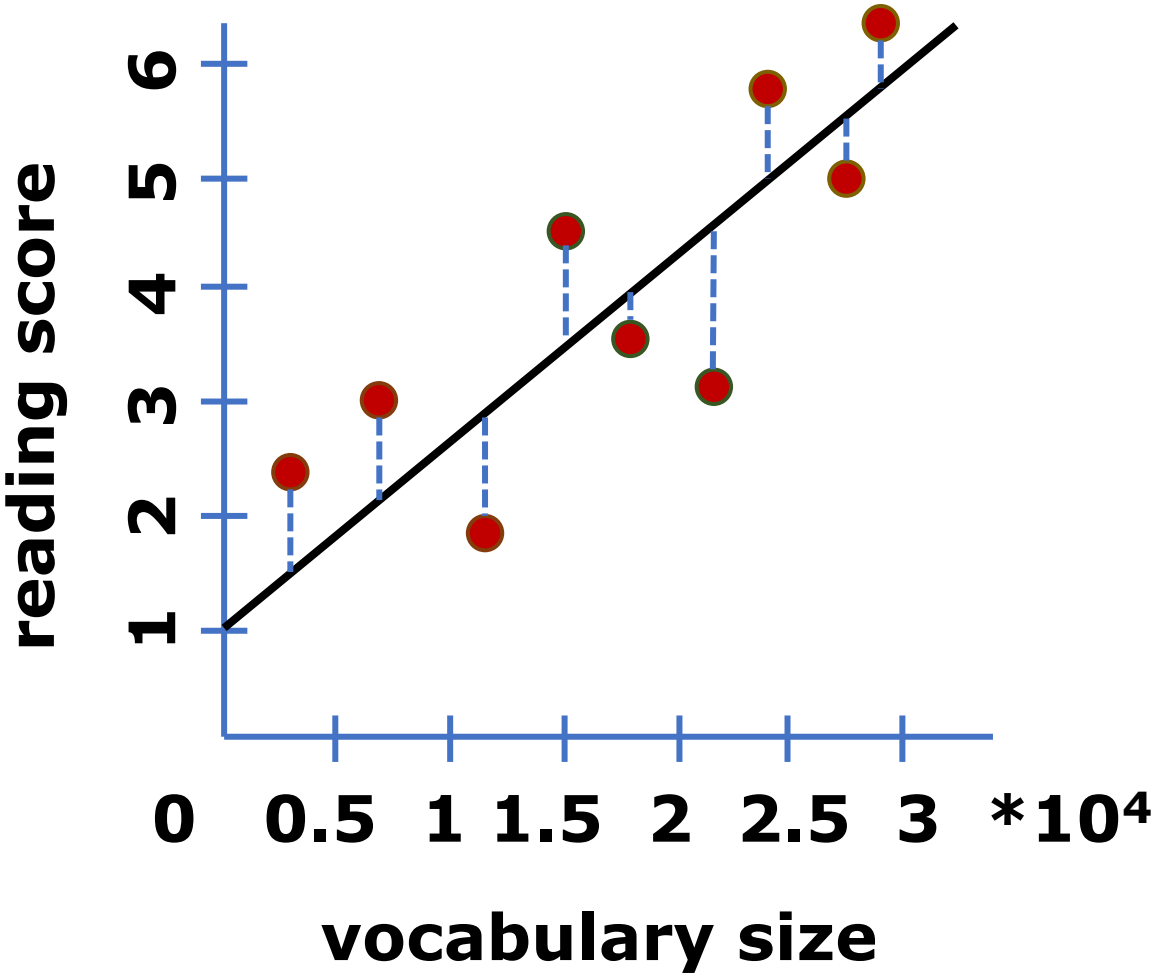
Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

Fundamentals of Statistics for Language Sciences LT2206

Jixing Li

Lecture 10: Multiple Regression

Simple linear regression



$y = 2x + 1$

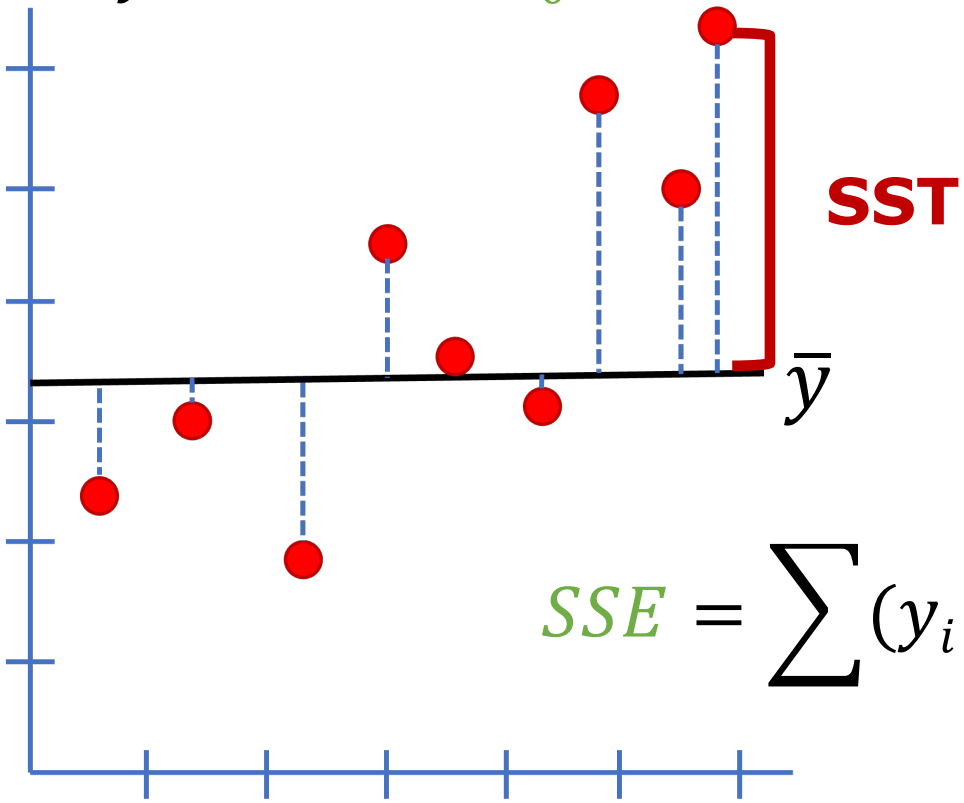
slope intercept

slope: vocabulary size increase by 10,000, reading score increases by 2 points on average.

intercept: the expected y when x=0, may or may not make sense

R^2

$y = 0 * x + b_0$

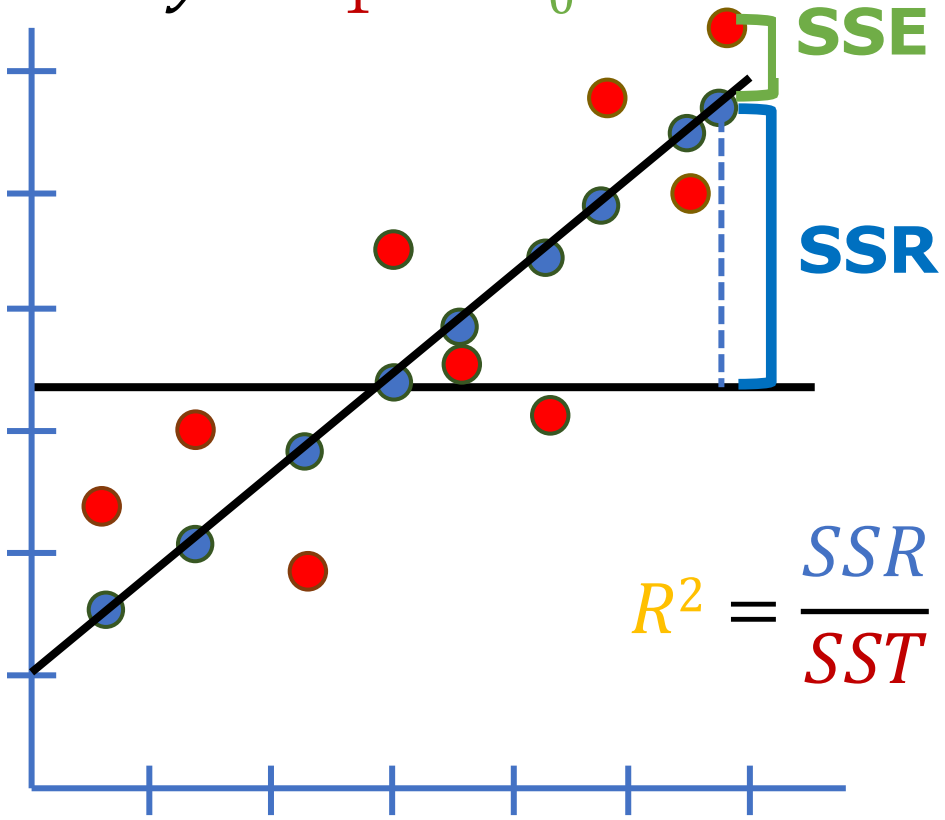


$$SSE = \sum (y_i - \hat{y})^2$$

x has no effect on y

$$SST = \sum (y_i - \bar{y})^2$$

$y = b_1 x + b_0$

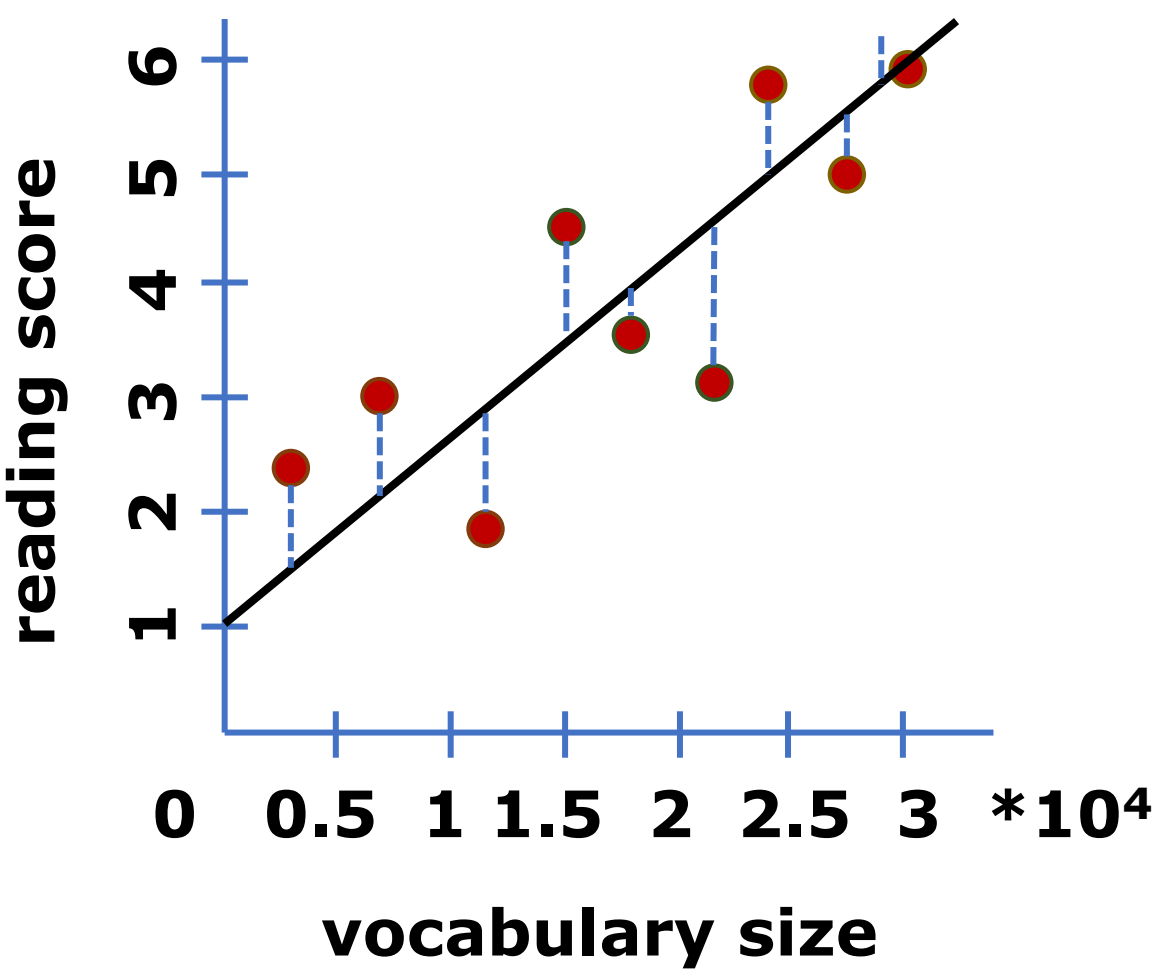


$$R^2 = \frac{SSR}{SST}$$

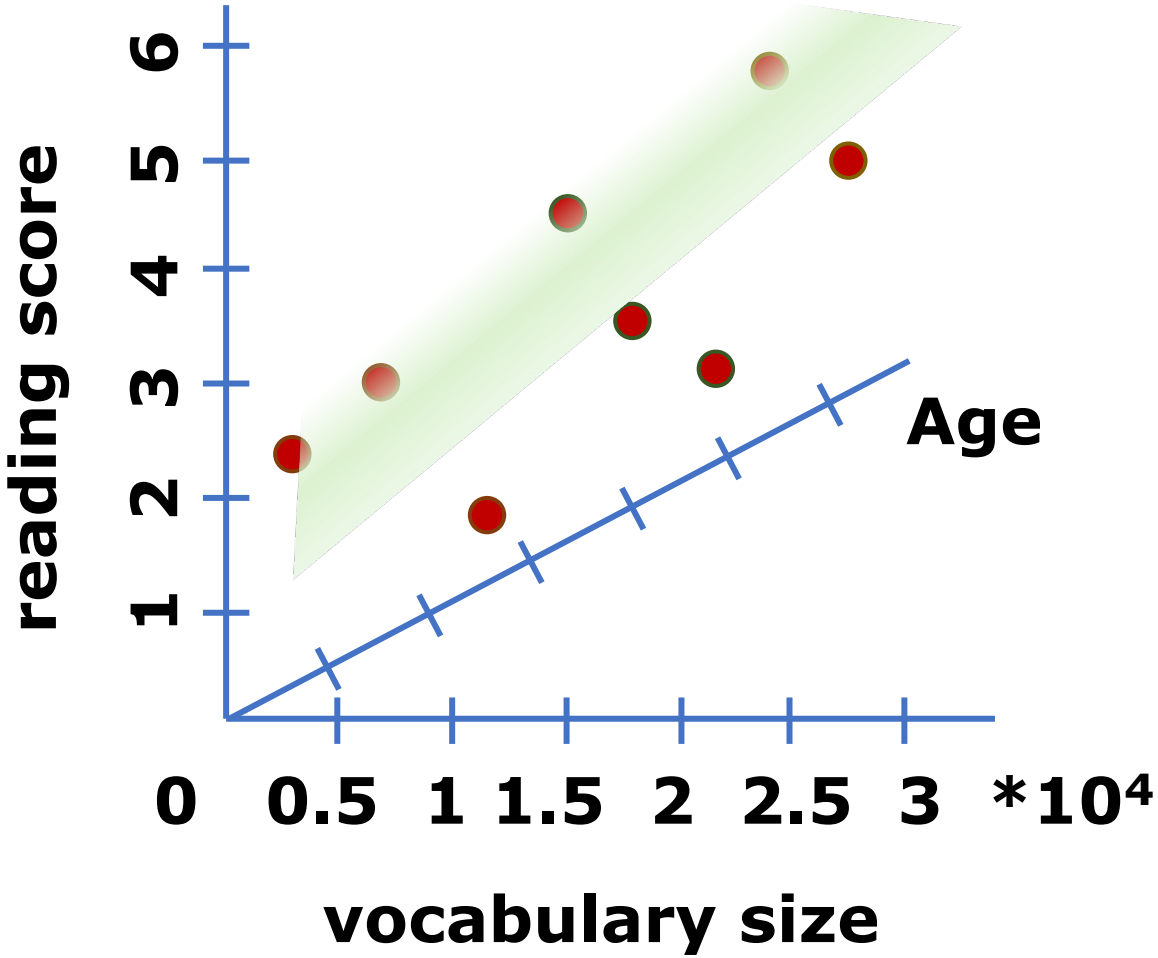
x has positive effect on y

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Multiple regression



$$y = b_1 x + b_0$$

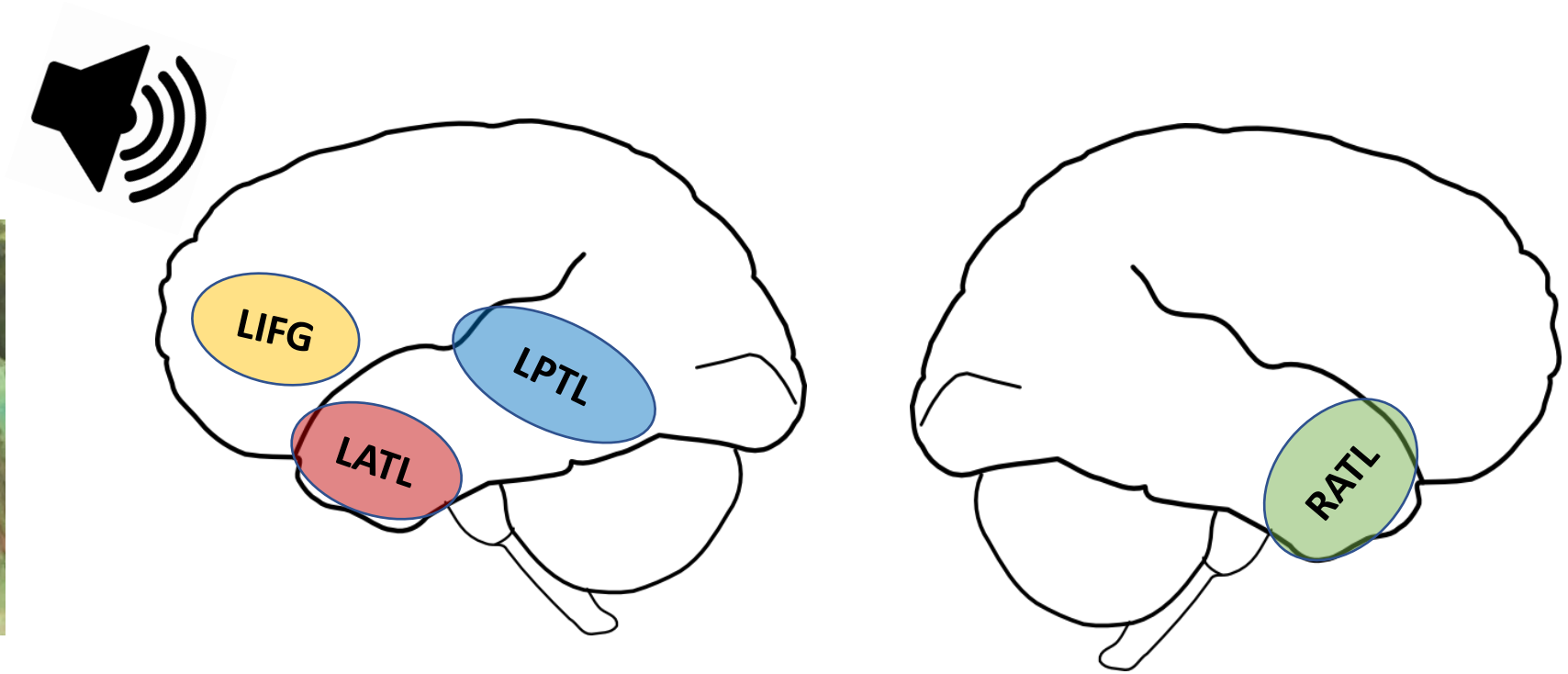


$$y = b_1 x_1 + b_2 x_2 + b_0$$

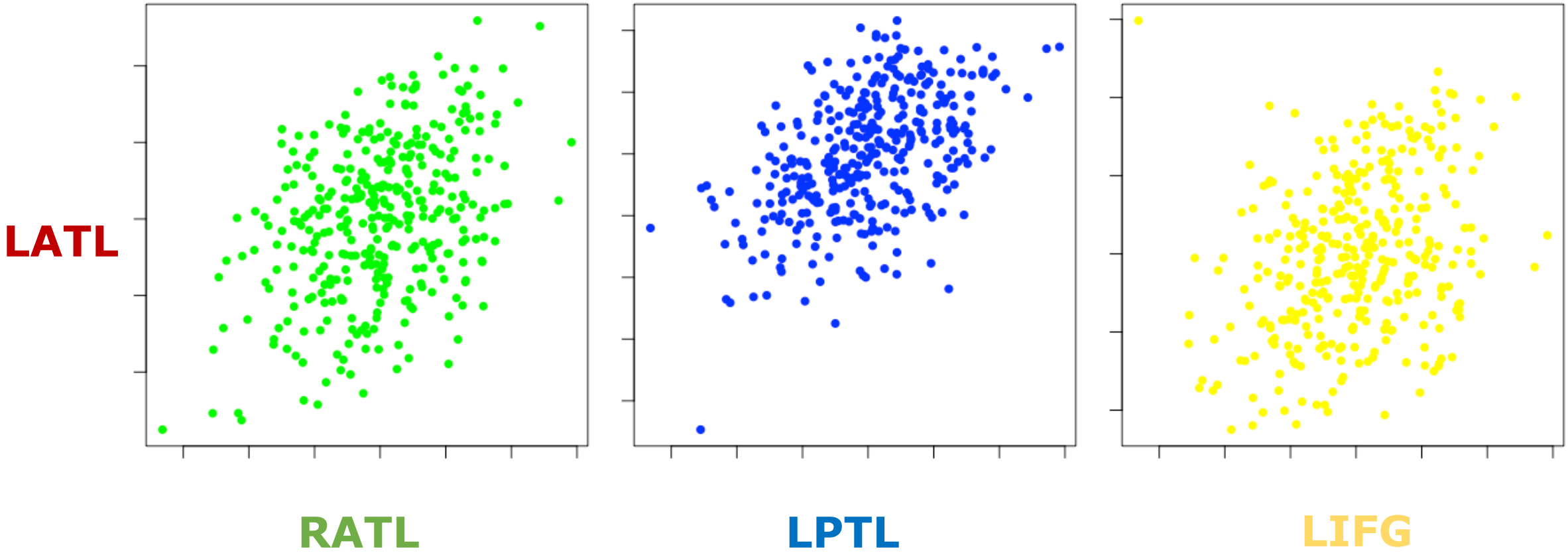
Example: Alice dataset

Participants listened to the first chapter of *Alice in Wonderland* in the fMRI scanner.

Question: How is the brain activity in RATL, LPTL and LIFG affect the brain activity in LATL?



Plot data first



All three brain regions seem to have a linear relationship with LATL

One-variable model

LATL~RATL

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.115e-12  4.828e-02   0.000        1
RATL        3.980e-01  4.835e-02   8.232 3.42e-15 ***

Residual standard error: 0.9186 on 360 degrees of freedom
Multiple R-squared:  0.1584,    Adjusted R-squared:  0.1561
F-statistic: 67.77 on 1 and 360 DF,  p-value: 3.419e-15
```

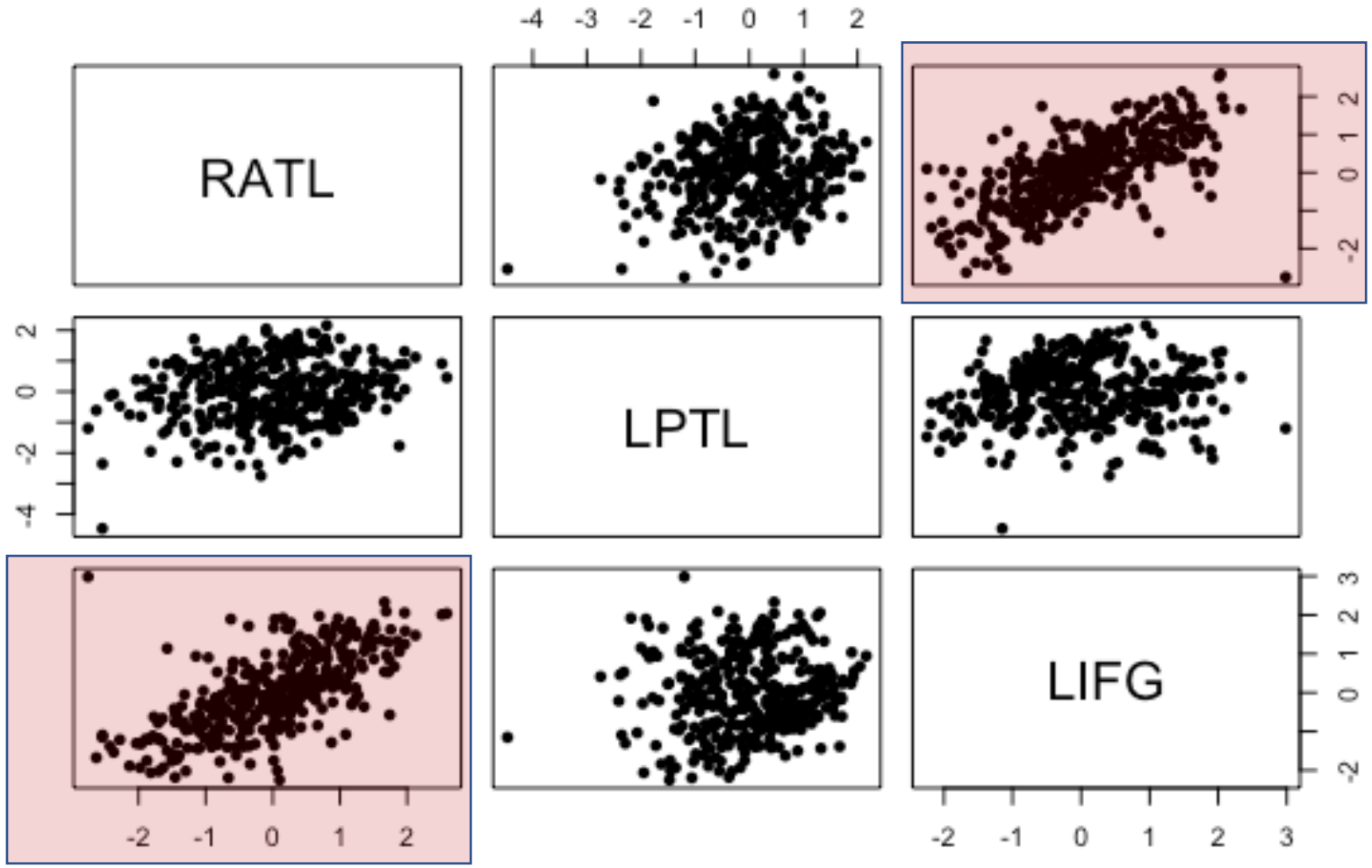
LATL~LIFG

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.295e-11  5.000e-02   0.000        1
LIFG        3.122e-01  5.007e-02   6.234 1.27e-09 ***

Residual standard error: 0.9513 on 360 degrees of freedom
Multiple R-squared:  0.09745,   Adjusted R-squared:  0.09494
F-statistic: 38.87 on 1 and 360 DF,  p-value: 1.27e-09
```

LATL~LPTL

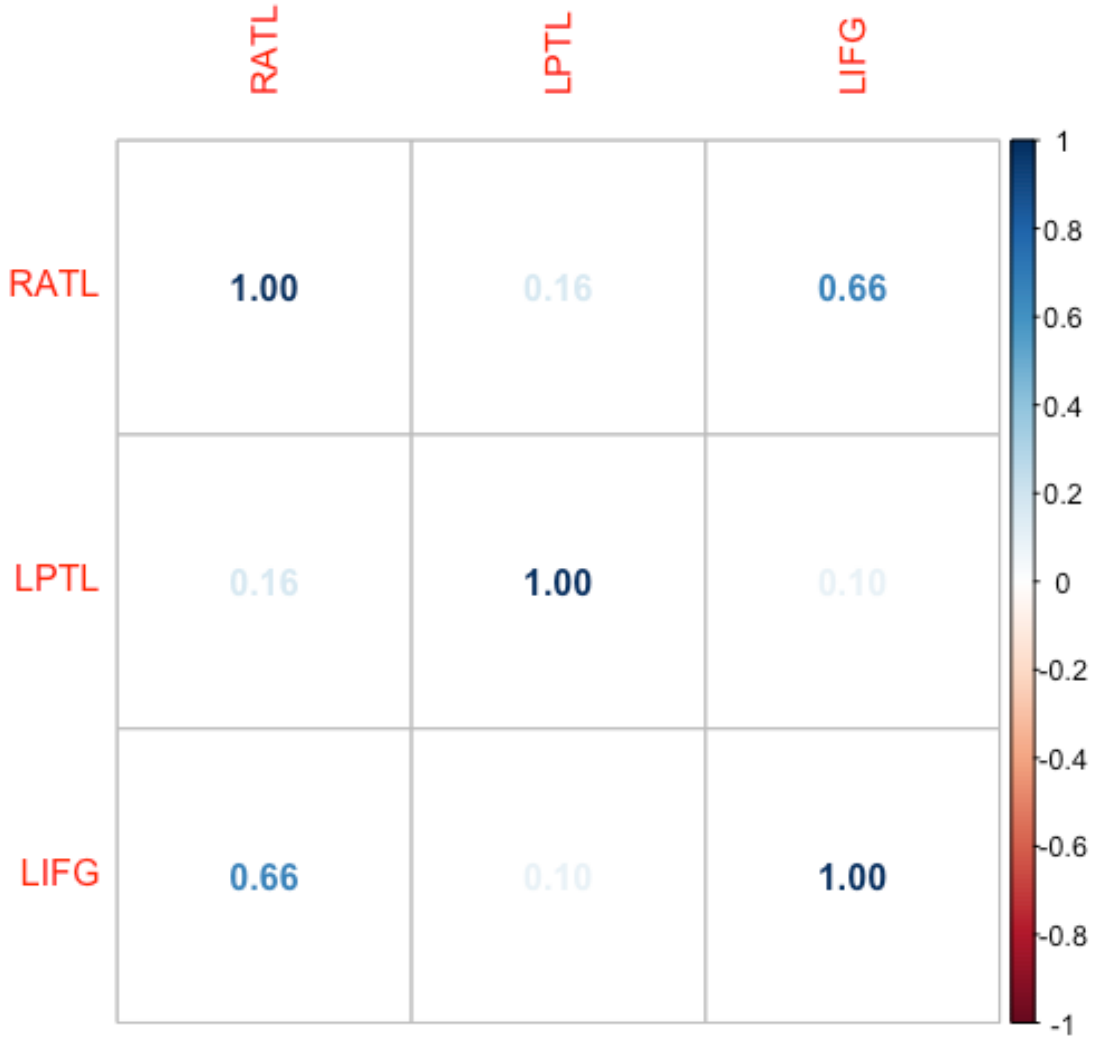
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.381e-11  4.582e-02    0.00        1
LPTL        4.921e-01  4.588e-02   10.72   <2e-16 ***

Residual standard error: 0.8718 on 360 degrees of freedom
Multiple R-squared:  0.2421,    Adjusted R-squared:   0.24
F-statistic:   115 on 1 and 360 DF,  p-value: < 2.2e-16
```

Correlation among independent variables?



RATL correlated with LIFG

Correlation matrix



$p < 2.2e\text{-}16$

Multicollinearity:
Two or more independent variables are correlated

→ We cannot be sure which variable explains the variance in the dependent variable

Two-variable model

LATL~RATL + LPTL

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.487e-12  4.261e-02   0.000        1
RATL        3.271e-01  4.324e-02   7.564  3.3e-13 ***
LPTL        4.392e-01  4.324e-02  10.158  < 2e-16 ***

Residual standard error: 0.8108 on 359 degrees of freedom
Multiple R-squared:  0.3463,    Adjusted R-squared:  0.3427
F-statistic: 95.09 on 2 and 359 DF,  p-value: < 2.2e-16
```

LATL~ RATL + LIFG

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.988e-12  4.823e-02   0.000    1.000
RATL        3.405e-01  6.437e-02   5.289 2.14e-07 ***
LIFG        8.706e-02  6.437e-02   1.353    0.177

Residual standard error: 0.9176 on 359 degrees of freedom
Multiple R-squared:  0.1627,    Adjusted R-squared:  0.158
F-statistic: 34.88 on 2 and 359 DF,  p-value: 1.438e-14
```

LATL~LPTL + LIFG

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.307e-11  4.368e-02   0.000        1
LPTL        4.665e-01  4.395e-02  10.615  < 2e-16 ***
LIFG        2.676e-01  4.395e-02   6.089 2.93e-09 ***

Residual standard error: 0.8311 on 359 degrees of freedom
Multiple R-squared:  0.313,     Adjusted R-squared:  0.3092
F-statistic:  81.8 on 2 and 359 DF,  p-value: < 2.2e-16
```

The full model

LATL~RATL + LPTL + LIFG

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.451e-12  4.250e-02   0.000   1.0000
RATL        2.636e-01  5.723e-02   4.605 5.73e-06 ***
LPTL        4.403e-01  4.313e-02  10.208  < 2e-16 ***
LIFG        9.581e-02  5.674e-02   1.689   0.0921 .

Residual standard error: 0.8087 on 358 degrees of freedom
Multiple R-squared:  0.3515,    Adjusted R-squared:  0.346
F-statistic: 64.67 on 3 and 358 DF,  p-value: < 2.2e-16
```

Model comparison

Model	*F*	*p*	R^2	R^2 adjusted	VIF
RATL	67.77	3.419e-15	0.158	0.156	1
LPTL	115	< 2.2e-16	0.242	0.24	1
LIFG	38.87	1.27e-09	0.097	0.095	1
RATL+LPTL	95.09	< 2.2e-16	0.346	0.343	1.03
RATL+LIFG	34.88	1.438e-14	0.163	0.158	1.78
LPTL+LIFG	81.8	< 2.2e-16	0.313	0.309	1.01
RATL+LPTL+LIFG	64.67	< 2.2e-16	0.352	0.346	1.81,1.03,1.78

Adjusted R^2

Adjusted R^2 is a corrected goodness-of-fit measure for linear models.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 does not decrease when the number of variables increases;
Additional variable usually will account for some variance, if not 0.

→ **need a standardized R^2:**

$$R^2\,adjusted = 1 - \frac{R^2\,(n-1)}{n-k-1}$$

n: number of observations
k: number of model parameters

→ As the number of parameters increases, adjusted R^2 decreases if R^2 does not increase significantly

Variance Inflation Factor (VIF)

VIF provides a measure of multicollinearity among the independent variables in a multiple regression model.

$$VIF_i = \frac{1}{1-R_i^2}$$

Regressing the i_{th} variable on the other variables: How much variance in the i_{th} variable can be explained by other variables

$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

$x1 = \beta_0 + \beta_1x_2 + \beta_2x_3 \rightarrow R_1^2$

VIF < 3 is usually good; VIF > 10 indicates high collinearity.

Best model?

Model	*F*	*p*	R^2	R^2 adjusted	VIF
RATL	67.77	3.419e-15	0.158	0.156	1
LPTL	115	< 2.2e-16	0.242	0.24	1
LIFG	38.87	1.27e-09	0.097	0.095	1
RATL+LPTL	95.09	< 2.2e-16	0.346	0.343	1.03
RATL+LIFG	34.88	1.438e-14	0.163	0.158	1.78
LPTL+LIFG	81.8	< 2.2e-16	0.313	0.309	1.01
RATL+LPTL+LIFG	64.67	< 2.2e-16	0.352	0.346	1.81,1.03,1.78

Testing whether the more complex model is significantly better at capturing the data than the simpler model.

```
Model 1: LATL ~ RATL + LPTL
Model 2: LATL ~ RATL + LPTL + LIFG
  Res.Df RSS Df Sum of Sq   F Pr(>F)
1    359 236
2    358 234  1       1.9 2.9   0.09 .
```