Simple Linear Regression

BIOST 515/518

Discussion - Week 1

DSST Study

Features of this study

- Cohort (observational) study
- Participants over 65 years old
- Generally healthy
- Lasted 11 years
- Looking for risk factors for cardiovascular and cerebrovascular disease

Scientific Questions

- What are correlates of decreased cognitive function?
- What associations exist between measurements of cognitive function and the available data on
  - participant demographics?
  - behavior?
  - and various clinical and laboratory measures of organ system functioning?

DSST data

- Response (dependent) variables
  - DSST
  - MMMSE
- Predictor (independent) variables
  - Demographics: **age**, **sex**, gender, height, weight
  - Behavior: smoking, alcohol consumption
  - Biological measures: blood pressure, kidney function, etc.

Descriptive Summary: Code

```
data <- read.csv("../dsst.txt", sep="") #read in data

summary(data$dsst) #summarize dsst
summary(data$mmmse)
summary(data$age)
table(data$male) #number of males (0=female, 1=male)

#scatter plot
plot(data$age,data$dsst,xlab="age",ylab="DSST")
plot(data$age,data$mmmse,xlab="age",ylab="MMMSE")

#boxplots
boxplot(data$dsst~data$male,
        names=c("female","male"),ylab="DSST")
boxplot(data$mmmse~data$male,
        names=c("female","male"),ylab="MMMSE")
```

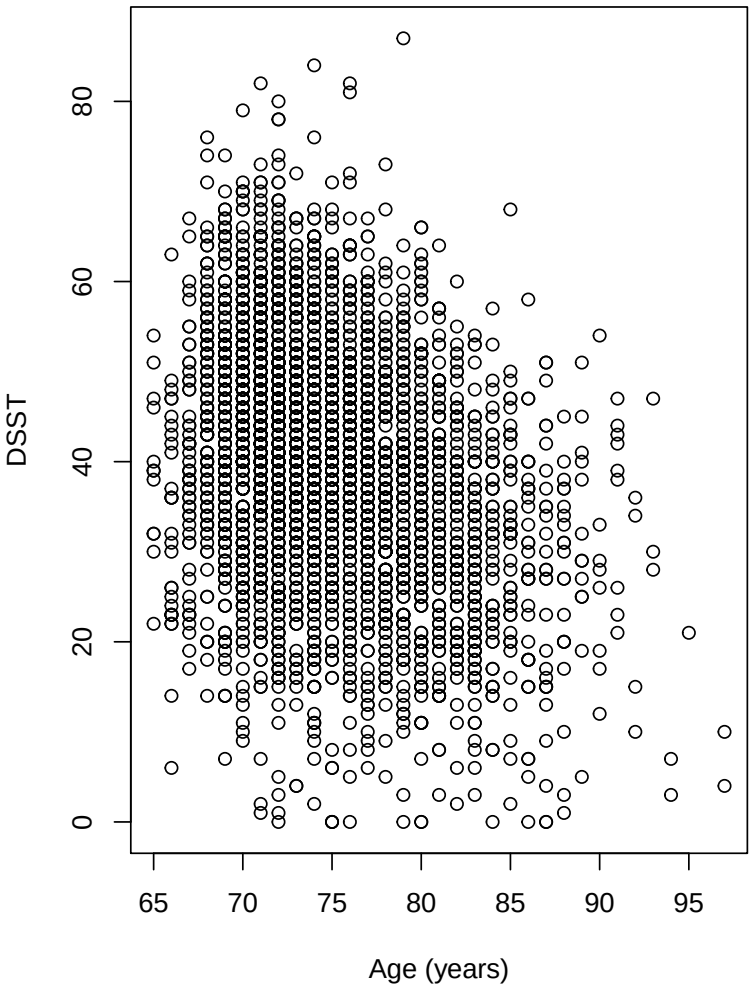
Descriptive Summary: Results

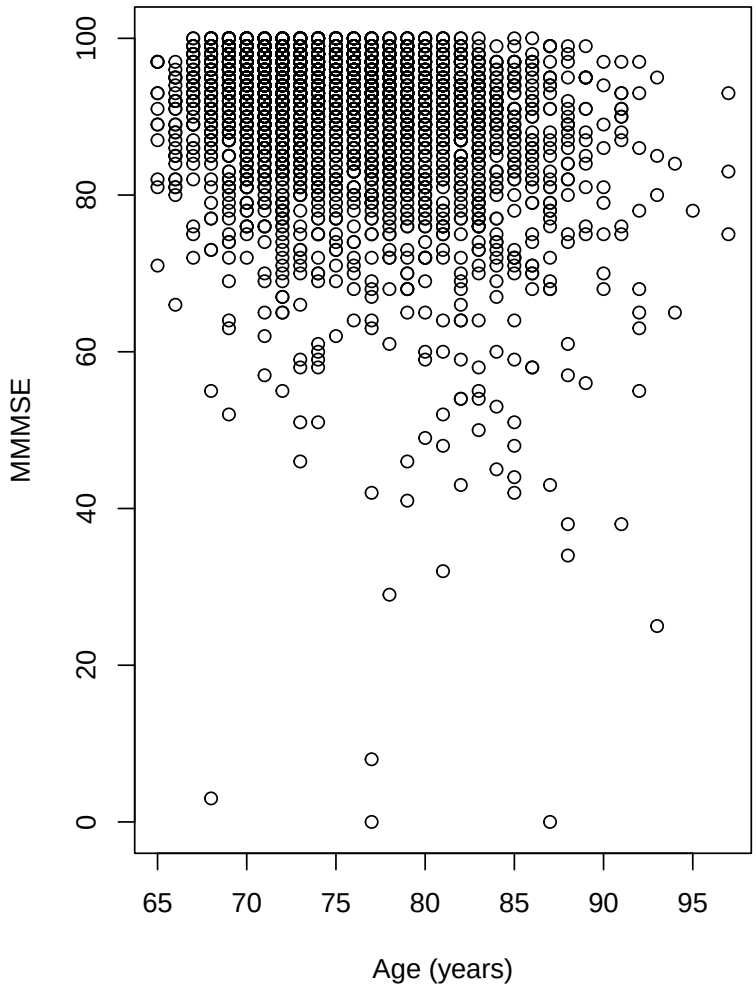
Sample size: $n = 3660$

Variable	Mean (SD)	Range	Missing
DSST	39.4 (13.6)	0-87	118
MMMSE	90.8 (9.1)	0-100	11
Age (years)	75.1 (5.2)	65-97	0
Male (yes = 1, no = 0)[1]	1527 (42%)	0, 1	0

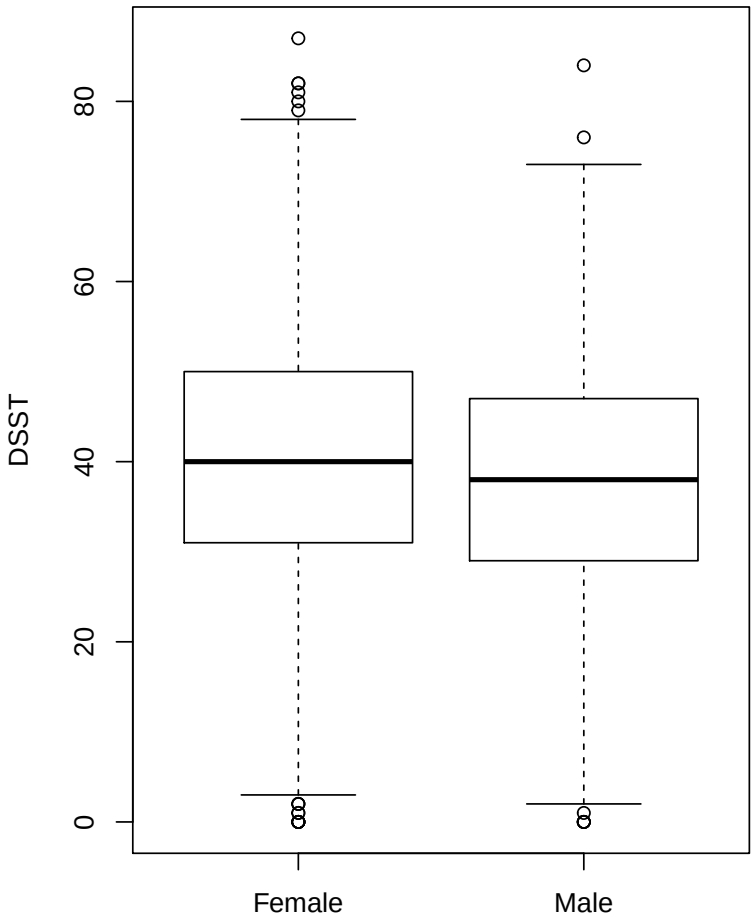
[1]: *Count (Percent) reported.*

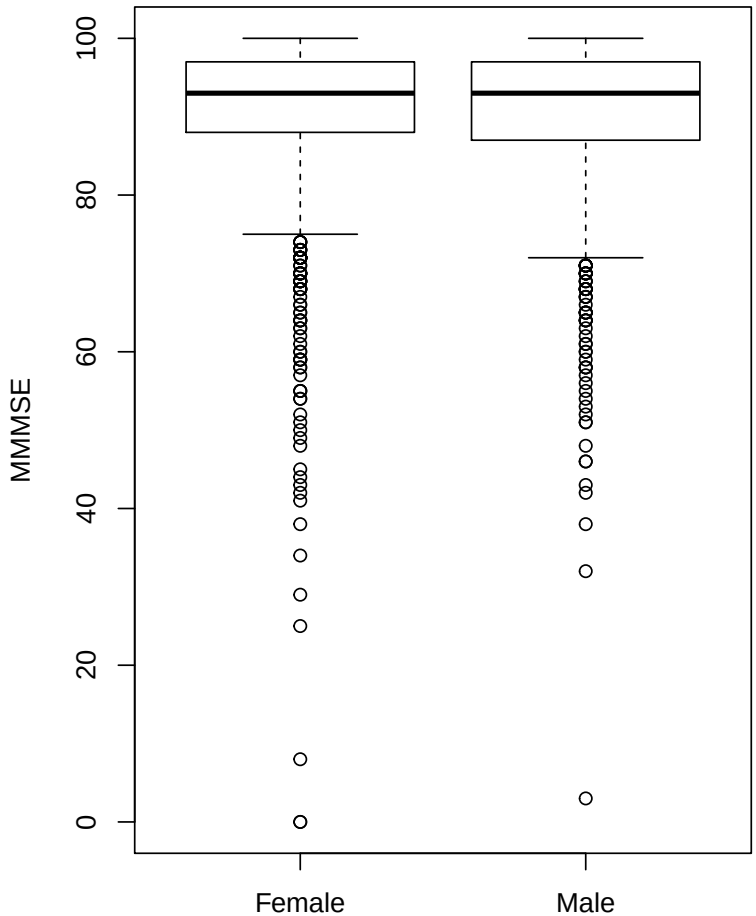
Descriptive Summary: Results





Descriptive Summary: Results





Regression Model 1: Code

```
mod1 <- lm(dsst~age, data=data)
summary(mod1)
```

```
##
## Call:
## lm(formula = dsst ~ age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.829  -8.513   0.171   8.676  50.992
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.70753    3.19394   31.84   <2e-16 ***
## age          -0.83164    0.04251  -19.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.92 on 3540 degrees of freedom
##   (118 observations deleted due to missingness)
## Multiple R-squared:  0.09757,    Adjusted R-squared:  0.09732
## F-statistic: 382.8 on 1 and 3540 DF,  p-value: < 2.2e-16
```

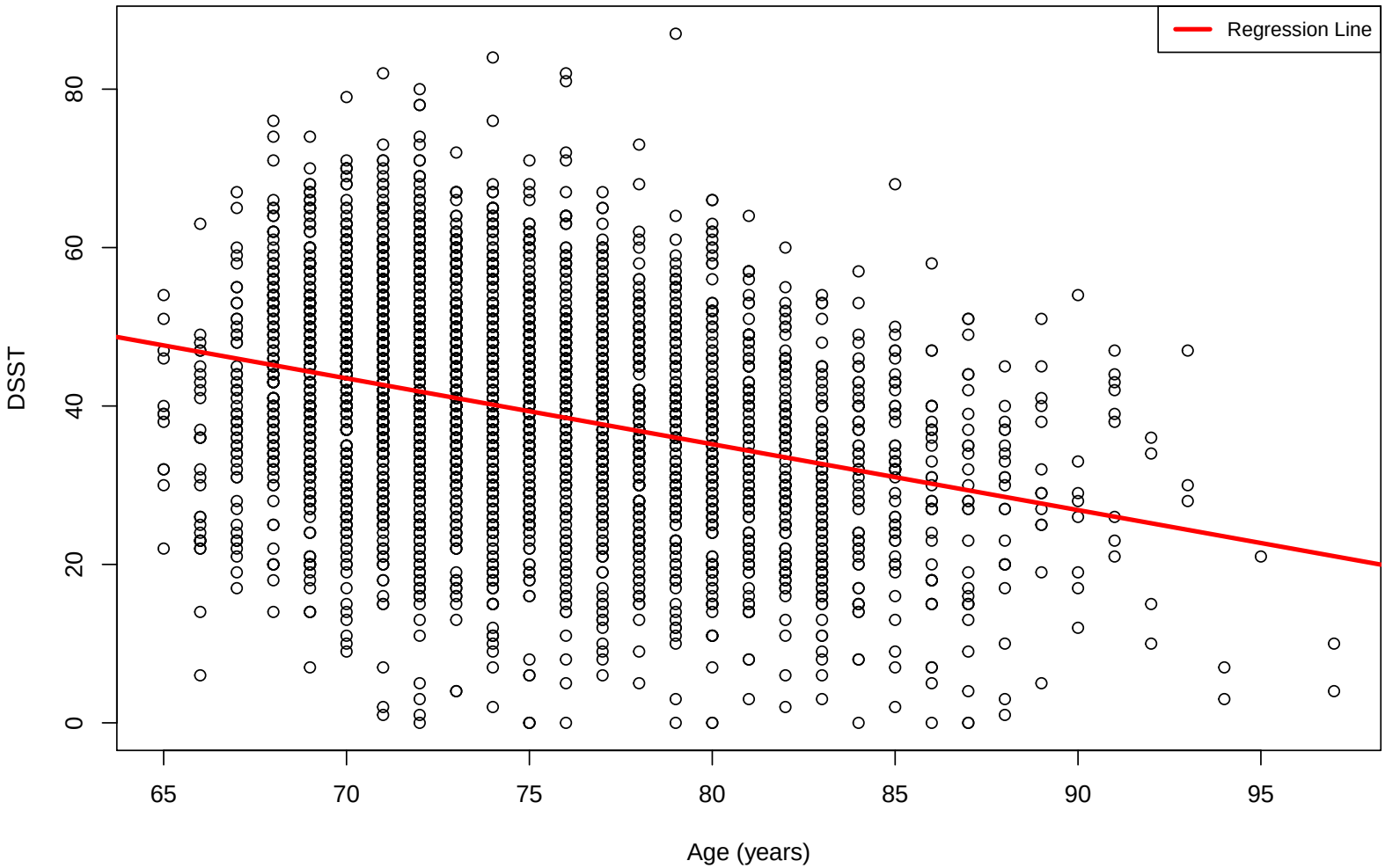
Regression Model 1: Code

```
confint.default(mod1)
```

```
##                  2.5 %      97.5 %
## (Intercept) 95.4475149 107.9675480
## age         -0.9149551  -0.7483263
```

```
#plots regression line on top of scatter plot
plot(data$age, data$dsst, xlab="Age", ylab="DSST")
abline(mod1, lwd=3, col="red")
legend("topright", c("Regression Line"),
       lwd=3, col="red", cex=0.8)
```

Regression Model 1: Results



Regression Model 1: Results

We estimate that for each 1 year difference in age between two populations the mean DSST is 0.83 lower in the older group. The 95% **confidence interval** suggests that observed mean DSST scores between -0.75 and -0.91 lower per year are not unusual. We found the relationship between age and DSST score to be significant, with a **p-value** less than 0.001.

Regression Model 2: Code (Regression)

```
mod2 <- lm(dsst~male,data=data)
summary(mod2)
```

```
##
## Call:
## lm(formula = dsst ~ male, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.555  -9.555   0.300   9.445  46.445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.5554     0.2976 136.262  < 2e-16 ***
## male         -2.8554     0.4609  -6.195 6.49e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.52 on 3540 degrees of freedom
##   (118 observations deleted due to missingness)
## Multiple R-squared:  0.01073,    Adjusted R-squared:  0.01045
## F-statistic: 38.38 on 1 and 3540 DF,  p-value: 6.489e-10
```

Regression Model 2: Results

We estimate that mean DSST score is 2.9 points lower in males than females. The 95% **confidence interval** suggests that observed differences between 2.0 and 3.8 lower in males are not unusual. We found the relationship between age and DSST to be significant, with a **p-value** less than 0.001.

Regression vs T-test

Since gender is a binary variable, we could also perform a t-test comparing mean DSST between males and females.

Regression vs T-test

```
t.test(data$dsst[data$male==1],data$dsst[data$male==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$dsst[data$male == 1] and data$dsst[data$male == 0]
## t = -6.2226, df = 3228.8, p-value = 5.519e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.755090 -1.955671
## sample estimates:
## mean of x mean of y
##  37.70007  40.55545
```

Regression vs T-test

We see that the estimated difference in means in the same, and reach the same conclusion as in simple linear regression.