Week 7: Interpretting Regression Equations

[brouwern@gmail.com](mailto:brouwern@gmail.com)

February 23, 2017

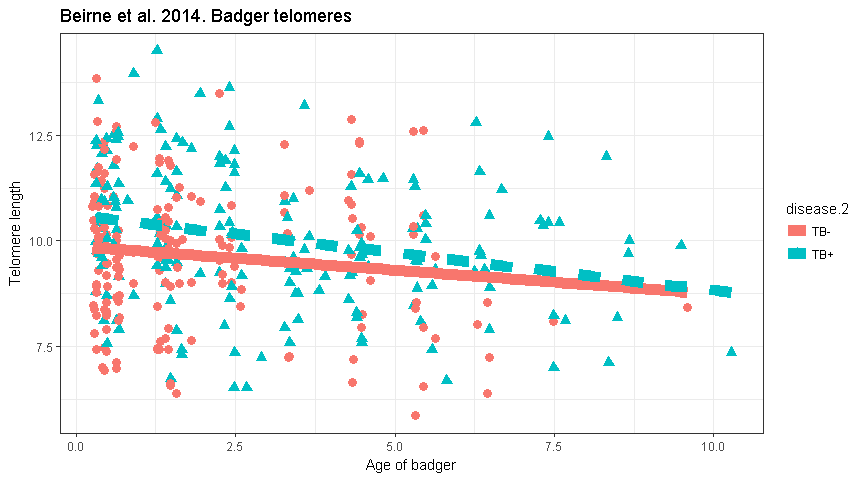
## Understanding Regression Equations

## Today's Goals

* Interpret slopes & intercepts reported by R's regression output
* Understand how to interpret regression equations
* Compare regression equations for different forms of models
  + Null model
  + y ~ continous predictor
  + y ~ categorical predictor
  + y ~ continous + categorical
  + y ~ continous\*categorical
* Understanding "interactions"
* Introduction to models w/ 2 continuous x variable in them

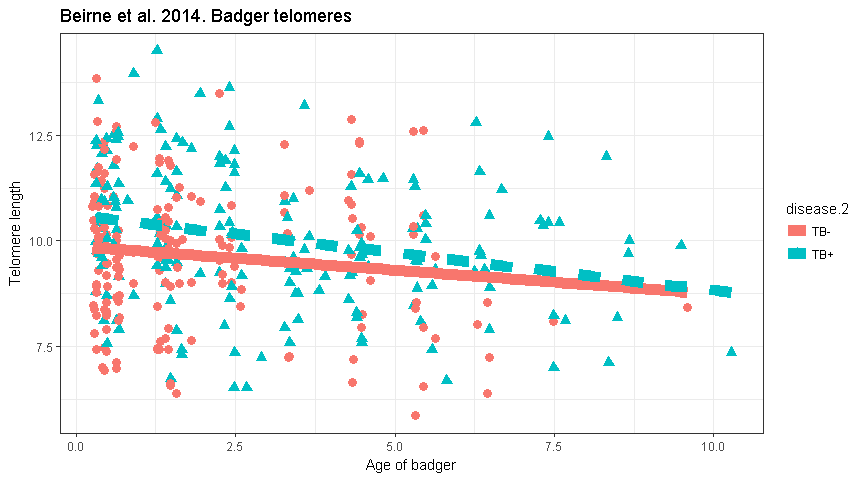
### Data for today:

* Beirne et al. 2014. Age-Related Declines & Disease-Associated Variation in Immune Cell Telomere Length in a Wild Mammal. PLoS ONE.



* Telomeres typically shorten as organisms age: but not always!
* Stressful conditions degrade telomere
  + disease?, migration?, stats classes?

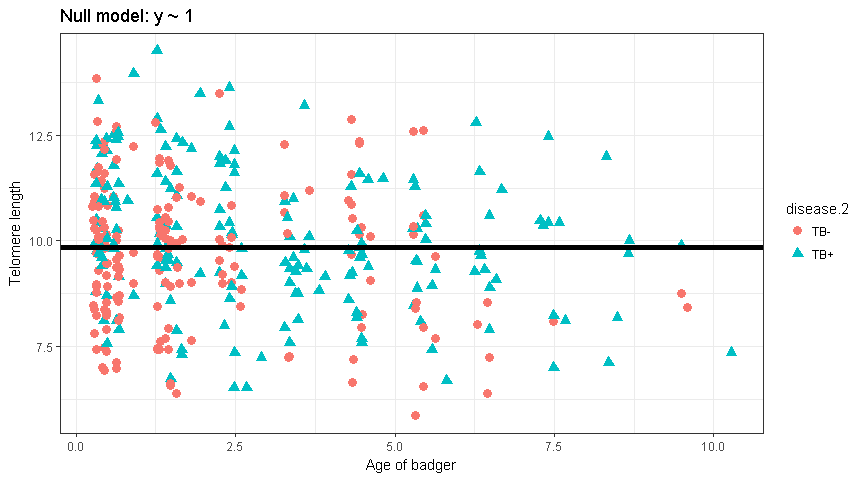
Data for today:



* Response variable (y): **relative telomere length** (via qPCR)
* Continuous Predictor variable (x): **badger age**
  + (self-reported by badger)
* Categorical predictor variable (x): badger **TB status**
  + **TB-**: uninfected
  + **TB+**: positive (symptomatic & asymp.)

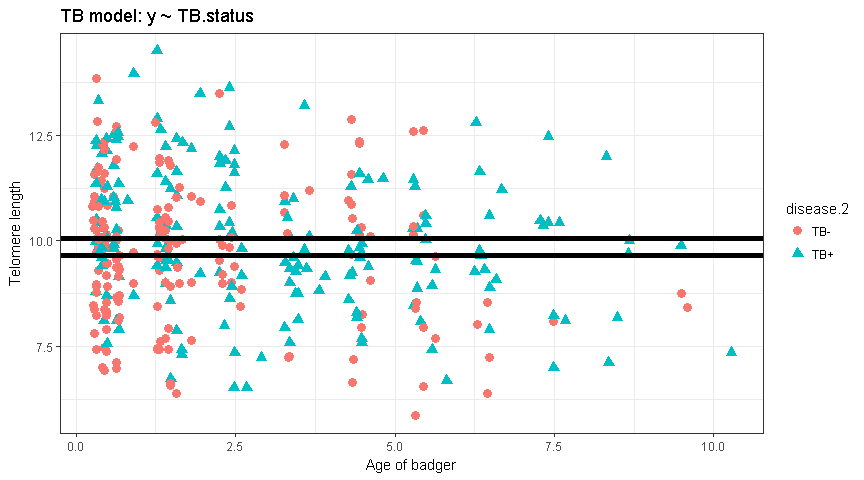
Possible relationships between telomeres, age & TB

### Model 1: Telomere length does not change w/age or differ by disease status



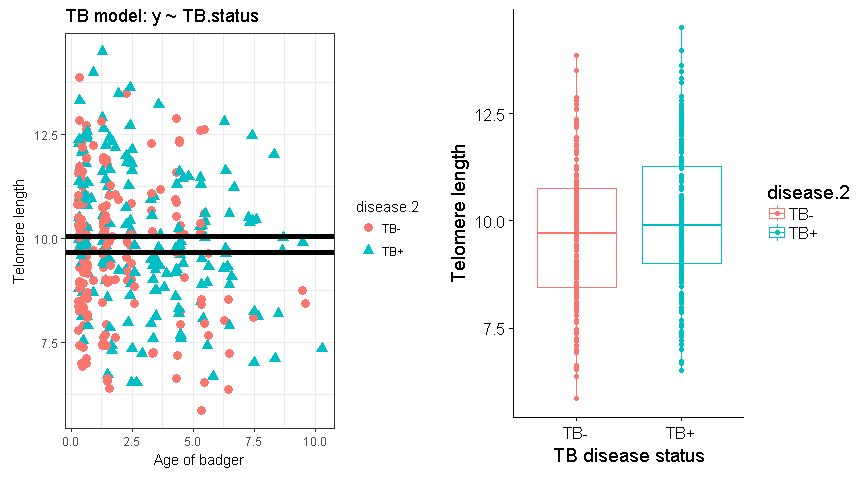
* "Null" model = Flat line
* R code: y ~ 1
* Best **predictor** of telomere length of new badger: mean of current sample

### Model 2: Telomere length depends on disease status (TB- vs TB+)



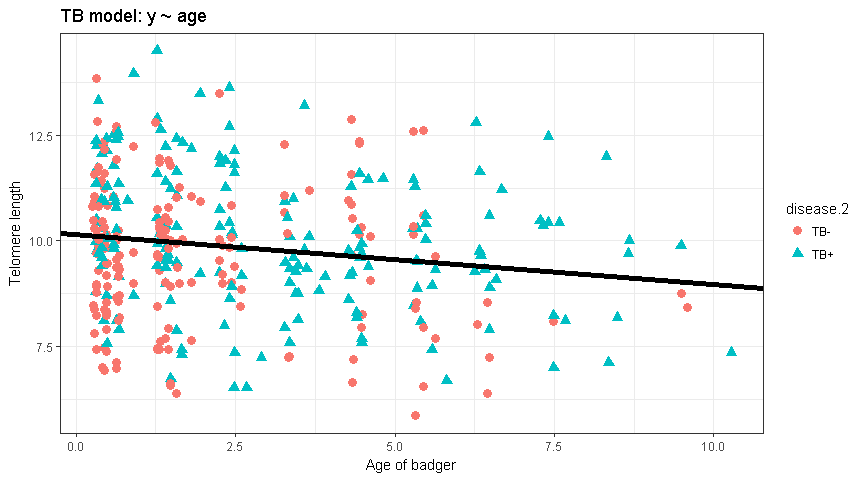
* "Null" model = Flat line
* Model w/just a categorical variable = **2 flat lines**
* Hypothesis: there is an average negative effect of acquiring TB
  + Perhaps initial immune response causes initial stress & telomere damage
  + … w/no subsequent damage after initial infection
* Best predictor of telomere length of new badger: TB status

### Normally we'd just look at the data like this



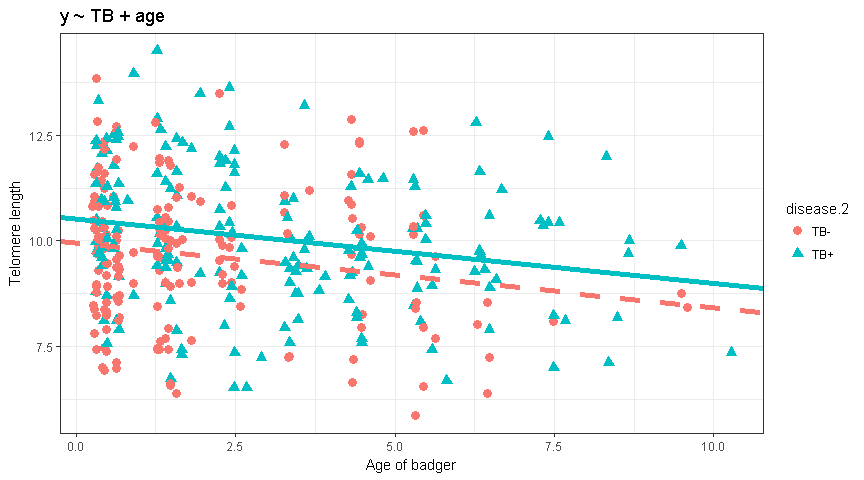
Note that on average, TB+ animals have LONGER telomeres!

### Model 3: Telomere length depends on age



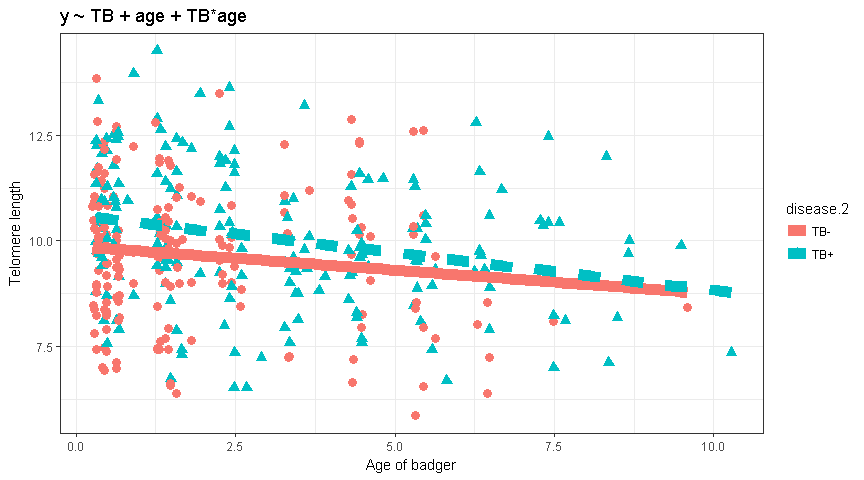
* Typical regression situation
* Hypothesis: telomeres degrade w/age
* TB status does not impact degradation rate
  + No average reduction when infection occurs
  + TB infection does not increase rate of infection

### Model 4: Telomeres degrade w/ age AND there's a discrete effect of being TB-



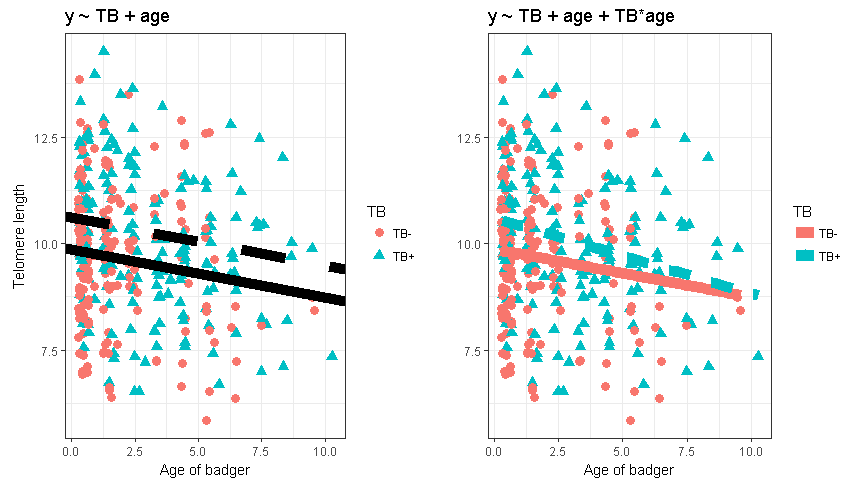
* Hypothesis: there is an average negative effect of acquiring TB
  + Perhaps initial immune response causes initial telomere damage, w/no subsequent effect; telomeres also degrade with age
* Best predictor of telomere length of new badger: TB status AND age
* One way of looking at model: **"after controling for age, there is a difference in telomere length between TB- and TB+ animals"**

### Model 5: "Interaction" between age and TB status



* "Multiple" regression situation
* Hypothesis: telomeres degrade w/age
* AND **rate** of degradation increases when TB+

How do models 4 and 5 differ?



\* How do they represent different STAT hypotheses?

\* How do they represent different BIOLOGICAL hypotheses?

### Summary of models to fit to data

I am going to fit a series of models of the form: "telomere length ~ ..."

|  |  |  |  |
| --- | --- | --- | --- |
| model.names | formula | variable.types | lines |
| null | ~ 1 |  | 1 |
| TB | ~ TB | categorical (TB) | 1 |
| age | ~ age | continuous (age) | 1 |
| age+TB | ~ age + TB | cat+cont | 2 |
| age.x.TB | ~ age + TB + age\*TB | cat\*cont | 2 |

### Null model

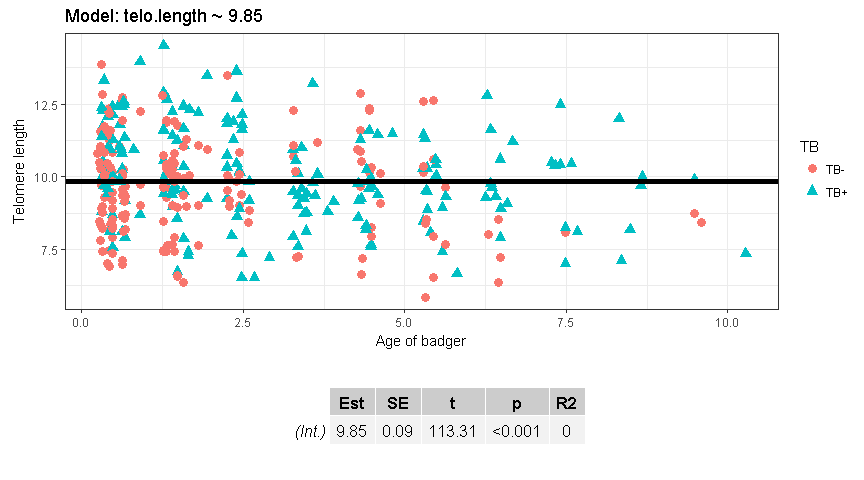
#Fit null  
m.null <- lm(telo.length ~ 1,   
 data = bad)  
  
#Output of null  
summary(m.null)

##   
## Call:  
## lm(formula = telo.length ~ 1, data = bad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9941 -1.1157 -0.0207 1.0922 4.6435   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
**## (Intercept) 9.85295 0.08696 113.3 <2e-16 \*\*\***## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.652 on 360 degrees of freedom

* Formula "~1" means "model w/ just an intercept term"
* This model will be defined just by the overall mean of the the telomere lengths (y variable)

### Plot null

#### Raw data + null model



Model: telo.length ~ mean(telo.length)

~ 9.85

### TB model

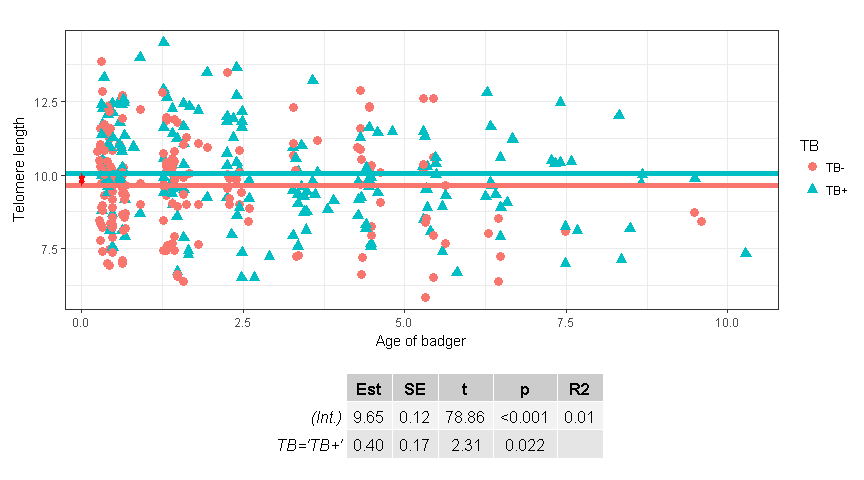
* Model telomere length vs **disease status** (TB- vs TB+)

m.TB <- lm(telo.length ~ TB,   
 data = bad)  
  
summary(m.TB)

##   
## Call:  
## lm(formula = telo.length ~ TB, data = bad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7943 -1.1183 -0.0191 1.1746 4.4448   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
**## (Intercept) 9.6531 0.1224 78.857 <2e-16 \*\*\* ##  
## TBTB+ 0.3985 0.1729 2.305 0.0217 \* ##**

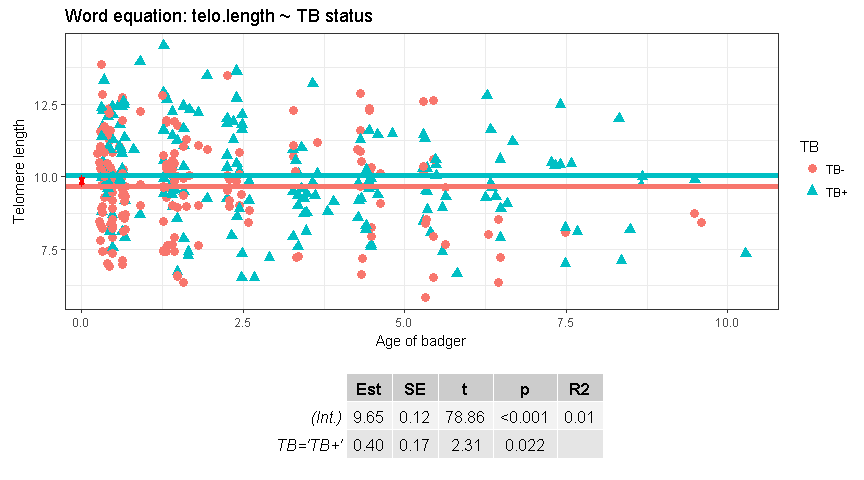
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.642 on 359 degrees of freedom  
## Multiple R-squared: 0.01459, Adjusted R-squared: 0.01184   
## F-statistic: 5.314 on 1 and 359 DF, p-value: 0.02172

#### Plot TB model (a)

What equation(s) define these lines? 

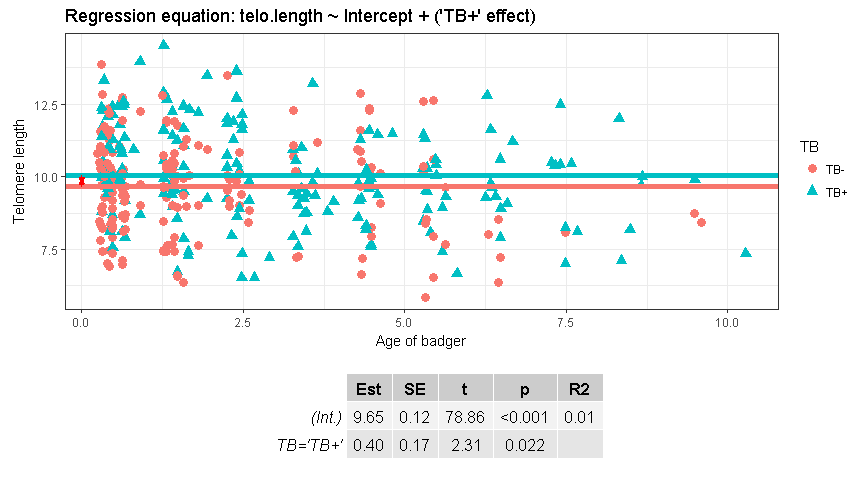
* What equations define these lines?
* What is the meaning of the 0.40 estimate (“**TBTB+” row in R output)**

#### Plot TB model (b)



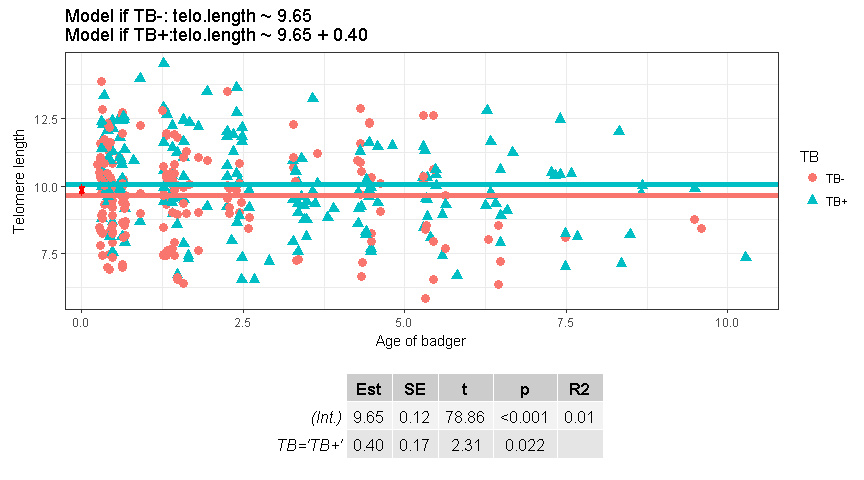
* Models often expressed as word equations
* Similar to R code
* Intercept is implicit
* Could write "telo.length ~ Intercept + TB effect"
* or "telo.length ~ 1 + TB effect"

Plot TB model (c)



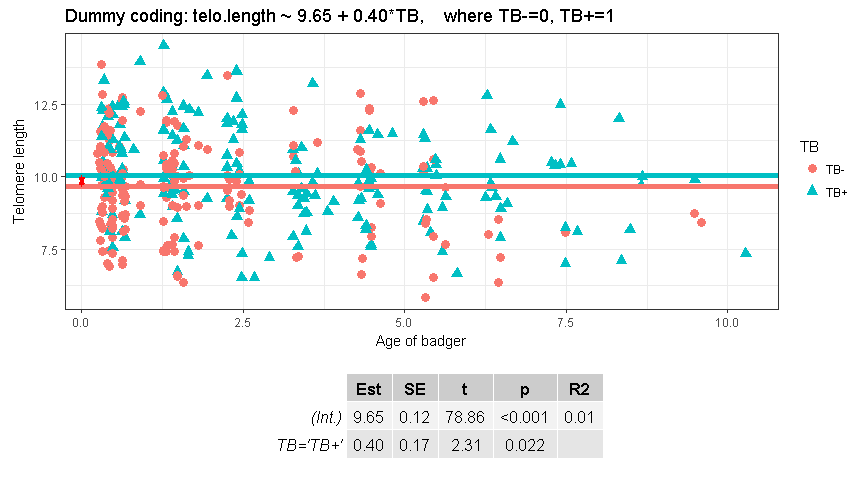
* Regression models set one of the factor levels as the "baseline"
  + In R it done alphabetically; can be reset with levels()
* This acts as the intercept
* The effect of being 'TB+' is defined as:
  + intercept + TB effect
* There are therefore essentially 2 equations:
  + Equation for TB-, defined just by intercept
  + Equation for TB+, defined by intercept and TB+effect

Plot TB model as 2 equations



Separate equation for each line

#### Plot TB model with dummy coding



* The underlying math uses "dummy variables" / "dummy coding"
* We use words as our factor / categorical variables; R converts them to 0s and 1s
* For TB-: telo.length ~ 9.65 + 0.40\*0
* ~ 9.65
* For TB+: telo.length ~ 9.65 + 0.40\*1
* ~ 10.05

### Test dummy coding

Set up dummy coding using ifelse() function

#Convert to 0/1 with ifelse()  
bad$TB.dummy <- ifelse(bad$TB == "TB-",0,1)  
  
  
#Look at output as table  
table(bad$TB.dummy,  
 bad$TB)

##   
## TB- TB+  
## 0 180 0  
## 1 0 181

### Test dummy coding

Fit the models

#Original model  
m.TB <- lm(telo.length ~ TB,   
 data = bad)  
  
  
#Dumming coding  
m.TB.dummy <- lm(telo.length ~ TB.dummy,   
 data = bad)

Check the output

|  |  |  |
| --- | --- | --- |
|  | (Intercept) | TBTB+ |
| **original** | 9.653 | 0.3985 |
| **dummy** | 9.653 | 0.3985 |

### Alternative parameterization: means

* R's default is "effects" parameterization
* Statitical focus is whether the "effect" is equal to 0.0
* We can have R fit seperate intercepts for each group
* Add "-1" to lm() formula

## Original Model  
### Note I've add the +1 for intercept  
### This is normally implicit and adding by R automatically  
**m.TB.orig <-** lm**(telo.length ~ 1+ TB,   
 data = bad)**  
  
## Means paramterization  
### add "-1"  
**m.TB.means <-** lm**(telo.length ~ -1 + TB,   
 data = bad)**

### Compare "Effects" and "means" parameterization

|  |  |  |
| --- | --- | --- |
|  | (Intercept) | TBTB+ |
| **original** | 9.653 | 0.399 |
|  | TBTB- | TBTB+ |
| **dummy** | 9.653 | 10.052 |

(Intercept) + ‘TB+’ effect

= 9.653 + 0.399

= 10.052 = TB+ intercept

# Interpretting the coefficients from a “means parameterization”

summary(m.TB.means)

##   
## Call:  
## lm(formula = telo.length ~ -1 + TB, data = bad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7943 -1.1183 -0.0191 1.1746 4.4448   
##   
## Coefficients:  
**## Estimate Std. Error t value Pr(>|t|)   
## TBTB- 9.6531 0.1224 78.86 <2e-16 \*\*\* ###  
## TBTB+ 10.0517 0.1221 82.34 <2e-16 \*\*\* ###**  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.642 on 359 degrees of freedom  
## Multiple R-squared: 0.9731, Adjusted R-squared: 0.973   
## F-statistic: 6499 on 2 and 359 DF, p-value: < 2.2e-16

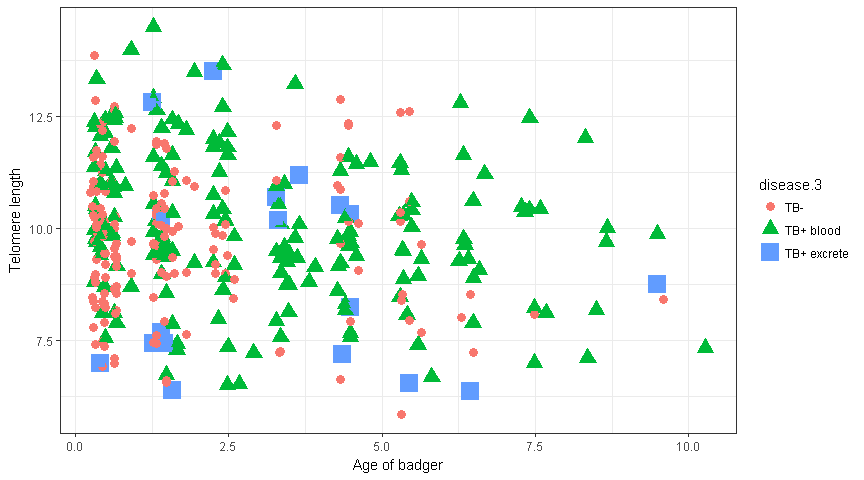
* Why are both p values so low?
* What hypothesis is being tested?
  + NB: p for TB+ "effect" was only p = 0.0217

### What if there are more than 2 levels to factor variables

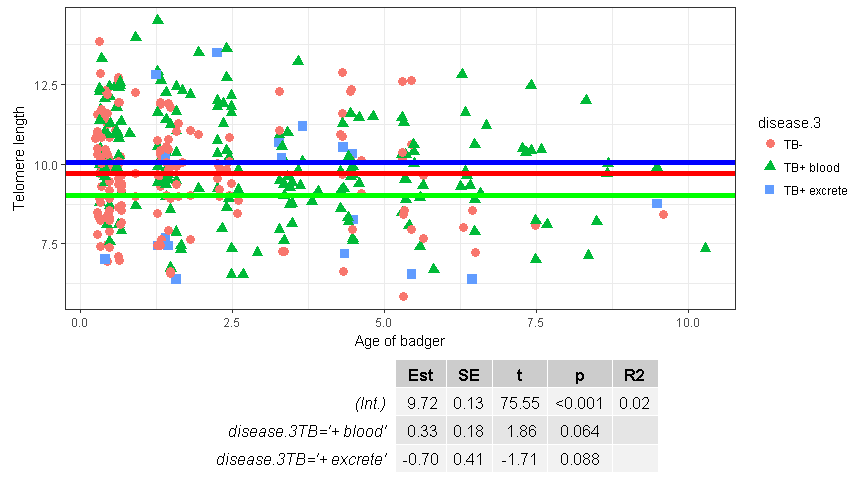
## **TB- TB+ blood TB+ excrete**   
## 162 181 18

* Original study had 3 "levels" to their categorical / factor vaiable
* TB-: No evidence of disease
* TB+ blood: carrier of TB
* TB+ excrete: shedding virus through feces, mucous

### Plot all 3 categories



### Plot all 3 categories and 3 lines



### Age model

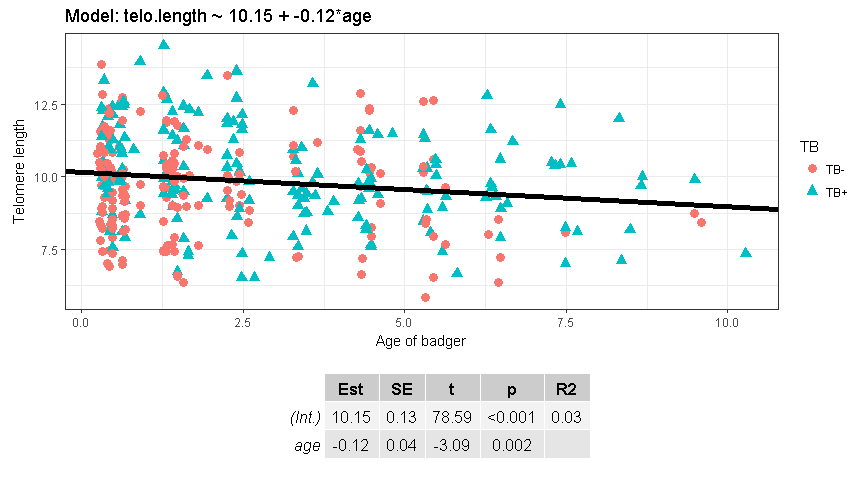
* telomere length deceases with age

### Fit age model

m.age <- lm(telo.length ~ age,   
 data = bad)  
  
summary(m.age)

##   
## Call:  
## lm(formula = telo.length ~ age, data = bad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.663 -1.046 -0.032 1.104 4.496   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
**## (Intercept) 10.1506 0.1292 78.587 < 2e-16 \*\*\* ##  
## age -0.1182 0.0383 -3.087 0.00218 \*\* ##**  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.633 on 359 degrees of freedom  
## Multiple R-squared: 0.02586, Adjusted R-squared: 0.02314   
## F-statistic: 9.53 on 1 and 359 DF, p-value: 0.002179

### Plot age model



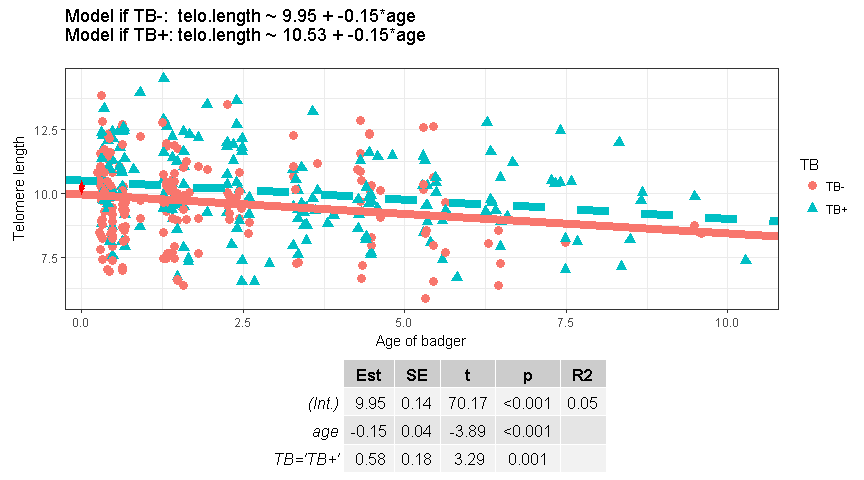
## "ANCOVA": age + TB model

m.age.plus.TB <- lm(telo.length ~ age + TB,   
 data = bad)  
  
summary(m.age.plus.TB)

##   
## Call:  
## lm(formula = telo.length ~ age + TB, data = bad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6299 -1.0551 -0.0565 1.0675 4.1649   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
**## (Intercept) 9.94605 0.14175 70.165 < 2e-16 \*\*\*  
## age -0.15233 0.03918 -3.888 0.00012 \*\*\* #age slope  
## TBTB+ 0.57918 0.17583 3.294 0.00109 \*\* #’TB+’ effect**## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.611 on 358 degrees of freedom  
## Multiple R-squared: 0.05452, Adjusted R-squared: 0.04923   
## F-statistic: 10.32 on 2 and 358 DF, p-value: 4.387e-05

* We can think of this model 2 ways
  + 1. Does telomere length depend on age AND TB
    2. Is there an effect of TB on telomeres AFTER "controlling" for age

### Plot "ANCOVA"



‘TB-‘ Intercept = 9.95

‘TB+’ Intercept = 9.95 + 0.58

### "ANCOVA" w/ an interaction: age\*TB

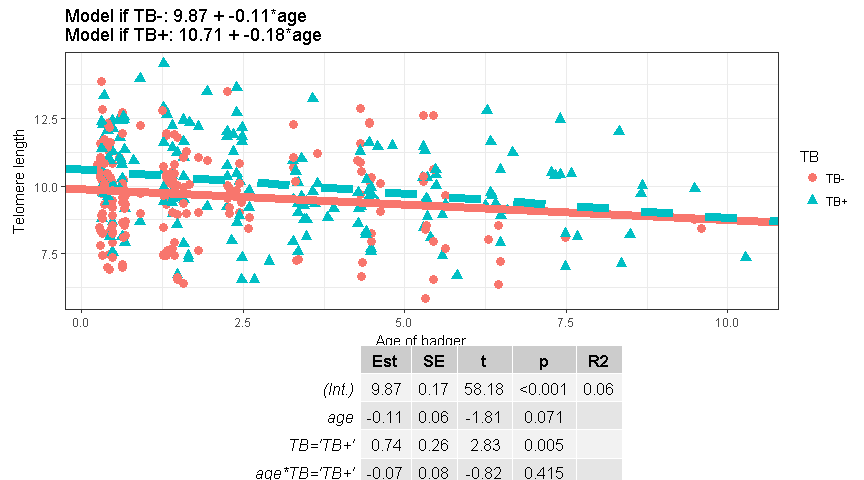
* Model the "interaction" between and TB
* full model: "**age + TB + age\*TB**"
* can just use "**age\*TB**" and R expands out the model

#Full formula  
m.age.x.TB <- lm(telo.length ~ **age + TB + age\*TB**,   
 data = bad)  
  
#Equivalent  
## R expands out the terms   
## of the model  
m.age.x.TB <- lm(telo.length ~ **age\*TB**,   
 data = bad)

summary(m.age.x.TB)

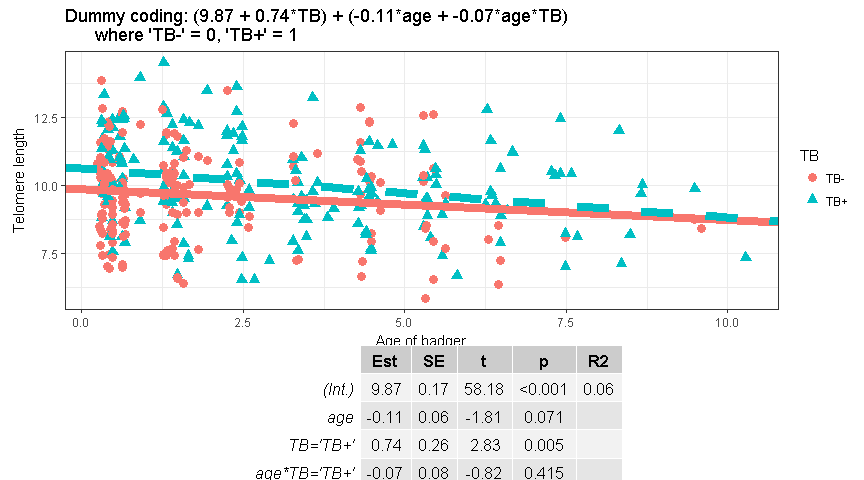
##   
## Call:  
## lm(formula = telo.length ~ age \* TB, data = bad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6462 -1.0919 -0.0537 1.0999 4.1173   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
**## (Intercept) 9.87000 0.16966 58.176 < 2e-16 \*\*\*  
## age -0.11279 0.06230 -1.810 0.07109 . #slope  
## TBTB+ 0.73576 0.26021 2.828 0.00495 \*\* #’TB+’ effect on int.  
## age:TBTB+ -0.06546 0.08015 -0.817 0.41468 #’TB+’ effect on slope.**## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.612 on 357 degrees of freedom  
## Multiple R-squared: 0.05628, Adjusted R-squared: 0.04835   
## F-statistic: 7.096 on 3 and 357 DF, p-value: 0.0001211

### Plot ANCOVA w/interaction

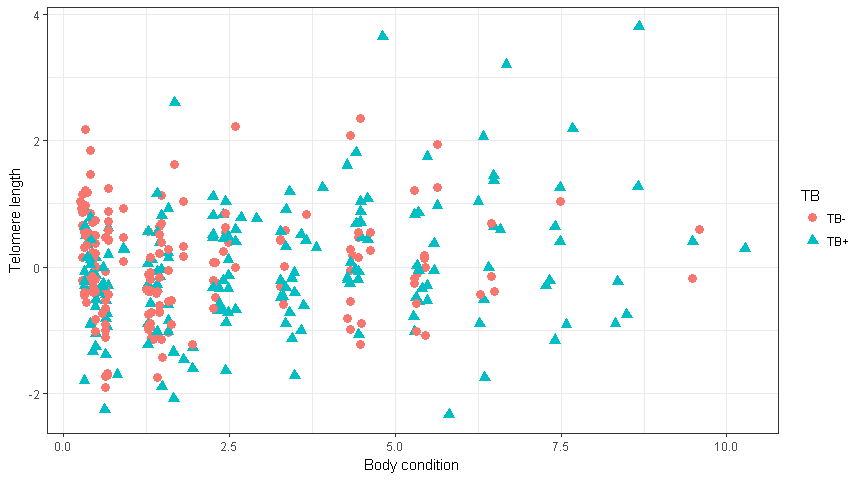


* Slope is **steeper** (more negative) if TB+
  + (-0.11 + -0.07)\*age
  + (-0.18)\*age
* **What would it be with "dummy" coding? …**

### ANCOVA equations w/ "dummy" coding



### What about 2 continuous variables?



summary(bad$condition)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.32400 -0.55150 -0.07301 0.00000 0.55140 3.80000

* New x variable Body condition = a measure of overall health

### Model w/ 2 continous covariates

# Fit the model  
m.age.condition <- lm(telo.length ~ 1 + age + condition,   
 data = bad)  
  
  
summary(m.age.condition)

##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.18030 0.12999 78.314 < 2e-16 \*\*\*  
## age -0.13003 0.03882 -3.350 0.000895 \*\*\*  
## condition 0.16529 0.09689 1.706 0.088892 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.629 on 358 degrees of freedom  
## Multiple R-squared: 0.03371, Adjusted R-squared: 0.02831   
## F-statistic: 6.245 on 2 and 358 DF, p-value: 0.002158

\*\* Model has\*\*

* an intercept
* a slope for "age"
* a slope for "condition"
* These slopes are often called "partial regression coefficients"
* Have to think carefully how to plot them…