... there is no real trick.
- Some suggest amputation.

Collinearity

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Collinearity

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.526 -23.005   2.164  24.950  53.982
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 528.297729 135.312947   3.904 0.000426 ***
## Age           0.519504   0.408039   1.273 0.211593
## Weight        0.004271   0.311720   0.014 0.989149
## Ht           -4.211905   0.999056  -4.216 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.49 on 34 degrees of freedom
## Multiple R-squared:  0.6562, Adjusted R-squared:  0.6258
## F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

Collinearity

- We only kept one of the 6 strongly correlated variable and have a model almost as good (according to the R^2 value).
- It's not a flattering property for Linear models. Ideally we would like a model that understand the correlation and use it to strengthed it's prediction!
- It leads to inference problems as well :
- *We pick height as the simplest to measure. We are not claiming that the other predictors are not associated with the response, just that we do not need them all to predict the response*
- How to actually amputate the data is a complicated question.
- We will cover Model and Variables selection next week as our last topic of the semester!

Collinearity

- It is at least important to check for collinearity.
- Document/mention it.
- Understand it's effect on parameters estimates.
- Proceed with amputation if we have a good way to select some predictors that are deemed more interesting.

Conclusion

- Interactions are AMAZING. They must be added to most MLR models as they can capture structures that would be missed otherwise.
- That are complicated to interpret but that really is our job. Your collaborators will have trouble understanding interaction, you need to be able to explain them.
- Using many predictors to model the response is an improvement and a great extension to the Simple Linear Regression. Most of the model checking procedure is the same but one new problem might arise: collinearity.
- Collinearity is caused by highly correlated predictors and greatly increases the variance of the estimates.
- We have to check for collinearity whenever we fit a MLR model.

Practice Problems

- A Modern Approach to Regression with R ch.6 : 1,2,3,4 and 5 Solutions here

External Sources

- Craig Burkett's Chapter 8 Notes
- Linear Models with R ch.5

Test2

- Rooms : A-T BA1160, W-Z BA1170 (first letter on your student ID)
- Duration : 2h50
- Nothing is allowed, no calculators, notes nor cheat sheets.
- Extremely similar to Test#1 in style. (Should be reassuring to you)
- Types of questions : Some conceptual and interpretation questions (30%), some math (30%) and some R output (40%).
- Topics :
  - Predictive inference (25%)
  - Model checking (25%)
  - Multiple linear regression (50%)