

Conceptual Introduction to Linear Regression

Do you like drawing lines?

We're going to draw some lines.

Do you like drawing lines?

We're going to draw some lines.



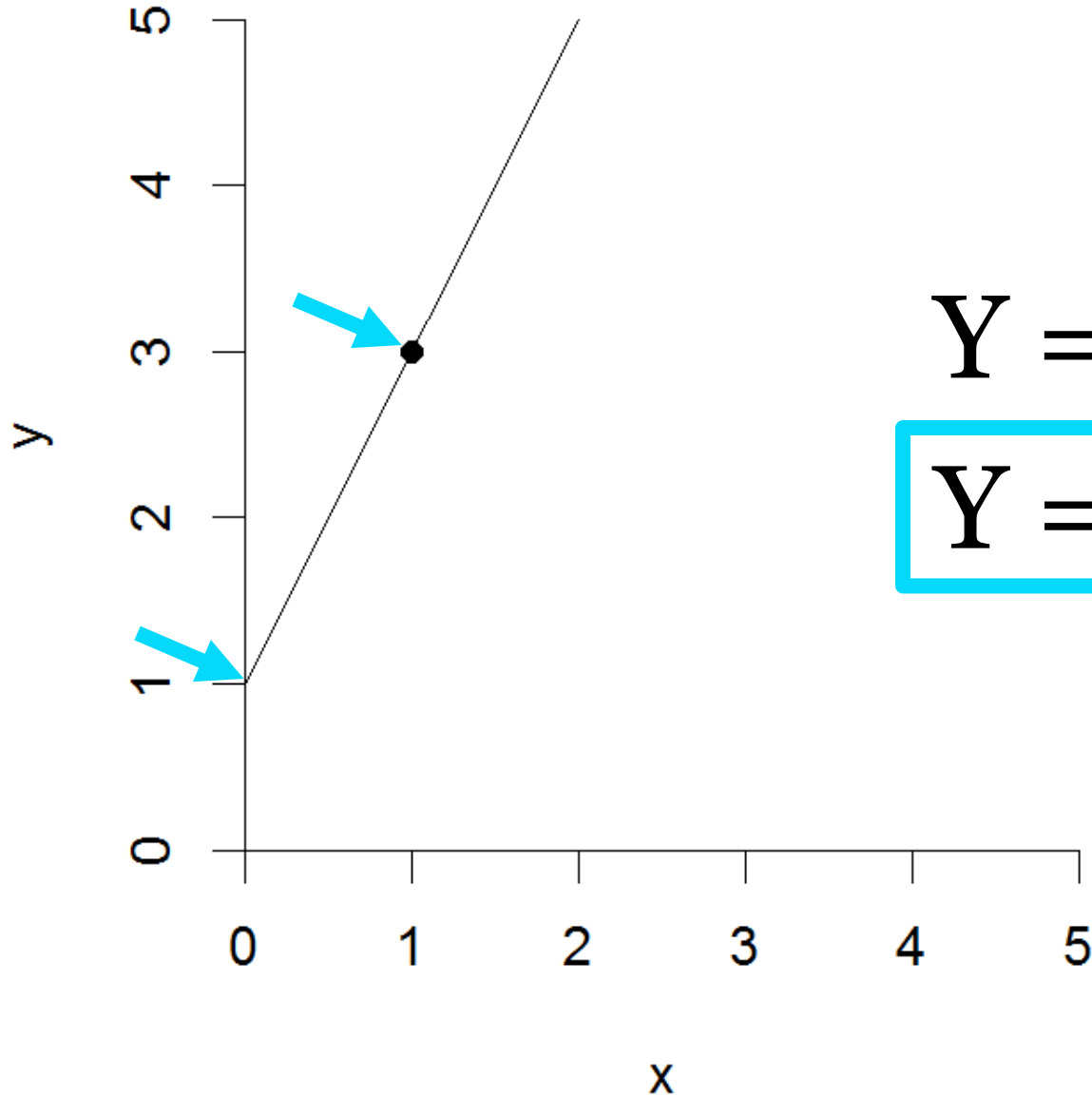
Do you like drawing lines?

We're going to draw some lines.

Part 0

What is a linear equation?

Algebra: Linear Equations

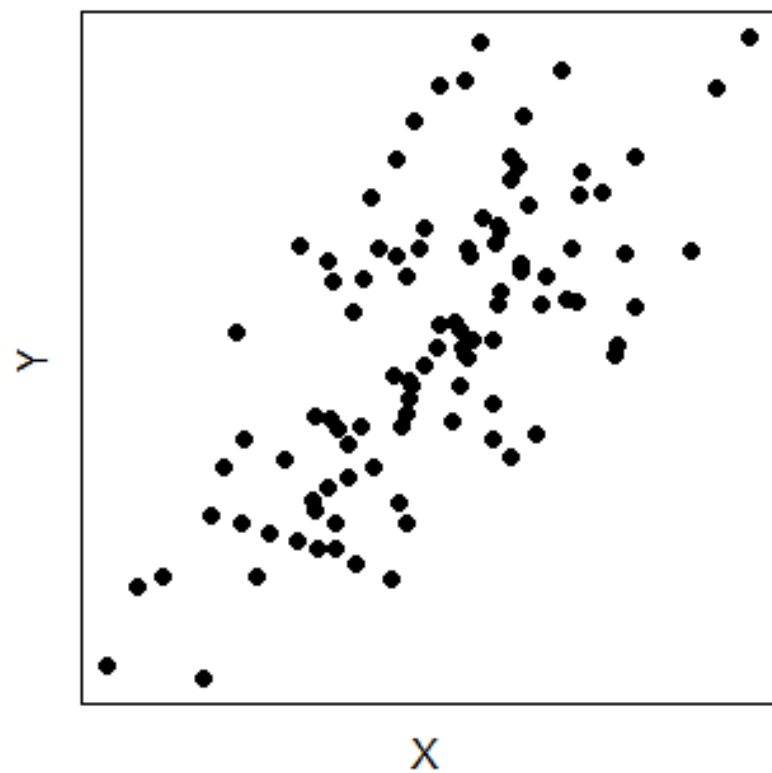


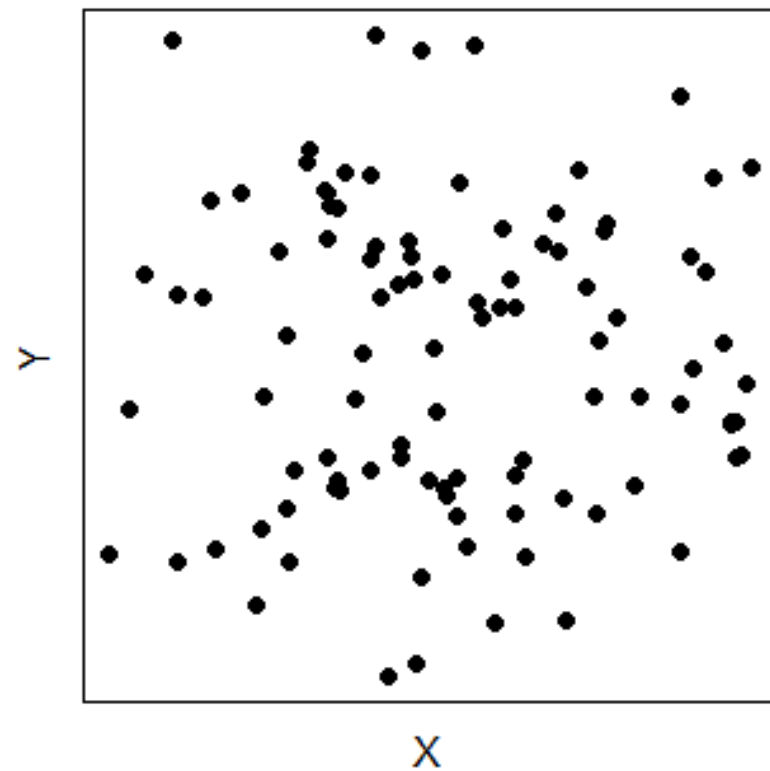
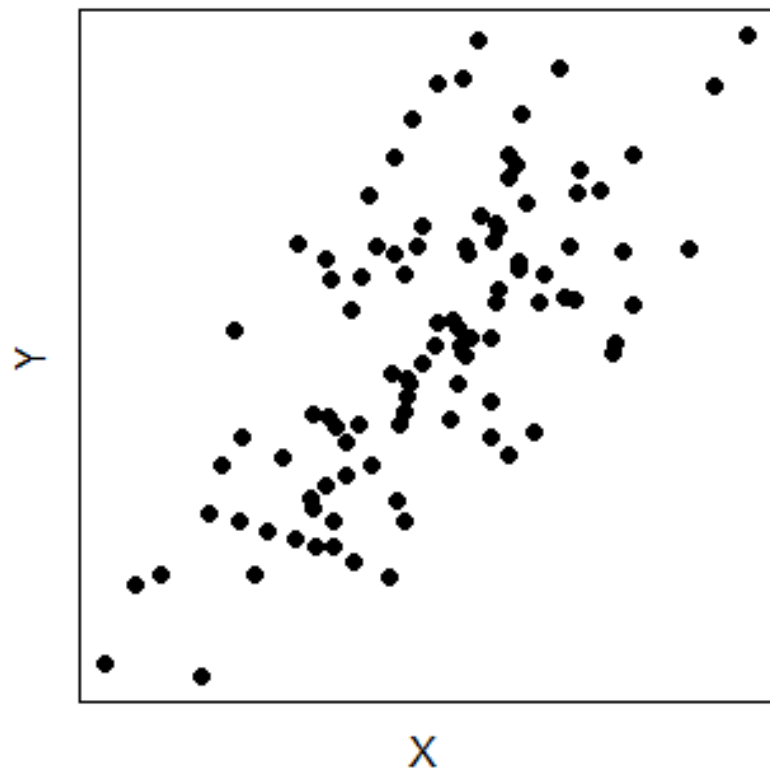
$$Y = 2X + 1$$

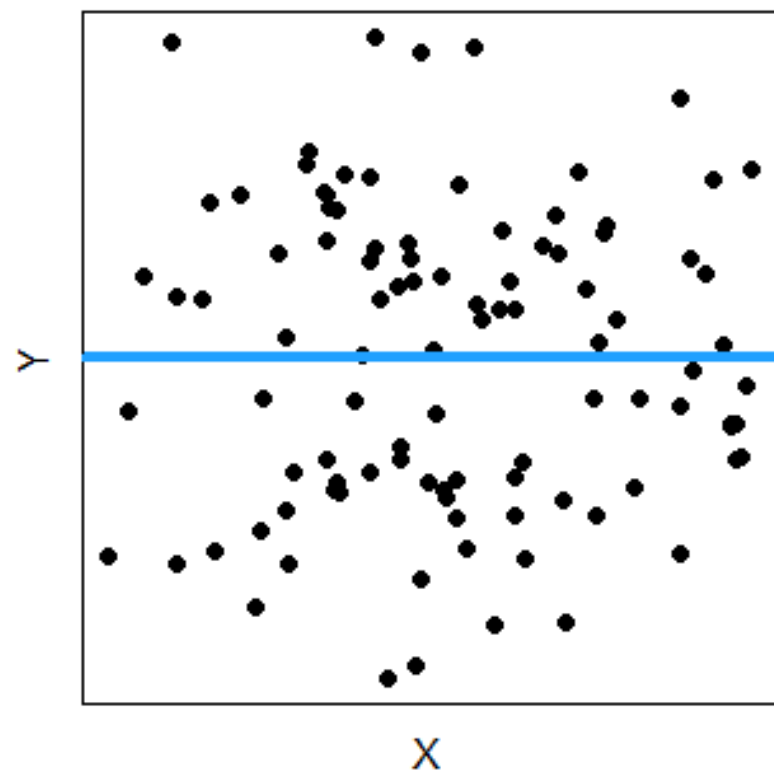
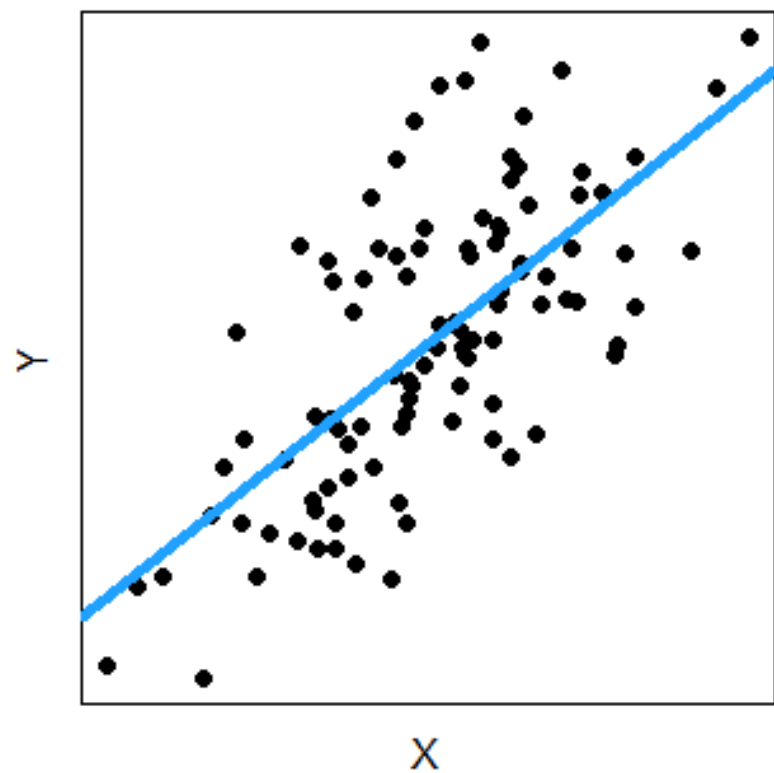
$$Y = 1 + 2X$$

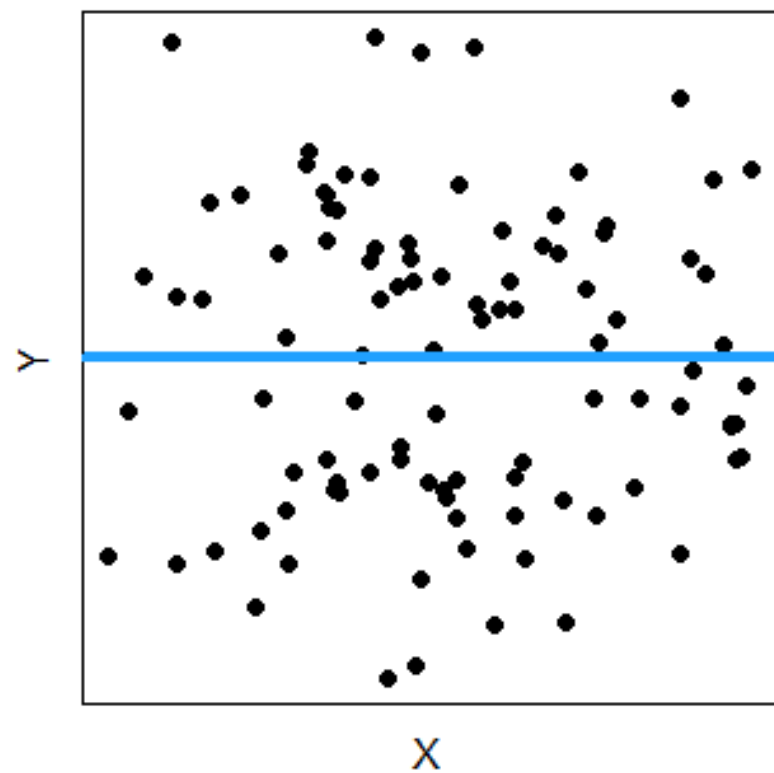
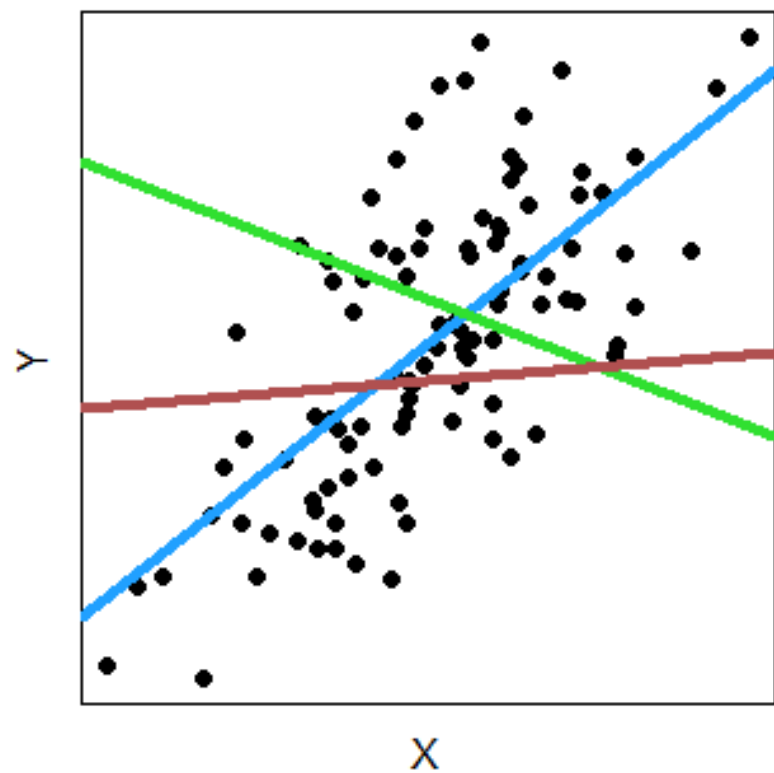
Part 1

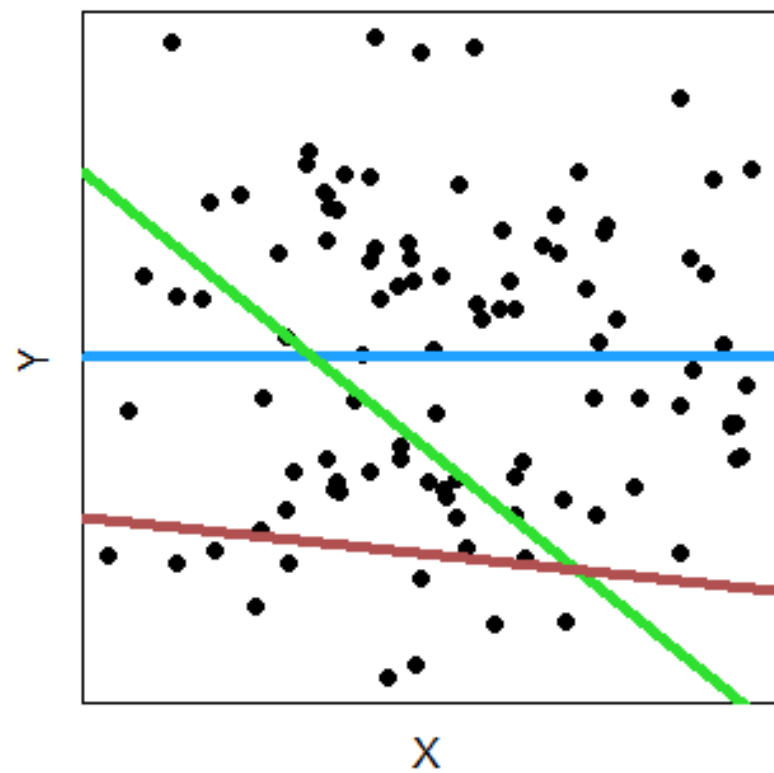
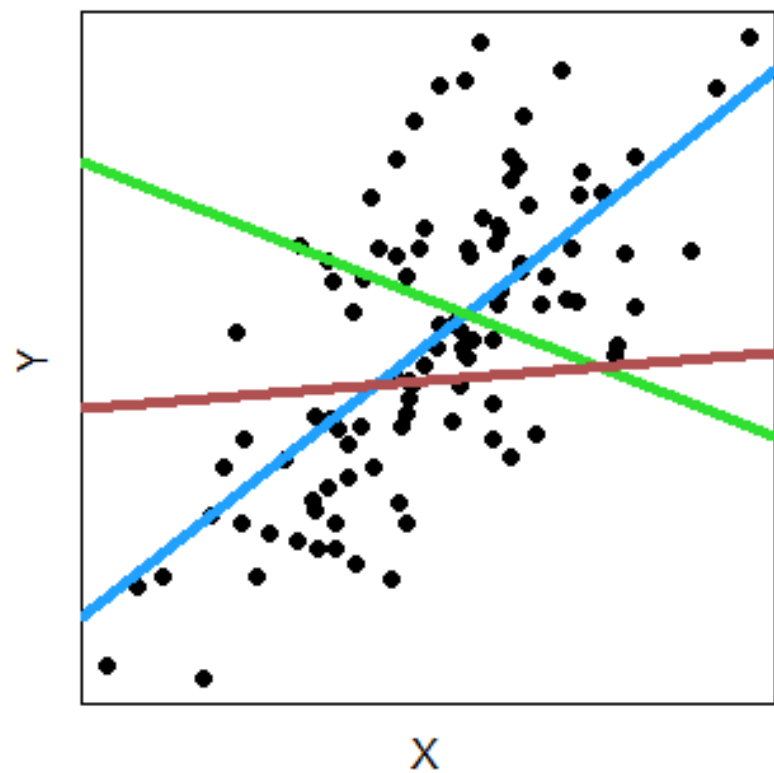
What is ordinary least squares (OLS) regression?











Simple Linear Regression

$$Y = b_0 + b_1X$$

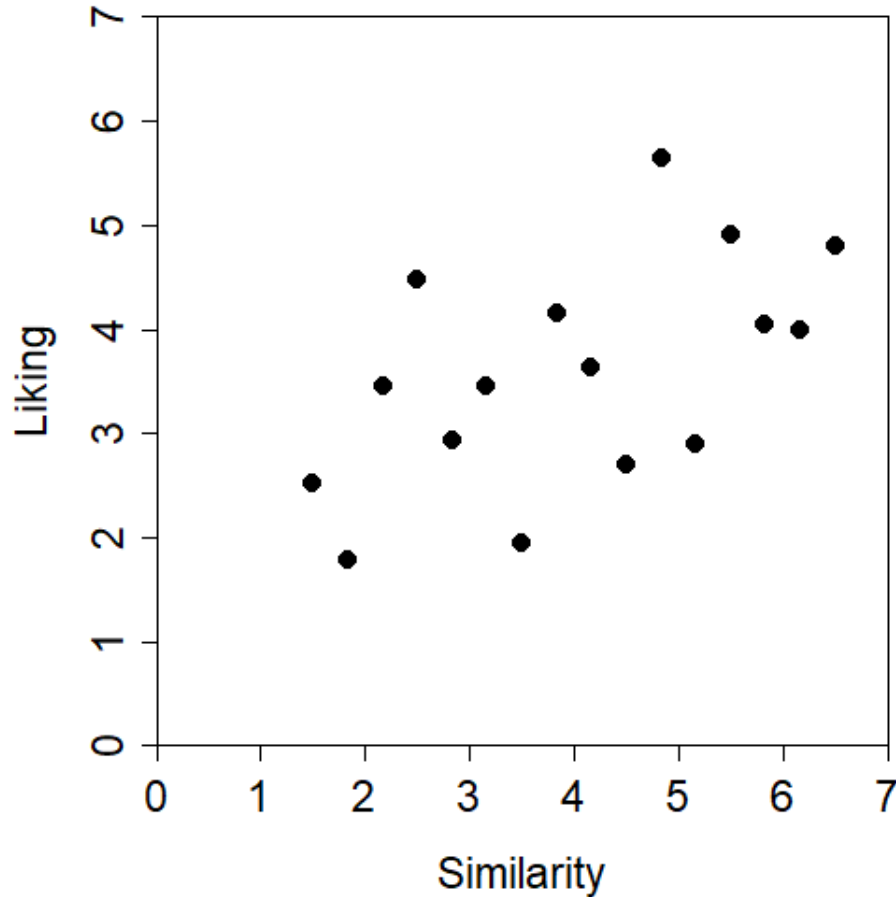


intercept



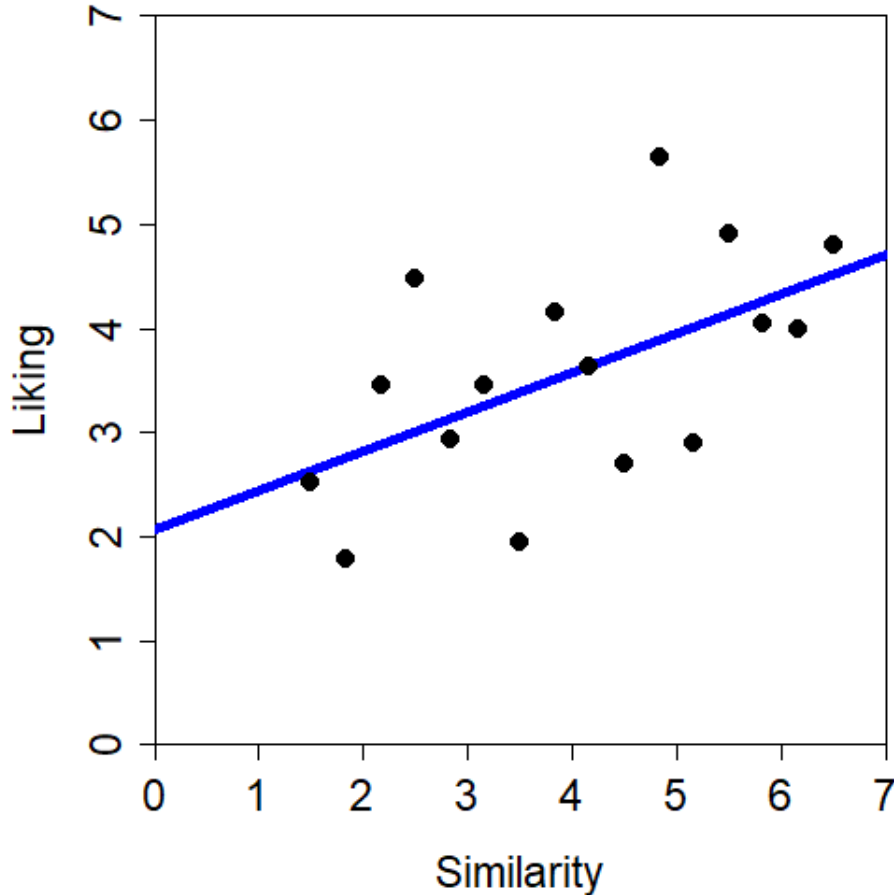
slope

Simple Linear Regression



$$Y = b_0 + b_1X$$

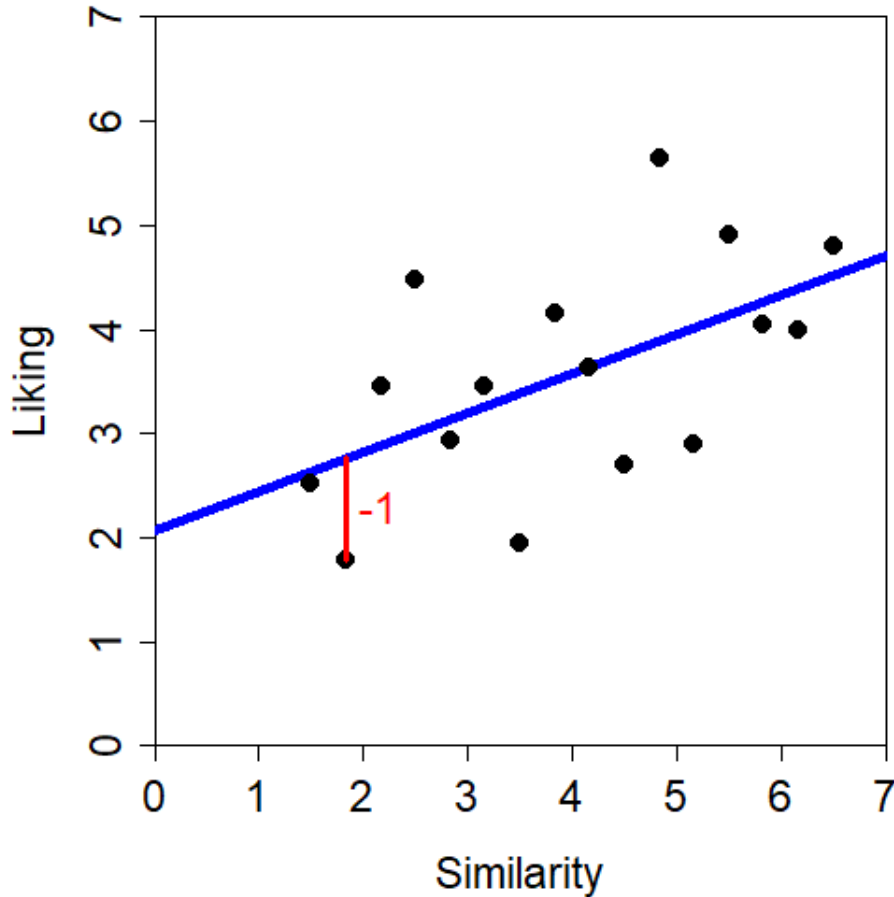
Simple Linear Regression



$$Y = b_0 + b_1X$$

$$y = 2.1 + 0.4x + e$$

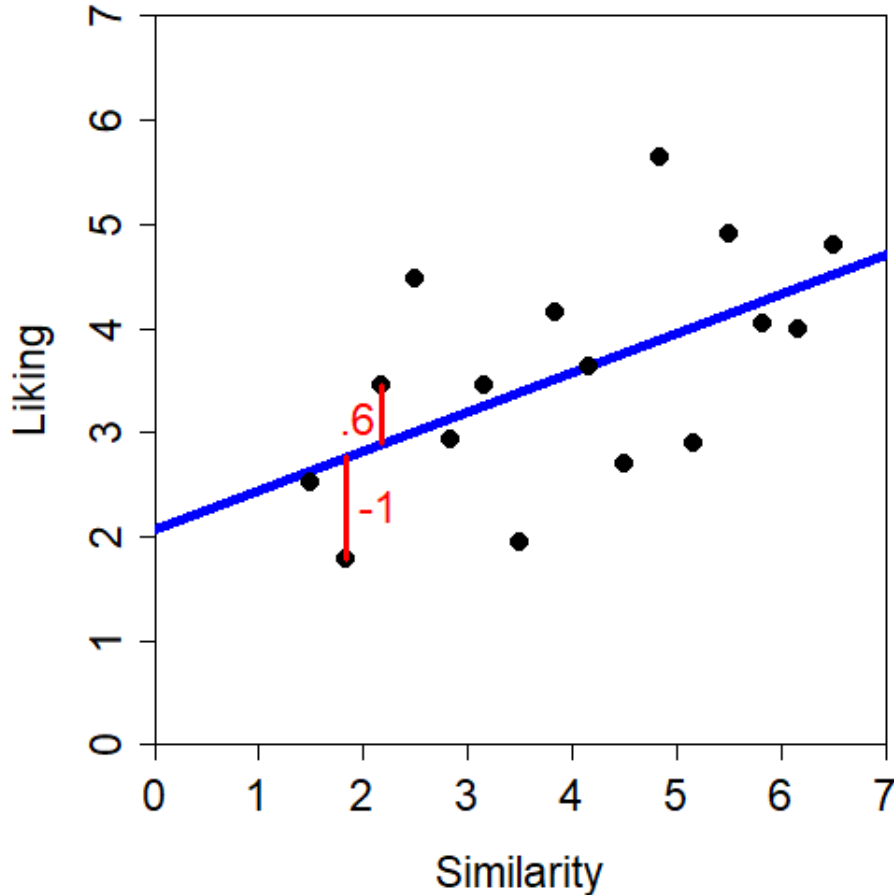
Simple Linear Regression



$$Y = b_0 + b_1X$$

$$y = 2.1 + 0.4x + e$$

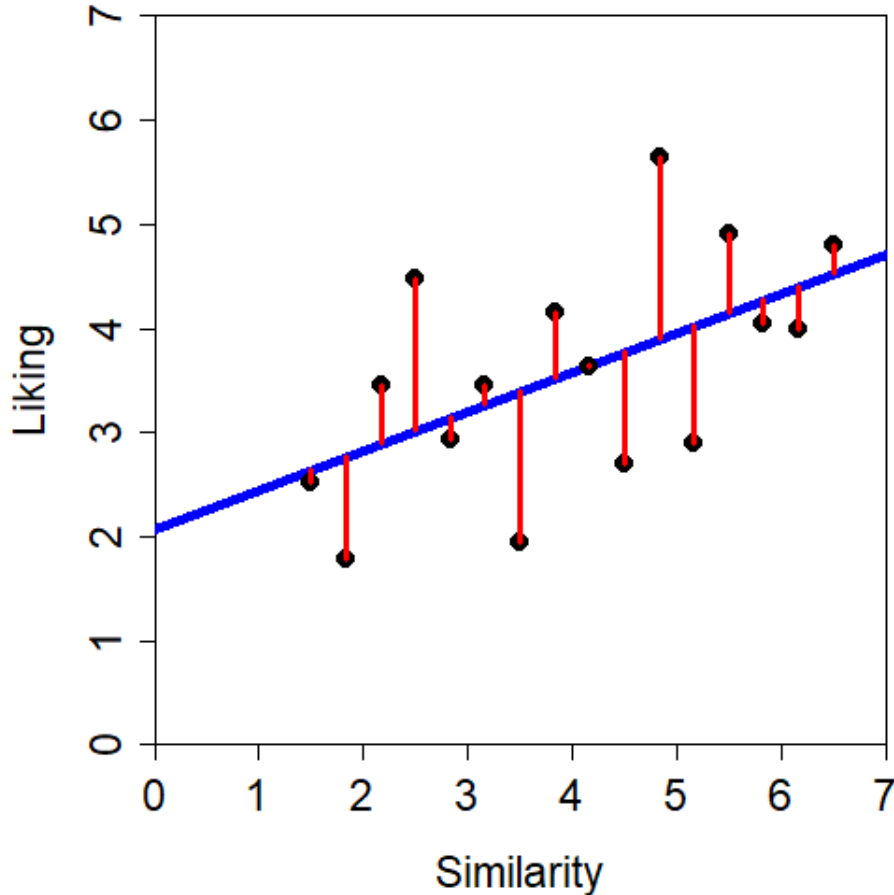
Simple Linear Regression



$$Y = b_0 + b_1X$$

$$y = 2.1 + 0.4x + e$$

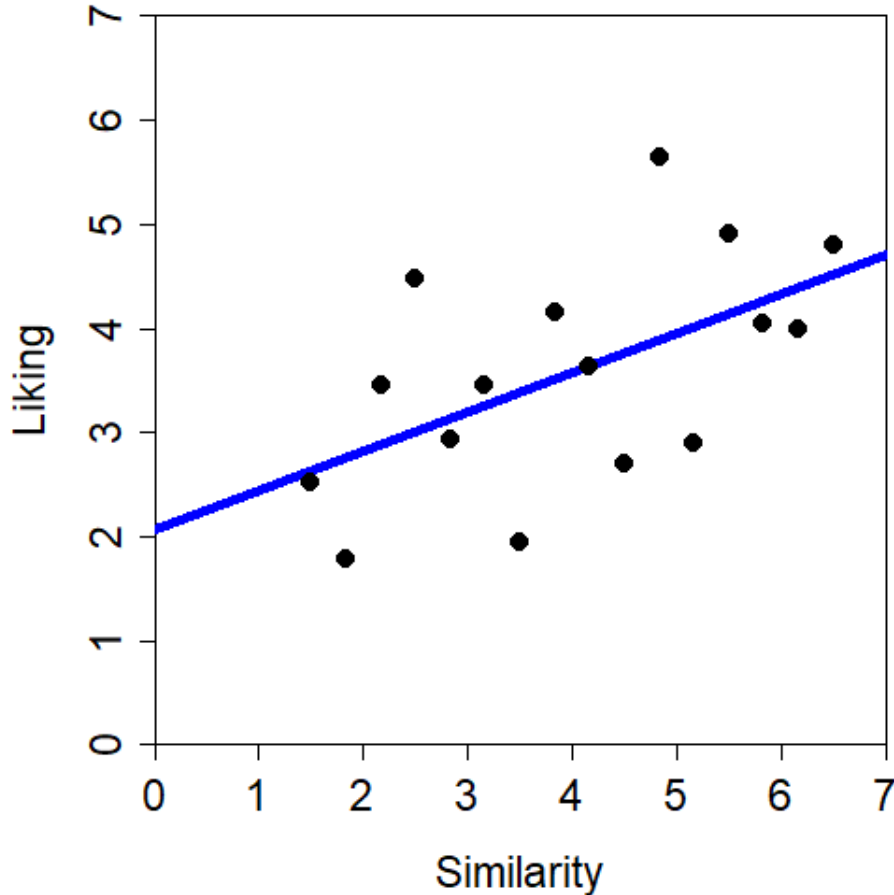
Simple Linear Regression



$$Y = b_0 + b_1X$$

$$y = 2.1 + 0.4x + e$$

Simple Linear Regression



$$Y = b_0 + b_1X$$

$$y = 2.1 + 0.4x + e$$

Simple Linear Regression

$$Y = b_0 + b_1X$$

Simple Linear Regression

$$Y = b_0 + b_1X$$



response/
criterion/
dependent
variable



predictor/
independent
variable

Simple Linear Regression

$$Y = b_0 + b_1X$$

For participant 7...

$$y_7 = b_0 + b_1x_7 + e_7$$

Simple Linear Regression

For participant i ...

$$y_i = b_0 + b_1x_i + e_i$$

Simple Linear Regression

$$Y = b_0 + b_1X$$

Multiple Linear Regression

$$Y = b_0 + b_1X_1 + b_2X_2$$

Multiple Linear Regression

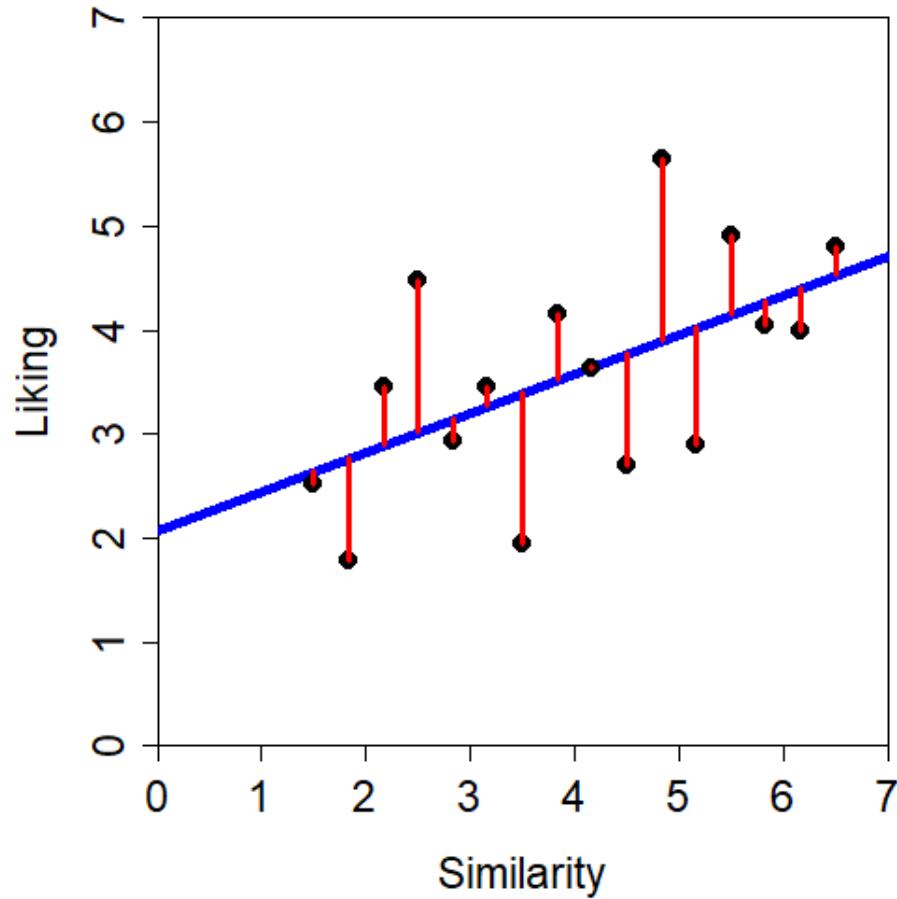
$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Multiple Linear Regression

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

$$y_i = b_0 + b_1x_{1,i} + b_2x_{2,i} + b_3x_{3,i} + e_i$$

$$Y = b_0 + b_1X$$



$$Y = b_0 + b_1X_1 + b_2X_2$$

Regression Plane

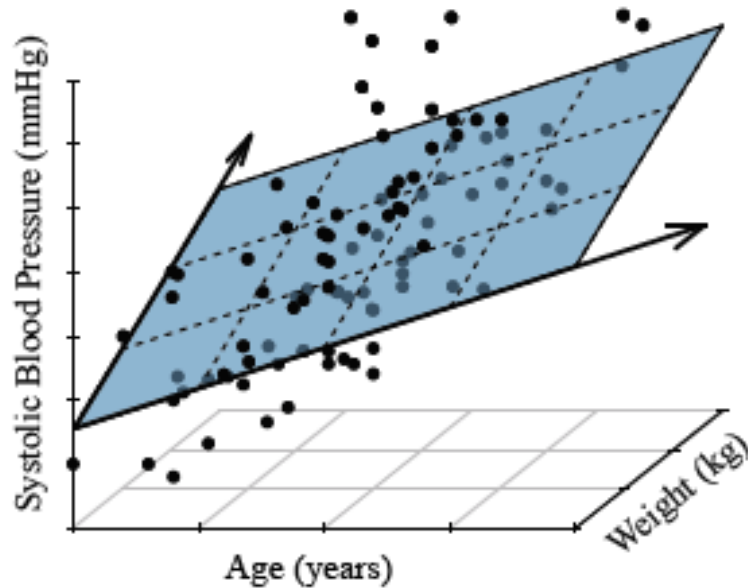


Figure 2.25: Systolic blood pressure linearly increases with age, but also with bodyweight. A line in two directions forms a plane.

Residuals

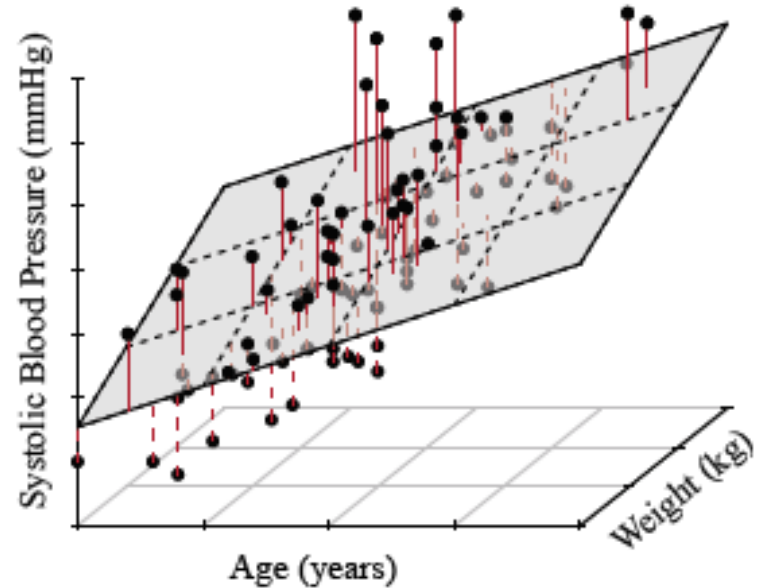
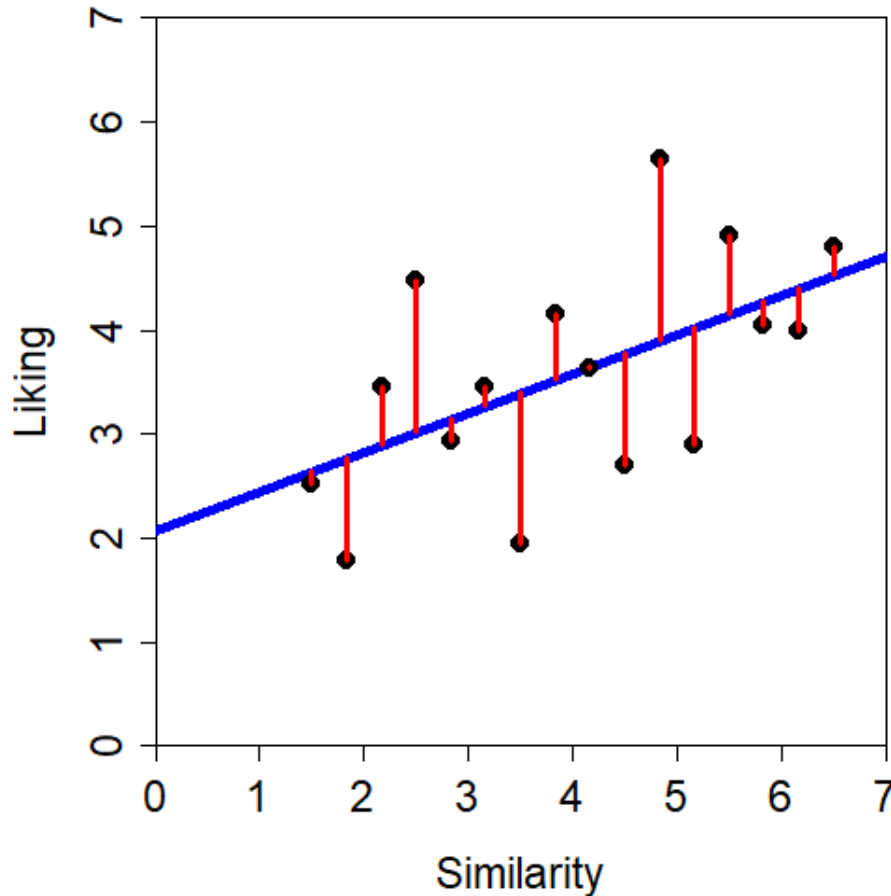


Figure 2.26: The residuals of figure 2.25 are the vertical distances to the plane. Negative residuals are indicated by dashed linepieces.

Image by Frans Rodenburg (frans-rodensburg)

<https://stackoverflow.com/questions/47344850/scatterplot3d-regression-plane-with-residuals>

One predictor



Two predictors

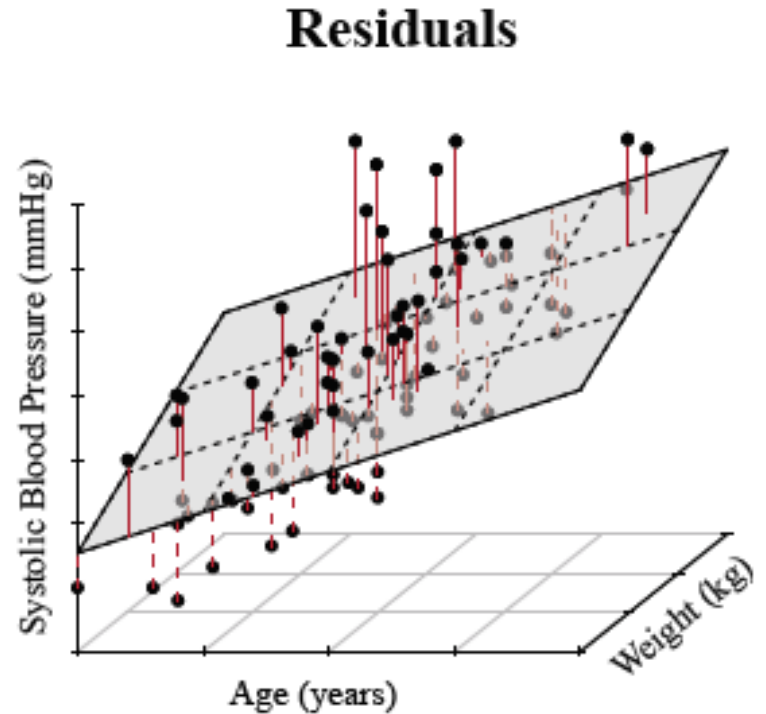


Figure 2.26: The residuals of figure 2.25 are the vertical distances to the plane. Negative residuals are indicated by dashed linepieces.

Image by Frans Rodenburg (frans-rodensburg)

<https://stackoverflow.com/questions/47344850/scatterplot3d-regression-plane-with-residuals>

Multiple Linear Regression

$$\text{Fitted value} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$\text{Observed value} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

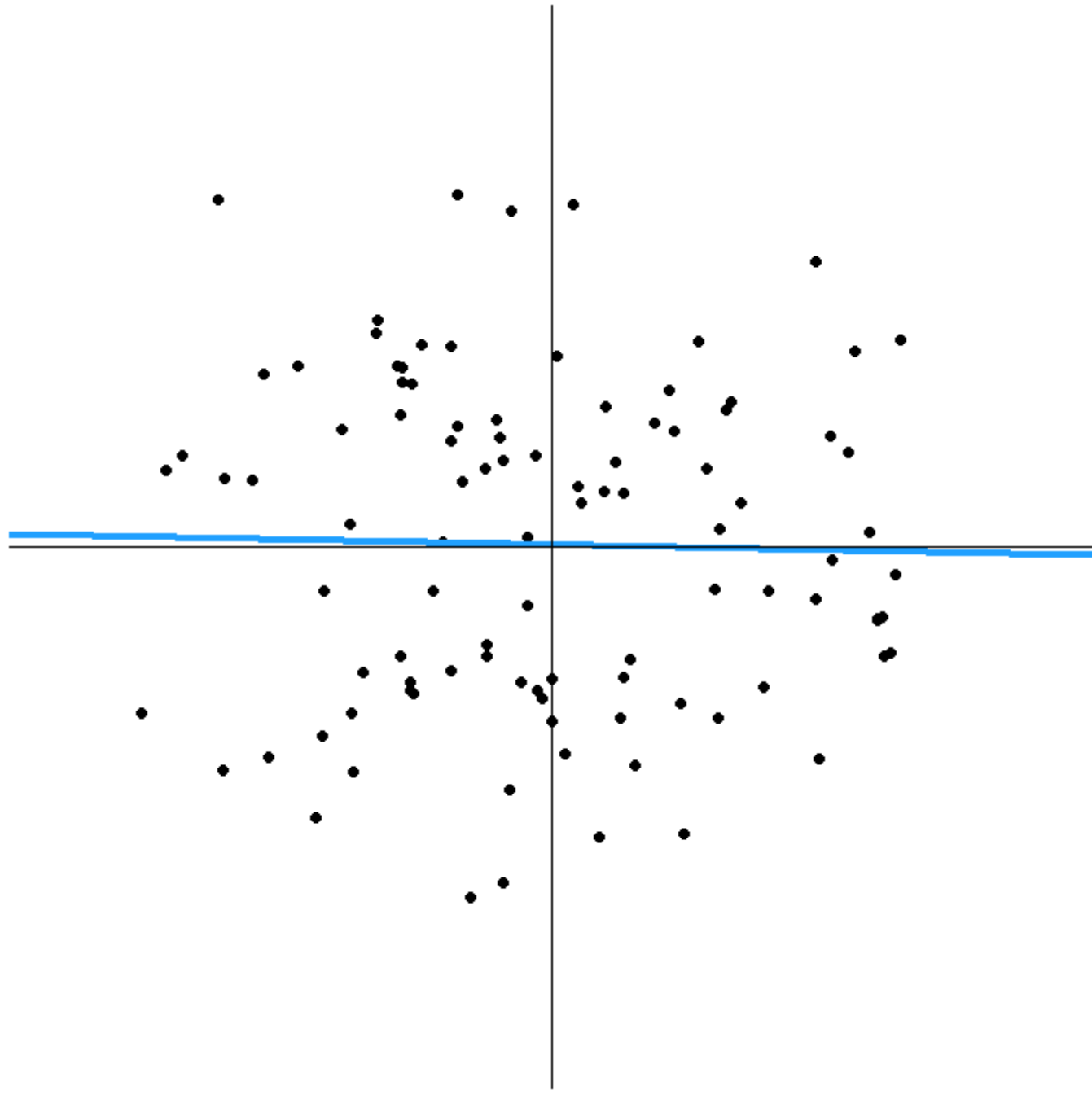
regression

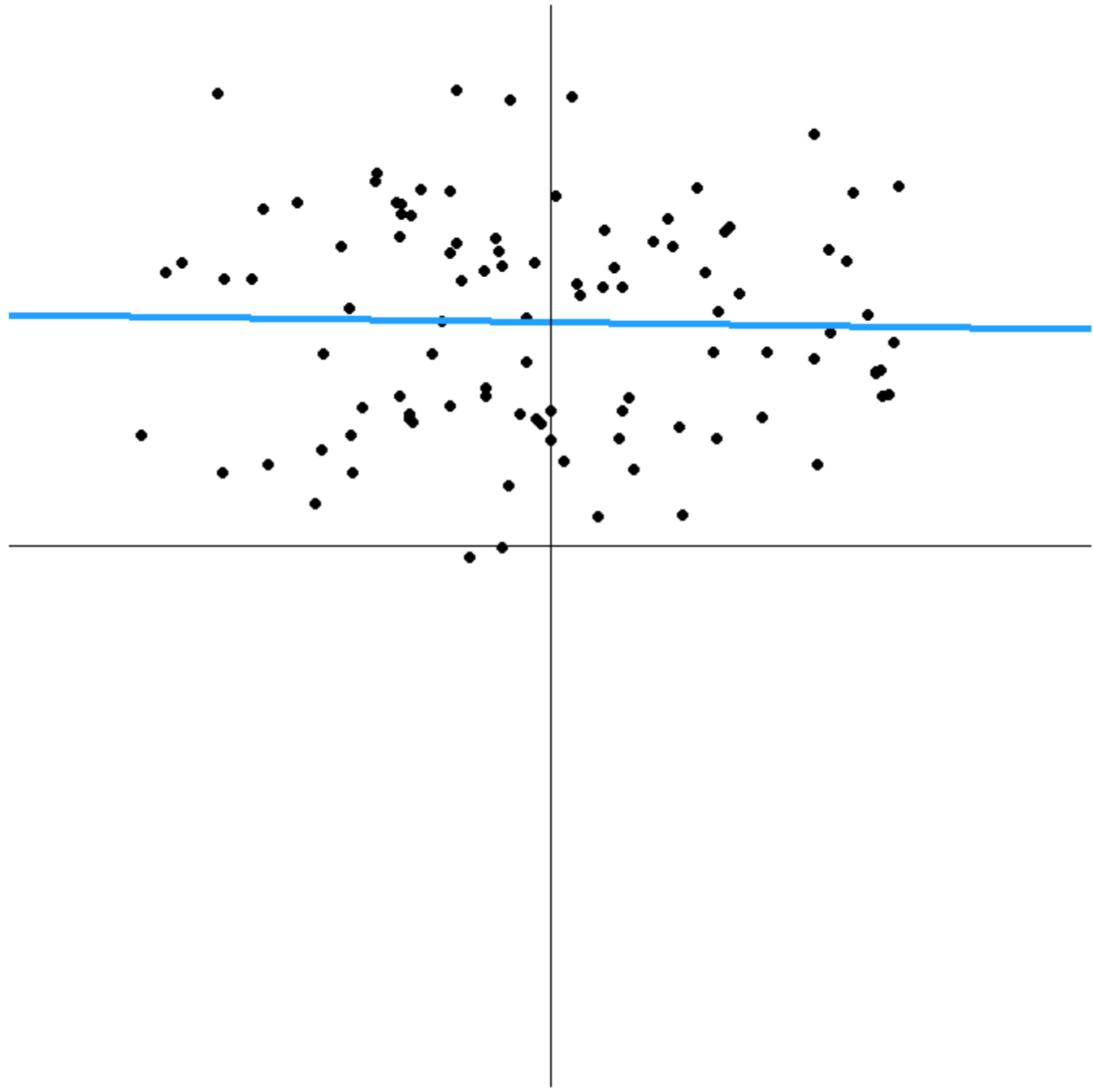
linear regression

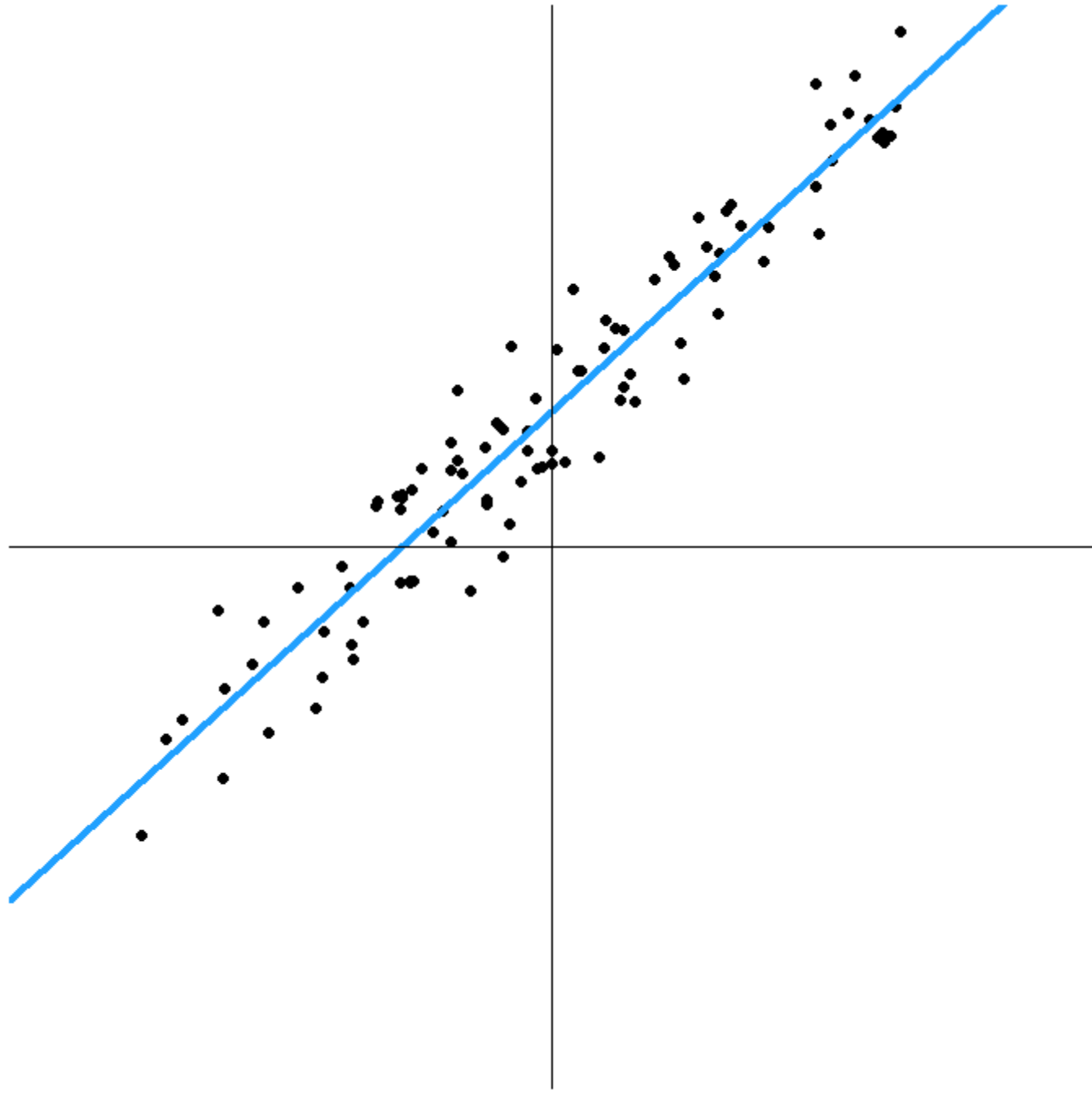
ordinary least squares (OLS) regression

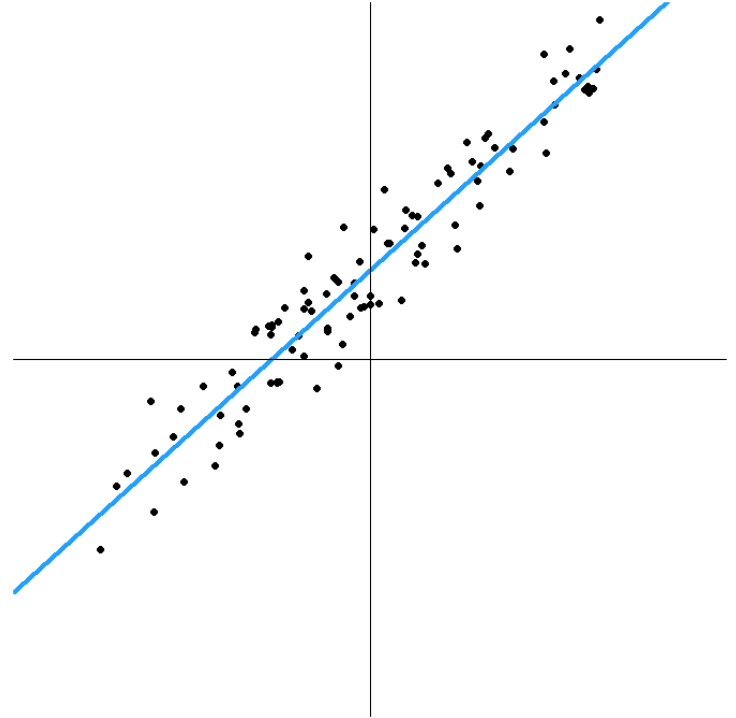
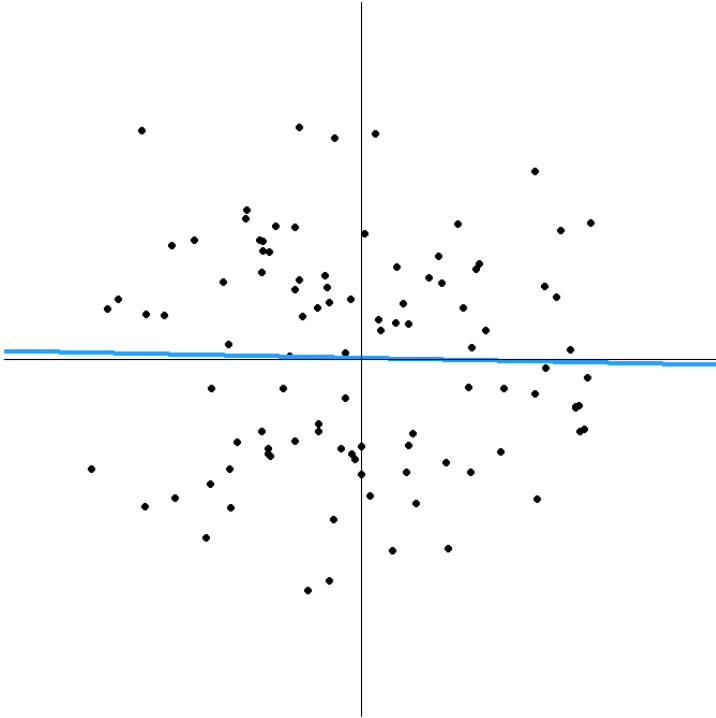
Part 2

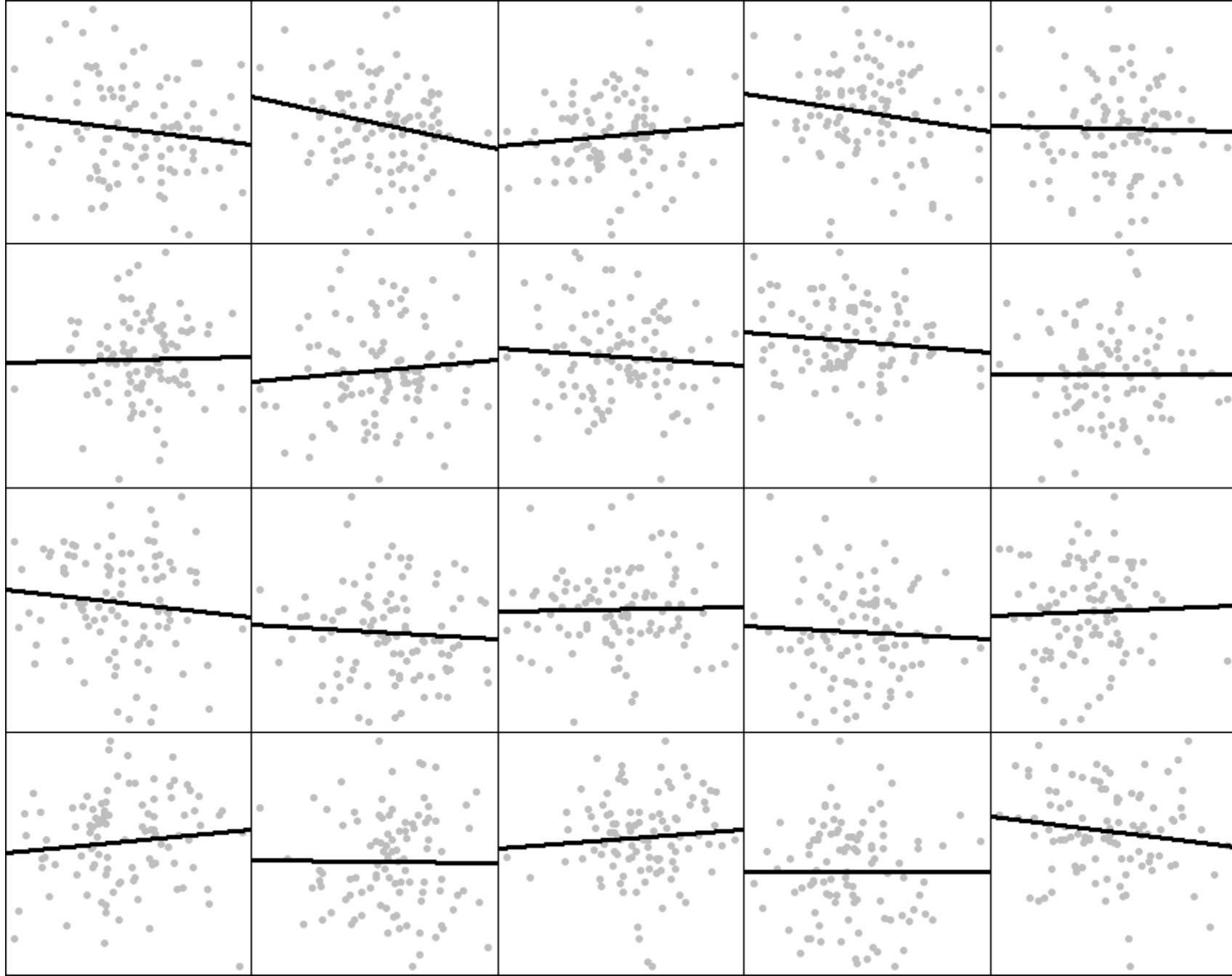
p -values for the
slope and intercept

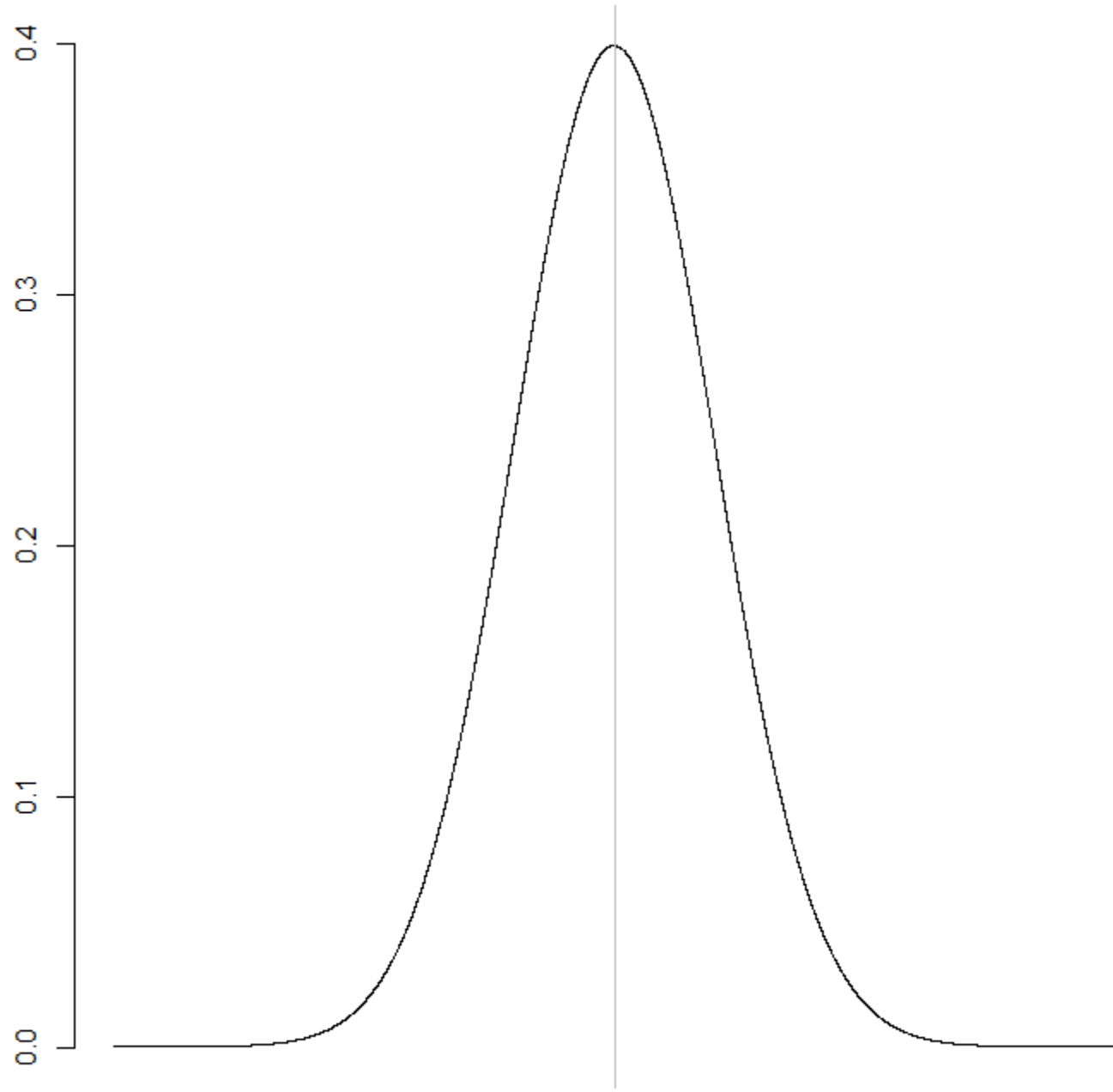


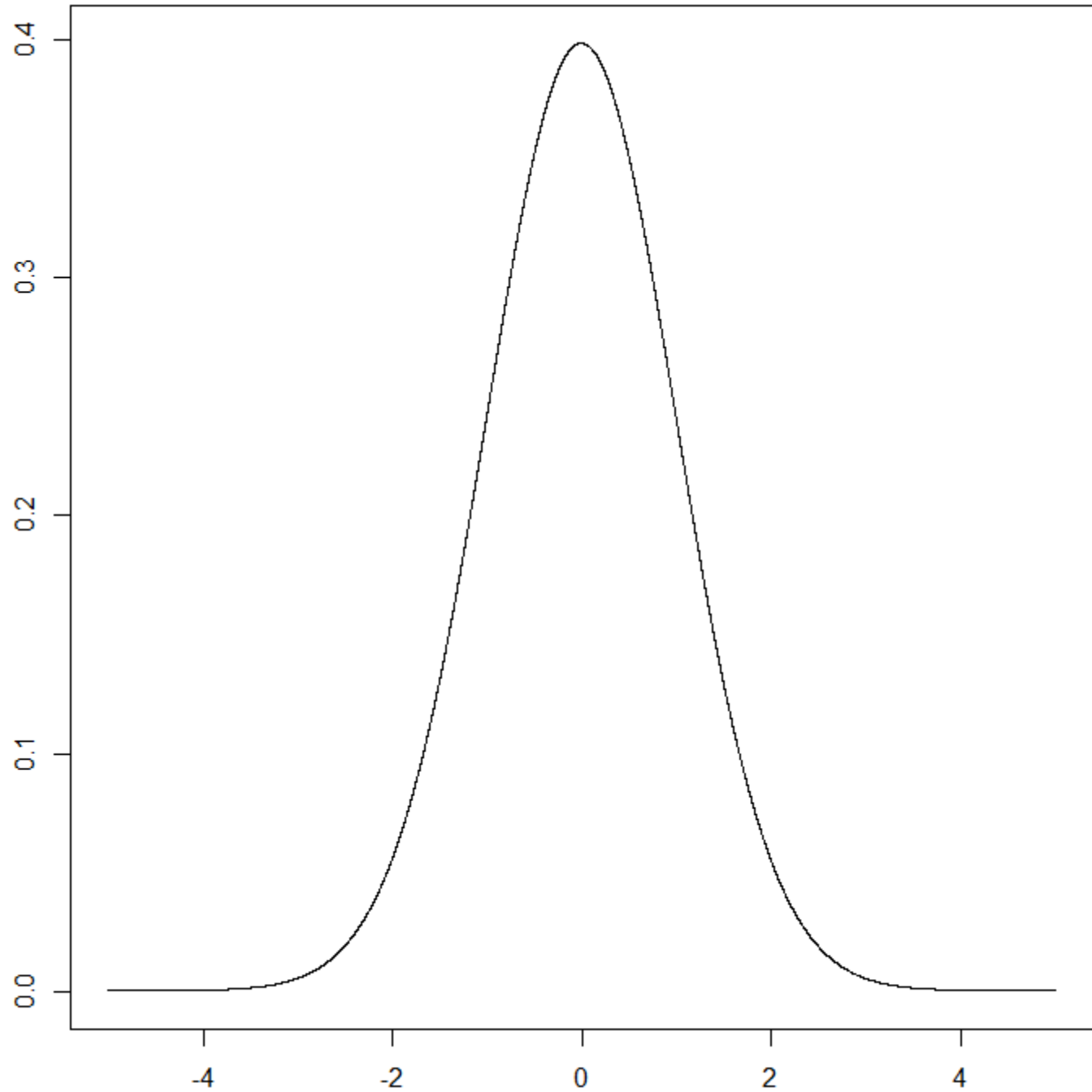


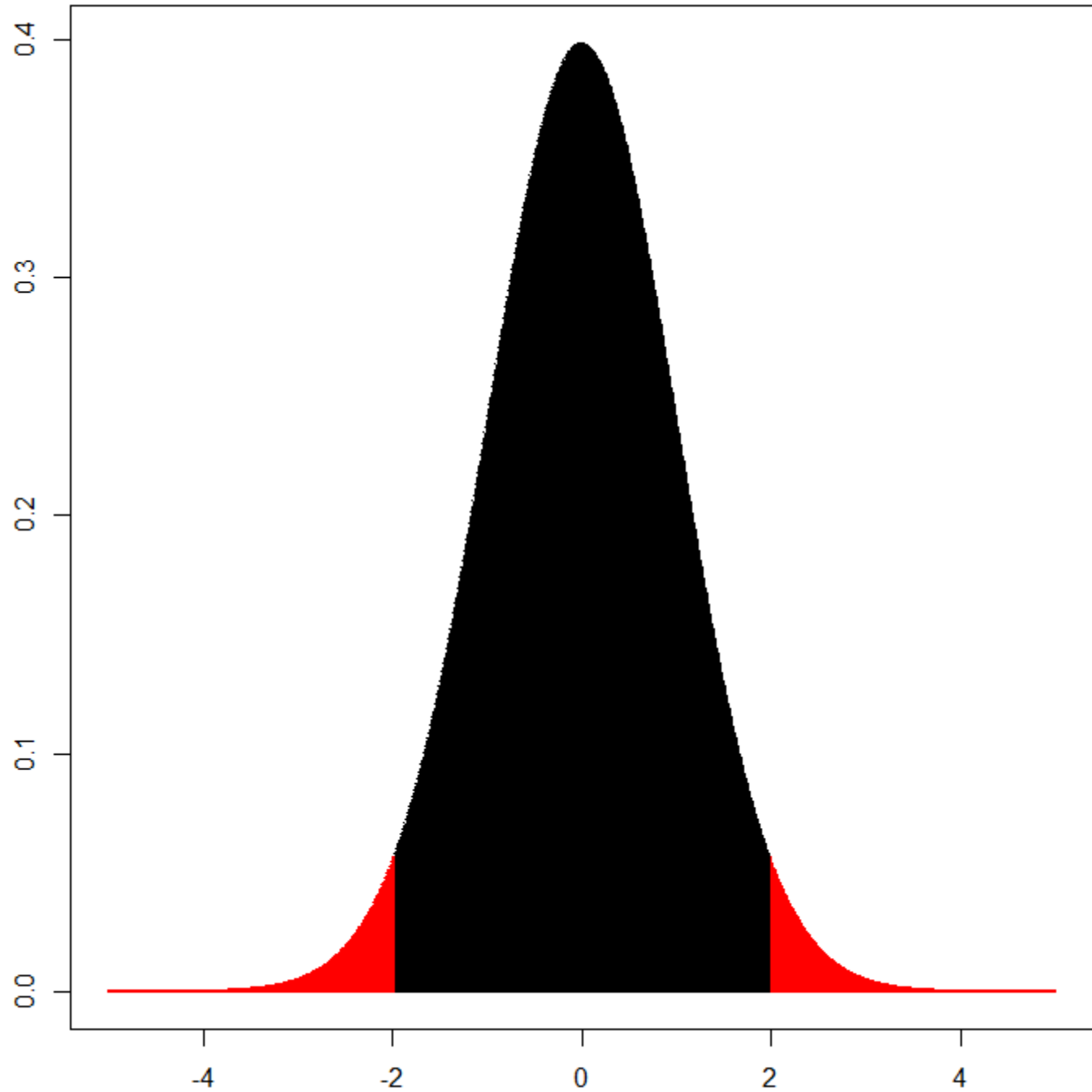


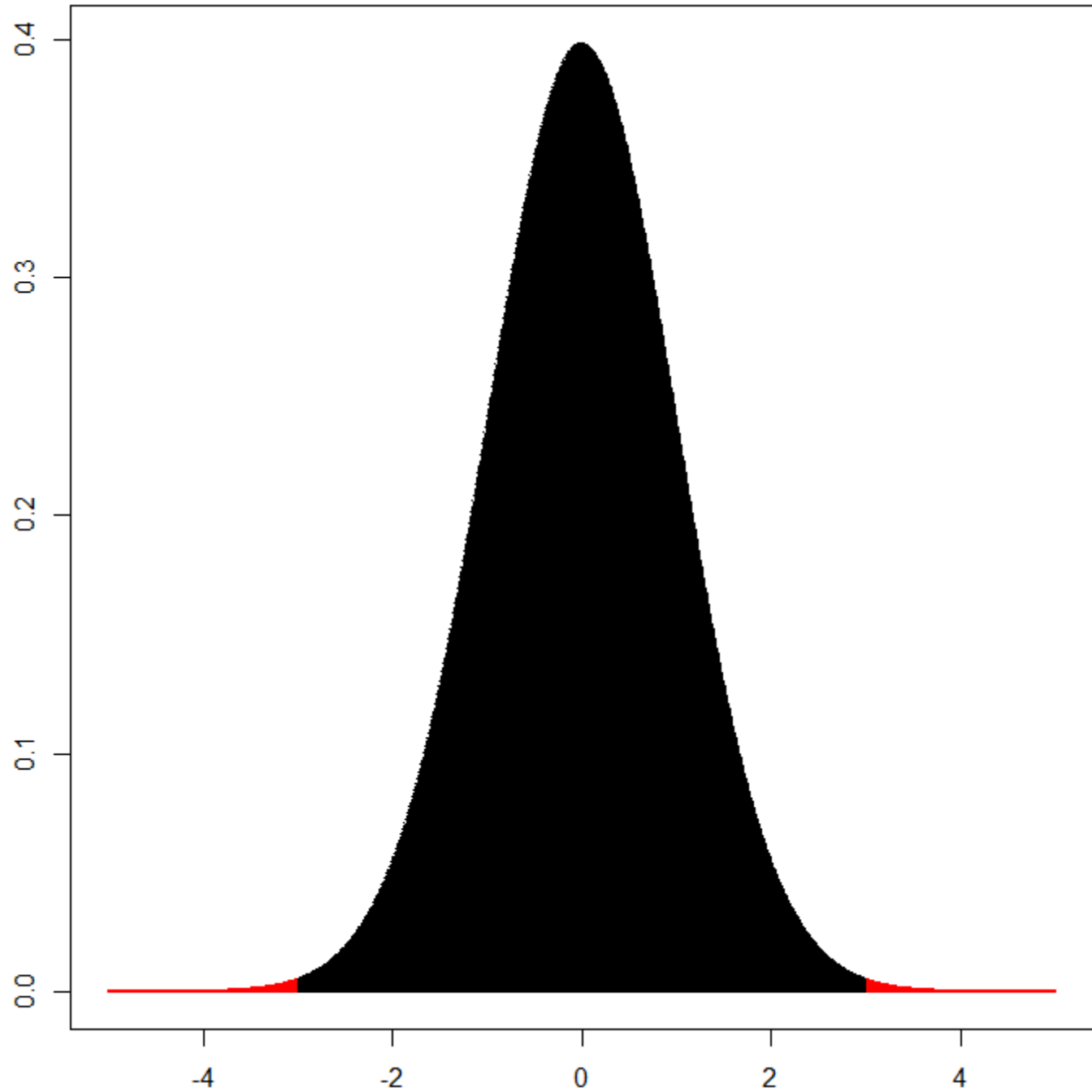


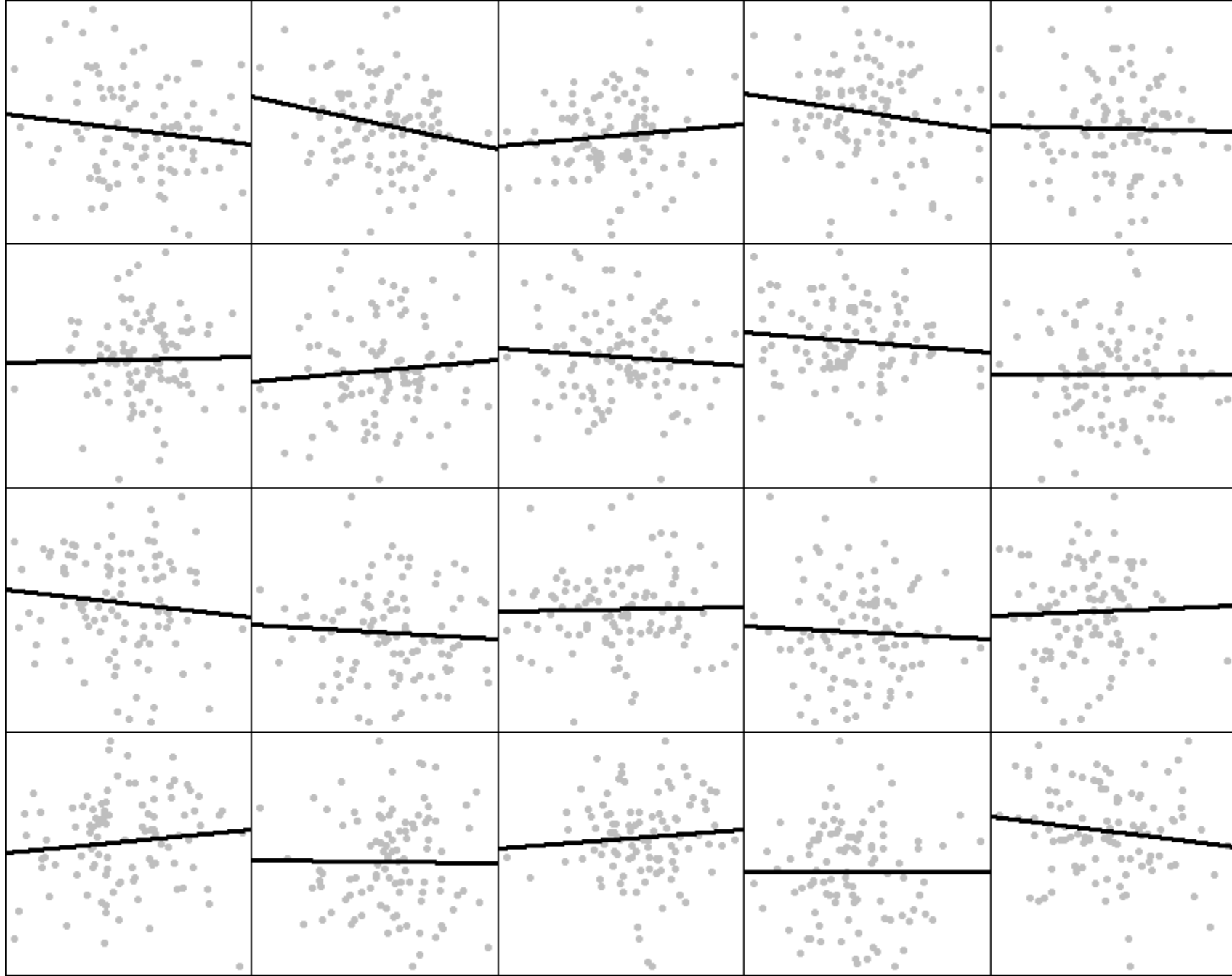










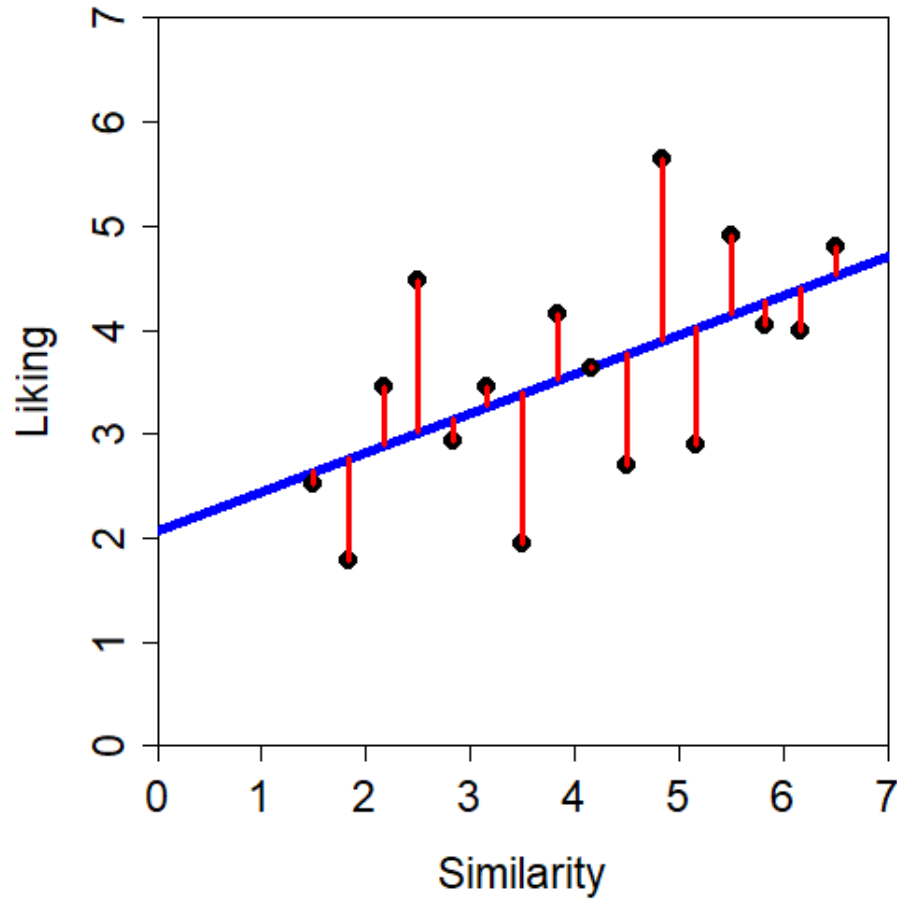


Part 3

R^2

(R-squared)

$$Y = b_0 + b_1X$$

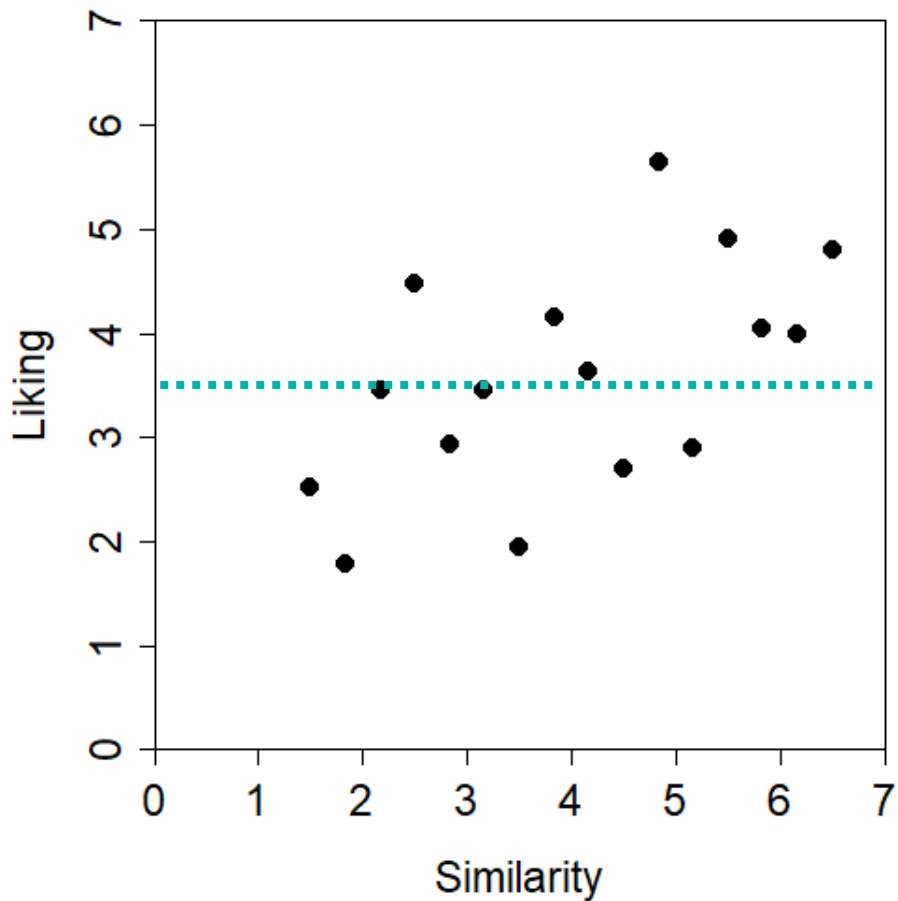


A quick detour: Mean and variance

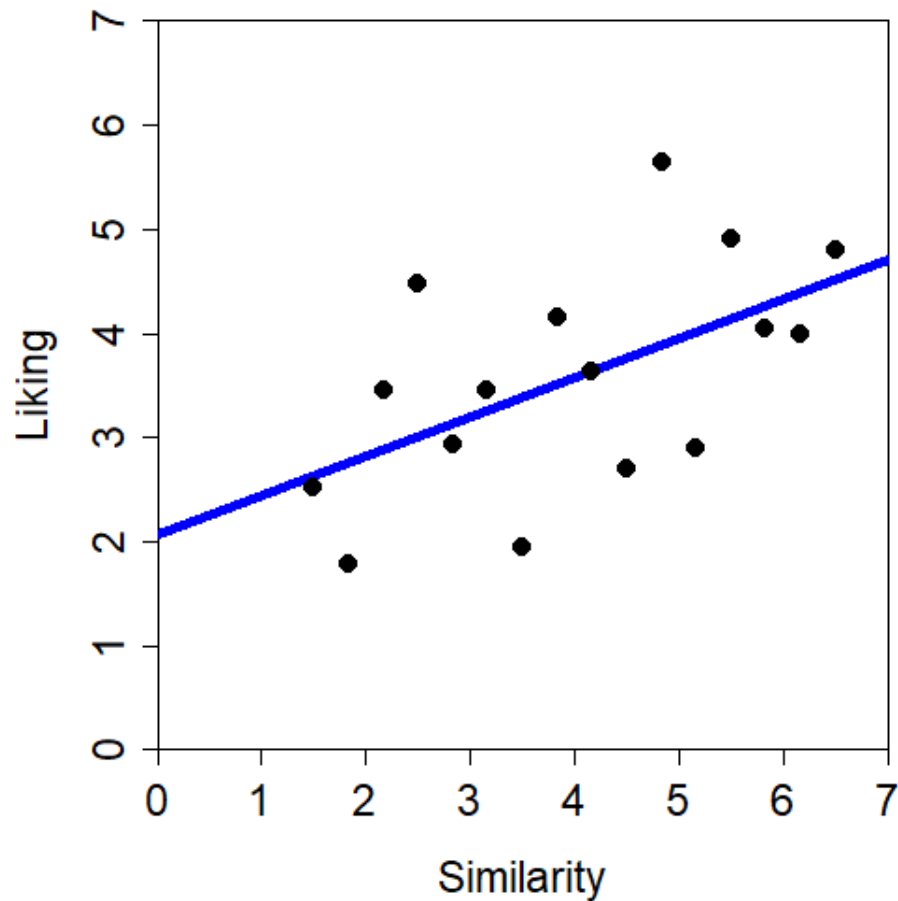
Mean $\bar{y} = \frac{\Sigma(Y)}{n}$

Variance $s^2 = \frac{\Sigma(Y - \bar{y})^2}{n-1}$

$$Y = \text{mean}$$



$$Y = b_0 + b_1X$$



$$\frac{\Sigma(\mathbf{e}^2)}{\Sigma(\mathbf{Y} - \bar{\mathbf{y}})^2}$$

R^2

$$R^2 = 1 - \frac{\Sigma(e^2)}{\Sigma(Y - \bar{y})^2}$$

R^2

$$R^2 = 1 - \frac{\Sigma(Y - \text{fitted value})^2}{\Sigma(Y - \text{mean})^2}$$

R^2

$R^2 =$ “coefficient of determination”

Adjusted R^2 is better for many kinds of model comparison.

Part 4

Example R output

Effect of practice on math test

Simple experiment:

0 to 120 minutes of practice time

Math test percentage score

Effect of practice on math test

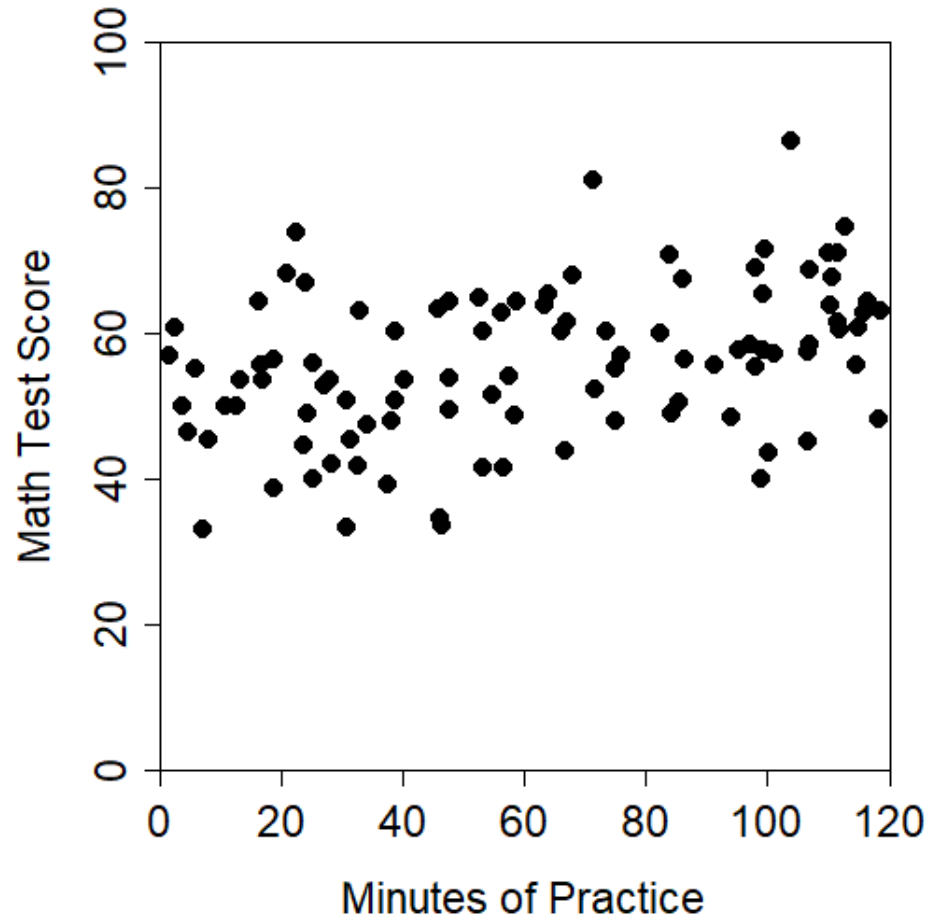
```
# Set random seed so example always comes out the same  
set.seed(5)
```

```
# Set graphics parameters so the plot elements are big  
par(mar=c(3.5,3.5,1,1),cex=1.4)
```

```
# Generate some random data  
practice = runif(100,0,120)  
mathscore = rnorm(100,50,10) + practice/10
```

```
# Plot math score by practice time  
plot(mathscore~practice,xlim=c(0,120),ylim=c(0,100),  
      mgp=c(2,.5,0),tcl=-.3,xaxs="i",yaxs="i",pch=16,  
      ylab="Math Test Score",xlab="Minutes of Practice")
```

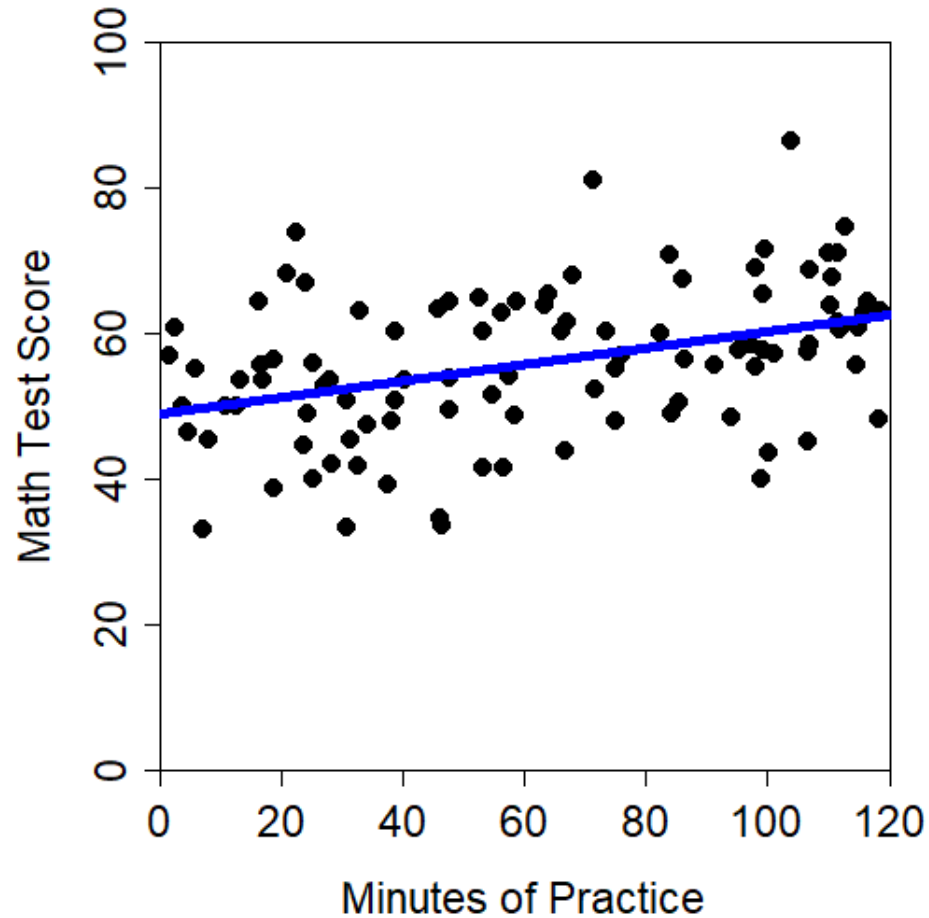
Effect of practice on math test



Effect of practice on math test

```
# Add regression line  
abline(lm(mathscore~practice), col="blue", lwd=5)
```

Effect of practice on math test



Effect of practice on math test

```
# Fit simple regression model and get summary  
summary(lm(mathscore~practice))
```

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~
```

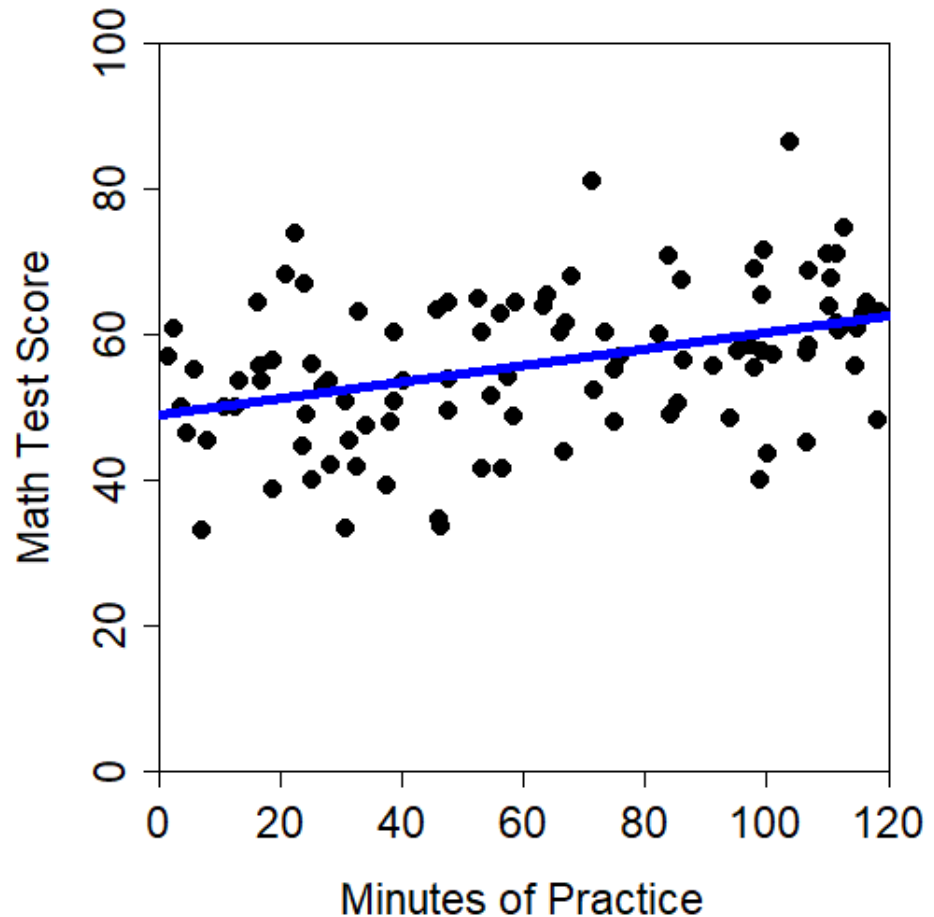
Residuals:

Min	1Q	Median
-20.6462	-5.3216	-0.261

Coefficients:

	Estimate	Std. Error
(Intercept)	49.03019	1
practice	0.11275	(

Signif. codes: 0 '***' (



Residual standard error:

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

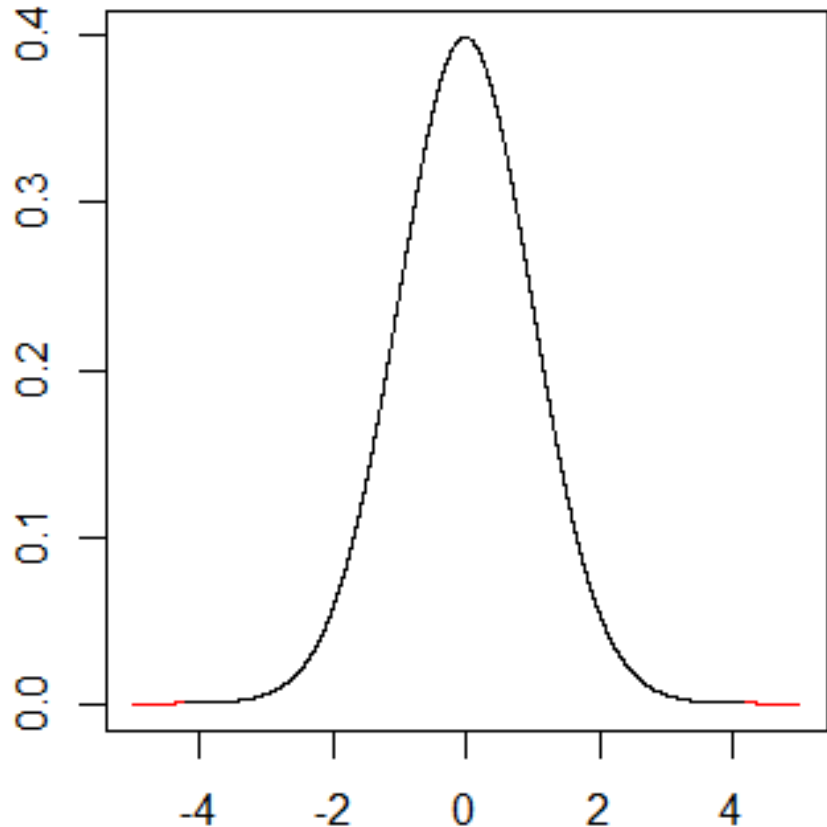
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05



model and get summary
actice))

practice)

n	3Q	Max
5	7.1163	25.6459

	Error	t value	Pr(> t)	
.93133	25.387	< 2e-16	***	
practice	0.11275	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom
Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439
F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

Effect of practice on math test

```
> # Confidence interval
> confint(lm(mathscore~practice))
                2.5 %      97.5 %
(Intercept) 45.19752774 52.8628509
practice      0.05948496  0.1660086
```

Part 5

Practice with linear
transformations

```
> # Fit simple regression model and get summary
> summary(lm(mathscore~practice))
```

Call:

```
lm(formula = mathscore ~ practice)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.03019	1.93133	25.387	< 2e-16	***
practice	0.11275	0.02684	4.201	5.87e-05	***

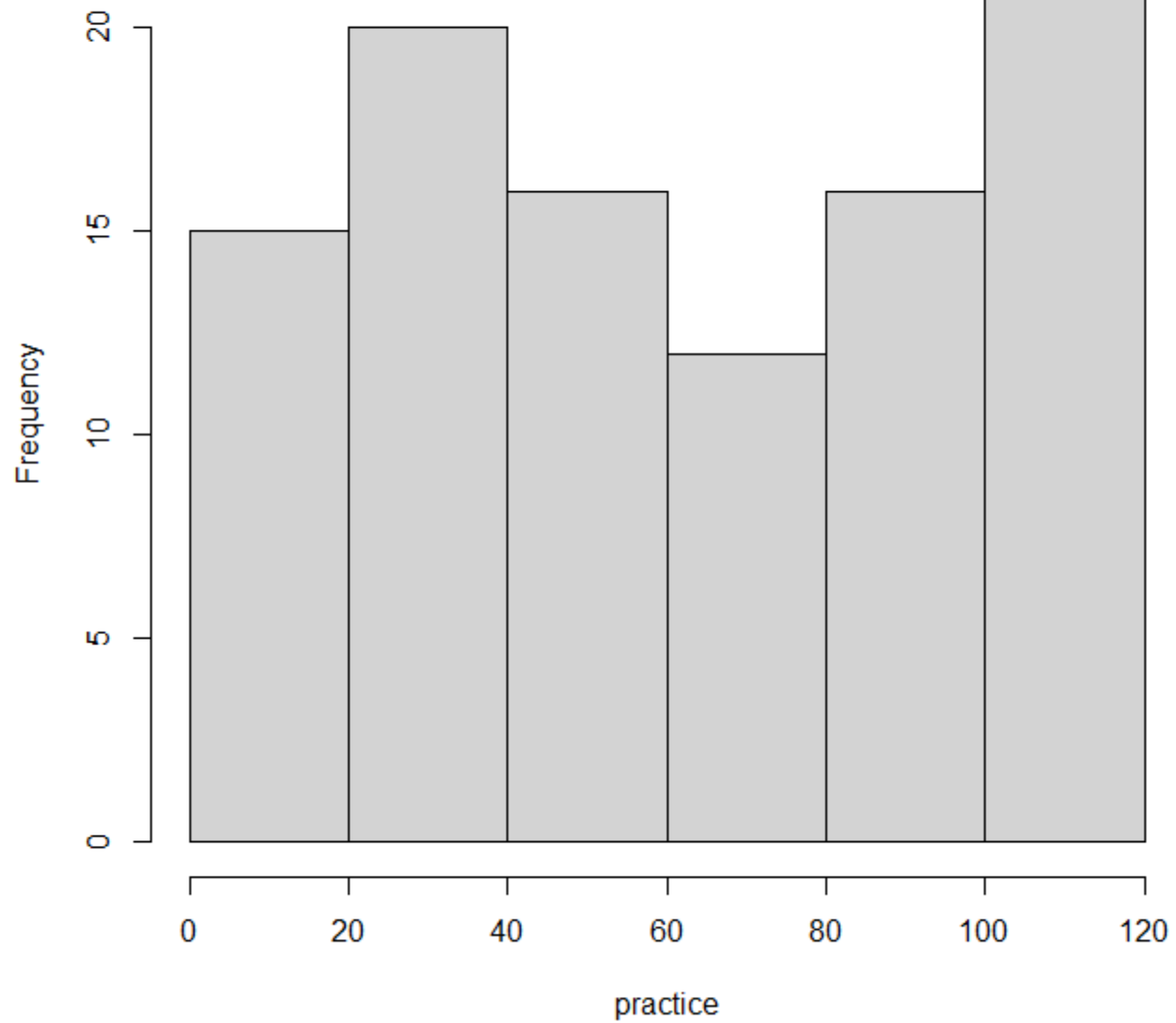
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.706 on 98 degrees of freedom

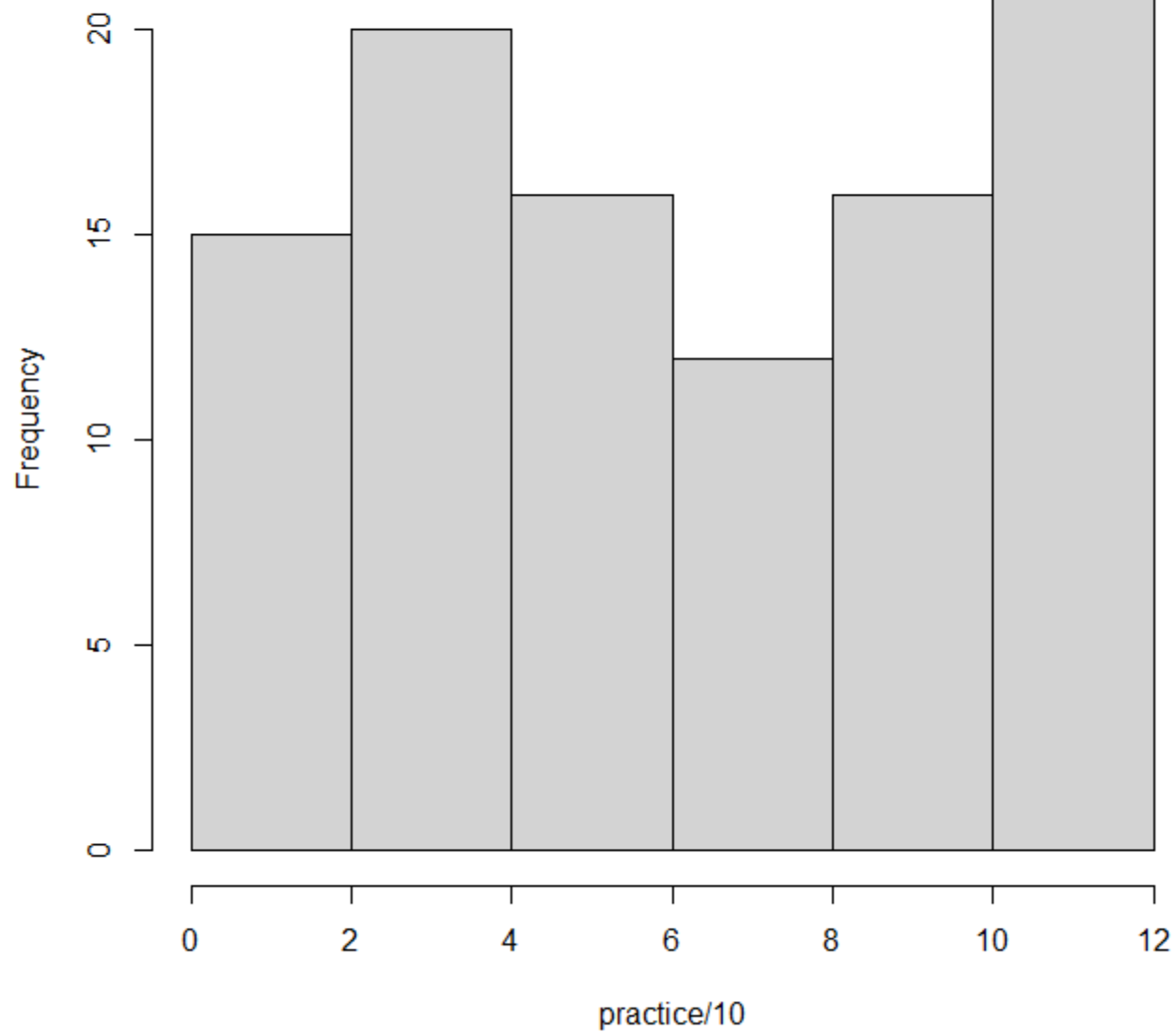
Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

Histogram of practice



Histogram of practice/10



```
> # Represent practice in ten-minute increments
> practice10 = practice/10
> summary(lm(mathscore~practice10))
```

Call:

```
lm(formula = mathscore ~ practice10)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.0302	1.9313	25.387	< 2e-16	***
practice10	1.1275	0.2684	4.201	5.87e-05	***

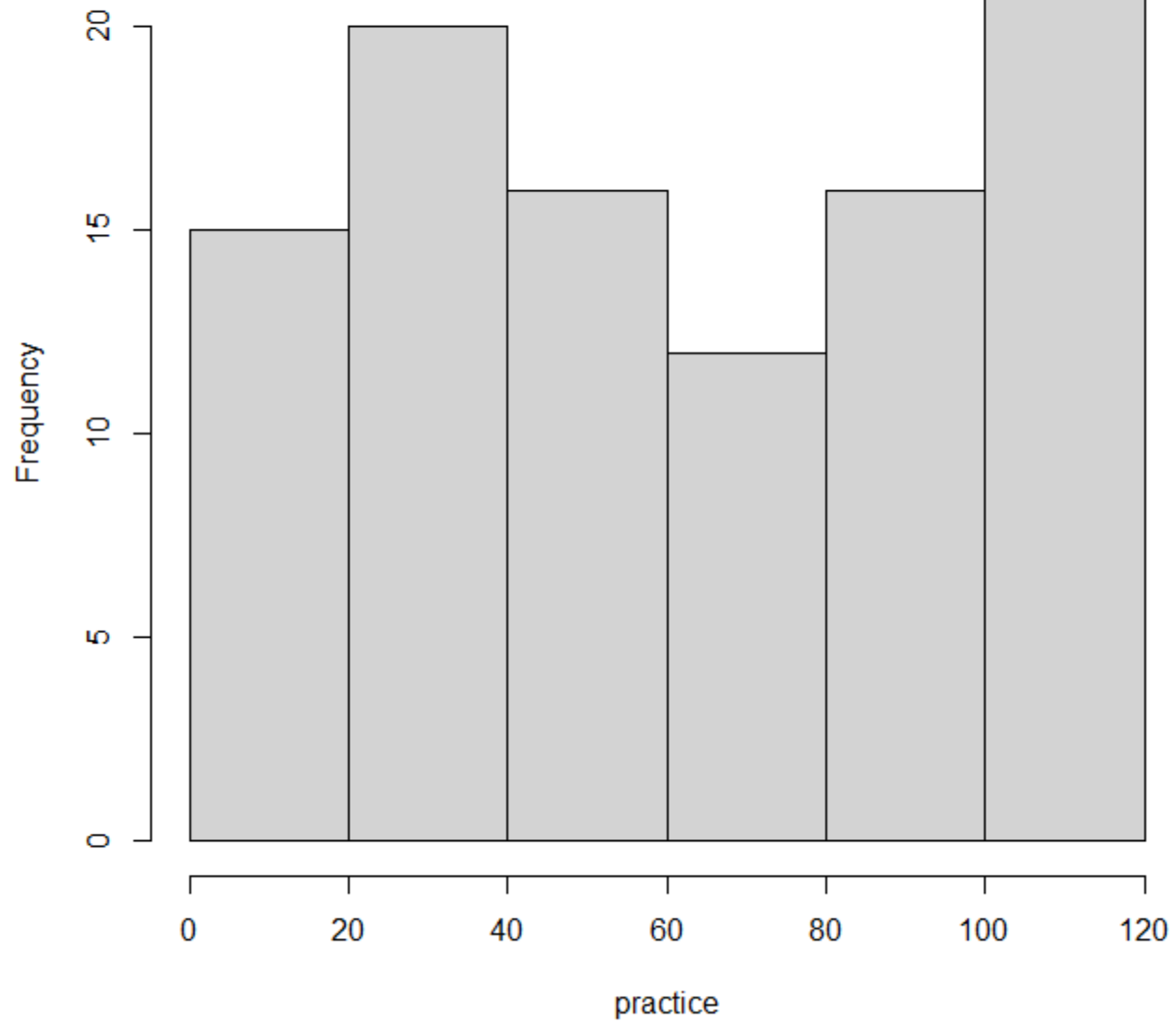
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.706 on 98 degrees of freedom

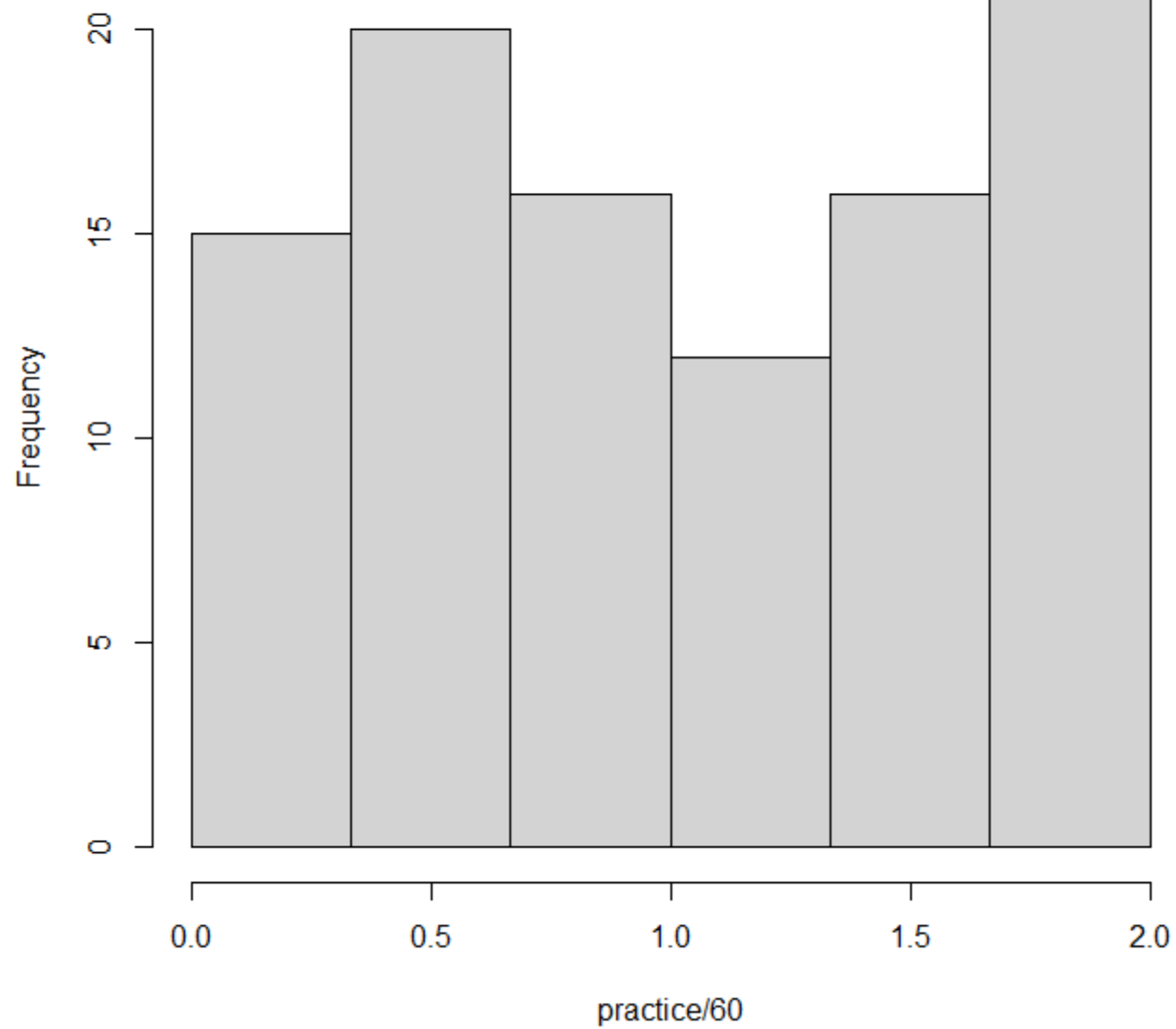
Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

Histogram of practice



Histogram of practice/60



```
> # Represent practice in one-hour increments
> practicehrs = practice/60
> summary(lm(mathscore~practicehrs))
```

Call:

```
lm(formula = mathscore ~ practicehrs)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.030	1.931	25.387	< 2e-16	***
practicehrs	6.765	1.610	4.201	5.87e-05	***

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

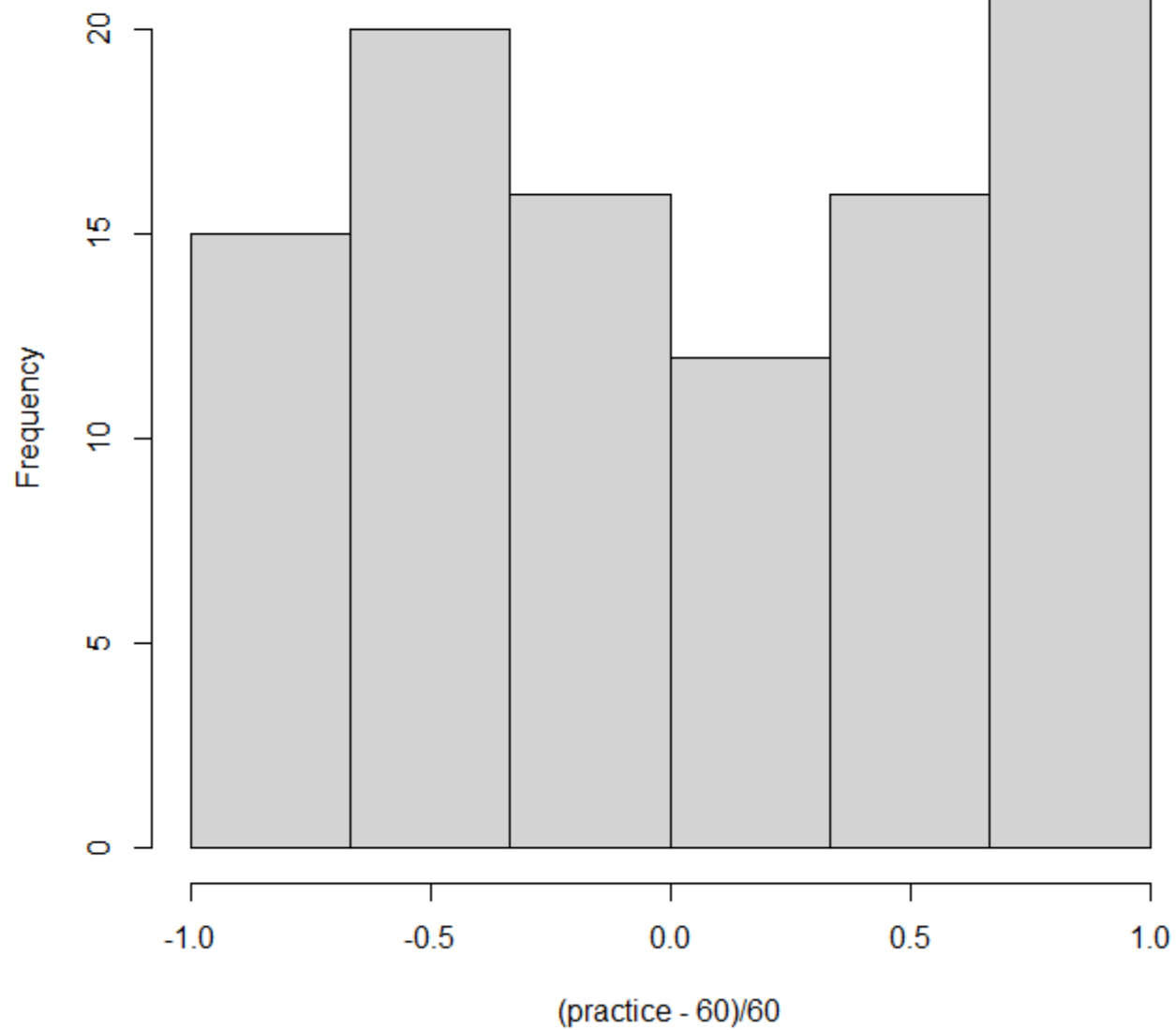
Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Set intercept at one hour  
> practiceover1hr = (practice-60)/60  
> summary(lm(mathscore~practiceover1hr))
```

Histogram of $(\text{practice} - 60)/60$



```
> # Set intercept at one hour
> practiceover1hr = (practice-60)/60
> summary(lm(mathscore~practiceover1hr))
```

Call:

```
lm(formula = mathscore ~ practiceover1hr)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	55.7950	0.9724	57.380	< 2e-16 ***
practiceover1hr	6.7648	1.6104	4.201	5.87e-05 ***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

A quick detour: Standardizing a variable

$$\text{Standardized } Y = \frac{Y - \text{mean}(Y)}{\text{sd}(Y)}$$

```
> # Z-score practice variable
> practiceZ = (practice - mean(practice)) / sd(practice)
> summary(lm(mathscore~practiceZ))
```

Call:

```
lm(formula = mathscore ~ practiceZ)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.6462	-5.3216	-0.2615	7.1163	25.6459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	56.0444	0.9706	57.744	< 2e-16	***
practiceZ	4.0977	0.9755	4.201	5.87e-05	***

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

Residual standard error: 9.706 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> # Z-score both variables
> mathscoreZ = (mathscore - mean(mathscore)) /
+             sd(mathscore)
> summary(lm(mathscoreZ~practiceZ))
```

Call:

```
lm(formula = mathscoreZ ~ practiceZ)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.96819	-0.50730	-0.02493	0.67839	2.44480

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.751e-16	9.252e-02	0.000	1
practiceZ	3.906e-01	9.299e-02	4.201	5.87e-05 ***

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

Residual standard error: 0.9252 on 98 degrees of freedom

Multiple R-squared: 0.1526, Adjusted R-squared: 0.1439

F-statistic: 17.65 on 1 and 98 DF, p-value: 5.868e-05

```
> cor.test(mathscore,practice)
```

```
        Pearson's product-moment correlation
```

```
data:  mathscore and practice
```

```
t = 4.2008, df = 98, p-value = 5.868e-05
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
 0.2103514 0.5452154
```

```
sample estimates:
```

```
      cor
```

```
0.3906299
```

$r = .39$

Pearson correlation (the usual kind of correlation)

The correlation is the slope between standardized versions of both variables.

$$**r = .39**$$

One standard deviation of additional practice time was associated with scores on the math test that were .39 standard deviations higher.

For more on this concept:

<https://saraemilyburke.com/stats/correlationZscores.html>

Part 6

Another way of thinking
about correlation and R^2

Sum of products

X	Y
-2	-1
-1	-2
0	2
1	0
2	1

Sum of products

X	Y	X*Y
-2	-1	2
-1	-2	2
0	2	0
1	0	0
2	1	2

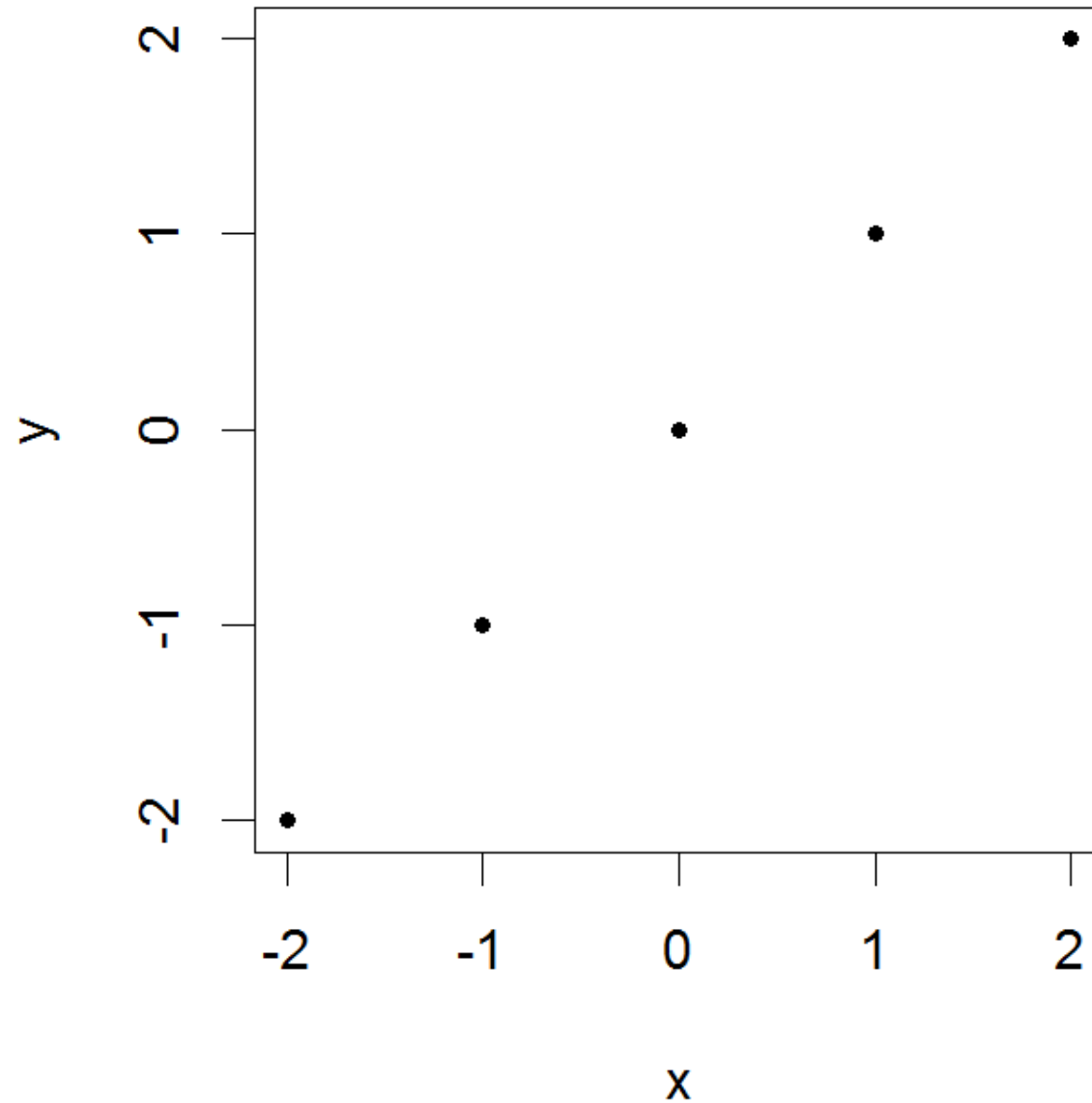
$$\underline{\text{Sum}(X*Y) =}$$

6

Sum of products

X	Y	X*Y
-1	-2	
2	1	
0	2	
1	0	
-2	-1	

Perfect correlation



Perfect correlation

X	Y	X*Y
-2	-2	4
-1	-1	1
0	0	0
1	1	1
2	2	4

Sum(X*Y)

10

Perfect correlation

X	Y	X*Y
-2	-2	4
-1	-1	1
0	0	0
1	1	1
2	2	4

Sum / n-1

2.5

Covariance

$$\frac{\Sigma(X - \bar{x})(Y - \bar{y})}{n-1}$$

Variance

X	X	$(X-\bar{x}) * (X-\bar{x})$
-2	-2	4
-1	-1	1
0	0	0
1	1	1
2	2	4

Sum / n-1

2.5

Covariance

$$\text{Covariance} \quad \frac{\underline{\underline{\Sigma(X - \bar{x})(Y - \bar{y})}}}{\mathbf{n-1}}$$

$$\text{Variance} \quad \frac{\underline{\underline{\Sigma(X - \bar{x})^2}}}{\mathbf{n-1}}$$

$$\frac{\underline{\underline{\Sigma(X - \bar{x})(X - \bar{x})}}}{\mathbf{n-1}}$$

Correlation

Covariance of X and Y

$$\sqrt{\text{Var}(X) * \text{Var}(Y)}$$

R^2

$R^2 = r^2 =$ “coefficient of determination”

R^2

$R^2 = r^2 =$

proportion of variance explained

R^2

$(\text{Covariance of X and Y})^2$

$\text{Var}(X) * \text{Var}(Y)$

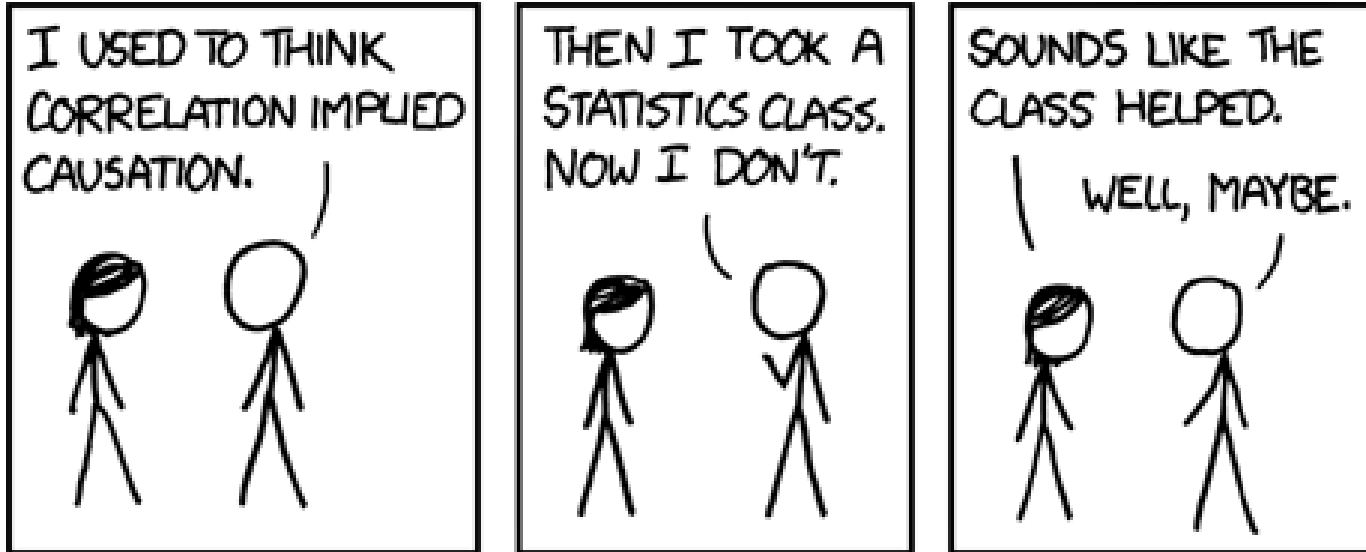
R^2 $\text{Cov}(X, Y) * \text{Cov}(X, Y)$

 $\text{Var}(X) * \text{Var}(Y)$

Part 7

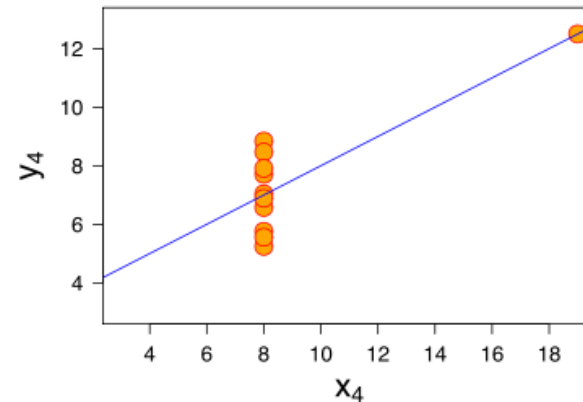
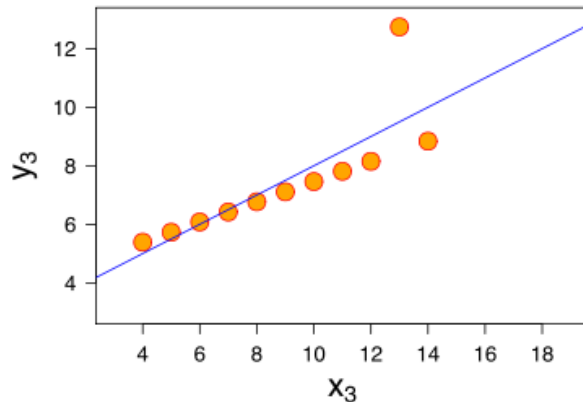
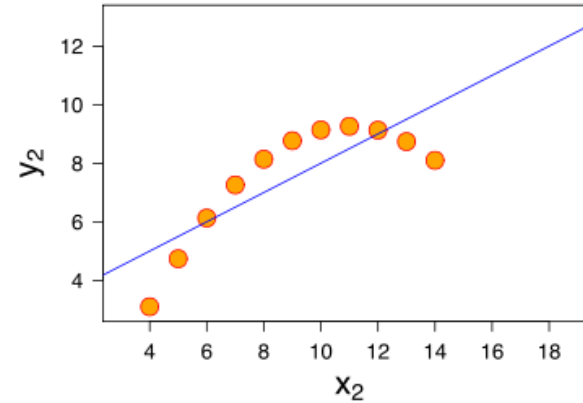
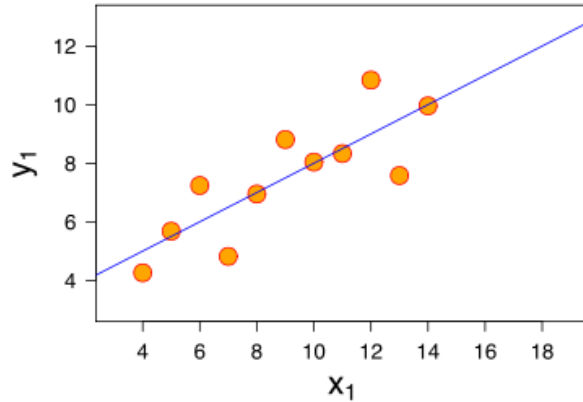
Caveats

Causation?



Copied for noncommercial purposes from xkcd.com/552 under a Creative Commons Attribution-NonCommercial 2.5 License

Linearity?



Anscombe's quartet

Anscombe, Francis J. (1973) Graphs in statistical analysis. *American Statistician*, 27, 17-21.
Image available at en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg
under the GNU General Public License

Part 8

Example with a binary
predictor

Effect of coffee on math test

Simple experiment:

Coffee or decaf

Math test percentage score

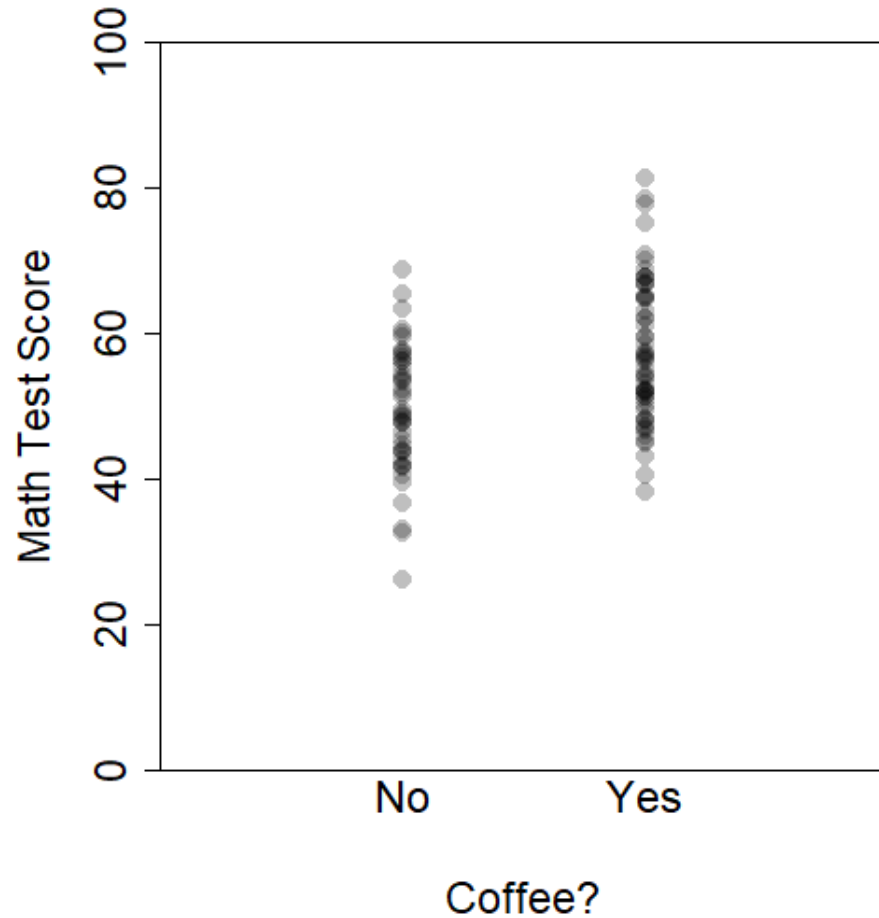
Effect of coffee on math test

```
# Next example -- coffee and math test

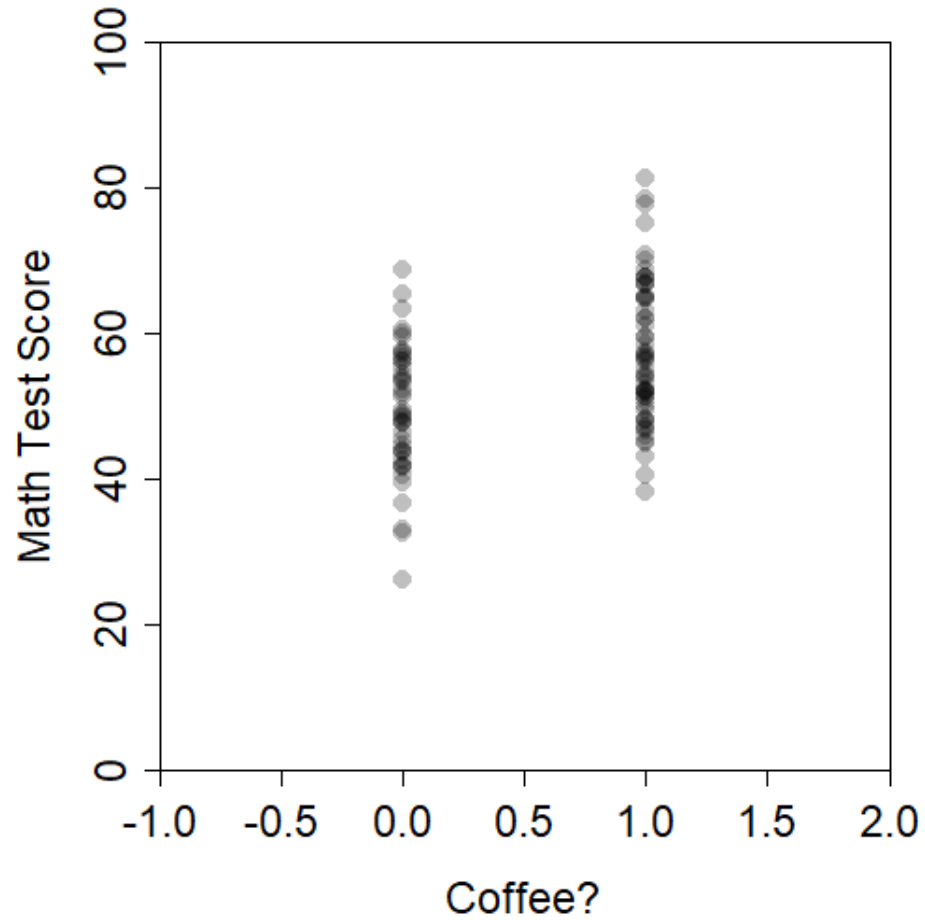
# Generate random data
set.seed(4)
coffee = rbinom(100,1,.5)
mathscore = rnorm(100,50,10) + coffee*8

# Plot math score by coffee
plot(mathscore~coffee,xlim=c(-1,2),ylim=c(0,100),
     mgp=c(2,.5,0),tcl=-.3,xaxs="i",yaxs="i",
     pch=16,col="#00000040",xaxt="none",
     ylab="Math Test Score",xlab="Coffee?")
mtext("No",1,0,at=0,cex=1.5)
mtext("Yes",1,0,at=1,cex=1.5)
```

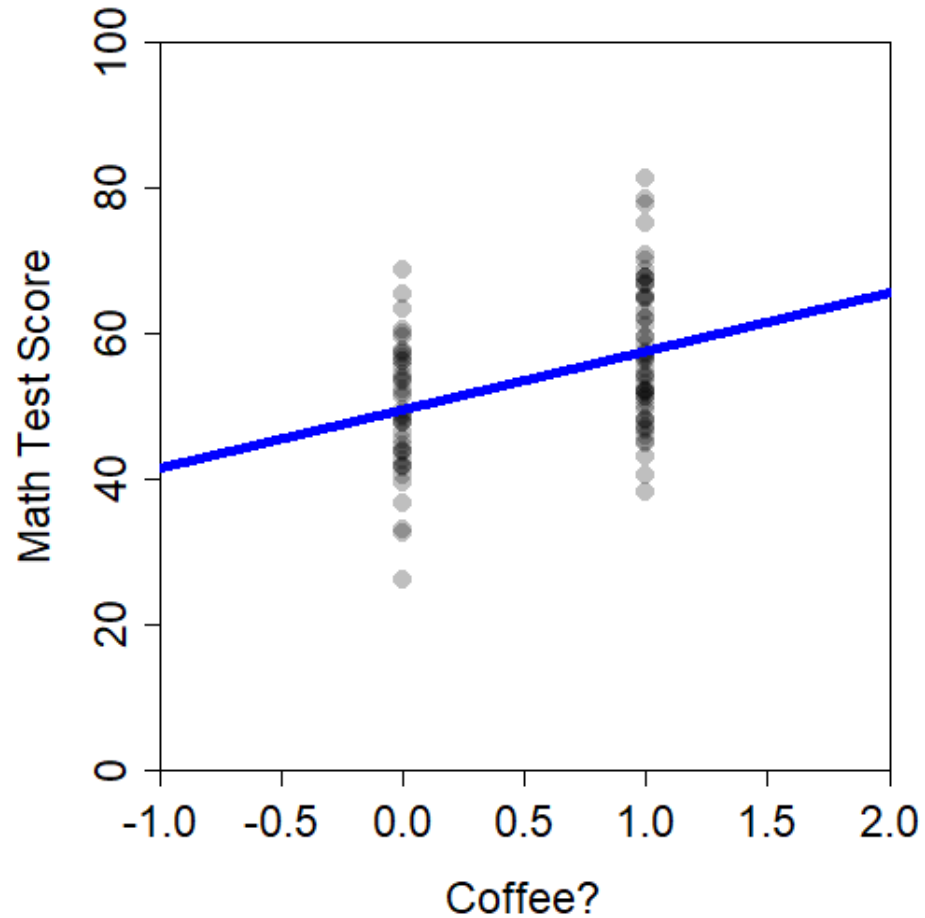
Effect of coffee on math test



Effect of coffee on math test



Effect of coffee on math test



```
> # Regression model
> summary(lm(mathscore~coffee))
```

Call:

```
lm(formula = mathscore ~ coffee)
```

Residuals:

Min	1Q	Median	3Q	Max
-23.4160	-6.3891	-0.7449	7.2595	23.6984

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	49.594	1.448	34.248	< 2e-16 ***
coffee	8.011	1.918	4.177	6.42e-05 ***

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

Residual standard error: 9.496 on 98 degrees of freedom

Multiple R-squared: 0.1511, Adjusted R-squared: 0.1425

F-statistic: 17.45 on 1 and 98 DF, p-value: 6.418e-05

```
> # Regression model
> summary(lm(mathscore~coffee))
```

Call:

```
lm(formula = mathscore ~ coffee)
```

Residuals:

Min	1Q	Median	3Q	Max
-23.4160	-6.3891	-0.7449	7.2595	23.6984

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.594	1.448	34.248	< 2e-16	***
coffee	8.011	1.918	4.177	6.42e-05	***

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

Residual standard error: 9.496 on 98 degrees of freedom

Multiple R-squared: 0.1511, Adjusted R-squared: 0.1425

F-statistic: 17.45 on 1 and 98 DF, p-value: 6.418e-05

```
> # Student's t-test
> t.test(mathscore~coffee,var.equal=T)
```

Two Sample t-test

```
data:  mathscore by coffee
```

```
t = -4.1768, df = 98, p-value = 6.418e-05
```

```
alternative hypothesis: true difference in means is not equal
```

```
95 percent confidence interval:
```

```
-11.817560  -4.205013
```

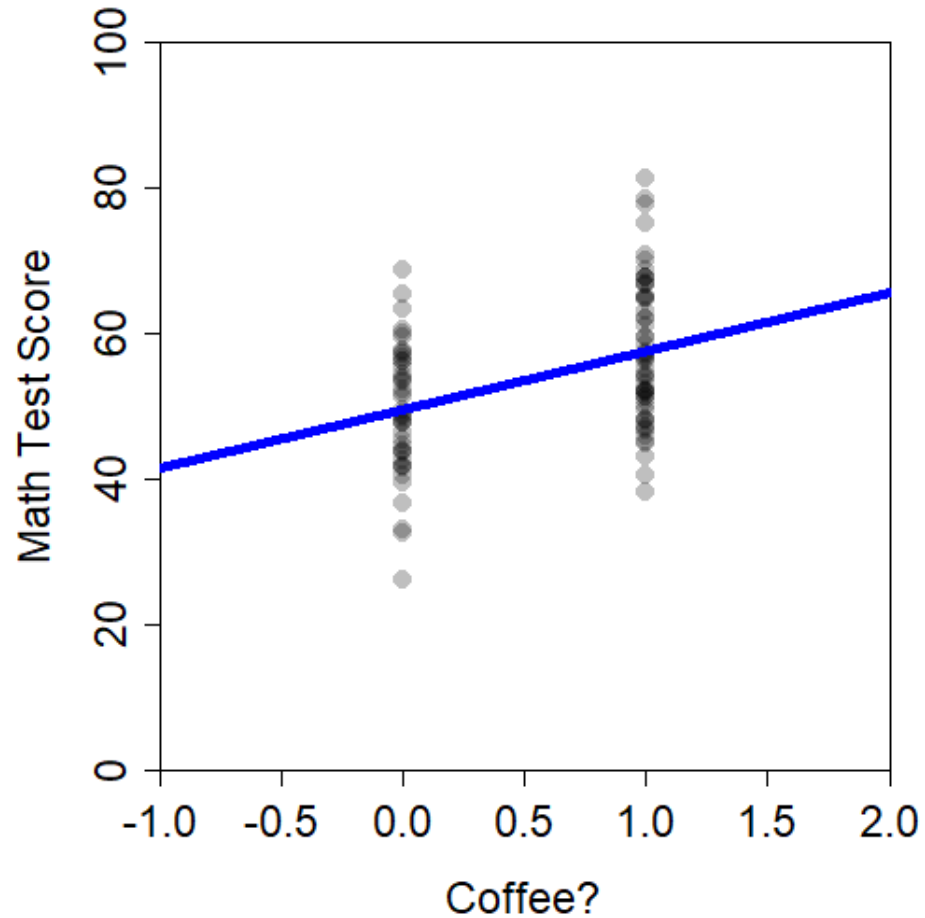
```
sample estimates:
```

```
mean in group 0 mean in group 1
```

```
49.59353          57.60481
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	49.594	1.448	34.248	< 2e-16	***
coffee	8.011	1.918	4.177	6.42e-05	***

Effect of coffee on math test



Conceptual Introduction to Linear Regression

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Use the `lm()` function in R

Focus on the slopes (**b**)

Rescaling variables preserves
their relationships

Only describes straight lines