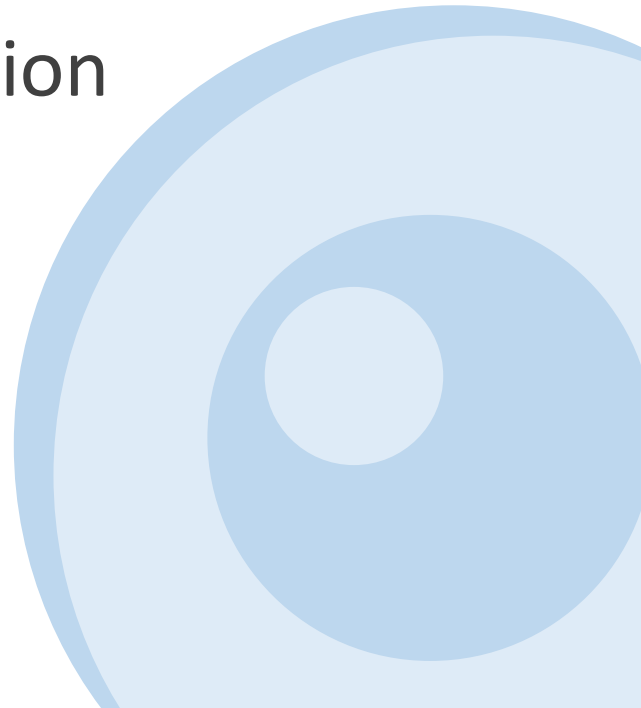


PSYC 193R:

Data Analysis & Visualization with R

Lecture 13: Simple Correlation / Regression

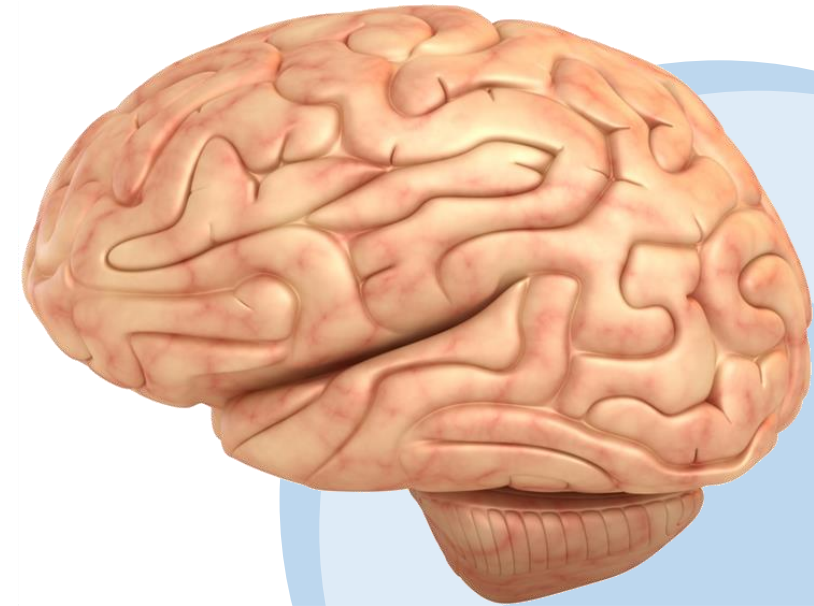
Jonas S.-H. Lau [silau -et- ucsd.edu]
psyc193r.ucsd.edu



Brain Size and Intelligence

People have long suspected that people with larger brains are smarter. To test the idea, a group of scientists recruited a number of healthy participants for a study. The scientists scanned the participants' brains using MRI, and also gave them a standardized intelligence test.

Note: they are only testing whether there is a relationship between brain size and the intelligence. They are *not* testing whether size of the brain *affects* intelligence.



Sources:

<https://www.scientificamerican.com/article/does-size-matter-for-brains/>

http://www.livescience.com/images/i/000/039/847/original/shutterstock_101000707.jpg

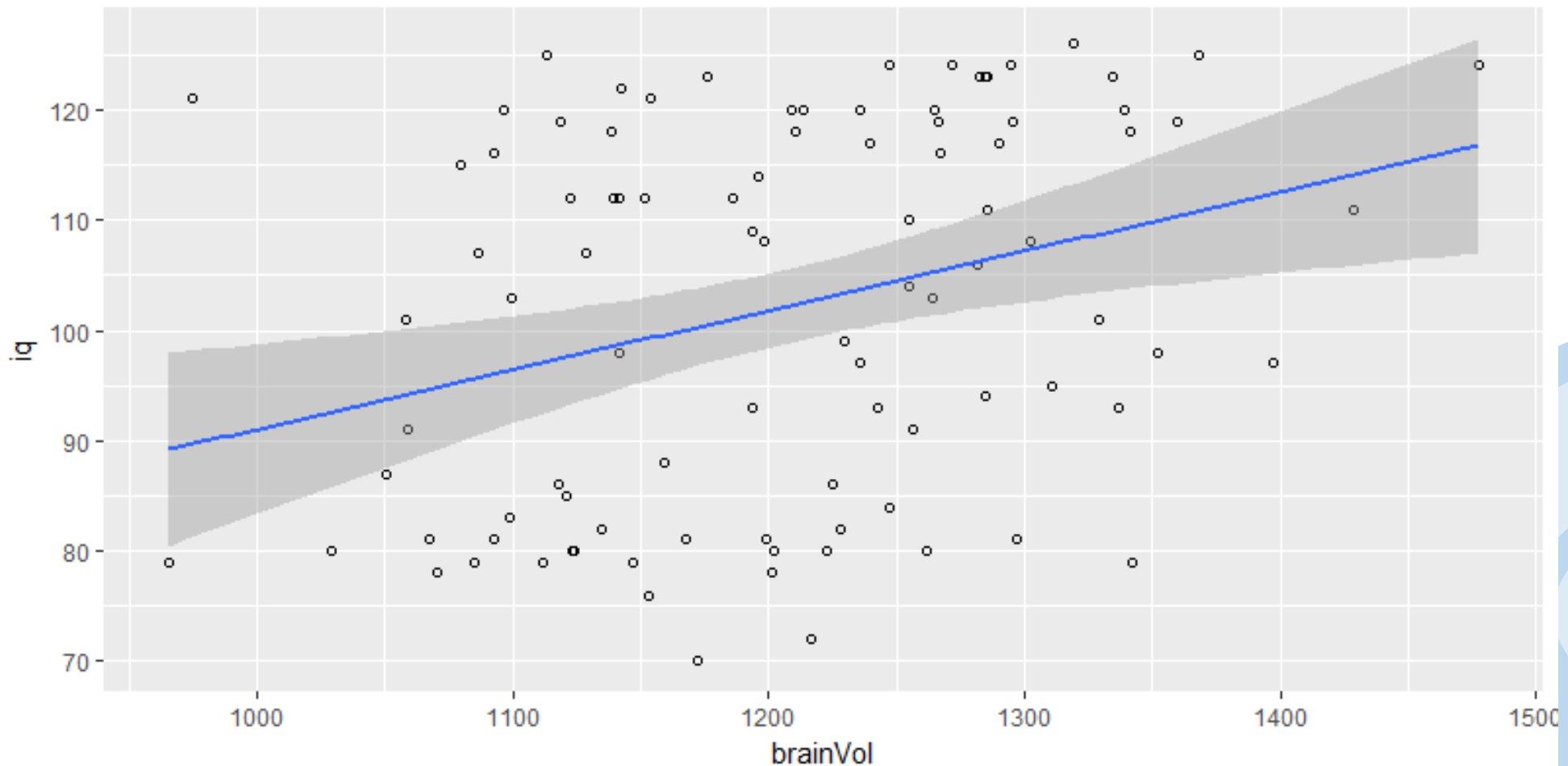
Brain Size and Intelligence: The Design

- What was the research question? What was the target population?
- Was it an experiment, quasi-experiment, or an observational study?
- What were the IV(s) and DV(s)?
- How were the IV(s) and DV(s) operationalized?
(also, scale of measurement?)
- Was it a between / within-subjects design?
- What are the expected results?
If the manipulation does not work:
If the manipulation works:



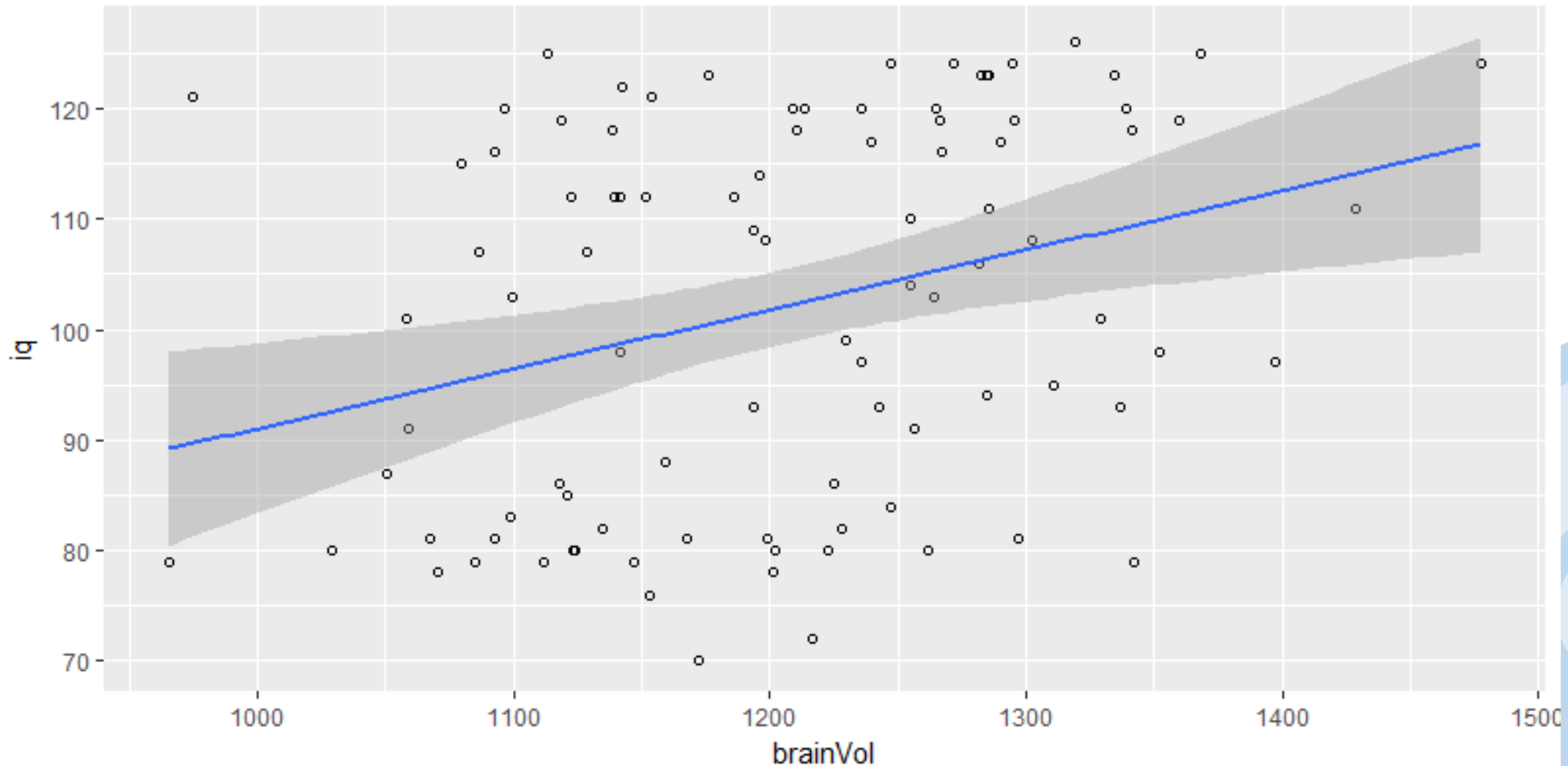
Goal 1: *linear* relationship between variables

$$r = \frac{SP_{XY}}{\sqrt{(SS_X)(SS_Y)}}$$



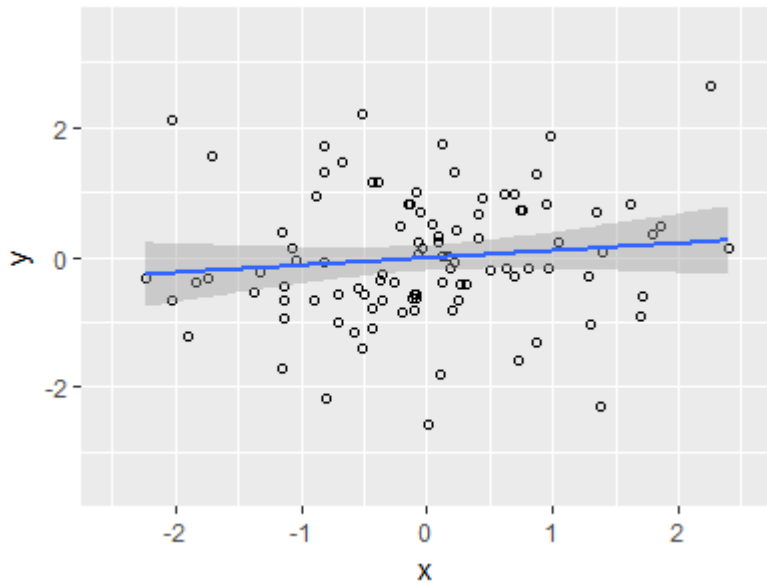
Goal 1: *linear* relationship between variables

$$SS_X = \frac{\Sigma(X - M_X)^2}{n} \quad SS_Y = \frac{\Sigma(Y - M_Y)^2}{n} \quad SP_{XY} = \Sigma XY - \frac{\Sigma X \times \Sigma Y}{n} \quad r = \frac{SP_{XY}}{\sqrt{(SS_X)(SS_Y)}}$$

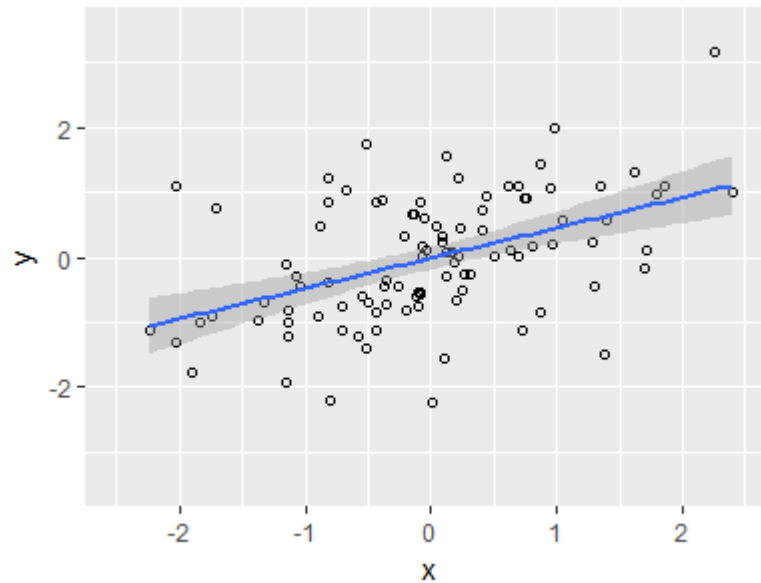


Correlation coefficient r ranges from -1 to +1

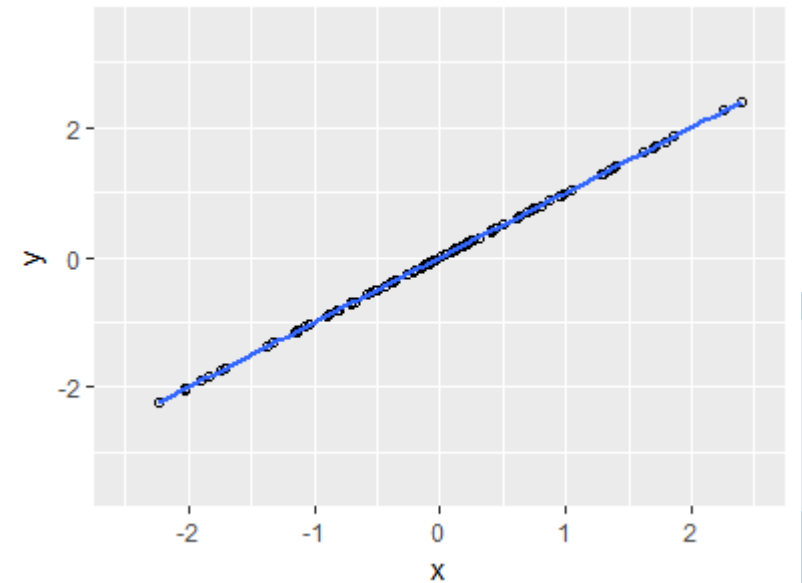
$r = +0.15$



$r = +0.50$

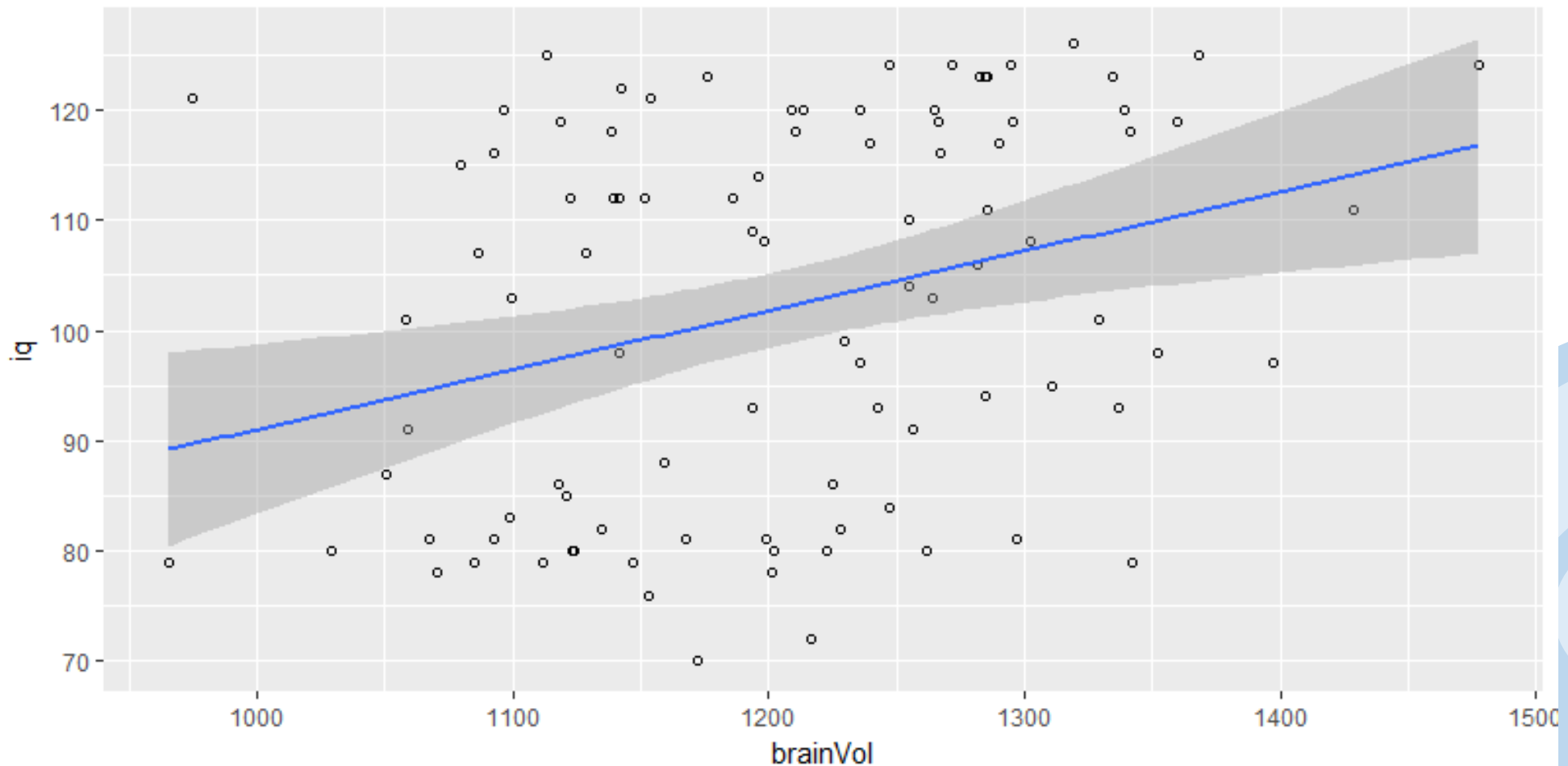


$r = +1.0$



Goal 2: describe the relationship in a math form

$$\hat{Y} = bX + a$$



Import Data

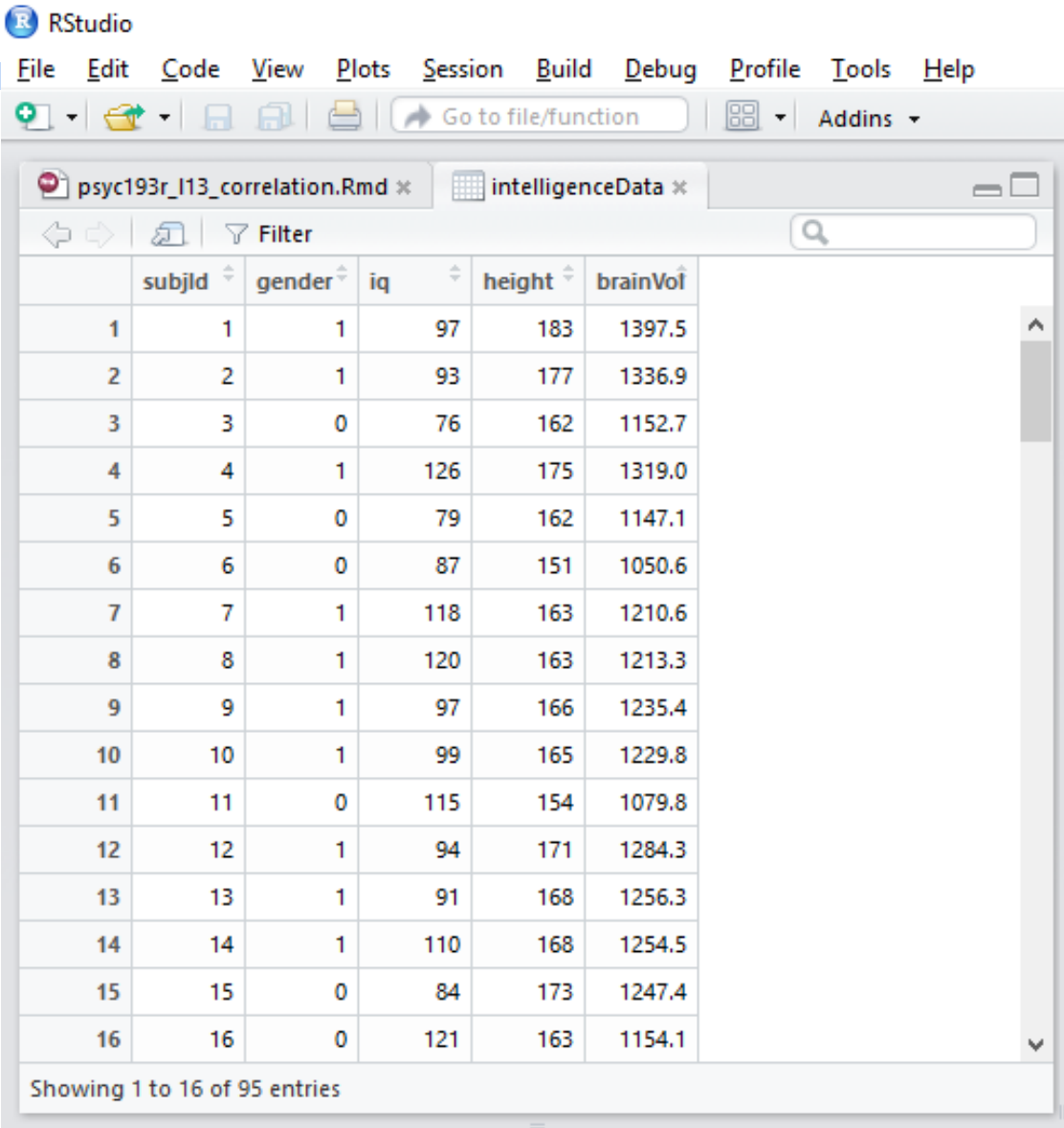
- Download the data file:
https://psyc193r.ucsd.edu/data/psyc193r_intelligence_data.csv
- Tell R where your data file is:
Session → Set Working Directory → Choose Directory
`setwd()`
- Check the *comma separated values* (csv) file with MS Excel to ensure correct format
- Import the file into R as a data frame, saving it to “intelligenceData” variable
`... = read.csv("file_name.csv")`

Ideal Data Structure

“Long” data structure with three variables, one row represents an entry from a participant

For the columns:

- One for participants identification
- One for the IV
- One for the DV



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

psyc193r_l13_correlation.Rmd x intelligenceData x

Filter

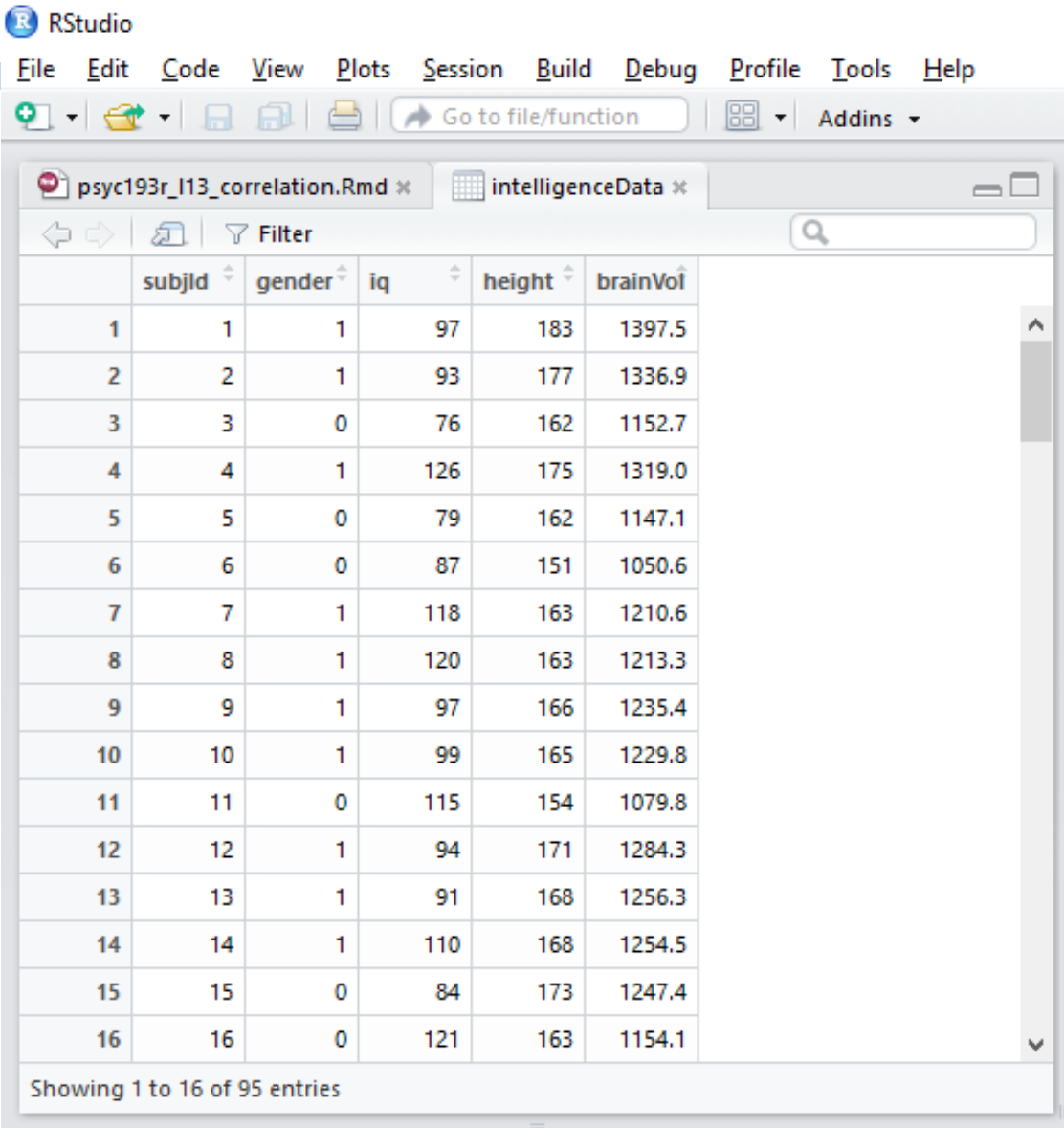
	subjId	gender	iq	height	brainVol
1	1	1	97	183	1397.5
2	2	1	93	177	1336.9
3	3	0	76	162	1152.7
4	4	1	126	175	1319.0
5	5	0	79	162	1147.1
6	6	0	87	151	1050.6
7	7	1	118	163	1210.6
8	8	1	120	163	1213.3
9	9	1	97	166	1235.4
10	10	1	99	165	1229.8
11	11	0	115	154	1079.8
12	12	1	94	171	1284.3
13	13	1	91	168	1256.3
14	14	1	110	168	1254.5
15	15	0	84	173	1247.4
16	16	0	121	163	1154.1

Showing 1 to 16 of 95 entries

Ideal Data Structure

Relevant columns:

- subjId: participant identification
- gender: 0 -- female; 1 -- male
- iq: score on a standardized intelligence test
- height: height of the participant
- brainVol: brain volume acquired from MRI scan



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

psyc193r_l13_correlation.Rmd x intelligenceData x

Filter

	subjId	gender	iq	height	brainVol
1	1	1	97	183	1397.5
2	2	1	93	177	1336.9
3	3	0	76	162	1152.7
4	4	1	126	175	1319.0
5	5	0	79	162	1147.1
6	6	0	87	151	1050.6
7	7	1	118	163	1210.6
8	8	1	120	163	1213.3
9	9	1	97	166	1235.4
10	10	1	99	165	1229.8
11	11	0	115	154	1079.8
12	12	1	94	171	1284.3
13	13	1	91	168	1256.3
14	14	1	110	168	1254.5
15	15	0	84	173	1247.4
16	16	0	121	163	1154.1

Showing 1 to 16 of 95 entries

Data Formatting / Processing / Cleaning

- Factorize subjId

```
intelligenceData$subjId = factor(intelligenceData$ ____ )
```

- Coerce iq / brainVol to a numeric variable

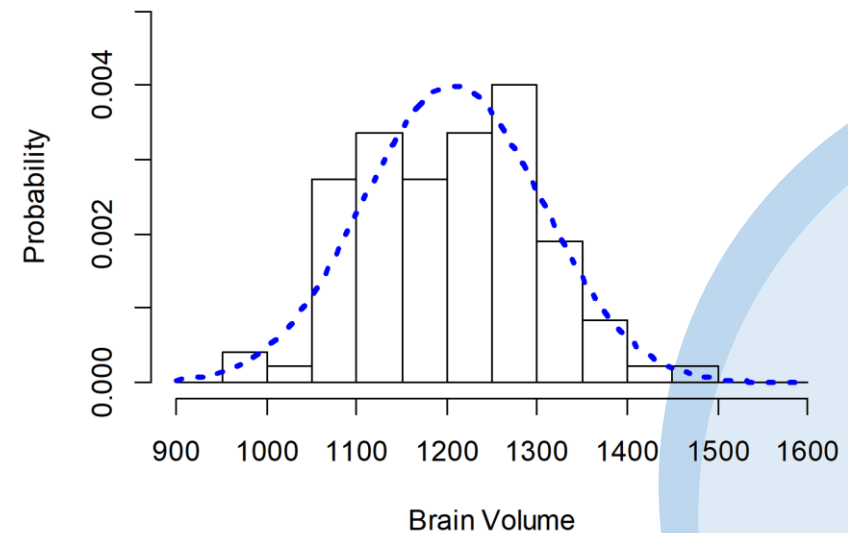
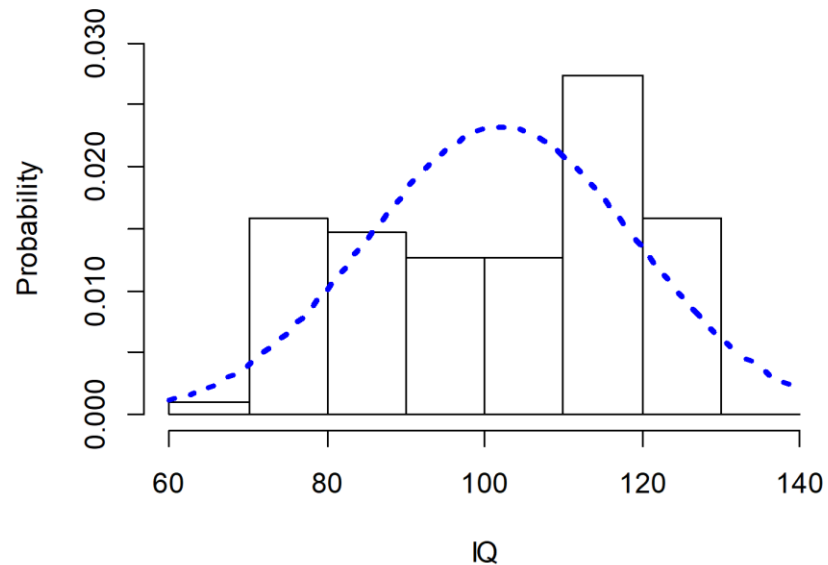
```
intelligenceData$iq = as.numeric(intelligenceData$ ____ )  
intelligenceData$brainVol = as.numeric(intelligenceData$ ____ )
```



Checking Assumptions (1)

- Normality Assumptions → $n = 30$
 - Outliers
- Plot a histogram for each variable

```
hist(intelligenceData$ ____ )  
hist(intelligenceData$ ____ )
```

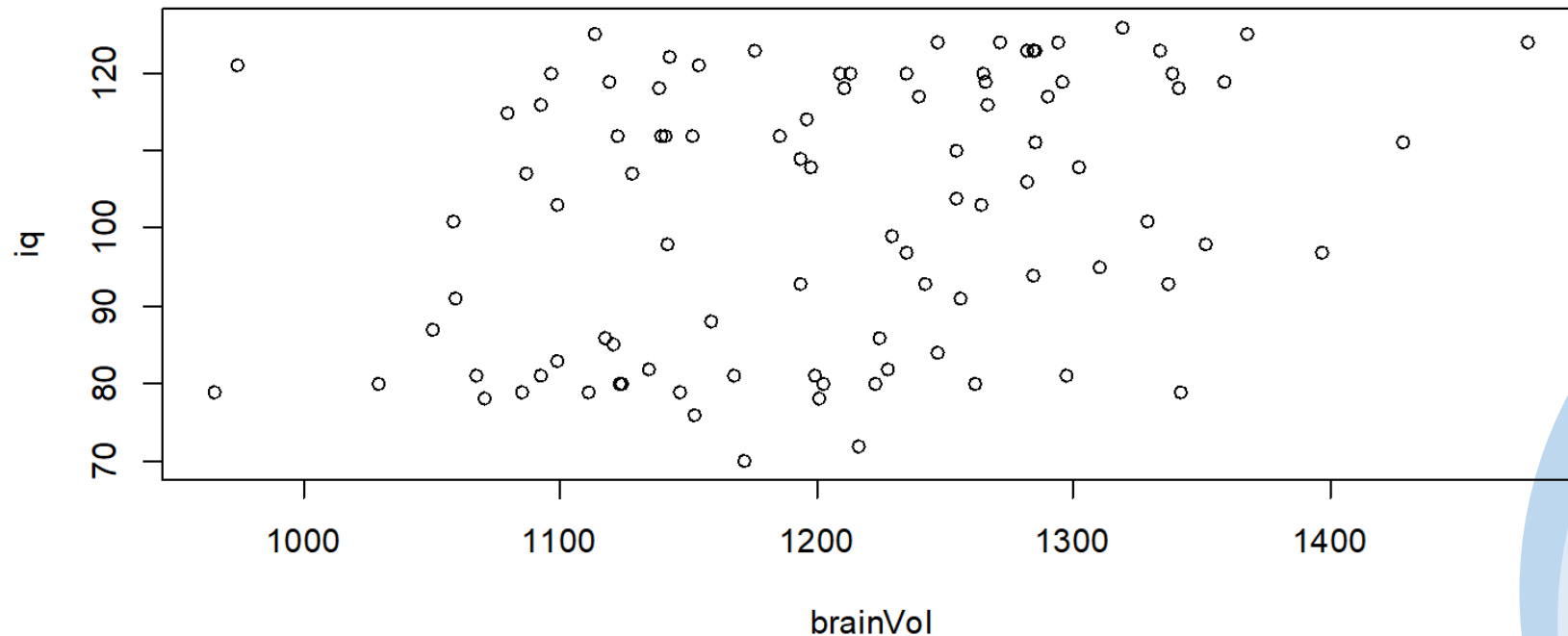


Checking Assumptions (2)

- Outliers

Plot a scatterplot to visually inspect outliers

```
with(intelligenceData, plot( x , y ))
```



Correlation & Effect Size

$$SS_X = \frac{\Sigma(X - M_X)^2}{n} \quad SS_Y = \frac{\Sigma(Y - M_Y)^2}{n} \quad SP_{XY} = \Sigma XY - \frac{\Sigma X \times \Sigma Y}{n} \quad r = \frac{SP_{XY}}{\sqrt{(SS_X)(SS_Y)}}$$

- Or use `cor( ____, ____, )`

- We would test r against $\rho = 0$

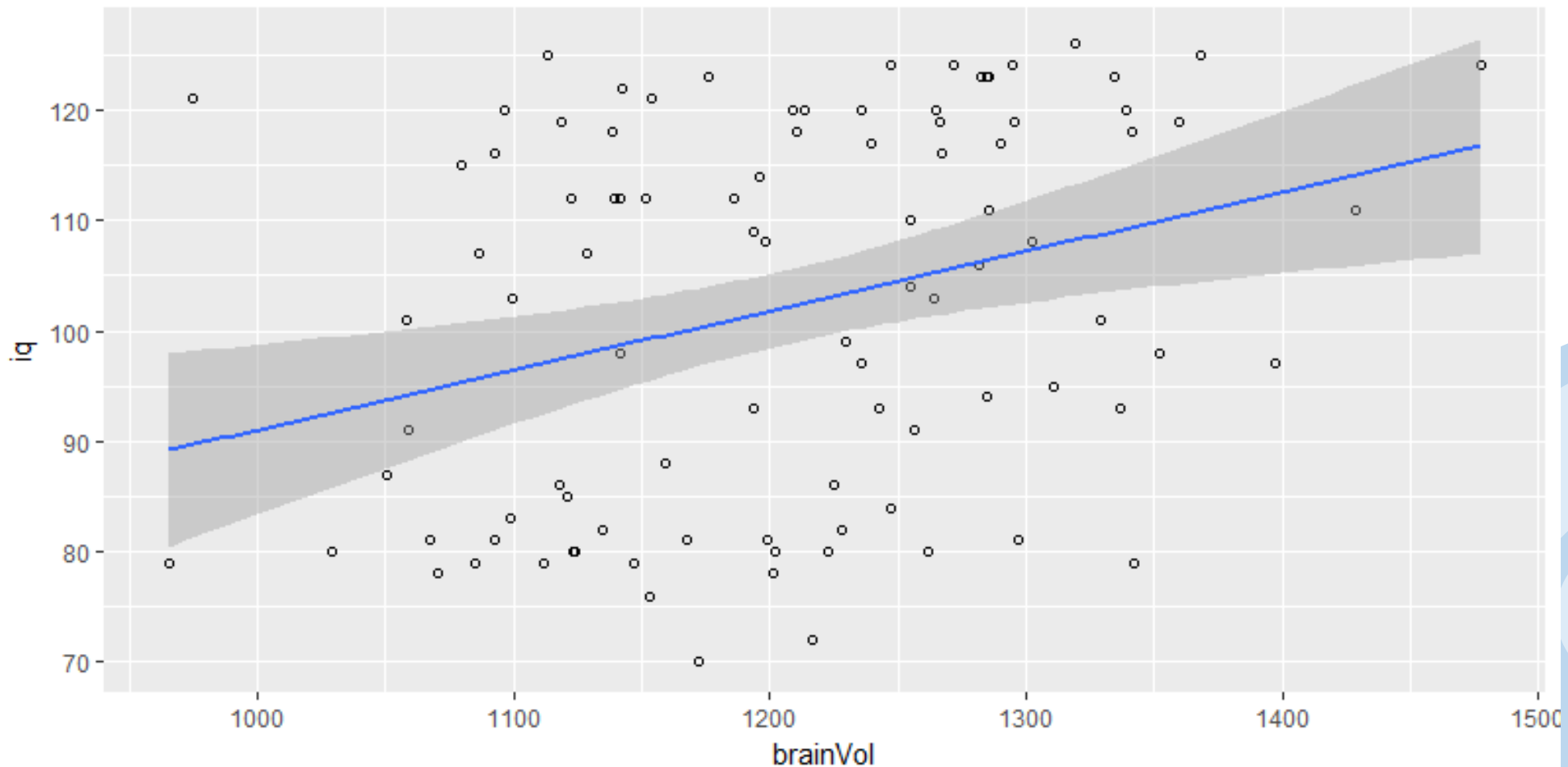
$H_0: \rho = 0$
 $H_1: \rho \neq 0$ } In summary() output

$$t_{sample} = \frac{r - \rho}{\sqrt{\frac{(1 - r^2)}{(n - 2)}}}$$

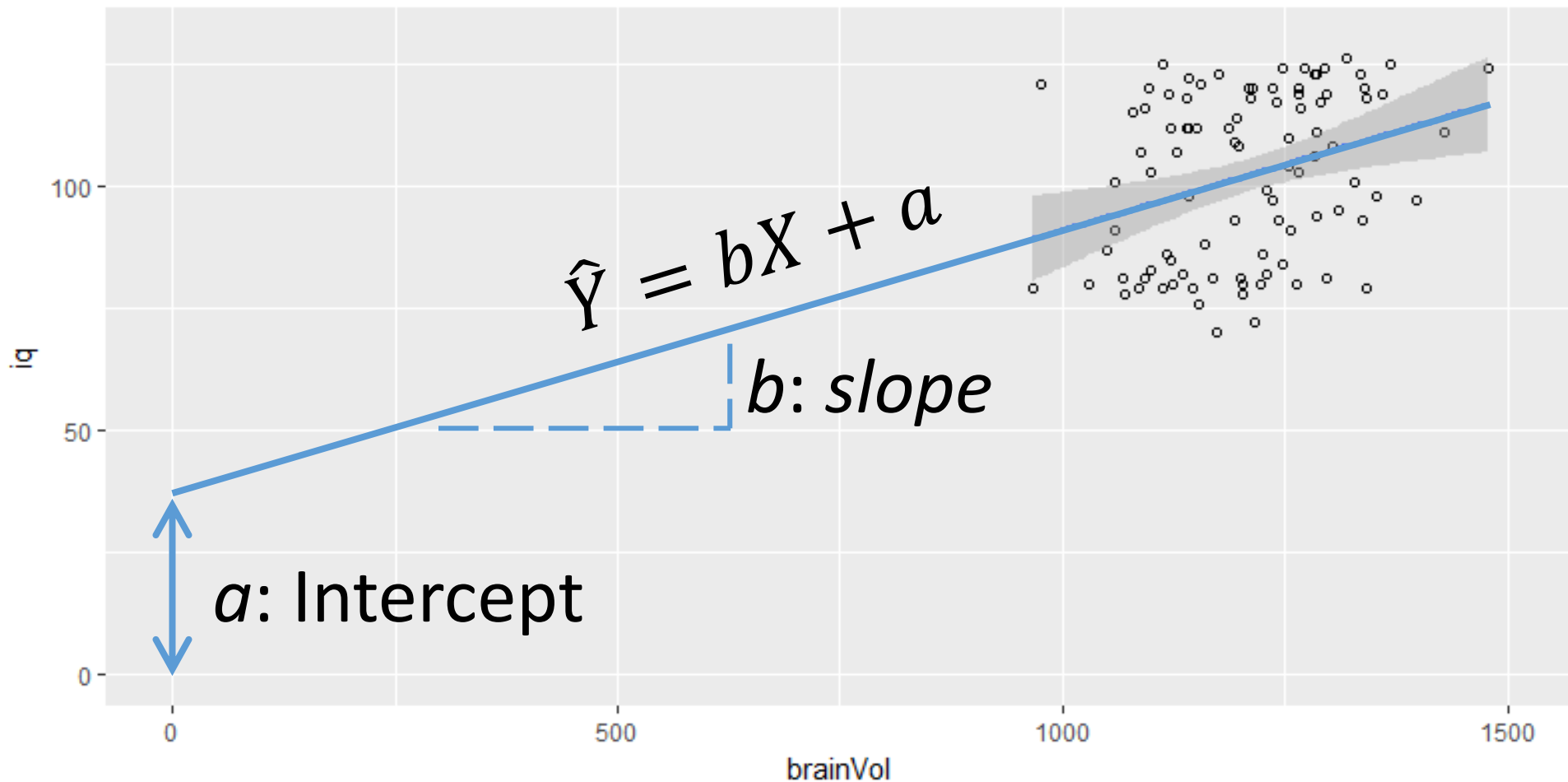
- Effect Size (r^2) in Correlation: % variance explained by the model
 $\therefore$ ____ % of the variance in intelligence is explained by brain volume.

Regression: relationship between the variables

$$\hat{Y} = bX + a$$



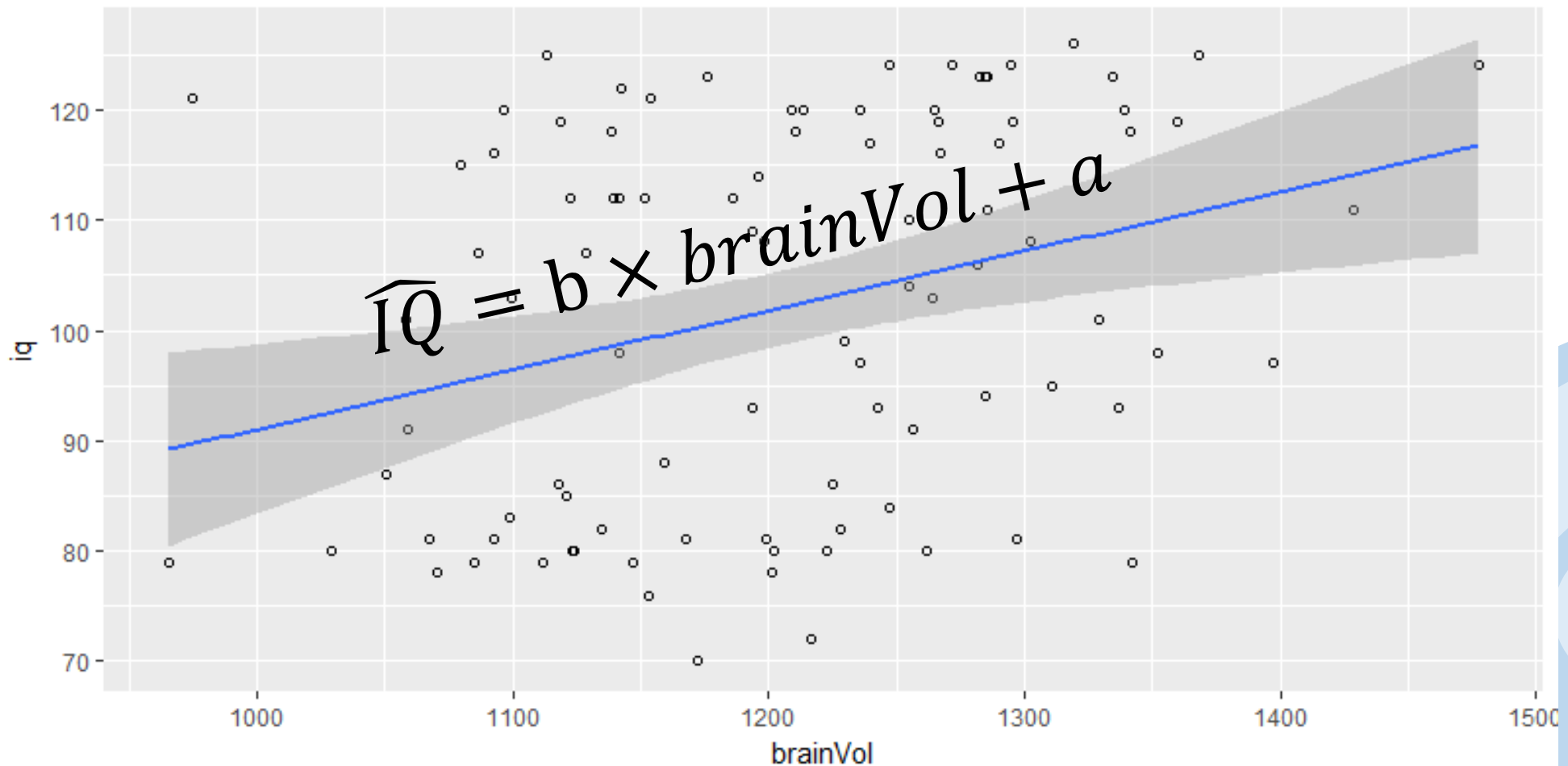
Regression: relationship between the variables



Regression: relationship between the variables

$$b = \frac{SP}{SS_X}$$

$$a = M_Y - b \times M_X$$



Regression: relationship between the variables

- Build a linear model (lm)

```
iq.lm = lm( ___ ~ ___, data = intelligenceData)
```

- Read out the summary of the model

```
summary(iq.lm)
```

call:

```
lm(formula = iq ~ brainVol, data = intelligenceData)
```

Residuals:

Min	1Q	Median	3Q	Max
-30.6817	-14.4198	0.7136	14.0750	31.3254

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.24801	20.49625	1.817	0.07239 .
brainVol	0.05379	0.01694	3.176	0.00203 **

coefficient

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.4 on 93 degrees of freedom

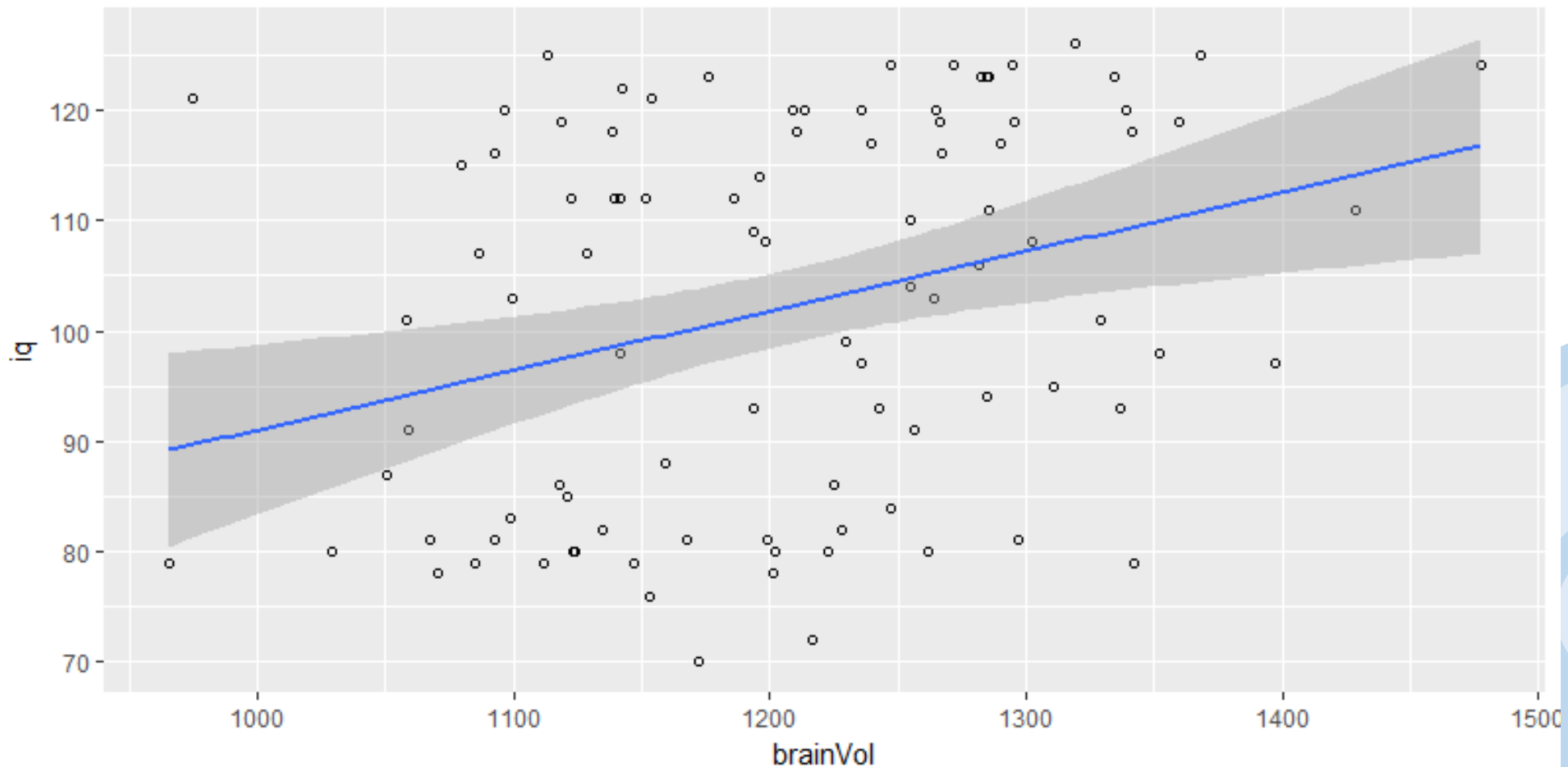
$df = n - 2$

Multiple R-squared: 0.09783, Adjusted R-squared: 0.08813

F-statistic: 10.08 on 1 and 93 DF, p-value: 0.002029

Regression: relationship between the variables

$$\widehat{IQ} = b \times brainVol + a$$



Prediction

If we meet a new friend with a brain volume of 1250 cm³, and we want to know how intelligent he is.

- Make a new data frame for the score

```
newScore = data.frame(brainVol = 1250)
```

- Use the predict function

```
predict(iq.lm, newScore, interval = "prediction")
```

	fit	lwr	upr
1	104.4892	71.70793	137.2705

Visualization

- Scatterplot only

```
# Import the ggplot2 package
library(ggplot2)

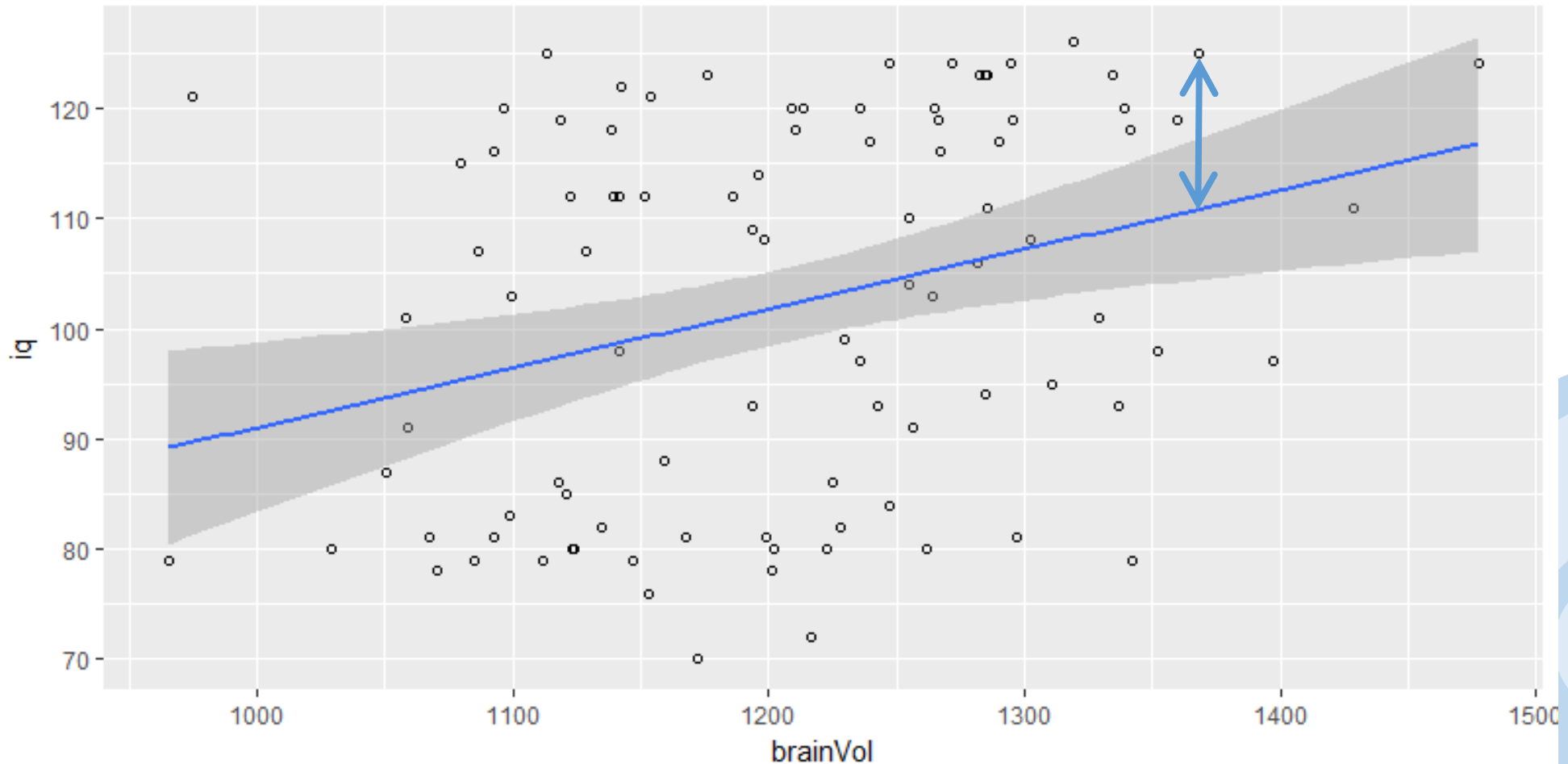
# Plot a scatterplot
ggplot(data = dataset, aes(x= iv, y= dv )) +
  geom_point(shape=1)      # Use hollow circles
```

- Scatterplot with fitted curve

```
# Plot a scatterplot with geom_point
ggplot(data = intelligenceData, aes(x=brainVol, y=iq)) +
  geom_point(shape=1) +      # Use hollow circles
  geom_smooth(method=lm)     # Add linear regression line
                              # (by default includes 95% confidence region)
```

Residuals

- A residual is the difference between its actual Y value and its expected value, i.e., $y - \hat{y}$



Residual Plot: post-fit checking

- A residuals is the difference between its actual Y value and its expected value, i.e., $y - \hat{y}$
- Calculated the $\hat{y}$:

```
iq.res = resid( the_lm_model )
```

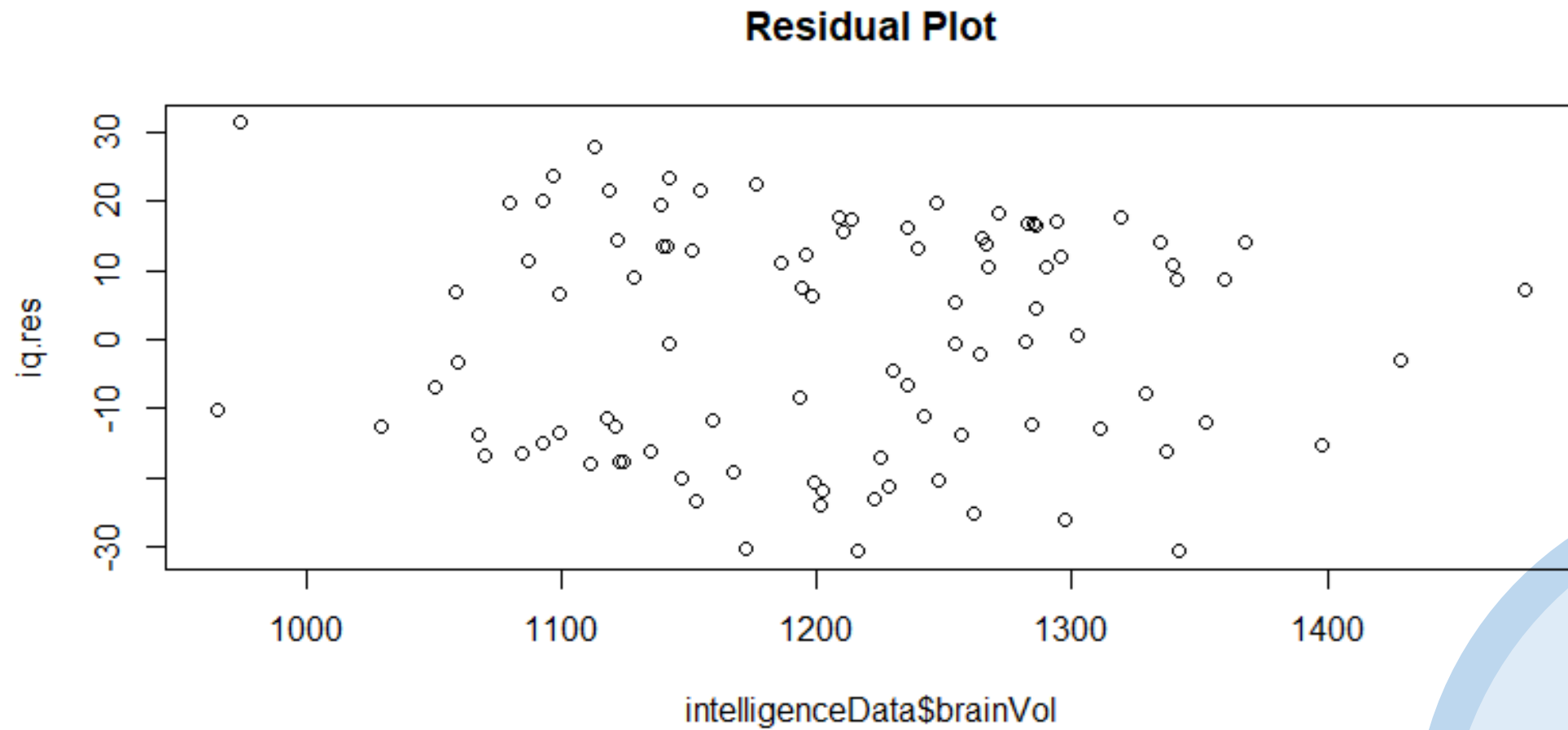
- Plot the residual graph:

```
plot( x, iq.res, main = "Residual Plot")
```

- The points should distribute randomly (like a cloud)



Residual Plot



Wrap up: Correlation / Regression

- Assumptions checking: Linearity / Normality / Outliers
- Correlation Coefficient / Hypothesis Testing
- Variance explained by the model / Effect Size
- Regression Line
- Predictions
- Visualization
- Residuals / Post-fit checking: Residual Plot

