STA302 - Lecture 6

Cedric Beaulac

May 28, 2019

Introduction

Announcements

- You will receive your grades today. (The average is 73.5, is it symetrical and all! That's great! I'm great!)
- New office hours schedule begin next week. (Mon-Wed 5-7)
- It gives you an extre office hour before test#2!

Today's plan

- Today :
  - Review of the model and diagnostic
  - Transformations
  - Dummy variables (and its effect on the intercept)
  - Multiple linear regression

Review of the model

Review

- We have established this basic model: $y_i = \beta_0 + \beta_1 x_i + e_i$, where e_i are independantly ditributed $\sim N(0, \sigma^2)$.
- $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ which implies $\mathbf{y} \sim N(\mathbf{X}\beta, I\sigma^2)$.
- We estimate β with $\hat{\beta}_{MLE} = (\mathbf{X}^T\mathbf{X})^1\mathbf{X}^T\mathbf{y}$

Review

- We have also establish a list of important assumptions related to this model.
- The most constraining asusmptions is with respect to the errors e_i. In our model we assumed $e_i \sim N(0, \sigma^2)$. Most model checking for the errors are based on the residuals (the observed errors) $\hat{e}_i = \hat{y}_i - y_i$.
- By plotting the residuals against variables (the responses for example) we can check the *constant variance* assumption.
- And we can use QQ-plots to check the normality of the residuals.
- We must also check for uncorrelatedness.

Review

- We also spent some times talking about unusual observations; outlier, leverage points and influential points.
- Leverage points are points whose x-value is far from other observed x-values.
- Outliers are points whose y-value is far from other observed y-values.
- Influential points are observations that drastically change the parameters estimates. A outlier with large leverage is a good exemple.
- Looking at the plots of y against x, the fitted line is helpful.
- The Cook's distance as well.

Transformation

Transformation

- Well, assumptions are violated! The promises of a good model are gone (the cake is a lie).
- This is the end of linear model.
- ....
- Not so fast!

Transformation

- We can use transformations to fix 2 problems :
  - Non-constant variance
  - Non-linearity
- There is no guarantee that the transformation will work.
- STA303 is all about transformations (through link functions)

Transformation

- Recall the varius plots of residuals against response introduced last lecture and notice how different are the violations.

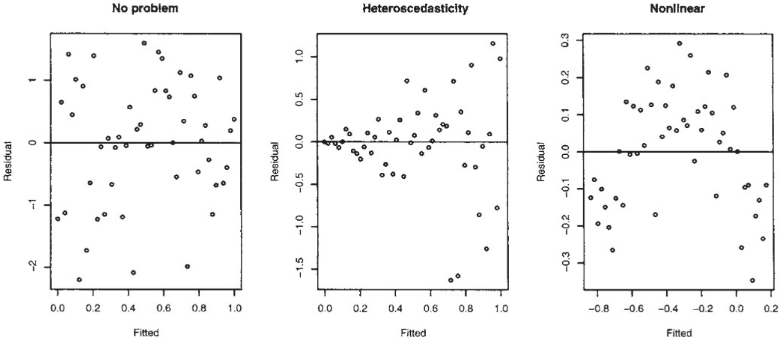


Figure 1: Residuals against fitted

Transformation

- When there is a clear pattern, we need a new model (next week and STA303).
- But when the variance is *exploding* we can consider a transformation.
- Typically we will raise y to a power between 0 and 1 or apply a logarithmic transformation.

$$\log(y_i) = \beta_0 + \beta_1 x_i + e_i$$

Transformation

- This common transformation reduces large values of y and tends to fix some non-constant variance issues.
- One might notice that :

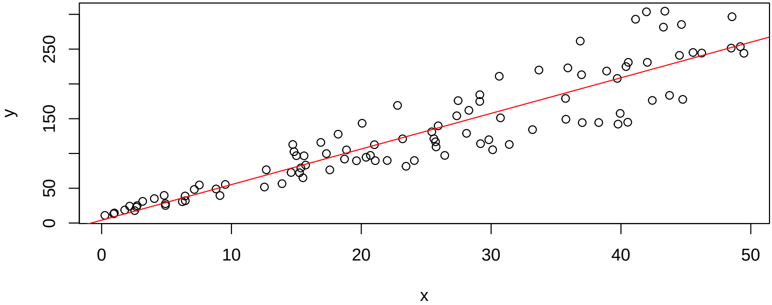
$$\begin{aligned} \log(y_i) &= \beta_0 + \beta_1 x_i + e_i \\ \Rightarrow y_i &= \exp(\beta_0)\exp(\beta_1 x_i)\exp(e_i) \end{aligned}$$

- If the error is multiplicative with $\exp(\beta_0 + \beta_1 x_i)$ on the y scale then maybe this is why the variance was *exploding*.

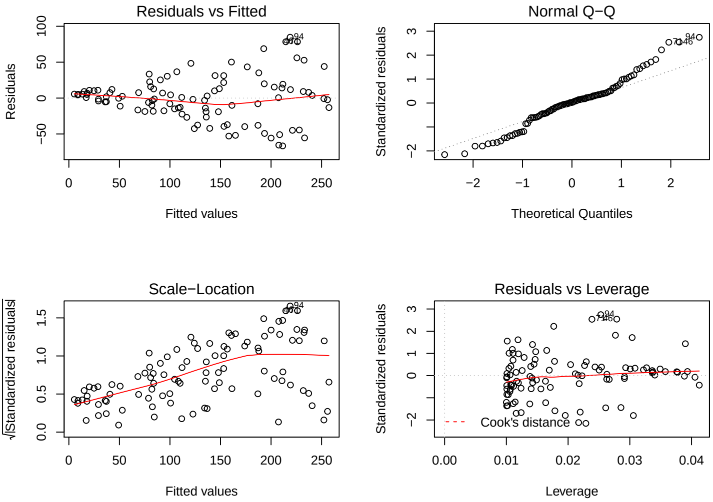
Transformation

- Is it still a linear model ?
- Yes $\log(y)$ has a linear relationship with x.
- Let's give it a shot!

Transformation

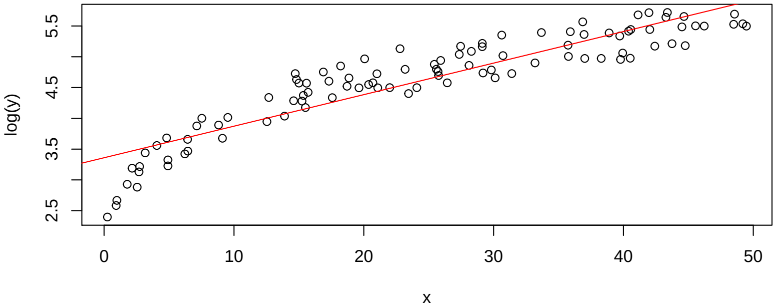


Transformation

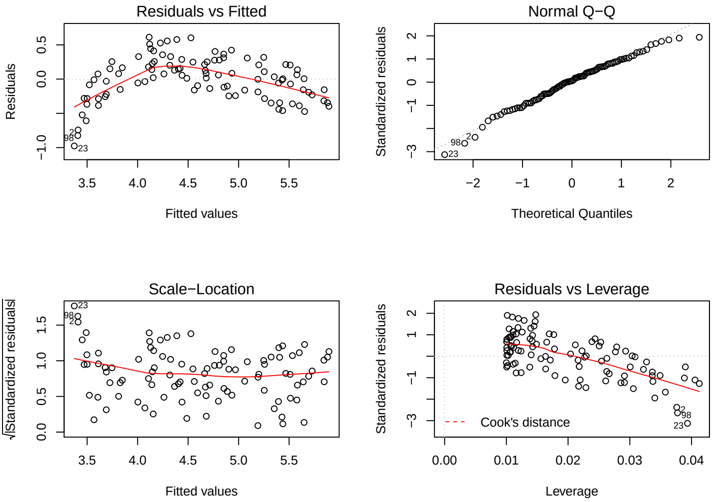


Transformation

- Alright, the variance is exploding let's transform the data!



Transformation



Transformation

- But I'm telling you it works sometimes!
- Are you convinced ? I'm not.
- I didn't want to teach transformations.
- Rigourous approaches aren't really convincing. People tend to try multiple transformations until they find one. It's ok I guess... (I think it is better when we find a automatic way (using the data)to select the transformation)
- Let's introduce the Box-Cox transformation. Is is most used procedure for automatic transformation *selection*.

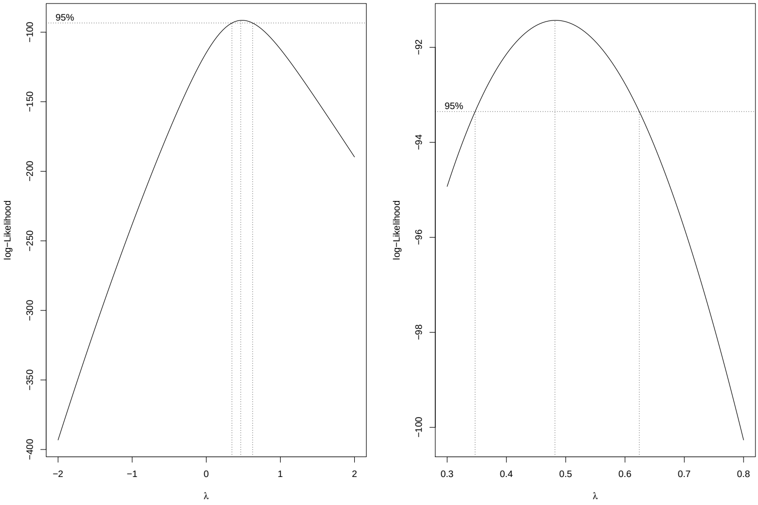
Transformation : Box-Cox

- Let's consider a family of possible transformation $g_\lambda(y)$.

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

- Let's select λ according to the model achieving the highest log-likelihood.
- Let's not compute all of those values on our own and use R instead.

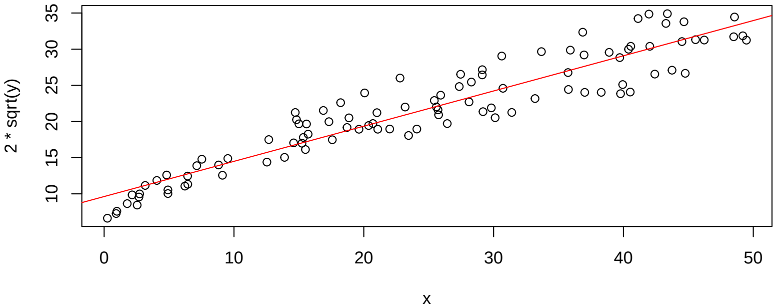
Transformation : Box-Cox



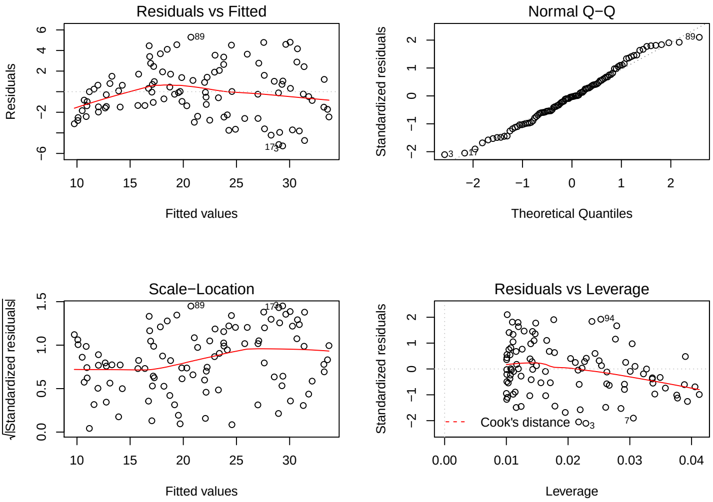
Transformation : Box-Cox

- If explaining the model is important, you should round λ to the nearest interpretable value.
- Let's pick $\lambda = 0.5$ and thus use the following transformed model :$\sqrt{y} = \beta_0 + \beta_1 x + e$.

Transformation : Box-Cox



Transformation : Box-Cox



Transformation : Conclusion

- Transforming the response (or the predictors) might help with violated assumptions.
- We have no guarentee it will work. But honestly, it works from time to time (textbook examples and exercises) and it's worth a try.
- It makes the model less interpretable.
- It is a debatable approach, but we had to talk about it.
- We are not done with transformations. We will discuss polynomial fit later as a special transformation to modify the model itself to fix residuals with a clear pattern.

Dummy variable and introduction to Multiple Linear Regression

Dummy variables

- Reminder : Go slowly!
- A set of dummy variables is a set of binary variables established to represent a categorical variable.
- It is a set of indicator variables.
- It allows to fit a parameter for every possible categories of a categorical variable.

Dummy variables

- For a simple categorical variable representing two groups : A or B, we can represent it with a binary variable such that :

Groupe	x
A	0
B	1

Dummy variables

- We could fit a linear regression on the dummy variable with our usual model :

$$y = \beta_0 + \beta_1 x + e$$

,

where $e \sim N(0, \sigma^2)$.

In this model we have $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$. Remember x is a binary variables so this model implies :

$$\mathbf{E}(y) = \begin{cases} \beta_0 & \text{if } x = 0 \text{ (Group = A)} \\ \beta_0 + \beta_1 & \text{if } x = 1 \text{ (Group = B)} \end{cases}$$

Dummy variables

- So $\mu_A = \beta_0$ and $\mu_B = \beta_0 + \beta_1$.
- This is a simple model that only consist of *two different intercepts*. There is no *slope* because there is no continuous predictor.
- Remember the t-test for β_1 tests the null $H_0 : \beta_1 = 0$
- This is equivalent to testing $H_0 : \mu_A = \mu_B$.
- The response is normally distributed and we assumed a fixed variance σ for all observations.
- This is EXACTLY the two sample t-tests established in lecture 2! (You're suppose to be surprised and/or impressed)

Dummy variables

```r
A <- rnorm(n=100,2,5)
B <- rnorm(n=100,0,5)
t.test(A,B, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  A and B
## t = 2.628, df = 198, p-value = 0.009262
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4752019 3.3322475
## sample estimates:
## mean of x mean of y
##  2.106674  0.202949
```

Dummy variables

```r
y <- c(A,B)
x <- c(rep(0,100),rep(1,100))
summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5981  -3.4061  -0.1971   4.0492  14.0484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1067     0.5122   4.113 5.73e-05 ***
## x            -1.9037     0.7244  -2.628  0.00926 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.122 on 198 degrees of freedom
## Multiple R-squared:  0.03371,    Adjusted R-squared:  0.02883
## F-statistic: 6.906 on 1 and 198 DF,  p-value: 0.009262
```

Dummy variables

- *Ok ok, so you're telling me we could use a dummy variable linear model to replace t.test ?* Yes! They are equivalent tests.
- Guess what ? We can also totally replace ANOVA!

Dummy variables

- For a simple categorical variable representing three (or T) groups : A, B and C we can represent it with a 2 (or T-1) binary variables such that :

Group	x_1	x_2
A	0	0
B	1	0
C	0	1

- Let's fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$.
- Heu... How do we do that ?

Multiple Linear Regression

Multiple Linear Regression

- Suppose we have more than one predictor. Let say p.
- The new model is $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + e$.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ . \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & ... & x_{1,p} \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n,1} & ... & x_{n,p} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ \beta_p \end{bmatrix}$$

- The model is $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{e} \sim MVN(0, I\sigma^2)$ (a vector of independant Normal variables with mean 0 and variance σ^2).
- It leads to $\mathbf{y} \sim N(\mathbf{X}\beta, I\sigma^2)$.

Multiple Linear Regression

- Let's get the parameters using maximum likelihood.

$$l(\theta|x_i) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- Maximizing this term with respect to β is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{d}{d\beta}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \hat{\beta}$$

Multiple Linear Regression

- *Ok ok, so you're telling me Multiple Linear Regression is exactly like Simple Linear regression ?* Yes! This is why we used the matrix notation in the first place.
- We talked about ANOVA earlier, let's take a look at it.

Multiple Linear Regression

- For a simple categorical variable representing three (or T) groups : A, B and C we can represent it with a 2 (or T-1) binary variables such that :

Groupe	x_1	x_2
A	0	0
B	1	0
C	0	1

- with $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$.

Multiple Linear Regression

- $\mu_A = \beta_0$, $\mu_B = \beta_0 + \beta_1$ and $\mu_C = \beta_0 + \beta_2$.
- So the individual test $H_0 : \mu_B = 0$ checks if Group B is different from Group A.
- To check if all the groups are the same (the predictors has no effect) we would need to check $H_0 : \beta_1 = \beta_2 = 0$.
- This test is also performed by R, let's take a look.

Multiple Linear Regression and ANOVA

```r
A <- rnorm(n=100,2,5)
B <- rnorm(n=100,0,5)
C <- rnorm(n=100,1,2)
y <- c(A,B,C)
x <- c(rep(0,100),rep(1,100),rep(2,100))
anova(lm(y~as.factor(x)))
```

```
## Analysis of Variance Table
##
## Response: y
##               Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(x)   2  476.6  238.31  13.152 3.367e-06 ***
## Residuals    297 5381.7   18.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Linear Regression and ANOVA

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1502  -2.2129  -0.2248   2.6778  16.3632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2093     0.4257   5.190 3.90e-07 ***
## x1           -3.0428     0.6020  -5.054 7.56e-07 ***
## x2           -1.0683     0.6020  -1.775    0.077 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.257 on 297 degrees of freedom
## Multiple R-squared:  0.08136,    Adjusted R-squared:  0.07517
## F-statistic: 13.15 on 2 and 297 DF,  p-value: 3.367e-06
```

Multiple Linear Regression

- This test checks if $\beta_1 = \beta_2 = ... = \beta_p = 0$.
- It's the same as ANOVA $\frac{SSreg/p-1}{SSE/n-p}$ (*new notation).
- Actually the bottom part of the R output is an analysis of sum of squares.(Remember R^2 is also there)
- The test checks if the fitted model is an improvements over $\bar{y}$.

Sum of squares revisited

- SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sum across all observations of squarred distance between the observation and the base prediction (the mean)
- SSG=SSreg $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the sum across all observations of squarred distance between the regression prediction and the base prediction.
- It is the *explained* variation. How much of our explanation takes us away from the base prediction.
- SSE $= \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ is the sum across all observations of squarred distance between the observation and the regression prediction.
- It is the *unexplained* variation. How much our model prediction is away from the true observation. This distance is unexplained by the model.

Multiple Linear Regression

- So MLR is a great, it works for both continuous and categorical predictors.
- It is easy to use.
- It was easy to extend the estimation for more than one predictor.
- The tests and such are easy to interpret.

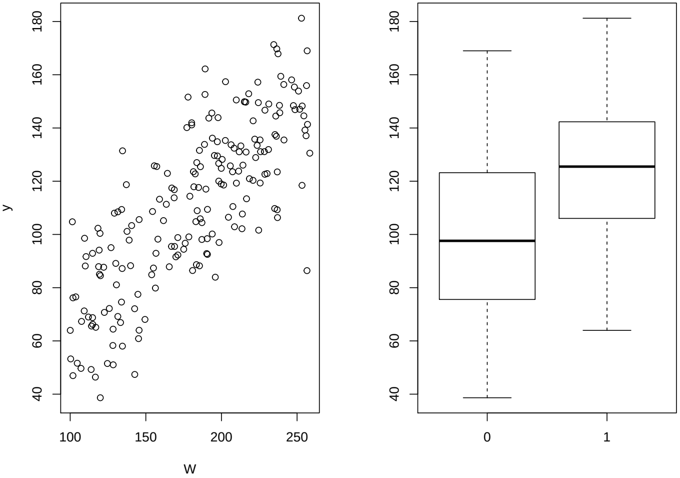
Multiple Linear Regression

- Let's introduce one last example.
- What if we have one continuous and one categorical predictor and use the simple model we defined.
- *Let's predict the weight given the height and the biological gender*
- y: the response is the weight. x_1 is the height and x_2 is a dummy variable representing biological gender ($0 =$ male, $1 =$ female).

Multiple Linear Regression

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$
- For a fix weight x
  - If the observation is a male, then $\mathbf{E}(y) = \beta_0 + \beta_1 x$
  - If the observation is a female then $\mathbf{E}(y) = \beta_0 + \beta_1 x + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x$.
- So the categorical predictor actually moves the intercept.

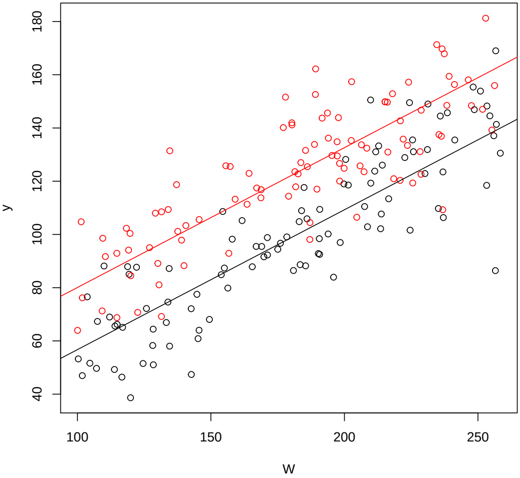
Multiple Linear Regression



Multiple Linear Regression

```
##
## Call:
## lm(formula = y ~ W + G)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.569 -10.748   0.374  10.173  36.102
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.17913    4.47897   0.933    0.352
## W            0.52547    0.02341  22.450   <2e-16 ***
## G           23.34012    2.12160  11.001   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15 on 197 degrees of freedom
## Multiple R-squared:  0.7648, Adjusted R-squared:  0.7624
## F-statistic: 320.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression



Multiple Linear Regression

- Here we assumed the two predictors were independent. We assumed their effect were additive (not multiplicative).
- But what if the groups have 2 different slopes ?
- These things are called interactions (the gender interacts with the weight in its effect on height)
- We will introduce this topic next lecture.

Conclusion

- Transormation of variables is the usual recommandation when assumptions are not respected.
- Sometimes it works (look into the textbooks) sometimes it doesn't (in the slides).
- The idea of transformation will be extended in STA303.
- We can use dummy variables to represent categorical variables in the linear regression set up.
- The tests are equivalent and easy to interpret.
- We seeminglessly extended our model to multiple predictors The intuition is preserved.

Practice Problems

- A Modern Approach to Regression with R ch.3 problem : 4,5 (solutions)
- Alison Gibbs' additional chapter 3 practice problems : 2 (here)
- A Modern Approach to Regression with R problem of p. 138

External Sources

- A Modern Approach to Regression with R ch.3
- Linear Models with R ch.7
- A Modern Approach to Regression with R ch.2 (section 2.6)