Linear regression

Methods 1, E2021 - Lecture 7
Tuesday 26/10/2021
Fabio Trecca


QUIZ
TIME

Quiz time

- What is a quasi-experiment?
- What is an independent measures design?
- What is a repeated measures design?
- $$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
- $$\frac{\overline{D} - \mu_D}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Statistical tests for this semester (1)

- **Correlation**
  - descriptive assessment/inferential test of the relation between two continuous variables
- **T-test**
  - inferential test of whether two means are significantly different from each other / whether one mean is significantly different from a hypothesized mean
- **(Simple/multiple) linear regression**
  - inferential test predicting a continuous outcome from one or more continuous or categorical predictors

Statistical tests for this semester (2)

…participant *i* from a sample of participants

our statistical model

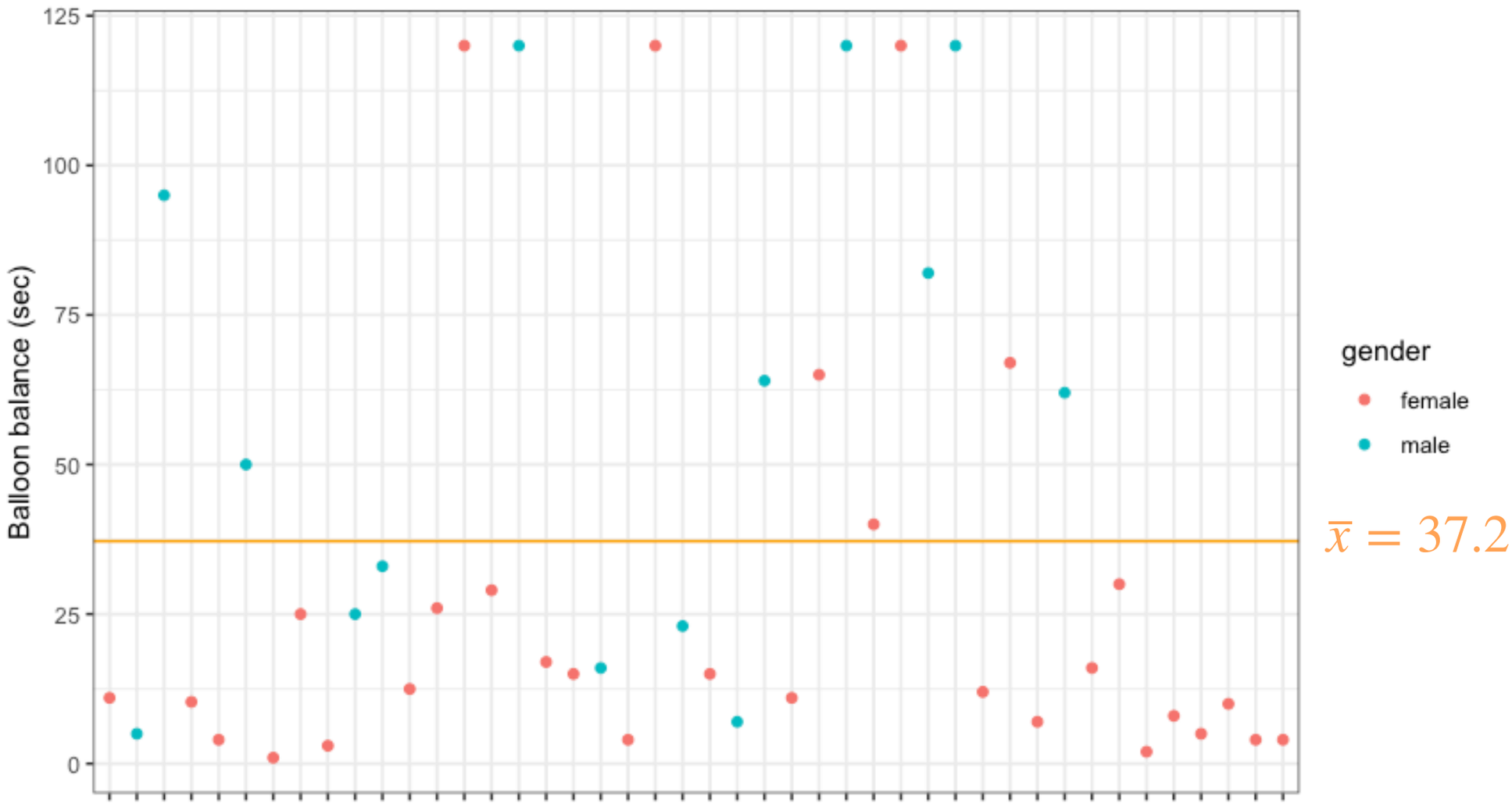
$Outcome_i = model + error_i$

a certain observation/ data point collected from…

predicted by

and some error (= deviation from the model) for each participant *i*

Mean as simplest model ("null model")



Error as measure of model fit

- Deviance:

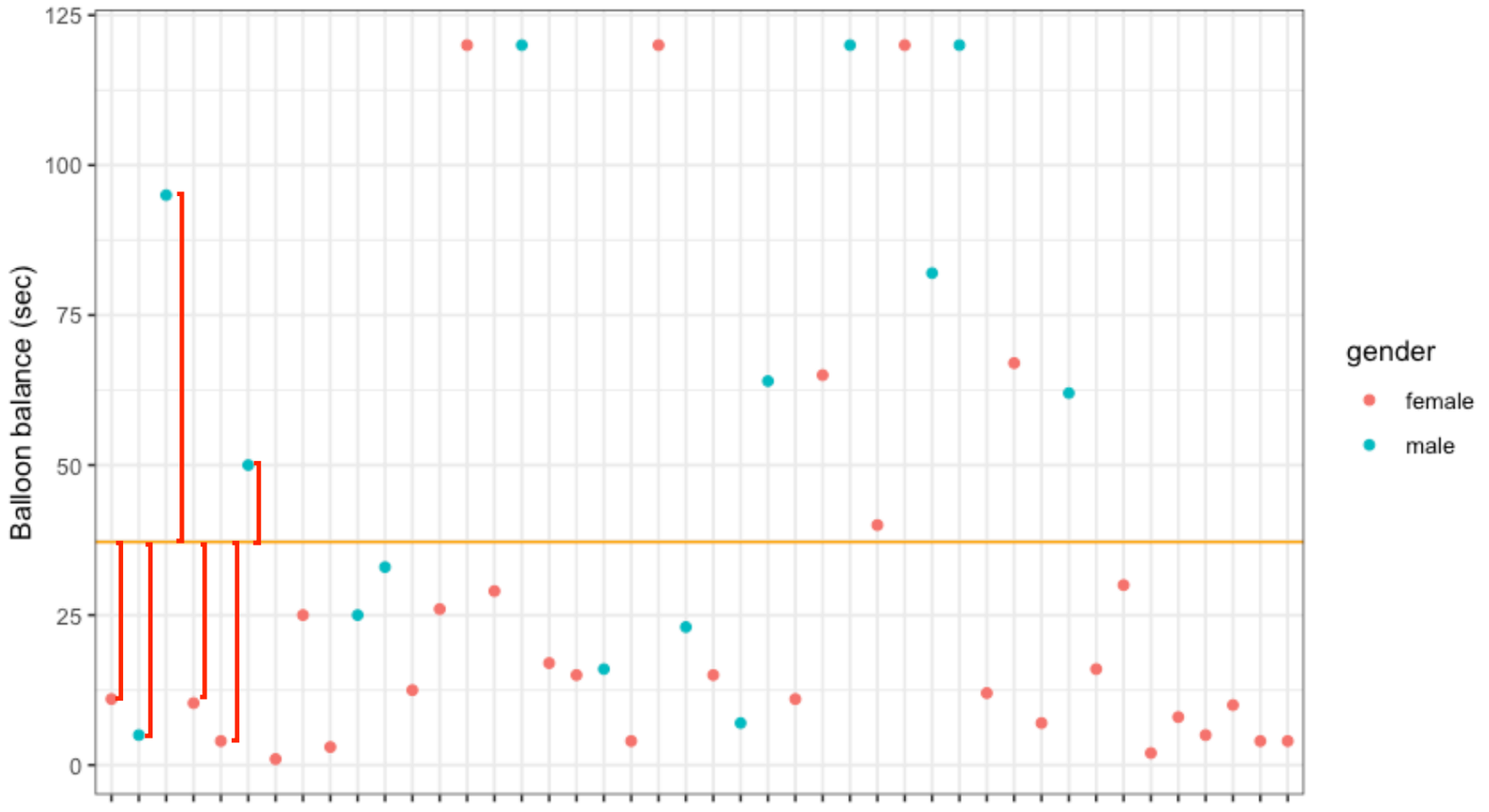
$$SS = \sum (x_i - \bar{x})^2$$

- Variance:

$$s^2 = \frac{SS}{N-1}$$

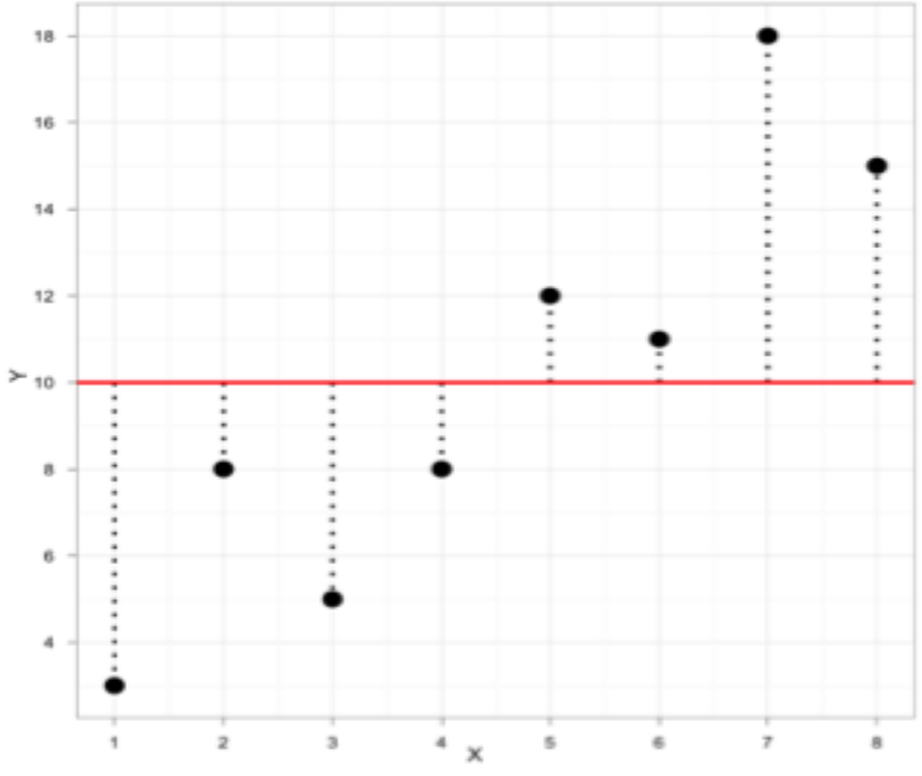
- Standard deviation:

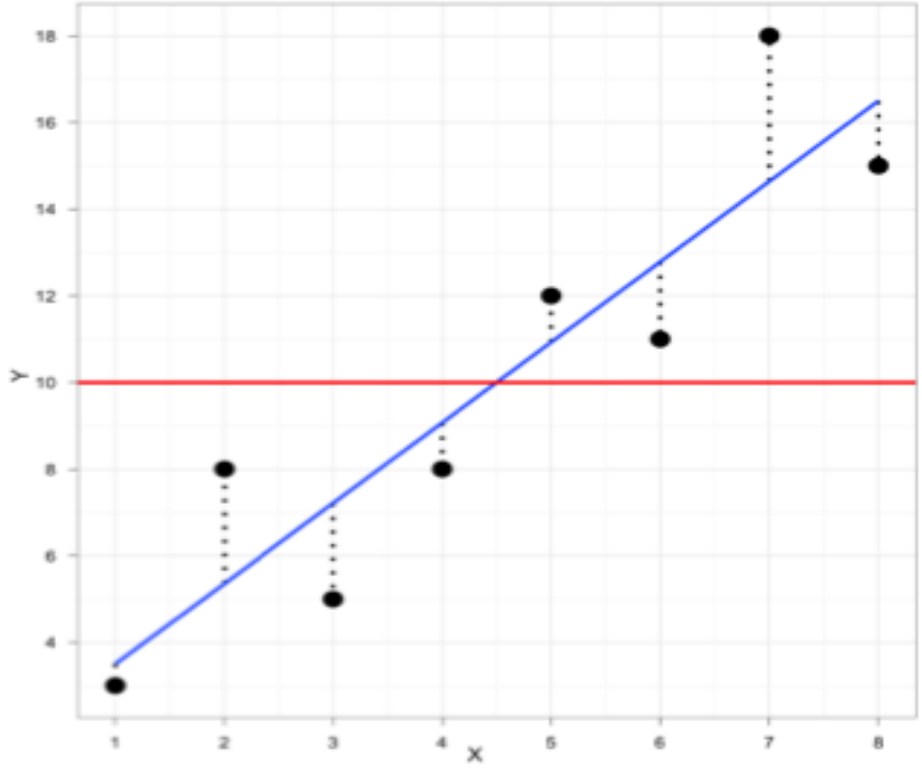
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$$



The regression line: A better model?

- Is there a better model other than the mean to summarize our data?
- The regression line is often a better model (if there is a relationship between the variables)





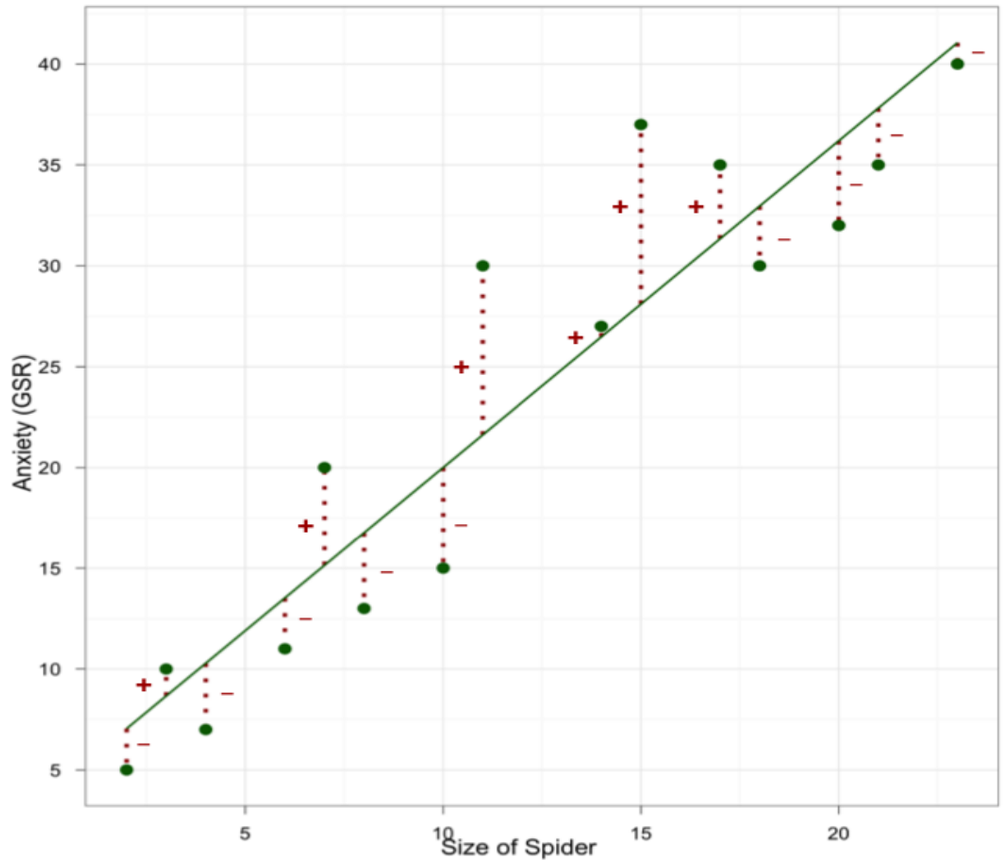
The regression line

- The line that minimizes the vertical distances between the model and the data points

- = The line that minimizes the sum of the squares of the error (residuals)

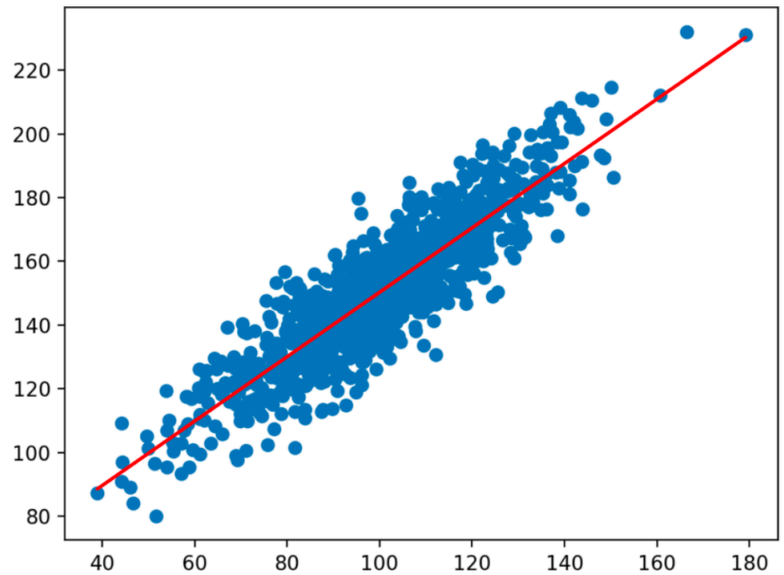
- $SS = \sum (x_i - \bar{x})^2$ (see Lecture 3)

- Ordinary least squares method

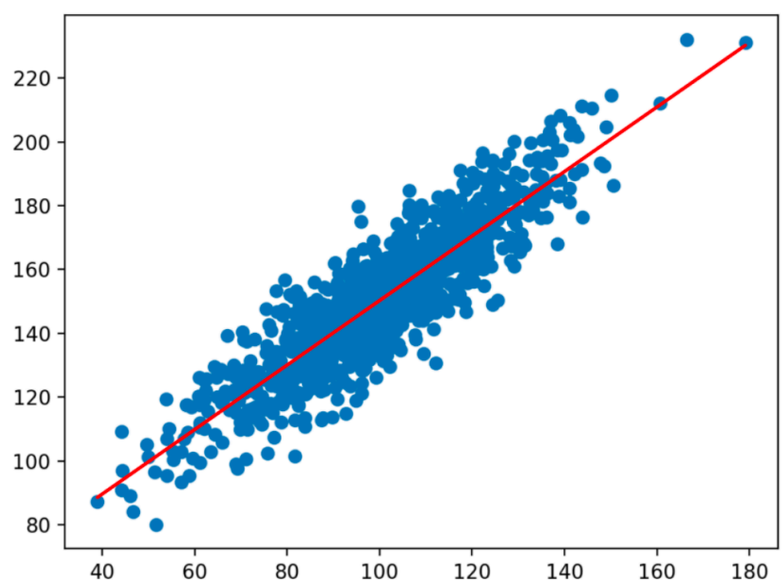


Correlation vs linear regression

- **Correlation**
  - Implies association in a quasi-experimental setting
  - No directionality effect between variables (order doesn't matter)



- **Regression**
  - Implies causation in an experimental setting
  - Directionality: the outcome variable is explained by the predictor(s)



Fitting the regression line

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Null model: $Y_i = \bar{x} + \varepsilon_i$

Fitting the regression line

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

The i-th value of the outcome Y that I am trying to predict

Fitting the regression line

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Regression coefficient for the intercept of the regression line
= the value of the outcome variable Y when the predictor variable X has value 0

The *i*-th value of the outcome Y that I am trying to predict

Fitting the regression line

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Regression coefficient for the slope of the regression line = the value by which Y changes for each increment of X

Regression coefficient for the intercept of the regression line = the value of the outcome variable Y when the predictor variable X has value 0

The *i*-th value of the outcome Y that I am trying to predict

Fitting the regression line

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

The i-th value of the predictor variable X

Regression coefficient for the slope of the regression line = the value by which Y changes for each increment of X

Regression coefficient for the intercept of the regression line = the value of the outcome variable Y when the predictor variable X has value 0

The i-th value of the outcome Y that I am trying to predict

Fitting the regression line

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

The residual error for the *i*-th observation

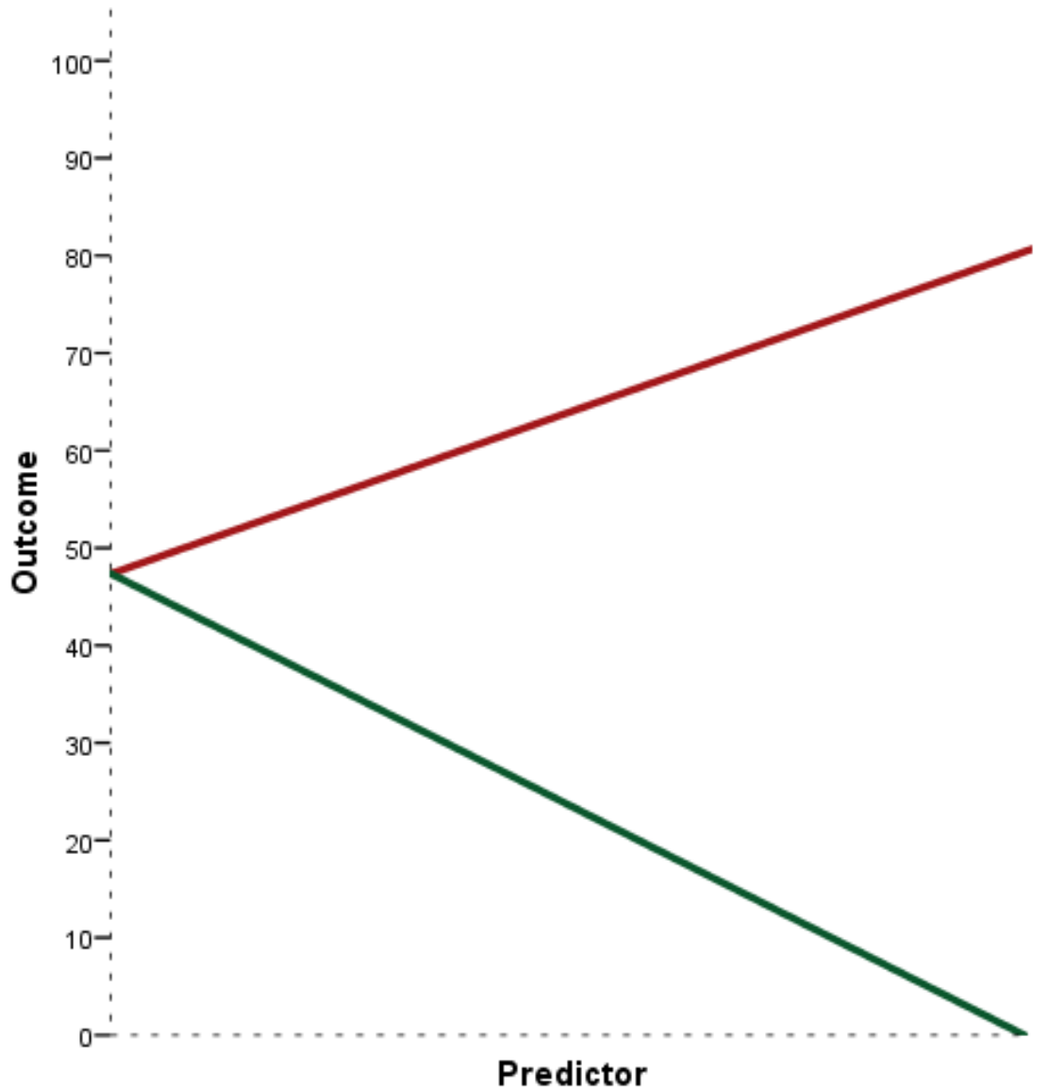
The *i*-th value of the predictor variable X

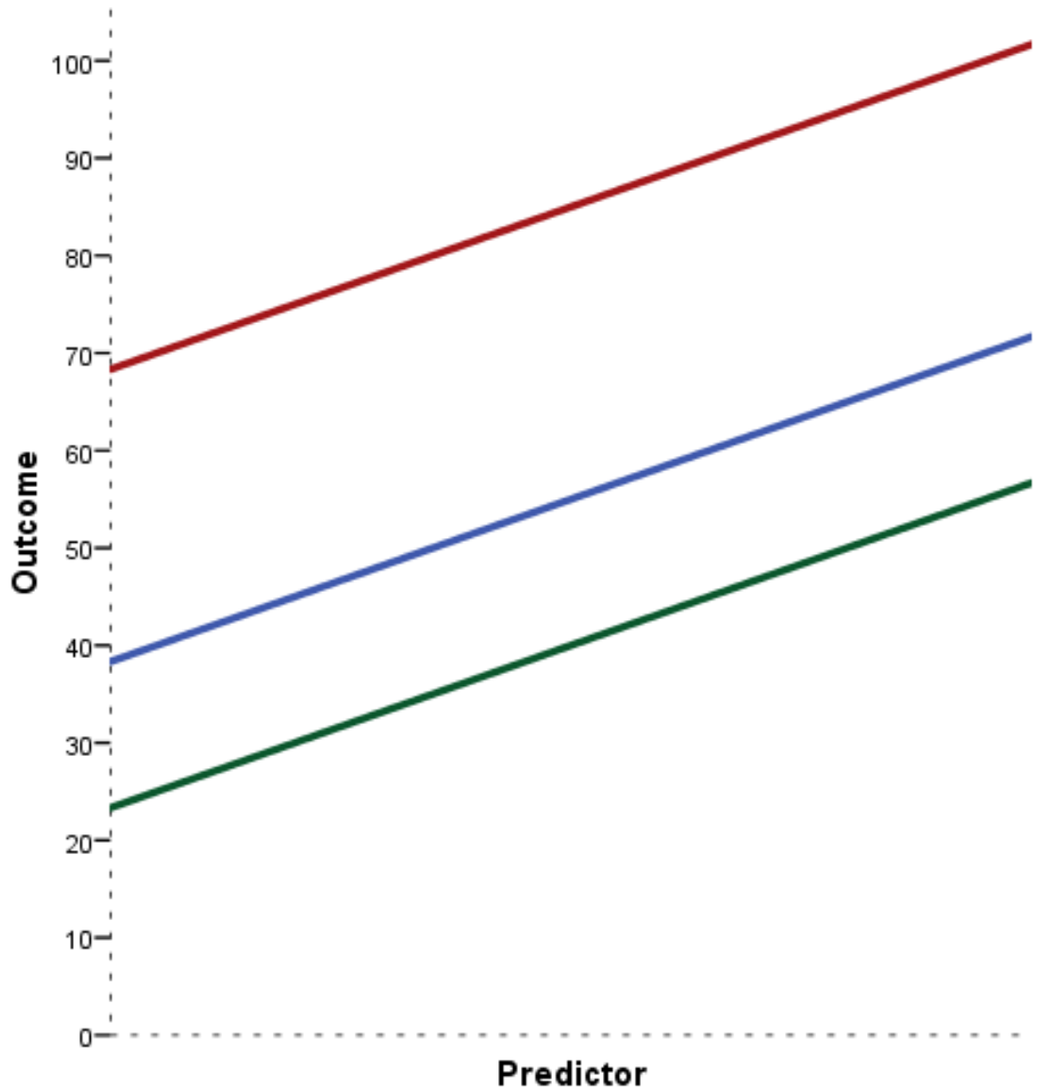
Regression coefficient for the slope of the regression line = the value by which Y changes for each increment of X

Regression coefficient for the intercept of the regression line = the value of the outcome variable Y when the predictor variable X has value 0

The *i*-th value of the outcome Y that I am trying to predict

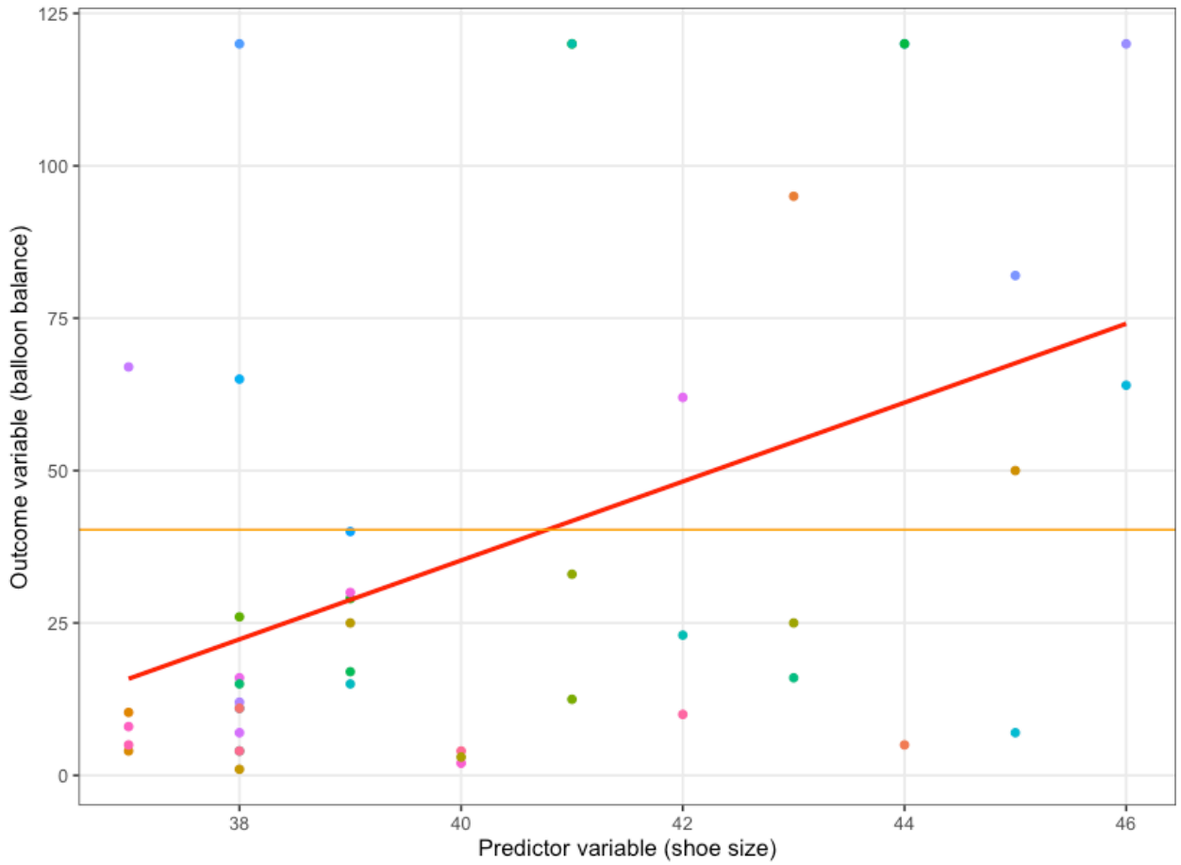
Intercepts and slopes





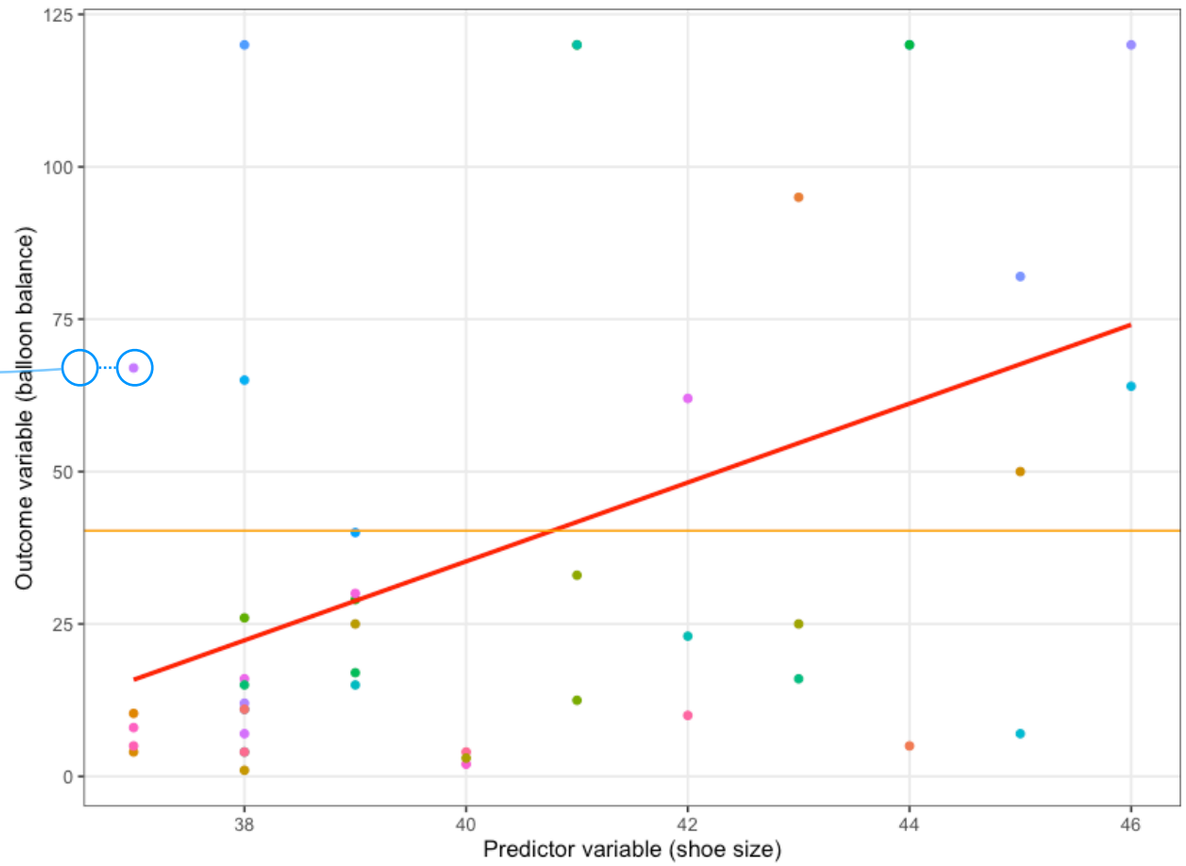
Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



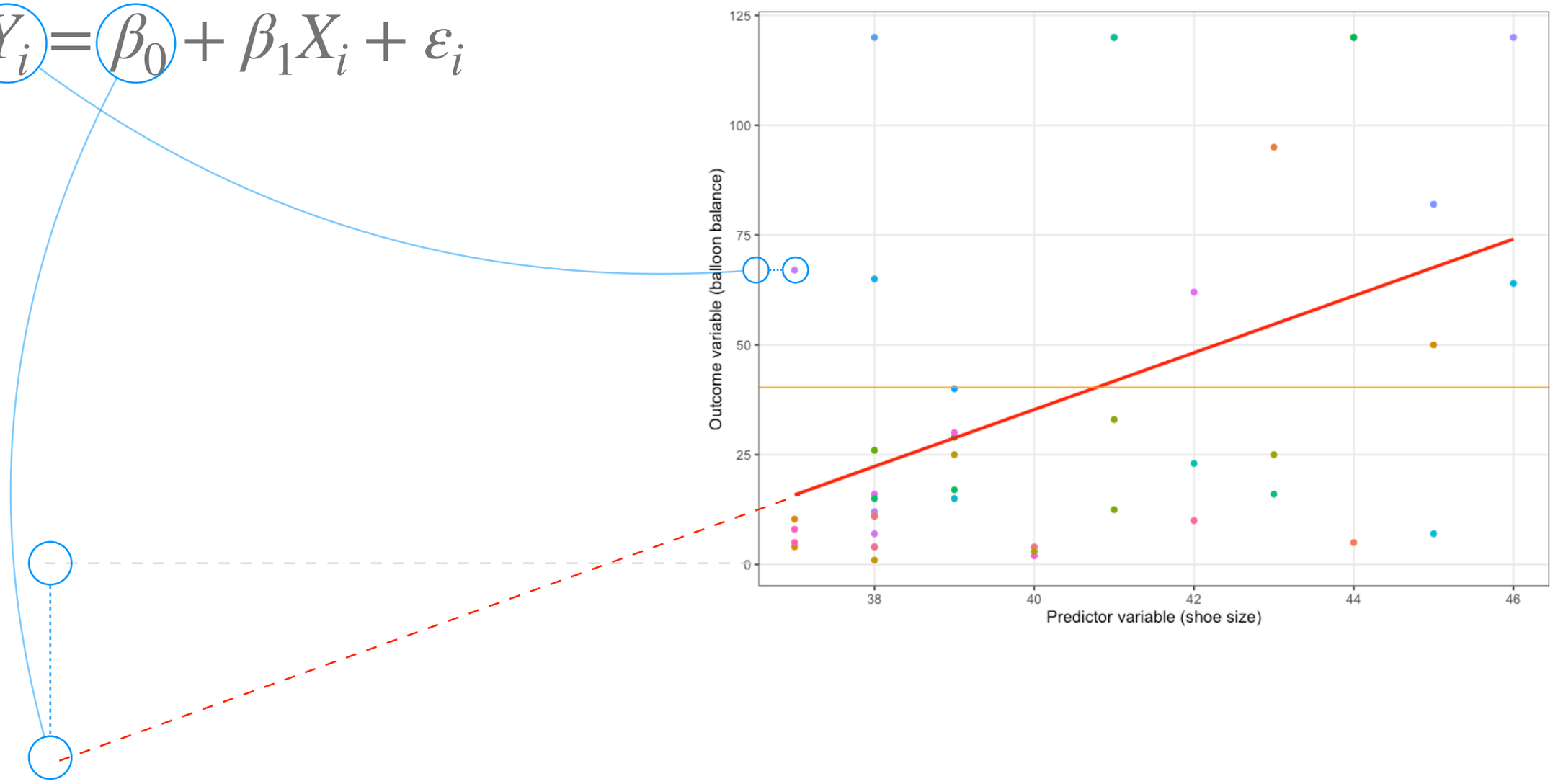
Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



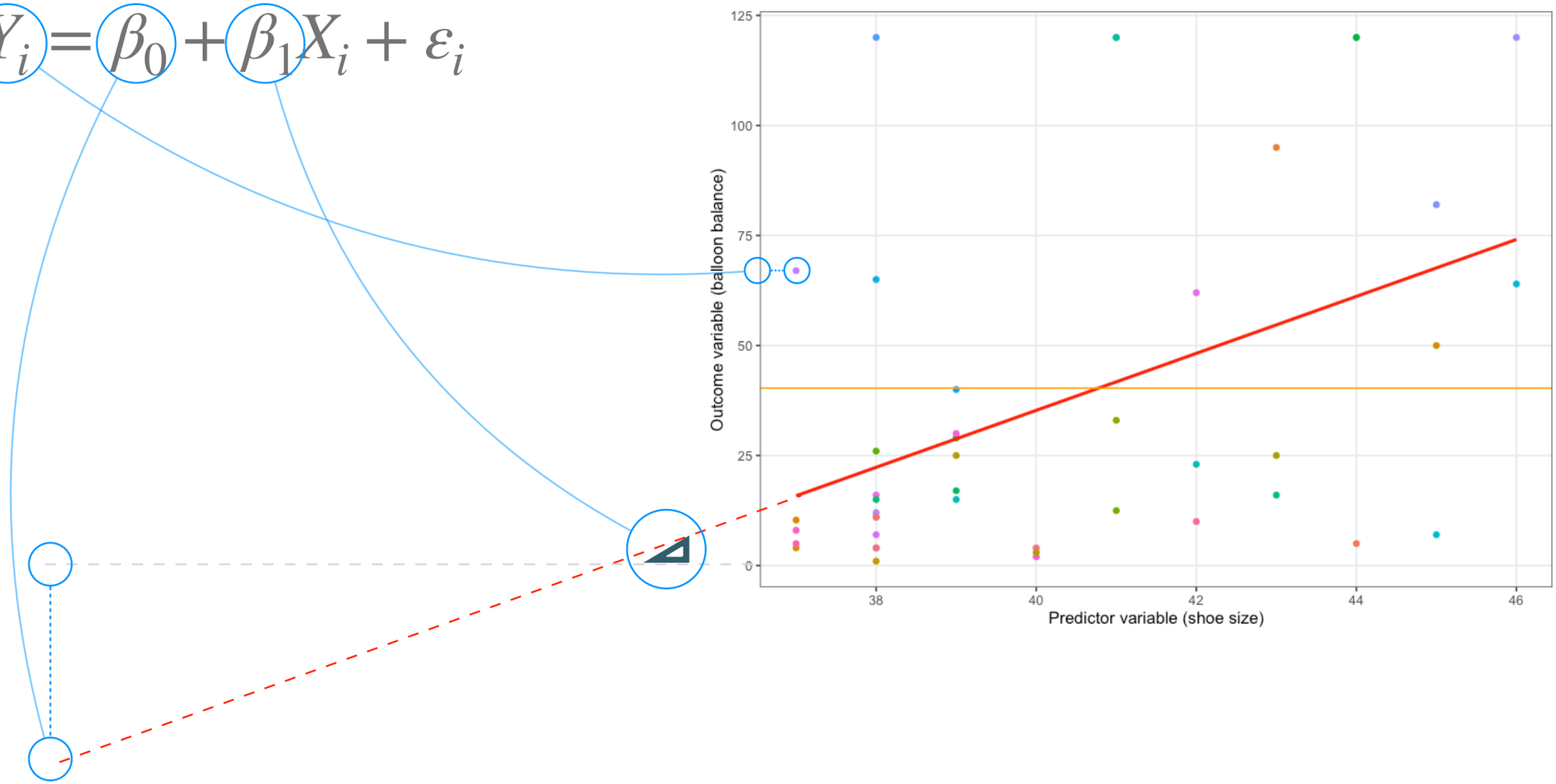
Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



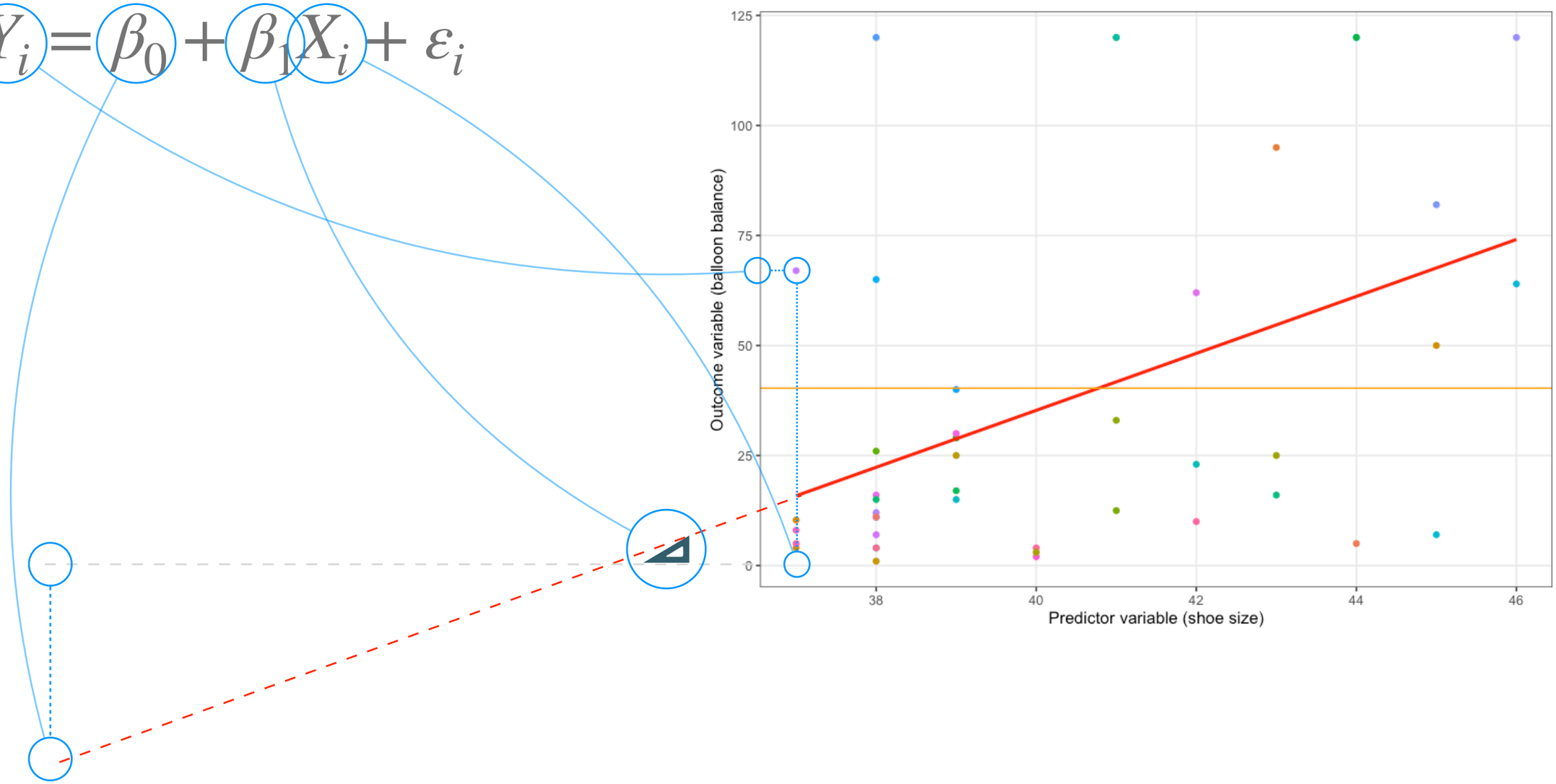
Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



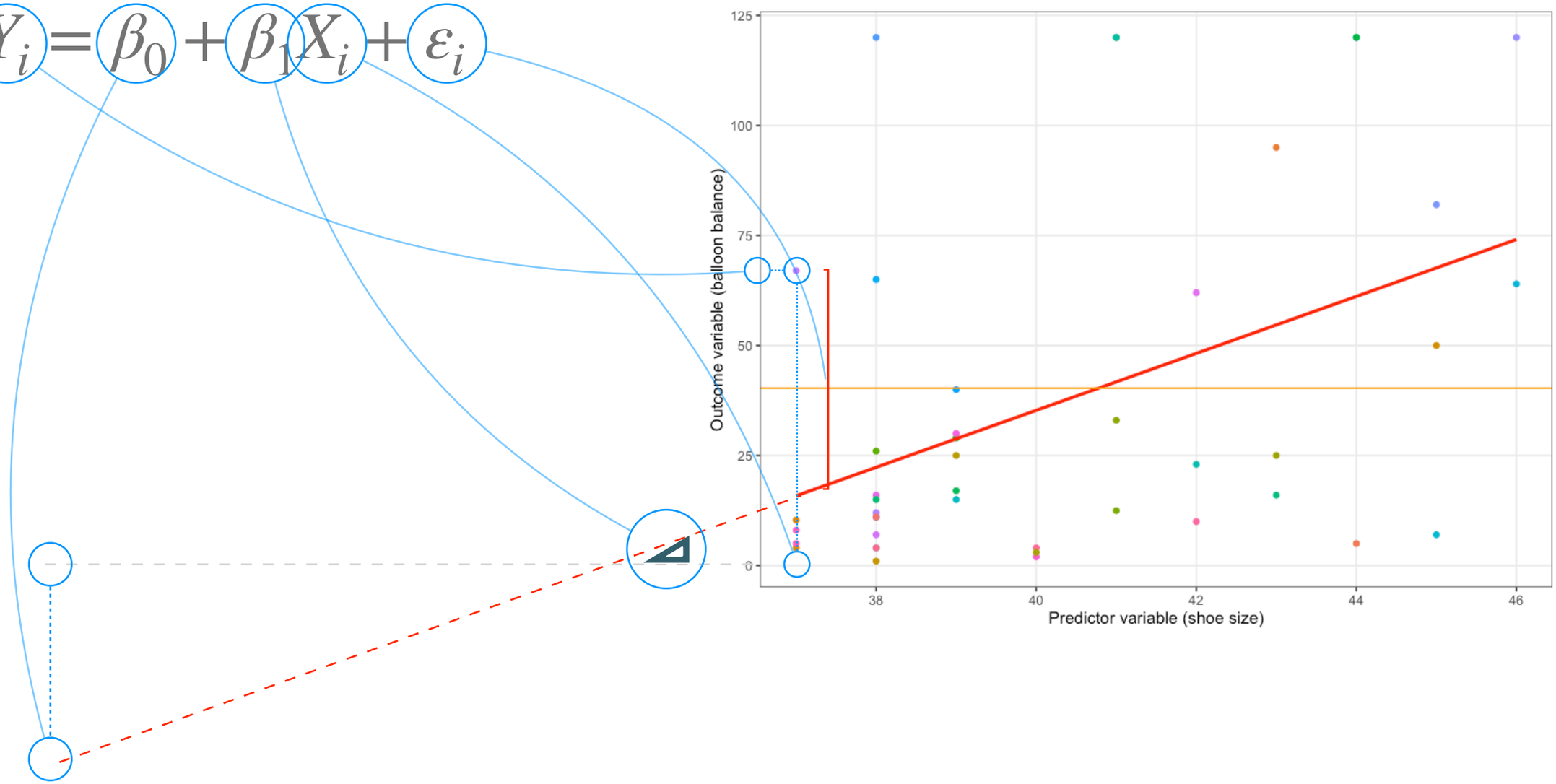
Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



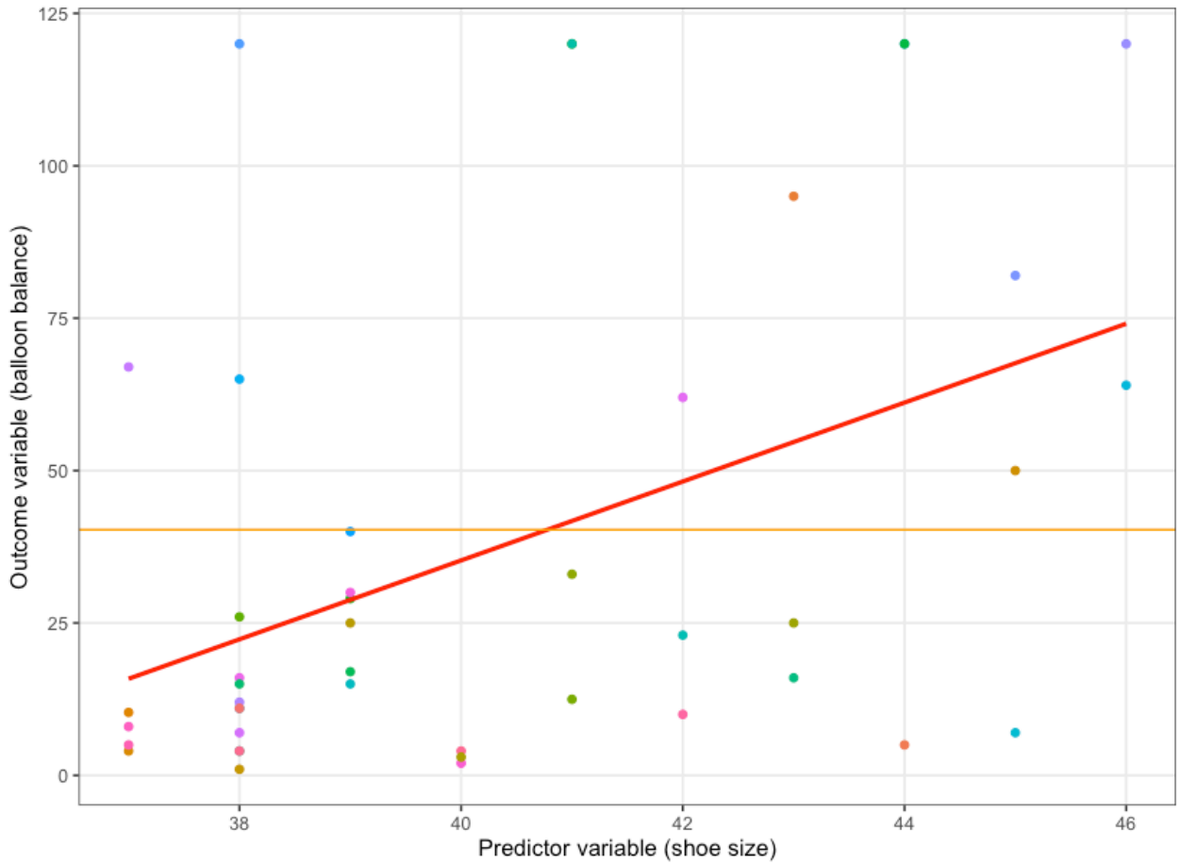
Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



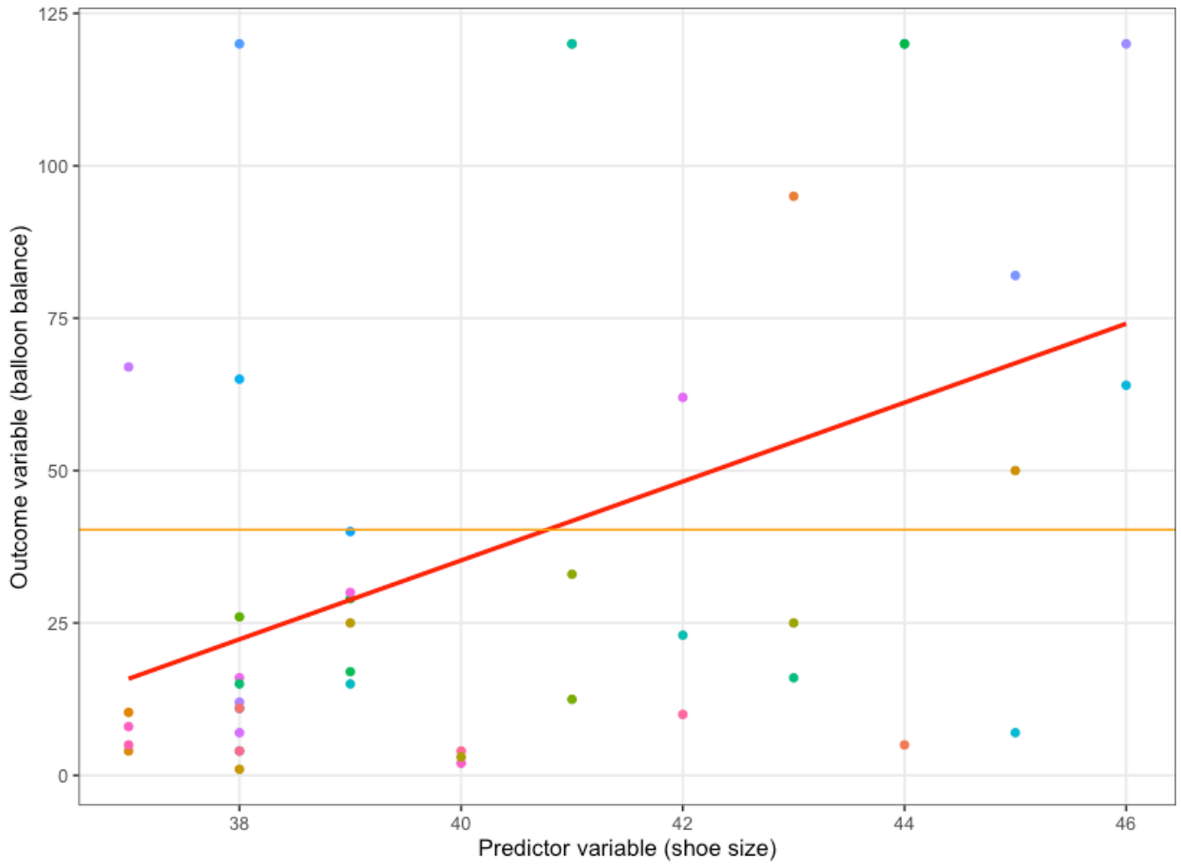
Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i =$
- $= -223.612 + 6.472 X_i + \varepsilon_i$



Interpreting the regression formula

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i =$
- $= -223.612 + 6.472 X_i + \varepsilon_i$



- $X_{Fabio} = 44$
- $Y_{Fabio} = -223.612 + 6.472 \times 44 + \varepsilon_{Fabio}$
- $Y_{Fabio} = 61.146$

Regression analysis output

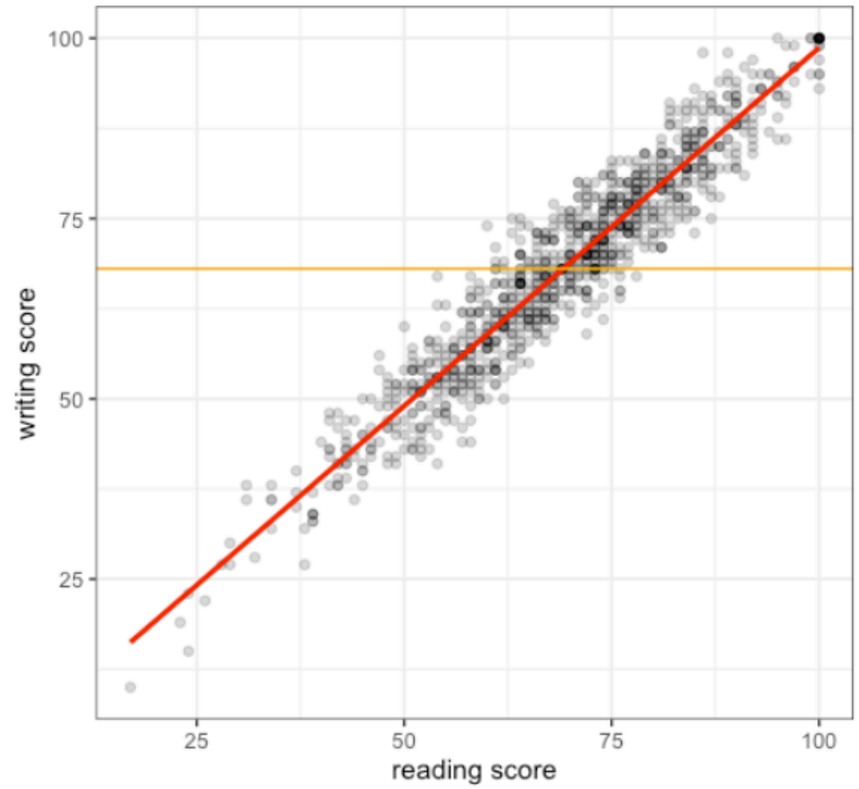
- **Estimation of model parameters:**
  - intercept ($\hat{\beta}_0$) and slope ($\hat{\beta}_1$)
  - how does the model answer my research question?
- **Model fit:**
  - how much of the variance in our outcome variable is explained by our predictor variable(s)?
  - is the quality of my model high or low?

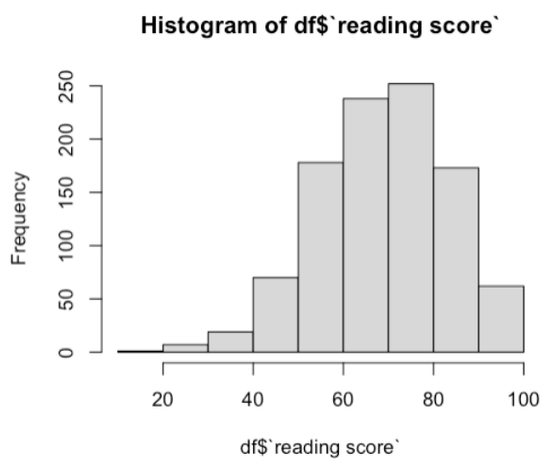
lm() function in R

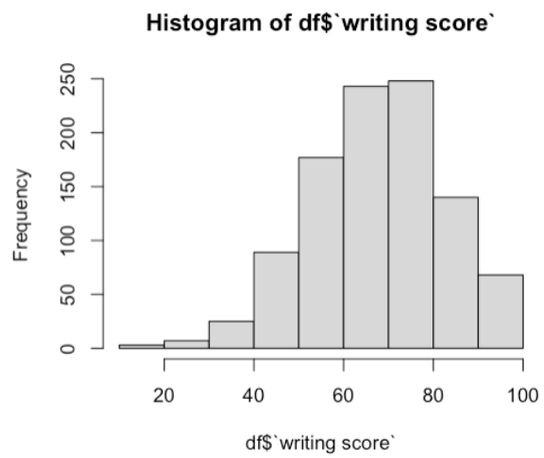
- `lm(outcome ~ predictor, data = mydata)`
- `summary(lm(outcome ~ predictor, data = mydata))`

Regression with continuous predictors

```
> summary(lm(`writing score` ~ `reading score`, data = df))

Call:
lm(formula = `writing score` ~ `reading score`, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.667554   0.693792  -0.962    0.336
`reading score`  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113, Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```







Regression with continuous predictors

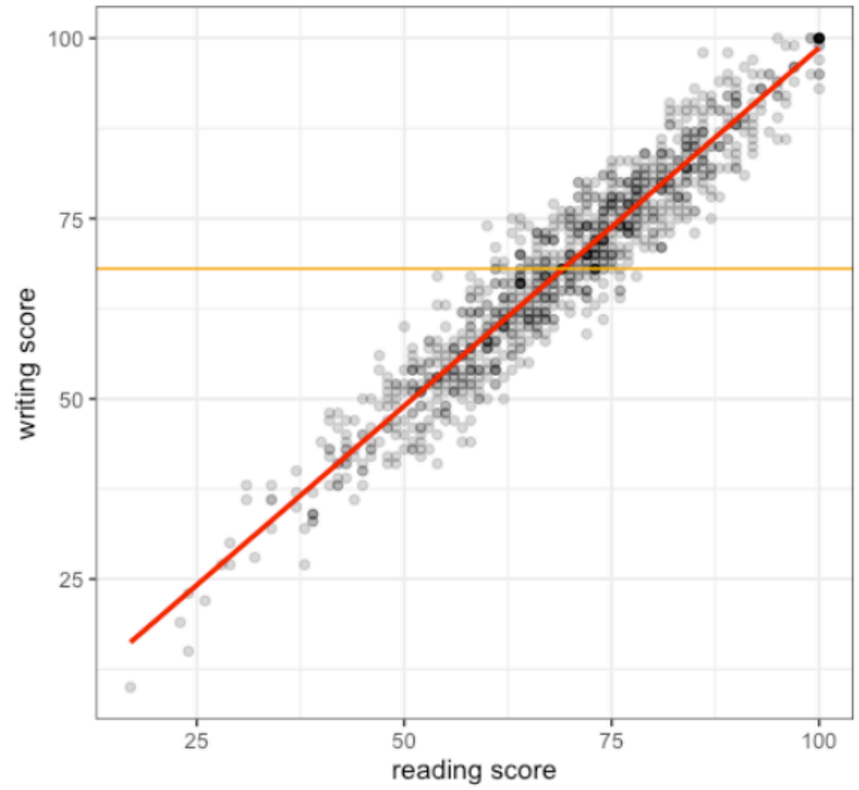
```
> summary(lm(`writing score` ~ `reading score`, data = df))
```

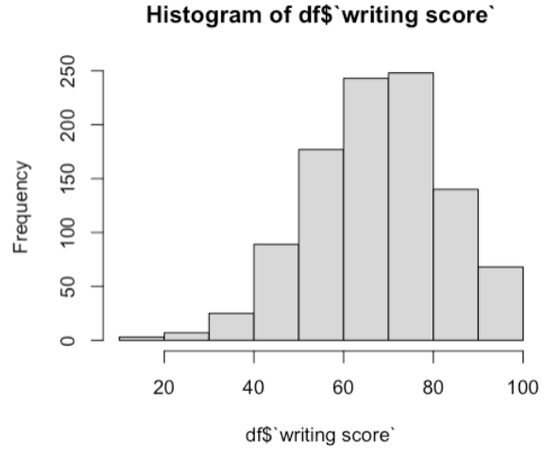
```
Call:
lm(formula = `writing score` ~ `reading score`, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.667554   0.693792  -0.962    0.336
`reading score`  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model parameters

```
Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113, Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Model fit





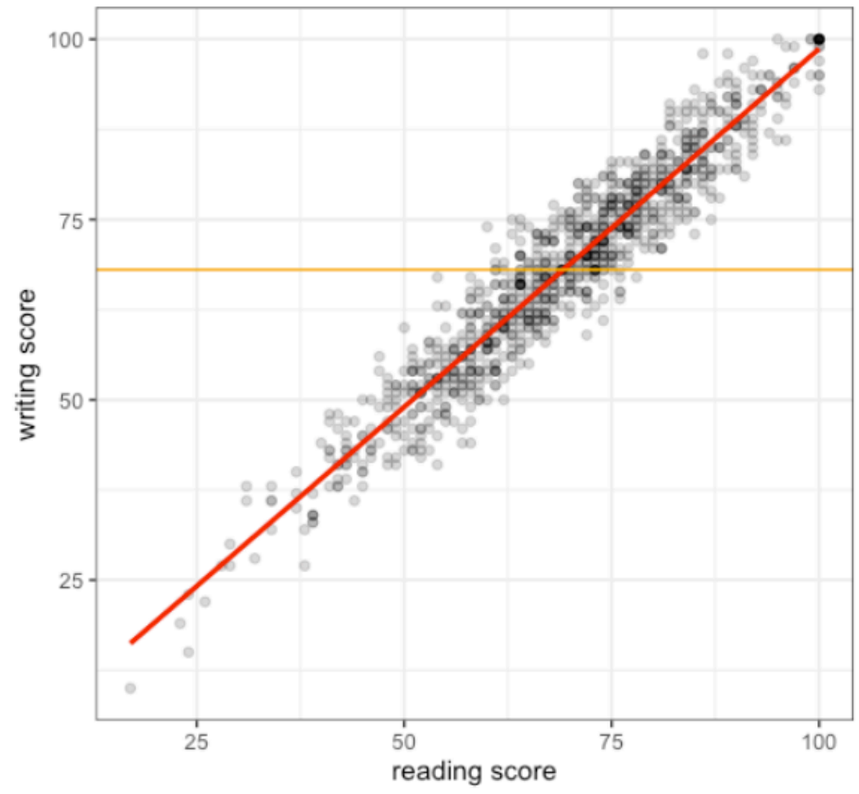
Regression with continuous predictors

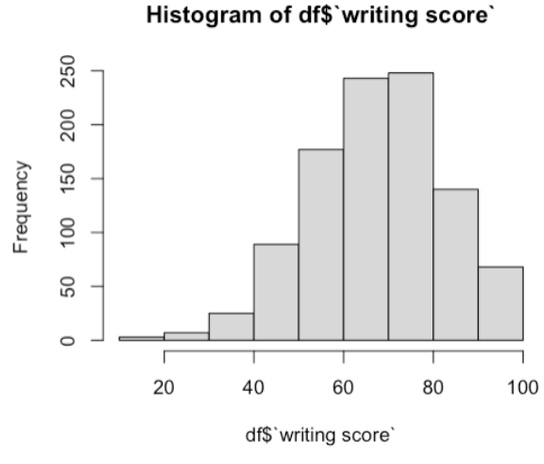
```
> summary(lm(`writing score` ~ `reading score`, data = df))

Call:
lm(formula = `writing score` ~ `reading score`, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.667554   0.693792  -0.962    0.336
`reading score`  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113,  Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Model parameters

Inferential test statistics

$\hat{\beta}_0$ $\hat{\beta}_1$

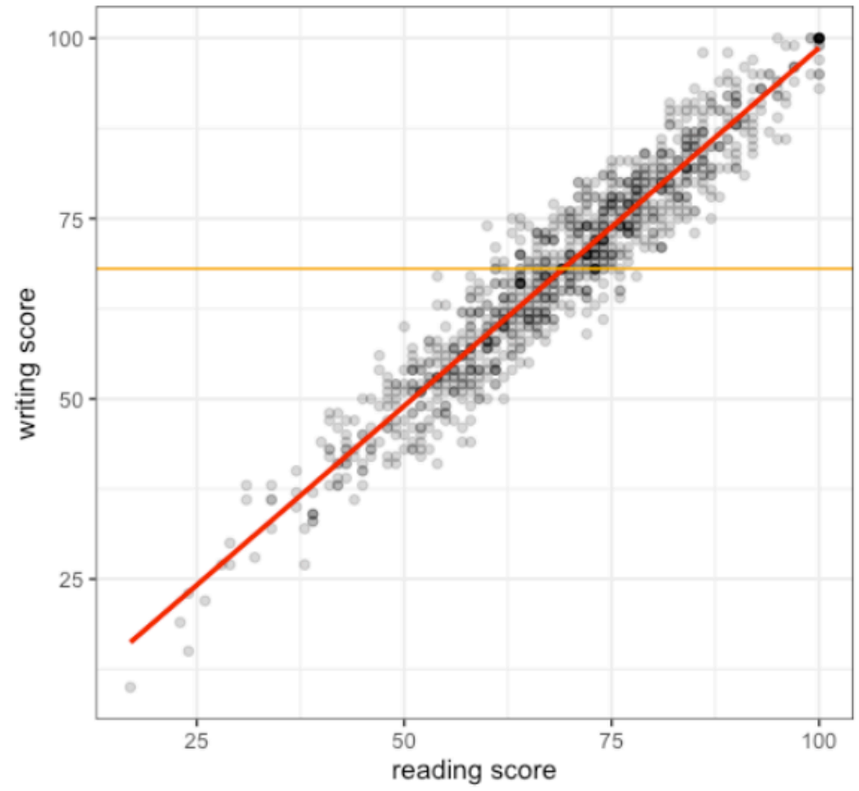


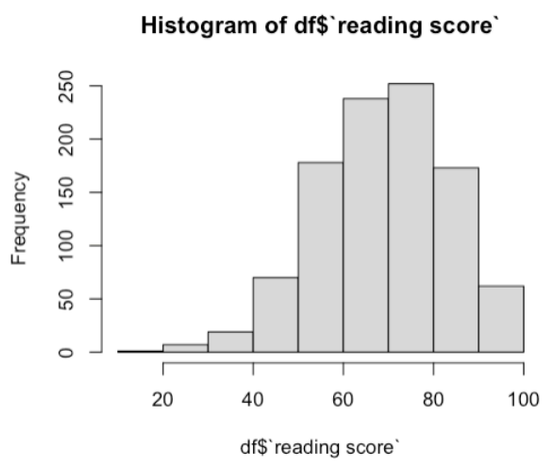


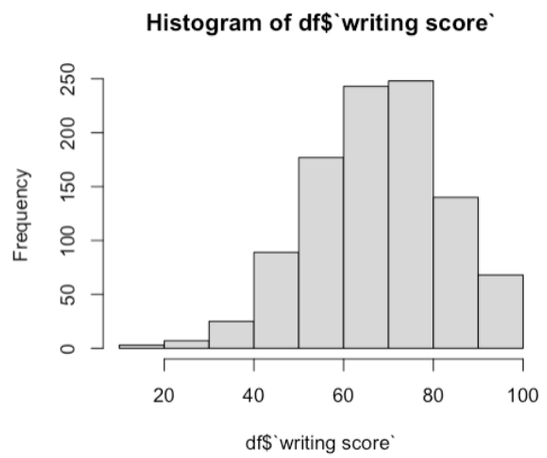
Regression with continuous predictors

```
> summary(lm(`writing score` ~ `reading score`, data = df))

Call:
lm(formula = `writing score` ~ `reading score`, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.667554   0.693792  -0.962    0.336
`reading score`  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113, Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Model fit



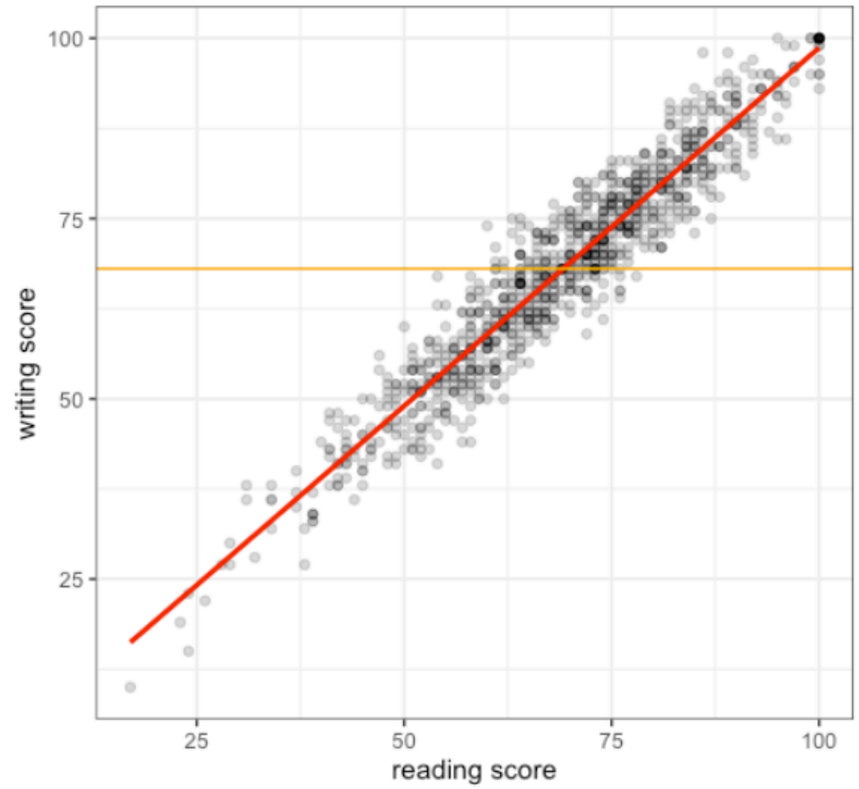




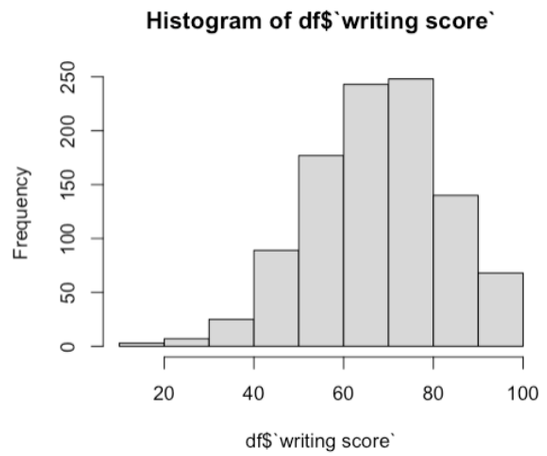
Regression with continuous predictors

```
> summary(lm(`writing score` ~ `reading score`, data = df))

Call:
lm(formula = `writing score` ~ `reading score`, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.667554   0.693792  -0.962    0.336
`reading score`  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113, Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Model fit



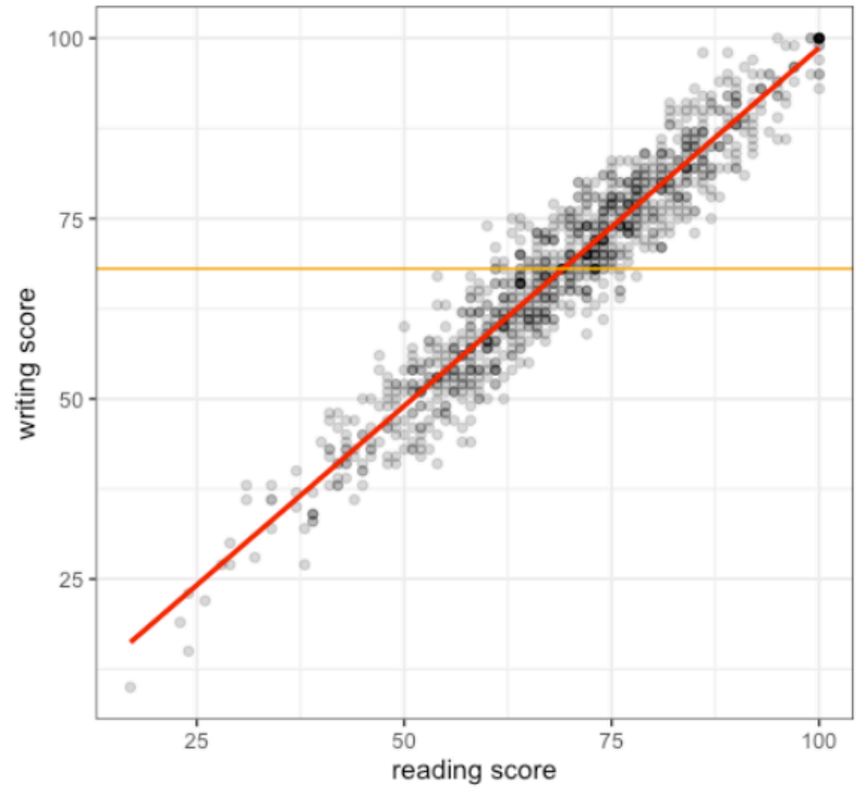

```
> cor.test(df$`writing score`, df$`reading score`)

        Pearson's product-moment correlation

data:  df$`writing score` and df$`reading score`
t = 101.23, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9487506 0.9597921
sample estimates:
      cor
0.9545981
```



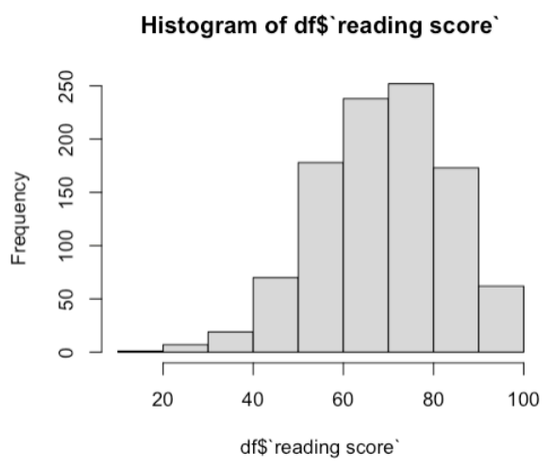
Regression with continuous predictors

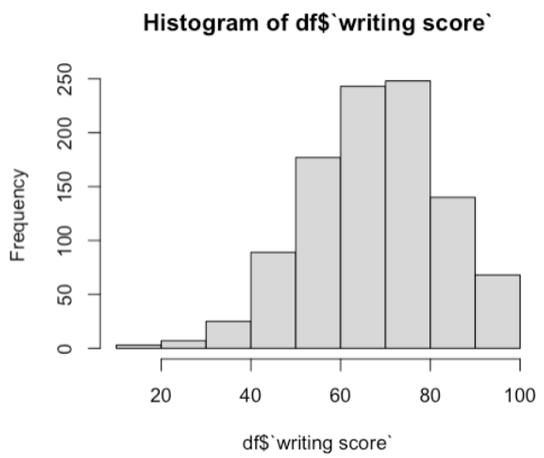
```
> summary(lm(`writing score` ~ `reading score`, data = df))

Call:
lm(formula = `writing score` ~ `reading score`, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.667554   0.693792  -0.962    0.336
`reading score`  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113,  Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Model fit




```
> cor.test(df$`writing score`, df$`reading score`)

        Pearson's product-moment correlation

data:  df$`writing score` and df$`reading score`
t = 101.23, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9487506 0.9597921
sample estimates:
      cor
0.9545981
```



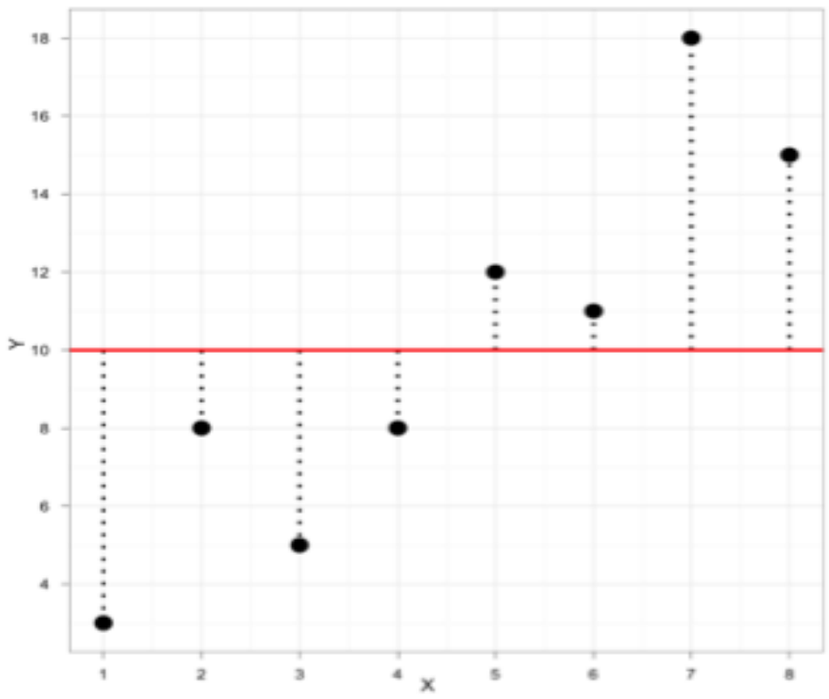


Recap: The model parameters

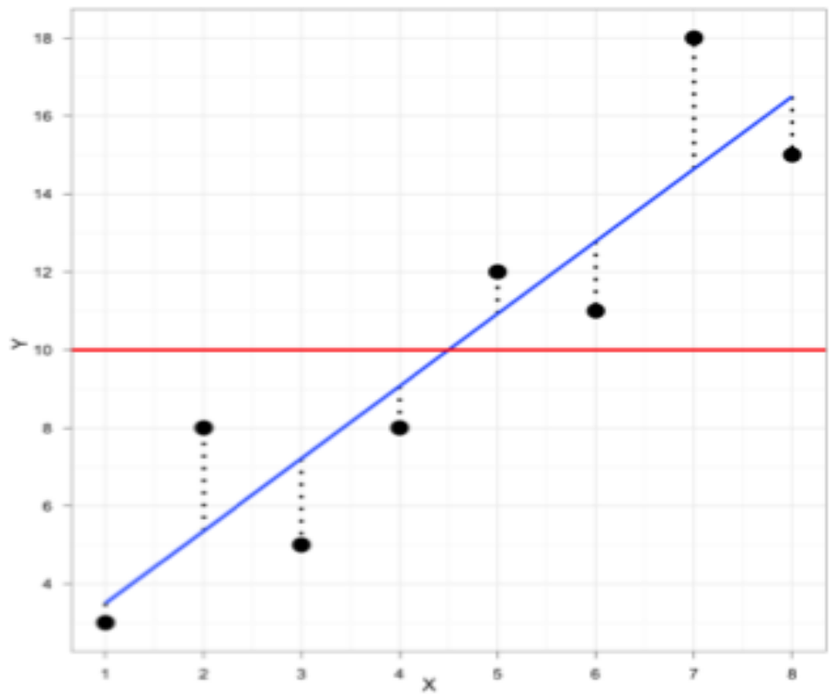
- For each predictor variable in our model, we get:
  - **A $\hat{\beta}_0$ value aka. "intercept":**
    what is the value of Y when X is zero
  - **A $\hat{\beta}_1$ value aka. "slope" or regression coefficient:**
    how much does Y change for each one increment on X
  - **SE of the betas:**
    a measure of how good our estimates of the betas ($\hat{\beta}_0$ and $\hat{\beta}_1$) are in relation to the "true" population value of the betas (β_0 and β_1)
  - **The t-value of beta:**
    how far from zero is beta on a *t* distribution, measured as the ratio between systematic and unsystematic variance
  - **A p-value for the t-value of beta:**
    the probability of the t-value given the degrees of freedom if H_0 is correct

Model fit (1)

- How well does the model fit the data?
- The mean of Y is our best model if X is unknown (“null-model”)
- Does our regression model improve the null model?



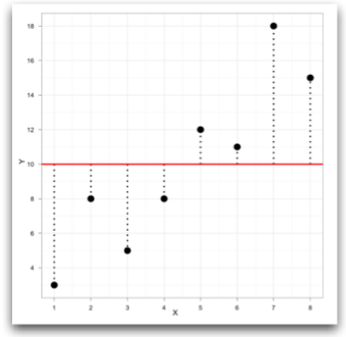
Null model



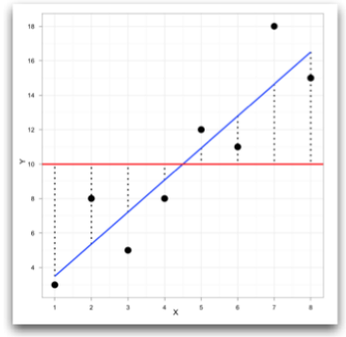
Model with predictor X

Model fit (2)

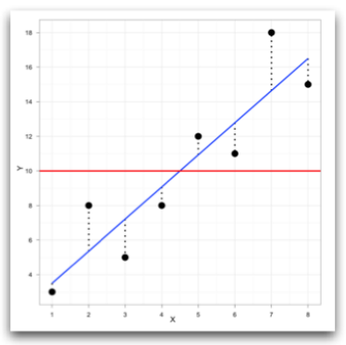
- We compare the regression model to the null-model:
- **Total sum of squares (SS_T):**
  - The ‘error’ of the null-model (variability between scores and the mean)



- **Model sum of squares (SS_M):**
  - The squared difference between the prediction of the model and the null-model
  - The predicted deviation from the mean (once we know X)

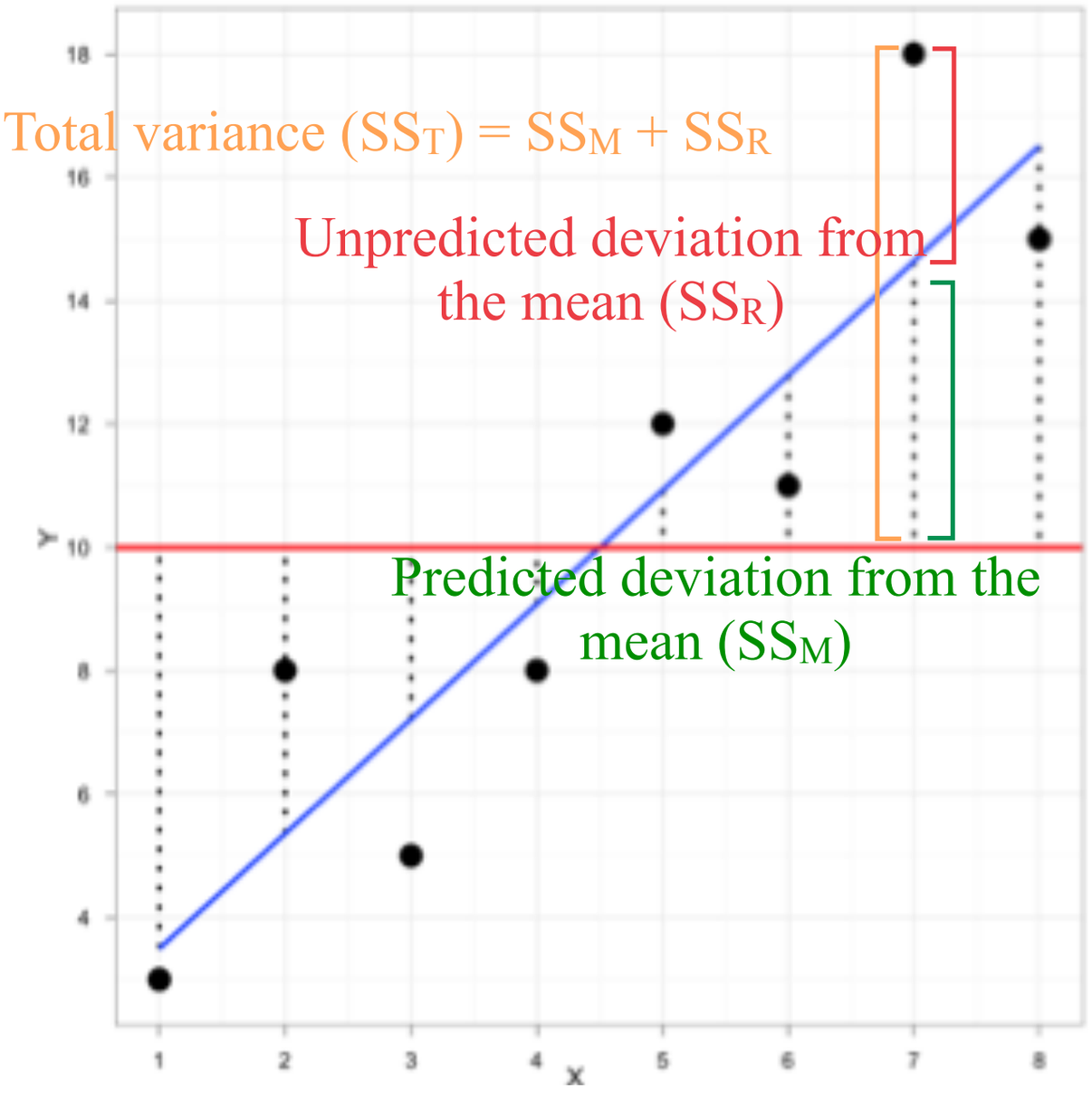


- **Residual sum of squares (SS_R):**
  - The ‘unexpected’ deviation from the model
  - Variability between the regression model and the actual data
  - How well does the regression model account for the data?



Model fit (3)

- E.g.:
- predicted value of Y when X is 7 = 14.6
- observed value = 18
- 18 -14.6 = 3.4 (unexplained variance)



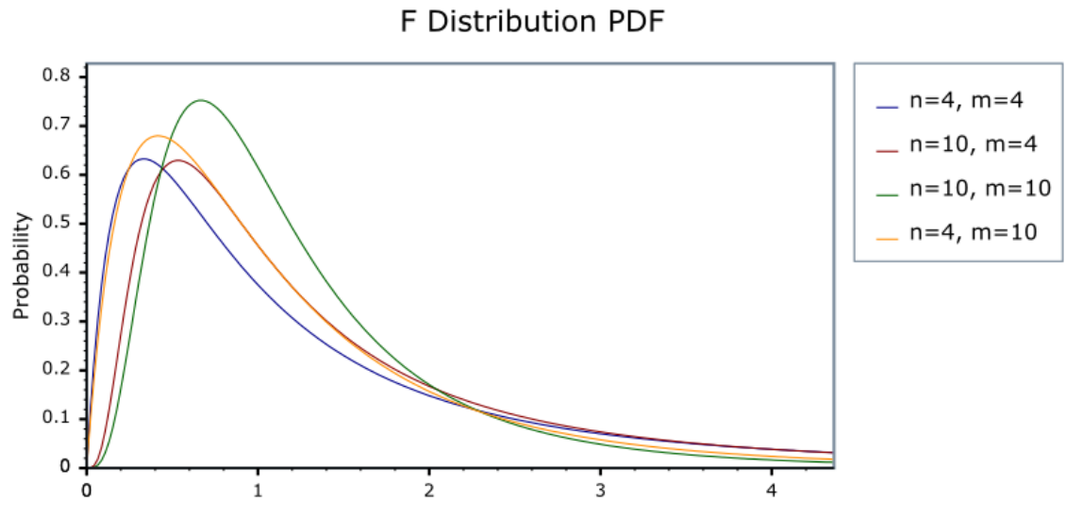
Model fit (4)

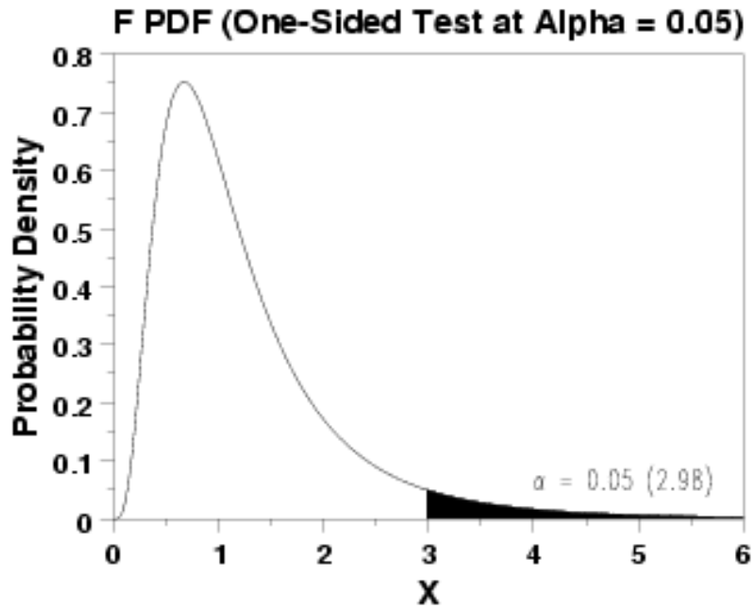
- How much of the total variance in the data is the non-null model capturing?
  - $R^2 = \frac{SS_M}{SS_T}$

- From lecture 5: Coefficient of determination R^2
  - R2(Sarah, Mother) = 0.78 * 0.78 = 0.61
  - →61% of the total variance in our data is explained by the relationship between Sarah and her mom's MLUs

The F-distribution

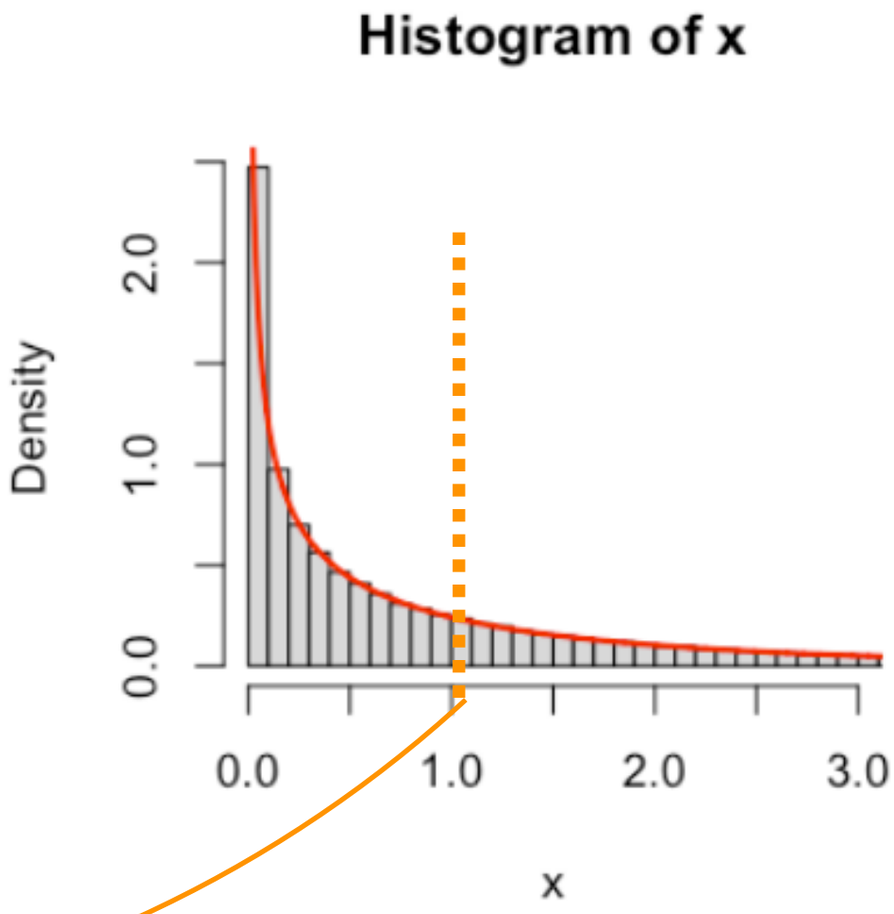
- If model > null model, then SS_M should be high relative to SS_R
- $$F = \frac{SS_M}{SS_R}$$
- *F* value (and *P*(F)): test full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable
- F(n, m), where
  - n = DF for N
  - m = DF for predictors





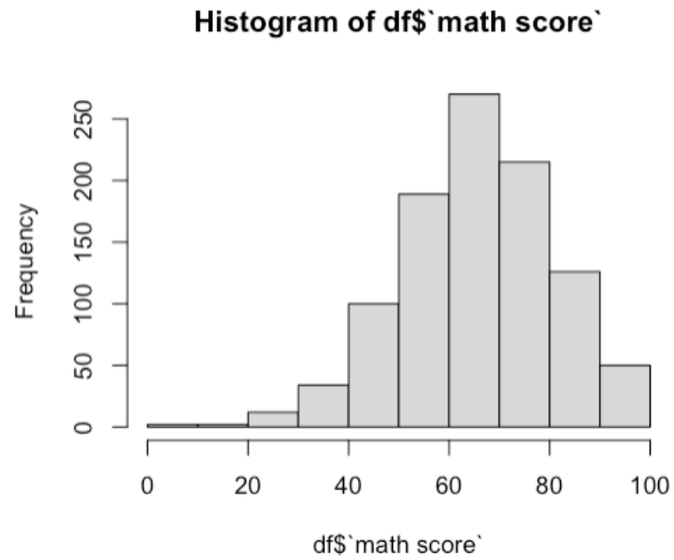
Back to the R output

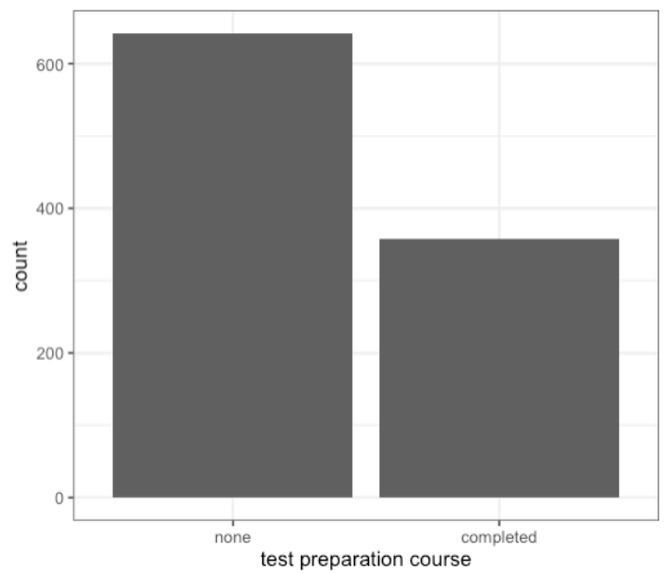
```
> summary(lm(`writing score` ~ `reading score`, data = df))

Call:
lm(formula = `writing score` ~ `reading score`, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9573  -2.9573   0.0363   3.1026  15.0557

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.667554   0.693792  -0.962    0.336
`reading score`  0.993531   0.009814 101.233   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113,   Adjusted R-squared:  0.9112
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```



Regression with categorical predictors

```
> summary(lm(`math score` ~ `test preparation course`, data = df))

Call:
lm(formula = `math score` ~ `test preparation course`, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-64.078 -10.078  -0.078   9.922  35.922

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       64.0779     0.5892 108.752  < 2e-16 ***
`test preparation course`completed   5.6176     0.9848   5.705 1.54e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.93 on 998 degrees of freedom
Multiple R-squared:  0.03158,	Adjusted R-squared:  0.03061
F-statistic: 32.54 on 1 and 998 DF,  p-value: 1.536e-08
```





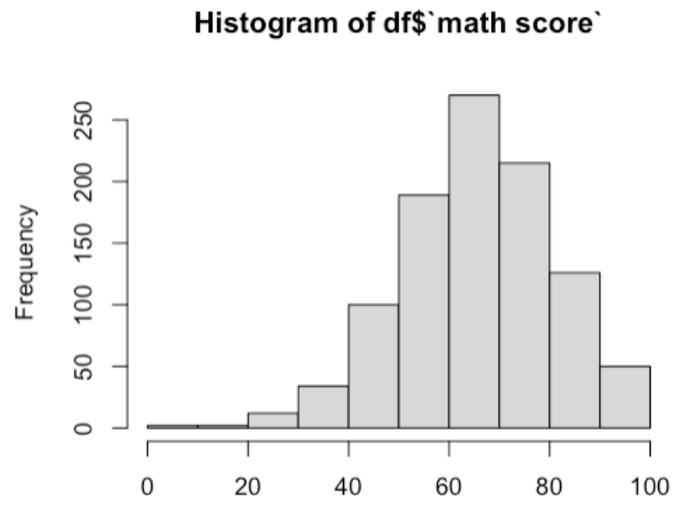
Regression with categorical predictors

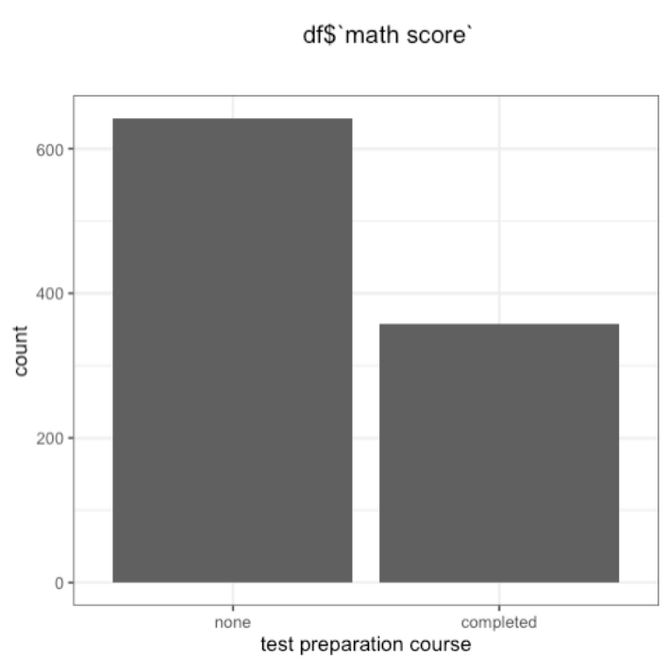
```
> summary(lm(`math score` ~ `test preparation course`, data = df))

Call:
lm(formula = `math score` ~ `test preparation course`, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-64.078 -10.078  -0.078   9.922  35.922

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       64.0779     0.5892 108.752  < 2e-16 ***
`test preparation course`completed   5.6176     0.9848   5.705 1.54e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model parameters

```
Residual standard error: 14.93 on 998 degrees of freedom
Multiple R-squared:  0.03158,Adjusted R-squared:  0.03061
F-statistic: 32.54 on 1 and 998 DF,  p-value: 1.536e-08
```

Model fit





Regression with categorical predictors

```
> summary(lm(`math score` ~ `test preparation course`, data = df))

Call:
lm(formula = `math score` ~ `test preparation course`, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-64.078 -10.078  -0.078   9.922  35.922

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        64.0779     0.5892 108.752  < 2e-16 ***
`test preparation course`completed  5.6176     0.9848   5.705 1.54e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.93 on 998 degrees of freedom
Multiple R-squared:  0.03158,Adjusted R-squared:  0.03061
F-statistic: 32.54 on 1 and 998 DF,  p-value: 1.536e-08
```

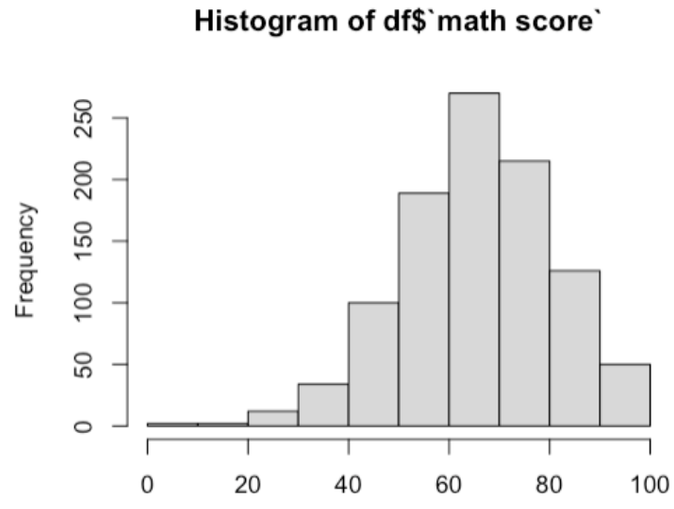
Model parameters

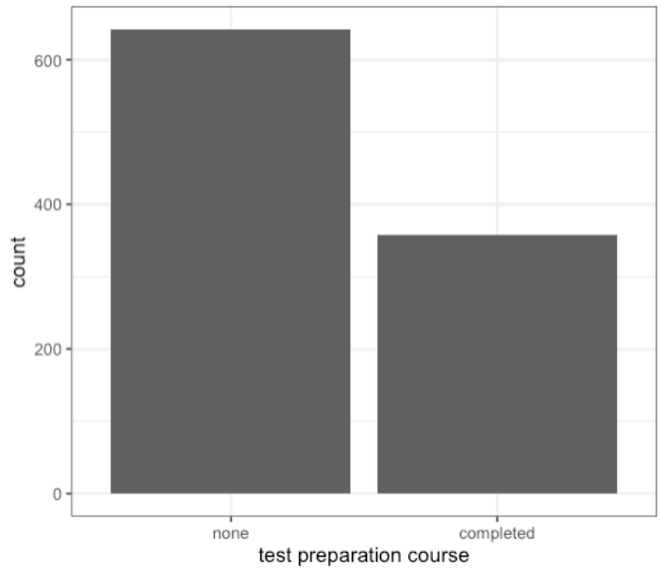
Model fit

Inferential test statistics

$\hat{\beta}_0$

$\hat{\beta}_1$



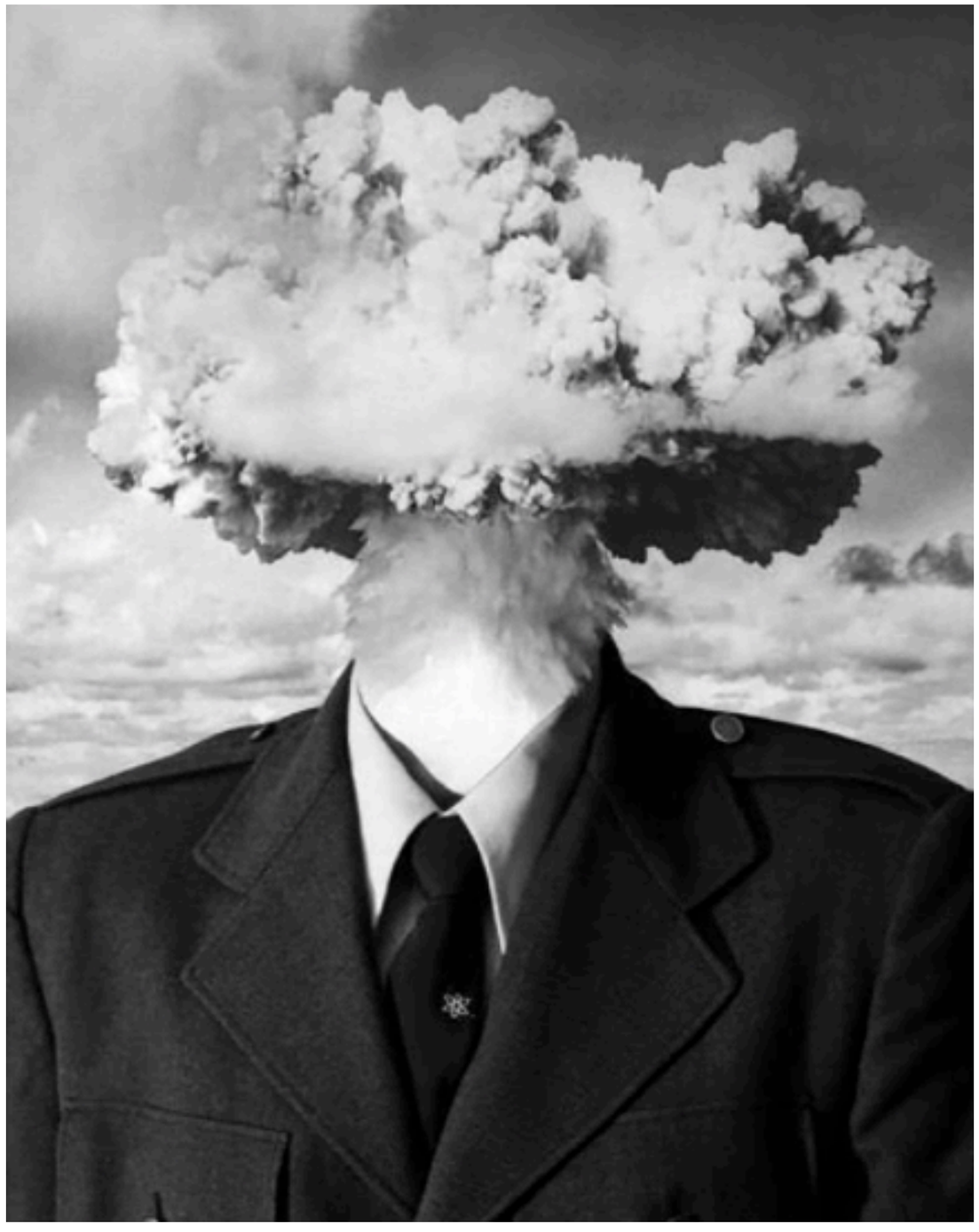


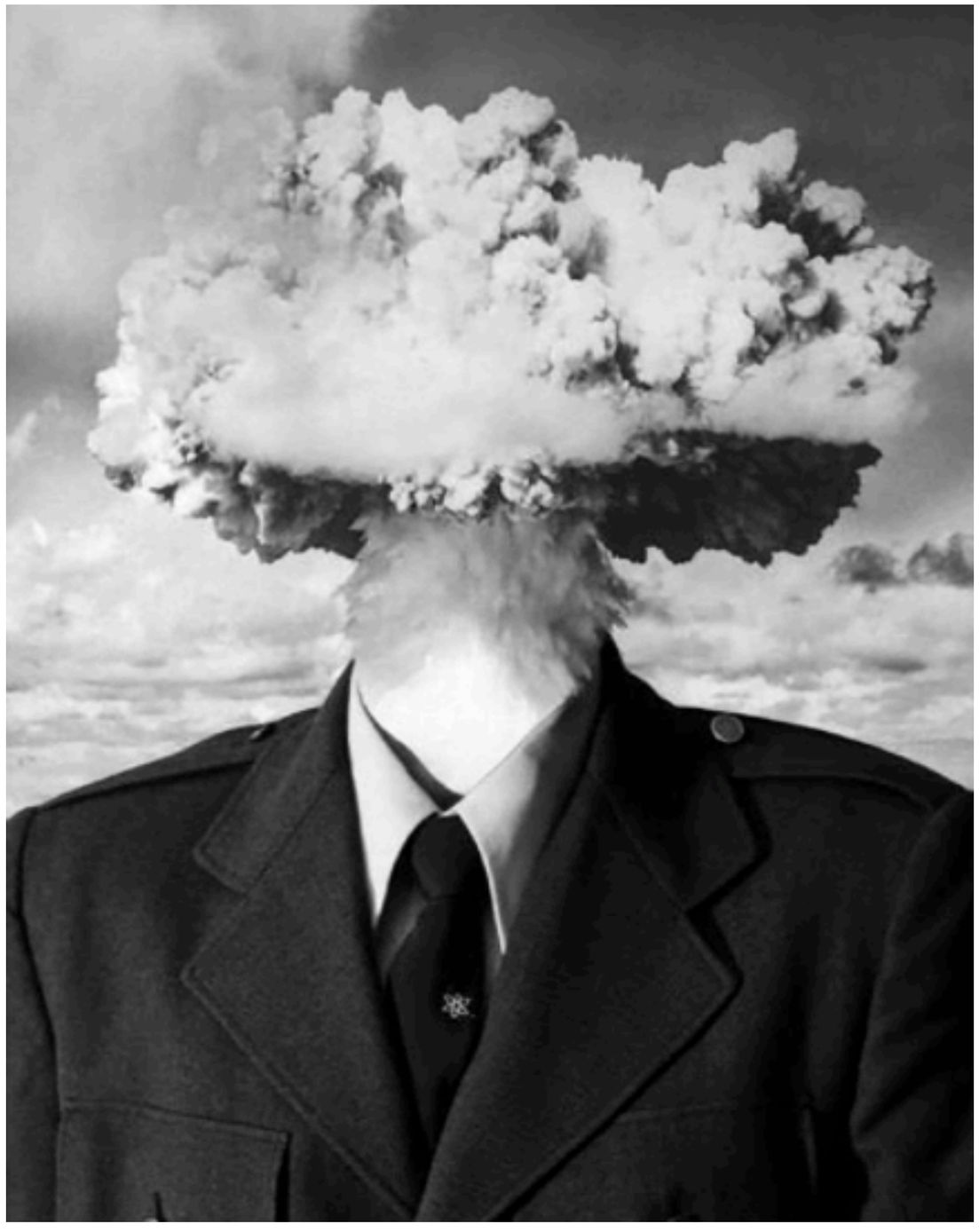
Regression vs. t-test

- When the predictor is a categorical variable with only two levels, regression is equivalent to t-test:

```
> summary(lm(`math score` ~ `test preparation course`, data = df))

Call:
lm(formula = `math score` ~ `test preparation course`, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-64.078 -10.078  -0.078   9.922  35.922

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      64.0779     0.5892 108.752  < 2e-16 ***
`test preparation course`completed  5.6176     0.9848   5.705 1.54e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.93 on 998 degrees of freedom
Multiple R-squared:  0.03158,      Adjusted R-squared:  0.03061
F-statistic: 32.54 on 1 and 998 DF,  p-value: 1.536e-08
```

```
> t.test(`math score` ~ `test preparation course`, data = df)

        Welch Two Sample t-test

data:  math score by test preparation course
t = -5.787, df = 770.08, p-value = 1.043e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.523257 -3.712041
sample estimates:
     mean in group none mean in group completed
               64.07788                69.69553
```





Reporting the results

- Beta values answer our research question directly:
    - "Age significantly predicts the number of words the child understands in early childhood, β = 26.82, SE = 0.68, t = 38.93, p < .0001"
- If comparing two contrasting models, it's better to report model fit statistics:
    - The number of words known by the child in early childhood is well predicted by age, $F(1, 1839) = 1516$, $p < .0001$, adjusted $R^2 = .45$"

Thursday

- Chilling out with linear regression

- We will learn how to run linear regression models, how to understand the outputs, and finally how to interpret our findings