universität wien

BAK3: Introduction to Quantitative Methods

Week 12: Multiple Linear Regression

Leonardo Carella

universität wien

The Plan for today

- Statistics:
  - Recap: Simple (Bivariate) Linear Regression.
  - Multiple Linear Regression.

    Disclaimer: This is a lot. It's okay if you don't get everything the first time.
- Coding in R:
  - Multiple Linear Regression and the logic of 'controlling'.
  - Categorical predictors.

universität wien

Simple Linear Regression

- Predicts Y as a linear function of X, plus some chance error ε:

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

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

- α and β represent the intercept and slope of the regression line.

- 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.

universität wien

Simple Linear Regression

- We actually **estimate** $\hat{\alpha}$ and $\hat{\beta}$ from our observed data, by figuring out the "best" line to fit through our data:

Predicted (or "fitted") values of Y

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

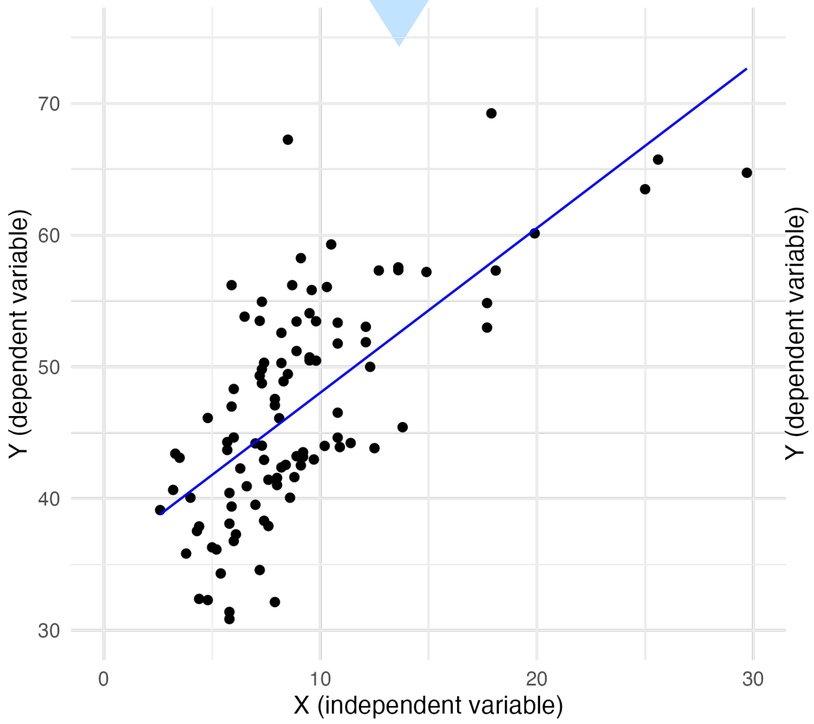
Estimates for the intercept and slope

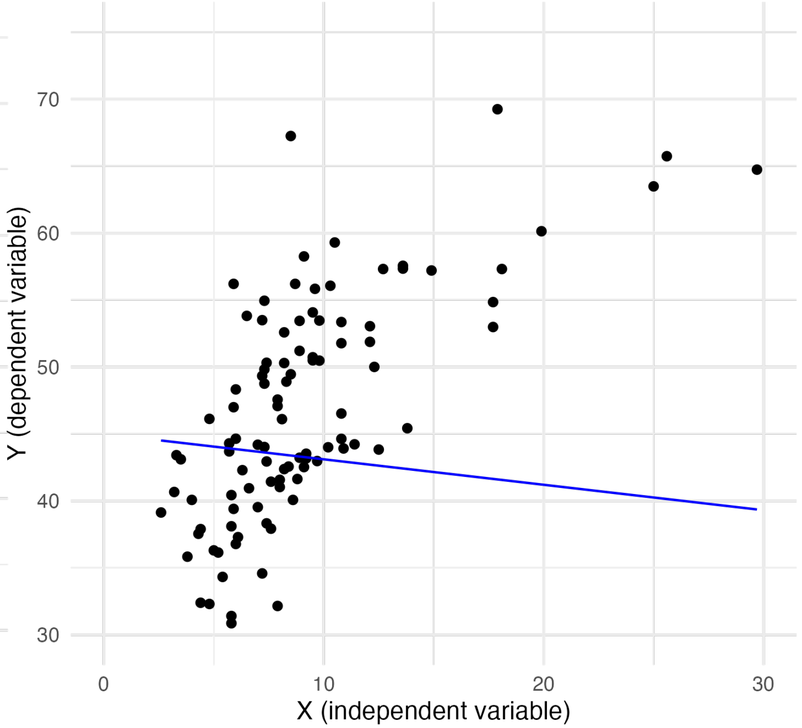
- How do we pick the "best" line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$?
- Ordinary Least Squares: we choose $\hat{\alpha}$ and $\hat{\beta}$ so that they minimise the **sum of squared residuals**, where residuals are $Y_i - \hat{Y}_i$.

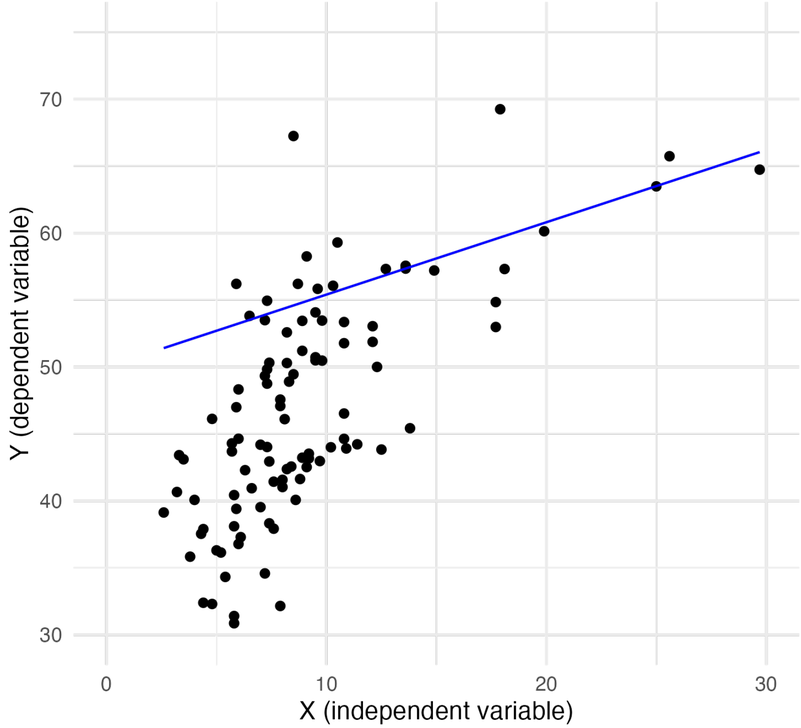
universität wien

Visually…

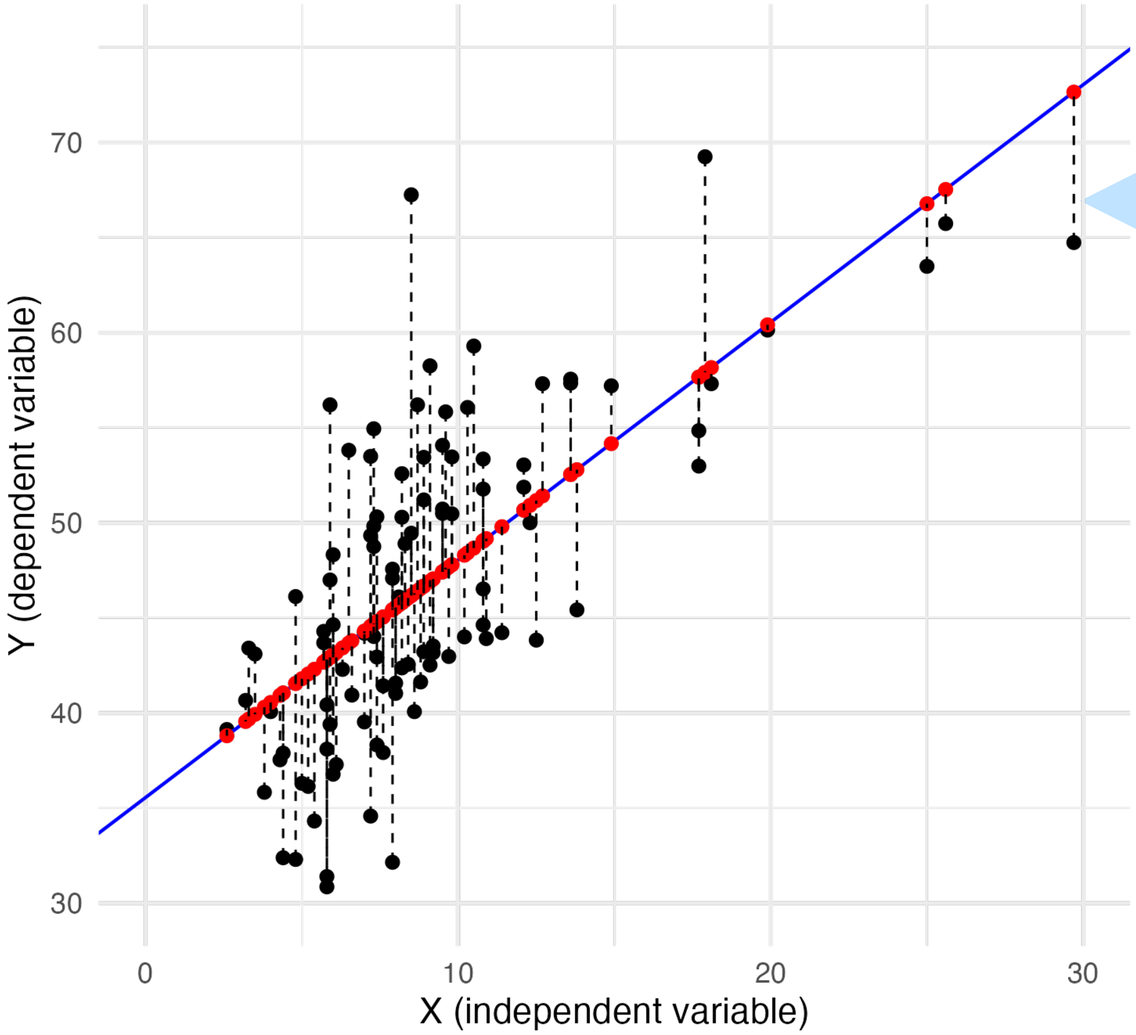
Of all possible lines, we pick this one…







Visually…



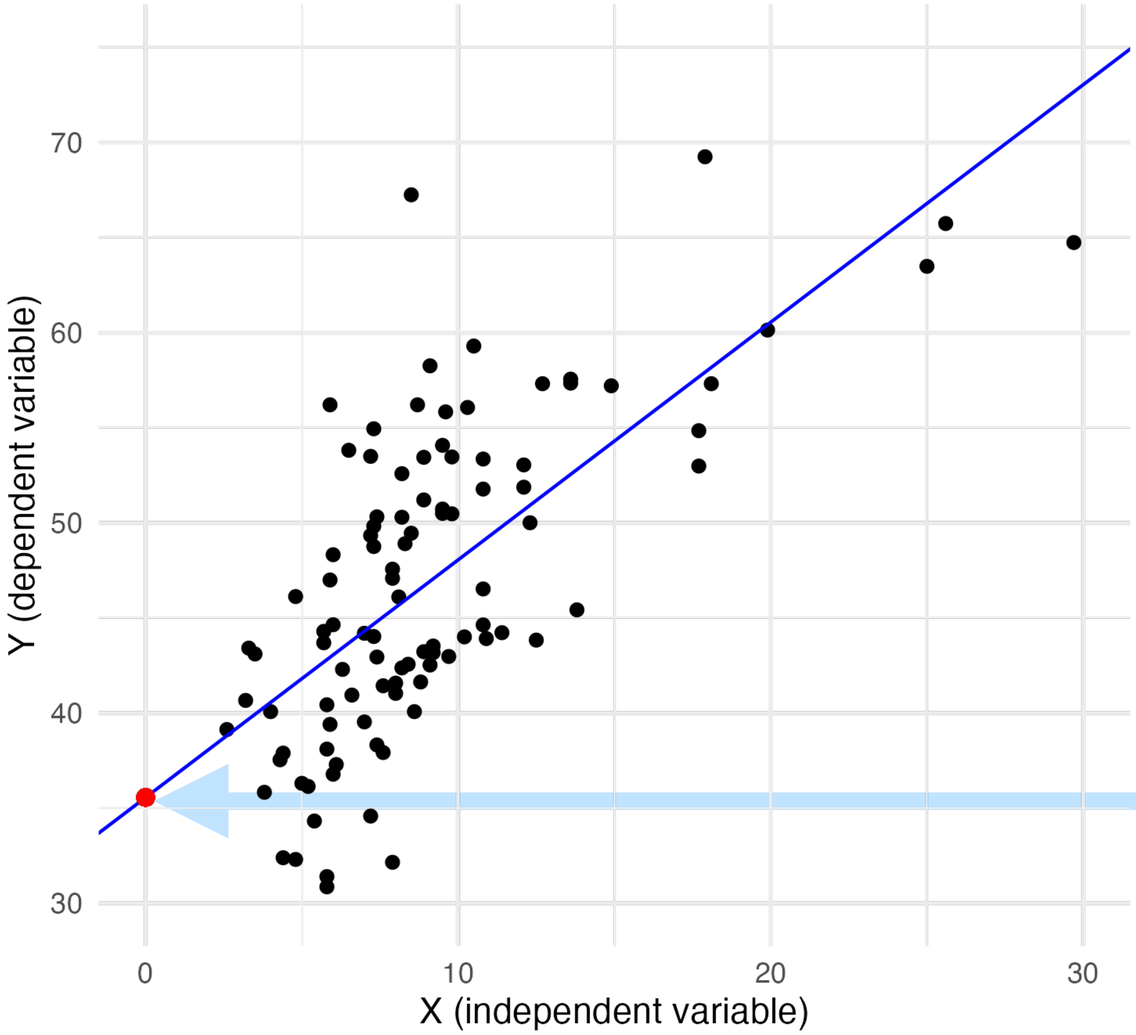
…because if we take all the difference between the observed values of Y (in black) and the predicted values $\hat{Y}$ (in red)…

…Then we **square** these differences (the residuals), so they all become positive…

…and we sum them all up, this specific line returns the **minimum sum of squared residuals**.

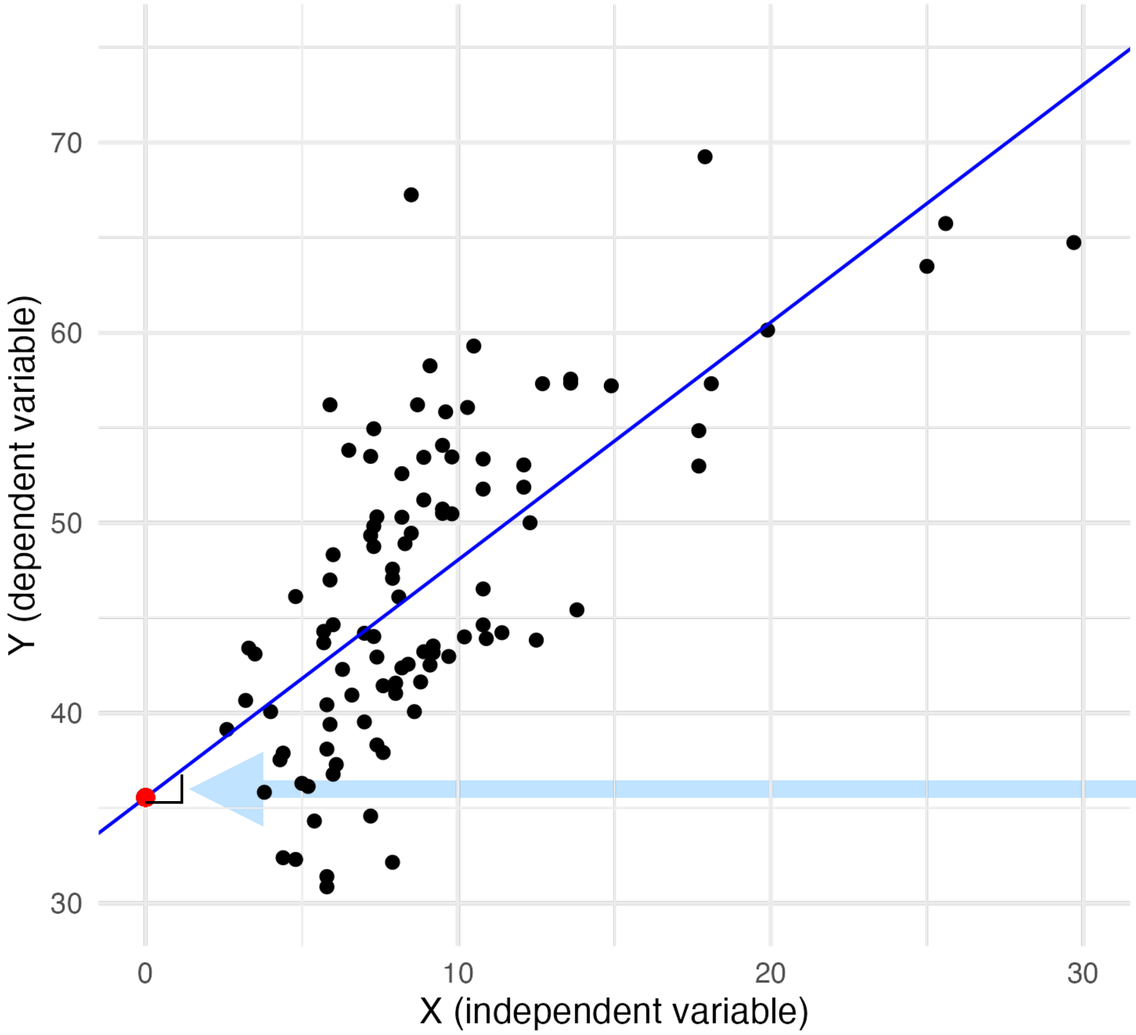
universität wien

Visually…



The intercept $\hat{\alpha}$ is the predicted value of Y when X is zero. In this case, about 35.

Visually…



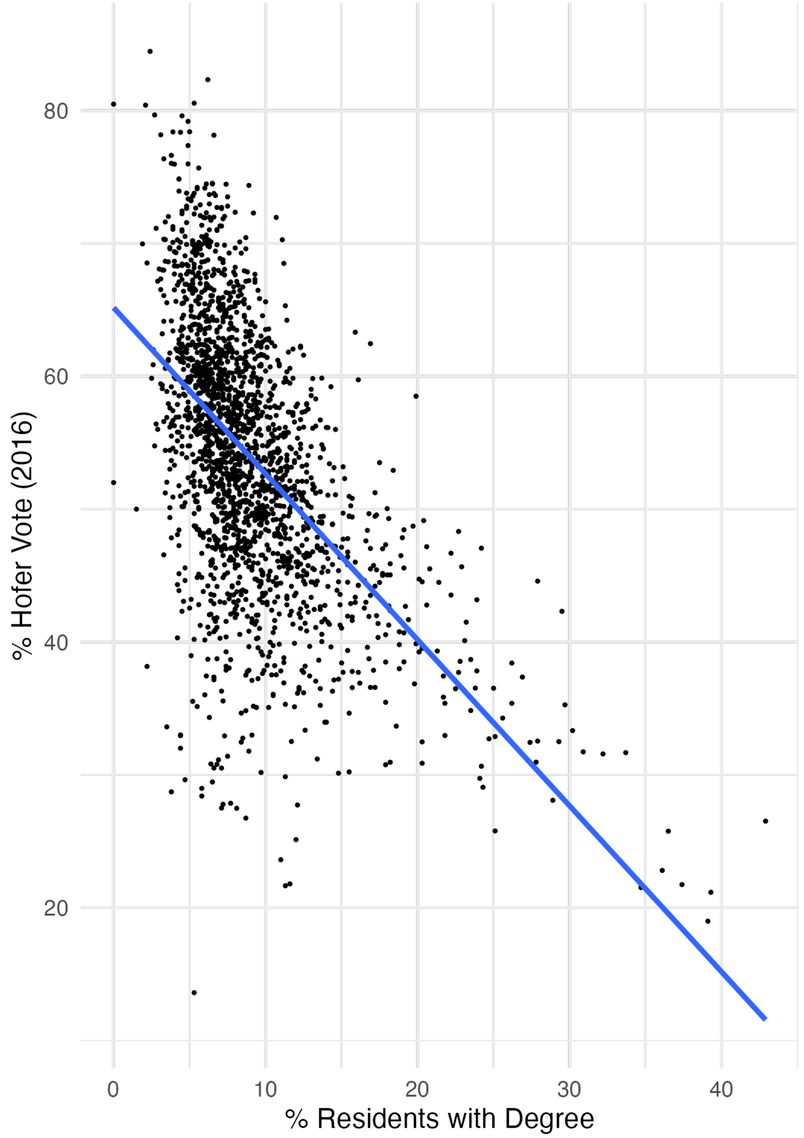
The slope $\hat{\beta}$ is the predicted change in Y as we increase X by 1. In this case, it's 1.25

universität wien

Hofer Vote $= \alpha + \beta(\text{Pct. Degree}) + \varepsilon$

```
> model_hofer <- lm(data = austria,
  pct_hofer ~ pct_degree)
> stargazer(model_hofer, type = "text")
```

```
================================================
                         Dependent variable:
                     ---------------------------
                              pct_hofer
------------------------------------------------
pct_degree                    -1.250***
                               (0.041)
Constant                      65.165***
                               (0.406)
------------------------------------------------
Observations                    2,122
R2                              0.308
Adjusted R2                     0.308
Residual Std. Error       8.266 (df = 2120)
F Statistic           945.242*** (df = 1; 2120)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```



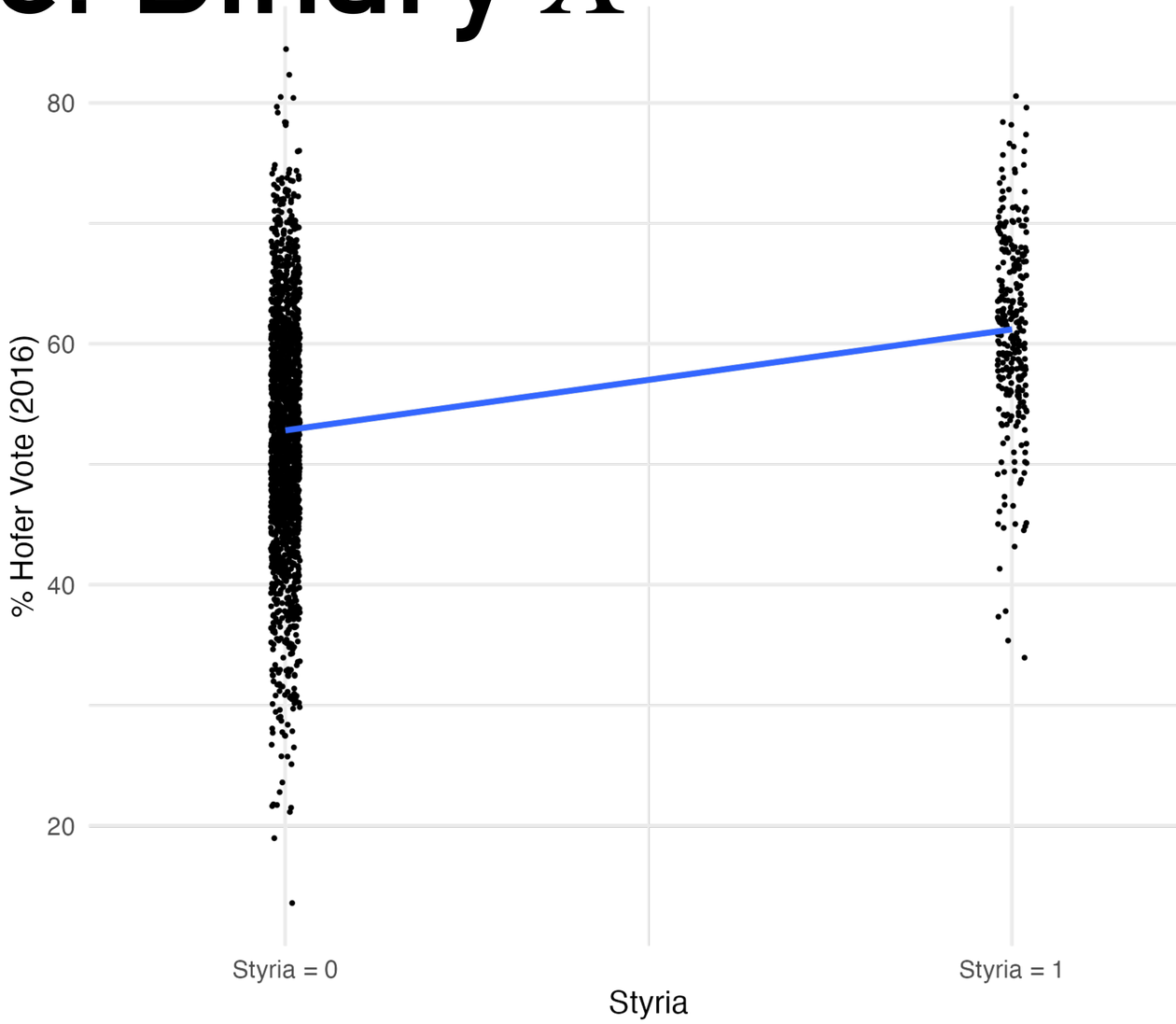
Special Case: Binary X

- Consider X as a 0-1 binary variable for 'Styria' (Steiermark):
  - $X = 1$ if the town is in Styria
  - $X = 0$ if the town is **not** in Styria
- How do we make sense of Hofer Vote $= \alpha + \beta(\text{Styria}) + \varepsilon$?

	Dependent variable:
	pct_hofer
styria	**8.382*****
	(0.604)
Constant	**52.824*****
	(0.222)
Observations	2,122
R2	0.083
Adjusted R2	0.083
Residual Std. Error	9.517 (df = 2120)
F Statistic	192.506*** (df = 1; 2120)
Note:	*p<0.1; **p<0.05; ***p<0.01

Special Case: Binary X

- Consider X as a 0-1 binary variable for ‘Styria’ (Steiermark):
  - $X = 1$ if the town is in Styria
  - $X = 0$ if the town is **not** in Styria
- How do we make sense of Hofer Vote $= \alpha + \beta(\text{Styria}) + \varepsilon$?



Special Case: Binary X

- Interpretation:
- "The predicted percentage of Hofer vote is 8.38 points higher for municipalities in Styria than elsewhere", or…
- "On average, a municipality being in Styria is associated with a 8.38-point increase in support for Hofer"

	Dependent variable:
	pct_hofer
styria	**8.382*****
	(0.604)
Constant	**52.824*****
	(0.222)
Observations	2,122
R2	0.083
Adjusted R2	0.083
Residual Std. Error	9.517 (df = 2120)
F Statistic	192.506*** (df = 1; 2120)
Note:	*p<0.1; **p<0.05; ***p<0.01

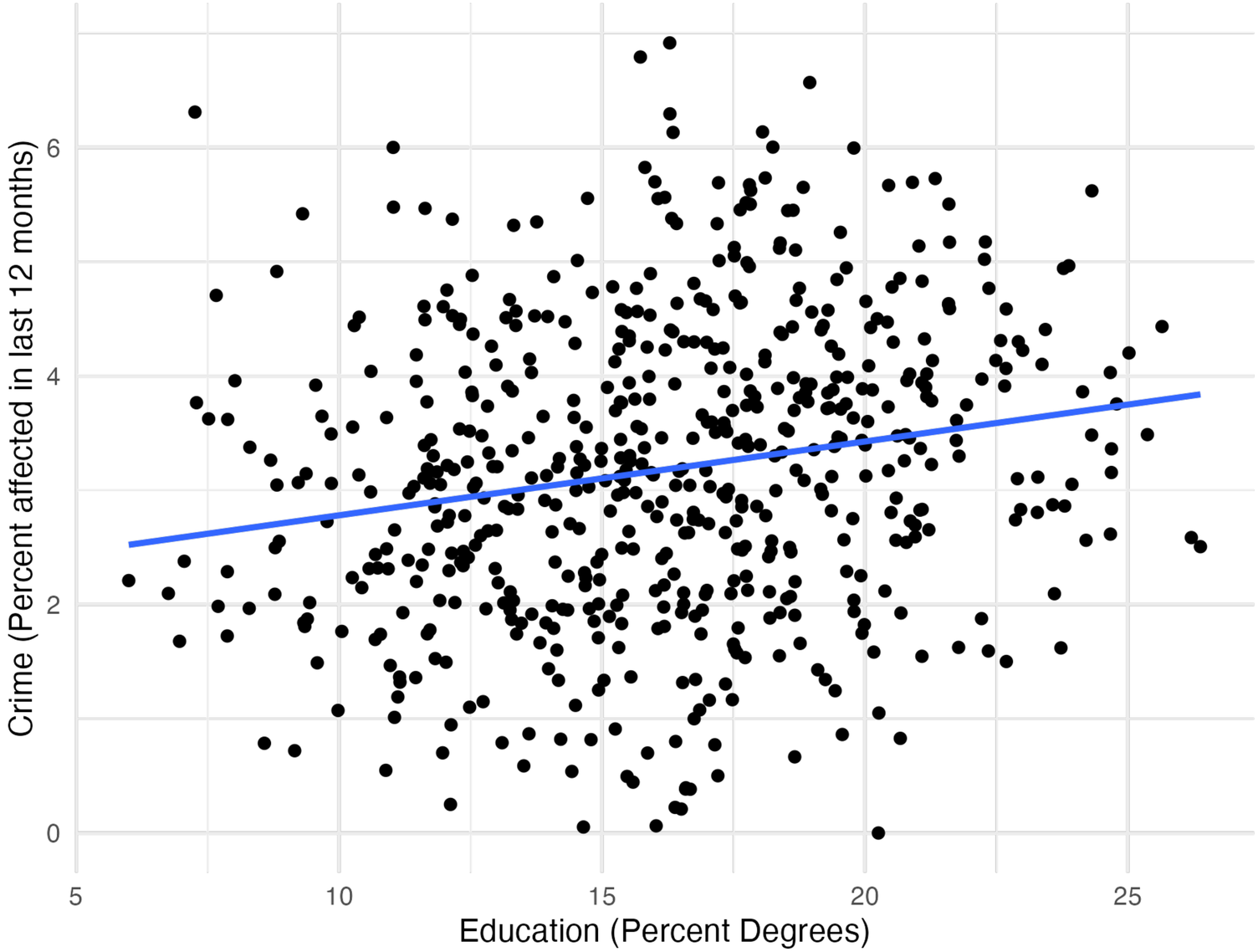
universität wien

Multiple Linear Regression

- The logic: predicting Y as a function of $X_1, X_2, X_3 \ldots$ instead of just one X.
- Why might we want to do that?
  - **Prediction**: richer models can give us more precise guesses for the value of Y.
  - **Description**: describe the **relationship** between X_1 and Y, **conditional** *on* X_2 — a.k.a. 'adjusting' for X_2, *or* 'controlling' for X_2, or holding X_2 constant. This allows us to reduce **omitted variable bias**, which can lead to inaccurate conclusions.
  - **Causation**: effect of X_1 on Y 'holding all else equal'. Extremely (impossibly?) ambitious, requires very heroic assumptions (more on this next week).

universität wien

Conditional Relationships and Omitted Variable Bias



universität wien

Multiple Linear Regression

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

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

- For instance:

$$\text{Pct. Hofer}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Styria}_i + \varepsilon_i$$

- 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 is minimised.
- Where the sum of squared residuals is $\sum (\hat{\varepsilon}_i)^2 = \sum (Y_i - \hat{Y}_i)^2$.

universität wien

Multiple Linear Regression in R

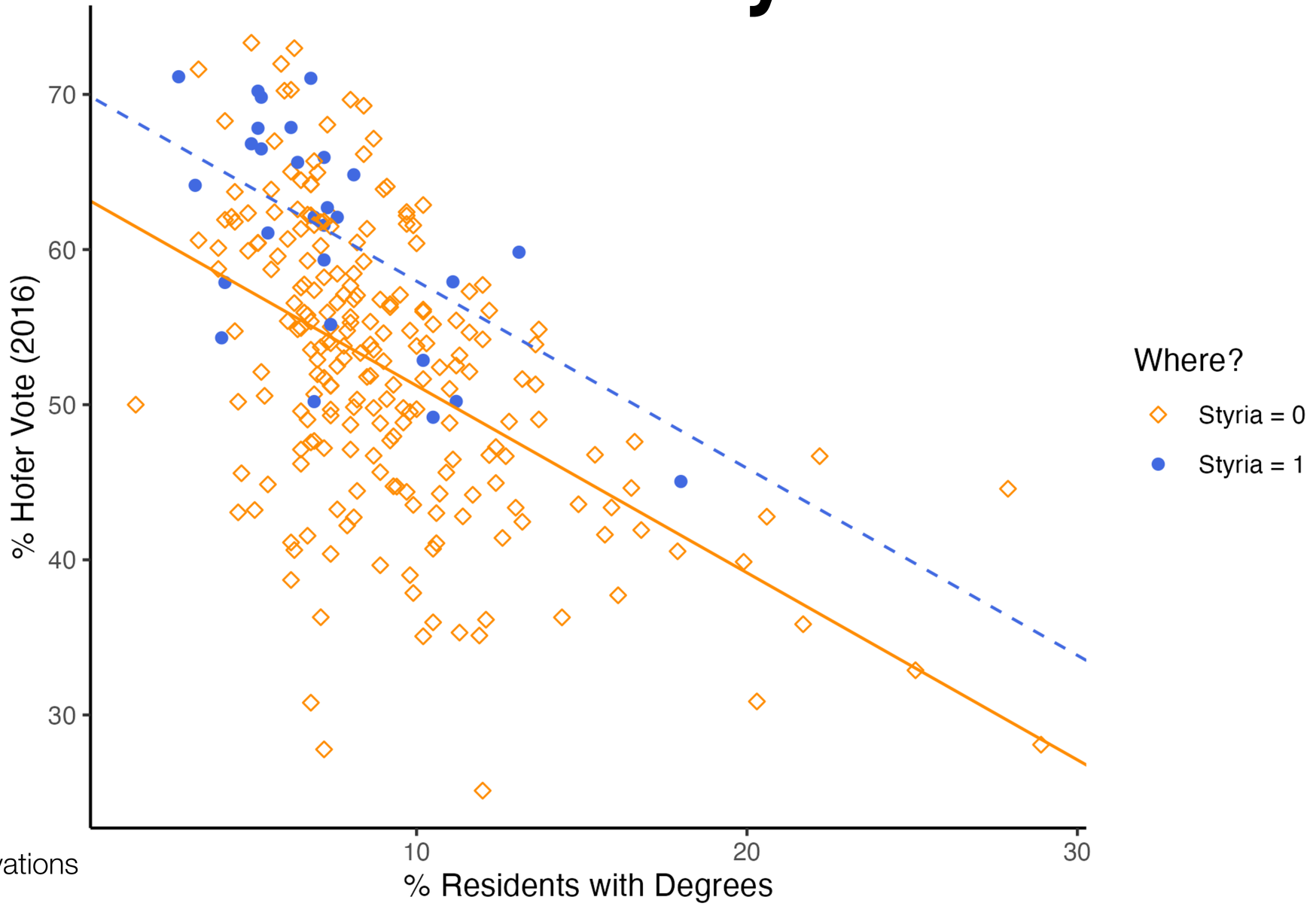
```
> model <- lm(pct_hofer ~ pct_degree + styria, data = austria)
> summary(model)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.74925    0.40525  157.31   <2e-16 ***
pct_degree  -1.19270    0.03933  -30.32   <2e-16 ***
styria       6.68544    0.50767   13.17   <2e-16 ***
```

- When `styria = 0`
  - Hofer Vote = **63.74** **−1.19** × Pct Degrees + **6.68** × (0)
    Hofer Vote = **63.74** **−1.19** × Pct Degrees
- When `styria = 1`
  - Hofer Vote = **63.74** **−1.19** × Pct Degrees + **6.68** × (1)
    Hofer Vote = **70.43** **−1.19** × Pct Degrees

universität wien

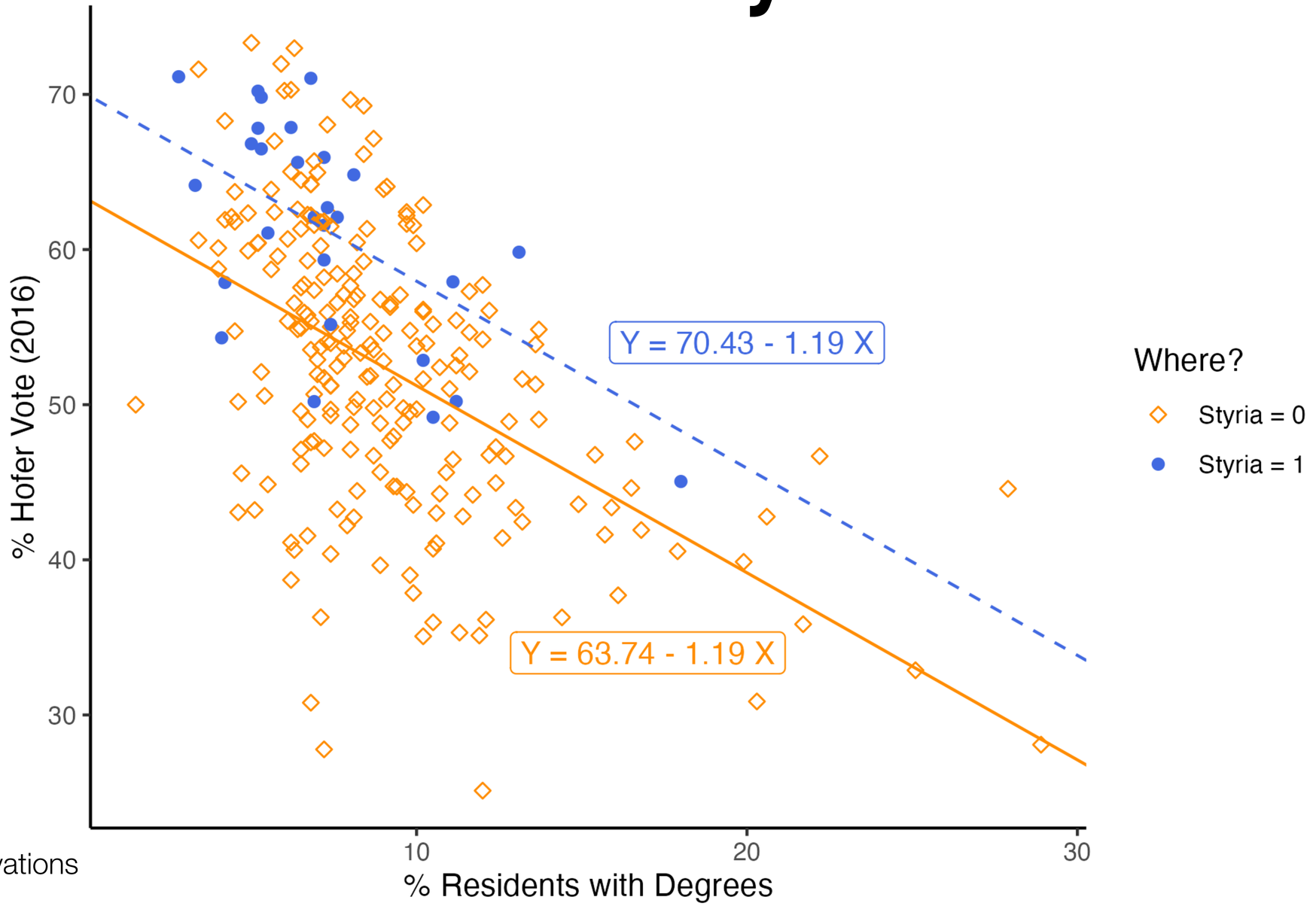
Visually…



- Note: only 250 observations plotted, for clarity.

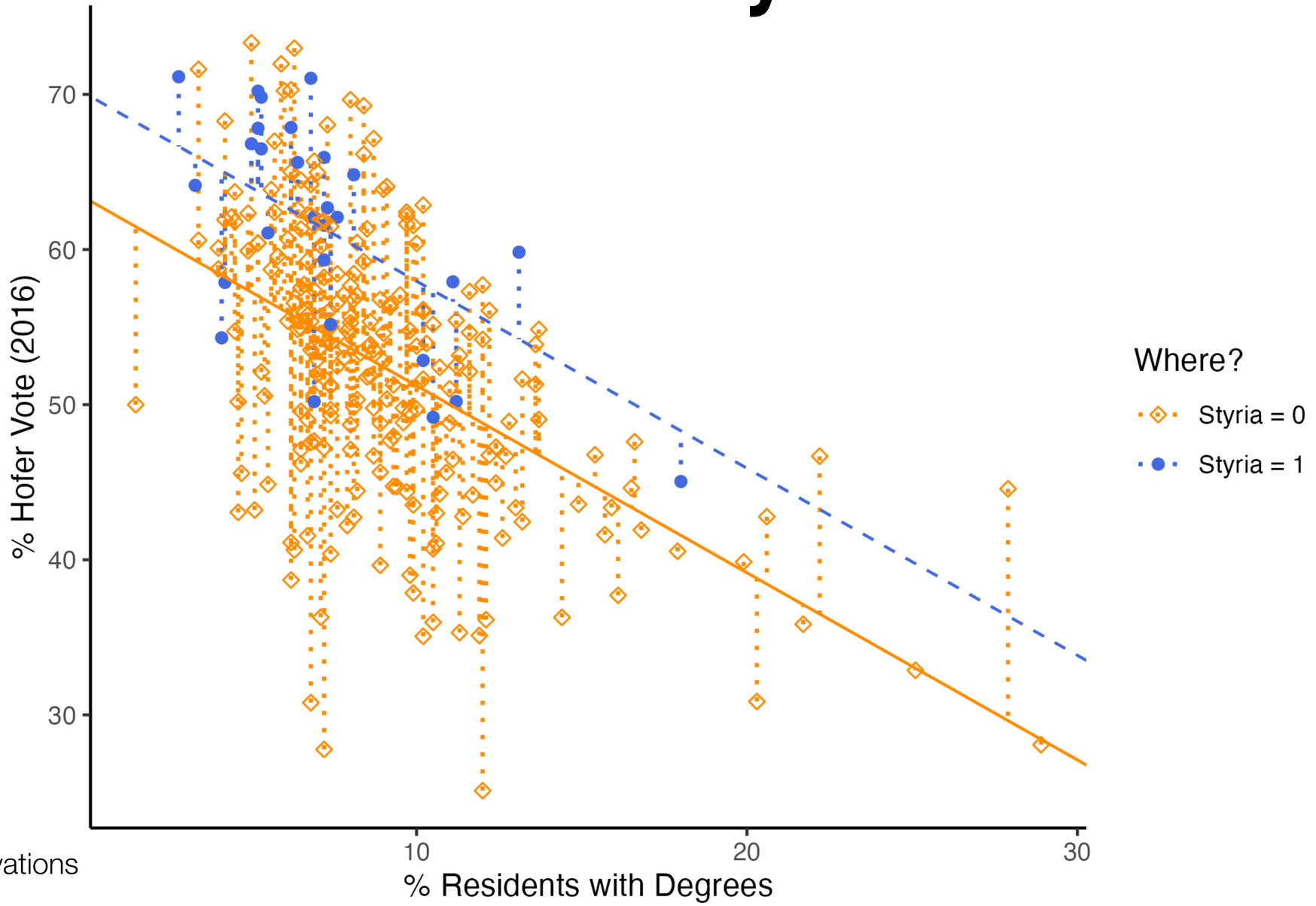
universität wien

Visually…



- Note: only 250 observations plotted, for clarity.

Visually…



- Note: only 250 observations plotted, for clarity.

universität wien

Multiple Linear Regression

- Now let's try with two numerical predictors:

$$\text{Pct. Hofer}_i = \alpha + \beta_1 \text{Pct. Degrees}_i + \beta_2 \text{Pct. Over 65}_i + \varepsilon_i$$

```
> model <- lm(pct_hofer ~ pct_degree + pct_over65, data = austria)
> summary(model)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.69855    1.05178  55.809  < 2e-16 ***
pct_degree  -1.23784    0.04028 -30.730  < 2e-16 ***
pct_over65   0.33326    0.05008   6.654 3.62e-11 ***
```

universität wien

Multiple Linear Regression

```
> model <- lm(pct_hofer ~ pct_degree + pct_over65, data = austria)
> summary(model)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.69855    1.05178  55.809  < 2e-16 ***
pct_degree  -1.23784    0.04028 -30.730  < 2e-16 ***
pct_over65   0.33326    0.05008   6.654 3.62e-11 ***
```

- $\hat{\alpha}$ **(the intercept)** = predicted Hofer vote when Pct. Degree = 0 and Pct. Over 65 = 0.
- $\hat{\beta}_1$ = predicted change in Hofer vote associated with a one-point increase in the **percentage of residents with a degree**, holding the percentage of residents over 65 years of age constant.
- $\hat{\beta}_2$ = predicted change in Hofer vote associated with a one-point increase in the **percentage of residents over 65 years of age,** holding the percentage of residents with a degree constant.

universität wien

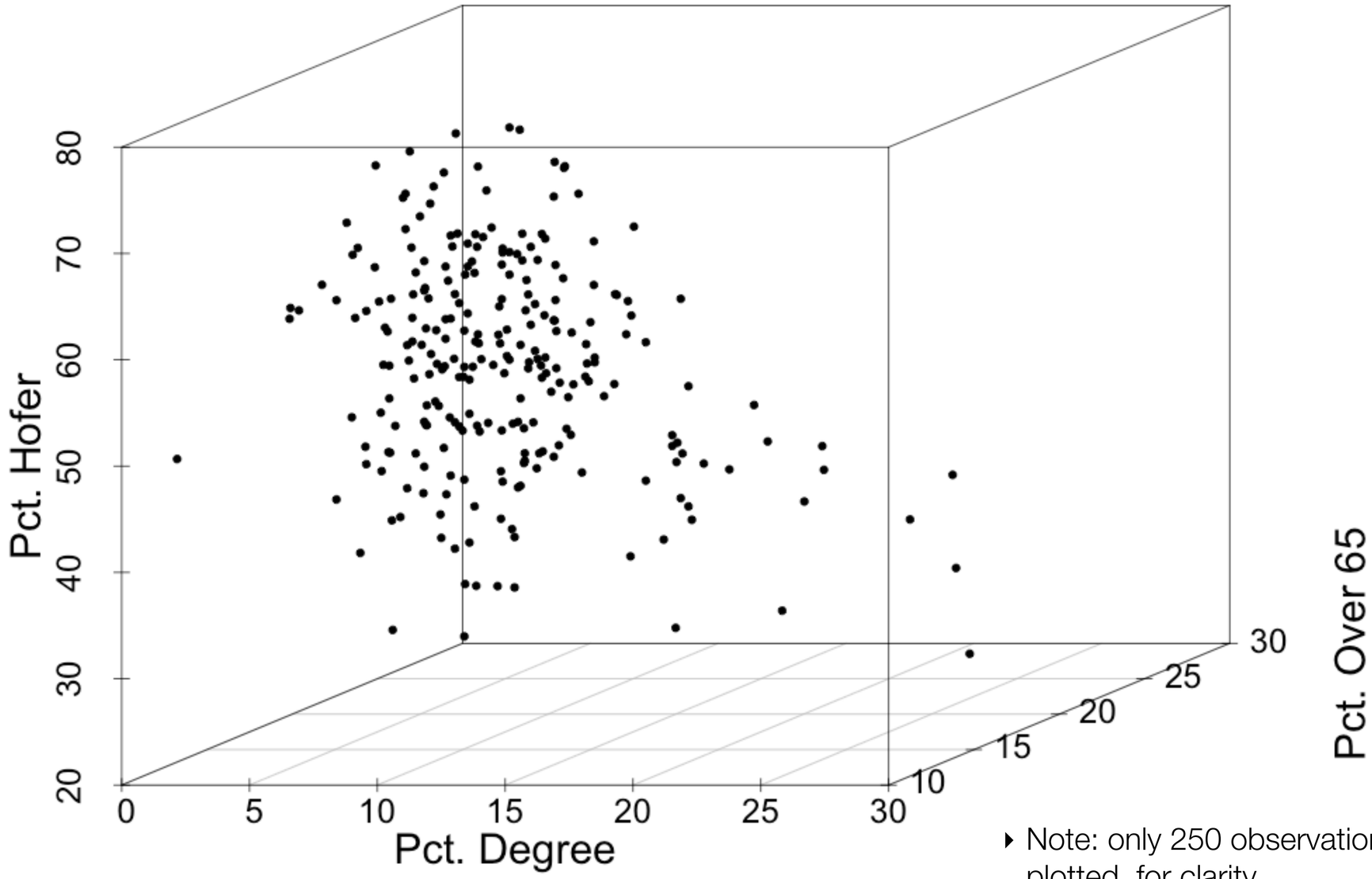
Visually…

Imagine your observations in a 3-dimensional space…

Where ‘Pct. Degree’ is the ‘width’ dimension…

‘Pct. Over 65’ is the ‘depth’ dimension…

…and ‘Pct. Hofer’ is the ‘height’ dimension



- Note: only 250 observations plotted, for clarity.

universität wien

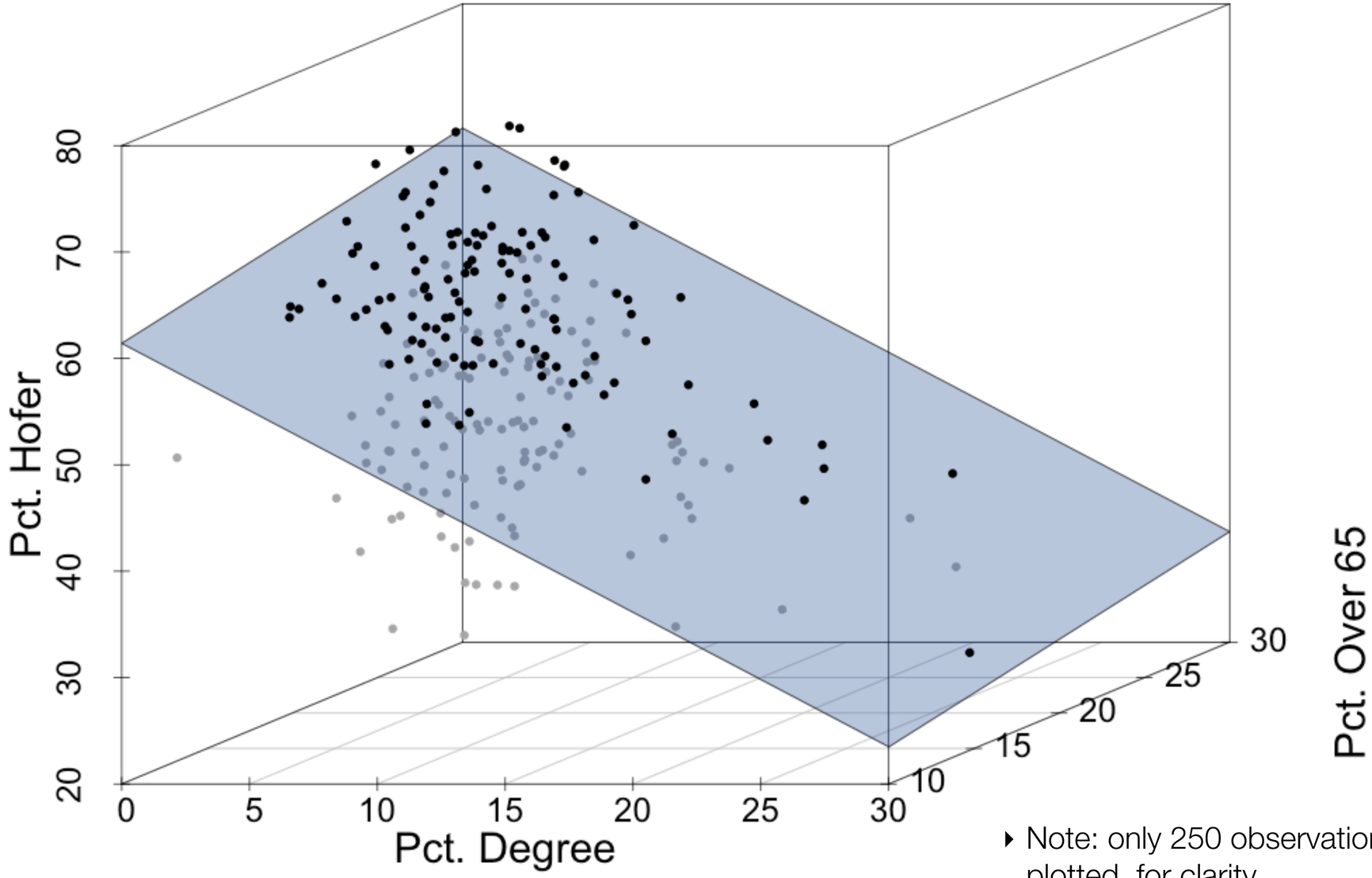
Visually…

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

can be represented geometrically as a 2-dimensional **plane**…

… where the slope of 'Pct Degree' (β_1) is the inclination of the plane on the 'width' dimension…

…and the slope of 'Pct Over 65' (β_2) is the inclination of the plane on the 'depth' dimension.



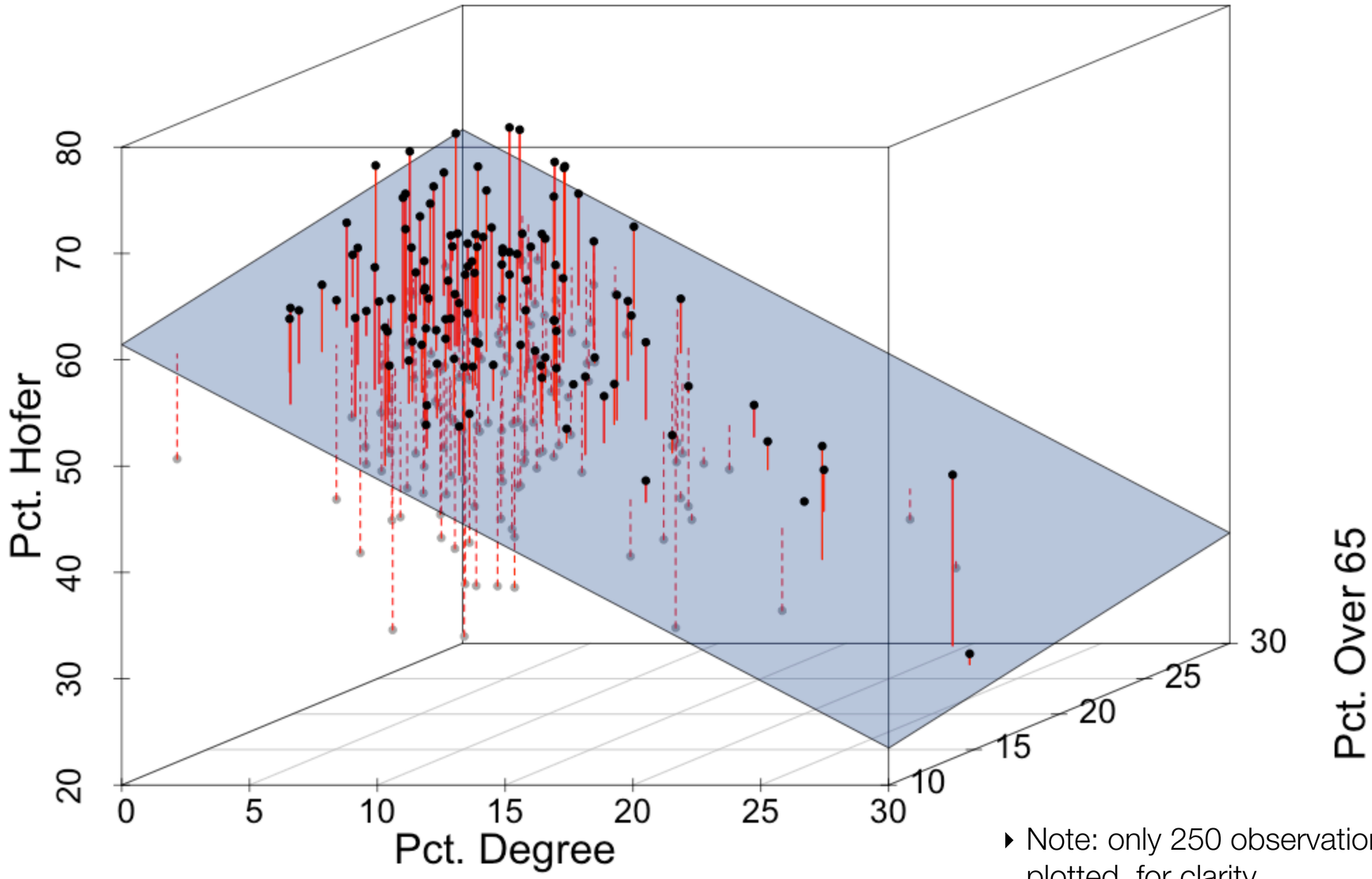
- Note: only 250 observations plotted, for clarity.

Visually…

Why this plane and not other possible planes?

Because this is the plane that minimises the sum of squared residuals

Where residuals are the distance between the observed values of Y and the predicted values $\hat{Y}$, which lie on the plane.



- Note: only 250 observations plotted, for clarity.

universität wien

Multiple Linear Regression

- A one-point increase in the percentage of residents with a degree, is associated with a **1.24-percentage point decrease** in predicted Hofer vote, holding the percentage of residents over 65 years of age constant.

- A one-point increase in the percentage of residents with over 65 years of age is associated with a **0.33-percentage point increase** in predicted Hofer vote, holding the percentage of residents with degrees constant.

```
> model <- lm(data = austria, pct_hofer ~
pct_degree + pct_over65)
> stargazer(model, type = "text")

================================================
                         Dependent variable:
                     ---------------------------
                               pct_hofer
------------------------------------------------
pct_degree                     -1.238***
                                (0.040)
pct_over65                     0.333***
                                (0.050)
Constant                       58.699***
                                (1.052)
------------------------------------------------
Observations                     2,122
R2                               0.323
Adjusted R2                      0.322
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

More Predictors!

- Same story, more Xs:

$$\text{Pct. Hofer} = \alpha + \beta_1 \text{Pct. Degrees} + \beta_2 \text{Pct. Over 65}$$
$$+ \beta_3 \text{Median Wage} + \beta_4 \text{Styria} + \varepsilon$$

- Harder to interpret geometrically: “fitting hyperplanes through multi-dimensional clouds of data points” (?). But the interpretation of the coefficients in terms of **conditional relationships** (not causal effects!) remains valid:

- $\hat{\beta}_3$ is the predicted change in Hofer vote percentage associated with a one-Euro increase in median wages, holding education, age and location (Styria/non-Styria) constant.

- $\hat{\beta}_4$ is the predicted difference in Hofer vote percentage between towns in Styria and elsewhere, holding education, age and median wages constant.

universität wien

```
> model <- lm(data = austria, pct_hofer ~
pct_degree + pct_over65 + median_wage + styria)
> stargazer(model, type = "text")

================================================
                         Dependent variable:
                     ---------------------------
                              pct_hofer
------------------------------------------------
pct_degree                    -1.127***
                               (0.060)
pct_over65                    0.237***
                               (0.049)
median_wage                    -0.0001
                              (0.0001)
styria                        6.239***
                               (0.513)
Constant                      61.862***
                               (2.145)
------------------------------------------------
Observations                     2,122
R2                               0.368
Adjusted R2                      0.367
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Predicted value of Y for an observation with '0' on all Xs

Share of variance in Y explained by the ***whole*** model

universität wien

Categorical Predictors

- So far, our Xs have been numerical or 0-1 binary variables. What if we want to control for a categorical (i.e. nominal or ordinal) variable?
- Trick: a categorical variable with n categories can be recoded as $n - 1$ binary variables.
  - Two categories Styria/non-Styria → one 0-1 variable.
  - Rural / Suburban / Urban → two 0-1 variables:
    Rural/non-Rural, Suburban/non-Suburban.
  - Married/Divorced/Single/Widowed → three 0-1 variables.
- R does this automatically when we pass a categorical predictors in the `lm()` function.

- For instance, if we have a categorical predictor for ‘marital status’ with four categories (Married/Divorced/Single/Widowed), we can run…

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

```
> model <- lm(data = ess, life_satisf ~ marital_status)
> summary(model)

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 **
```

- The intercept represents the predicted value of Y for ‘married’ people: those who have ‘0’ on the binary variables ‘divorced’, ‘single’ and ‘widowed’.

- The slopes represent the difference-in-means between each category and ‘married’

- For instance, if we have a categorical predictor for ‘marital status’ with four categories (Married/Divorced/Single/Widowed), we can run…

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

```
> model <- lm(data = ess, life_satisf ~ marital_status)
> summary(model)

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 **
```

- The ‘omitted’ category (married) is known as the **reference category.**

- It doesn’t matter which category you omit. The same predictions will be the same.

universität wien

```
> model <- lm(data = austria, pct_hofer ~ federal_state)
> stargazer(model, type = "text", single.row = TRUE)
```

```
================================================
                         Dependent variable:
                     ---------------------------
                               pct_hofer
------------------------------------------------
federal_stateKärnten          -1.439 (0.938)
federal_stateNiederösterreich -6.495*** (0.705)
federal_stateOberösterreich   -10.769*** (0.729)
federal_stateSalzburg         -5.009*** (0.966)
federal_stateSteiermark        0.116 (0.782)
federal_stateTirol            -11.264*** (0.786)
federal_stateVorarlberg       -21.764*** (1.032)
federal_stateWien             -29.038*** (1.798)
Constant                      61.090*** (0.619)
------------------------------------------------
Observations                        2,122
R2                                  0.339
Adjusted R2                         0.337
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

universität wien

Why Include Covariates?

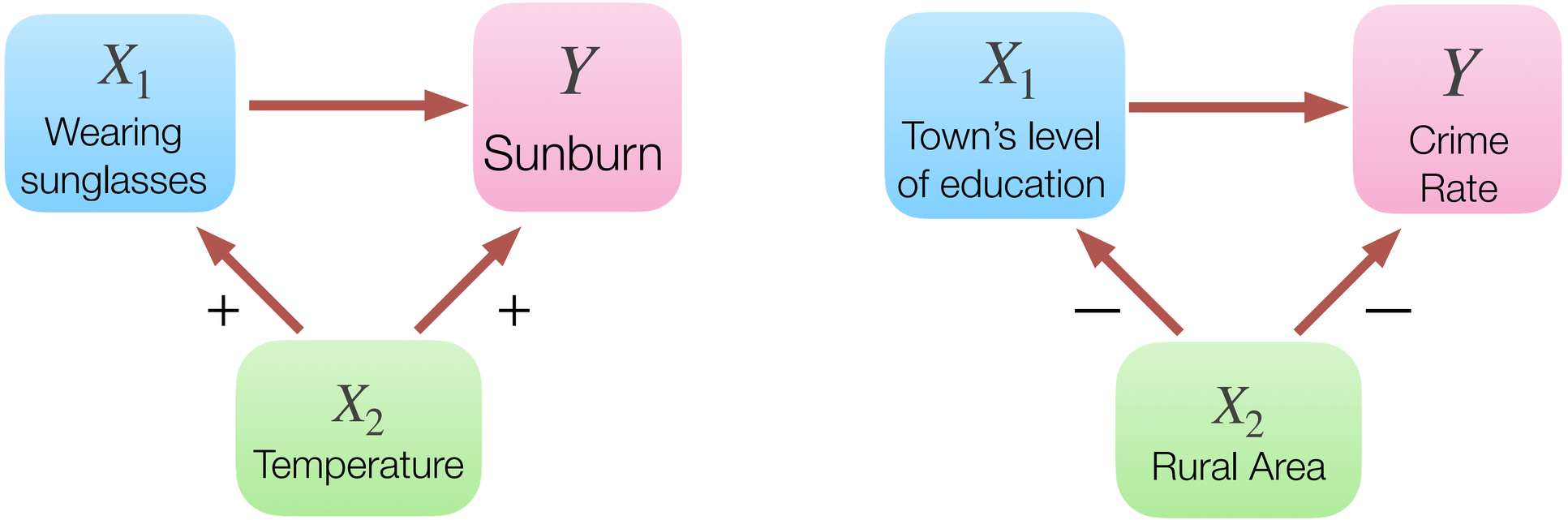
- A reason why we use multiple regression is that it allows to estimate the relationship between X_1 and Y controlling for, or adjusting for other observed variables X_2, X_3 etc. — i.e., keeping them at a constant value.

- If X_2, X_3 etc. are associated both with X_1 and with Y, then adding them to the model will **change our estimate** of the association between X_1 and Y .

- So, we want to choose as 'covariates' (or 'control variables') some variables (X_2, X_3) that are potential **confounders** of the relationship between X_1 and Y. This ensures that the estimated relationship between X_1 and Y is not driven by these confounders, thereby reducing **omitted variable bias**.

Why Include Covariates?

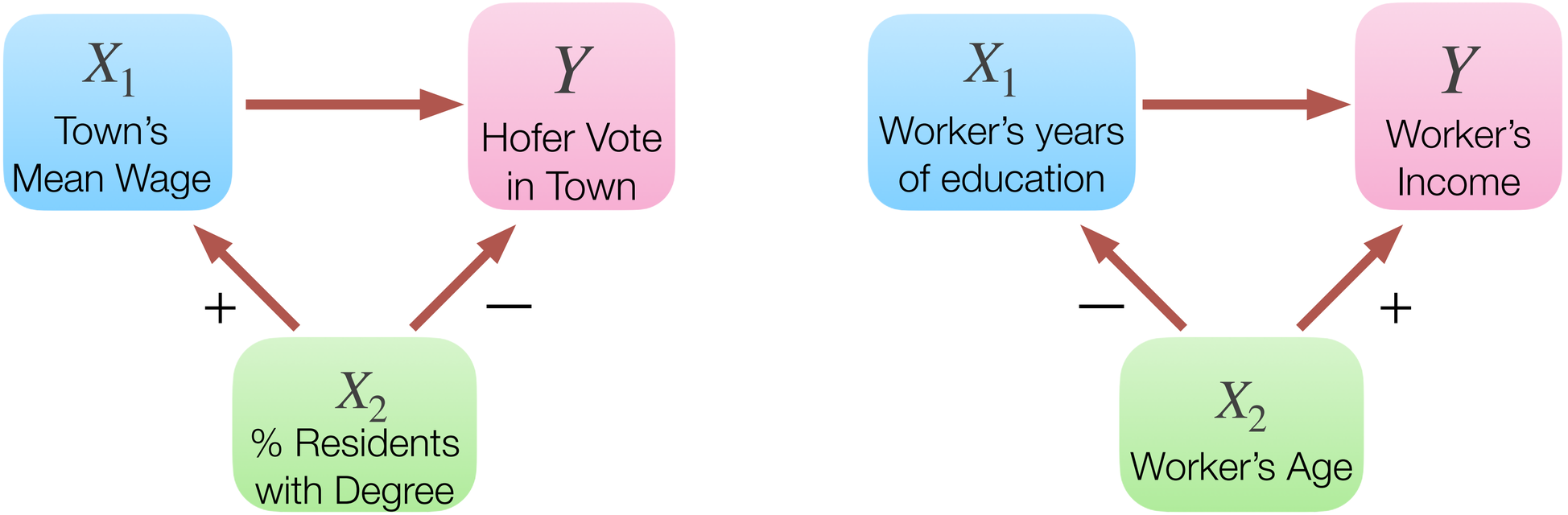
- If we have a model like $Y = \alpha + \beta_1 X_1$ and we include a covariate X_2 so that the model becomes $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$, the slope coefficient β_1 will **go down** if X_2 is associated to **both** X_1 and Y in the **same** direction.



This means that our original estimate for β_1 was positively biased ('too high').

Why Include Covariates?

- The slope coefficient β_1 will instead **go up** if the covariate X_2 is positively associated to X_1 and negatively associated to Y, or vice versa:



This means that our original estimate for β_1 was negatively biased ('too low').

universität wien

Summing Up…

- Multiple linear regression estimates **conditional relationships** that take into account how many independent variables (predictors) relate to each other and to the dependent variable (outcome).
- We are normally interested in interpreting **a single slope coefficient** as the "predicted change in Y associated with a one-unit increase in X, holding covariates constant".
- Categorical variables can be included as 0-1 binary variables: each category's coefficient is the "predicted average difference between units in that category and the reference category, holding covariates constant".