

# regression with categorical explanatory variables

poverty vs. region

explanatory variable: region

$$\widehat{poverty} = 11.17 + 0.38 \text{ region} : west$$

for eastern states  
plug in 0 for  $x$

$$\widehat{poverty} = 11.17 + 0.38 \times 0 = 11.17$$

reference level

for western states  
plug in 1 for  $x$

$$\widehat{poverty} = 11.17 + 0.38 \times 1 = 11.55$$



## slope and intercept

$$\widehat{poverty} = 11.17 + 0.38 \text{ region : west}$$

- ▶ intercept: The model predicts an 11.17% average poverty percentage in eastern states.
  - ▶ This is the value we get if we plug in 0 for the explanatory variable
- ▶ Slope: The model predicts that the average poverty percentage in western states is 0.38% higher than in the eastern states.

Next, we use a new region variable (**region4**) with four levels: northeast, midwest, west, south. Write the linear regression model based on the regression output below.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

$$\widehat{\% \text{ in poverty}} = 9.50 + 0.03 \text{ reg4:mw} + 1.79 \text{ reg4:w} + 4.16 \text{ reg4:s}$$



What is the reference level of the **region4** variable:  
 northeast, midwest, west, south?

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

Calculate the predicted poverty rate for western states.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

$$\begin{aligned}\widehat{\% \text{ in poverty}} &= 9.50 + 0.03 \cancel{\text{reg4:mw}} + 1.79 \cancel{\text{reg4:w}} + 4.16 \cancel{\text{reg4:s}} \\ &= 9.50 + 0 + 1.79 + 0 \\ &= 11.29\end{aligned}$$