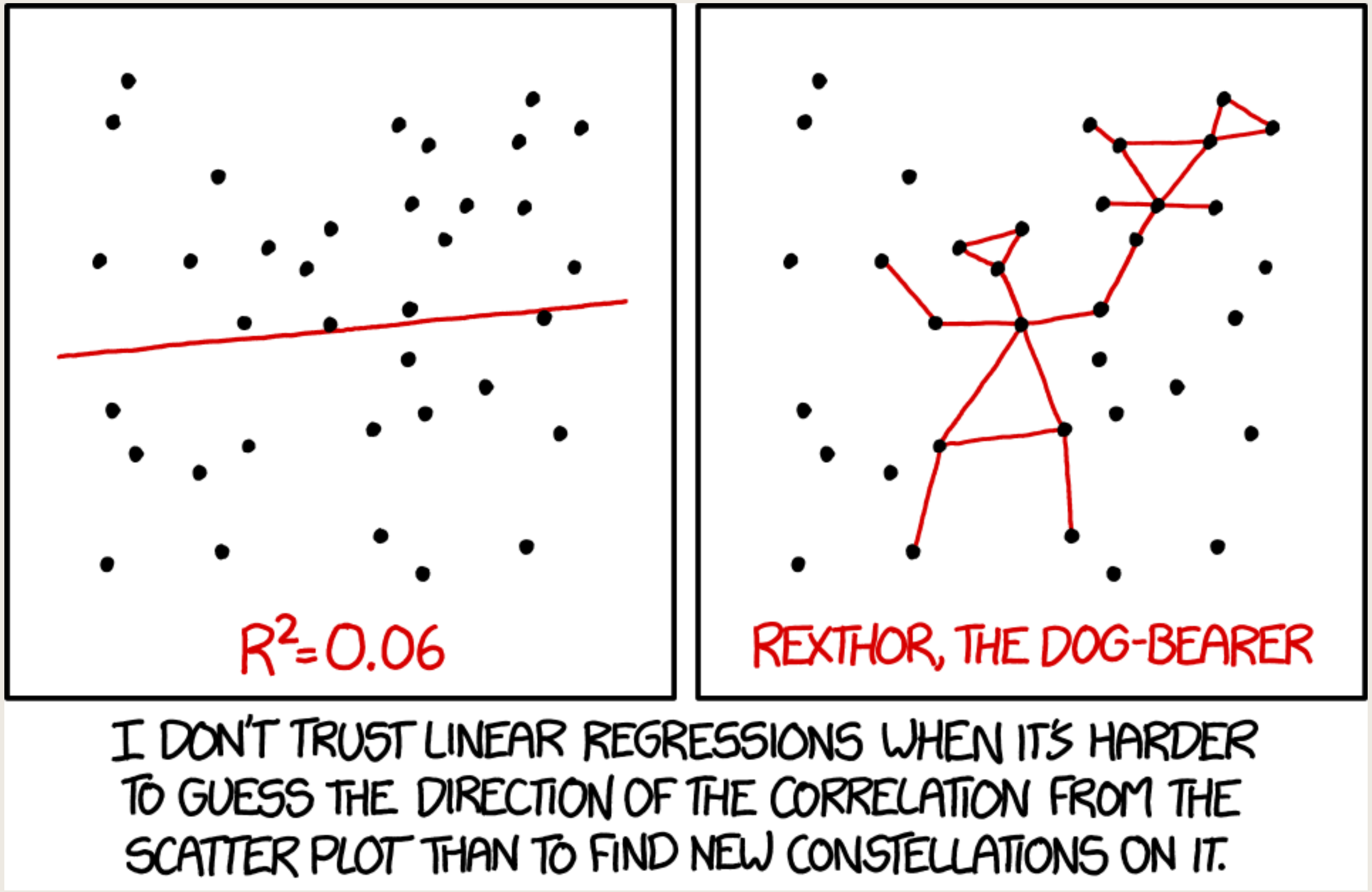


Linear Regression

Introduction to Statistics

The Plan for Today

The Plan for Today

- From *Simple* to *Multiple/Multivariable* Linear Regression (OLS)
  - Predicting Y as a linear function of $X_1, X_2, X_3 \ldots$
  - Goodness of fit (R^2).

The Plan for Today

- From *Simple* to *Multiple/Multivariable* Linear Regression (OLS)
    - Predicting Y as a linear function of $X_1, X_2, X_3 \ldots$
    - Goodness of fit (R^2).
- OLS Assumptions
    - Four conditions for *unbiased* estimation with OLS.
    - Next week: two additional assumptions for *efficient* estimation.

Disclaimers

Disclaimers

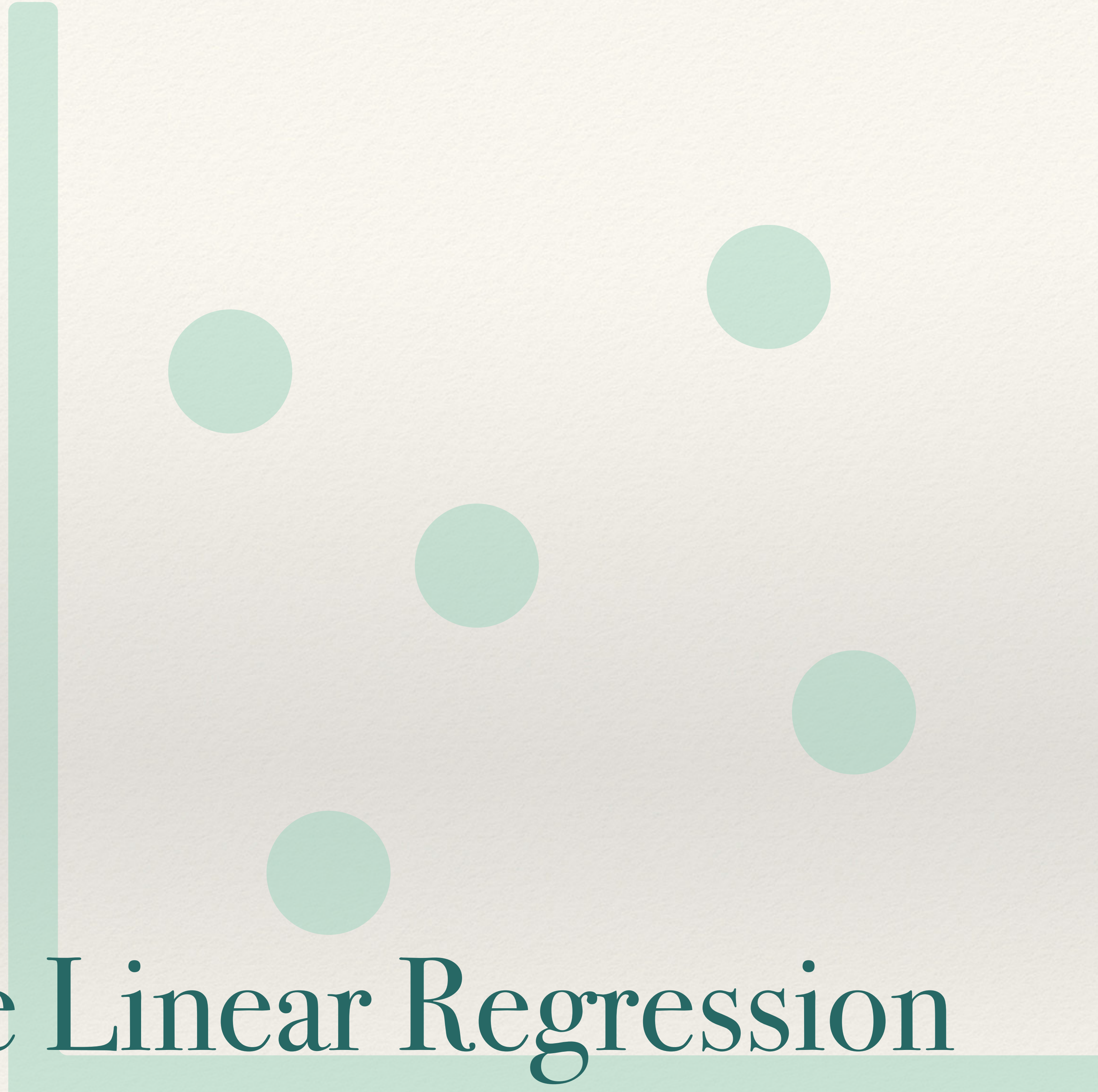
- This is **a lot**. It's okay not to get everything first time.
  - We'll revisit regressions over the next two weeks.

Disclaimers

- This is **a lot**. It's okay not to get everything first time.
  - We'll revisit regressions over the next two weeks.
- We won't be able to follow all the math.
  - Our aim: develop intuitive/geometric understanding of simplest cases; generalise from there.

Disclaimers

* This is **a lot**. It's okay not to get everything first time.
  * We'll revisit regressions over the next two weeks.
* We won't be able to follow all the math.
  * Our aim: develop intuitive/geometric understanding of simplest cases; generalise from there.
* **Lots of assumptions** 'under the hood'.
  * Our aim: understand pitfalls and limitations of OLS. More in the lab + next week on diagnostics and potential remedies.



Simple Linear Regression

Simple Linear Regression **Model**

Simple Linear Regression **Model**

- Assumes a **model of the data-generating process** where Y is a linear function of X, plus some chance error ϵ:

Simple Linear Regression **Model**

- Assumes a **model of the data-generating process** where Y is a linear function of X, plus some chance error ϵ:

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

Simple Linear Regression **Model**

- Assumes a **model of the data-generating process** where Y is a linear function of X, plus some chance error ϵ:

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

- This is a **model,** a mathematical representation of our assumption that there is a linear relationship between X and Y.

Simple Linear Regression **Model**

- Assumes a **model of the data-generating process** where Y is a linear function of X, plus some chance error ϵ:

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

- This is a **model,** a mathematical representation of our assumption that there is a linear relationship between X and Y.
- α and β represent the **true**, **unknown** intercept and slope of the line of best fit. These are often called **parameters**.

Simple Linear Regression **Model**

* Assumes a **model of the data-generating process** where Y is a linear function of X, plus some chance error ϵ:

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

* This is a **model,** a mathematical representation of our assumption that there is a linear relationship between X and Y.
* α and β represent the **true**, **unknown** intercept and slope of the line of best fit. These are often called **parameters**.
* ϵ_i represents the chance error: $\alpha + \beta X_i$ will not return the exact value of Y_i but each observation will fall somewhere below or above the line. Assumption: **this discrepancy is random**.

Simple Linear Regression with OLS

Simple Linear Regression with OLS

- OLS = Ordinary Least Squares. It’s an **estimator**, like e.g. the sample means and the sample proportion.

Simple Linear Regression with OLS

- OLS = Ordinary Least Squares. It's an **estimator**, like e.g. the sample means and the sample proportion.
- Goal of OLS: estimating α and β in the population from a sample.

Simple Linear Regression with OLS

- OLS = Ordinary Least Squares. It's an **estimator**, like e.g. the sample means and the sample proportion.
- Goal of OLS: estimating α and β in the population from a sample.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

Simple Linear Regression with OLS

- OLS = Ordinary Least Squares. It's an **estimator**, like e.g. the sample means and the sample proportion.
- Goal of OLS: estimating α and β in the population from a sample.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- $\hat{Y}$ = **fitted values**, our prediction of Y for each observation.

Simple Linear Regression with OLS

- OLS = Ordinary Least Squares. It's an **estimator**, like e.g. the sample means and the sample proportion.
- Goal of OLS: estimating α and β in the population from a sample.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- $\hat{Y}$ = **fitted values**, our prediction of Y for each observation.
- $\hat{\alpha}$ and $\hat{\beta}$ = **coefficients**; our estimate of intercept and slope.

Simple Linear Regression with OLS

- OLS = Ordinary Least Squares. It's an **estimator**, like e.g. the sample means and the sample proportion.
- Goal of OLS: estimating α and β in the population from a sample.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- $\hat{Y}$ = **fitted values**, our prediction of Y for each observation.
- $\hat{\alpha}$ and $\hat{\beta}$ = **coefficients**; our estimate of intercept and slope.
- Coefficients are notated with a 'hat' because they are **estimates**, not the 'real' parameters; fitted values also come with a 'hat' because they depend on $\hat{\alpha}$ and $\hat{\beta}$.

Simple Linear Regression with OLS

Simple Linear Regression with OLS

- OLS estimates a line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ that minimises the **sum of squared residuals**.

Simple Linear Regression with OLS

- OLS estimates a line $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ that minimises the **sum of squared residuals**.
- **The residual** for observation i is $\hat{\epsilon}_i \equiv Y_i - \hat{Y}_i$: the difference between the actual, observed value of Y_i and our prediction.

Simple Linear Regression with OLS

- OLS estimates a line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ that minimises the **sum of squared residuals**.
- **The residual** for observation i is $\hat{\epsilon}_i \equiv Y_i - \hat{Y}_i$: the difference between the actual, observed value of Y_i and our prediction.
- So, OLS computes…

Simple Linear Regression with OLS

- OLS estimates a line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ that minimises the **sum of squared residuals**.
- **The residual** for observation i is $\hat{\epsilon}_i \equiv Y_i - \hat{Y}_i$: the difference between the actual, observed value of Y_i and our prediction.
- So, OLS computes…

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

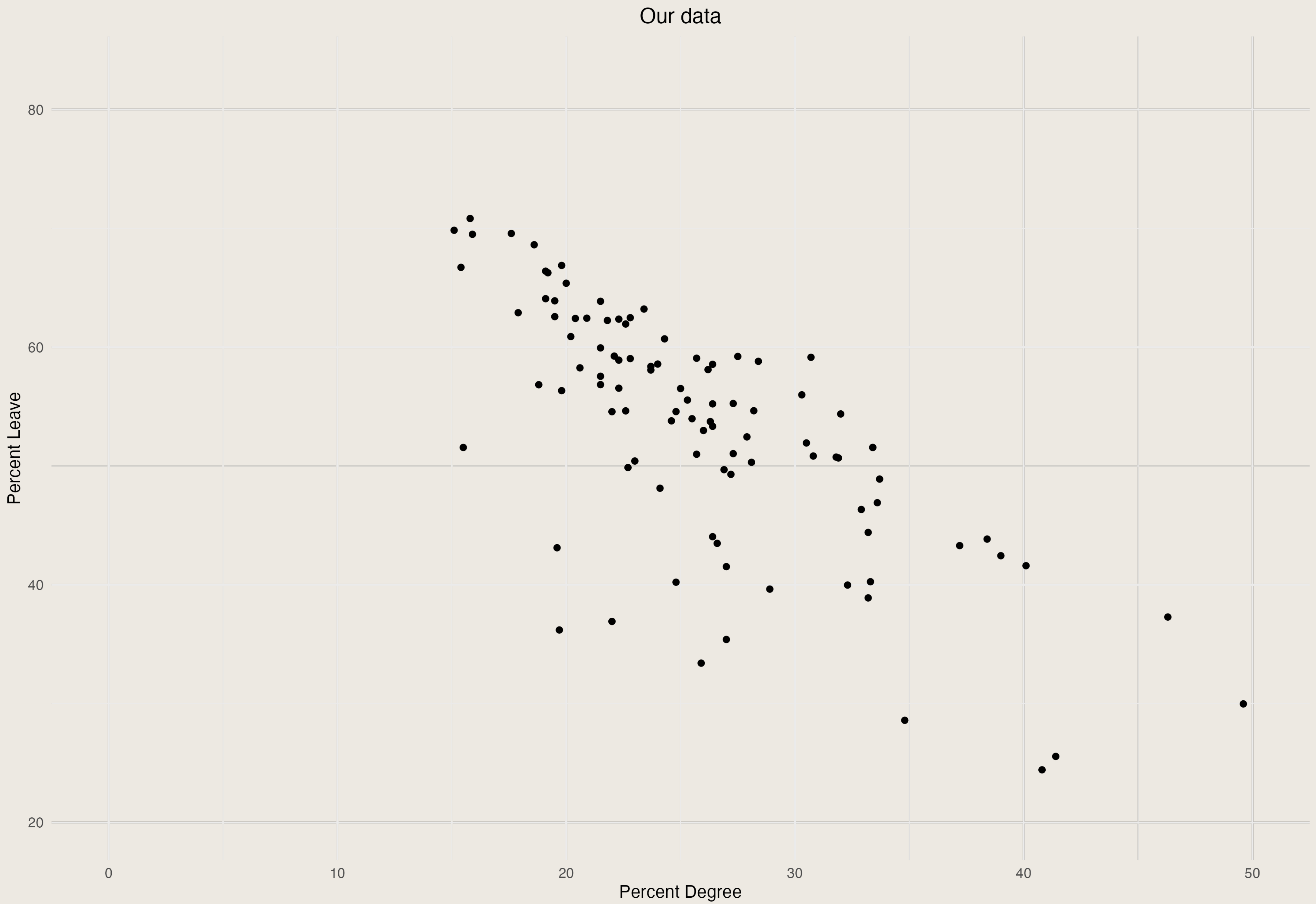
Simple Linear Regression with OLS

- OLS estimates a line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ that minimises the **sum of squared residuals**.
- **The residual** for observation i is $\hat{\epsilon}_i \equiv Y_i - \hat{Y}_i$: the difference between the actual, observed value of Y_i and our prediction.
- So, OLS computes…

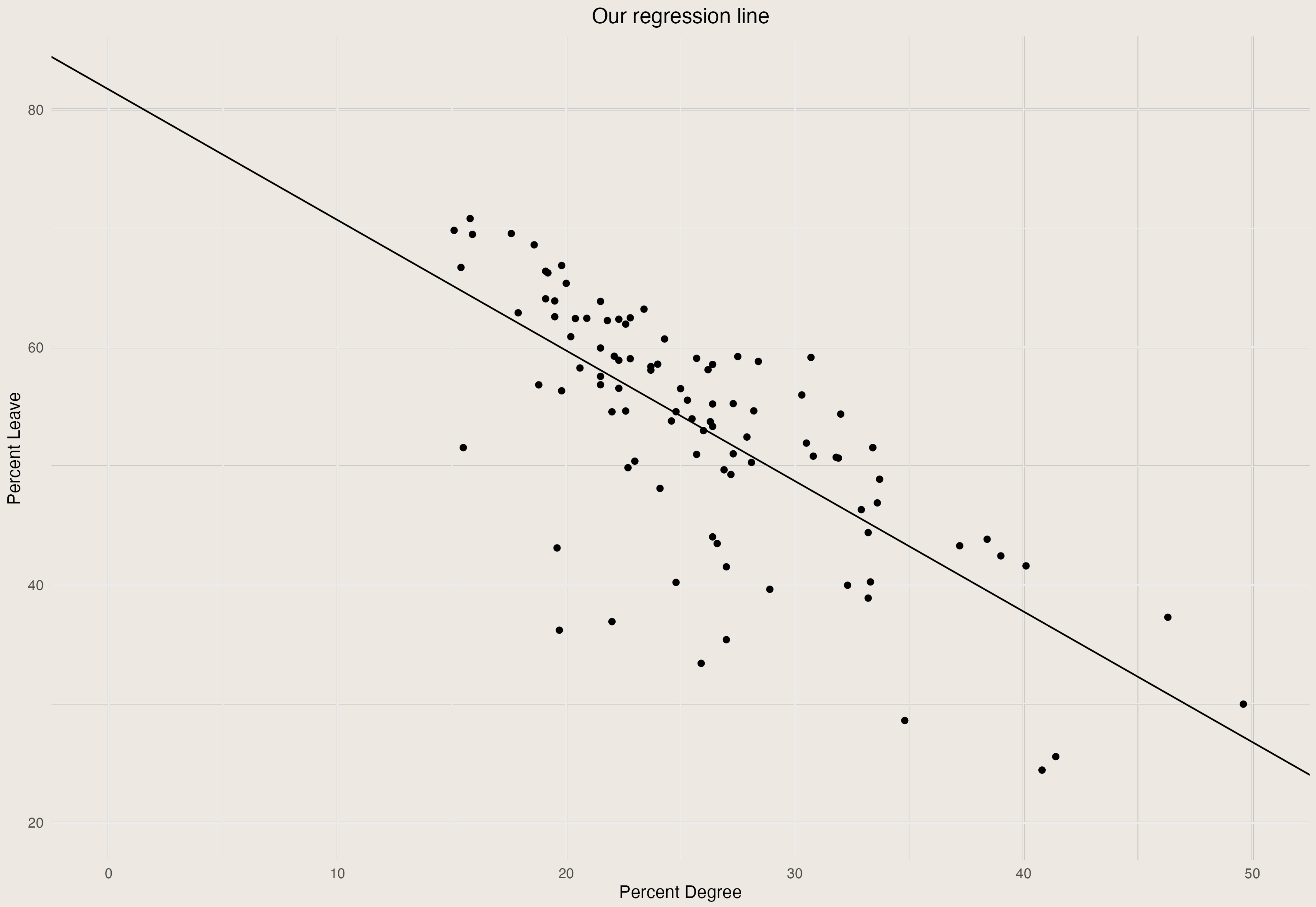
$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (\hat{\epsilon}_i)^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

- You can solve for $\hat{\alpha}$ and $\hat{\beta}$ with calculus (but we'll let R do it for us!)

Simple OLS: Geometric Interpretation

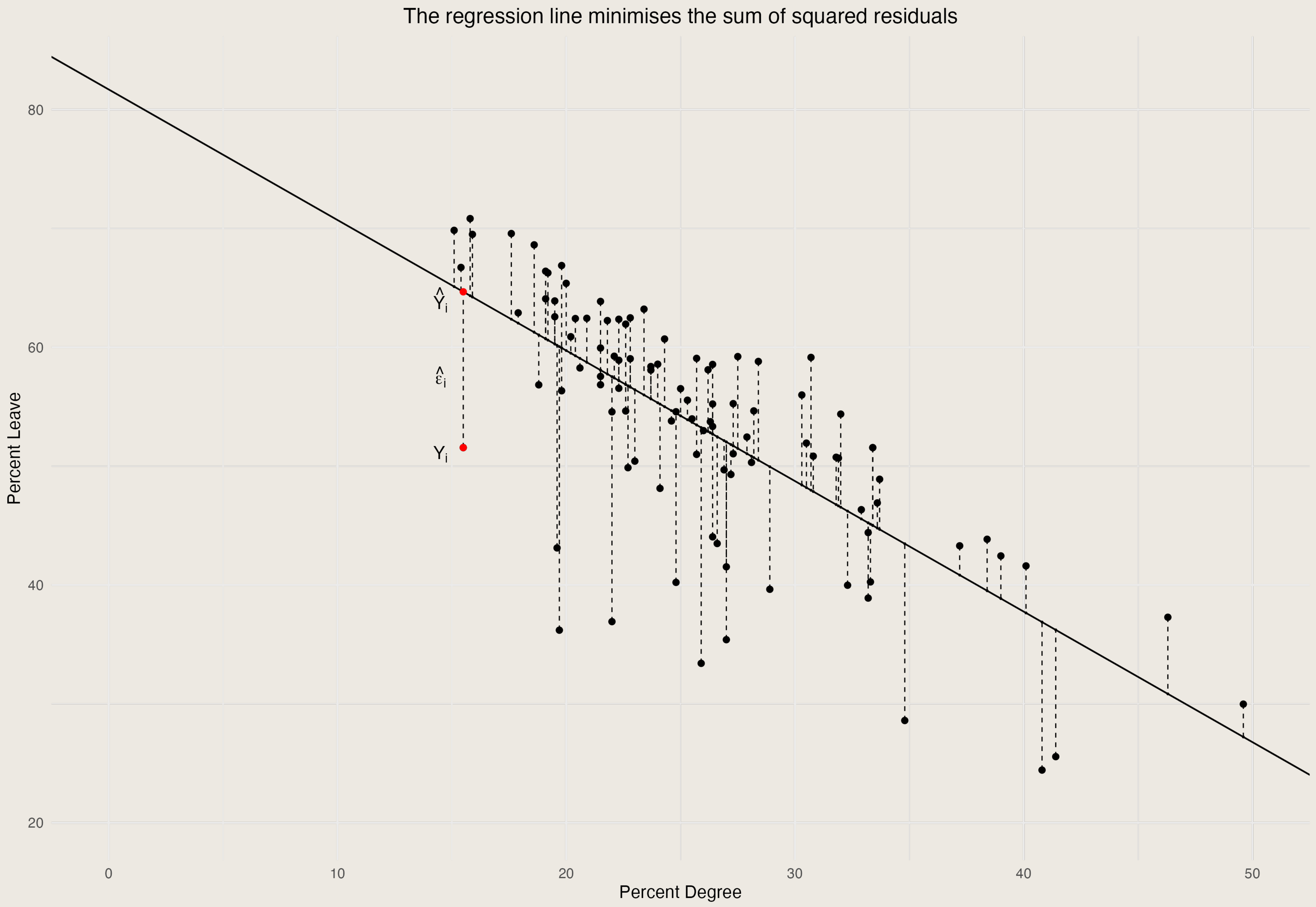


Simple OLS: Geometric Interpretation

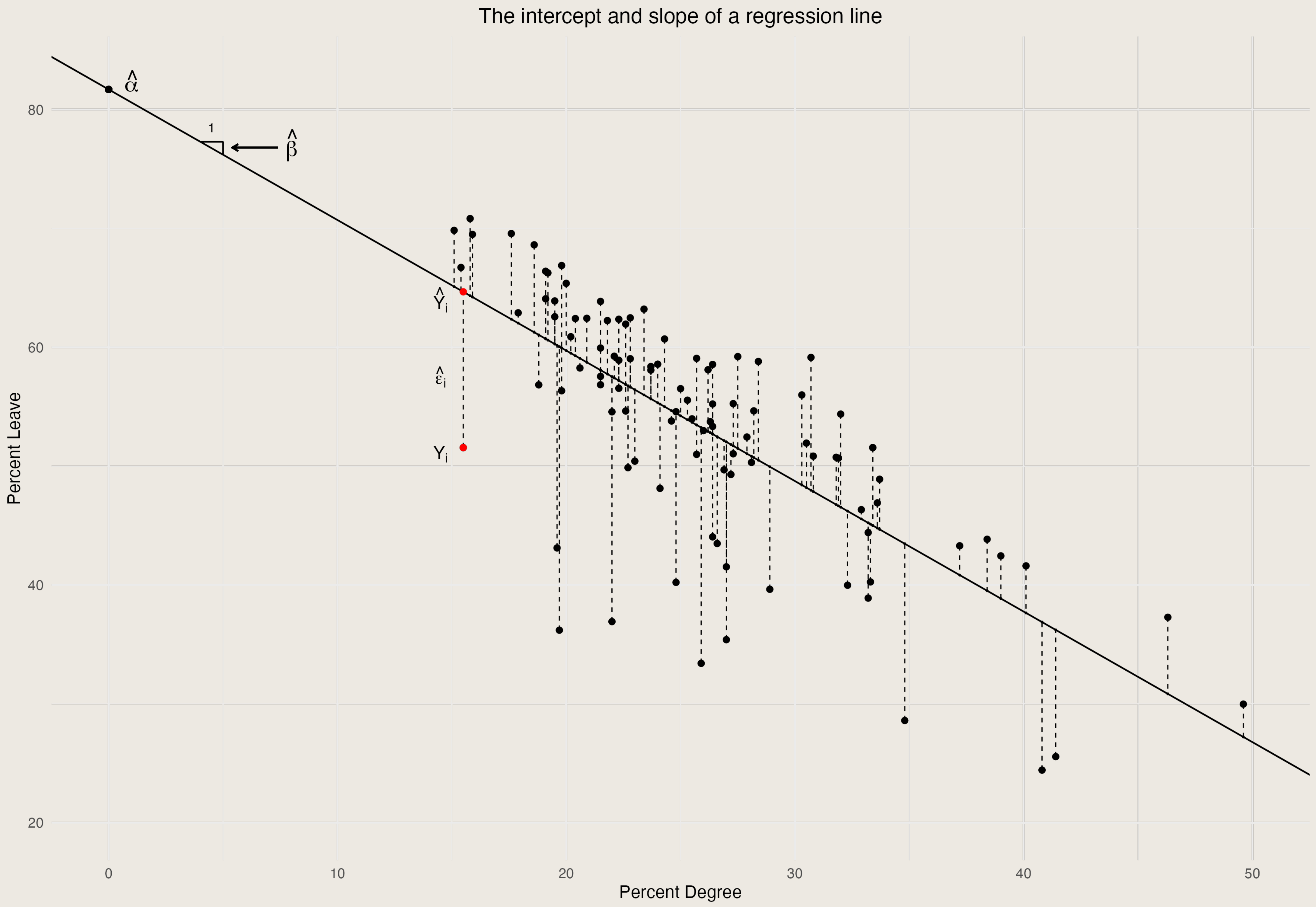


Simple OLS: Geometric Interpretation

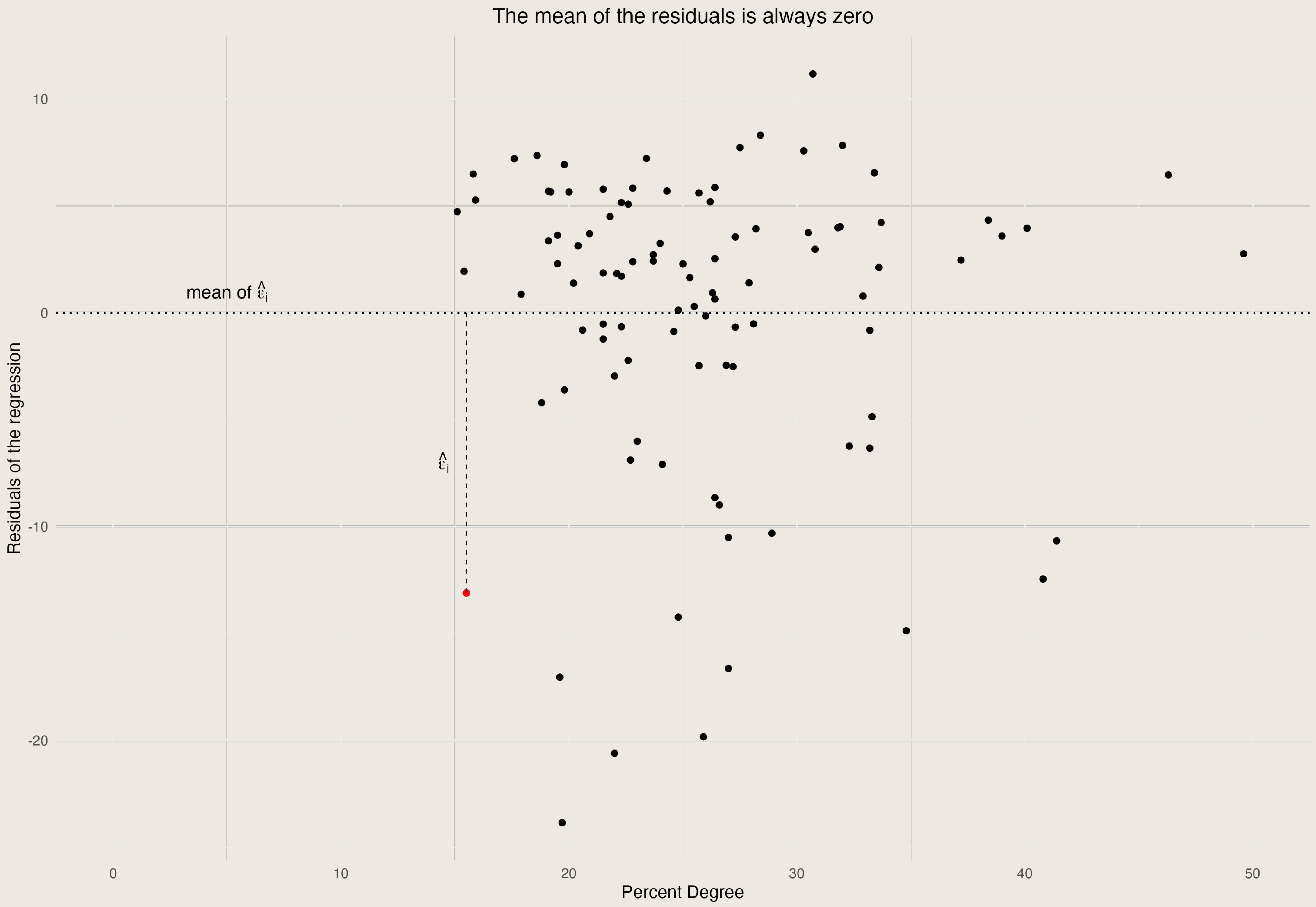
Simple OLS: Geometric Interpretation



Simple OLS: Geometric Interpretation



Simple OLS: Geometric Interpretation



Simple OLS in R

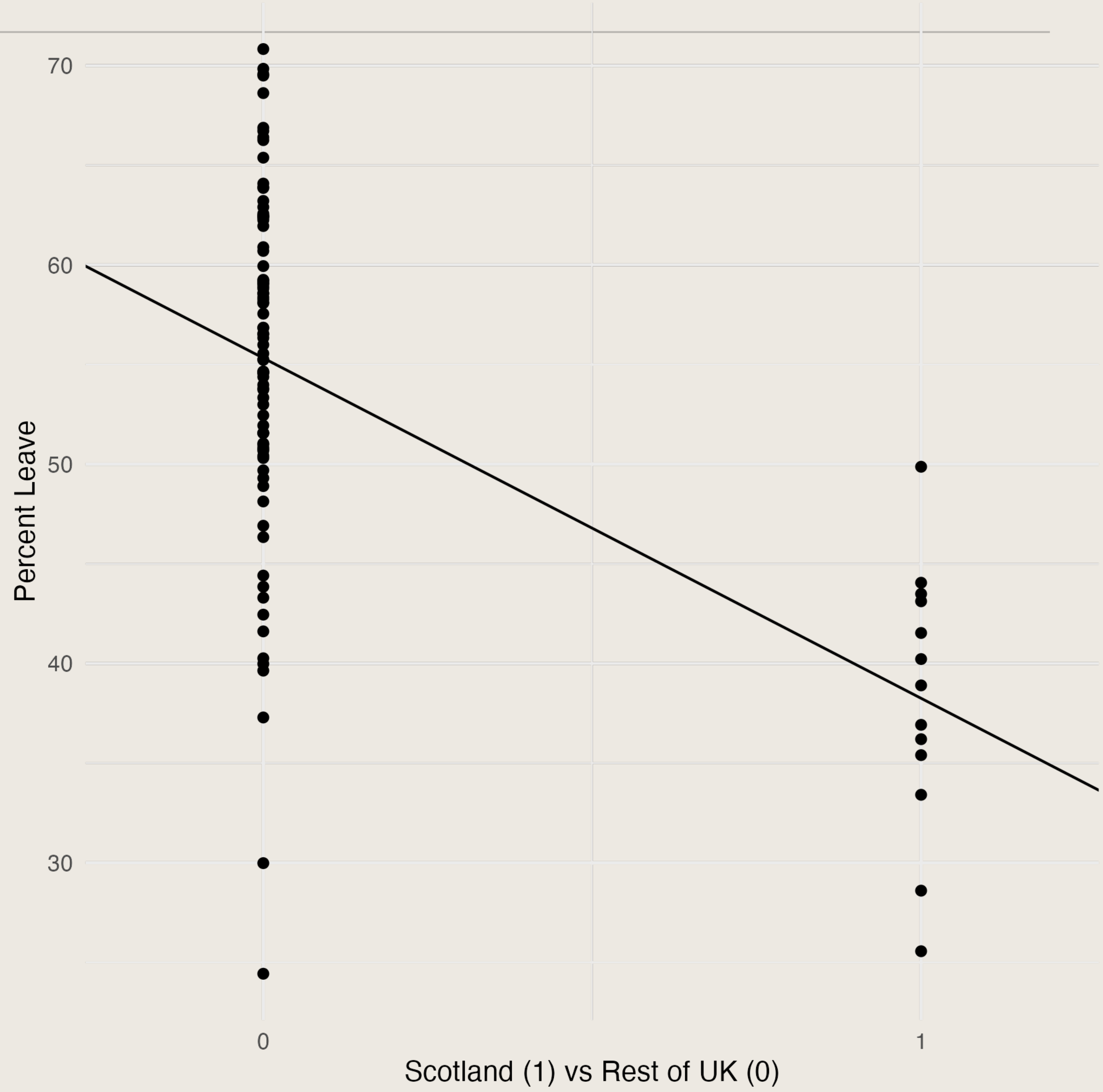
```
model1 <- lm(data = brexit, percent_leave ~ percent_degree)
summary(model1)

##
## Call:
## lm(formula = percent_leave ~ percent_degree, data = brexit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.855  -2.462   2.203   4.819  11.175
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     81.6906     2.8560   28.60   <2e-16 ***
## percent_degree  -1.0982     0.1063  -10.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.099 on 98 degrees of freedom
## Multiple R-squared:  0.5214, Adjusted R-squared:  0.5165
## F-statistic: 106.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

Simple OLS: Special Case

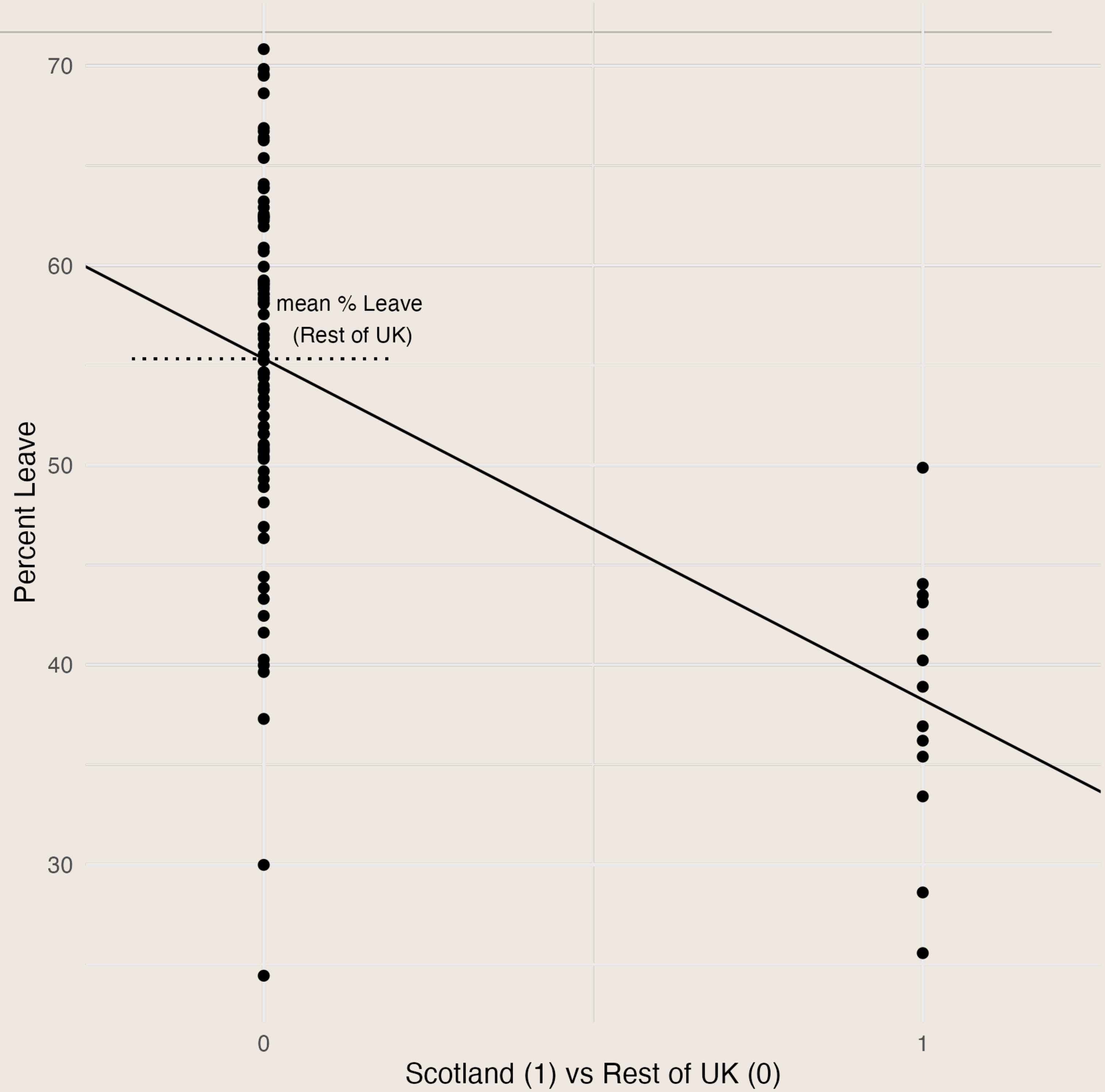
Simple OLS: Special Case

- When X is a 0-1 binary variable:



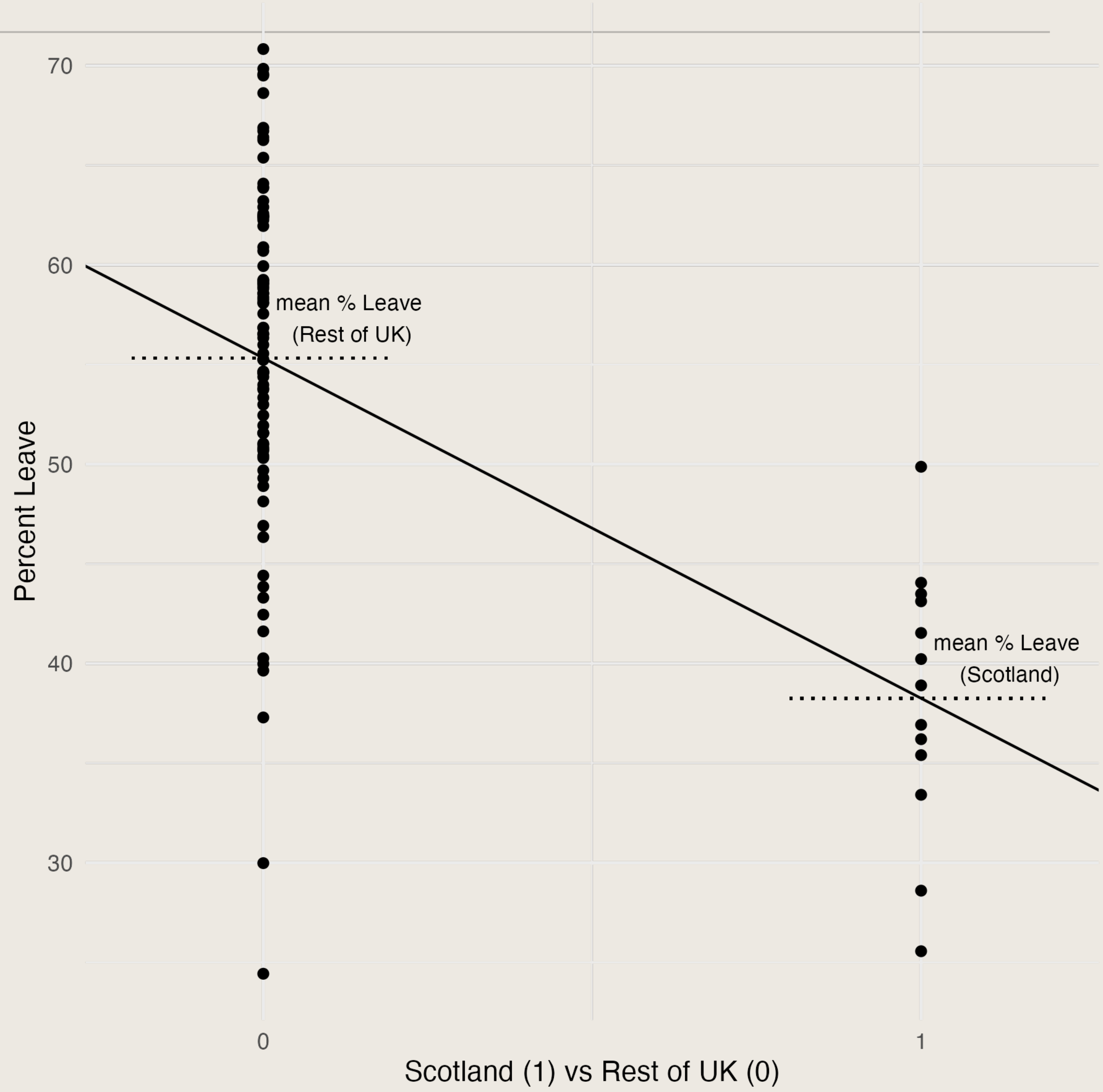
Simple OLS: Special Case

- When X is a 0-1 binary variable:
  - α is the mean of X for $X = 0$.



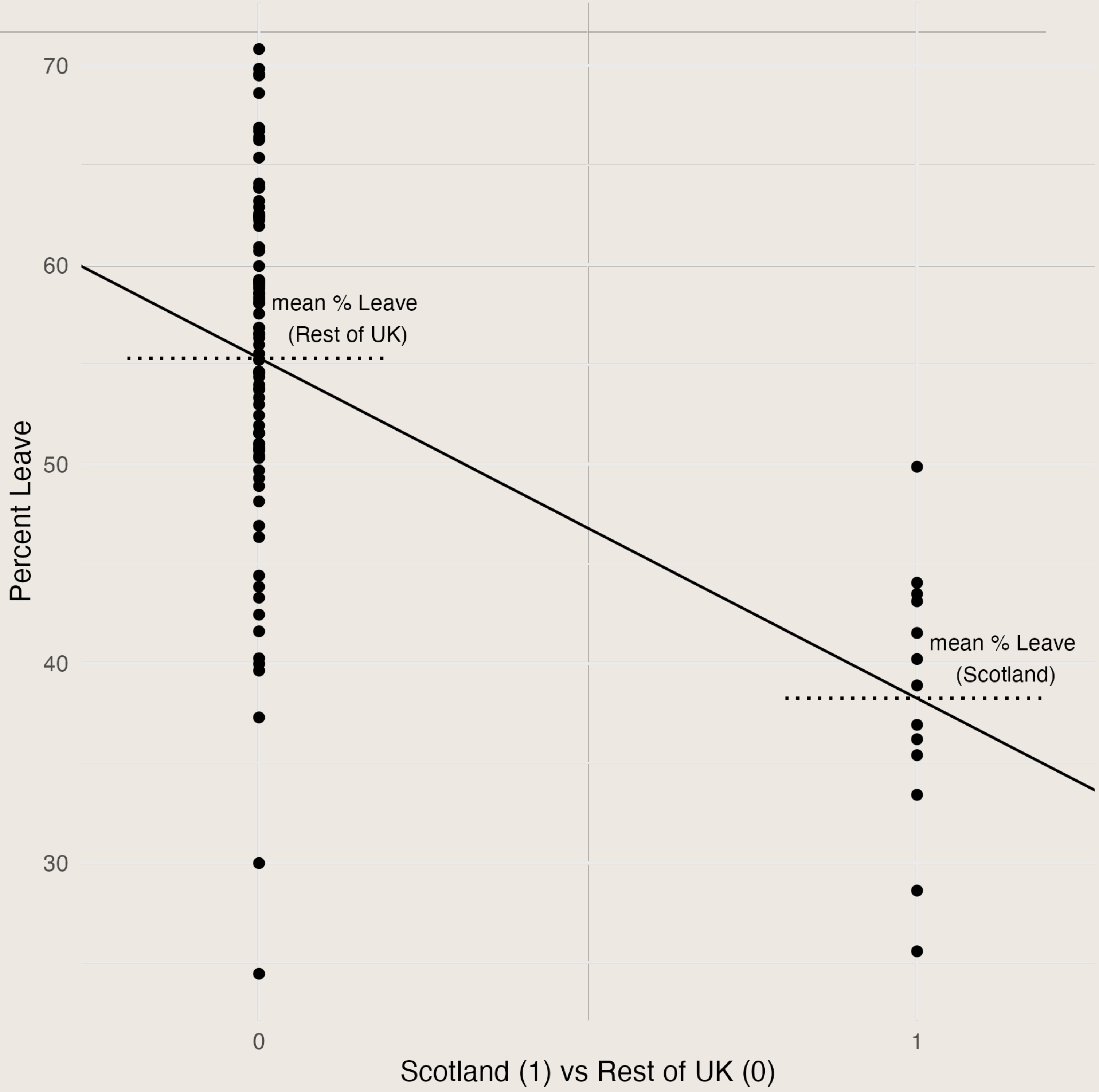
Simple OLS: Special Case

- When X is a 0-1 binary variable:
  - α is the mean of X for $X = 0$.
  - $\alpha + \beta$ is the mean for $X = 1$.



Simple OLS: Special Case

- When X is a 0-1 binary variable:
  - α is the mean of X for $X = 0$.
  - $\alpha + \beta$ is the mean for $X = 1$.
  - β is the difference-in-means.



Simple OLS: Special Case in R

```
model2 <- lm(data = brexit, percent_leave ~ scotland)
model2
```

```
##
## Call:
## lm(formula = percent_leave ~ scotland, data = brexit)
##
## Coefficients:
## (Intercept)     scotland
##       55.33       -17.07
```

```
brexit %>% group_by(scotland) %>%
summarise(mean_pct_leave = mean(percent_leave)) %>%
mutate(diff_in_means = mean_pct_leave - lag(mean_pct_leave))
```

```
## # A tibble: 2 × 3
##   scotland mean_pct_leave diff_in_means
##      <dbl>          <dbl>         <dbl>
## 1        0           55.3          NA
## 2        1           38.3         -17.1
```

Regression: why bother?

Regression: why bother?

* **Prediction:** make **guesses** for out-of-sample observations — e.g. constituency-level Brexit vote.

Regression: why bother?

* **Prediction:** make **guesses** for out-of-sample observations — e.g. constituency-level Brexit vote.
* **Description**: describe the **relationship** between an explanatory variable X and an outcome variable Y.

Regression: why bother?

* **Prediction:** make **guesses** for out-of-sample observations — e.g. constituency-level Brexit vote.
* **Description**: describe the **relationship** between an explanatory variable X and an outcome variable Y.
* **Causal Inference**: estimate the **effect** of X on Y — *only possible under <u>very</u> strong assumptions!*

Multiple Linear Regression

Why add independent variables?

Why add independent variables?

* **Prediction:** richer models give us more precise in-sample guesses and *can get us* to better out-of-sample guesses too (though not necessarily).

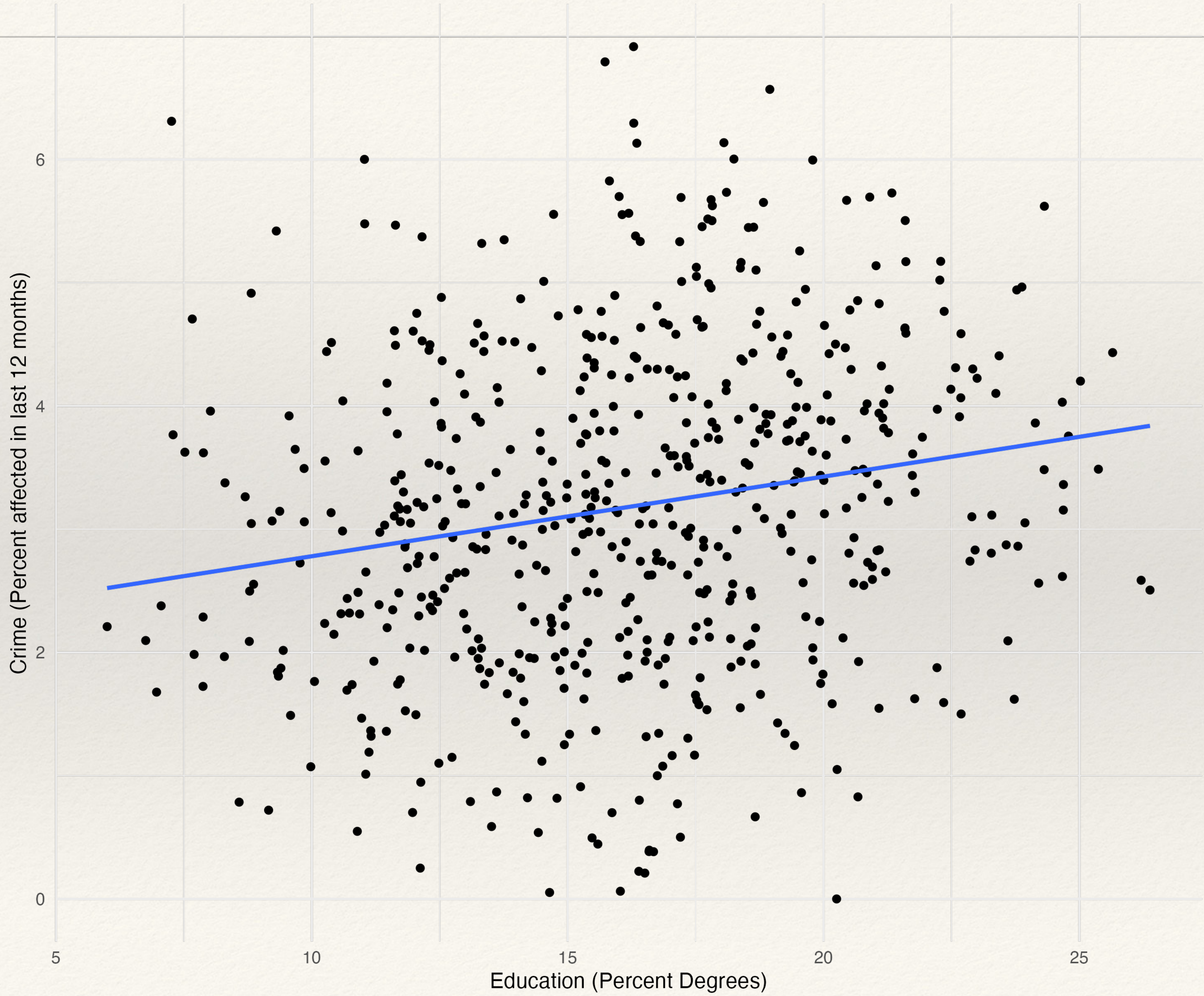
Why add independent variables?

* **Prediction:** richer models give us more precise in-sample guesses and *can get us* to better out-of-sample guesses too (though not necessarily).
* **Description**: describe the **relationship** between X and Y, *conditional on* Z — or 'controlling' for Z.

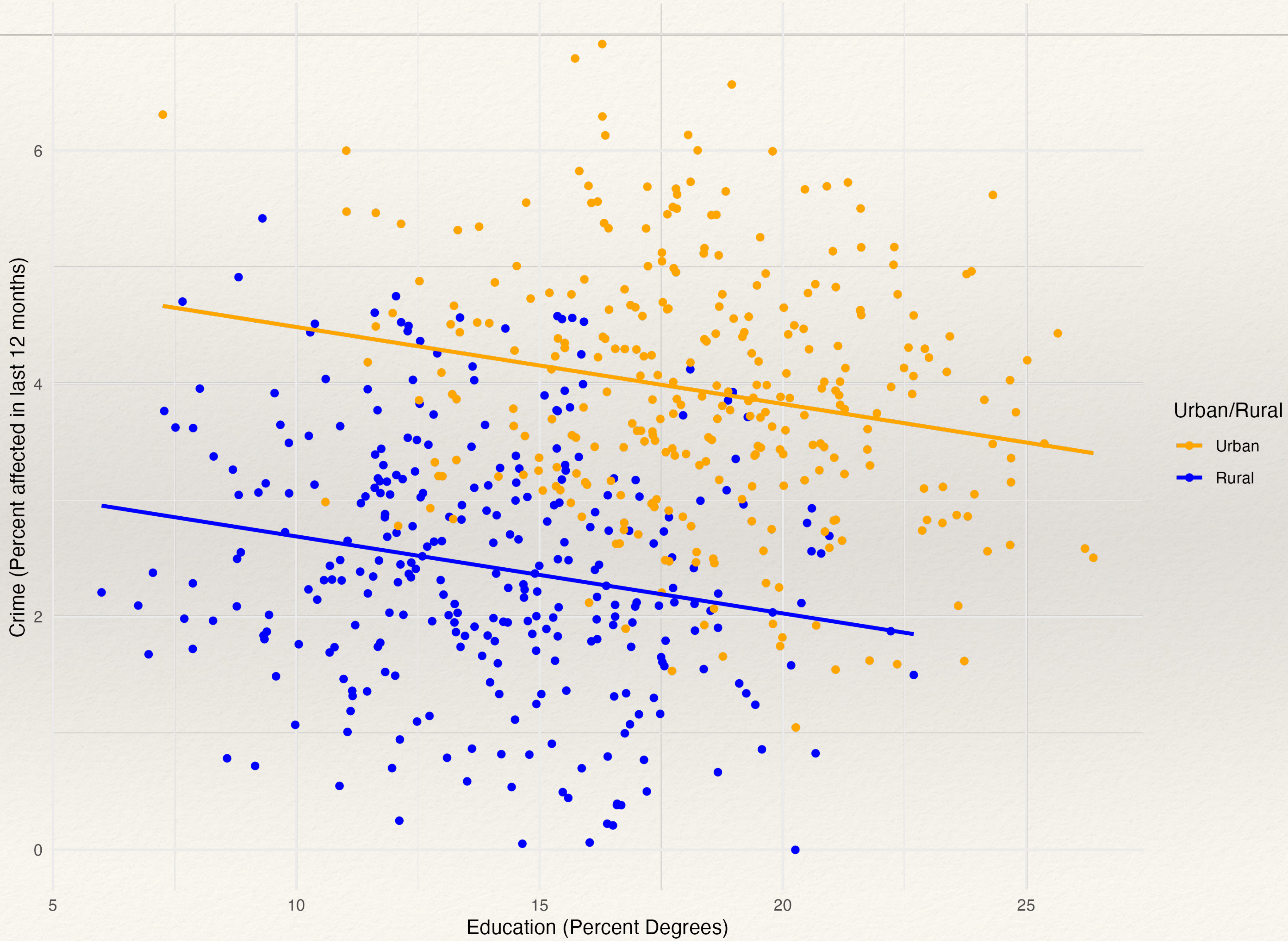
Why add independent variables?

- **Prediction:** richer models give us more precise in-sample guesses and *can get us* to better out-of-sample guesses too (though not necessarily).
- **Description**: describe the **relationship** between X and Y, *conditional on* Z — or 'controlling' for Z.
- **Causal Inference**: account for *confounders* to model counterfactual outcomes — effect of X on Y 'holding all else equal'. Again, requires *very strong assumptions*.

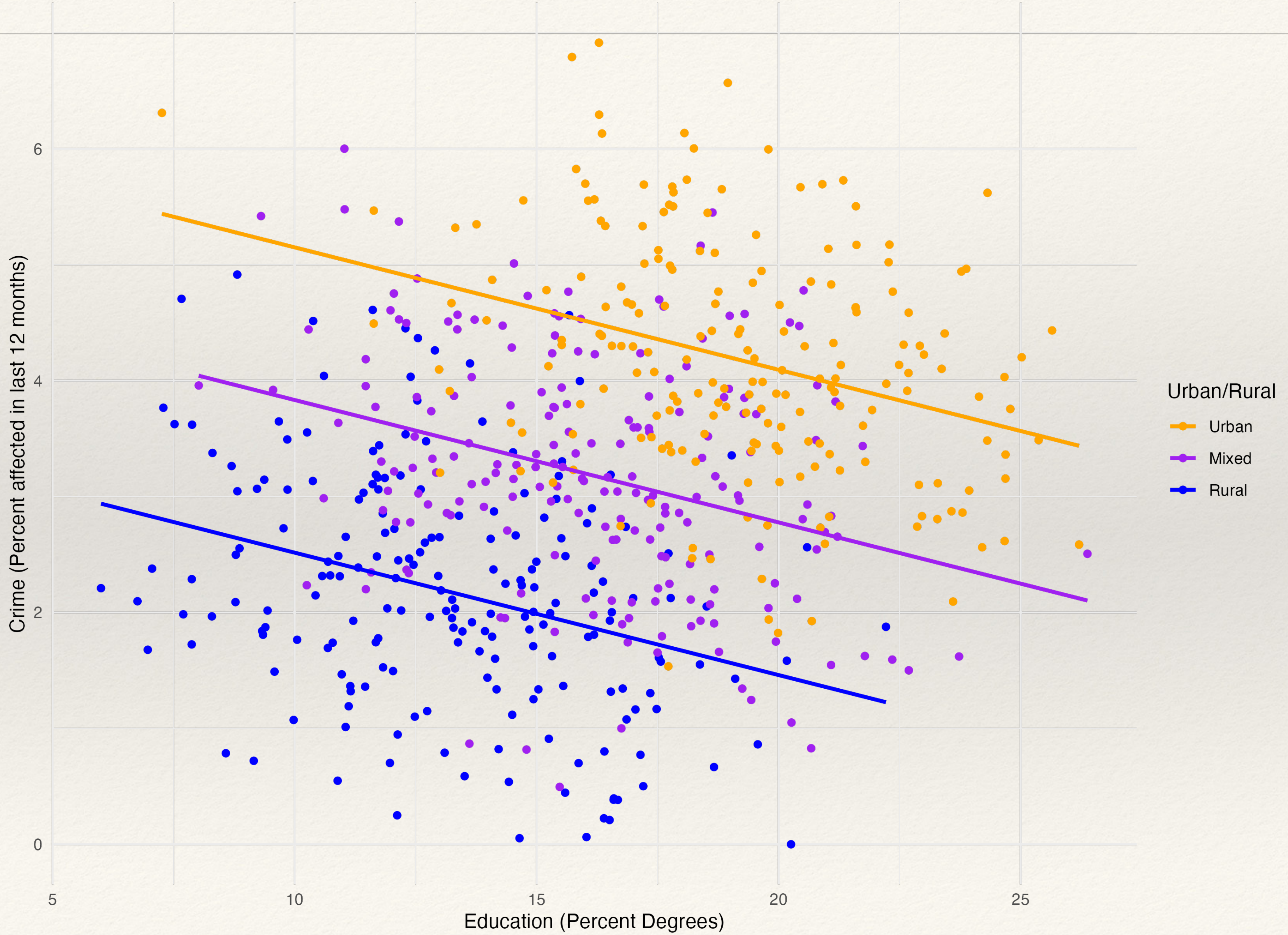
Conditional Relationships: Simpson's Paradox



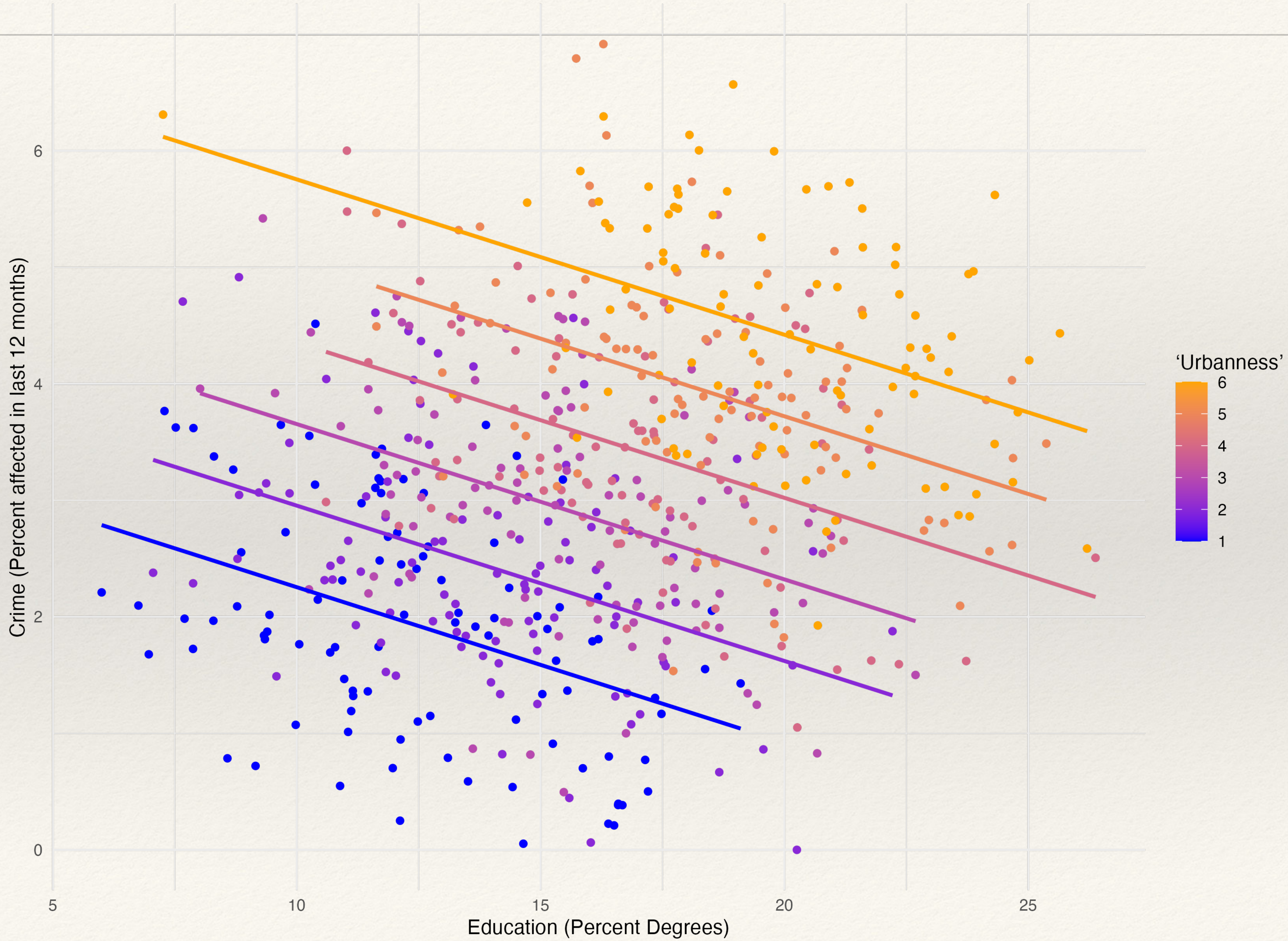
Conditional Relationships: Simpson's Paradox



Conditional Relationships: Simpson's Paradox



Conditional Relationships: Simpson's Paradox



Multiple OLS with Two Predictors

Multiple OLS with Two Predictors

* Our model of reality: Y as a **linear, additive** function of X_1 and X_2:

Multiple OLS with Two Predictors

- Our model of reality: Y as a **linear, additive** function of X_1 and X_2:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Multiple OLS with Two Predictors

* Our model of reality: Y as a **linear, additive** function of X_1 and X_2:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

* For instance:

$$\text{Pct. Leave}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Scotland}_i + \epsilon_i$$

Multiple OLS with Two Predictors

* Our model of reality: Y as a **linear, additive** function of X_1 and X_2:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

* For instance:

$$\text{Pct. Leave}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Scotland}_i + \epsilon_i$$

* Same least-square solution as the bivariate case:
  * Choose $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ so that in $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ the sum of squared residuals $\sum_{i=1}^{n} (\hat{\epsilon}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ is minimised.

Multiple OLS in R

```
model3 <- lm(data = brexit, percent_leave ~ percent_degree + scotland)
model3

##
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree        scotland
##         82.576          -1.053         -15.888
```

Multiple OLS in R

```
model3 <- lm(data = brexit, percent_leave ~ percent_degree + scotland)
model3

##
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree        scotland
##         82.576          -1.053         -15.888
```

- When Scotland = 1

Multiple OLS in R

```
model3 <- lm(data = brexit, percent_leave ~ percent_degree + scotland)
model3

##
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree        scotland
##         82.576          -1.053         -15.888
```

- When Scotland = 1
  - Pct. Leave ($\hat{Y}$) = 82.6 – 1.05 · (Pct. Degrees) – 15.9 · (1)

Multiple OLS in R

```
model3 <- lm(data = brexit, percent_leave ~ percent_degree + scotland)
model3

##
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree        scotland
##         82.576          -1.053         -15.888
```

- When Scotland = 1
  - Pct. Leave ($\hat{Y}$) = 82.6 – 1.05 · (Pct. Degrees) – 15.9 · (1)
  - Pct. Leave = 66.7 – 1.05 · (Pct. Degrees)

Multiple OLS in R

```r
model3 <- lm(data = brexit, percent_leave ~ percent_degree + scotland)
model3

##
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree        scotland
##         82.576          -1.053         -15.888
```

- When Scotland = 1
  - Pct. Leave ($\hat{Y}$) = 82.6 – 1.05 · (Pct. Degrees) – 15.9 · (1)
  - Pct. Leave = 66.7 – 1.05 · (Pct. Degrees)
- When Scotland = 0

Multiple OLS in R

```
model3 <- lm(data = brexit, percent_leave ~ percent_degree + scotland)
model3

##
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree        scotland
##         82.576          -1.053         -15.888
```

- When Scotland = 1
  - Pct. Leave ($\hat{Y}$) = 82.6 – 1.05 · (Pct. Degrees) – 15.9 · (1)
  - Pct. Leave = 66.7 – 1.05 · (Pct. Degrees)
- When Scotland = 0
  - Pct. Leave = 82.6 – 1.05 · (Pct. Degrees) – 15.9 · (0)

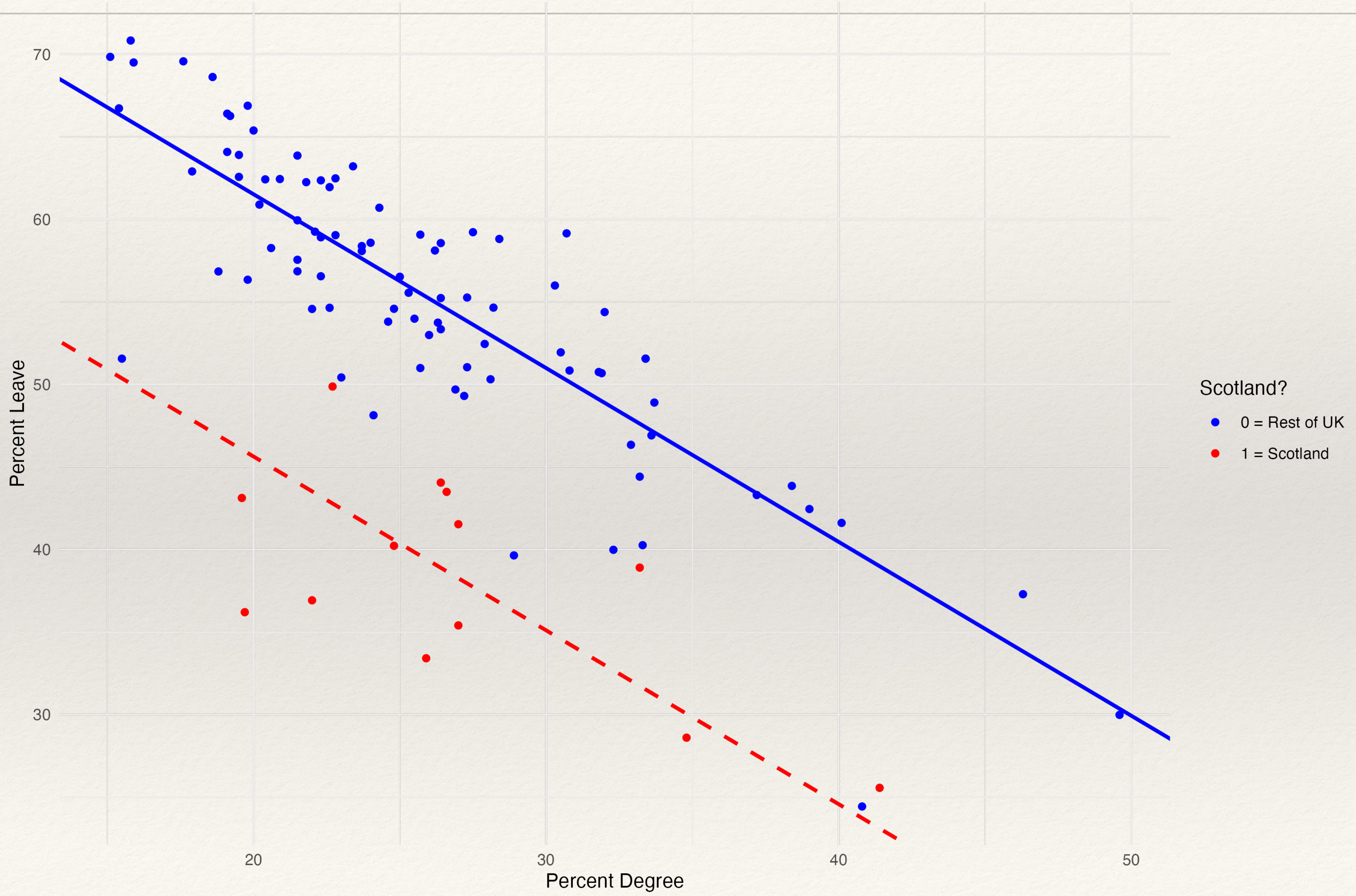
Multiple OLS in R

```
model3 <- lm(data = brexit, percent_leave ~ percent_degree + scotland)
model3

##
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree        scotland
##         82.576          -1.053         -15.888
```

- When Scotland = 1
  - Pct. Leave ($\hat{Y}$) = 82.6 – 1.05 · (Pct. Degrees) – 15.9 · (1)
  - Pct. Leave = 66.7 – 1.05 · (Pct. Degrees)
- When Scotland = 0
  - Pct. Leave = 82.6 – 1.05 · (Pct. Degrees) – 15.9 · (0)
  - Pct. Leave = 82.6 – 1.05 · (Pct. Degrees)

Geometric Interpretation



Multiple OLS with Two Interval Predictors

```
## Call:
## lm(formula = percent_leave ~ percent_degree + median_age, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree      median_age
##        67.3170         -1.0783          0.3296
```

Multiple OLS with Two Interval Predictors

- $\text{Pct. Leave}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Median Age}_i + \epsilon_i$

```
## Call:
## lm(formula = percent_leave ~ percent_degree + median_age, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree      median_age
##        67.3170         -1.0783          0.3296
```

Multiple OLS with Two Interval Predictors

- $\text{Pct. Leave}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Median Age}_i + \epsilon_i$

```
## Call:
## lm(formula = percent_leave ~ percent_degree + median_age, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree      median_age
##        67.3170         -1.0783          0.3296
```

- α = predicted value of Y when $X_1 = 0$ and $X_2 = 0$

Multiple OLS with Two Interval Predictors

- $\text{Pct. Leave}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Median Age}_i + \epsilon_i$

```
## Call:
## lm(formula = percent_leave ~ percent_degree + median_age, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree      median_age
##        67.3170         -1.0783          0.3296
```

- α = predicted value of Y when $X_1 = 0$ and $X_2 = 0$
- β_1 = change in Y associated with a one-percentage point increase in Pct. Degrees, *holding Median Age constant*.

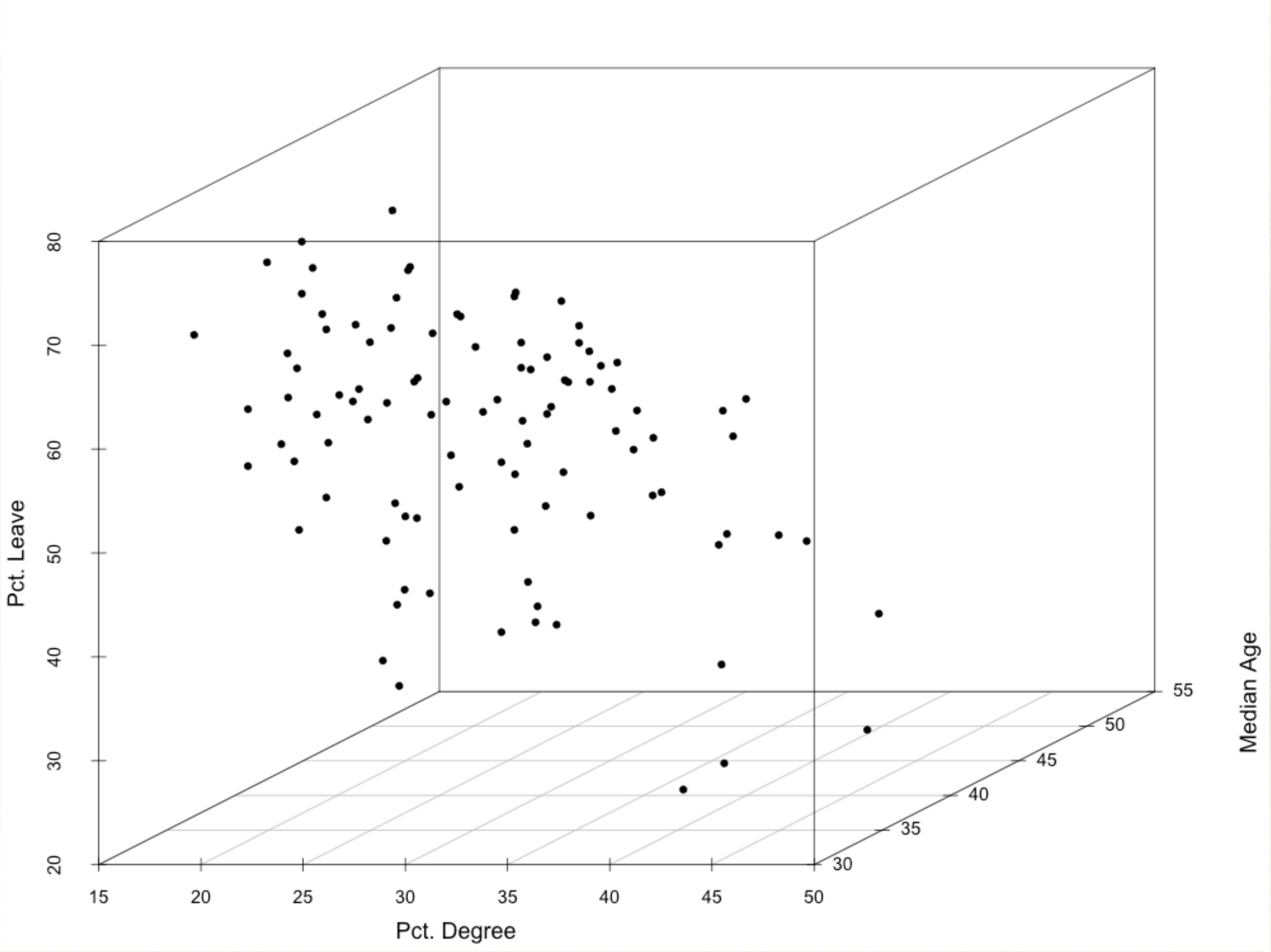
Multiple OLS with Two Interval Predictors

- $\text{Pct. Leave}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Median Age}_i + \epsilon_i$

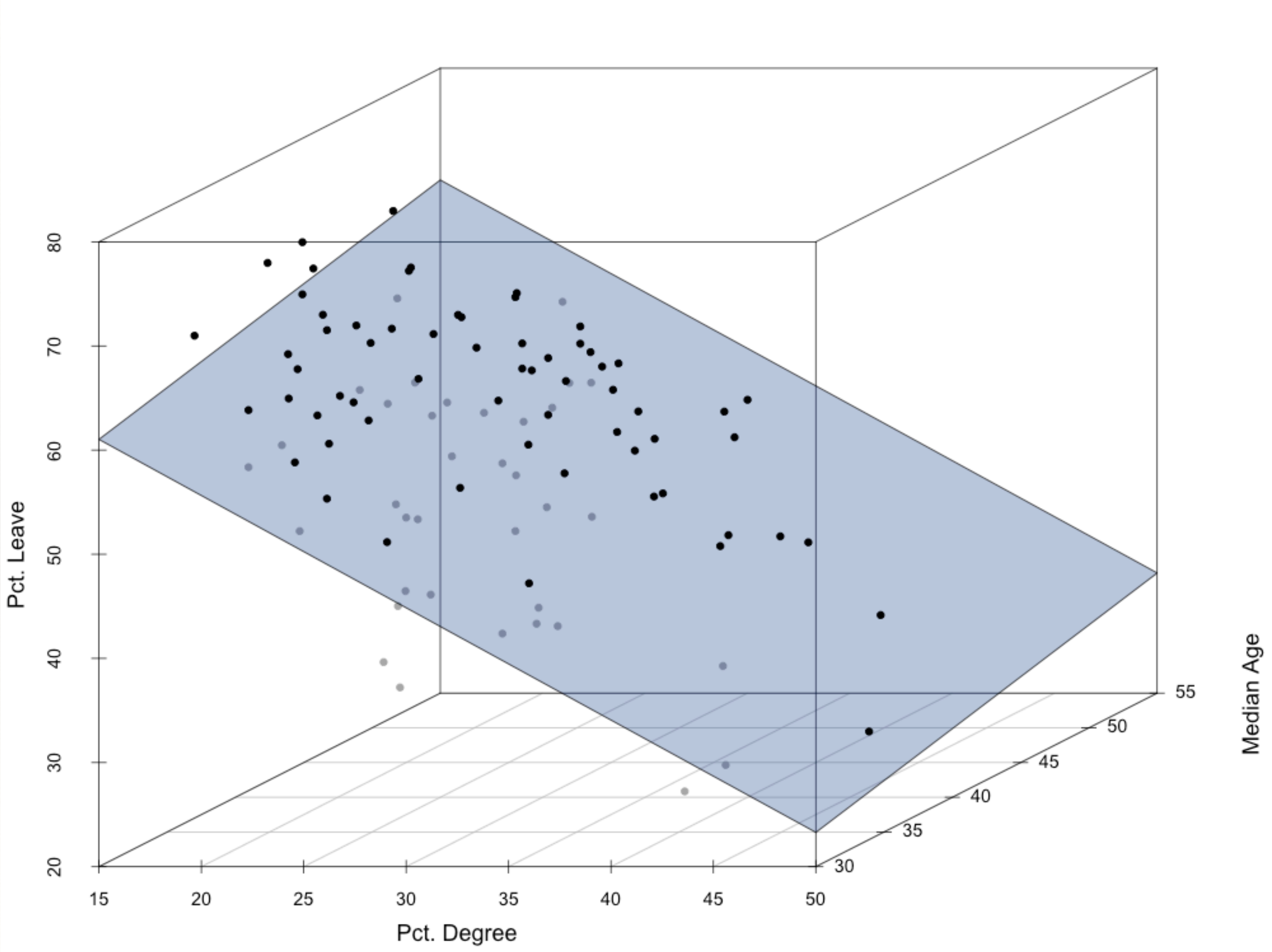
```
## Call:
## lm(formula = percent_leave ~ percent_degree + median_age, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree       median_age
##        67.3170         -1.0783           0.3296
```

- α = predicted value of Y when $X_1 = 0$ and $X_2 = 0$
- β_1 = change in Y associated with a one-percentage point increase in Pct. Degrees, *holding Median Age constant*.
- β_2 = change in Y associated with a one-year increase in Median Age, *holding Percentage of Residents with Degrees constant*.

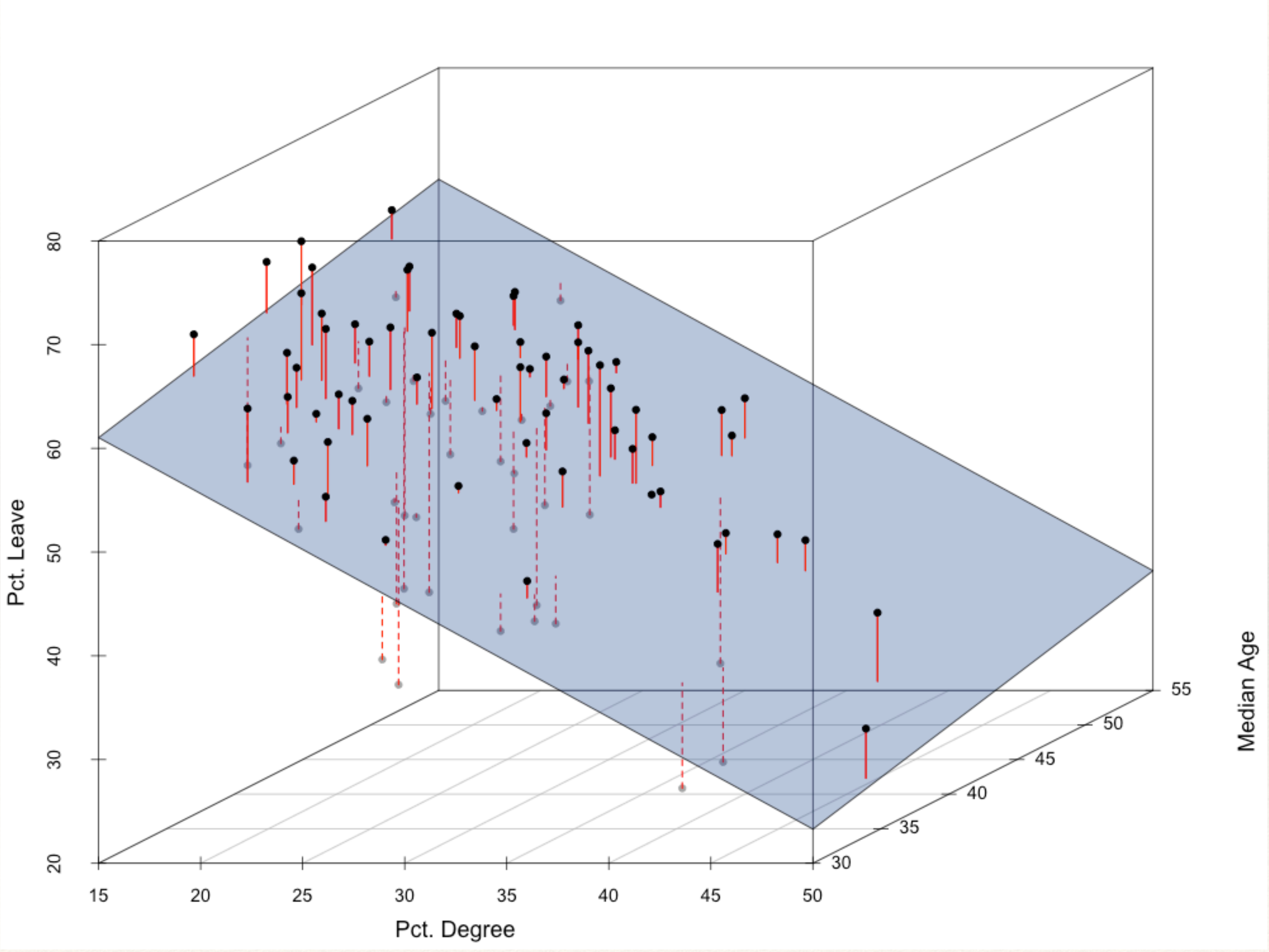
Geometric Interpretation



Geometric Interpretation



Geometric Interpretation



‘Partialing Out’ Interpretation

'Partialing Out' Interpretation

* Consider the model $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$.

‘Partialing Out’ Interpretation

* Consider the model $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$.
* What do we mean when we say that β_1 is the change in Y associated with a change in X, *controlling for* Z?

‘Partialing Out’ Interpretation

- Consider the model $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$.
- What do we mean when we say that β_1 is the change in Y associated with a change in X, *controlling for Z*?
- One way to think about it: β_1 is the ‘effect’ of

‘Partialing Out’ Interpretation

- Consider the model $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$.
- What do we mean when we say that β_1 is the change in Y associated with a change in X, *controlling for* Z?
- One way to think about it: β_1 is the ‘effect’ of
  - The component of X that is *not* explained by Z, on

‘Partialing Out’ Interpretation

* Consider the model $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$.
* What do we mean when we say that β_1 is the change in Y associated with a change in X, *controlling for* Z?
* One way to think about it: β_1 is the ‘effect’ of
    * The component of X that is *not* explained by Z, on
    * The component of Y that is *not* explained by Z.

‘Partialing Out’ Interpretation

‘Partialing Out’ Interpretation

- Multiple OLS coefficient can be obtained via separate binary regressions (Frisch-Waugh-Lovell theorem):

‘Partialing Out’ Interpretation

* Multiple OLS coefficient can be obtained via separate binary regressions (Frisch-Waugh-Lovell theorem):
    * Regress X on Z, extract the residuals $\hat{\epsilon}_x$: this is the component of X that is <u>not</u> explained by Z.

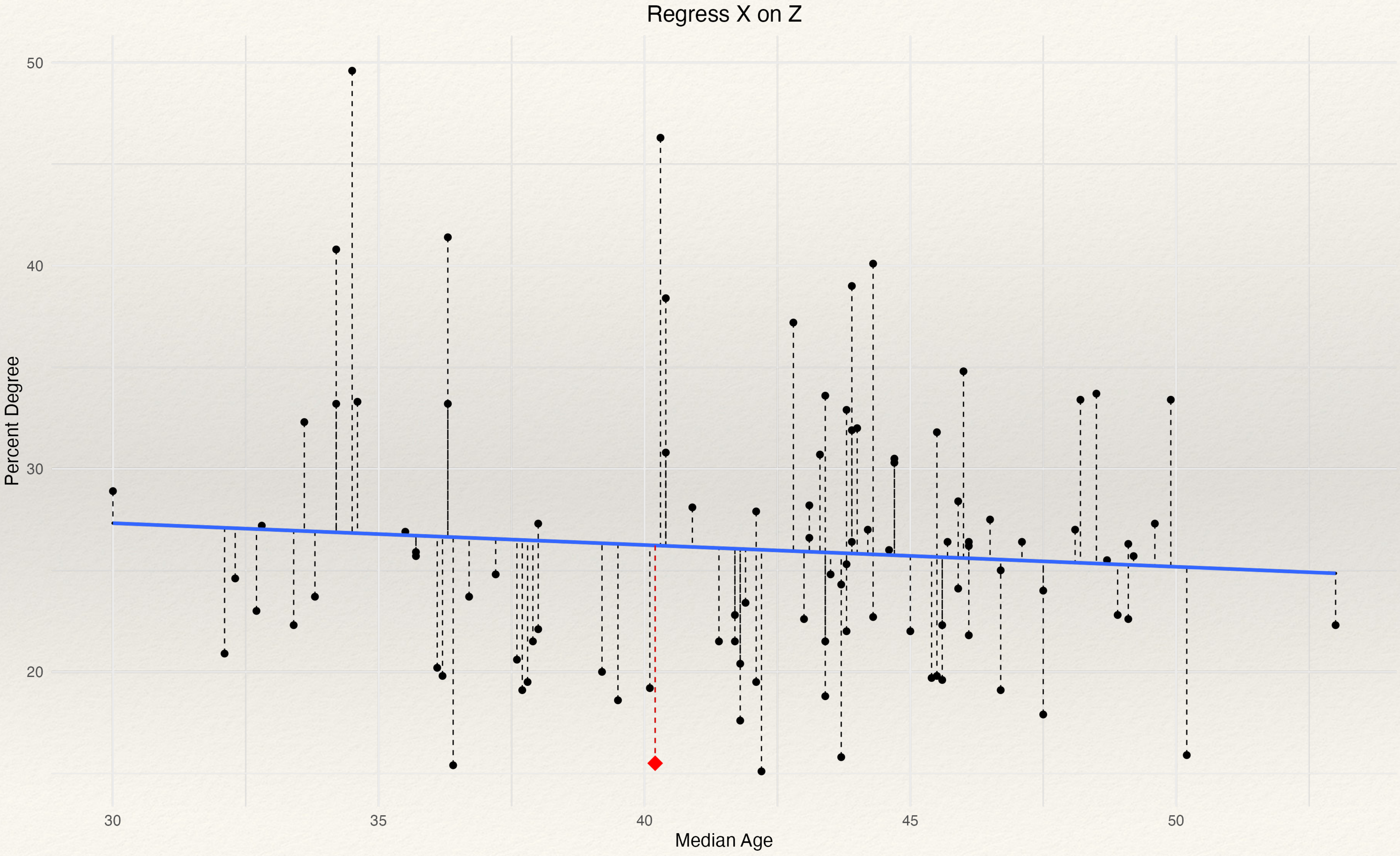
'Partialing Out' Interpretation

- Multiple OLS coefficient can be obtained via separate binary regressions (Frisch-Waugh-Lovell theorem):
  - Regress X on Z, extract the residuals $\hat{\epsilon}_x$: this is the component of X that is <u>not</u> explained by Z.
  - Regress Y on Z, extract the residuals $\hat{\epsilon}_y$: this is the component of Y that is <u>not</u> explained by Z.

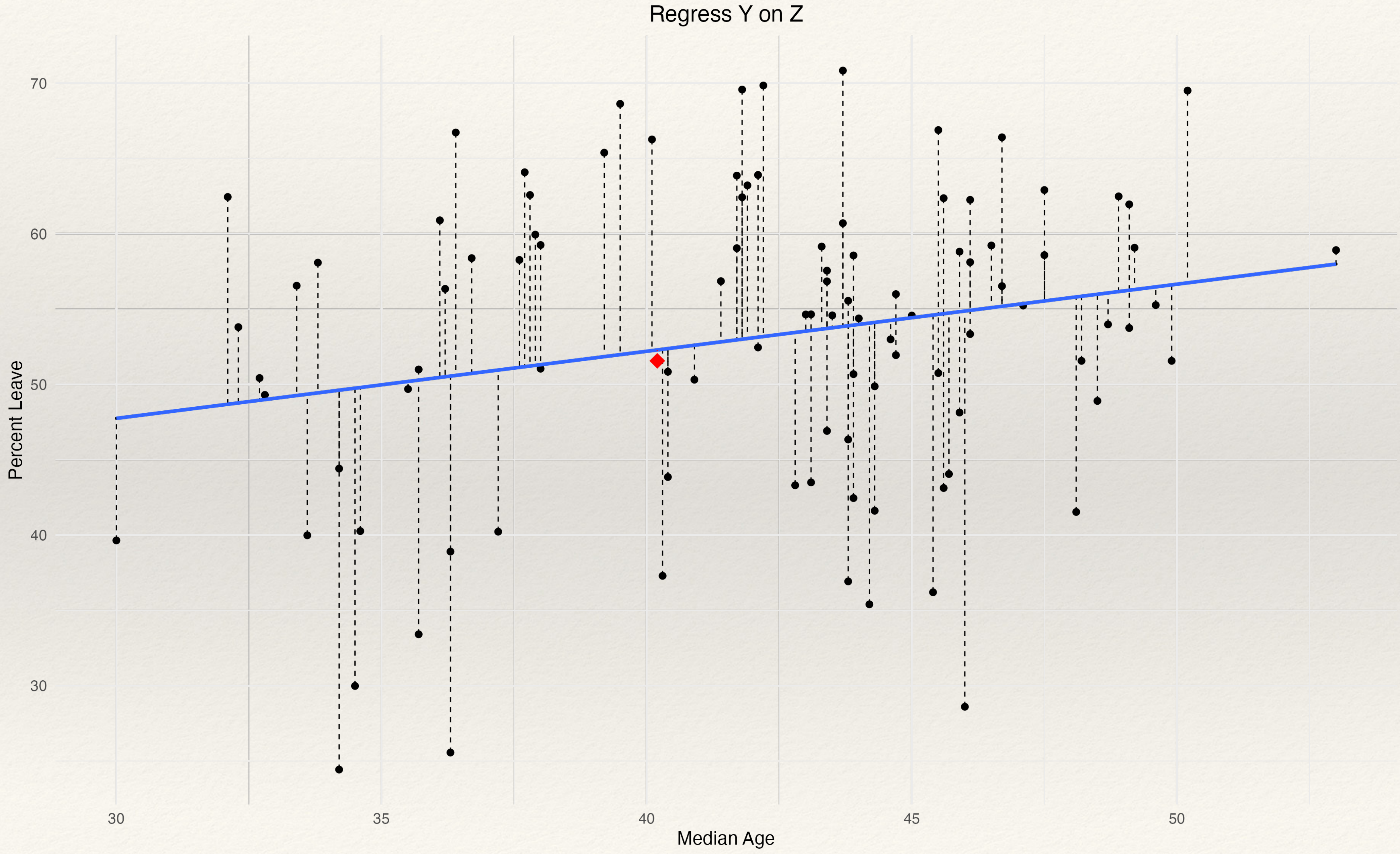
‘Partialing Out’ Interpretation

- Multiple OLS coefficient can be obtained via separate binary regressions (Frisch-Waugh-Lovell theorem):
  - Regress X on Z, extract the residuals $\hat{\epsilon}_x$: this is the component of X that is <u>not</u> explained by Z.
  - Regress Y on Z, extract the residuals $\hat{\epsilon}_y$: this is the component of Y that is <u>not</u> explained by Z.
  - Regress $\hat{\epsilon}_y$ on $\hat{\epsilon}_x$ — obtain β_1.

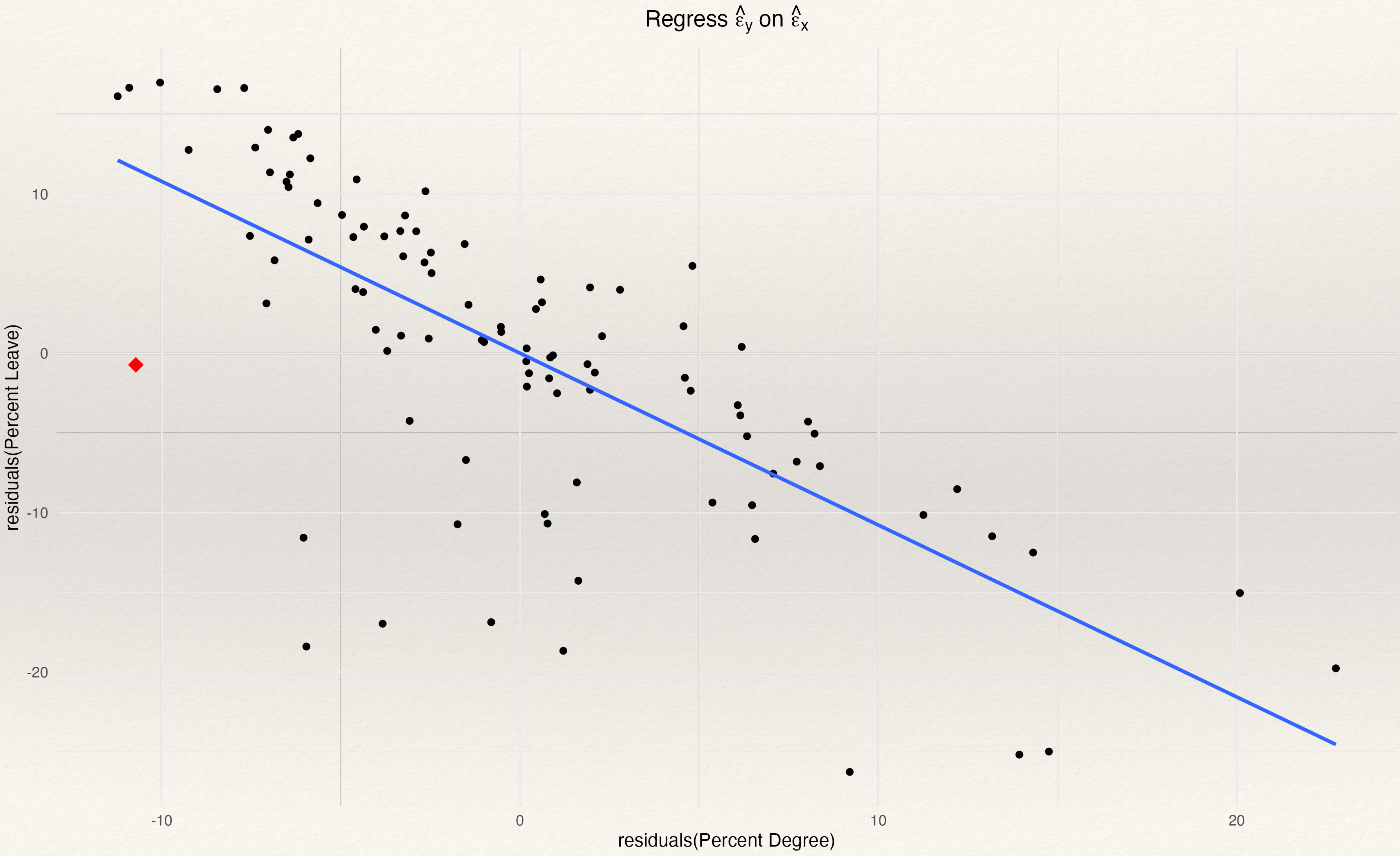
$$\text{Pct. Leave} = \alpha + \beta_1 \text{Pct. Degree} + \beta_2 \text{Median Age} + \epsilon$$



$$\text{Pct. Leave} = \alpha + \beta_1 \text{Pct. Degree} + \beta_2 \text{Median Age} + \epsilon$$



Pct. Leave = $\alpha + \beta_1$Pct. Degree + β_2Median Age + ϵ



‘Partialing Out’ Interpretation

```
# Multiple OLS
    model4 <- lm(data = brexit, percent_leave ~ percent_degree + median_age)
    coef(model4)[2]

## percent_degree
##      -1.078349

# Regress X on Z, extract residuals
    residuals_degree <- residuals(lm(data = brexit, percent_degree ~ median_age))
# Regress Y on Z, extract residuals
    residuals_leave <- residuals(lm(data = brexit, percent_leave ~ median_age))
# Regress residuals of Y on residuals of X
    residuals_regression <- lm(residuals_leave ~ residuals_degree)
    coef(residuals_regression)[2]

## residuals_degree
##        -1.078349
```

More Predictors!

More Predictors!

- Same story, more *X*s:

More Predictors!

- Same story, more Xs:
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_n X_n + \epsilon$

More Predictors!

- Same story, more Xs:
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_n X_n + \epsilon$

$$\text{Pct. Leave} = \alpha + \beta_1 \text{Pct. Degrees} + \beta_2 \text{Scotland}$$
$$+\beta_3 \text{Median Age} + \beta_4 \text{Weekly Earnings} + \epsilon$$

More Predictors!

- Same story, more Xs:
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_n X_n + \epsilon$

$$\text{Pct. Leave} = \alpha + \beta_1 \text{Pct. Degrees} + \beta_2 \text{Scotland}$$
$$+ \beta_3 \text{Median Age} + \beta_4 \text{Weekly Earnings} + \epsilon$$

More Predictors!

- Same story, more *X*s:
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_n X_n + \epsilon$

$$\text{Pct. Leave} = \alpha + \beta_1 \text{Pct. Degrees} + \beta_2 \text{Scotland}$$
$$+ \beta_3 \text{Median Age} + \beta_4 \text{Weekly Earnings} + \epsilon$$

- Harder to interpret geometrically: "fitting hyperplanes through multi-dimensional clouds of data points" (?)

More Predictors!

- Same story, more Xs:
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_n X_n + \epsilon$

$$\text{Pct. Leave} = \alpha + \beta_1 \text{Pct. Degrees} + \beta_2 \text{Scotland}$$
$$+ \beta_3 \text{Median Age} + \beta_4 \text{Weekly Earnings} + \epsilon$$

- Harder to interpret geometrically: "fitting hyperplanes through multi-dimensional clouds of data points" (?)
- Partialing out interpretation: β_1 as the effect of the component of X_1 that is uncorrelated with X_2, X_3, X_4 on the component of Y that is uncorrelated with X_2, X_3, X_4.

Multiple OLS in R

```
## Call:
## lm(formula = percent_leave ~ percent_degree + scotland + median_age +
##     median_earnings, data = brexit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2901  -2.2878   0.5524   2.7745   9.5145
##
## Coefficients:
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      56.225003   5.329834  10.549   < 2e-16 ***
## percent_degree   -1.203281   0.086124 -13.971   < 2e-16 ***
## scotland        -16.079284   1.228432 -13.089   < 2e-16 ***
## median_age        0.380043   0.082719   4.594 0.0000133 ***
## median_earnings   0.027371   0.009446   2.897   0.00467 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.122 on 95 degrees of freedom
## Multiple R-squared:  0.8436, Adjusted R-squared:  0.837
## F-statistic: 128.1 on 4 and 95 DF,  p-value: < 2.2e-16
```

Multiple OLS in R

```
stargazer(model5, type = "text", single.row = T)

##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                            percent_leave
## -----------------------------------------------
## percent_degree          -1.203*** (0.086)
## scotland               -16.079*** (1.228)
## median_age              0.380*** (0.083)
## median_earnings         0.027*** (0.009)
## Constant                56.225*** (5.330)
## -----------------------------------------------
## Observations                   100
## R2                            0.844
## Adjusted R2                   0.837
## Residual Std. Error       4.122 (df = 95)
## F Statistic         128.105*** (df = 4; 95)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Categorical Predictors

Categorical Predictors

* So far, our Xs have been interval or 0-1 binary variables. What if we want to control for a categorical variable?

Categorical Predictors

* So far, our Xs have been interval or 0-1 binary variables. What if we want to control for a categorical variable?
* Trick: a categorical variable with n categories can be recoded as $n - 1$ binary variables.

Categorical Predictors

* So far, our Xs have been interval or 0-1 binary variables. What if we want to control for a categorical variable?
* Trick: a categorical variable with n categories can be recoded as $n - 1$ binary variables.
    * Scotland/non-Scotland $\rightarrow$ one 0-1 variable

Categorical Predictors

- So far, our Xs have been interval or 0-1 binary variables. What if we want to control for a categorical variable?
- Trick: a categorical variable with n categories can be recoded as $n - 1$ binary variables.
  - Scotland / non-Scotland $\rightarrow$ one 0-1 variable
  - Scotland / Wales / Rest of UK $\rightarrow$ two 0-1 variables: Scotland / non-Scotland, Wales / non-Wales.

Categorical Predictors

* So far, our Xs have been interval or 0-1 binary variables. What if we want to control for a categorical variable?
* Trick: a categorical variable with n categories can be recoded as $n - 1$ binary variables.
  * Scotland / non-Scotland → one 0-1 variable
  * Scotland / Wales / Rest of UK → two 0-1 variables: Scotland / non-Scotland, Wales / non-Wales.
  * Married / Divorced / Single / Widowed → three 0-1 variables.

Categorical Predictors

* So far, our Xs have been interval or 0-1 binary variables. What if we want to control for a categorical variable?
* Trick: a categorical variable with n categories can be recoded as $n - 1$ binary variables.
  * Scotland/non-Scotland → one 0-1 variable
  * Scotland/Wales/Rest of UK → two 0-1 variables: Scotland/non-Scotland, Wales/non-Wales.
  * Married/Divorced/Single/Widowed → three 0-1 variables.
* R does this automatically when we pass a categorical predictors in the `lm()` function.

$$\text{Life Satisfaction (0-10)} = \alpha + \beta_1\text{Divorced} + \beta_2\text{Widowed} + \beta_3\text{Single} + \epsilon$$

```
##
## Call:
## lm(formula = life_satisf ~ marital_status, data = ess)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5697 -0.7800  0.4303  1.4303  3.2200
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.56966    0.06361 118.994  < 2e-16 ***
## marital_status divorced -0.78966    0.29810  -2.649  0.00814 **
## marital_status single   -0.57286    0.10413  -5.501 4.27e-08 ***
## marital_status widowed  -0.50299    0.15570  -3.231  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

$$\text{Life Satisfaction (0-10)} = \alpha + \beta_1 \text{Divorced} + \beta_2 \text{Widowed} + \beta_3 \text{Single} + \epsilon$$

```
##
## Call:
## lm(formula = life_satisf ~ marital_status, data = ess)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5697 -0.7800  0.4303  1.4303  3.2200
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.56966    0.06361 118.994  < 2e-16 ***
## marital_status divorced -0.78966    0.29810  -2.649  0.00814 **
## marital_status single   -0.57286    0.10413  -5.501 4.27e-08 ***
## marital_status widowed  -0.50299    0.15570  -3.231  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

- Why no coefficient for `married`?

Categorical Predictors

Categorical Predictors

- Regression coefficients for categorical predictors are interpreted relative to the ‘reference category’.

Categorical Predictors

- Regression coefficients for categorical predictors are interpreted relative to the 'reference category'.
- Mathematically, makes no difference which value you choose as 'reference category'. Sometimes, it makes sense to choose one *for presentational purposes*:

Categorical Predictors

* Regression coefficients for categorical predictors are interpreted relative to the ‘reference category’.
* Mathematically, makes no difference which value you choose as ‘reference category’. Sometimes, it makes sense to choose one *for presentational purposes*:
    * e.g. if I’m interested in how unemployment affects people’s attitudes, I will be more interested in the coefficient for `unemployed` relative to `employed` rather than e.g. `student` or `pensioner`.

Goodness of Fit in Simple OLS

Goodness of Fit in Simple OLS

- How good is our model?

Goodness of Fit in Simple OLS

- How good is our model?
- R^2 (coefficient of determination) = measure of 'fit'. The logic is comparing…

Goodness of Fit in Simple OLS

- How good is our model?
- R^2 (coefficient of determination) = measure of 'fit'. The logic is comparing…
- Our best guess of Y *without* the model: $\bar{Y}$ (the mean).

Goodness of Fit in Simple OLS

- How good is our model?
- R^2 (coefficient of determination) = measure of 'fit'. The logic is comparing…
- Our best guess of Y *without* the model: $\bar{Y}$ (the mean).
- Our best guess of Y *with* the model: $\hat{Y}$ (the fitted values).

Goodness of Fit in Simple OLS

Goodness of Fit in Simple OLS

- How far our predictions are from the observed *Y*s?

Goodness of Fit in Simple OLS

- How far our predictions are from the observed Ys?
- Before fitting the line $= \sum (Y_i - \bar{Y}_i)^2$

Goodness of Fit in Simple OLS

- How far our predictions are from the observed Ys?
- Before fitting the line = $\sum (Y_i - \bar{Y}_i)^2$
- After fitting the line = $\sum (Y_i - \hat{Y}_i)^2$

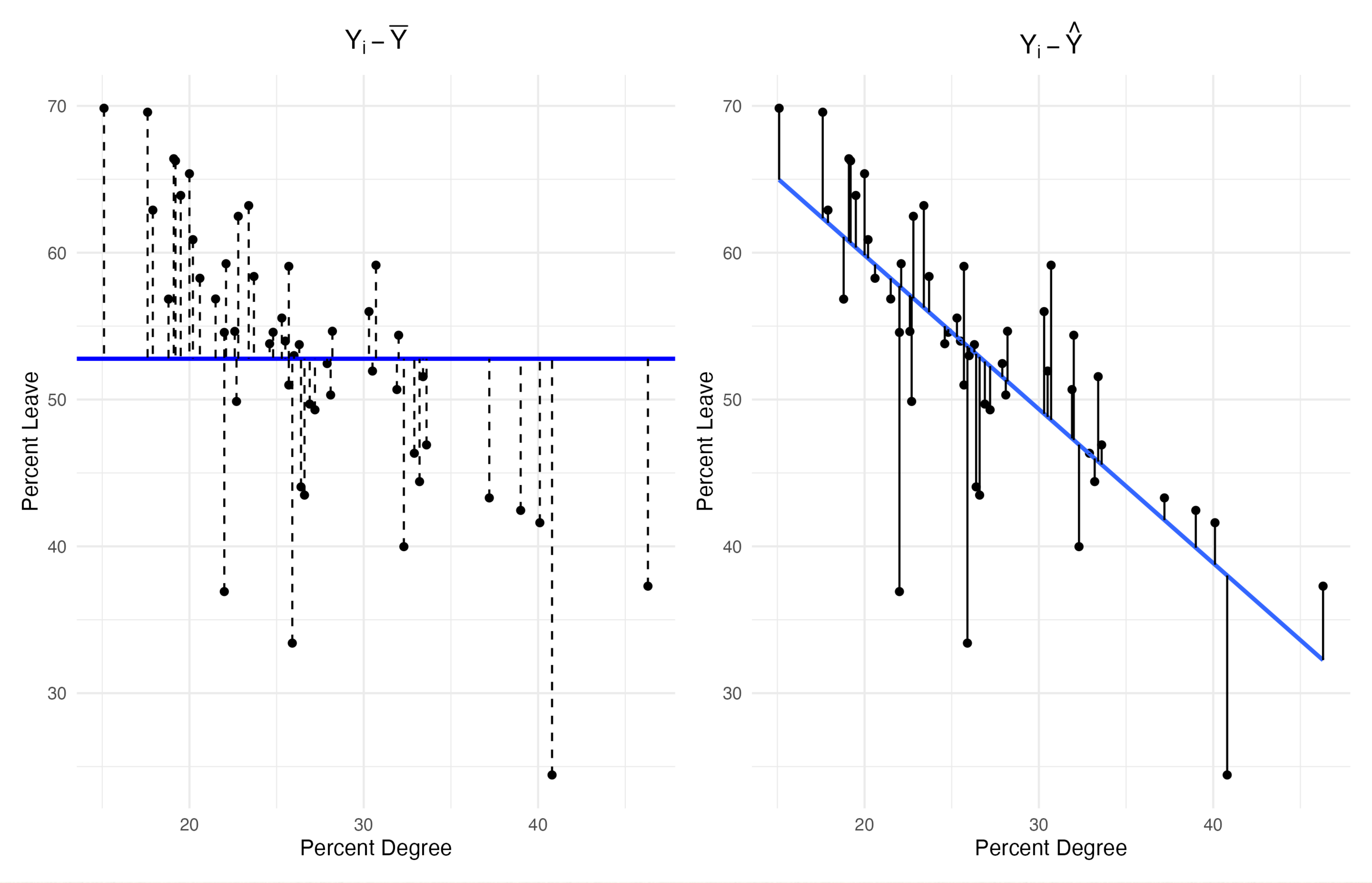
Goodness of Fit in Simple OLS

- How far our predictions are from the observed Ys?
- Before fitting the line $= \sum (Y_i - \bar{Y}_i)^2$
- After fitting the line $= \sum (Y_i - \hat{Y}_i)^2$
- So, the $R^2 = 1 - \dfrac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2}$

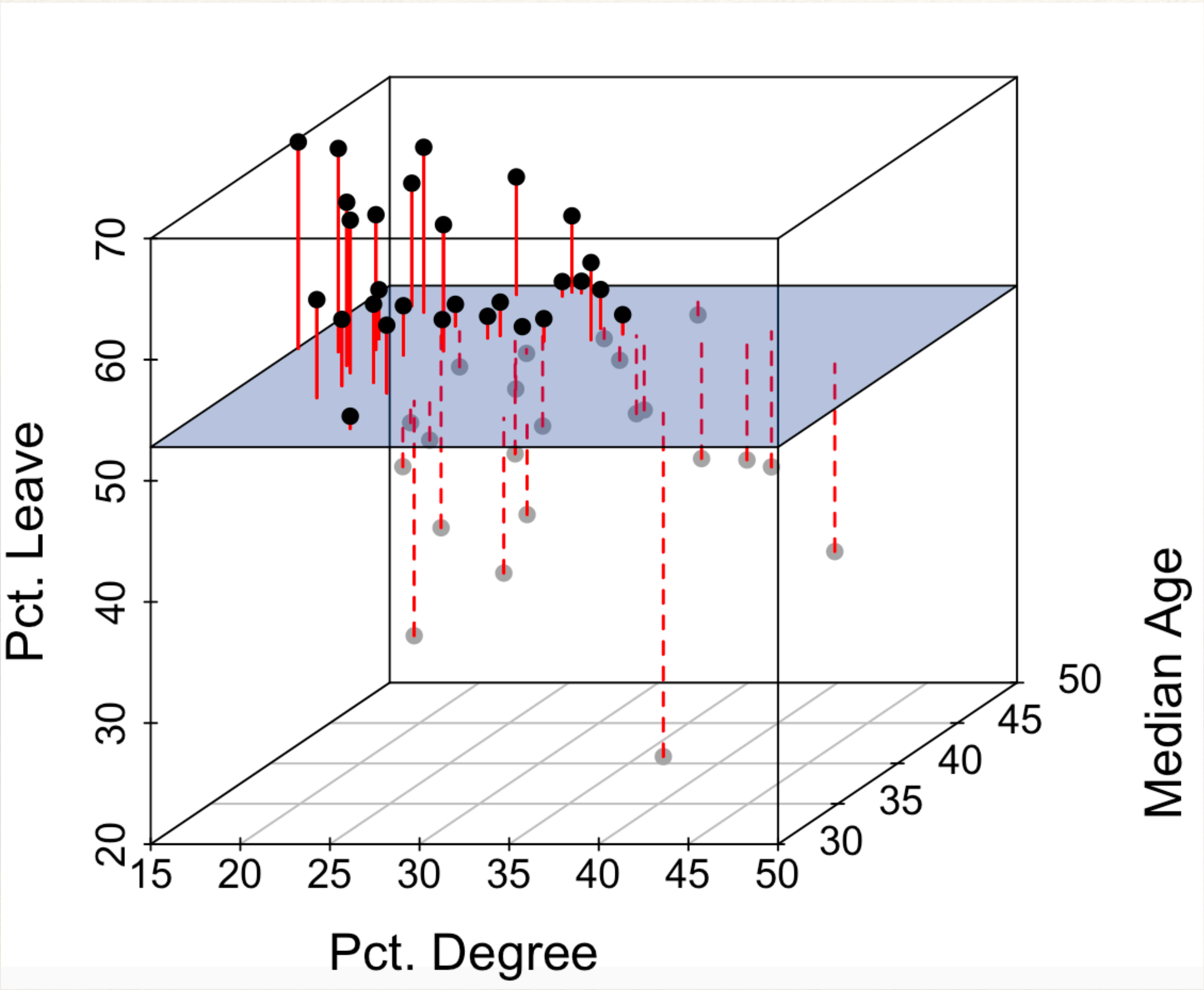
Goodness of Fit in Simple OLS

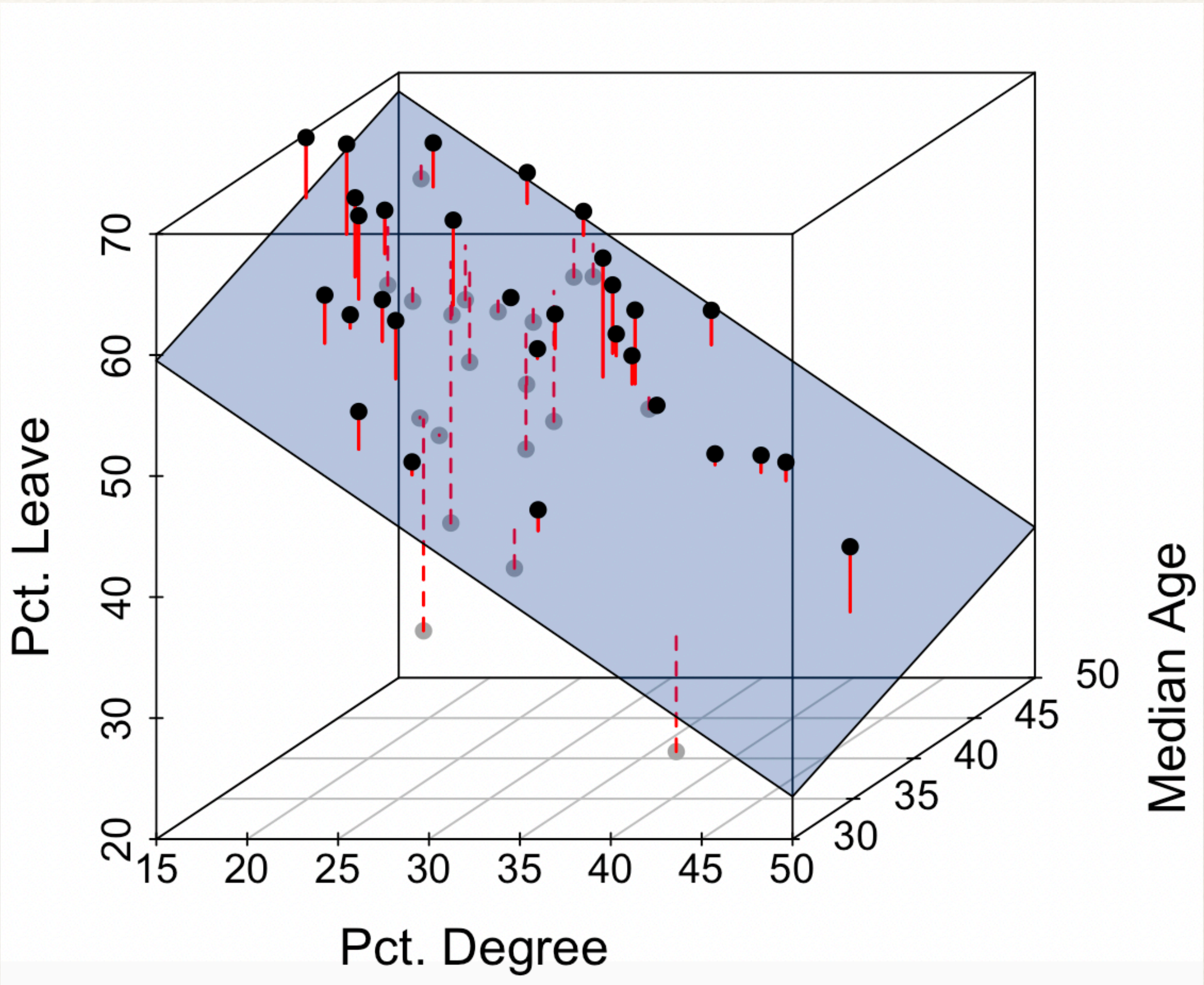
- How far our predictions are from the observed Ys?
- Before fitting the line = $\sum (Y_i - \bar{Y}_i)^2$
- After fitting the line = $\sum (Y_i - \hat{Y}_i)^2$
- So, the $R^2 = 1 - \dfrac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2}$
- Interpretation: value between 0 and 1 that tells us the **share of variance in *Y* explained by the model.**

Goodness of Fit in Simple OLS



Goodness of Fit in Multiple OLS





Goodness of Fit in Multiple OLS

Goodness of Fit in Multiple OLS

- When you add variables, your model will get better at predicting outcomes within the sample.

Goodness of Fit in Multiple OLS

- When you add variables, your model will get better at predicting outcomes within the sample.
- With n observation, we can have $n - 1$ binary variables, and we'll get exact predictions of Y for each observation.

Goodness of Fit in Multiple OLS

* When you add variables, your model will get better at predicting outcomes within the sample.
* With n observation, we can have $n - 1$ binary variables, and we'll get exact predictions of Y for each observation.
* $\hat{Y} = Y$ and R^2 will be 1. Perfect Fit. Hooray!

```
> lm(data = brexit, percent_leave ~ area)

Call:
lm(formula = percent_leave ~ area, data = brexit)

Coefficients:
                  (Intercept)                       areaAdur                       areaArun                  areaAshfield      areaBarking and Dagenham
                        38.90                          15.67                          23.58                         30.94                         23.54
                 areaBasildon                 areaBirmingham      areaBlackburn with Darwen                  areaBolsover                     areaBrent
                        29.72                          11.52                          17.44                         31.93                          1.36
                areaBrentwood                   areaBridgend                 areaBroxbourne                 areaBroxtowe                areaCanterbury
                        20.25                          15.74                          27.36                         15.75                         12.14
                  areaCardiff                  areaCherwell        areaCity of Edinburgh                    areaConwy                  areaCornwall
                         1.08                          11.41                         -13.34                         15.08                         17.62
                areaCotswold              areaCounty Durham          areaDerbyshire Dales              areaDundee City     areaEast Dunbartonshire
                        10.00                          18.65                          12.66                          1.32                        -10.30
            areaEast Lothian         areaEast Staffordshire                 areaEastleigh          areaForest of Dean                 areaGateshead
                        -3.50                          24.31                          13.55                         19.68                         17.95
                  areaGedling               areaGlasgow City                  areaGosport                 areaGravesham                 areaGreenwich
                        16.65                          -5.49                          24.96                         26.48                          5.51
   areaHammersmith and Fulham               areaHarborough                 areaHaringey                areaHartlepool                    areaHavant
                        -8.92                          11.85                         -14.47                         30.67                         23.46
 areaHerefordshire, County of                areaHertsmere                 areaHighland                areaInverclyde                   areaIpswich
                        20.32                          11.94                           5.15                         -2.70                         19.36
          areaIsle of Wight  areaKing's Lynn and West Norfolk               areaKnowsley                     areaLeeds                 areaLichfield
                        23.05                          27.50                          12.66                         10.79                         19.91
                    areaLuton                 areaManchester                   areaMedway                    areaMelton                 areaMid Devon
                        17.65                           0.74                          25.18                         19.21                         14.44
              areaMid Suffolk                 areaMid Sussex                    areaMoray         areaNeath Port Talbot      areaNewcastle upon Tyne
                        16.33                           8.01                          10.97                         17.94                         10.40
           areaNorth Ayrshire              areaNorth Norfolk       areaNorth Warwickshire  areaNorth West Leicestershire            areaNorthampton
                         4.22                          20.01                          27.98                         21.80                         19.48
        areaOadby and Wigston              areaPeterborough                 areaPlymouth                areaPortsmouth                     areaPowys
                        15.68                          21.99                          21.04                         19.18                         14.84
                  areaPurbeck                areaRushcliffe                 areaRyedale                  areaSandwell         areaScottish Borders
                        20.17                           3.55                          16.36                         27.82                          2.63
                   areaSefton                 areaSevenoaks                areaSheffield                  areaShepway         areaShetland Islands
                         9.23                          15.48                          12.09                         23.35                          4.59
        areaSouth Lanarkshire             areaSouth Ribble              areaSouthampton                areaSt Albans                  areaStafford
                        -1.98                          19.66                          14.90                         -1.61                         17.09
                areaStevenage        areaStratford-on-Avon                 areaTendring               areaTest Valley                 areaUttlesford
                        20.35                          12.66                          30.60                         13.04                         11.78
     areaVale of White Horse                   areaWaveney                 areaWaverley          areaWellingborough        areaWest Oxfordshire
                         4.40                          24.00                           2.71                         23.52                          7.44
                    areaWigan                    areaWoking            areaWolverhampton                  areaWorthing                   areaWrexham
                        25.00                           4.95                          23.67                         14.09                         20.14
```

Goodness of Fit in Multiple OLS

- When you add variables, your model will get better at predicting outcomes within the sample.
- With n observation, we can have $n - 1$ binary variables, and we'll get exact predictions of Y for each observation.
- $\hat{Y} = Y$ and R^2 will be 1. Perfect Fit. Hooray!

Goodness of Fit in Multiple OLS

- When you add variables, your model will get better at predicting outcomes within the sample.
- With n observation, we can have $n - 1$ binary variables, and we'll get exact predictions of Y for each observation.
- $\hat{Y} = Y$ and R^2 will be 1. Perfect Fit. Hooray!
- This is obviously silly:

Goodness of Fit in Multiple OLS

- When you add variables, your model will get better at predicting outcomes within the sample.
- With n observation, we can have $n - 1$ binary variables, and we'll get exact predictions of Y for each observation.
- $\hat{Y} = Y$ and R^2 will be 1. Perfect Fit. Hooray!
- This is obviously silly:
  - Our models are meant to simplify reality (*parsimony*).

Goodness of Fit in Multiple OLS

- When you add variables, your model will get better at predicting outcomes within the sample.
- With n observation, we can have $n - 1$ binary variables, and we'll get exact predictions of Y for each observation.
- $\hat{Y} = Y$ and R^2 will be 1. Perfect Fit. Hooray!
- This is obviously silly:
  - Our models are meant to simplify reality (*parsimony*).
  - Also, we can't make out-of-sample predictions.

Goodness of Fit in Multiple OLS

Goodness of Fit in Multiple OLS

- Adjusted R^2 penalises models with lots of predictors that explain little variation in Y:

Goodness of Fit in Multiple OLS

* Adjusted R^2 penalises models with lots of predictors that explain little variation in Y:

$$\text{Adj. } R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Goodness of Fit in Multiple OLS

- Adjusted R^2 penalises models with lots of predictors that explain little variation in Y:

$$\text{Adj. } R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

  - where p = no. of predictors.

Goodness of Fit in Multiple OLS

- Adjusted R^2 penalises models with lots of predictors that explain little variation in Y:

$$\text{Adj. } R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

  - where p = no. of predictors.

- Interpret Adjusted R^2 in multiple regression models. But maximising the R^2 is not your main goal!

```
## ===================================================================
##                                  Dependent variable:
##                      ----------------------------------------------
##                                      life_satisf
##                                (1)                      (2)
## -------------------------------------------------------------------
## age                           0.013                    0.013
## income_decile                 0.339                    0.336
## female                        0.519                    0.528
## religiosity                   0.094                    0.089
## years_education              -0.006                    -0.007
## star_signEarth                                         -0.009
## star_signFire                                          0.295
## star_signWater                                         -0.110
## favourite_number                                       0.006
## Intercept                     4.127                    4.095
## -------------------------------------------------------------------
## Observations                   210                      210
## R2                            0.214                    0.219
## Adjusted R2                   0.195                    0.184
## Residual Std. Error     1.983 (df = 204)         1.996 (df = 200)
## F Statistic         11.123*** (df = 5; 204) 6.228*** (df = 9; 200)
## ===================================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

Least Squares Assumptions: An Essential Checklist

‘BLUE’

‘BLUE’

- Gauss-Markov Theorem: OLS is the **B**est **L**inear **U**nbiased **E**stimator (**BLUE**) under **six assumptions**.
- **Unbiased**: OLS estimate will recover the population parameter in expectation (assumptions 1-4):

$$\mathrm{E}[\hat{\beta}] = \beta$$

- **Best**: out of all the possible unbiased estimators of a linear relationship, OLS is the one with least variance: i.e. it works comparatively well even in small samples (assumptions 5-6, next week).

Classical Linear Model Assumptions

Classical Linear Model Assumptions

1. Linearity

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. No Perfect Collinearity

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean (Exogeneity)

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean (Exogeneity)
5. *Constant variance of the error term (Homoskedasticity)*

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean (Exogeneity)
5. *Constant variance of the error term (Homoskedasticity)*
6. *Normality of the Error Term*

Classical Linear Model Assumptions

1. **Linearity**
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean (Exogeneity)
5. *Constant variance of the error term (Homoskedasticity)*
6. *Normality of the Error Term*

Linearity

Linearity

- The population regression model is linear in its parameters (“linear in the βs”).

Linearity

* The population regression model is linear in its parameters ("linear in the βs").
* Remember: we don't observe the population regression. We just **assume** that the relationships that generate our data are linear → regression is a **model** of reality.

Linearity

- The population regression model is linear in its parameters ("linear in the βs").
- Remember: we don't observe the population regression. We just **assume** that the relationships that generate our data are linear → regression is a **model** of reality.
- If linear relationships *cannot* be assumed:

Linearity

* The population regression model is linear in its parameters ("linear in the βs").
* Remember: we don't observe the population regression. We just **assume** that the relationships that generate our data are linear → regression is a **model** of reality.
* If linear relationships *cannot* be assumed:
  * Tricks to model non-linear relations in OLS (week 8).

Linearity

* The population regression model is linear in its parameters (“linear in the βs”).
* Remember: we don’t observe the population regression. We just **assume** that the relationships that generate our data are linear → regression is a **model** of reality.
* If linear relationships *cannot* be assumed:
    * Tricks to model non-linear relations in OLS (week 8).
    * Use non-linear regression instead (beyond this course).

Classical Linear Model Assumptions

1. Linearity
2. **Random Sampling**
3. No Perfect Collinearity
4. Zero Conditional Mean (Exogeneity)
5. *Constant variance of the error term (Homoskedasticity)*
6. *Normality of the Error Term*

Random Sampling

Random Sampling

* Random sample: $(Y_i, X_{1i}, X_{2i} \dots X_{ki})$ are sampled randomly from the population for $i = 1 ... i = N$.

Random Sampling

- Random sample: $(Y_i, X_{1i}, X_{2i} \ldots X_{ki})$ are sampled randomly from the population for $i = 1 ... \; i = N$.
- If sampling is non-random, our estimates will be biased, i.e. they will not recover the population parameters.

Random Sampling

* Random sample: $(Y_i, X_{1i}, X_{2i} \ldots X_{ki})$ are sampled randomly from the population for $i = 1 \ldots i = N$.
* If sampling is non-random, our estimates will be biased, i.e. they will not recover the population parameters.
* Related problem: non-random **missing data**.

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. **No Perfect Collinearity**
4. Zero Conditional Mean (Exogeneity)
5. *Constant variance of the error term (Homoskedasticity)*
6. *Normality of the Error Term*

No Perfect Collinearity

No Perfect Collinearity

* None of the independent variables is **constant**, and there are no **exact linear relationships** among the independent variables.

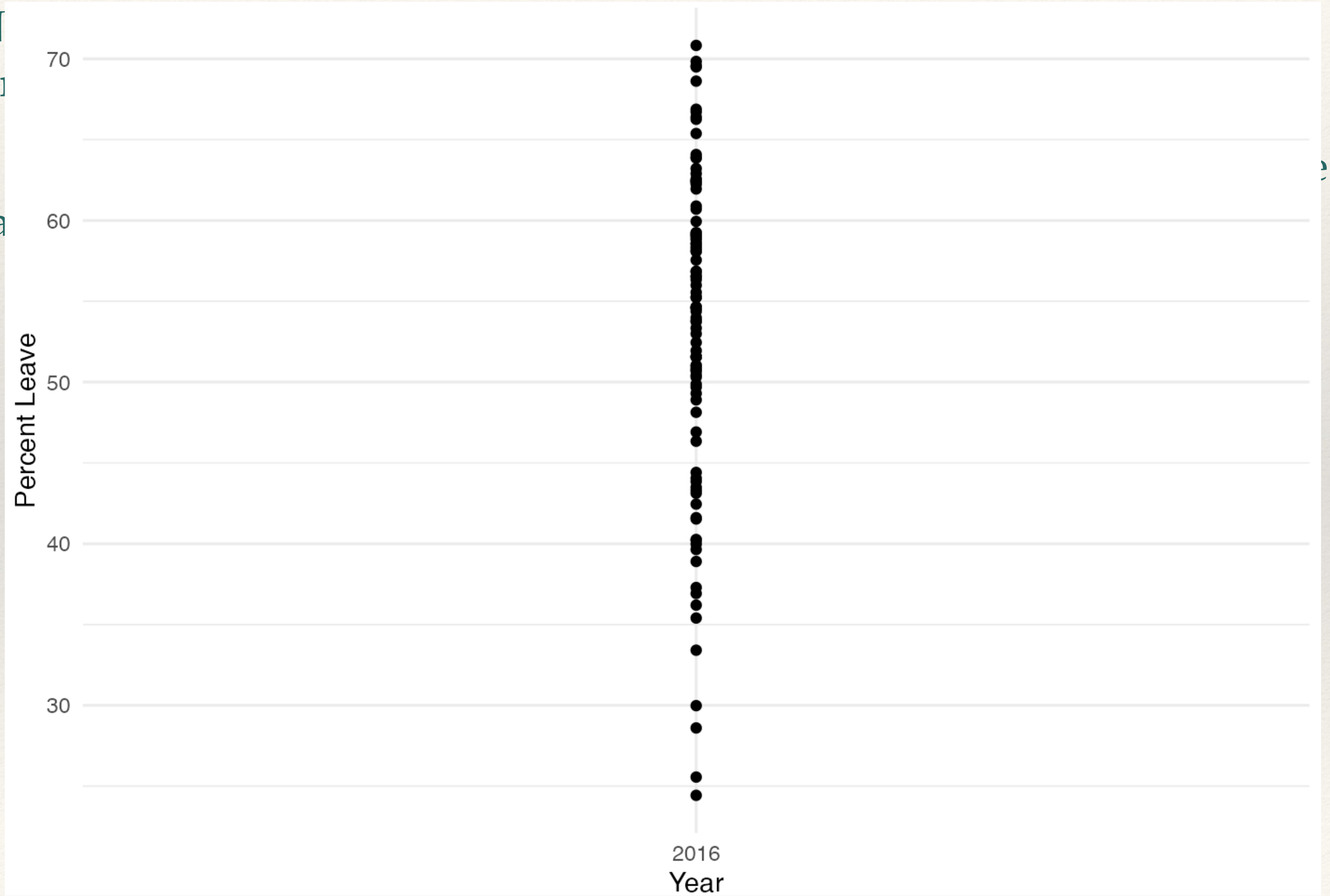
No Perfect Collinearity

* None of the independent variables is **constant**, and there are no **exact linear relationships** among the independent variables.
* If a variable is a constant: infinite regression lines solve OLS; impossible to say how changes in X affect Y because X doesn't change.

No Perfect Collinearity

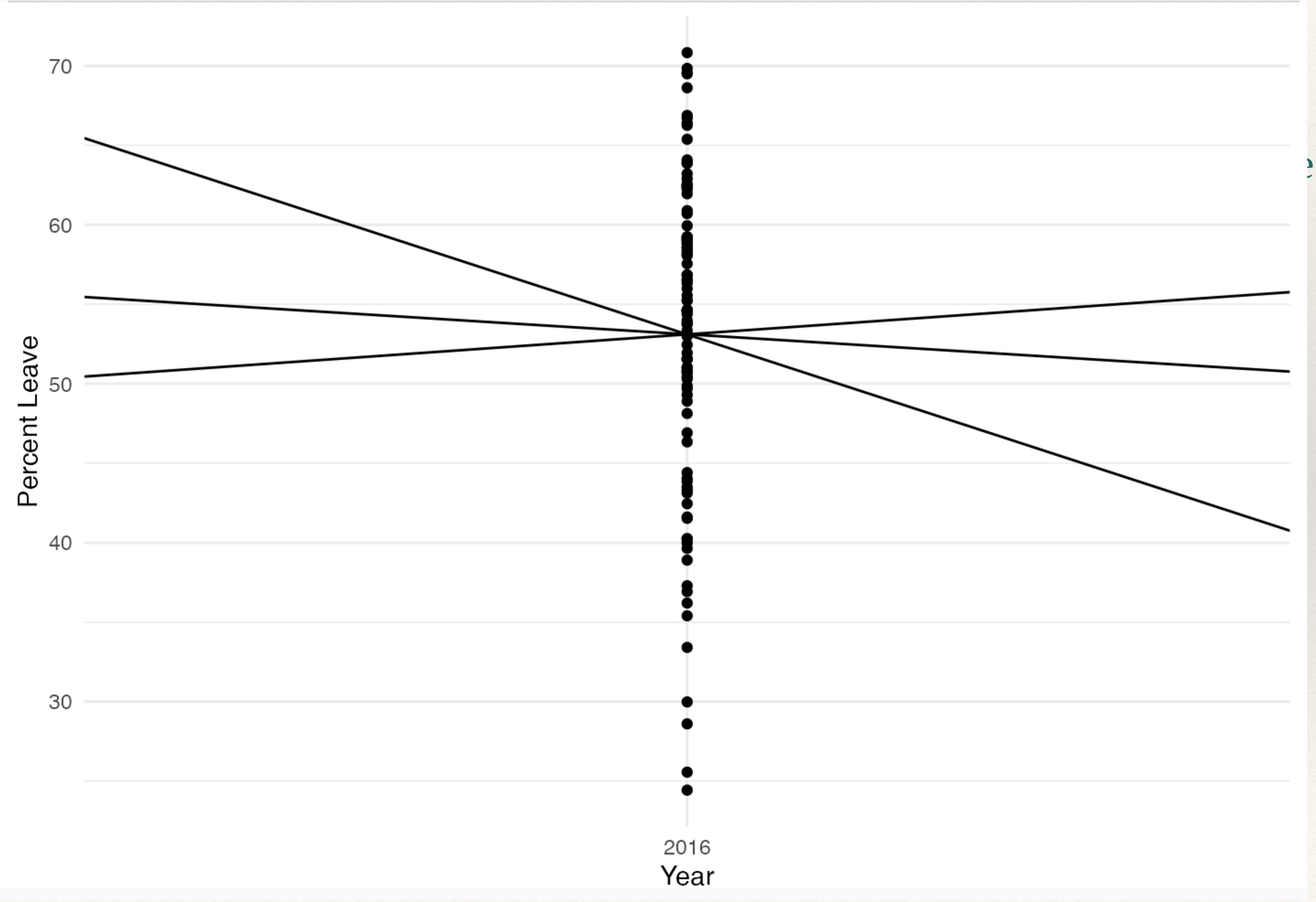
- N
  li

- If ... to
  sa



No Perfect Collinearity

* N
li
* If ... to
sa



No Perfect Collinearity

* None of the independent variables is **constant**, and there are no **exact linear relationships** among the independent variables.
* If a variable is a constant: infinite regression lines solve OLS; impossible to say how changes in X affect Y because X doesn't change.

No Perfect Collinearity

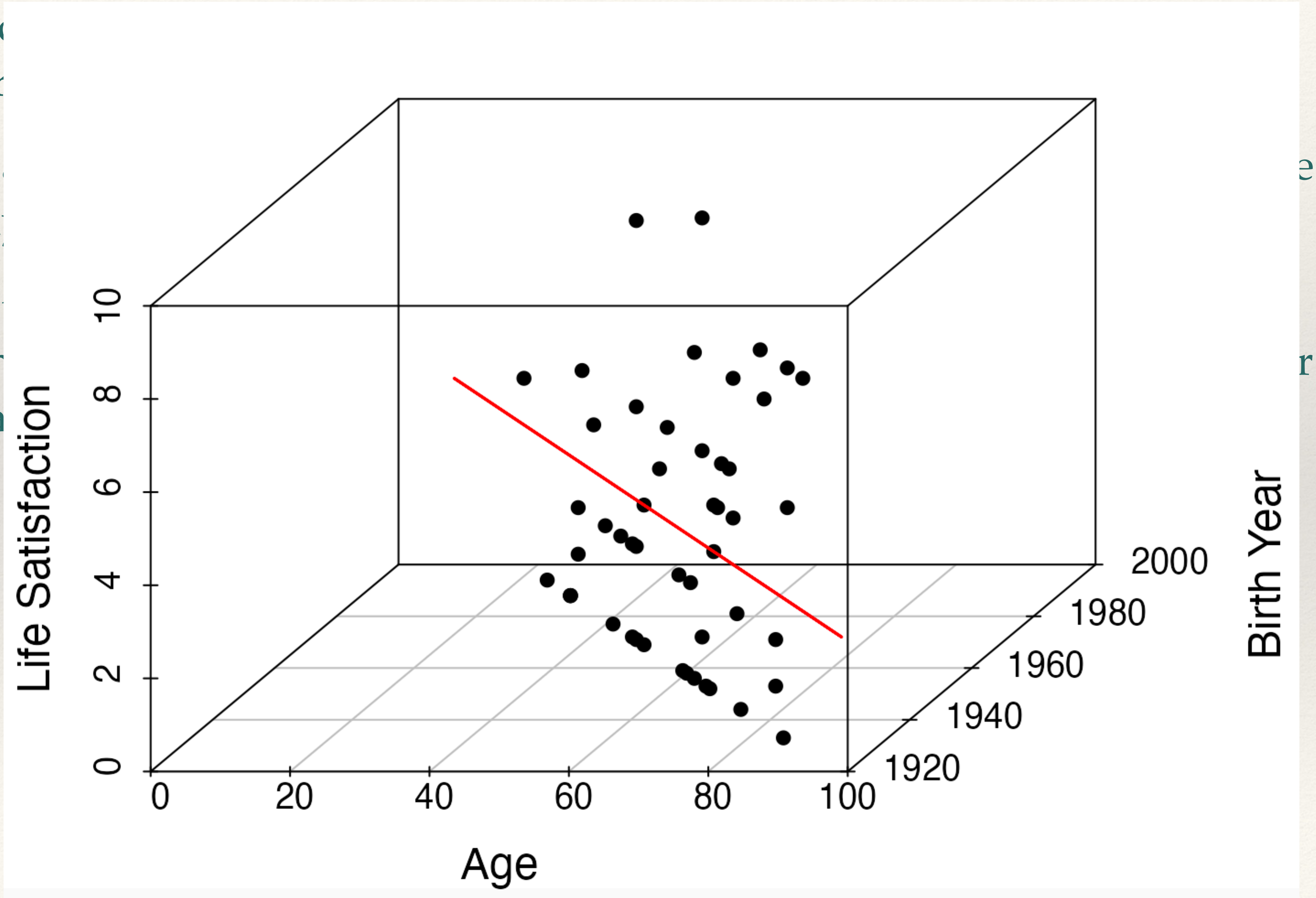
- None of the independent variables is **constant**, and there are no **exact linear relationships** among the independent variables.
- If a variable is a constant: infinite regression lines solve OLS; impossible to say how changes in X affect Y because X doesn't change.
- If two independent variables are a linear combination of each other: impossible to determine partial effects because X_1 perfectly accounts for variation in X_2, and vice versa.

No Perfect Collinearity

* None of the independent variables is **constant**, and there are no **exact linear relationships** among the independent variables.
* If a variable is a constant: infinite regression lines solve OLS; impossible to say how changes in X affect Y because X doesn't change.
* If two independent variables are a linear combination of each other: impossible to determine partial effects because X_1 perfectly accounts for variation in X_2, and vice versa.
  * Age and Birth year in cross-sectional data: Age = Current Year − Birth Year.

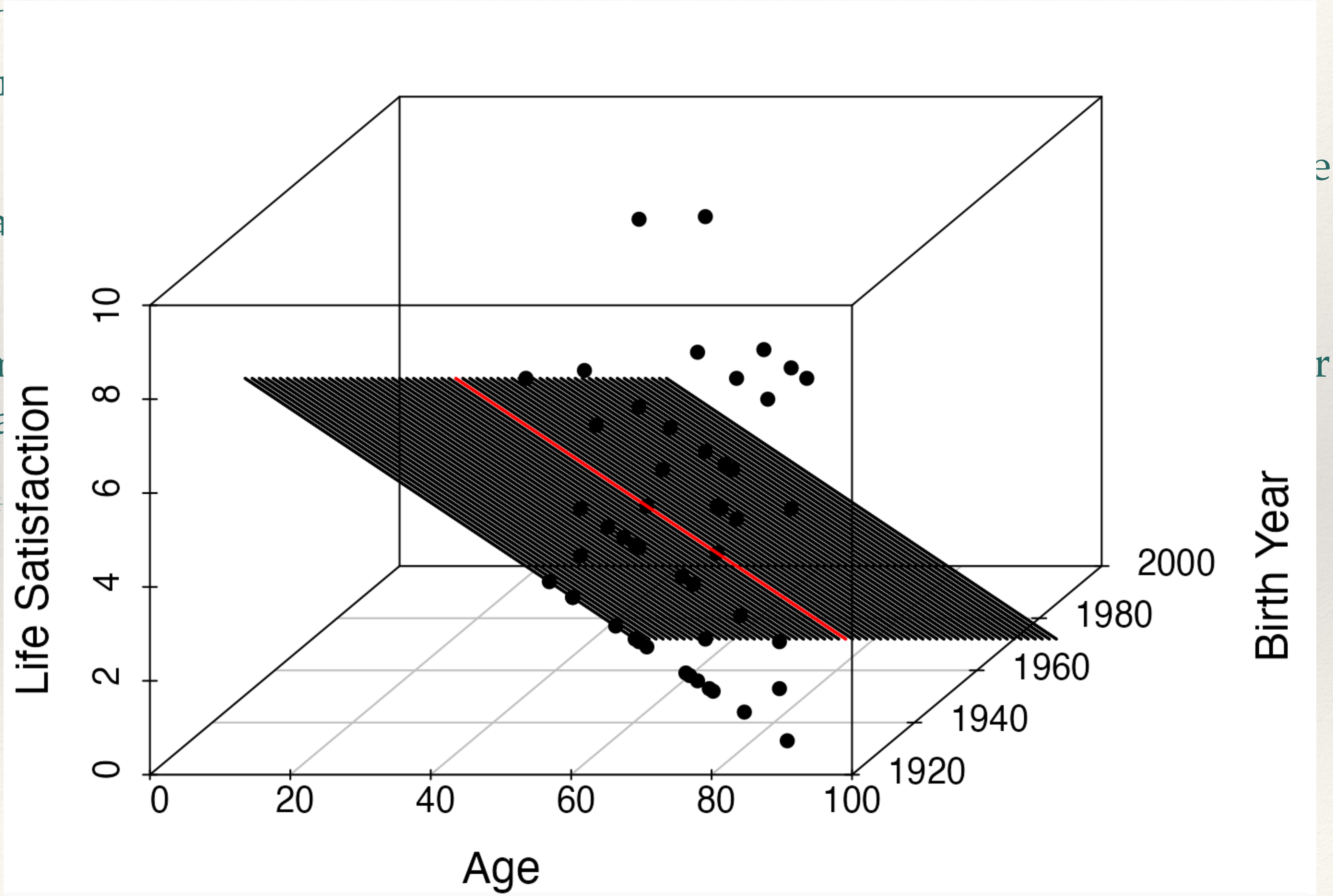
No Perfect Collinearity

* No
  lin
* If ... e to
  say
* If
  im ... r
  va
  *



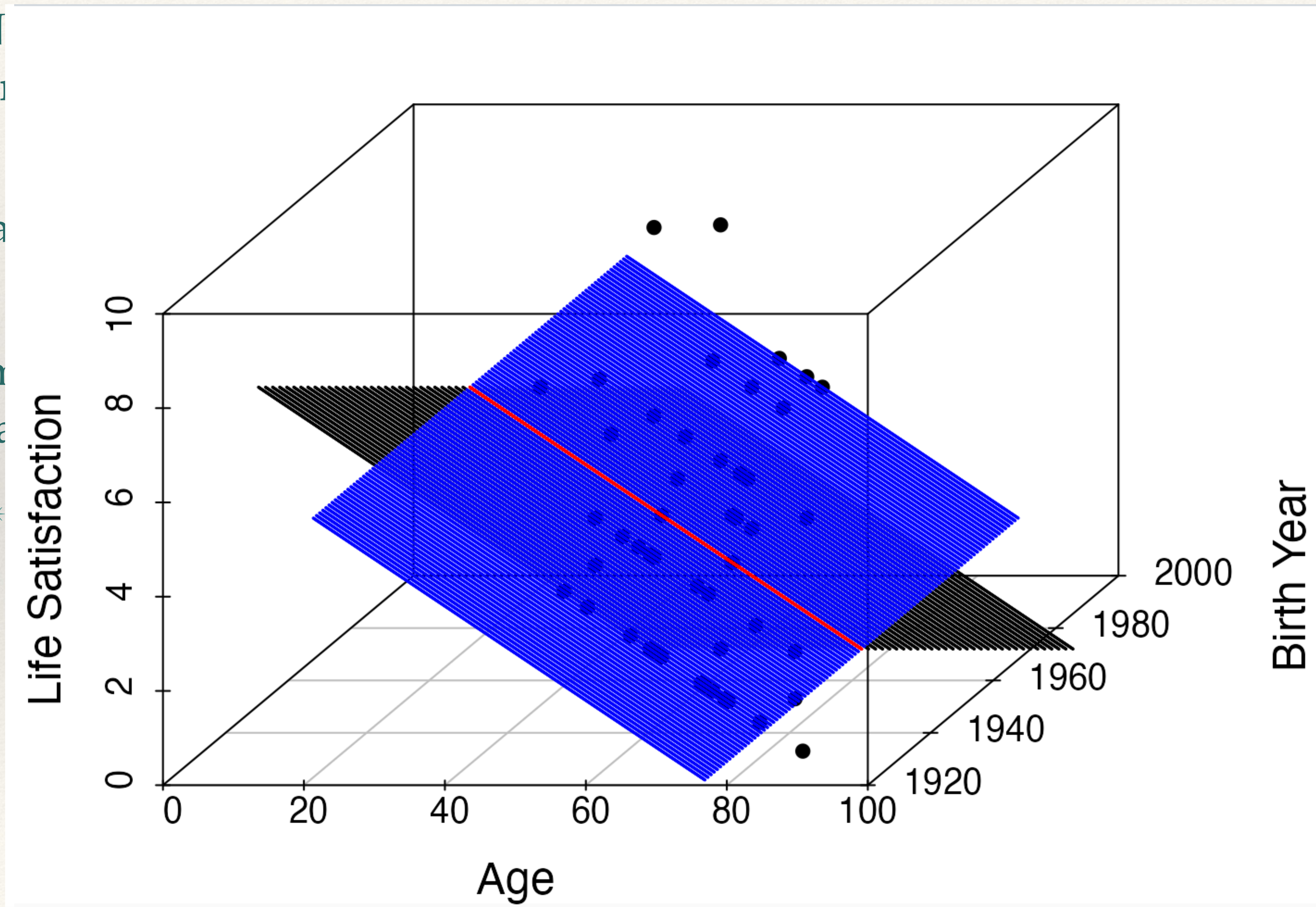
No Perfect Collinearity

* N
  li
* If … e to
  sa
* If … r
  in
  va
  *



No Perfect Collinearity

- N
  li
- If … e to
  sa
- If
  in … r
  va
  -



No Perfect Collinearity

- None of the independent variables is **constant**, and there are no **exact linear relationships** among the independent variables.
- If a variable is a constant: infinite regression lines solve OLS; impossible to say how changes in X affect Y because X doesn't change.
- If two independent variables are a linear combination of each other: impossible to determine partial effects because X_1 perfectly accounts for variation in X_2, and vice versa.
  - Age and Birth year in cross-sectional data:
    Age = Current Year − Birth Year.

No Perfect Collinearity

- None of the independent variables is **constant**, and there are no **exact linear relationships** among the independent variables.
- If a variable is a constant: infinite regression lines solve OLS; impossible to say how changes in X affect Y because X doesn't change.
- If two independent variables are a linear combination of each other: impossible to determine partial effects because X_1 perfectly accounts for variation in X_2, and vice versa.
  - Age and Birth year in cross-sectional data:
    Age = Current Year − Birth Year.
- Also the reason why we have $n - 1$ binary variables when we recode a categorical variable with n categories. For instance, Male (0-1) is a linear combination of Female (0-1):
  Male = −1(Female−1)

Multicollinearity

Multicollinearity

- **Multicollinearity:** independent variables are **very strongly**, but not perfectly, correlated with each other.

Multicollinearity

* **Multicollinearity:** independent variables are **very strongly**, but not perfectly, correlated with each other.
* This makes for **unstable estimates** across iterated samples from the population, **large S.E.** and **high p-values**.

Multicollinearity

- **Multicollinearity:** independent variables are **very strongly**, but not perfectly, correlated with each other.
- This makes for **unstable estimates** across iterated samples from the population, **large S.E.** and **high p-values**.
- **Not** a violation of OLS assumptions. Also, less well-defined issue than perfect collinearity. How strong is 'too strong'?

Multicollinearity

- **Multicollinearity:** independent variables are **very strongly**, but not perfectly, correlated with each other.
- This makes for **unstable estimates** across iterated samples from the population, **large S.E.** and **high p-values**.
- **Not** a violation of OLS assumptions. Also, less well-defined issue than perfect collinearity. How strong is 'too strong'?
- Diagnostic tool (VIF) in the lab. But you normally want to **avoid to include covariates that are tightly correlated**.

Multicollinearity

* **Multicollinearity:** independent variables are **very strongly**, but not perfectly, correlated with each other.
* This makes for **unstable estimates** across iterated samples from the population, **large S.E.** and **high p-values**.
* **Not** a violation of OLS assumptions. Also, less well-defined issue than perfect collinearity. How strong is 'too strong'?
* Diagnostic tool (VIF) in the lab. But you normally want to **avoid to include covariates that are tightly correlated**.
* Solutions: (1) increase the number of observations, (2) drop one of the variables affected, (3) Nothing. **OLS is still BLUE**.

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. No Perfect Collinearity
4. **Zero Conditional Mean (Exogeneity)**
5. *Constant variance of the error term (Homoskedasticity)*
6. *Normality of the Error Term*

Zero Conditional Mean (aka Exogeneity)

Zero Conditional Mean (aka Exogeneity)

- The population error ϵ has an expected value of zero (i.e. a mean across repeated samples) **given any values of the Xs**.

Zero Conditional Mean (aka Exogeneity)

- The population error ϵ has an expected value of zero (i.e. a mean across repeated samples) **given any values of the Xs**.
- Unexplained component of Y, which is modelled as part of the error term ϵ, should be uncorrelated with the Xs.

Zero Conditional Mean (aka Exogeneity)

- The population error ϵ has an expected value of zero (i.e. a mean across repeated samples) **given any values of the Xs**.
- Unexplained component of Y, which is modelled as part of the error term ϵ, should be uncorrelated with the Xs.
- In other words: **there are no un-modelled confounding variables**.

Zero Conditional Mean (aka Exogeneity)

* The population error ϵ has an expected value of zero (i.e. a mean across repeated samples) **given any values of the Xs**.

* Unexplained component of Y, which is modelled as part of the error term ϵ, should be uncorrelated with the Xs.

* In other words: **there are no un-modelled confounding variables**.

* By far the most important of OLS assumptions, and the one most often violated in practice.

Zero Conditional Mean (aka Exogeneity)

Zero Conditional Mean (aka Exogeneity)

* Trickiest violations:

Zero Conditional Mean (aka Exogeneity)

- Trickiest violations:
  - **Omitted Variable Bias**

Zero Conditional Mean (aka Exogeneity)

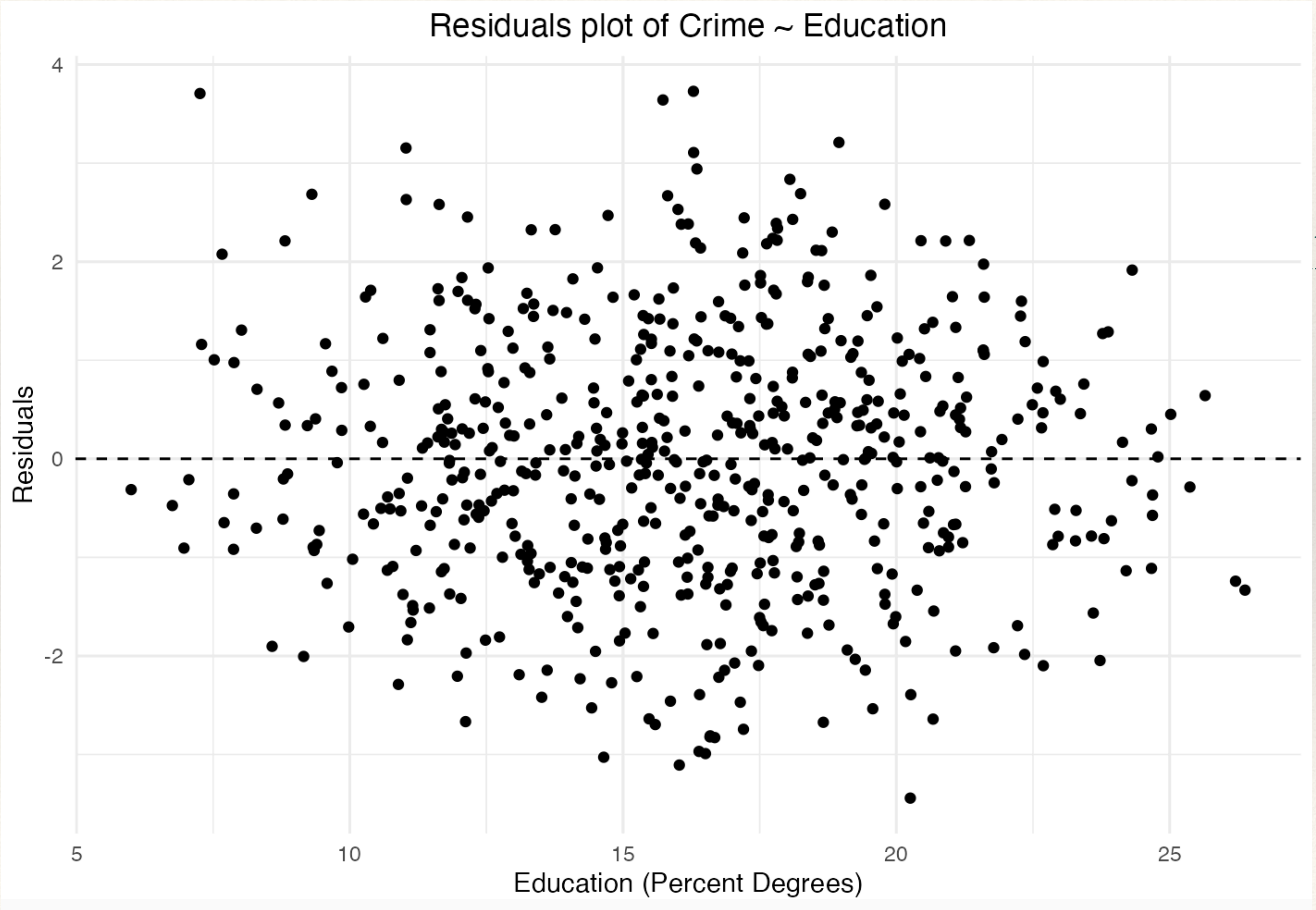
- Trickiest violations:
  - **Omitted Variable Bias**
    - If there is a Z correlated with Y *and* with X but is not modelled, then ϵ will be correlated with X.

Zero Conditional Mean (aka Exogeneity)

* Tr

  *

led,

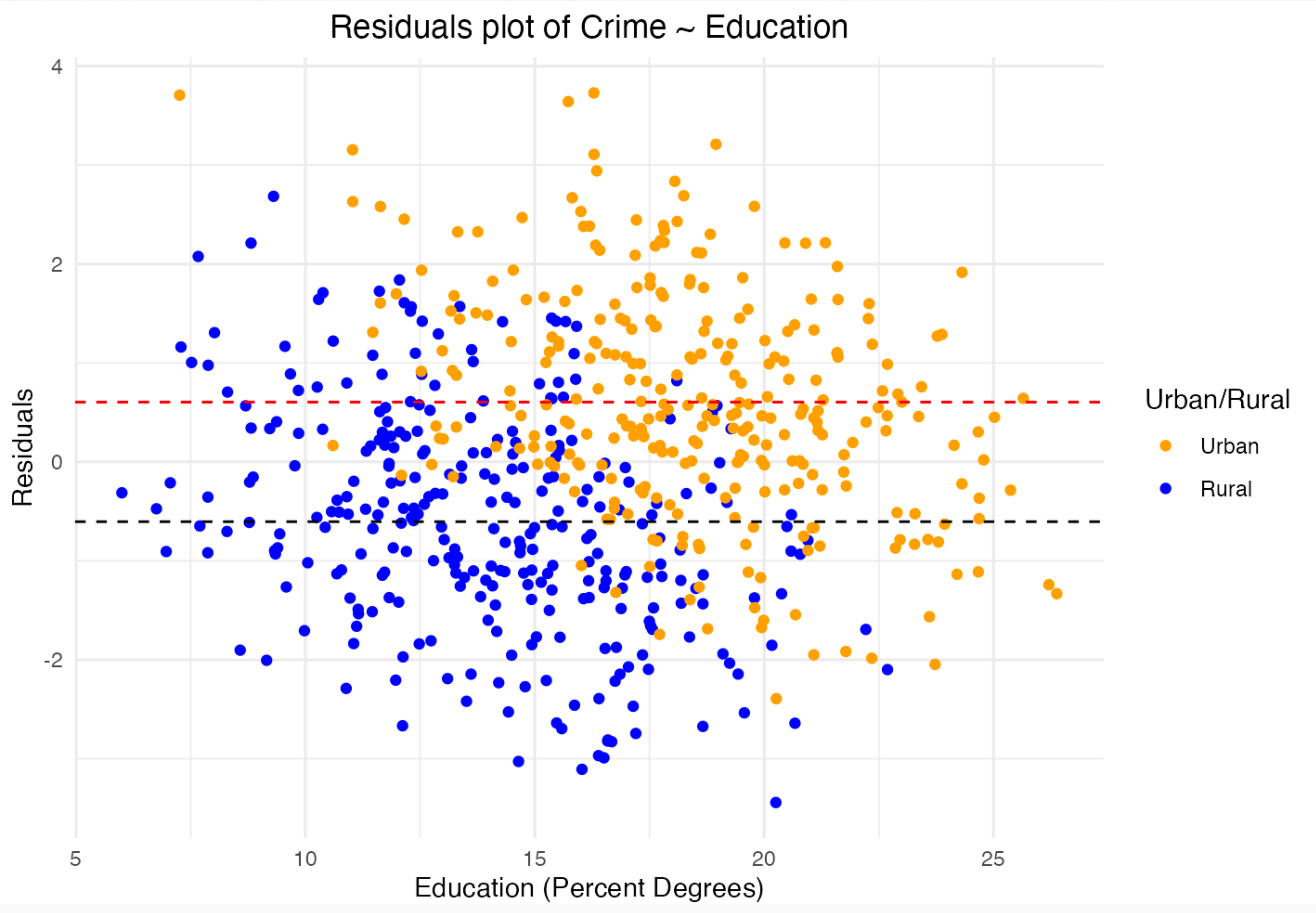


Zero Conditional Mean (aka Exogeneity)

* Tr

  *

led,



Zero Conditional Mean (aka Exogeneity)

- Trickiest violations:
  - **Omitted Variable Bias**
    - If there is a Z correlated with Y *and* with X but is not modelled, then ϵ will be correlated with X.
  - **Simultaneity/Reverse Causality**

Zero Conditional Mean (aka Exogeneity)

- Trickiest violations:
  - **Omitted Variable Bias**
    - If there is a Z correlated with Y *and* with X but is not modelled, then ϵ will be correlated with X.
  - **Simultaneity/Reverse Causality**
    - If an independent variable X is jointly produced with Y:

Zero Conditional Mean (aka Exogeneity)

- Trickiest violations:
  - **Omitted Variable Bias**
    - If there is a Z correlated with Y *and* with X but is not modelled, then ϵ will be correlated with X.
  - **Simultaneity/Reverse Causality**
    - If an independent variable X is jointly produced with Y:
    - $Y = \beta_0 + \beta_1 X + \epsilon$

Zero Conditional Mean (aka Exogeneity)

- Trickiest violations:
  - **Omitted Variable Bias**
    - If there is a Z correlated with Y *and* with X but is not modelled, then ϵ will be correlated with X.
  - **Simultaneity/Reverse Causality**
    - If an independent variable X is jointly produced with Y:
    - $Y = \beta_0 + \beta_1 X + \epsilon$
    - $X = \gamma_0 + \gamma_1 Y + \upsilon$

Zero Conditional Mean (aka Exogeneity)

- Trickiest violations:
  - **Omitted Variable Bias**
    - If there is a Z correlated with Y *and* with X but is not modelled, then ϵ will be correlated with X.
  - **Simultaneity/Reverse Causality**
    - If an independent variable X is jointly produced with Y:
    - $Y = \beta_0 + \beta_1 X + \epsilon$
    - $X = \gamma_0 + \gamma_1 Y + \upsilon$
    - $X = \gamma_0 + \gamma_1(\beta_0 + \beta_1 X + \epsilon) + \upsilon$

Zero Conditional Mean (aka Exogeneity)

Zero Conditional Mean (aka Exogeneity)

- Endogeneity is not a problem when you can randomly assign the treatment (e.g. RCT). But no easy fixes in observational research.

Zero Conditional Mean (aka Exogeneity)

- Endogeneity is not a problem when you can randomly assign the treatment (e.g. RCT). But no easy fixes in observational research.
- **Traditional approach**: control for all things that could be related to both the independent variable of interest and the dependent variable.

Zero Conditional Mean (aka Exogeneity)

- Endogeneity is not a problem when you can randomly assign the treatment (e.g. RCT). But no easy fixes in observational research.
- **Traditional approach**: control for all things that could be related to both the independent variable of interest and the dependent variable.
- **Design-based approaches**: identify an 'as-if-random' X. This is what a lot of contemporary trends in social science methods are all about (RDD, IV, D-in-D). Covered in *Causal Inference* course.

Zero Conditional Mean (aka Exogeneity)

- Endogeneity is not a problem when you can randomly assign the treatment (e.g. RCT). But no easy fixes in observational research.
- **Traditional approach**: control for all things that could be related to both the independent variable of interest and the dependent variable.
- **Design-based approaches**: identify an 'as-if-random' *X*. This is what a lot of contemporary trends in social science methods are all about (RDD, IV, D-in-D). Covered in *Causal Inference* course.
- **Sensitivity analysis**: accept possibility of omitted variables. How big should be the effect of the unobserved confounder(s) to make our relationship non-significant? Is it plausible?

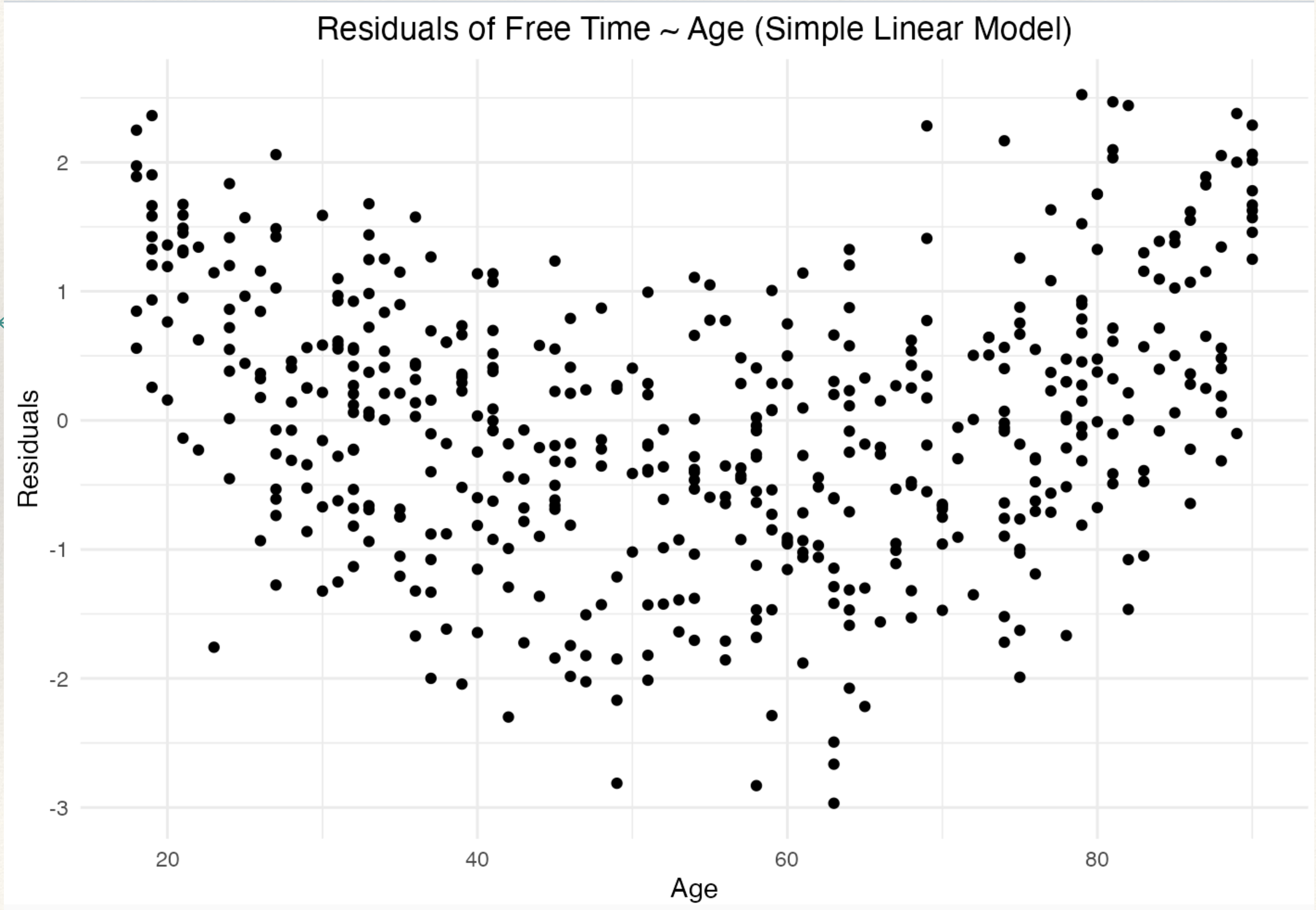
Zero Conditional Mean (aka Exogeneity)

Zero Conditional Mean (aka Exogeneity)

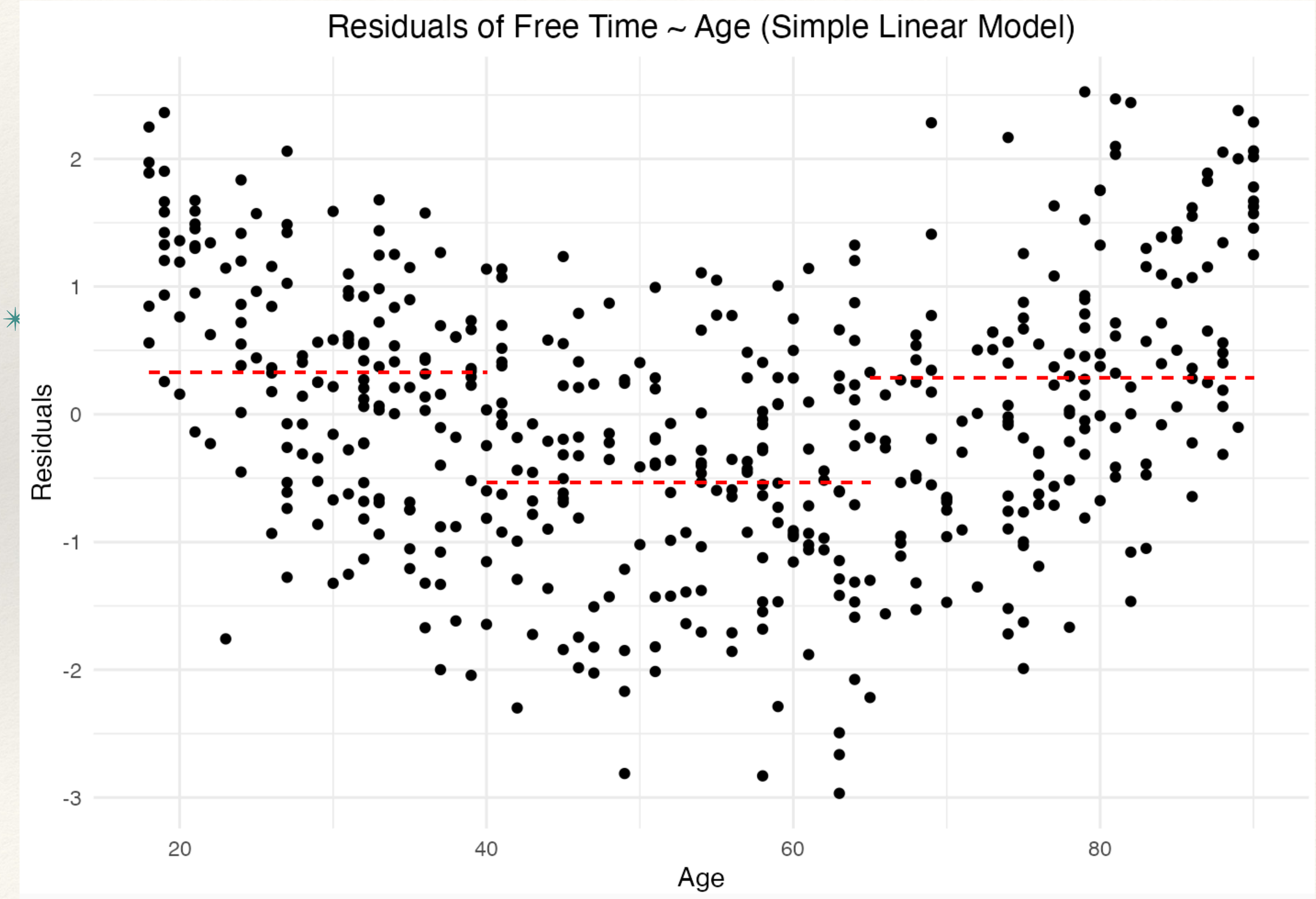
- Another violation: **un-modelled non-linearities**.

Zero Conditional Mean (aka Exogeneity)

Zero Conditional Mean (aka Exogeneity)



Zero Conditional Mean (aka Exogeneity)



Zero Conditional Mean (aka Exogeneity)

- Another violation: **un-modelled non-linearities**.
- We'll deal with some fixes for this next week (polynomial transformations).

What Variables Should I Control For?

What Variables Should I Control For?

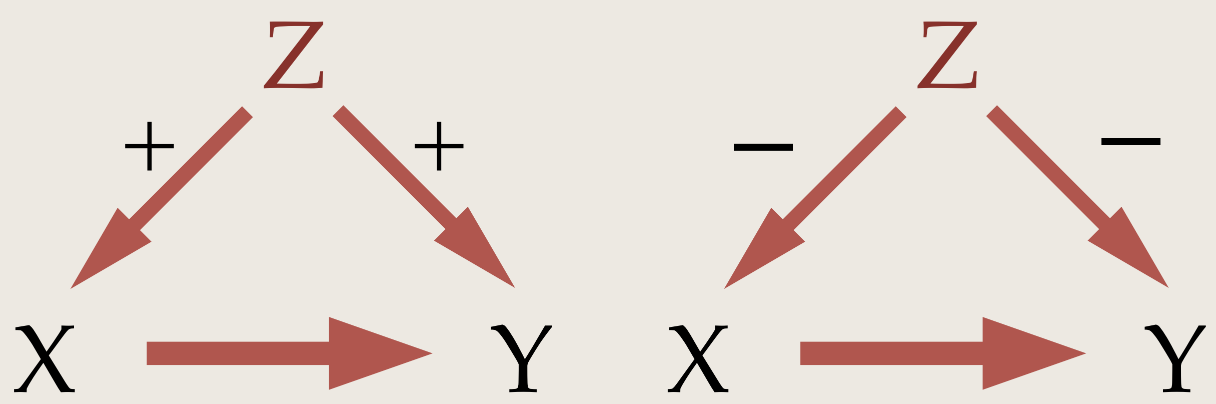
- Goal of ‘controlling’: *accounting for omitted variable bias*.

What Variables Should I Control For?

- Goal of ‘controlling’: *accounting for omitted variable bias*.
- Visually, close ’back doors’ to the causal path $X \rightarrow Y$

What Variables Should I Control For?

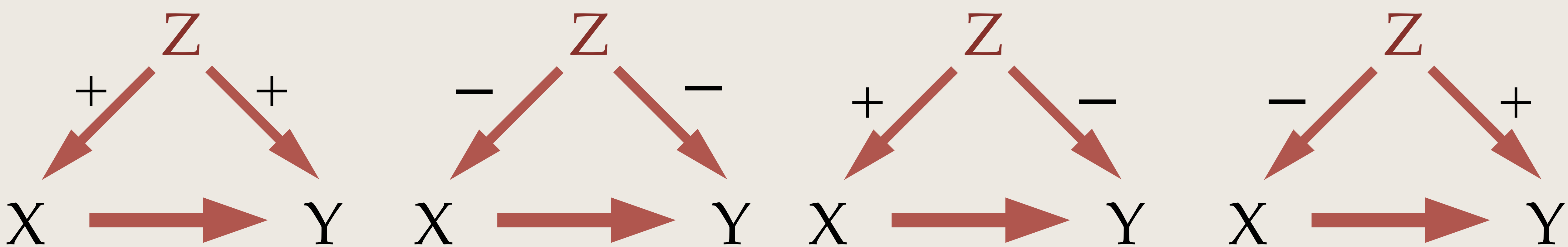
- Goal of 'controlling': *accounting for omitted variable bias*.
- Visually, close 'back doors' to the causal path $X \rightarrow Y$



Without controlling for *Z*, the ATE of *X* on *Y* is positively biased

What Variables Should I Control For?

- Goal of 'controlling': *accounting for omitted variable bias*.
- Visually, close 'back doors' to the causal path $X \to Y$

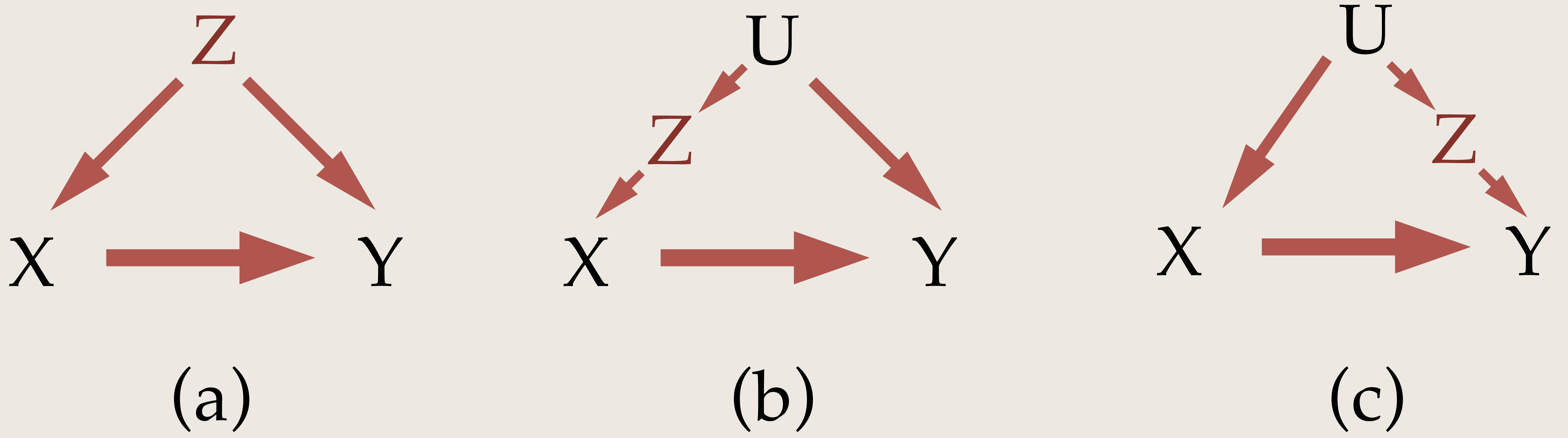


Without controlling for *Z*, the ATE of *X* on *Y* is positively biased

Without controlling for *Z*, the ATE of *X* on *Y* is negatively biased

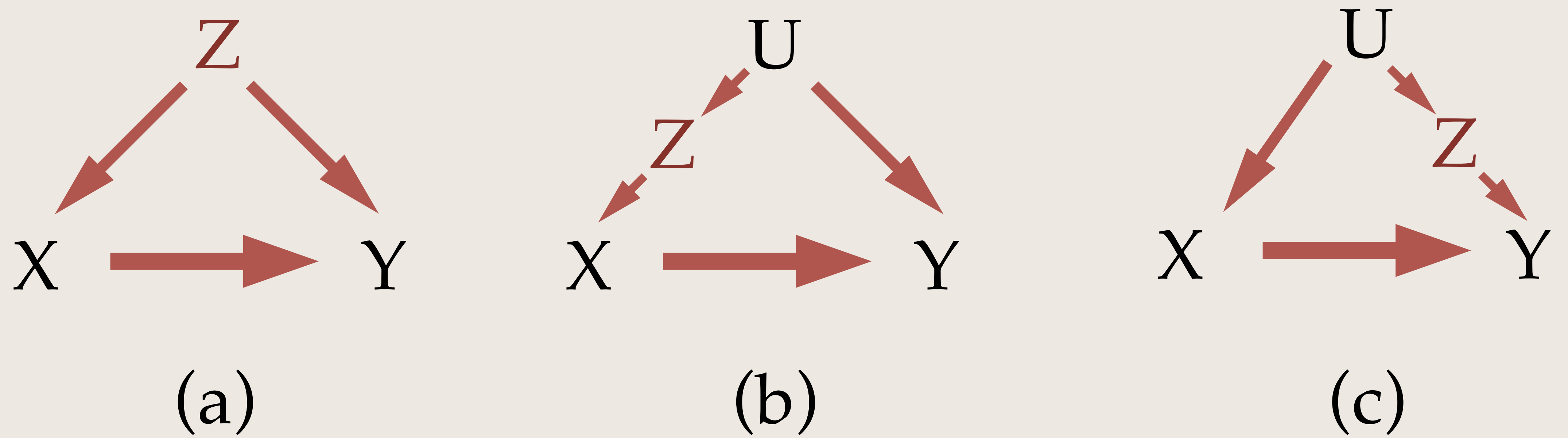
What Variables Should I Control For?

What Variables Should I Control For?



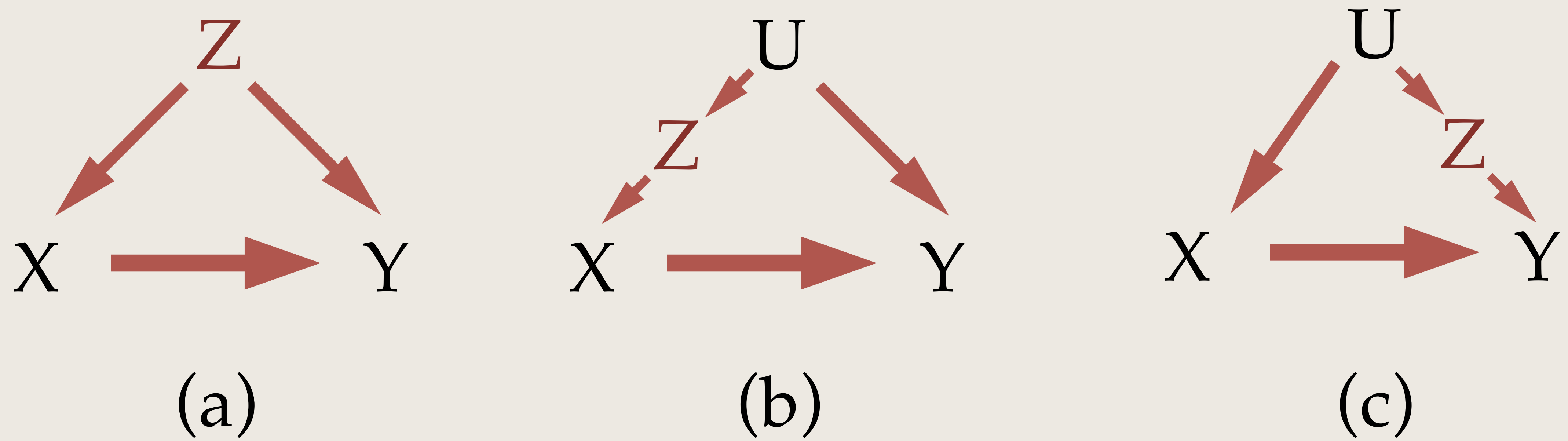
- Adapted from Cinelli et al (2022)

What Variables Should I Control For?



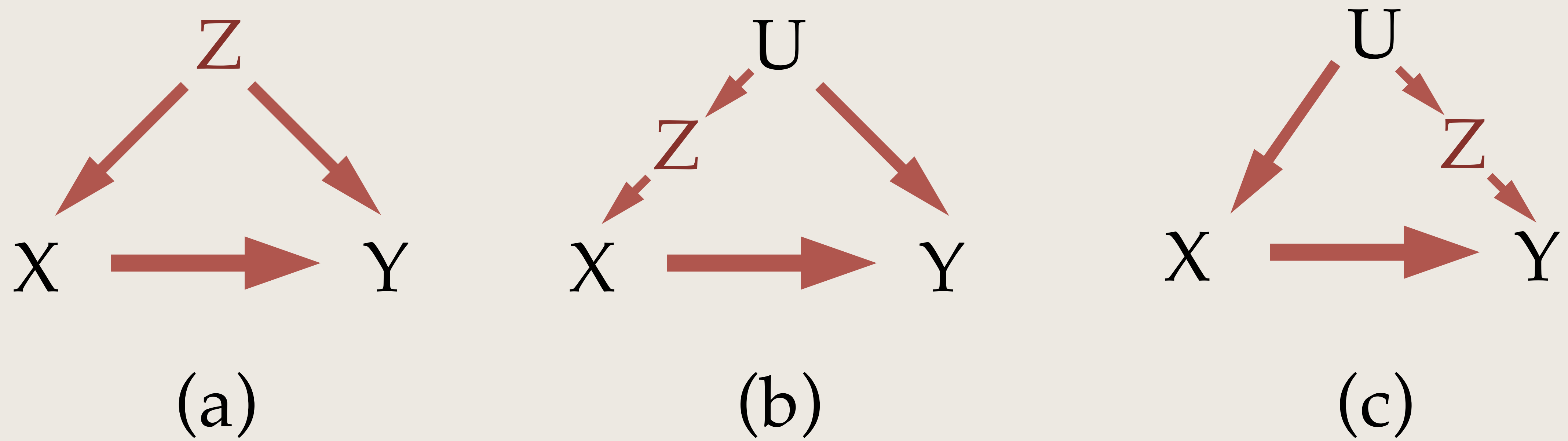
- Adapted from Cinelli et al (2022)
  - **Back-door criterion**: Z is a 'good control' if

What Variables Should I Control For?



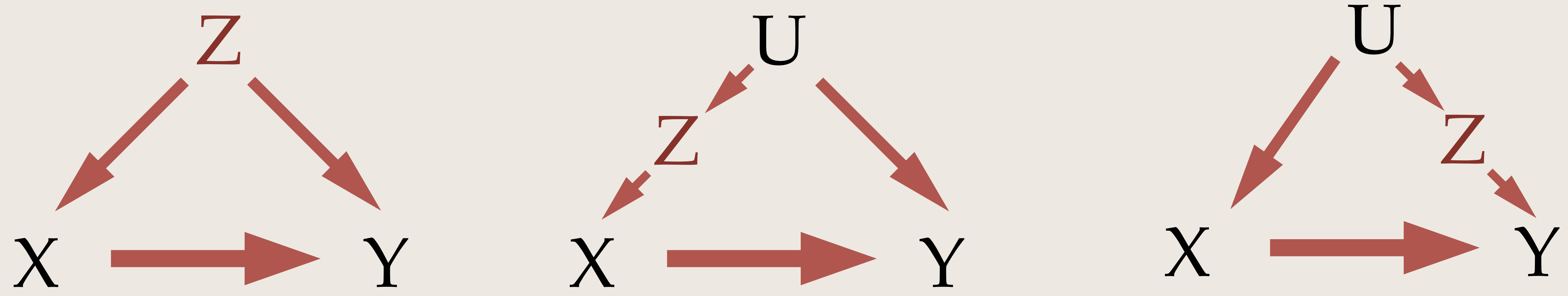
- Adapted from Cinelli et al (2022)
  - **Back-door criterion**: Z is a ‘good control’ if
    1. Z is not a descendant of X (not **post-treatment**), and

What Variables Should I Control For?



- Adapted from Cinelli et al (2022)
  - **Back-door criterion**: Z is a 'good control' if
    1. Z is not a descendant of X (not **post-treatment**), and
    2. Z blocks a path between X and Y **that contains an arrow into X**.

What Variables Should I Control For?



(a) (b) (c)

- Adapted from Cinelli et al (2022)
  - **Back-door criterion**: Z is a 'good control' if
    1. Z is not a descendant of X (not **post-treatment**), and
    2. Z blocks a path between X and Y **that contains an arrow into X**.
       - i.e. Z is a **common cause** of X and Y (*a*) or is the **mediator** of the relationship between an unobserved common cause U and either X or Y (respectively, *b* and *c*) .

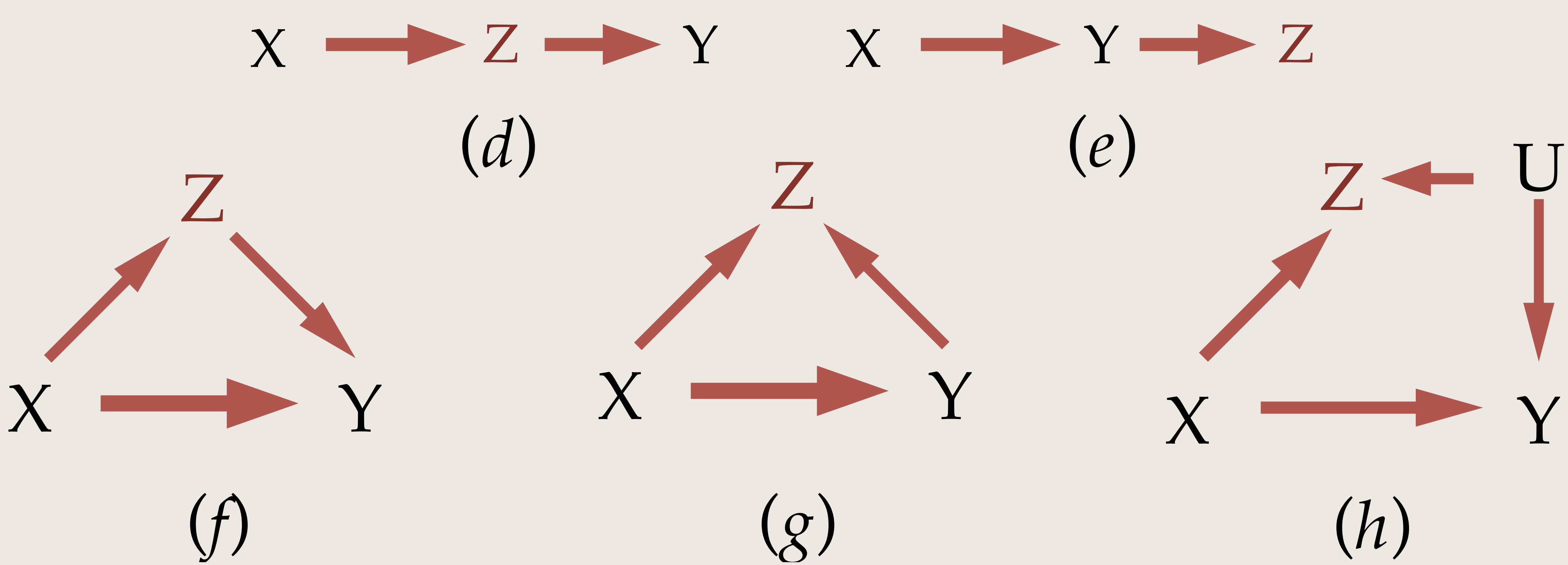
What Variables Should I **Not** Control For?

What Variables Should I **Not** Control For?

- If Z descends from of X (post-treatment variable): **bad idea**.

What Variables Should I **Not** Control For?

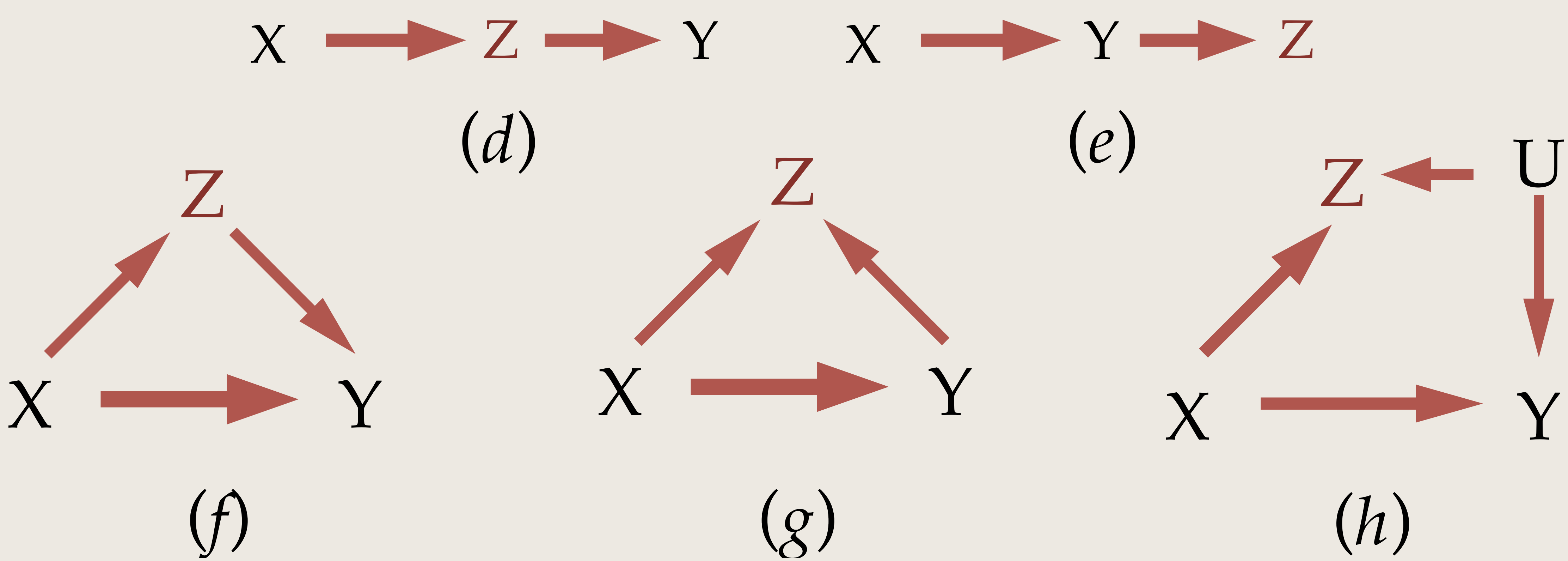
- If Z descends from of X (post-treatment variable): **bad idea**.



- Adapted from Cinelli et al (2022)

What Variables Should I **Not** Control For?

- If Z descends from of X (post-treatment variable): **bad idea**.
- These can: (1) **block the causal path** $X \rightarrow Y$ (*d*), (2) are **effects** of the outcome (*e*), or (3) **open a backdoor path** to a previously unbiased causal path (*f*, *g* and *h*).

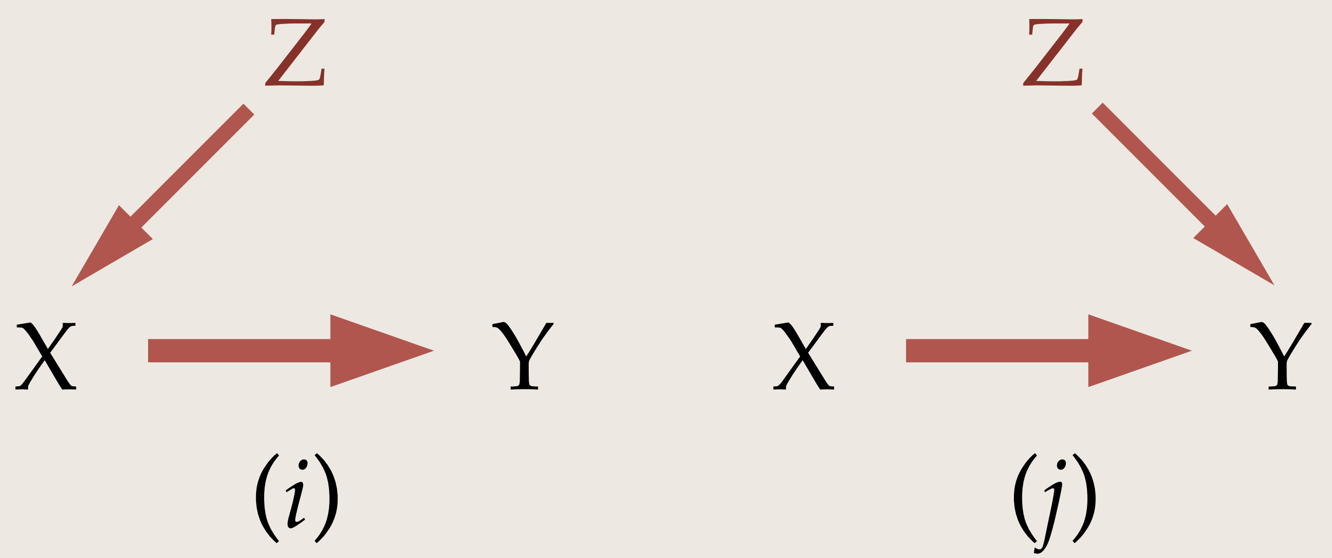


- Adapted from Cinelli et al (2022)

Control for **all** pre-treatment variables?

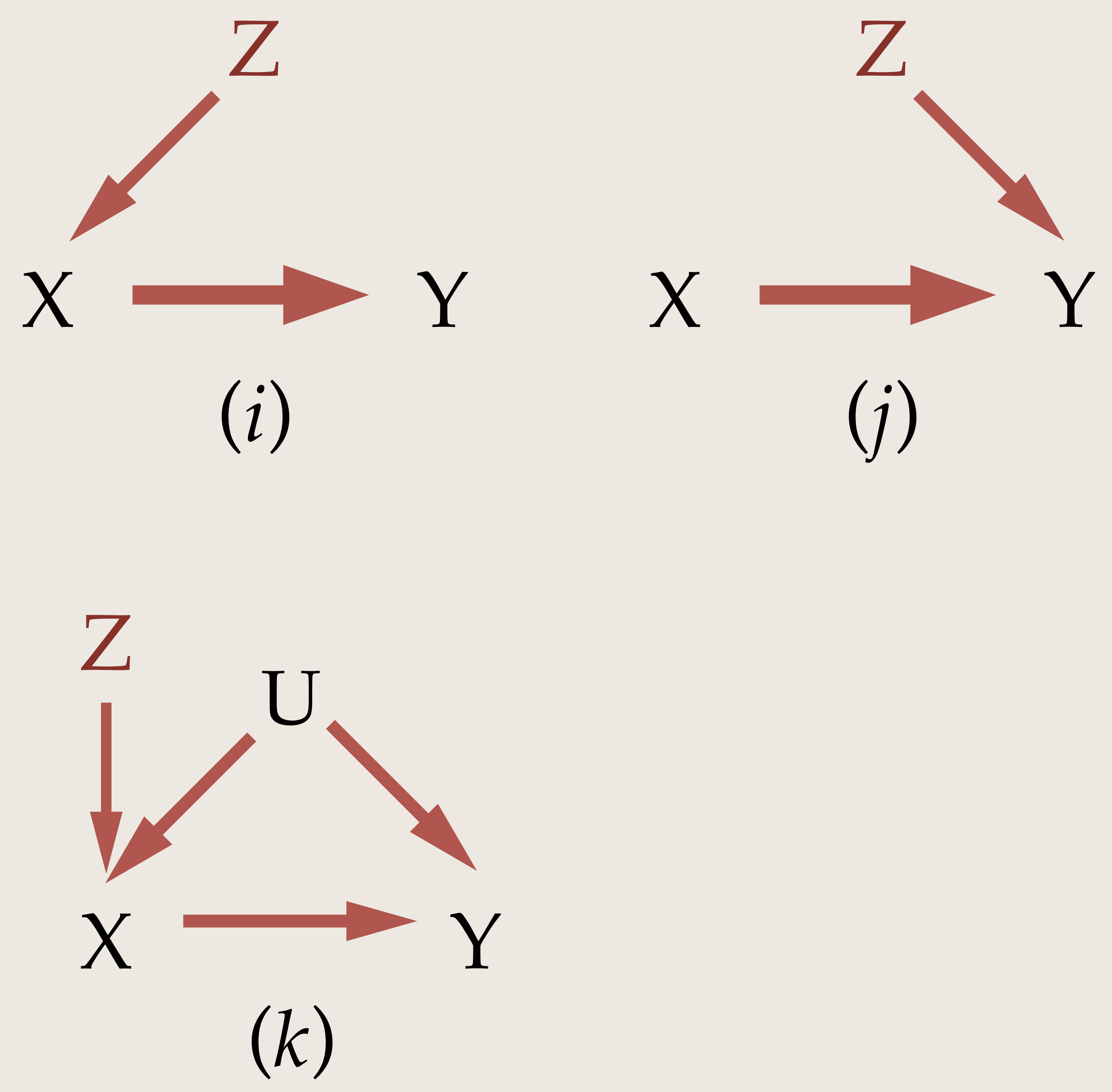
Control for **all** pre-treatment variables?

- Usually pre-treatment variables are good (*a*, *b* and *c*) or **neutral** (*i* and *j*).



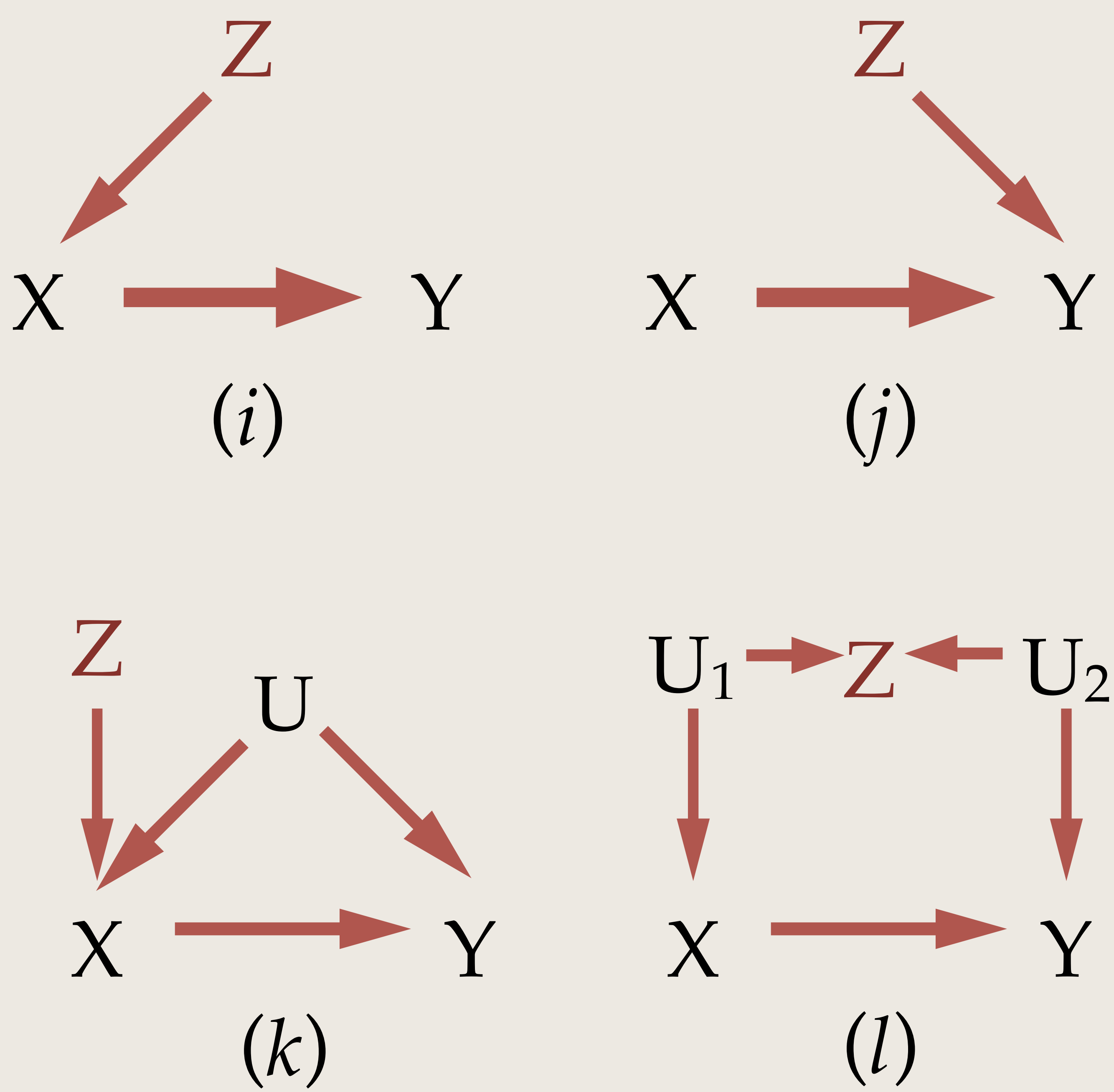
Control for **all** pre-treatment variables?

- Usually pre-treatment variables are good (*a*, *b* and *c*) or **neutral** (*i* and *j*).
- But **in presence of unobserved confounders**, 'pointless' control can make existing bias worse (*k*).



Control for **all** pre-treatment variables?

- Usually pre-treatment variables are good (*a*, *b* and *c*) or **neutral** (*i* and *j*).
- But **in presence of unobserved confounders**, 'pointless' control can make existing bias worse (*k*).
- Also, they can be a problem if they **open a backdoor path** (*l*, collider bias).

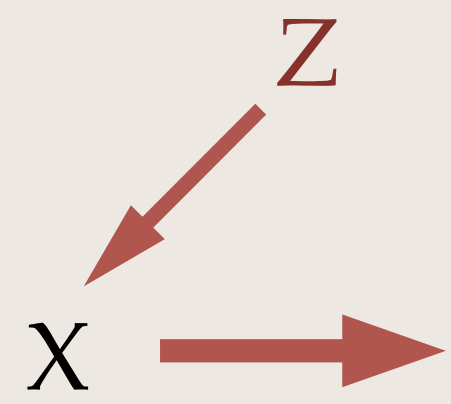


*

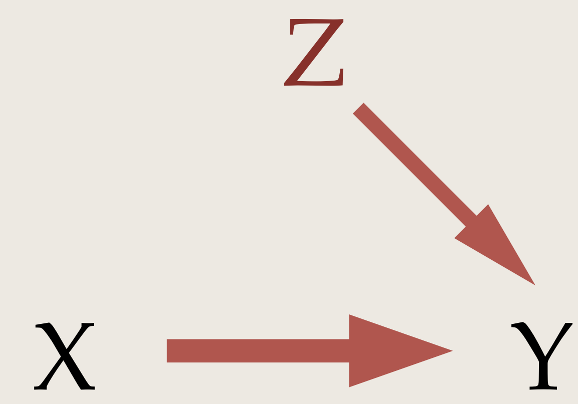
Adapted from Cinelli et al (2022)

Control for **all** pre-treatment variables?

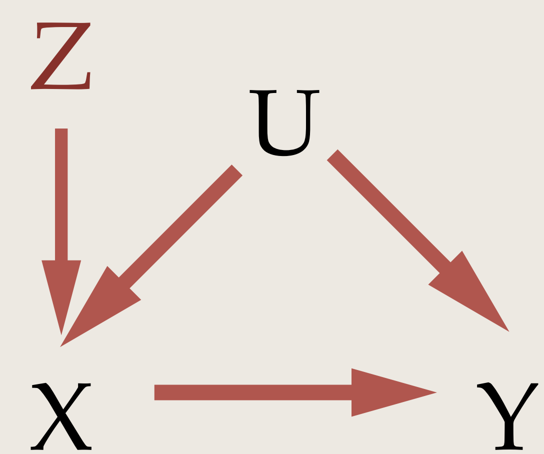
- Usually pre-treatment variables are good (*a*, *b* and *c*) or **neutral** (*i* and *j*).
- But **in presence of unobserved confounders**, ‘pointless’ control can make existing bias worse (*k*).
- Also, they can be a problem if they **open a backdoor path** (*l*, collider bias).
- Bottom line: **theory** should inform your choice of controls, not data availability.



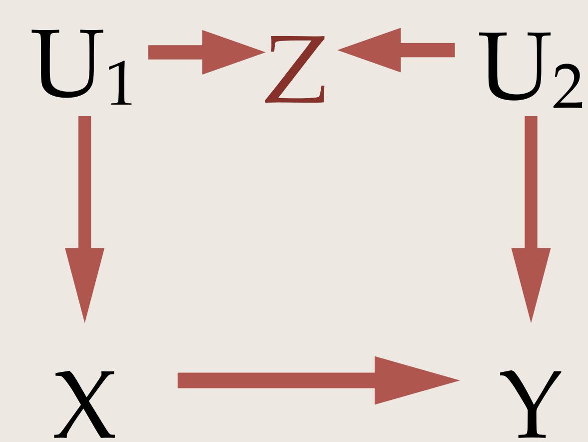
(*i*)



(*j*)



(*k*)



(*l*)

Adapted from Cinelli et al (2022)

Wrapping Up

Wrapping Up

* Linear regression estimates **conditional relationships** that take into account how many independent variables relate to each other and to the outcome.

Wrapping Up

* Linear regression estimates **conditional relationships** that take into account how many independent variables relate to each other and to the outcome.
* A flexible and powerful method, but not magic: strong assumptions are required. Most notably, that there are **no unobserved confounders**.

Wrapping Up

* Linear regression estimates **conditional relationships** that take into account how many independent variables relate to each other and to the outcome.
* A flexible and powerful method, but not magic: strong assumptions are required. Most notably, that there are **no unobserved confounders**.
* Next week: derive **measures of uncertainty of sample estimates**, and **test hypotheses** about the relationships existing in the population.

Thank you for your kind attention!

Leonardo Carella

leonardo.carella@nuffield.ox.ac.uk