Week 9: Linear Regression II

POP88162 Introduction to Quantitative Research Methods

Tom Paskhalis

Department of Political Science, Trinity College Dublin

Topics for Today

- Measure of fit, R^2
- Binary independent variables
- Multiple linear regression model
- Log transformation

Previously…

Review: Linear Regression Model

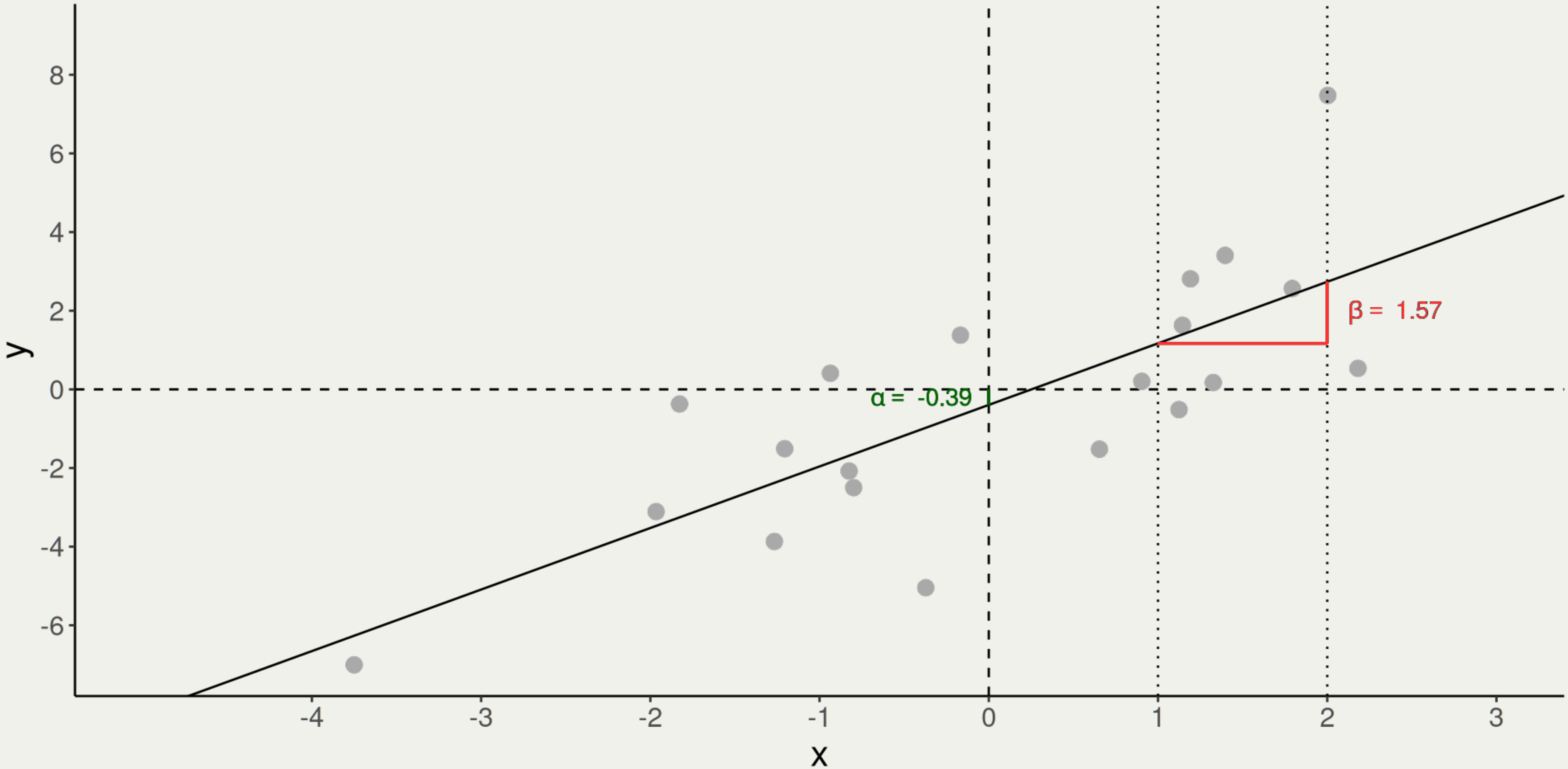
- We can express linear relationship (association) between two continuous variables with **bivariate linear regression model**:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where:

- Observations $i = 1, \ldots, n$
- Y is the dependent variable
- X is the independent variable
- α is the **intercept** or **constant**
- β is the **slope**
- ϵ_i is the **error term**

Review: Intercept and Slope



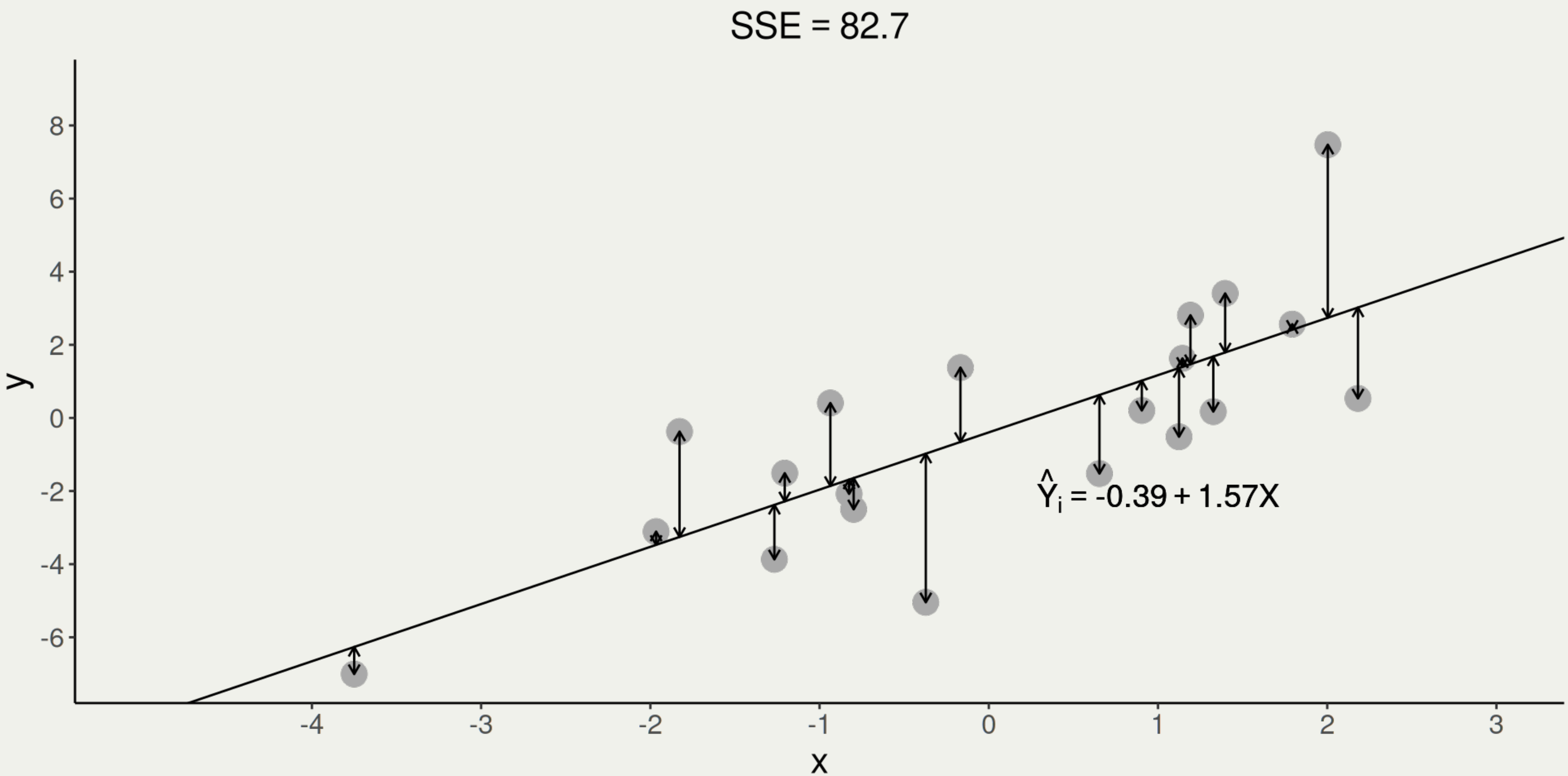
Review: Ordinary Least Squares

- The most common method of estimating the parameters of the linear regression model is the **ordinary least squares (OLS)** method.
- To pick the line that best fits the data OLS minimises the **sum of squared errors (SSE)**.
- This is the sum of squared differences between the actual values of each observation Y_i and the predicted value $\hat{Y}_i$.
- More formally a line that minisises the following expression is chosen:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\alpha} + \hat{\beta}X_i))^2$$

- SSE is also often called the **residual sum of squares (RSS)**.

Review: Geometric Interpretation



Review: t Test

- The test statistic for a single regression coefficient is:

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}}$$

where:

- $\hat{\beta}$ - is the estimated slope (coefficient)
- β_{H_0} - is the slope under H_0
- $\hat{\sigma}_{\hat{\beta}}$ - is the *standard error* of $\hat{\beta}$

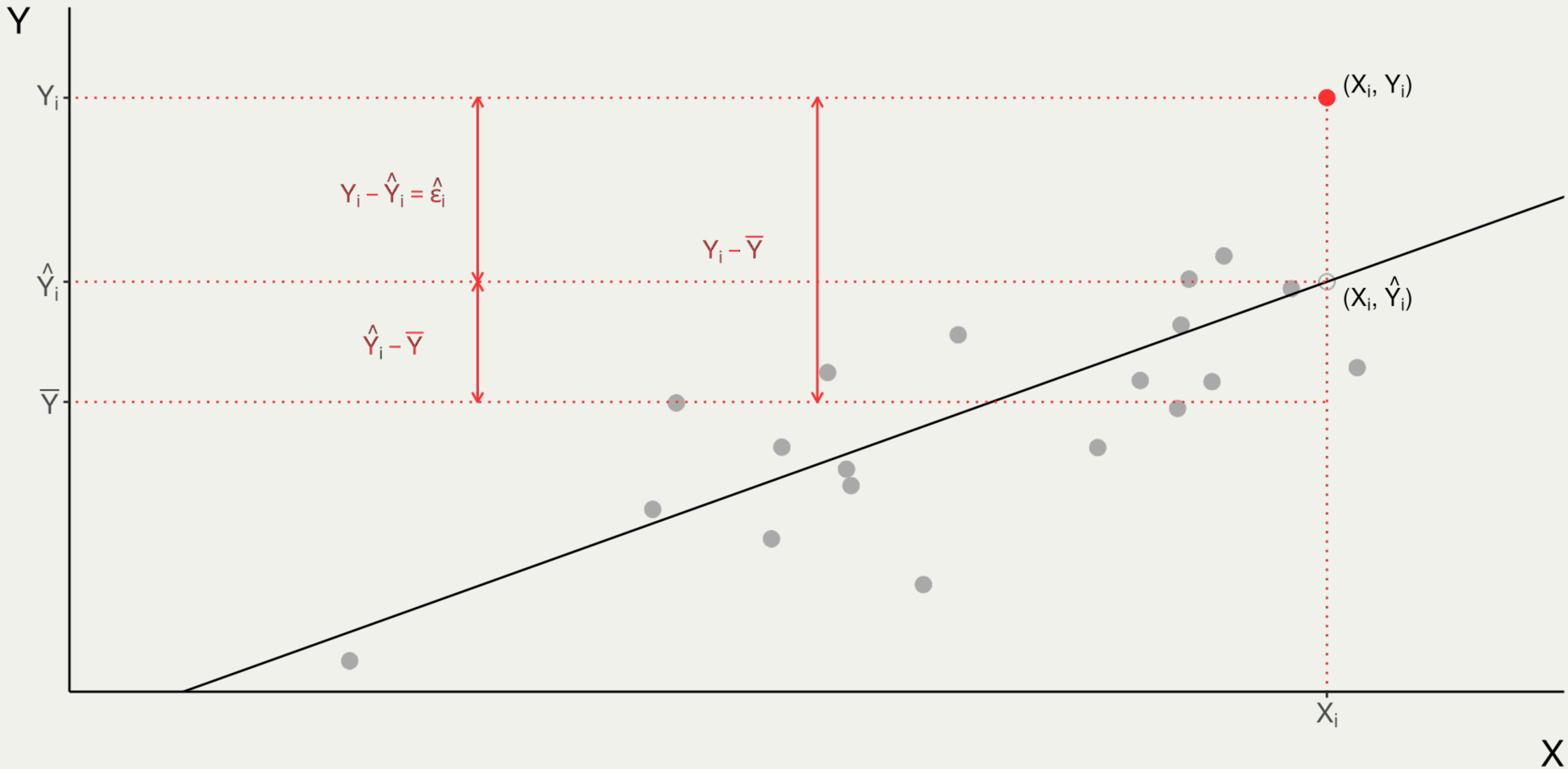
Note that in the very common case the null hypothesis is $\beta_{H_0} = 0$ the t-statistic simplifies to $t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$

Model Fit

Measure of Fit

- Apart from testing individual coefficients,
- We might want to assess the performance of our statistical model more broadly.
- **Model fit** (or **model performance**):
  - How tightly are the observations clustered around the line?
  - What proportion of the variation in the dependent variable can be explained by the independent variable?

Variation in Y



Total Variation in Y

- Previous plot graphically demonstrates that:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

- Squaring both sides of the equation and summing over all observations:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Total Sum of Squares

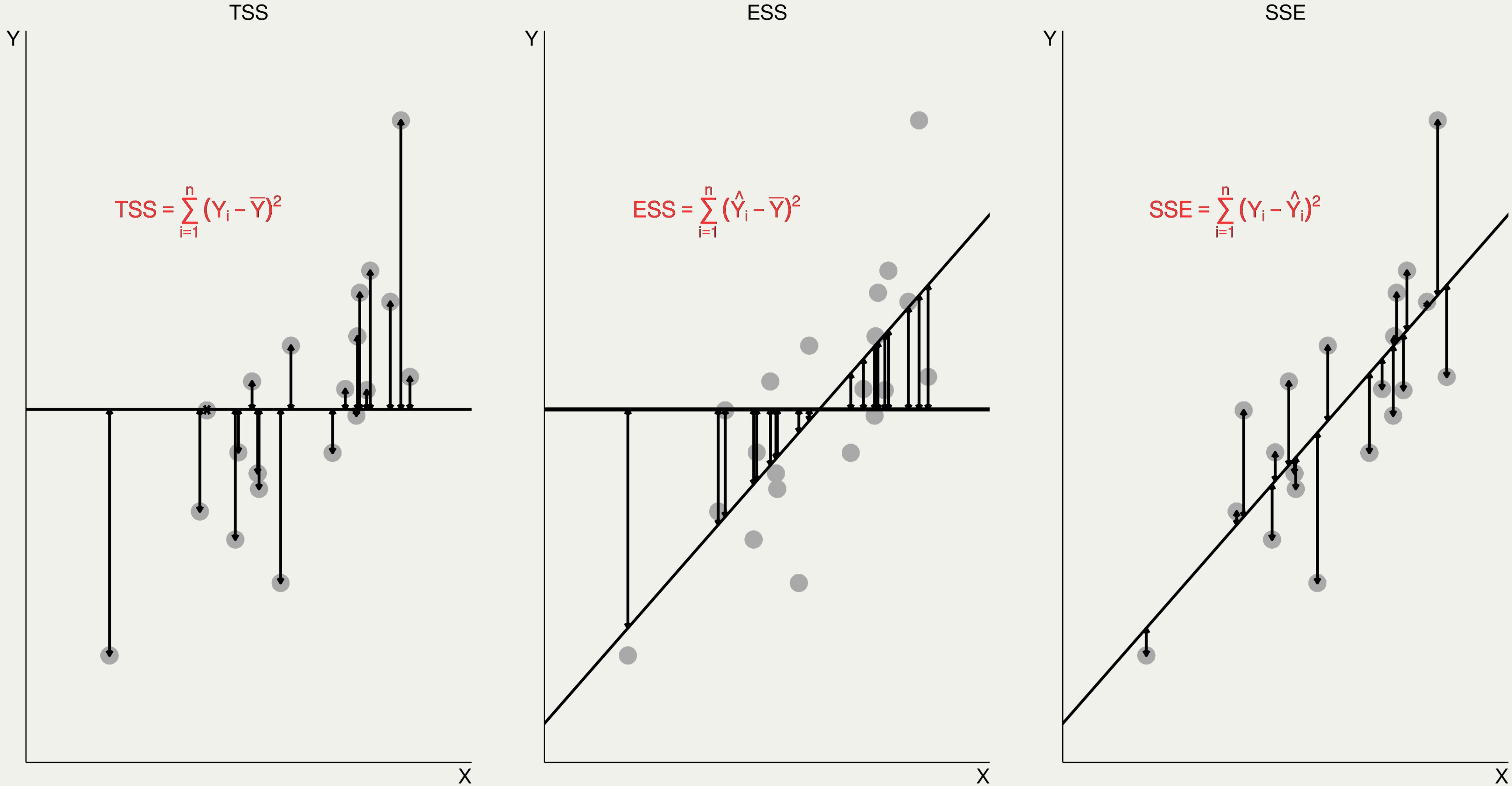
- Substituting each of the sums, the total variation in Y can be decomposed into:

$$TSS = SSE + ESS$$

where:

- TSS (**Total Sum of Squares**) equals $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- SSE (**Sum of Squared Errors**) equals $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- ESS (**Explained Sum of Squares**) equals $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

TSS vs ESS vs SSE



Model Fit

- R^2 (pronounced R-squared) - **coefficient of determination**.
- Provides a one number summary of model fit.
- Measure of the **proportional reduction in error** by the model.
- Proportional reduction relative to what?
  - Baseline prediction: $\bar{Y}$
  - Baseline prediction error: $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Model prediction: $\hat{Y}_i$
  - Model prediction error: $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - $TSS - SSE$ - reduction in prediction error by the model

R^2

- Thus R^2 can be expressed as:

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{ESS}{TSS} = 1 - \frac{SSE}{TSS}$$

- It shows the proportion of variation of Y explained by X.
- As with r^2 for correlation R^2 ranges between 0 and 1.
- If X explains all the variation in Y, then $R^2 = 1$.
- If X explains none of the variation in Y, then $R^2 = 0$.
- For simple bivariate linear regression model R is the absolute value of the correlation coefficient r.

Example: R^2

```r
lm_fit <- lm(gdp_per_capita ~ democracy_duration, data = democracy_gdp_2020)
```

```r
summary(lm_fit)
```

```

Call:
lm(formula = gdp_per_capita ~ democracy_duration, data = democracy_gdp_2020)

Residuals:
   Min     1Q Median     3Q    Max 
-44806  -8756  -4944   4820 163717 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         5051.44    2370.78   2.131   0.0345 *  
democracy_duration   182.22      35.15   5.185 5.99e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20900 on 173 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1345,	Adjusted R-squared:  0.1295 
F-statistic: 26.88 on 1 and 173 DF,  p-value: 5.995e-07
```

Overfitting

- Note that *coefficient of determination* provides an **in-sample** measure of fit.
    - I.e. R^2 shows how well your model predicts the data used to estimate it.
- It might tell you very little (or be outright misleading!) about **out-of-sample** fit:
    - Poor prediction (fit) to the new data.

Binary Predictors

Review: Difference in Means

- RQ: Do democracies last longer than autocracies?
- Data: political regimes in 2020
- Dependent variable (Y): Years of existence of the current regime (interval-scale)
- Independent variable (X): Type of political regime, autocracy/democracy (nominal-scale)
- Quantity of interest:

$$\bar{Y}_{X=autocracy} - \bar{Y}_{X=democracy} = \bar{Y}_{X=0} - \bar{Y}_{X=1}$$

Review: Inference for Difference in Means

- Statistical test:

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{se_{\bar{Y}_{X=0} - \bar{Y}_{X=1}}} = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{\sqrt{\frac{s^2_{X=0}}{n_{X=0}} + \frac{s^2_{X=1}}{n_{X=1}}}}$$

- First, find the difference in mean longevity between autocracies and democracies:

$$\bar{Y}_{X=0} - \bar{Y}_{X=1} = 54.61 - 45.05 = 9.56$$

Review: Inference for Difference in Means

- Second, calculate the standard error of the difference in the two means:

$$se_{\bar{Y}_{X=0}-\bar{Y}_{X=1}} \approx \sqrt{\frac{s^2_{X=0}}{n_{X=0}} + \frac{s^2_{X=1}}{n_{X=1}}} = \sqrt{\frac{2498.004}{77} + \frac{1521.604}{118}} = 6.73$$

- Third, conduct the statistical test by dividing the difference in means between two groups by the standard error:

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{\sqrt{\frac{s^2_{X=0}}{n_{X=0}} + \frac{s^2_{X=1}}{n_{X=1}}}} \approx \frac{9.56}{6.73} = 1.42$$

Review: Hypothesis Testing for Difference in Means

```r
(1 - pnorm(1.42)) * 2
```

```
[1] 0.1556077
```

```r
t.test(democracy_gdp_2020$democracy_duration ~ democracy_gdp_2020$democracy)
```

```

	Welch Two Sample t-test

data:  democracy_gdp_2020$democracy_duration by democracy_gdp_2020$democracy
t = 1.4198, df = 134.61, p-value = 0.158
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -3.75709 22.87617
sample estimates:
mean in group 0 mean in group 1 
       54.61039        45.05085 
```

Linear Regression with Binary X Variable

- Linear regression is a much more flexible tool than just measuring the association between two quantitative variables:
  - Y should always be (roughly) interval-scale and normally distributed
  - X can be essentially any level of measurement
- When X is **binary** (**dichotomous**) the estimate of β is, essentially, equivalent to difference-in-means estimate.
- **Binary** variables are nominal (categorical) variables that $= 1$ when an observation has a specific trait and $= 0$ otherwise.

Regression with Binary X

- Consider bivariate linear regression model where X is binary ($X_i = 0$ or 1):

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- When $X_i = 0$, we have $Y_i = \alpha + \epsilon_i$
  - The expected value (mean) of Y_i is α
- When $X_i = 1$, we have $Y_i = \alpha + \beta + \epsilon_i$
  - The expected value (mean) of Y_i is $\alpha + \beta$
- β represents the difference in the population means:
  - $\alpha - (\alpha + \beta) = -\beta$
- $\hat{\beta}$ is the estimated difference between the sample averages of Y_i in the two groups.

Example: Binary X

- Now let's estimate our $Democracy \rightarrow Longevity$ model in R:

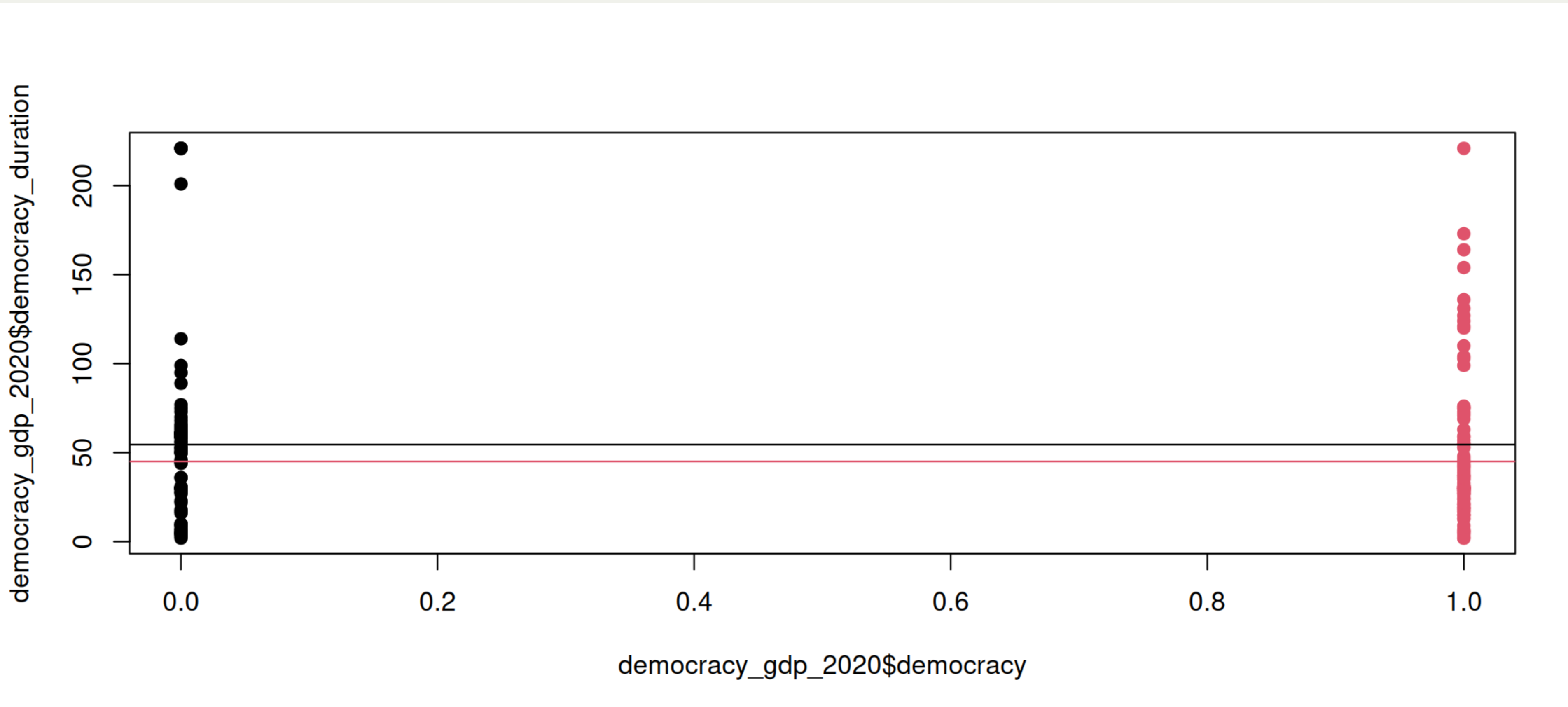
$$\widehat{Longevity}_i = \hat{\alpha} + \hat{\beta} Democracy_i$$

```r
lm_fit_1 <- lm(democracy_duration ~ democracy, data = democracy_gdp_2020)
summary(lm_fit_1)
```

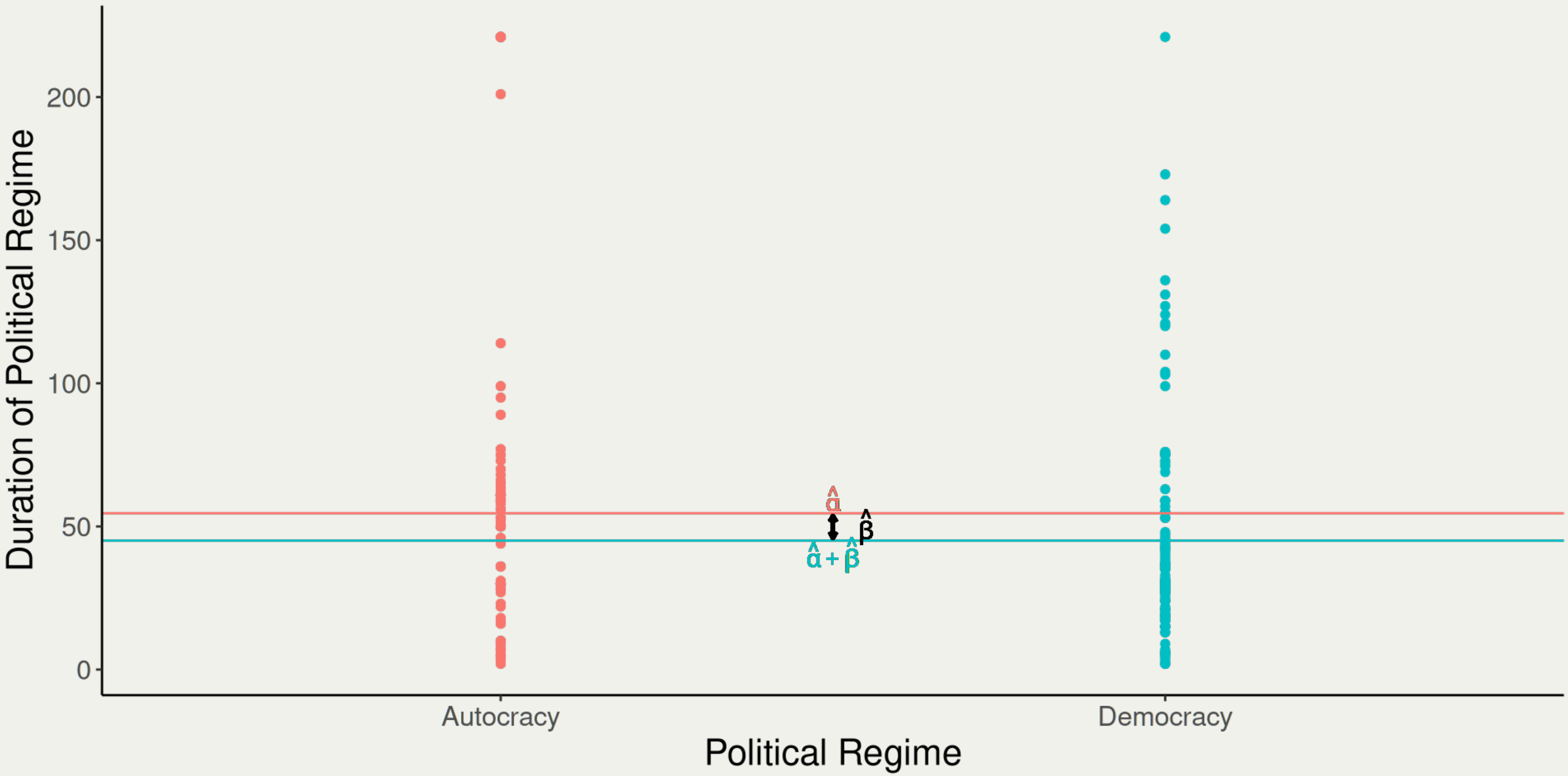
```

Call:
lm(formula = democracy_duration ~ democracy, data = democracy_gdp_2020)

Residuals:
    Min      1Q  Median      3Q     Max 
-52.610 -24.610 -10.051   7.949 175.949 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   54.610      4.975  10.976   <2e-16 ***
democracy     -9.560      6.396  -1.495    0.137    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.66 on 193 degrees of freedom
Multiple R-squared:  0.01144,	Adjusted R-squared:  0.00632 
F-statistic: 2.234 on 1 and 193 DF,  p-value: 0.1366
```

Example: Plotting Regression with Binary X

Plot | Code



Graphical Interpretation of Regression with Binary X



Interpretation of Linear Regression with Binary X

- What is a one-unit increase when X can only be 0 or 1?
  - It is simply the difference between $\bar{Y}_{X=0}$ and $\bar{Y}_{X=1}$
  - Here it is the difference in the expected longevity of autocracies ($X = 0$) and democracies ($X = 1$)
  - In 2020 democratic regimes, on average, tended to last 9.56 fewer years than autocratic.
- What is the expected value of Y when $X = 0$?
  - It is just the mean $\bar{Y}_{X=0}$.
  - Here it is the expected longevity (mean) of autocracies ($X = 0$)
  - In 2020 autocratic regimes, on average, tend to last 54.61 years.

Multiple Predictors

Multiple Linear Regression Model

- We can express the population **multiple (multivariate) linear regression model** as:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

where:

- Observations $i = 1, \ldots, n$
- Y is the dependent variable
- $X_{1i}, \ldots, X_{ki}$ are k independent variables
- α is the **intercept** or **constant**
- $\beta_1, \ldots, \beta_k$ are **coefficients**
- ϵ_i is the **error term**

Modelling the Expected Value

- Multiple linear regression model can alternatively be expressed as:

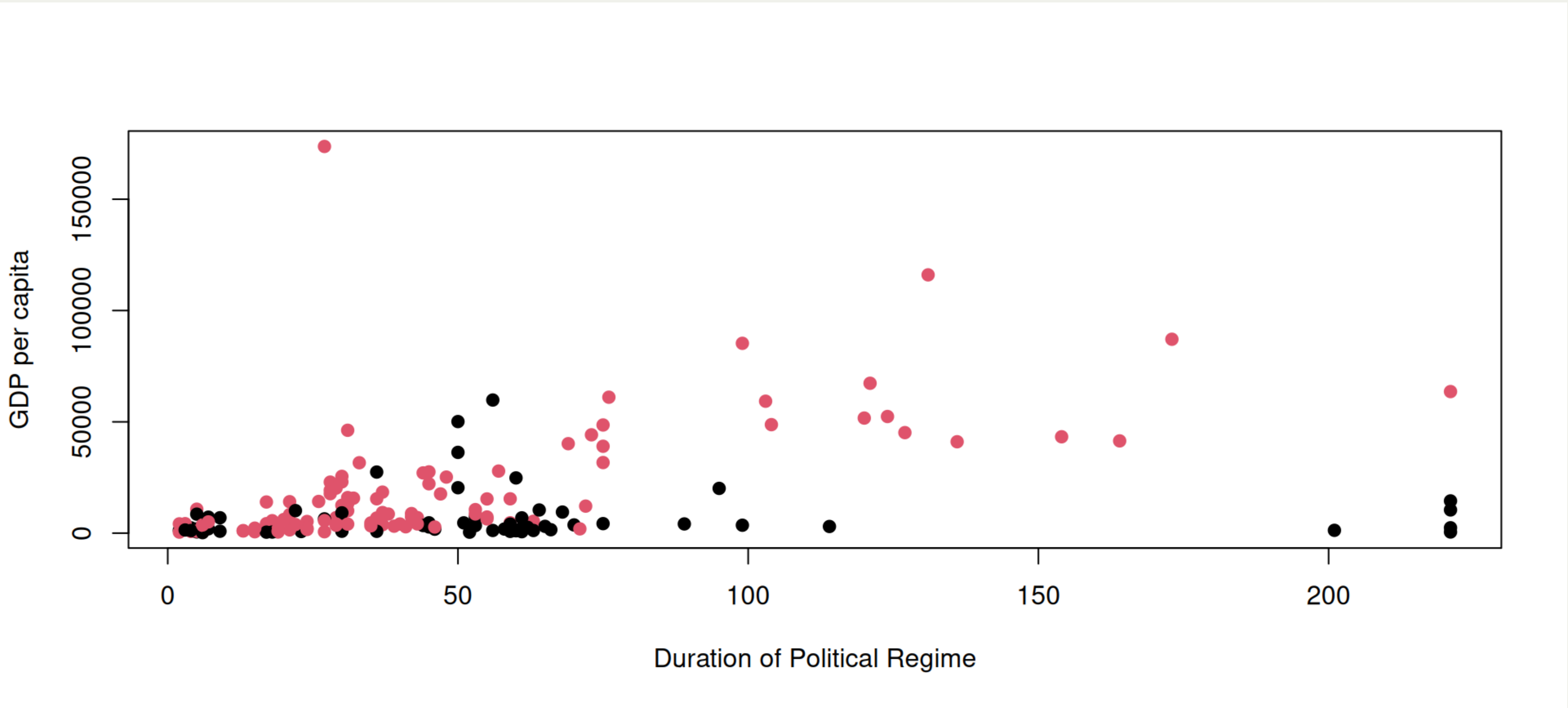
$$E(Y_i|X_{1i}, \dots, X_{ki}) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$$

where:

- $E(Y_i|X_{1i}, \dots, X_{ki})$ is the **conditional expected value** of Y_i given corresponding values of $X_{1i}, \dots, X_{ki}$ variables.
- This formulation emphasises that what we are modelling is the *mean* or *expected value* of our dependent variable for different value of our explanatory variables.

Example: Regime Longevity and GDP in 2020

Plot Code



Example: Regression Equation

- Here, the population regression model would be:

$$GDP_i = \alpha + \beta_1 Longevity_i + \beta_2 Democracy_i + \epsilon_i$$

- And our estimated model is:

$$\widehat{GDP_i} = \hat{\alpha} + \hat{\beta}_1 Longevity_i + \hat{\beta}_2 Democracy_i$$

Estimating Multiple Regression

- As with simple (bivariate) linear regression model, we estimate α, β_1 and β_2 by minimising the *sum of squared errors (SSE)*.
- We can re-write the formula for SSE from bivariate model to accommodate multiple independent variables:

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}))^2$$

- As before, ordinary least squares (OLS) method finds the $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$

Example: Estimating Multiple Regression

- Now let's estimate our $Democracy + Longevity \rightarrow GDP$ model in R:

$$\widehat{GDP_i} = \hat{\alpha} + \hat{\beta}_1 Democracy_i + \hat{\beta}_2 Longevity_i$$

```r
# Note the formula syntax: Y ~ X_1 + X_2
lm_fit_2 <- lm(gdp_per_capita ~ democracy + democracy_duration, data = democracy_gdp_2020)
lm_fit_2
```

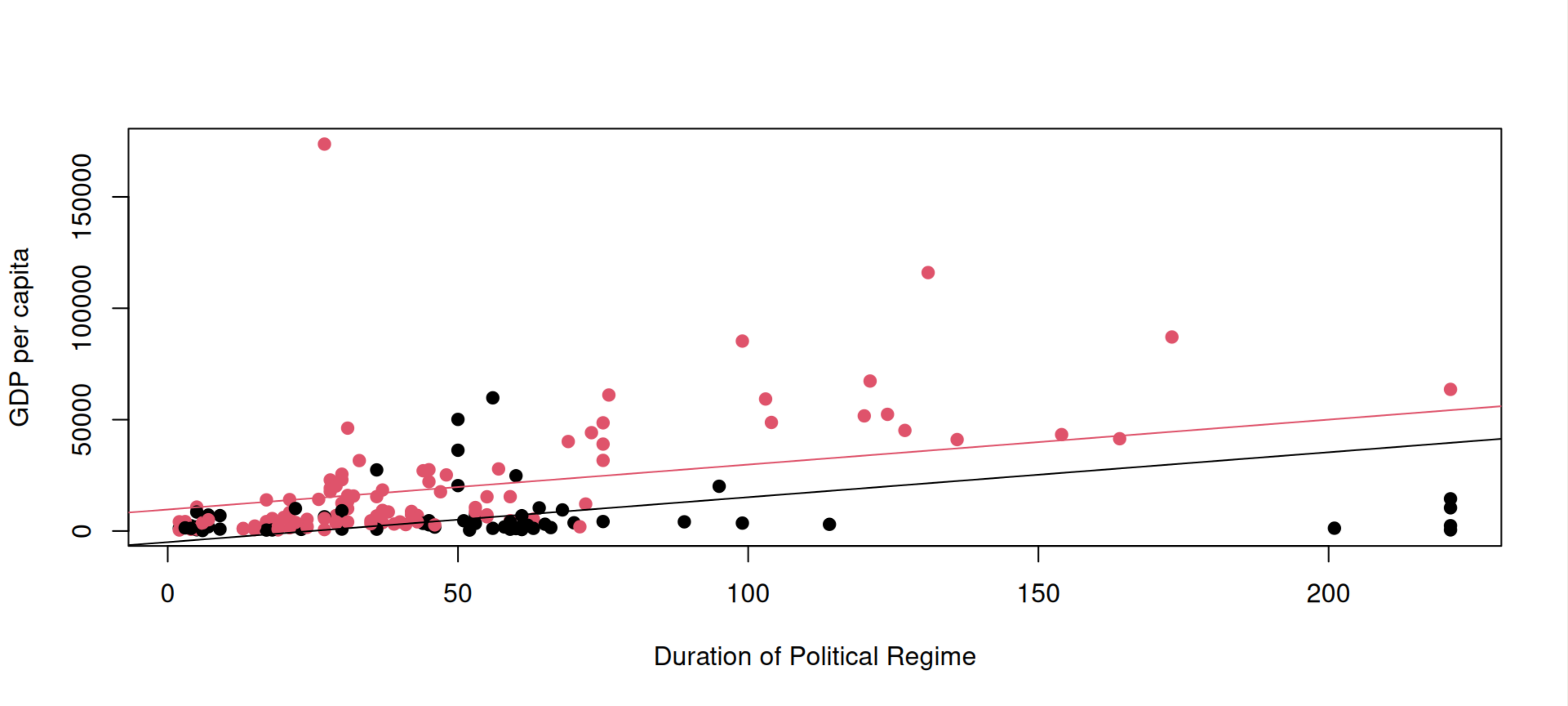
```

Call:
lm(formula = gdp_per_capita ~ democracy + democracy_duration,
    data = democracy_gdp_2020)

Coefficients:
       (Intercept)           democracy  democracy_duration
           -4971.8             14649.7               201.8
```

- In other words our OLS estimate of this model is:

$$\widehat{GDP_i} = -4971.8 + 14649.7 \times Democracy_i + 201.8 \times Longevity_i$$

Example: Plotting Fitted Multiple Regression

Plot | Code



Example: Interpreting Multiple Regression

- Let's interpret our fitted model:

$$\widehat{GDP_i} = -4971.8 + 14649.7 \times Democracy_i + 201.8 \times Longevity_i$$

- $\hat{\alpha} = -4971.8$ - the expected GDP per capita for an autocratic state where a political regime lasted 0 years is -4971.8 USD.
- $\hat{\beta}_1 = 14649.7$ - democratic political regimes are associated with a 14649.7 USD increase in GDP per capita, controlling for regime longevity.
- $\hat{\beta}_2 = 201.8$ - each additional year of political regime's longevity, on average, is associated with a 201.8 USD increase in GDP per capita, holding political regime constant.

Example: Significance Testing for Multiple Regression

```r
# Use `summary()` function to get a more detailed output about the fitted model
summary(lm_fit_2)
```

```

Call:
lm(formula = gdp_per_capita ~ democracy + democracy_duration, 
    data = democracy_gdp_2020)

Residuals:
   Min     1Q Median     3Q    Max 
-39100 -10196  -4907   5437 158563 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -4971.8     3076.8  -1.616    0.108    
democracy           14649.7     3089.0   4.742 4.41e-06 ***
democracy_duration    201.8       33.4   6.040 9.26e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19710 on 172 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.2346,	Adjusted R-squared:  0.2257 
```

Model Fit with Multiple Predictors

R^2 for Multiple Regression

- R^2 for the multiple linear regression model:
  - proportion of variation in Y explained by the model with variables $X_1, \ldots, X_k$
- But it mechanically increases as you add more variables to the model, which can result in overfitting.
- Implication:
  - Picking the model with the highest R^2 might be problematic
- Solution:
  - Penalise models with more explanatory variables

Adjusted R^2 for Multiple Regression

- Adjusted R^2 decreases regular R^2 for each additional predictor:

$$\text{Adjusted } R^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE}{TSS}$$

- Adjusted R^2 goes down if an added explanatory variable doesn't help predict.
- Adjusted R^2 is always smaller than regular R^2.
- We can interpret adjusted R^2 as the proportion of variation in Y explained by the model with variables $X_1, \ldots, X_k$, adjusted for the number of predictors.

Example: Adjusted R^2

```r
summary(lm_fit_2)
```

```
Call:
lm(formula = gdp_per_capita ~ democracy + democracy_duration,
    data = democracy_gdp_2020)

Residuals:
   Min     1Q Median     3Q    Max
-39100 -10196  -4907   5437 158563

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4971.8     3076.8  -1.616    0.108
democracy           14649.7     3089.0   4.742 4.41e-06 ***
democracy_duration    201.8       33.4   6.040 9.26e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19710 on 172 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.2346,    Adjusted R-squared:  0.2257
```

```r
# Note that the output of the summary() called on the regression model
# is just an R list.
summary(lm_fit_2)$r.squared
```

```
[1] 0.2345686
```

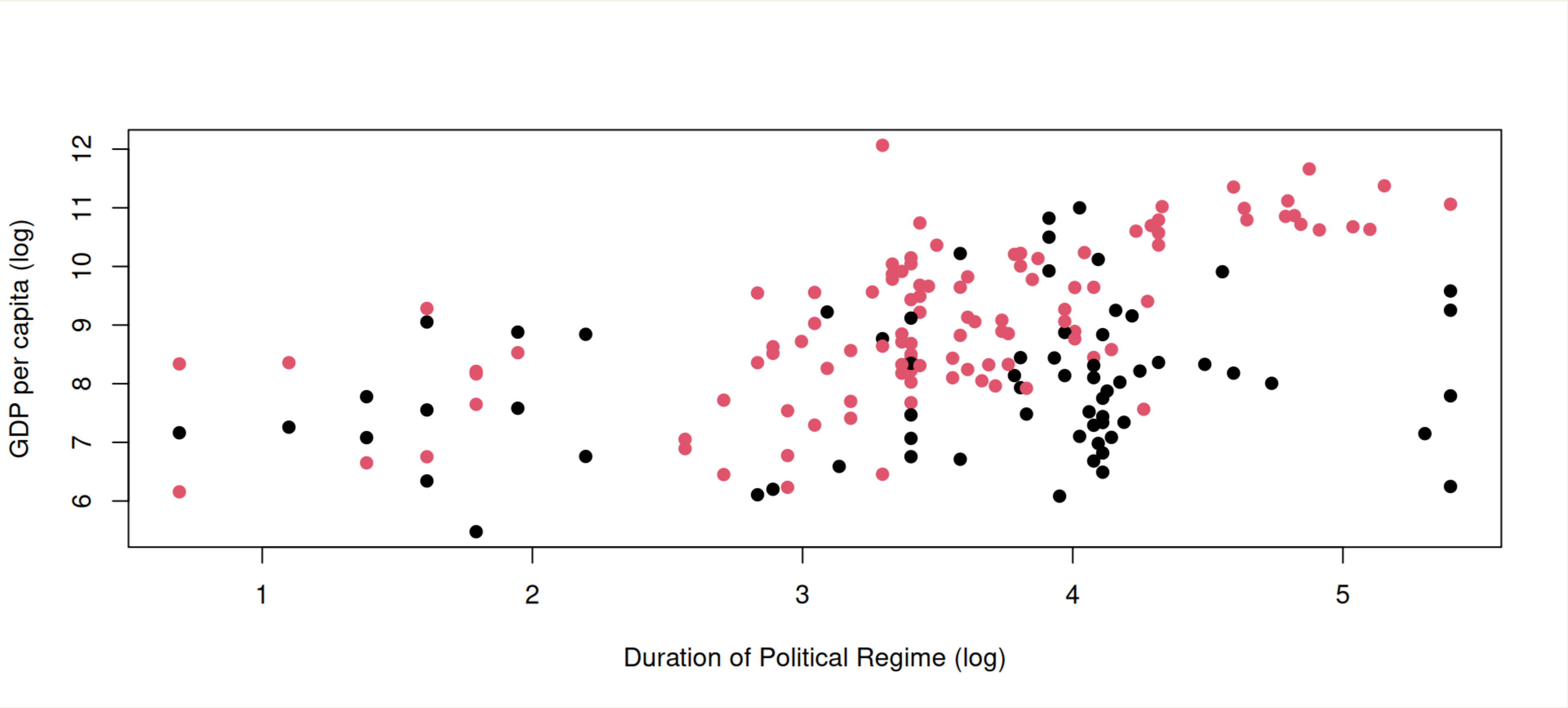
```r
summary(lm_fit_2)$adj.r.squared
```

```
[1] 0.2256682
```

Regression with Transformed Variables

Example: Log Transformation

Plot Code



Example: Multiple Regression with Log Transformation

- As we attempt to model the mean (expected value) of Y, for skewed distributions log transformation makes the variable distributions appear more normal.
- We can denote the population regression model as:

$$Log(GDP)_i = \alpha + \beta_1 log(Longevity)_i + \beta_2 Democracy_i + \epsilon_i$$

- And our estimated model is:

$$\widehat{Log(GDP)}_i = \hat{\alpha} + \hat{\beta}_1 log(Longevity)_i + \hat{\beta}_2 Democracy_i$$

- In economics this class of linear models (log DV and log IV) are known as **elasticity**.

Example: Estimating Regression with Log Transformation

- Now let's estimate our $log(Longevity) + Democracy \rightarrow GDP$ model in R:

$$\widehat{log(GDP)}_i = \hat{\alpha} + \hat{\beta}_1 log(Longevity)_i + \hat{\beta}_2 Democracy_i$$

```r
lm_fit_3 <- lm(log(gdp_per_capita) ~ log(democracy_duration) + democracy, data = democracy_gdp_2020)
lm_fit_3
```

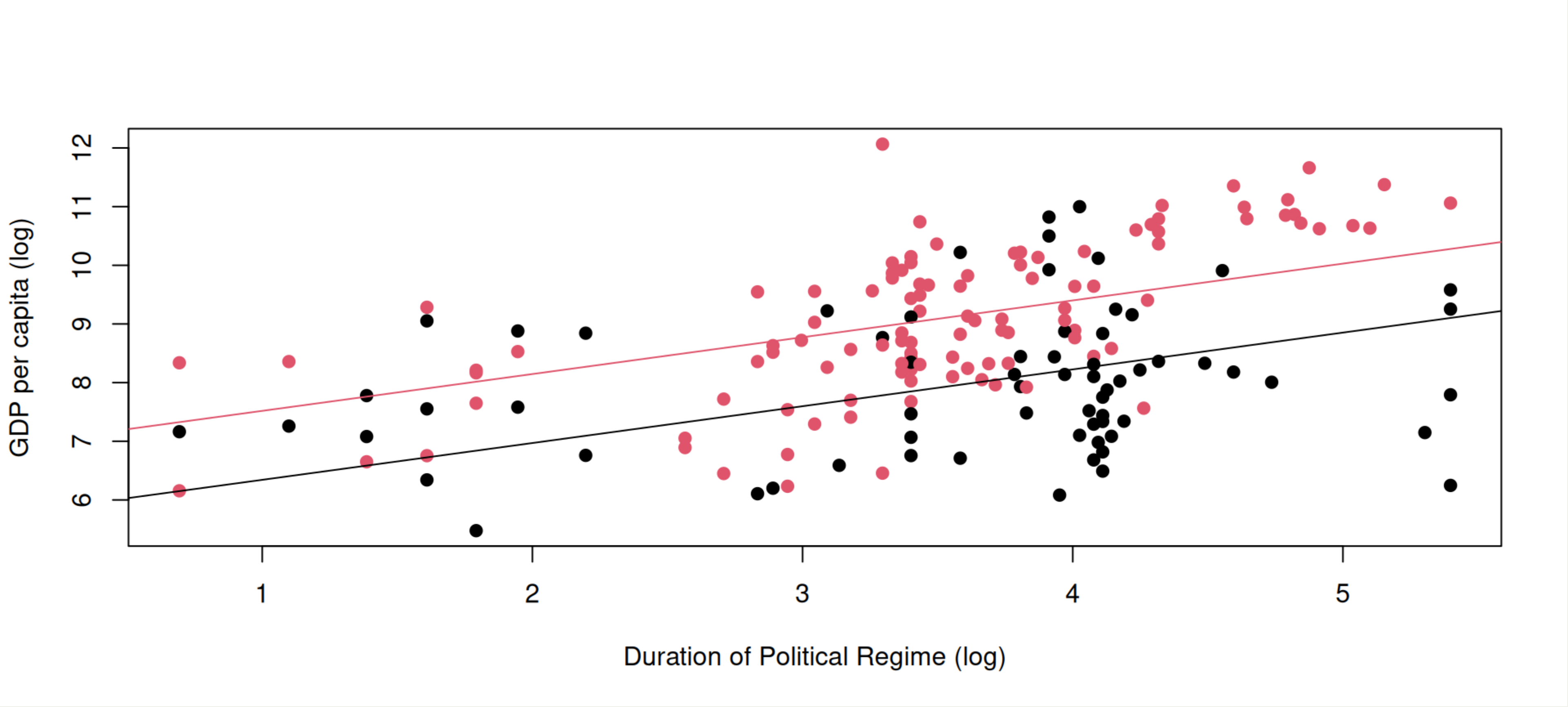
```

Call:
lm(formula = log(gdp_per_capita) ~ log(democracy_duration) + 
    democracy, data = democracy_gdp_2020)

Coefficients:
            (Intercept)  log(democracy_duration)                democracy  
                 5.7157                   0.6274                   1.1758  
```

- In other words our OLS estimate of this model is:

$$\widehat{log(GDP)}_i = 5.7157 + 0.6274 \times log(Longevity)_i + 1.1758 \times Democracy_i$$

Example: Plotting Fitted Regression Model

Plot | Code



Example: Interpreting Regression with Log Transformation

- Let's interpret our fitted model:

$$\widehat{log(GDP)}_i = 5.7157 + 0.6274 \times log(Longevity)_i + 1.1758 \times Democracy_i$$

- $\hat{\alpha} = 5.7157$ - the expected log GDP per capita for an autocratic state where a political regime lasted 1 year is 5.7157 USD.
- $\hat{\beta}_1 = 0.6274$ - each additional log year of political regime's longevity, on average, is associated with a 0.6274 USD increase in log GDP per capita, holding political regime constant.
    - I.e. increasing longevity by 1% is associated with increasing GDP per capita by 0.63%
- $\hat{\beta}_2 = 1.1758$ - democratic political regimes are associated with a 1.1758 USD increase in log GDP per capita, controlling for regime longevity.

Example: Significance Testing

```r
summary(lm_fit_3)
```

```
Call:
lm(formula = log(gdp_per_capita) ~ log(democracy_duration) +
    democracy, data = democracy_gdp_2020)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.85521 -0.87765 -0.07444  0.82037  3.10558 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              5.71571    0.34602  16.519  < 2e-16 ***
log(democracy_duration)  0.62745    0.08795   7.134 2.60e-11 ***
democracy                1.17576    0.17567   6.693 2.96e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.128 on 172 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.3441,    Adjusted R-squared:  0.3365 
```

Next

- Workshop:
  - RQ Presentations II
- 3 R Assignment due:
  - **08:59 Tuesday, 25 March**
- Next week:
  - Linear regression III