Exercise 4 - Correlation and Regression

Problem 1:

Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). Is there a statistically significant relationship between the scores?

Problem 2:

Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). What is the least squares regression equation with Sony 1 as the predictor variable? Draw the fitted line plot. Is there an indication of "regression to the mean"? Why?

Problem 3:

Consider the men's long jump in the Olympics (**longjump**). How strong is the relationship between Year and LongJump?

Problem 4:

Consider the following data set:

x	y
10	58
11	54
12	51
13	52
14	62
15	57
16	63
17	64
18	69
19	71
20	70

Find the least squares regression equation and use it to predict the y value for an observation with x=15

Solutions

Problem 1:

Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). Is there a statistically significant relationship between the scores?

Parameter: correlation coeffcient

Problem Test for independence

Method: pearson.test

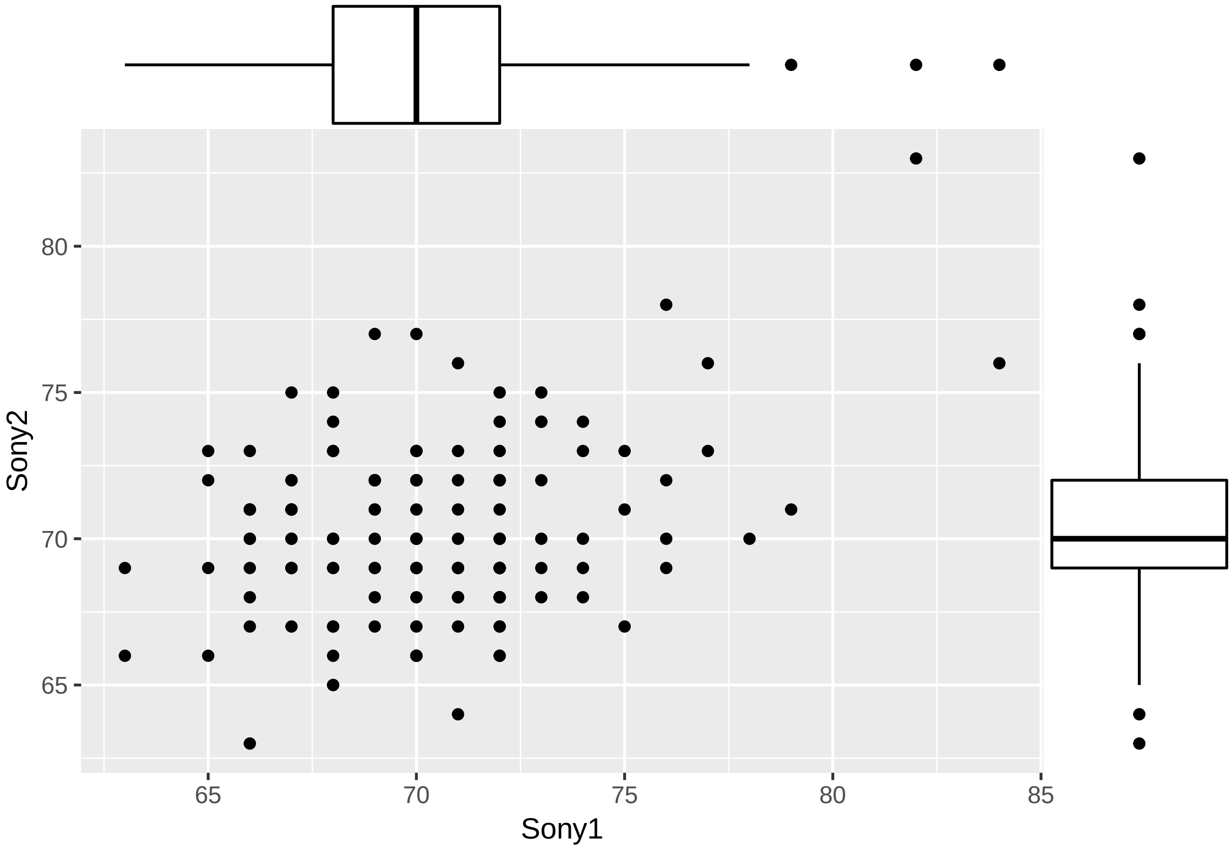
```
attach(golfscores)
```

1) Parameter: Pearson's correlation coefficient ρ
2) Method: Test for Pearson's correlation coefficient ρ
3) Assumptions: relationship is linear and that there are no outliers.
4) $\alpha = 0.05$

5) H$_0$: $\rho = 0$ (no relationship between Day of Year and Draft Number)
6) H$_a$: $\rho \neq 0$ (some relationship between Day of Year and Draft Number)
7) p = 0.000

```
pearson.cor(Sony1, Sony2, rho.null=0)
```

```
## Warning: Removed 37 rows containing missing values (geom_point).
```



```
## p value of test H0: rho=0 vs. Ha: rho <> 0:  0.000
```

8) $p < \alpha = 0.05$, so we reject the null hypothesis,
9) There is a statistically significant relationship between Day of Year and Draft Number.

Assumptions: boxplots and scatterplot show no outliers. No non-linear relationship.

Problem 2:

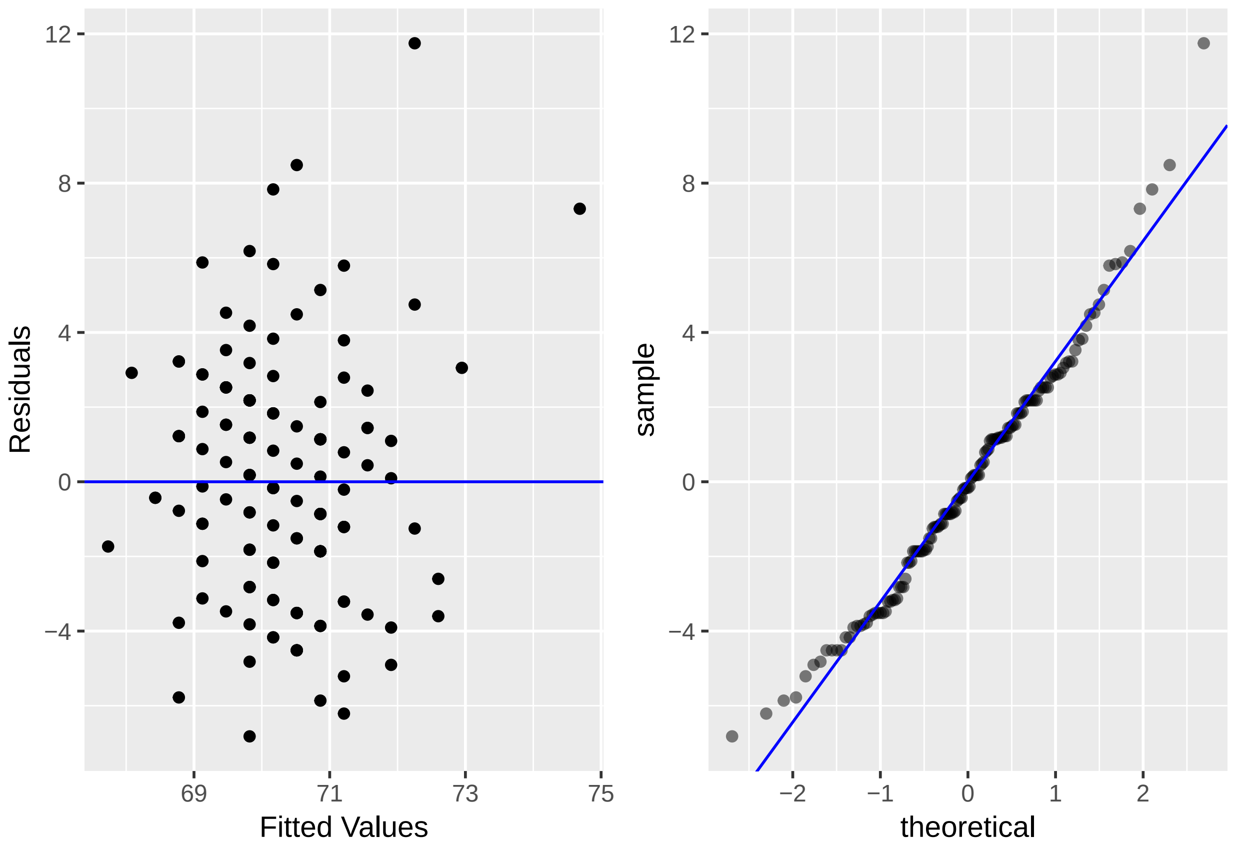
Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). What is the least squares regression equation with Sony 1 as the predictor variable? Draw the fitted line plot. Is there an indication of "regression to the mean"? Why?

Parameter: regression coefficients

Problem: find model

Method: slr

```r
slr(Sony2, Sony1)
```

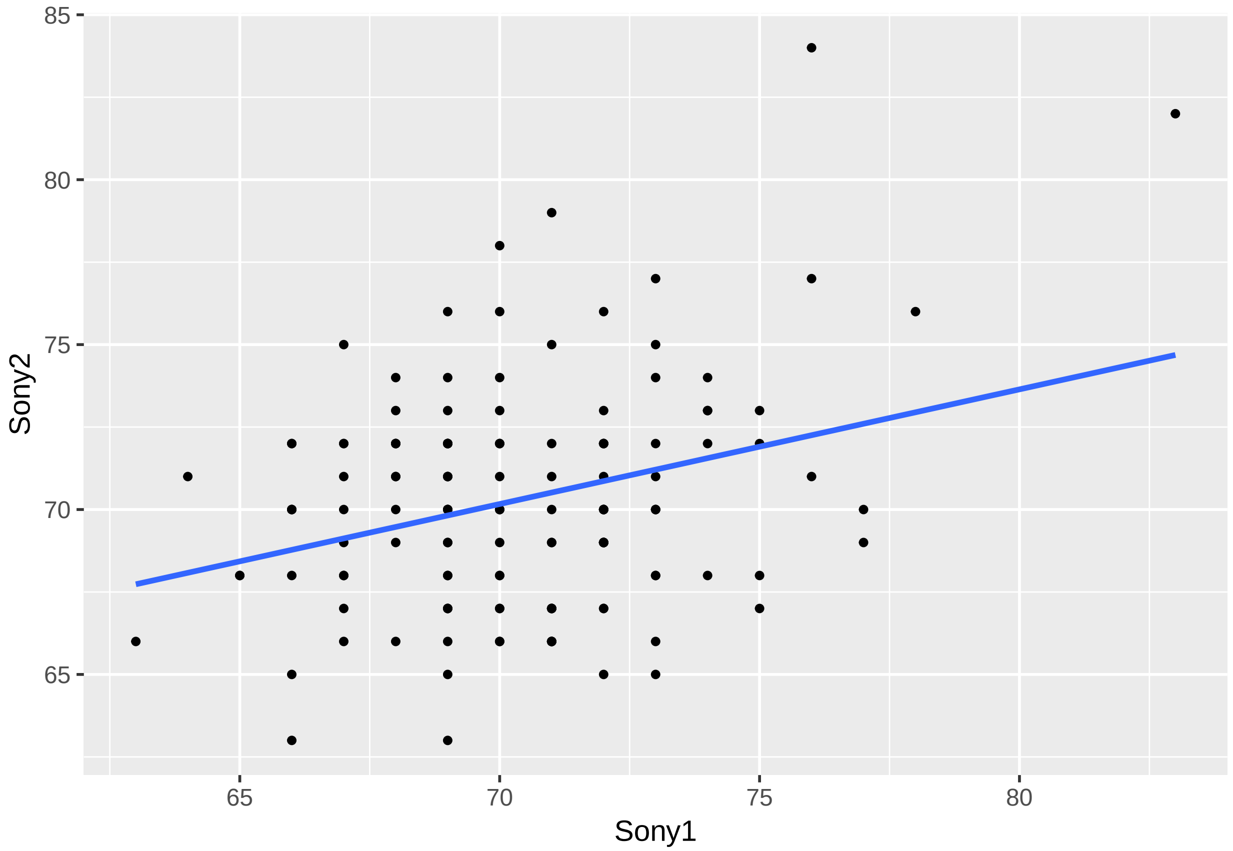


```
## The least squares regression equation is:
##  Sony2  = 45.836 + 0.348 Sony1
## R^2 = 9.27%
```

```
splot(y=Sony2, x=Sony1, add.line=1)
```

```
## Warning: Removed 37 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 37 rows containing missing values (geom_point).
```



the slope of the line (0.348) is between 0 and 1, so yes, there is an indication of regression to the mean.

Problem 3:

Consider the men's long jump in the Olympics (**longjump**). How strong is the relationship between Year and LongJump?

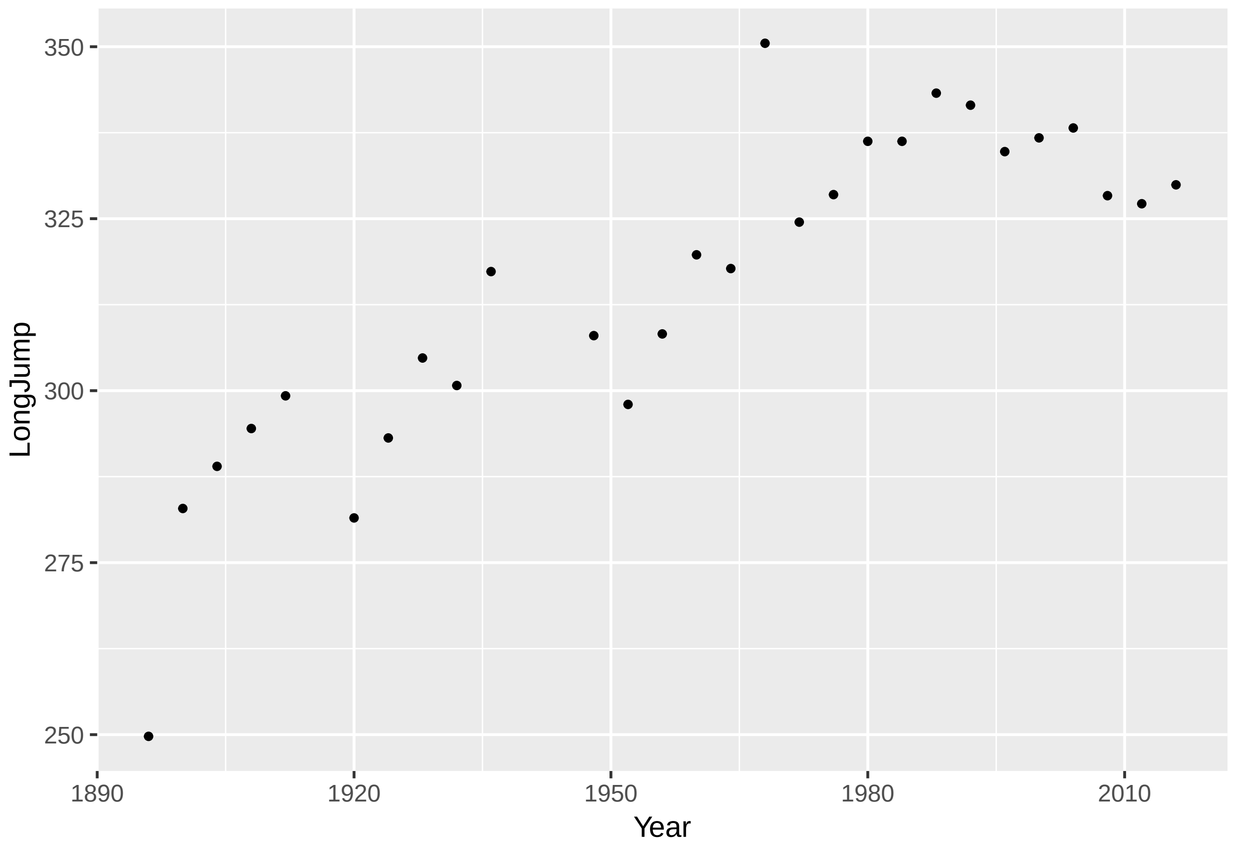
Parameter: correlation coeffcient

Problem: find correlation

Method: ????

the scatterplot of LongJump by Year shows a non-linear relationship, so we can't answer this question (want to know? come to ESMA3102!)

```
attach(longjump)
splot(LongJump, Year)
```



Problem 4:

Consider the following data set:

```
kable(p4data)
```

x	y
10	58
11	54
12	51
13	52
14	62
15	57
16	63
17	64
18	69
19	71
20	70

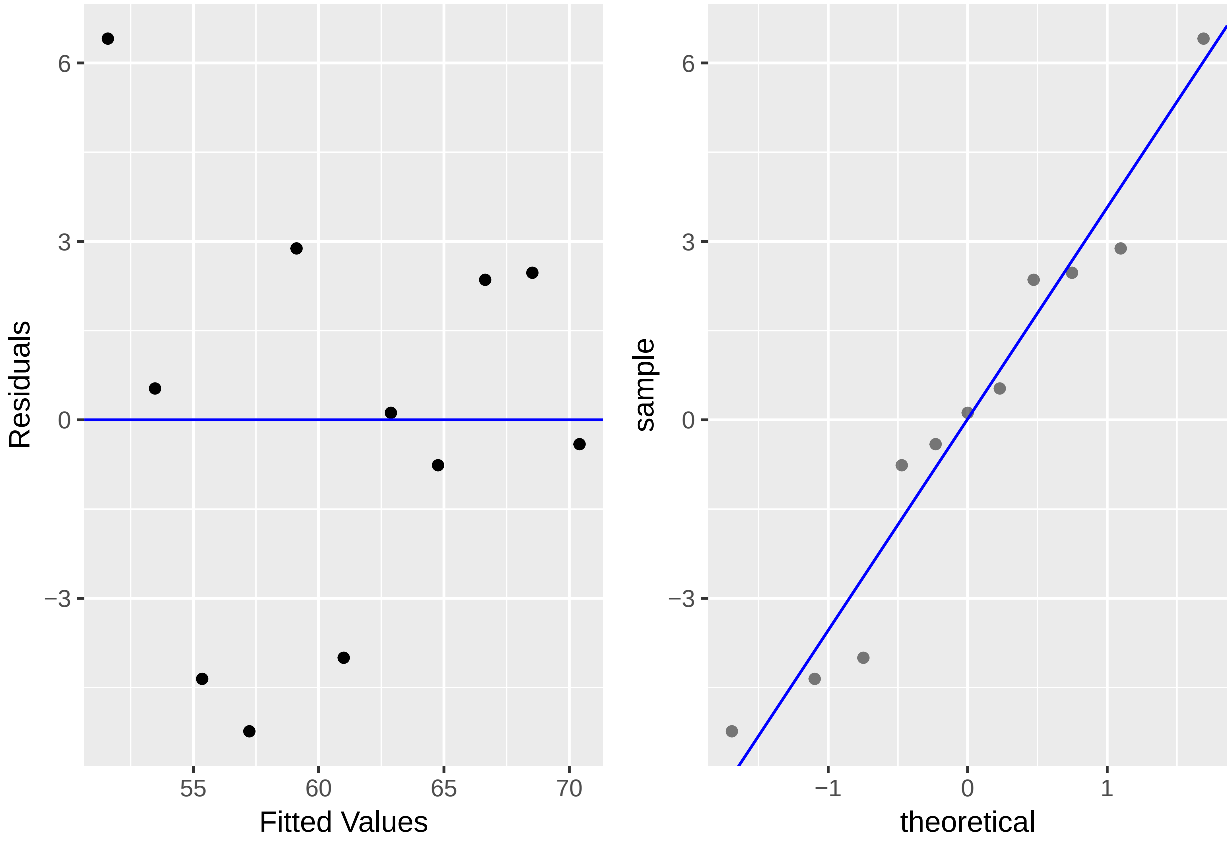
Find the least squares regression equation and use it to predict the y value for an observation with x=15

Parameter: regression coefficients

Problem: find model

Method: slr

```
slr(y=y, x=x)
```



```
## The least squares regression equation is:
##  y  = 32.773 + 1.882 x
## R^2 = 75.79%
```

```
32.773 + 1.882*15
```

```
## [1] 61.003
```

so y=61 is the prediction.