PUBL0055: Introduction to Quantitative Methods

Lecture 5: Regression II (Specification)

Michal Ovádek & Indraneel Sircar

Academic year 2024-25

Lecture Outline

Limits of Bivariate Linear Regression

Multiple Linear Regression

...with Categorical Variables

...with Interactions

Multiple and Adjusted R^2

Conclusion

Limits of Bivariate Linear Regression

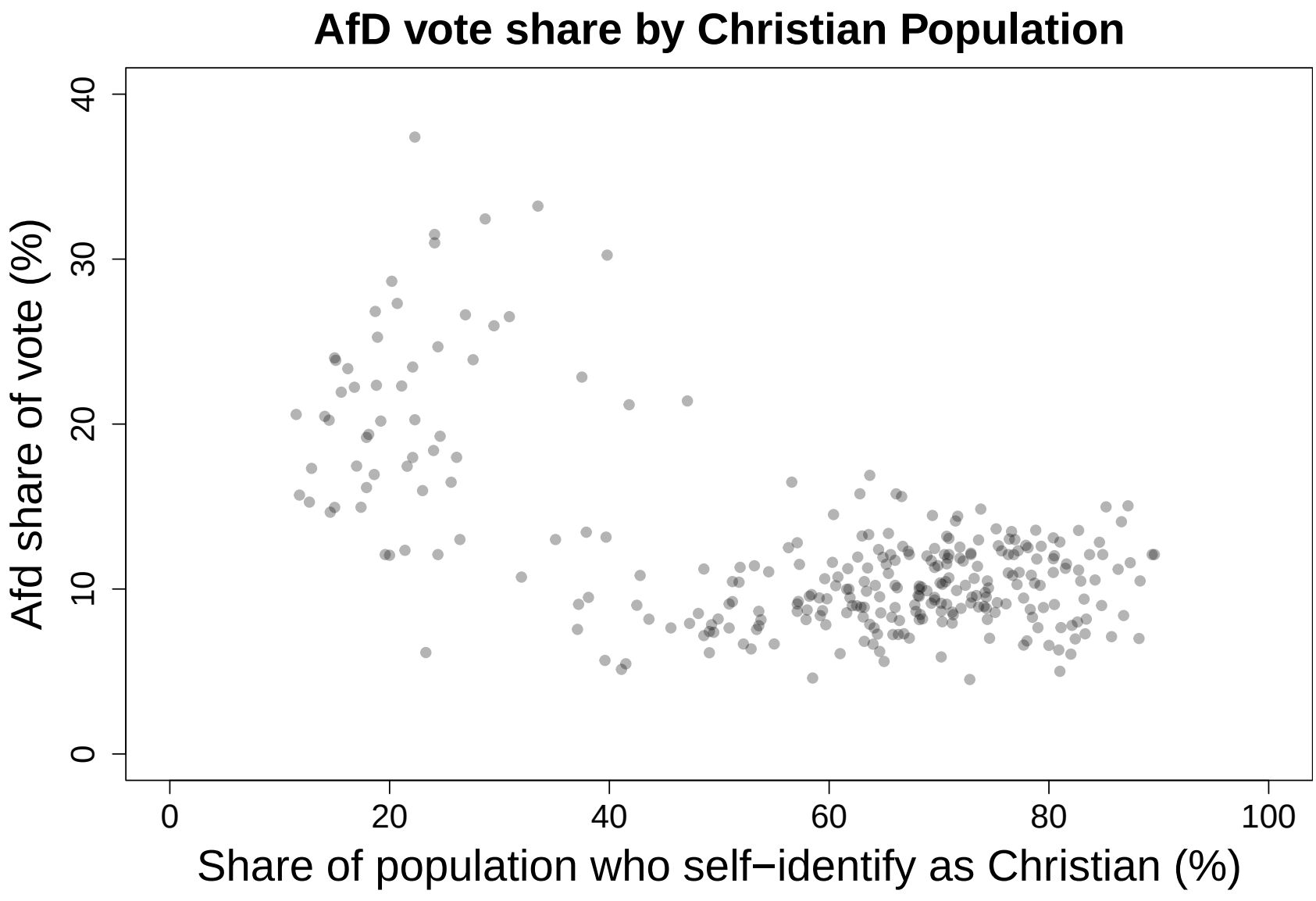
Motivating Example

Christianity and AfD vote share

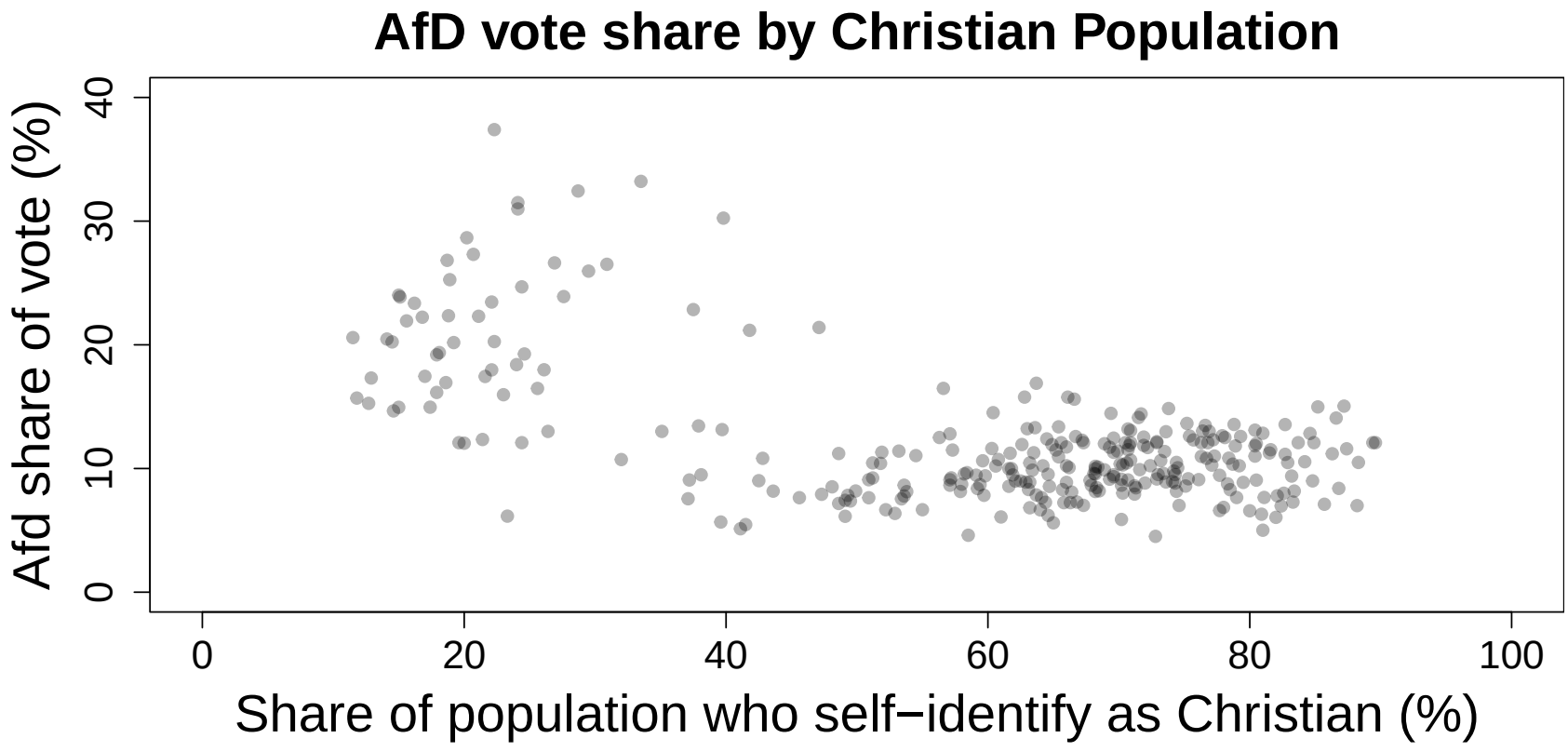
In the 2017 German Bundestag elections, many commentators noticed that the far-right AfD party received much stronger support in areas where there were fewer (self-identified) Christian citizens. We will evaluate the relationship between the "Christianity" of a region and AfD vote share by collecting data on the electoral outcomes of 299 electoral districts.

- Unit of analysis: 299 electoral districts
- Dependent variable (Y): AfD's share of the district vote
- Independent variable (X): Share of a district's population who identify as Christian.

Christianity and AfD vote share

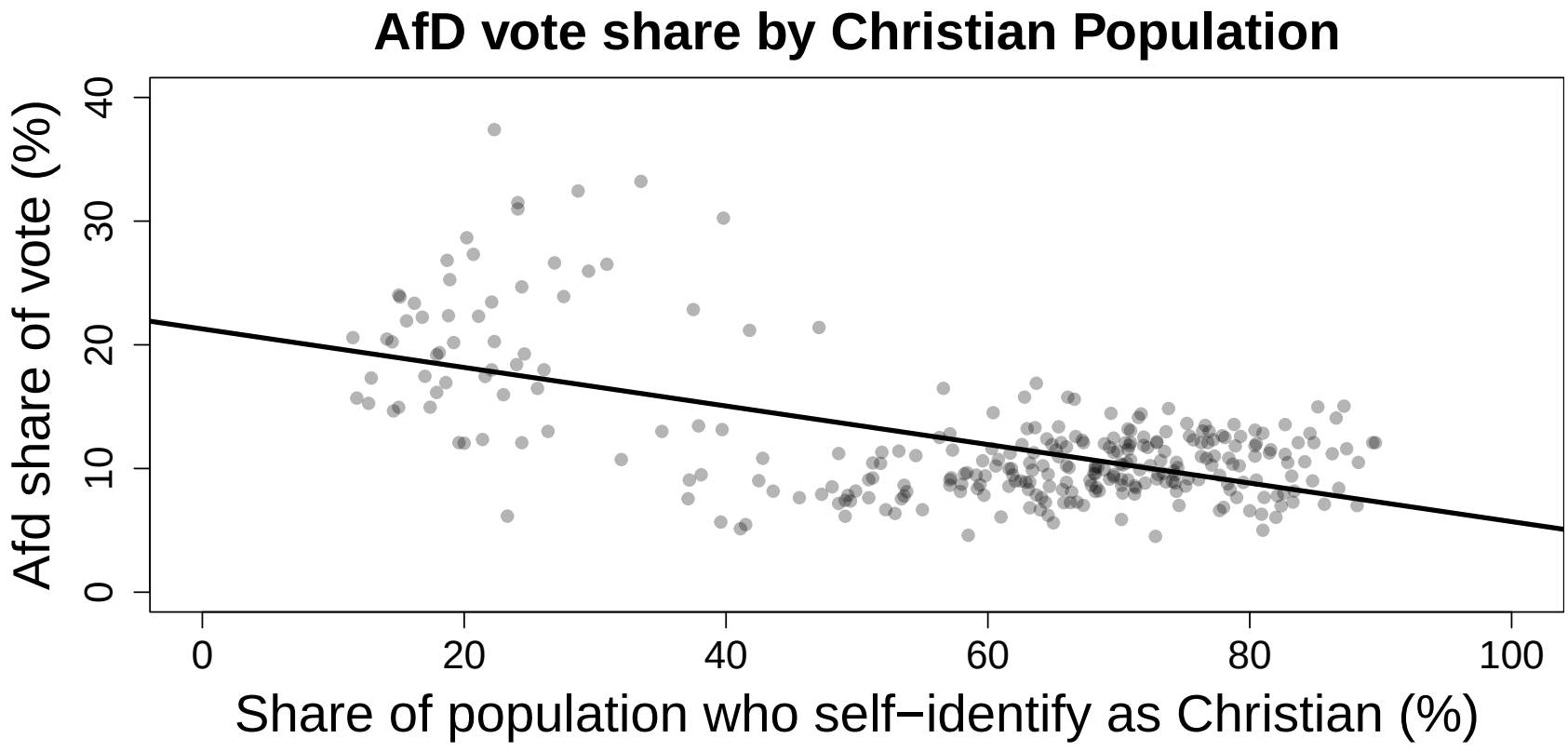


Simple Linear Regression



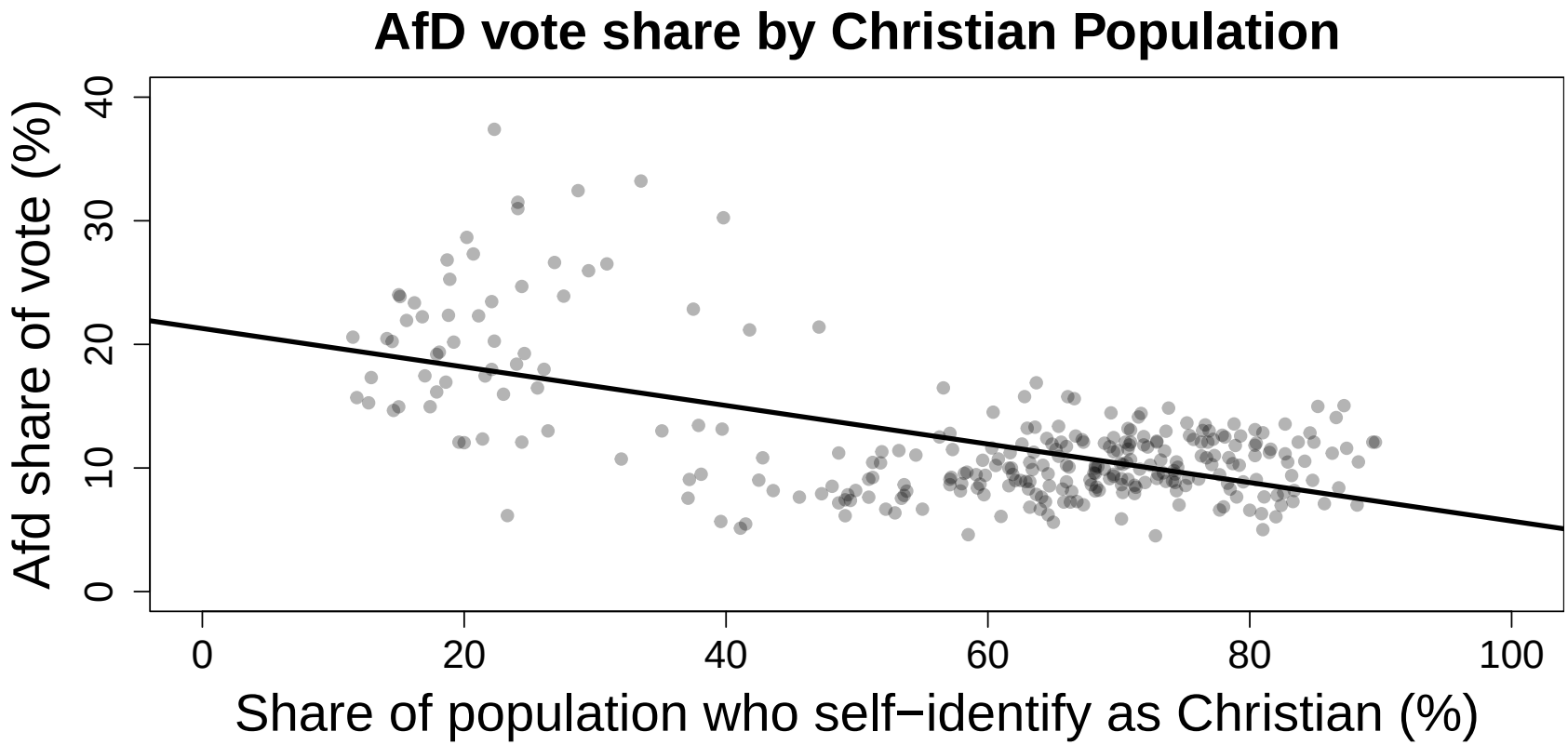
- We already know one way to analyse data like this.
  - We have a continuous dependent variable (AFD Share$_i$)
  - We have a continuous independent variable (% Christian$_i$)

Simple Linear Regression



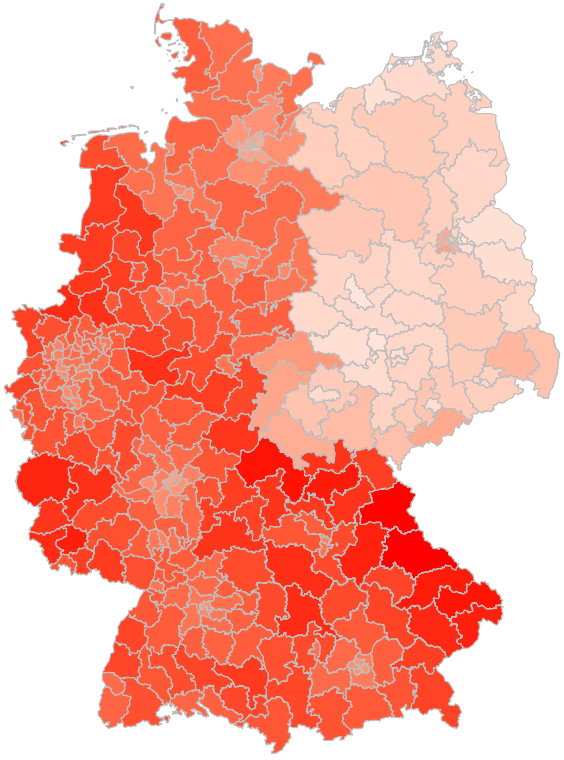
- We already know one way to analyse data like this.
  - We have a continuous dependent variable (AFD Share$_i$)
  - We have a continuous independent variable (% Christian$_i$)
  - $\rightarrow$ **simple/bivariate linear regression!**

Limits of Simple Linear Regression



- But what if understanding variation in AFD Share$_i$ requires paying attention to more than one variable at a time?
- What if % Christian$_i$ is not the only variable we want to consider?

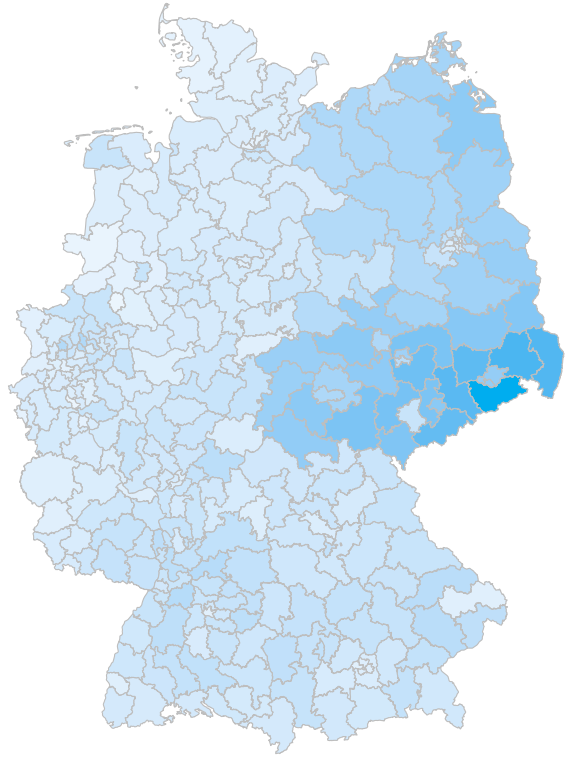
Share of Christians



Christian %

30 50 70

AfD vote share



AfD %

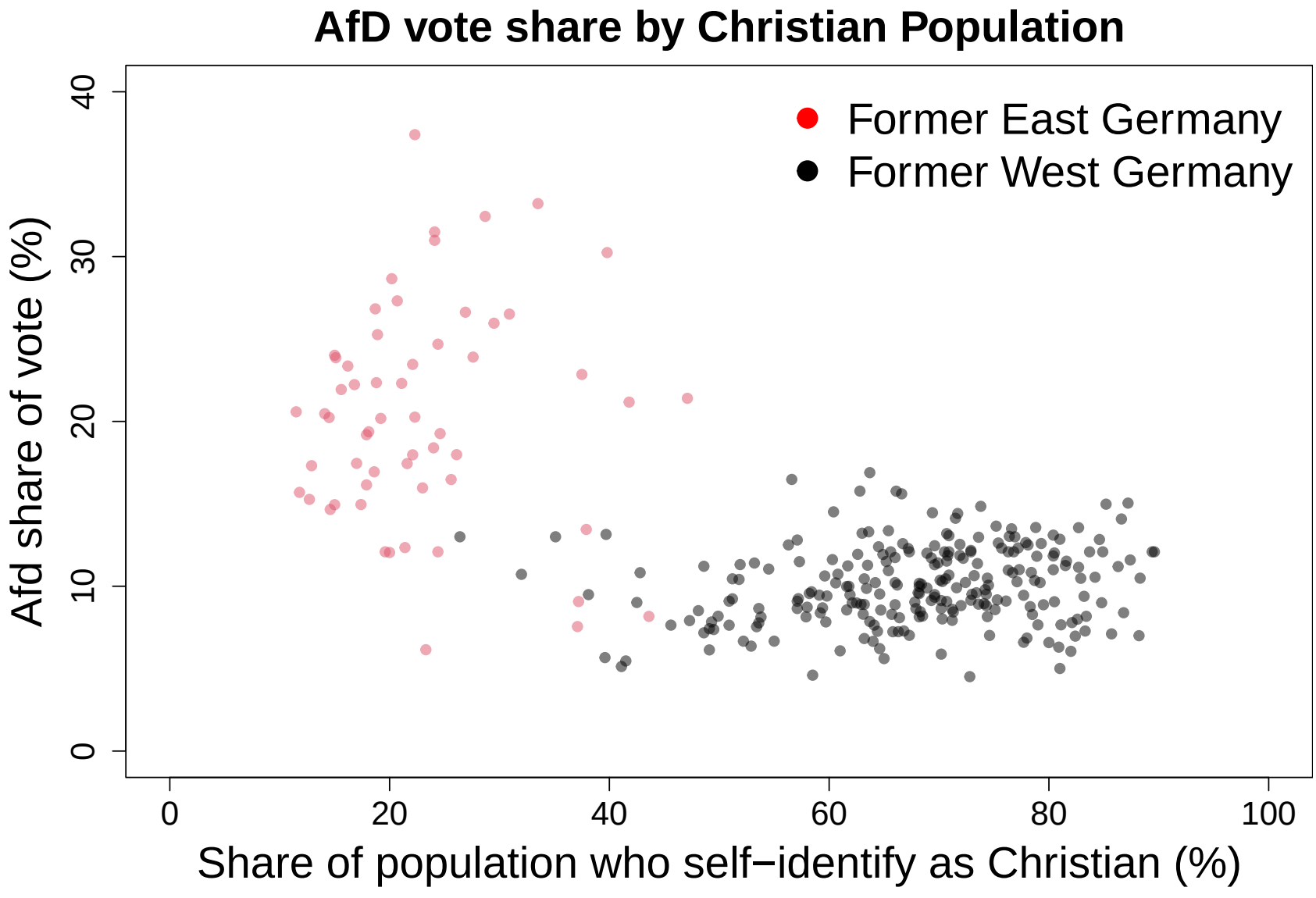
10 20 30

This and next Lecture

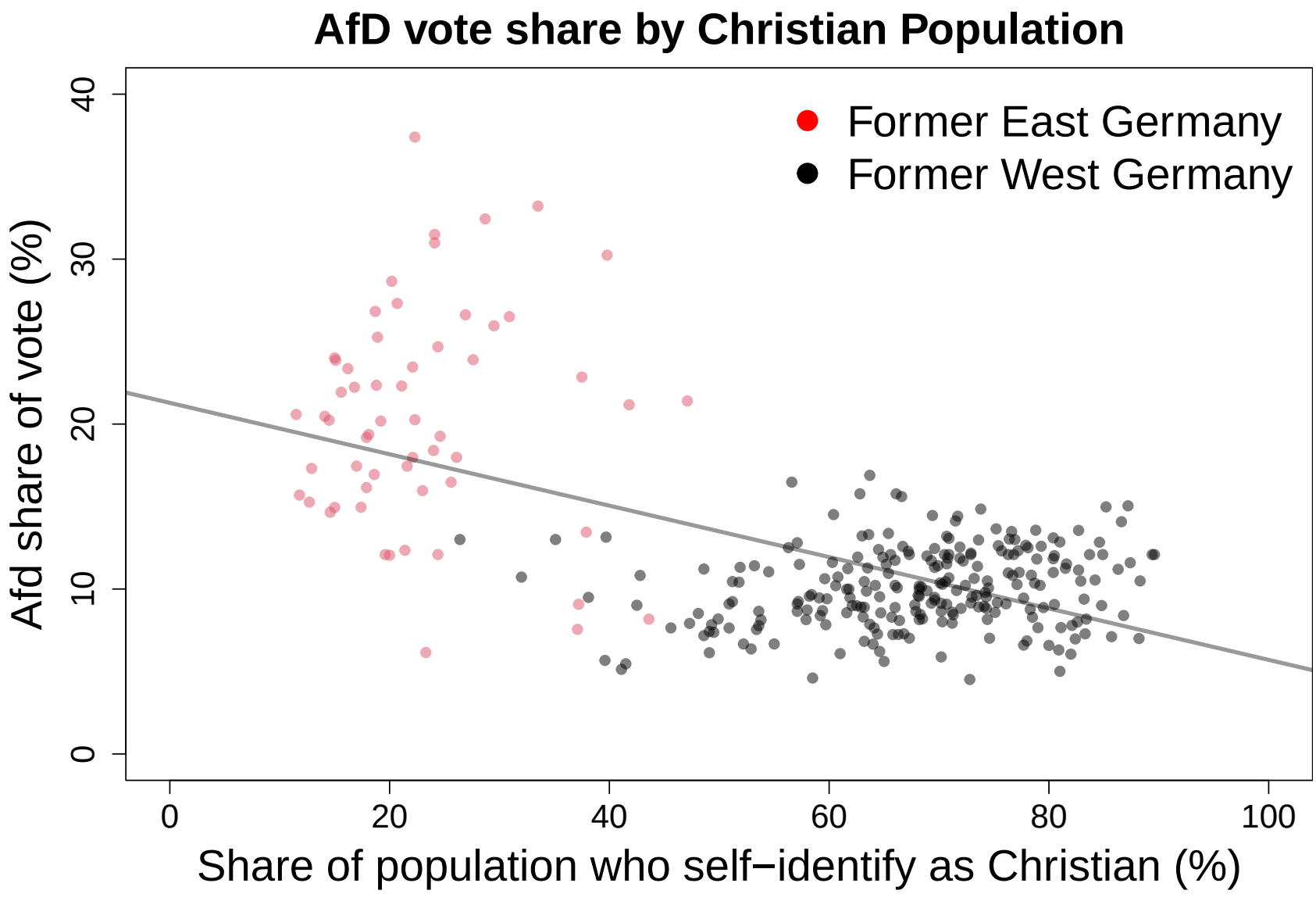
- This lecture we will explore the ways *multiple* linear regression can be used to describe variation in an outcome (like AfD vote share) using more than one explanatory variable.
  - Multiple regression specification for prediction
- Next lecture we will consider how to use Multiple Linear Regression to try establish causal claims about specific explanatory variables.
  - Multiple regression specification for causal inference

Multiple Linear Regression

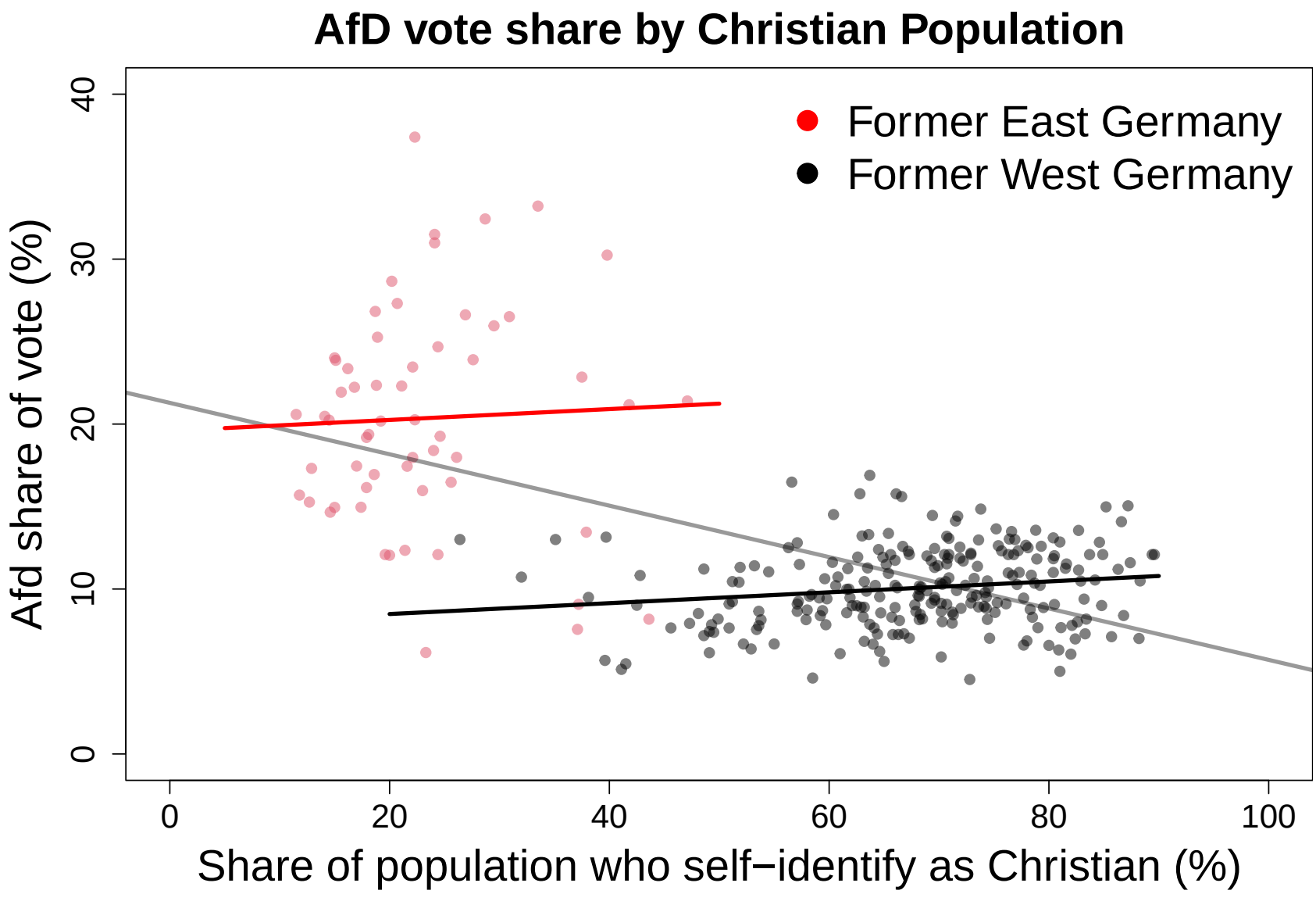
Example: Christianity and vote choice in Germany



Example: Christianity and vote choice in Germany



Example: Christianity and vote choice in Germany



Example: Christianity and vote choice in Germany

- AfD vote share is very different in East and West
- The Christian % is *also* very different in East and West
- If we estimate the relationship between Christianity and AfD vote within East and West Germany things look very different!

Moving beyond one variable

- Multiple regression provides a framework for describing variation in an outcome variable using more than one variable at once
  - Y: AfD vote share
  - X_1: Christian %
  - X_2: Former East

Multiple Regression

The multiple regression model is:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$$

- Observations $i = 1, ..., n$
- Y is the dependent variable
- $X_1, ..., X_k$ are k explanatory variables
- α is the intercept or constant
- $\beta_1, ..., \beta_k$ are the coefficients
- ϵ_i is the error term

Interpretation of $\beta_1, ..., \beta_k$

- Each β coefficient describes the association between the relevant independent variable and the dependent variable, controlling for other explanatory variables
  - These coefficients are known as partial associations
- Consider a model with three explanatory variables:

$$Y_i = (\alpha + \beta_1 X_1 + \beta_2 X_2) + \beta_3 X_3$$

Interpretation of $\beta_1, ..., \beta_k$

- Each β coefficient describes the association between the relevant independent variable and the dependent variable, controlling for other explanatory variables
  - These coefficients are known as partial associations
- Consider a model with three explanatory variables:

$$Y_i = (\alpha + \beta_1 X_1 + \beta_2 X_2) + \beta_3 X_3 = (\text{Others}) + \beta_3 X_3$$

Interpretation of $\beta_1, ..., \beta_k$

- Each β coefficient describes the association between the relevant independent variable and the dependent variable, controlling for other explanatory variables
  - These coefficients are known as partial associations
- Consider a model with three explanatory variables:

$$Y_i = (\alpha + \beta_1 X_1 + \beta_2 X_2) + \beta_3 X_3 = (\text{Others}) + \beta_3 X_3$$

  - Here, "(Others)" is the part of the model that depends on X_1 and X_2 but not on X_3

Interpretation of $\beta_1, ..., \beta_k$

- Each β coefficient describes the association between the relevant independent variable and the dependent variable, controlling for other explanatory variables
  - These coefficients are known as partial associations
- Consider a model with three explanatory variables:

$$Y_i = (\alpha + \beta_1 X_1 + \beta_2 X_2) + \beta_3 X_3 = (\text{Others}) + \beta_3 X_3$$

  - Here, "(Others)" is the part of the model that depends on X_1 and X_2 but not on X_3
- If X_3 increases by one unit, and X_1 and X_2 remain constant, the expected (i.e. *average*) value of Y will change by β_3 units

Interpretation of $\beta_1, ..., \beta_k$

Consider the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

Interpretation of $\beta_1, ..., \beta_k$

Consider the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

Now change X_1 by one unit (ΔX_1) and that will add something to Y

$$Y + \Delta Y = \alpha + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 \tag{2}$$

Interpretation of $\beta_1, ..., \beta_k$

Consider the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

Now change X_1 by one unit (ΔX_1) and that will add something to Y

$$Y + \Delta Y = \alpha + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 \tag{2}$$

What is the difference between (1) and (2)?

$$\begin{aligned} \Delta Y &= \beta_1 \cdot \Delta X_1 \\ \frac{\Delta Y}{\Delta X_1} &= \beta_1 \end{aligned}$$

Interpretation of $\beta_1, ..., \beta_k$

Consider the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

Now change X_1 by one unit (ΔX_1) and that will add something to Y

$$Y + \Delta Y = \alpha + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 \tag{2}$$

What is the difference between (1) and (2)?

$$\begin{aligned} \Delta Y &= \beta_1 \cdot \Delta X_1 \\ \frac{\Delta Y}{\Delta X_1} &= \beta_1 \end{aligned}$$

- β_1 is therefore the average change in Y associated with a 1-unit change in X_1 when X_2 stays constant.

Example: AfD vote share

- Dependent variable (Y): AfD vote share (continuous)
- 1st explanatory variable (X_1): Christian share (continuous)
- 2nd explanatory variable (X_2): East-West (binary, East $= 1$)

Example: AfD vote share

- Dependent variable (Y): AfD vote share (continuous)
- 1st explanatory variable (X_1): Christian share (continuous)
- 2nd explanatory variable (X_2): East-West (binary, East $= 1$)

$$\text{AfD Share}_i = \alpha + \beta_1 \cdot \text{Christian Share}_i + \beta_2 \cdot \text{East}_i$$

Example: AfD vote share

- Dependent variable (Y): AfD vote share (continuous)
- 1st explanatory variable (X_1): Christian share (continuous)
- 2nd explanatory variable (X_2): East-West (binary, East $= 1$)

$$\text{AfD Share}_i = \alpha + \beta_1 \cdot \text{Christian Share}_i + \beta_2 \cdot \text{East}_i$$

A one-unit increase in Christian Share is associated with a β_1 change in the average AfD Share, holding constant East-West location

Example: AfD vote share

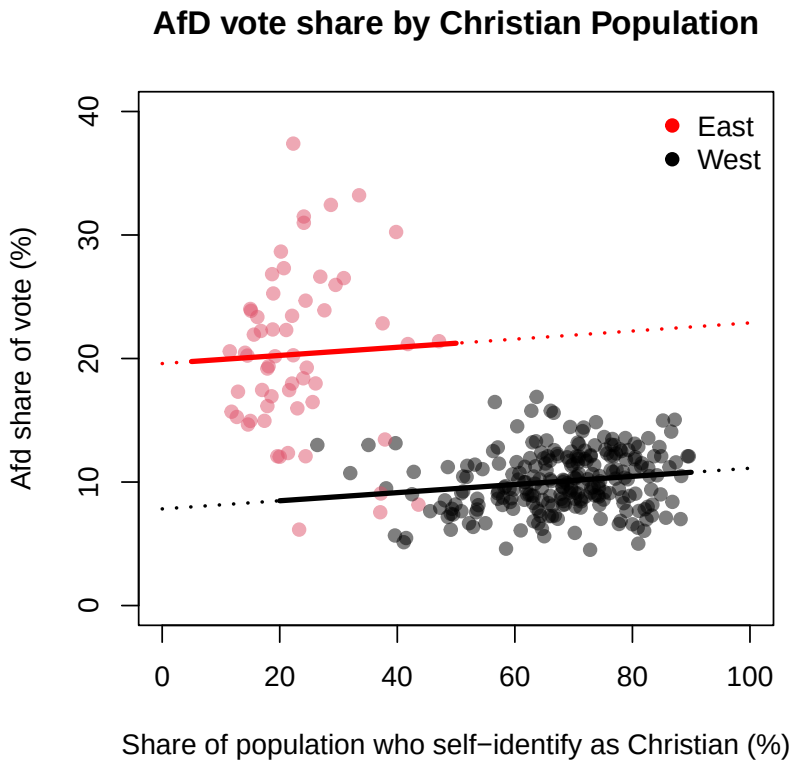
- Dependent variable (Y): AfD vote share (continuous)
- 1st explanatory variable (X_1): Christian share (continuous)
- 2nd explanatory variable (X_2): East-West (binary, East = 1)

$$\text{AfD Share}_i = \alpha + \beta_1 \cdot \text{Christian Share}_i + \beta_2 \cdot \text{East}_i$$

A one-unit increase in Christian Share is associated with a β_1 change in the average AfD Share, holding constant East-West location

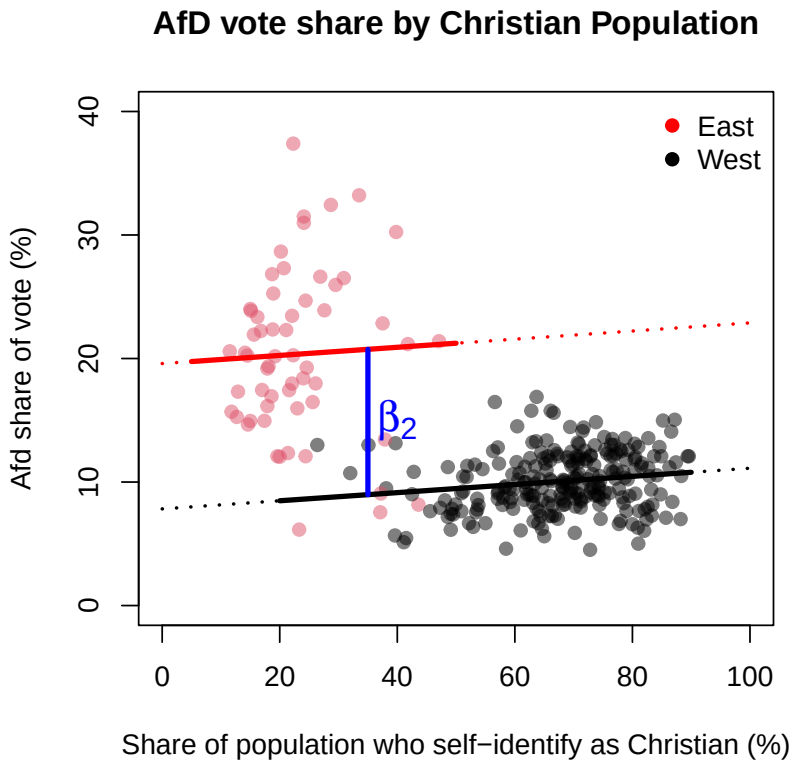
Eastern districts are associated with a β_2 change in the average AfD Share relative to Western districts, holding constant Christian Share

Interpretation



- Let's 'hold constant' our *East-West* explanatory variable
- β_1 describes the slope of the lines *within* East and West districts

Interpretation



- Let's 'hold constant' our *Christianity* explanatory variable
- β_2 describes the distance between lines

Multiple linear regression in R

```r
# our original model with one explanatory variable
linear_model_1 <- lm(AfD ~ christian, data = results)

# our new model, with two explanatory variables
linear_model_2 <- lm(AfD ~ christian + east, data = results)
```

Multiple linear regression in R

```
summary(linear_model_2)
```

```
##
## Call:
## lm(formula = AfD ~ christian + east, data = results)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2099  -1.8774  -0.0847   1.8863  17.0719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.82484    1.29957   6.021 5.12e-09 ***
## christian    0.03293    0.01883   1.749   0.0814 .
## eastTRUE    11.76672    0.99423  11.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.614 on 296 degrees of freedom
## Multiple R-squared:  0.5636, Adjusted R-squared:  0.5606
## F-statistic: 191.1 on 2 and 296 DF,  p-value: < 2.2e-16
```

Multiple linear regression output

	AfD
christian	0.033
	(0.019)
east	11.767*
	(0.994)
Constant	7.825*
	(1.300)
Observations	299
R^2	0.564
Note:	*p<0.05

- A one percentage point increase in the share of Christians is associated with a 0.03 point increase in the AfD vote share (percentage) *on average*,

Multiple linear regression output

	AfD
christian	0.033
	(0.019)
east	11.767*
	(0.994)
Constant	7.825*
	(1.300)
Observations	299
R^2	0.564
Note:	*p<0.05

- A one percentage point increase in the share of Christians is associated with a 0.03 point increase in the AfD vote share (percentage) *on average*, holding constant East-West location

Multiple linear regression output

	AfD
christian	0.033
	(0.019)
east	11.767*
	(0.994)
Constant	7.825*
	(1.300)
Observations	299
R^2	0.564
Note:	*p<0.05

- A one percentage point increase in the share of Christians is associated with a 0.03 point increase in the AfD vote share (percentage) *on average*, holding constant East-West location
- Eastern districts are associated with 11.8 percentage points higher AfD vote share *on average*,

Multiple linear regression output

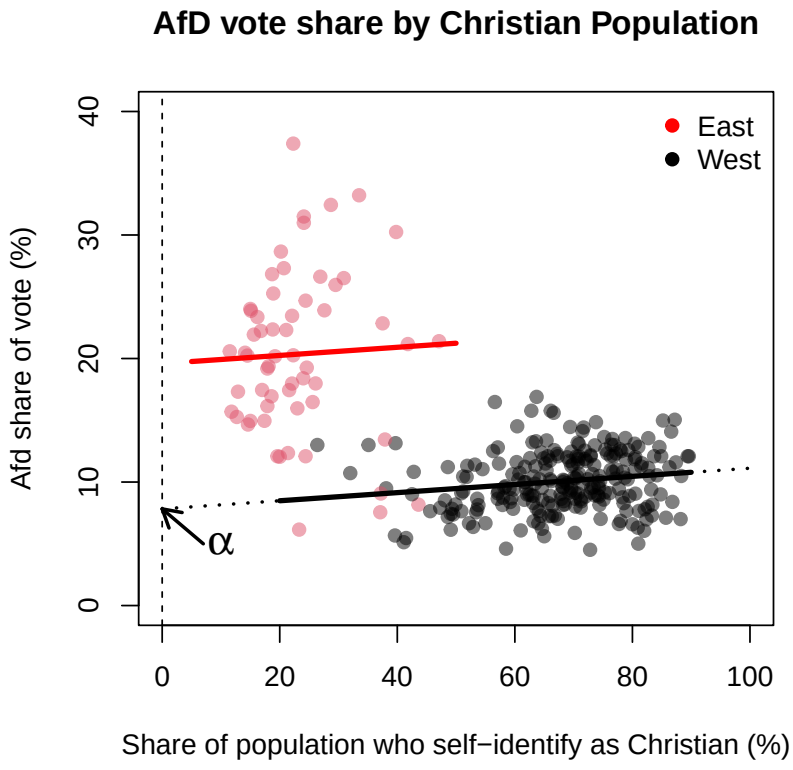
	AfD
christian	0.033
	(0.019)
east	11.767*
	(0.994)
Constant	7.825*
	(1.300)
Observations	299
R^2	0.564
Note:	*p<0.05

- A one percentage point increase in the share of Christians is associated with a 0.03 point increase in the AfD vote share (percentage) *on average*, holding constant East-West location
- Eastern districts are associated with 11.8 percentage points higher AfD vote share *on average*, holding constant the share of Christians

Interpretation of α

- Last week, the interpretation of $\hat{\alpha}$ was the average value of Y when $X = 0$
- Now we have more than one X, $\hat{\alpha}$ represents the average value of Y when **all** X variables are equal to zero
- As we add more and more independent variables, $\hat{\alpha}$ becomes less likely to be a quantity that has a substantively interesting interpretation

Interpretation of α



- Remember, our X_2 variable for East-West is equal to 1 for East districts and 0 for West districts
- $\hat{\alpha}$ is therefore the point at which the black line intersects the Y-axis

More than two independent variables

- We might think that the percentage of migrants in a district is *also* associated with AfD vote share

More than two independent variables

- We might think that the percentage of migrants in a district is *also* associated with AfD vote share
- If we want to incorporate this (continuous) variable, we have:

$$\text{AfD Share}_i = \alpha + \beta_1 \cdot \text{Christian}_i + \beta_2 \cdot \text{East}_i + \beta_3 \cdot \text{Migrant}_i + \epsilon_i$$

More than two independent variables

- We might think that the percentage of migrants in a district is *also* associated with AfD vote share
- If we want to incorporate this (continuous) variable, we have:

$$\text{AfD Share}_i = \alpha + \beta_1 \cdot \text{Christian}_i + \beta_2 \cdot \text{East}_i + \beta_3 \cdot \text{Migrant}_i + \epsilon_i$$

- ... and the interpretation remains the same:
  - β_k represents the average change in Y associated with a one-unit increase in X_k, holding all other explanatory variables constant

More than two independent variables

```
linear_model_3 <- lm(AfD ~ christian + east + migrantfraction, data = results)
library(texreg)
screenreg(list(linear_model_1, linear_model_2, linear_model_3),digits = 2)
```

```
##
## ============================================
##                  Model 1     Model 2     Model 3
## --------------------------------------------
## (Intercept)       21.29 ***    7.82 ***   11.78 ***
##                   (0.76)      (1.30)      (1.90)
## christian         -0.16 ***    0.03        0.00
##                   (0.01)      (0.02)      (0.02)
## eastTRUE                      11.77 ***    9.14 ***
##                               (0.99)      (1.35)
## migrantfraction                           -0.09 **
##                                           (0.03)
## --------------------------------------------
## R^2                0.36        0.56        0.58
## Adj. R^2           0.35        0.56        0.57
## Num. obs.         299         299         299
## ============================================
```

Fitted values

As before, we can calculate fitted values for our model:

- The fitted values ($\hat{Y}$) are:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$$

- Interpretation: The fitted values tell us the best guess for Y for specific values of $X_1, X_2, ..., X_k$

Fitted values

$$\widehat{\mathsf{AfD}}_i = \hat{\alpha} + \hat{\beta}_1 \mathsf{Christian}_i + \hat{\beta}_2 \mathsf{East}_i + \hat{\beta}_3 \mathsf{Migrant}_i$$

Fitted values

$$\begin{aligned}
\widehat{\mathsf{AfD}}_i &= \hat{\alpha} + \hat{\beta}_1 \mathsf{Christian}_i + \hat{\beta}_2 \mathsf{East}_i + \hat{\beta}_3 \mathsf{Migrant}_i \\
\widehat{\mathsf{AfD}}_i &= 11.78 + 0.004 \cdot \mathsf{Christian}_i + 9.14 \cdot \mathsf{East}_i - 0.09 \cdot \mathsf{Migrant}_i
\end{aligned}$$

Fitted values

$$\widehat{\text{AfD}}_i = \hat{\alpha} + \hat{\beta}_1 \text{Christian}_i + \hat{\beta}_2 \text{East}_i + \hat{\beta}_3 \text{Migrant}_i$$

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot \text{Christian}_i + 9.14 \cdot \text{East}_i - 0.09 \cdot \text{Migrant}_i$$

Question: What is the fitted value of AfD vote share for a district in the East, with 40% Christian and 5% migrant population?

Fitted values

$$\widehat{\text{AfD}}_i = \hat{\alpha} + \hat{\beta}_1 \text{Christian}_i + \hat{\beta}_2 \text{East}_i + \hat{\beta}_3 \text{Migrant}_i$$

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot \text{Christian}_i + 9.14 \cdot \text{East}_i - 0.09 \cdot \text{Migrant}_i$$

Question: What is the fitted value of AfD vote share for a district in the East, with 40% Christian and 5% migrant population?

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot 40 + 9.14 \cdot 1 - 0.09 \cdot 5$$

Fitted values

$$\begin{aligned}
\widehat{\text{AfD}}_i &= \hat{\alpha} + \hat{\beta}_1 \text{Christian}_i + \hat{\beta}_2 \text{East}_i + \hat{\beta}_3 \text{Migrant}_i \\
\widehat{\text{AfD}}_i &= 11.78 + 0.004 \cdot \text{Christian}_i + 9.14 \cdot \text{East}_i - 0.09 \cdot \text{Migrant}_i
\end{aligned}$$

Question: What is the fitted value of AfD vote share for a district in the East, with 40% Christian and 5% migrant population?

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot 40 + 9.14 \cdot 1 - 0.09 \cdot 5 = 20.63$$

Fitted values

$$\begin{aligned}
\widehat{\text{AfD}}_i &= \hat{\alpha} + \hat{\beta}_1 \text{Christian}_i + \hat{\beta}_2 \text{East}_i + \hat{\beta}_3 \text{Migrant}_i \\
\widehat{\text{AfD}}_i &= 11.78 + 0.004 \cdot \text{Christian}_i + 9.14 \cdot \text{East}_i - 0.09 \cdot \text{Migrant}_i
\end{aligned}$$

Question: What is the fitted value of AfD vote share for a district in the East, with 40% Christian and 5% migrant population?

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot 40 + 9.14 \cdot 1 - 0.09 \cdot 5 = 20.63$$

Question: What is the fitted value of AfD vote share for a district in the West, with 20% Christian population and 15% migrants?

Fitted values

$$\begin{aligned}
\widehat{\text{AfD}}_i &= \hat{\alpha} + \hat{\beta}_1 \text{Christian}_i + \hat{\beta}_2 \text{East}_i + \hat{\beta}_3 \text{Migrant}_i \\
\widehat{\text{AfD}}_i &= 11.78 + 0.004 \cdot \text{Christian}_i + 9.14 \cdot \text{East}_i - 0.09 \cdot \text{Migrant}_i
\end{aligned}$$

Question: What is the fitted value of AfD vote share for a district in the East, with 40% Christian and 5% migrant population?

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot 40 + 9.14 \cdot 1 - 0.09 \cdot 5 = 20.63$$

Question: What is the fitted value of AfD vote share for a district in the West, with 20% Christian population and 15% migrants?

$$\widehat{\text{AfD}}_i = 11.78 + 0.004 \cdot 20 + 9.14 \cdot 0 - 0.09 \cdot 15 = 10.47$$

...with Categorical Variables

Categorical Variables and "Qualitative Information"

- We have already seen that we can incorporate qualitative information by using dummy variables
  - Our `east` variable indicated whether a given district was located in (old) East Germany

Categorical Variables and "Qualitative Information"

- We have already seen that we can incorporate qualitative information by using dummy variables
  - Our `east` variable indicated whether a given district was located in (old) East Germany
- We can also include information for many groups
  - The 299 districts in Germany are clustered in 16 'regions'

Categorical Variables and Regression

- We do so by including a **set** of dummy variables for all groups (regions) except for one in the regression
- The category without a dummy is the reference or baseline category
- The coefficient of the category is the expected difference in Y between the category and the baseline
- The choice of baseline is arbitrary: the model is identical in substantive terms

Baseline / Dropping a level

- The choice of baseline is arbitrary but necessary
- If we included all categories (levels) of a categorical variable we would fall into the dummy variable trap
- Suppose a variable has one of three values: red, blue or green. After dummy encoding (one-hot encoding), if $red = 0$ and $green = 0$ then $blue$ must be 1.
- When one or several variables perfectly predict the value of another, this is called *perfect multicollinearity*
- OLS does not work in the presence of perfect multicollinearity ($X^T X$ is non-invertible, solutions are not uniquely defined)

Categorical Variables

- Example set of dummy variables for a categorical variable with four German regions: Hamburg, Bayern, Berlin, Brandenburg
- The reference category is Brandenburg:

Categorical Variables

- Example set of dummy variables for a categorical variable with four German regions: Hamburg, Bayern, Berlin, Brandenburg
- The reference category is Brandenburg:

	$X_{Hamburg}$	X_{Bayern}	X_{Berlin}
Bayern	0	1	0
Bayern	0	1	0
Berlin	0	0	1
Hamburg	1	0	0
Brandenburg	0	0	0
Bayern	0	1	0
⋮	⋮	⋮	⋮

- `R` will automatically convert any factor variable into a set of dummies, and will choose a baseline category

Categorical Variable Example

We will use the `region` variable from our data:

```
##
## BB BE BW BY HB HE HH MV NI NW RP SH SL SN ST TH
## 10 12 38 46  2 22  6  6 30 64 15 11  4 16  9  8
```

This shows the number of districts in each region in the data.

We can estimate a model with a categorical variable as before:

```
linear_model_4 <- lm(AfD ~ christian + region ,
                     data = results)
```

Categorical Variable Example

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  19.38519398 0.98617499 19.6569516 8.906315e-55
## christian     0.01948636 0.01956459  0.9960014 3.201033e-01
## regionBE     -7.88621573 1.23122444 -6.4051813 6.268269e-10
## regionBW     -9.19481347 1.39715459 -6.5810996 2.269137e-10
## regionBY     -9.99382116 1.46489653 -6.8222028 5.466454e-11
## regionHB    -10.76355393 2.29222915 -4.6956710 4.154088e-06
## regionHE     -9.78162918 1.38520152 -7.0615207 1.286179e-11
## regionHH    -11.89533599 1.52194233 -7.8158914 1.089623e-13
## regionMV     -1.60037325 1.47391164 -1.0858000 2.784948e-01
## regionNI    -11.90378488 1.36936299 -8.6929360 2.951986e-16
## regionNW    -11.25031787 1.34611890 -8.3575960 2.953719e-15
## regionRP    -10.56147953 1.58461659 -6.6650063 1.388249e-10
## regionSH    -13.04309714 1.45041829 -8.9926453 3.612896e-17
## regionSL    -11.64505844 2.06876178 -5.6289992 4.374645e-08
## regionSN      6.40523554 1.15243488  5.5580021 6.321840e-08
## regionST     -1.26241031 1.31281745 -0.9616038 3.370724e-01
## regionTH      2.49520321 1.36963016  1.8218080 6.954336e-02
```

- In our data, `region` is a categorical (factor) variable
- Brandenburg (BB) is the baseline category

Categorical Variable Example

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  19.38519398 0.98617499 19.6569516 8.906315e-55
## christian     0.01948636 0.01956459  0.9960014 3.201033e-01
## regionBE     -7.88621573 1.23122444 -6.4051813 6.268269e-10
## regionBW     -9.19481347 1.39715459 -6.5810996 2.269137e-10
## regionBY     -9.99382116 1.46489653 -6.8222028 5.466454e-11
## regionHB    -10.76355393 2.29222915 -4.6956710 4.154088e-06
## regionHE     -9.78162918 1.38520152 -7.0615207 1.286179e-11
## regionHH    -11.89533599 1.52194233 -7.8158914 1.089623e-13
## regionMV     -1.60037325 1.47391164 -1.0858000 2.784948e-01
## regionNI    -11.90378488 1.36936299 -8.6929360 2.951986e-16
## regionNW    -11.25031787 1.34611890 -8.3575960 2.953719e-15
## regionRP    -10.56147953 1.58461659 -6.6650063 1.388249e-10
## regionSH    -13.04309714 1.45041829 -8.9926453 3.612896e-17
## regionSL    -11.64505844 2.06876178 -5.6289992 4.374645e-08
## regionSN      6.40523554 1.15243488  5.5580021 6.321840e-08
## regionST     -1.26241031 1.31281745 -0.9616038 3.370724e-01
## regionTH      2.49520321 1.36963016  1.8218080 6.954336e-02
```

Controlling for the Christian share of the population, the AfD share in Berlin (BE) is 7.9 percentage points lower than in Brandenburg on average

...with Interactions

Interactions

Interactions

An **interaction** exists between two explanatory variables when the relationship between (either) one of them and the dependent variable depends on the value of the other.

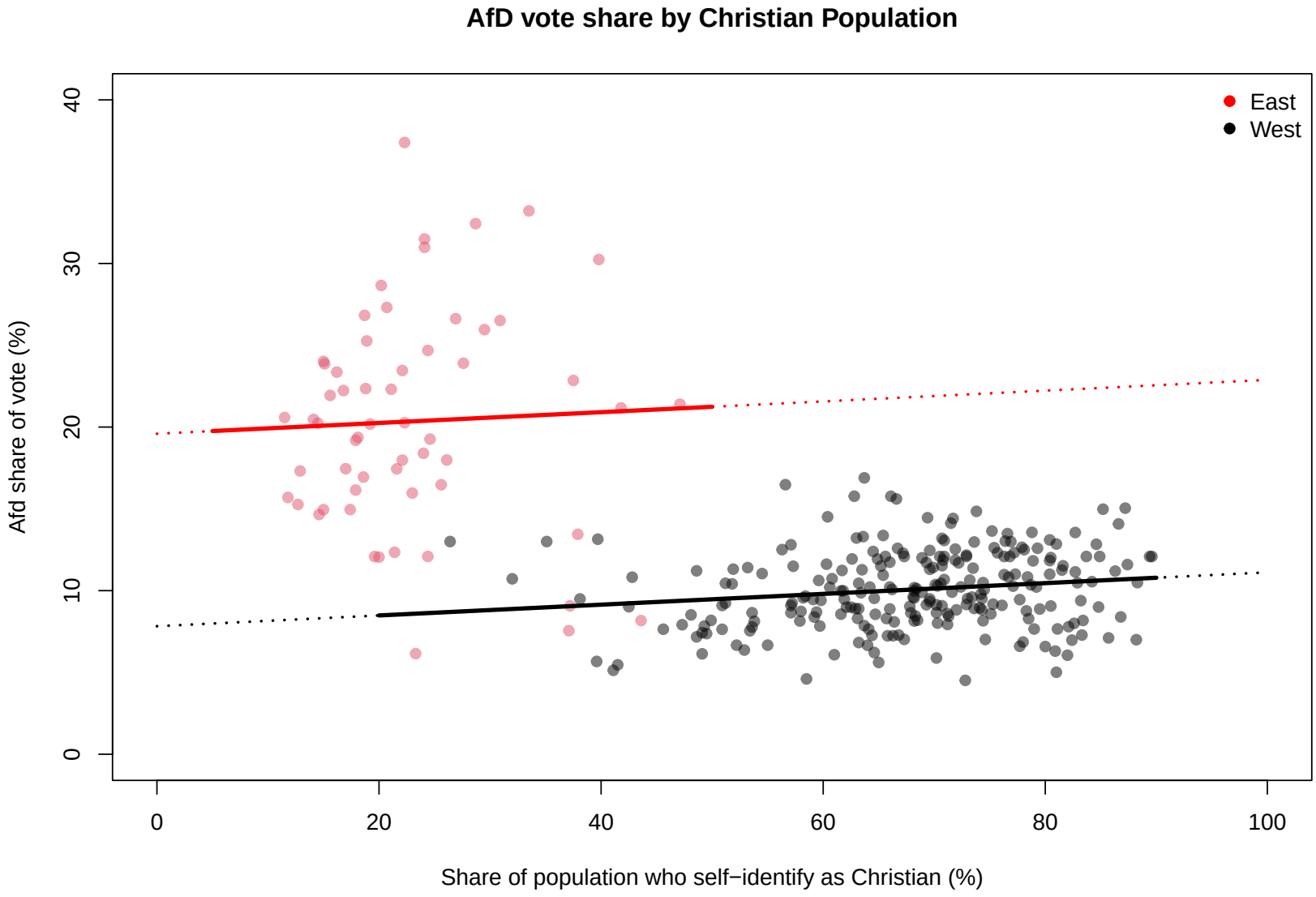
Interactions

Interactions

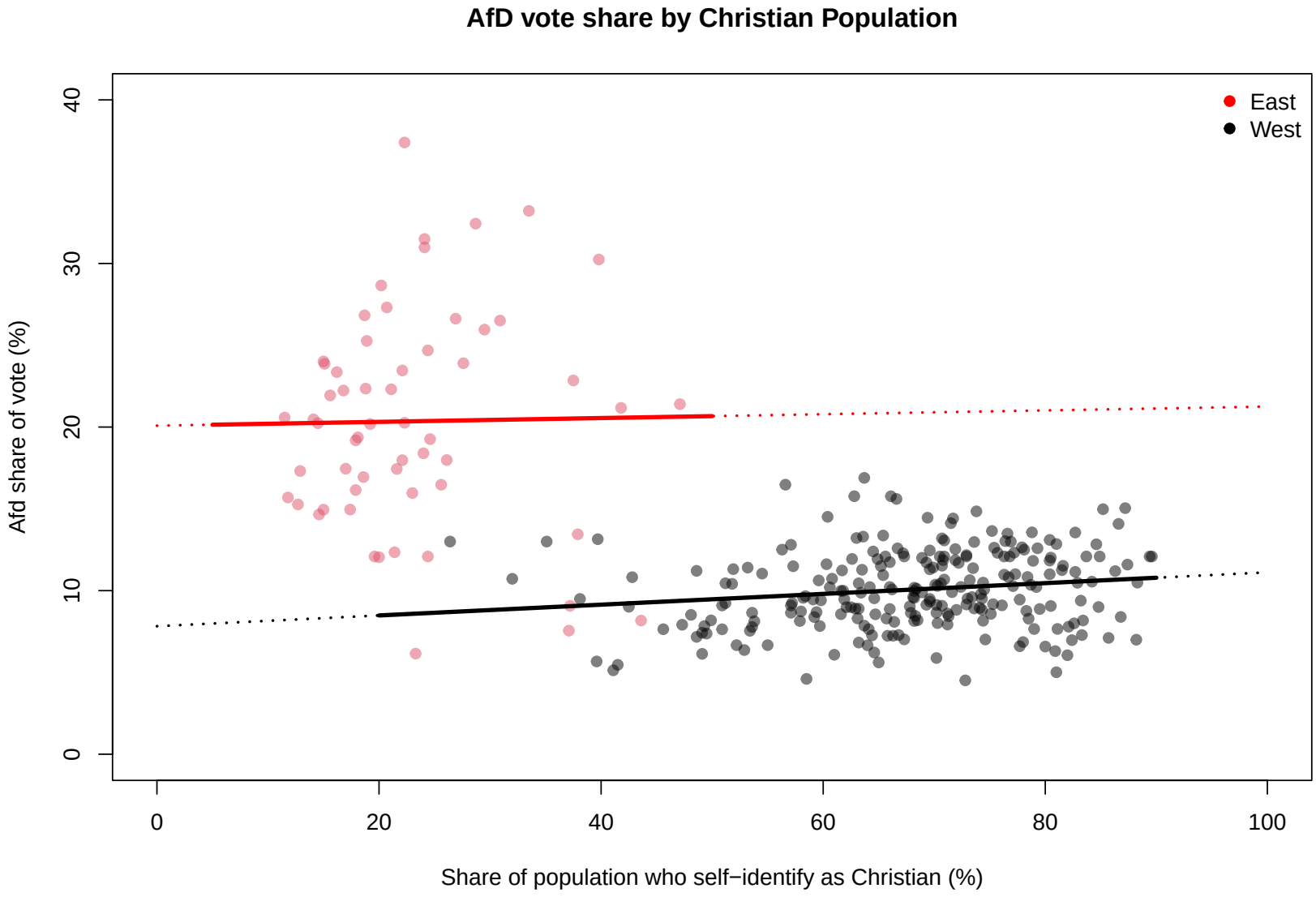
An **interaction** exists between two explanatory variables when the relationship between (either) one of them and the dependent variable depends on the value of the other.

- We can build this intuition into the linear regression model by including the **product** of two explanatory variables in our model

Visually (without an interaction)



Visually (with an interaction)



Example

Migrant population and AfD vote share

We are going to continue on with the same data set, but now focusing on whether the relationship between `migrantfraction` (X_1) and `afd` (Y) is **different** for East and West districts (X_2)

Conditional Associations

- The simple model we have been studying assumes 'constant associations' (i.e. the relationship between X and Y does not depend on other X's)

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Conditional Associations

- The simple model we have been studying assumes 'constant associations' (i.e. the relationship between X and Y does not depend on other X's)

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i}$$

- We can relax the assumption of constant association by adding the product of explanatory variables to a model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \cdot X_{2i} + \varepsilon_i$$

Conditional Associations

- The simple model we have been studying assumes 'constant associations' (i.e. the relationship between X and Y does not depend on other X's)

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i}$$

- We can relax the assumption of constant association by adding the product of explanatory variables to a model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \cdot X_{2i} + \varepsilon_i$$

- In our example, this would correspond to the following model:

$$\text{AfD} = \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east}$$

Three models

```
m1 <- lm(AfD ~ migrantfraction, data = results)
m2 <- lm(AfD ~ migrantfraction + east, data = results)
m3 <- lm(AfD ~ migrantfraction * east, data = results)
screenreg(list(m1,m2,m3))
```

```
##
## ============================================================
##                              Model 1     Model 2     Model 3
## ------------------------------------------------------------
## (Intercept)                   18.36 ***   12.13 ***    9.69 ***
##                               (0.60)      (0.66)      (0.66)
## migrantfraction               -0.34 ***   -0.10 ***    0.02
##                               (0.03)      (0.03)      (0.03)
## eastTRUE                                   8.91 ***   14.44 ***
##                                           (0.67)      (0.88)
## migrantfraction:eastTRUE                              -0.54 ***
##                                                       (0.06)
## ------------------------------------------------------------
## R^2                            0.32        0.58        0.66
## Adj. R^2                       0.32        0.57        0.66
## Num. obs.                    299         299         299
```

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\text{AfD} = \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east}$$

$$\text{AfD} = 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east}$$

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

$$\text{AfD} = \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0$$

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0 \\ \text{AfD} &= \underbrace{9.69}_{Intercept} + \underbrace{0.02}_{Slope} * \text{migrant} \end{aligned}$$

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0 \\ \text{AfD} &= \underbrace{9.69}_{Intercept} + \underbrace{0.02}_{Slope} * \text{migrant} \end{aligned}$$

- What is the estimate for the East (i.e., `east`$= 1$)?

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0 \\ \text{AfD} &= \underbrace{9.69}_{Intercept} + \underbrace{0.02}_{Slope} * \text{migrant} \end{aligned}$$

- What is the estimate for the East (i.e., `east`$= 1$)?

$$\text{AfD} = \alpha + \beta_1 * \text{migrant} + \beta_2 * 1 + \beta_3 * \text{migrant} * 1$$

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0 \\ \text{AfD} &= \underbrace{9.69}_{Intercept} + \underbrace{0.02}_{Slope} * \text{migrant} \end{aligned}$$

- What is the estimate for the East (i.e., `east`$= 1$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 1 + \beta_3 * \text{migrant} * 1 \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 + -0.54 * \text{migrant} \end{aligned}$$

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0 \\ \text{AfD} &= \underbrace{9.69}_{Intercept} + \underbrace{0.02}_{Slope} * \text{migrant} \end{aligned}$$

- What is the estimate for the East (i.e., `east`$= 1$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 1 + \beta_3 * \text{migrant} * 1 \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 + -0.54 * \text{migrant} \\ \text{AfD} &= 9.69 + 14.44 + (0.02 - 0.54) * \text{migrant} \end{aligned}$$

Interaction: Continuous and Dummy

What is the association between `migrantfraction` and `AfD`?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * \text{east} + \beta_3 * \text{migrant} * \text{east} \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 * \text{east} + -0.54 * \text{migrant} * \text{east} \end{aligned}$$

- What is the estimate for the West (i.e., `east`$= 0$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 0 + \beta_3 * \text{migrant} * 0 \\ \text{AfD} &= \underbrace{9.69}_{Intercept} + \underbrace{0.02}_{Slope} * \text{migrant} \end{aligned}$$

- What is the estimate for the East (i.e., `east`$= 1$)?

$$\begin{aligned} \text{AfD} &= \alpha + \beta_1 * \text{migrant} + \beta_2 * 1 + \beta_3 * \text{migrant} * 1 \\ \text{AfD} &= 9.69 + 0.02 * \text{migrant} + 14.44 + -0.54 * \text{migrant} \\ \text{AfD} &= 9.69 + 14.44 + (0.02 - 0.54) * \text{migrant} \\ \text{AfD} &= \underbrace{24.13}_{Intercept} - \underbrace{(0.52)}_{Slope} * \text{migrant} \end{aligned}$$

Interaction: Continuous and Dummy

East	Intercept	Slope
0 = west	α	β_1
	9.69	0.02
1 = east	$\alpha + \beta_2$	$\beta_1 + \beta_3$
	9.69 + 14.44 = 24.13	0.02 + -0.54 = -0.52

- Implication: the relationship between migrants and AfD vote share is **different** in East and West districts.

Interaction: Continuous and Dummy

East	Intercept	Slope
0 = west	α	β_1
	9.69	0.02
1 = east	$\alpha + \beta_2$	$\beta_1 + \beta_3$
	9.69 + 14.44 = 24.13	0.02 + -0.54 = -0.52

β_1

- Partial association between X_1 and Y when X_2 is equal to 0 (holding other things constant)

$\beta_1 + \beta_3$

- Partial association between X_1 and Y when X_2 is equal to 1 (holding other things constant)

Interaction: Continuous and Dummy

East	Intercept	Slope
0 = west	α	β_1
	9.69	0.02
1 = east	$\alpha + \beta_2$	$\beta_1 + \beta_3$
	9.69 + 14.44 = 24.13	0.02 + -0.54 = -0.52

β_1

- Describes the relationship between percentage of migrants and AfD vote share, for districts in West Germany

$\beta_1 + \beta_3$

- Describes the relationship between percentage of migrants and AfD vote share, for districts in East Germany

Interaction: Continuous and Dummy

East	Intercept	Slope
0 = west	α	β_1
	9.69	0.02
1 = east	$\alpha + \beta_2$	$\beta_1 + \beta_3$
	9.69 + 14.44 = 24.13	0.02 + -0.54 = -0.52

β_1

- In the **West**, increasing by one point the percentage of migrants in a district is associated with a 0.02 percentage point increase in AfD vote share, on average.

$\beta_1 + \beta_3$

- In the **East**, increasing by one point the percentage of migrants in a district is associated with a 0.52 percentage point decrease in AfD vote share, on average.

Interpreting the Constituents of the Interaction

	Model 1
(Intercept)	9.69
	(0.66)
migrantfraction	0.02
	(0.03)
east	14.44
	(0.88)
migrantfraction:east	−0.54
	(0.06)
Adj. R^2	0.66
Num. obs.	299

- β_1 (coefficient of `migrantfraction`) is the association between vote share and migration when the dummy `east` is 0, i.e. in western districts

Interpreting the Constituents of the Interaction

	Model 1
(Intercept)	9.69
	(0.66)
migrantfraction	0.02
	(0.03)
east	14.44
	(0.88)
migrantfraction:east	−0.54
	(0.06)
Adj. R^2	0.66
Num. obs.	299

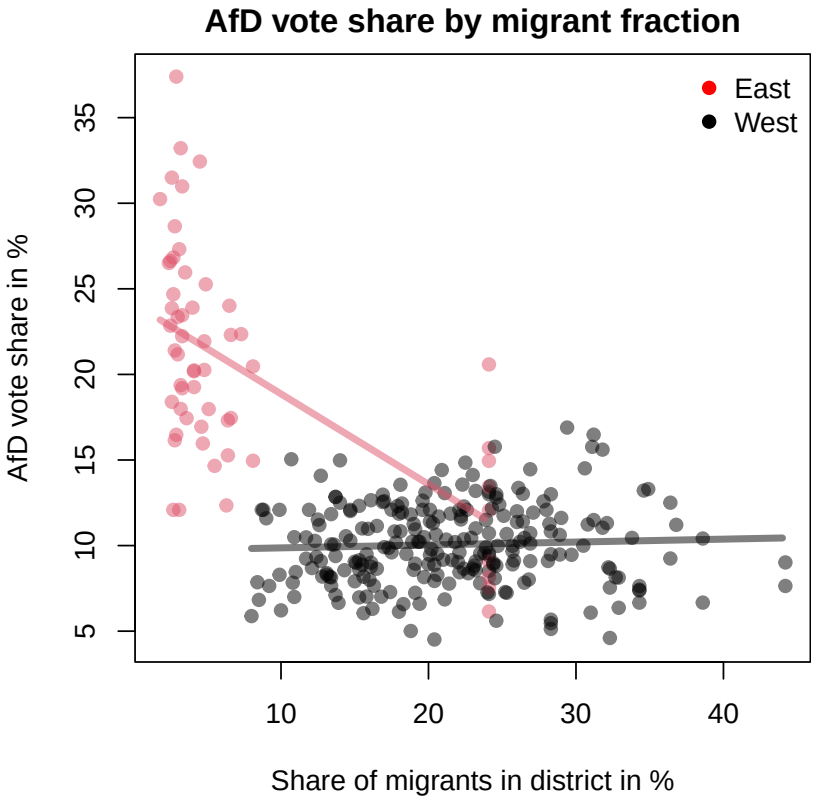
- β_1 (coefficient of `migrantfraction`) is the association between vote share and migration when the dummy `east` is 0, i.e. in western districts
- β_2 (coefficient of `east`) is the average difference between the east and the west when there are 0% migrants

Interpreting the Constituents of the Interaction

	Model 1
(Intercept)	9.69
	(0.66)
migrantfraction	0.02
	(0.03)
east	14.44
	(0.88)
migrantfraction:east	−0.54
	(0.06)
Adj. R^2	0.66
Num. obs.	299

- β_1 (coefficient of `migrantfraction`) is the association between vote share and migration when the dummy `east` is 0, i.e. in western districts
- β_2 (coefficient of `east`) is the average difference between the east and the west when there are 0% migrants
- β_3 (coefficient of the interaction) is the average difference in the association of `migrantfraction` and `AfD` between east and west

Interactions: Implication



- β_2 (the coefficient of `east`) is *not* the average difference between east and west
- β_1 (the coefficient of `migrantfraction`) is *not* an unconditional association of migration and AfD
- We do not have the general (unconditional) associations anymore

Multiple and Adjusted R^2

R^2 for the multiple regression model

R^2 is a useful general statistic:

- Simple linear regression
  - R^2 = proportion of the variance in Y explained by the model with variable X
- Multiple linear regression
  - R^2 = proportion of the variance in Y explained by the model with variables $X_1, ..., X_k$

Adjusting the R^2

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

- R^2 will almost always increase when we add a new X variable
- R^2 will never decrease when we add a new X variable

Adjusting the R^2

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

- R^2 will almost always increase when we add a new X variable
- R^2 will never decrease when we add a new X variable
- Implication:
  - Picking the model with the highest R^2 can be problematic
  - We need a measure that penalises using 'too many' X's

Adjusted R^2

$$adj.R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

where

- TSS (Total sum of squares) equals $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- SSR (Sum squared residuals) equals $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- k is the number of explanatory variables
- n is the number of observations

Adjusted R^2

$$adj.R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

Intuition

- adj. R^2 does not always increase when new X's are added
- adj. R^2 will always be smaller than R^2
- Interpretation is essentially the same as the 'normal' R^2:
  - $=$ proportion of the variance in Y explained by the model with variables $X_1, ..., X_k$, **adjusted** for the number of predictors

Adjusted R^2

$$adj.R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

- You don't need to know this formula! `R` will calculate it for you.

```
summary(m3)$r.squared
```

```
## [1] 0.6594207
```

```
summary(m3)$adj.r.squared
```

```
## [1] 0.6559571
```

Conclusion

What have we covered?

- Multiple regression
  - with categorical variables
  - with interactions of variables
- Multiple and adjusted R^2

We have not yet covered how to use regression to make/test causal claims, which is the subject of the next lecture.

Seminar

In seminars this week, you will learn about ...

1. Use of the `lm()` command to fit multiple linear regression models in R.
2. Use of the `screenreg()` and `htmlreg()` commands to compare differently specified multiple regression models.
3. Interpretation of categorical variables and interactions between variables.
4. R^2 and adjusted-R^2

There will be no homework assignment this week (Yay!),

Seminar

In seminars this week, you will learn about ...

1. Use of the `lm()` command to fit multiple linear regression models in R.
2. Use of the `screenreg()` and `htmlreg()` commands to compare differently specified multiple regression models.
3. Interpretation of categorical variables and interactions between variables.
4. R^2 and adjusted-R^2

There will be no homework assignment this week (Yay!), because you will have your midterm assessment (less Yay).