Regression

Expert Testimony

Predicting a numerical outcome variable

- What salary should I expect, *given my relevant characteristics*?
- What demand for transformers should I expect, *given a price I set*?
- What price should I expect for a hotel, *given its location and rating*?

Recipe

Given X, expect Y.

Predictions are never exact

- Cannot make perfectly accurate prediction $\rightarrow$ learn to live with *error*.
- Capture the *mean* of the outcome.
- Minimize prediction error.

Our first prediction

I expect the mean salary, mean demand, mean price.

$$\hat{Y}_{\text{mine}} = E(Y_{\text{all}})$$

"regression to the mean" $\rightarrow$ **regression**

This does not hold relevant characteristics fixed.

Our second prediction

Take the mean of Y for the group of cases with exactly the same X as mine.

$$\hat{Y}_{\text{mine}} = E(Y) \text{ for cases where } X = X_{\text{mine}}$$

This is called **conditional mean**, or mean of Y **conditional on** X:

$$\hat{Y}_{\text{mine}} = E(Y|X = X_{\text{mine}})$$

Can hold multiple things fixed

$$\widehat{\text{wage}}_{\text{mine}} = E(\text{wage}|\text{occupation} = \text{economist}, \text{eye color} = \text{blue})$$

How to compute this?

If all Xs are categorical:

1. select comparison group with exact same Xs
2. compute sample mean within this group

(Pivot table, AVG() ... GROUP BY ...)

What price for a 3-star hotel in Favoriten?

Limits of pivot tables

Often no or few exact matches with $X = X_{\text{mine}} \rightarrow$ noisy prediction.

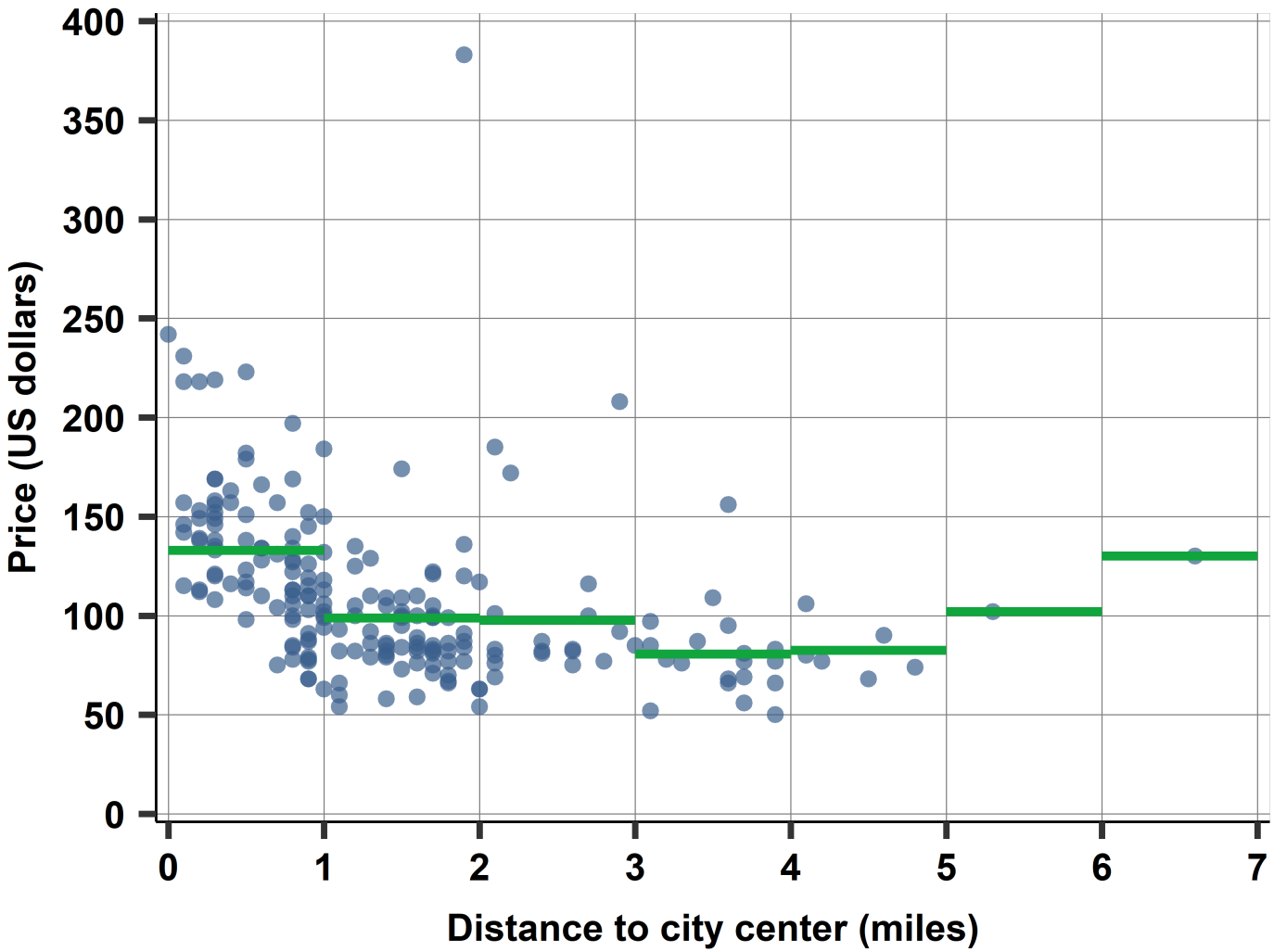
Particularly if X is a numerical variable.

What price for a 3-star hotel in Favoriten with a rating of 3.7?

Two solutions

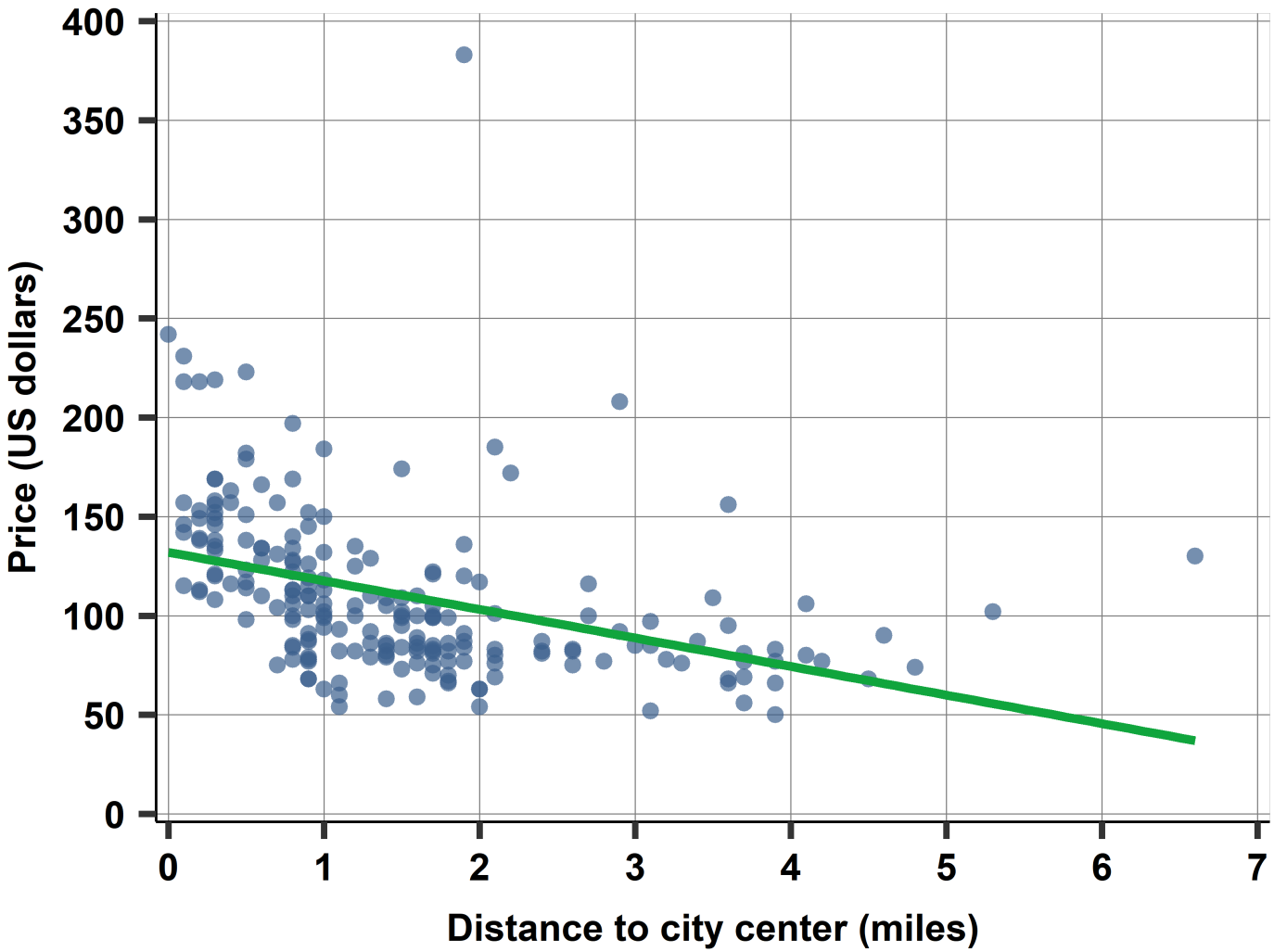
1. Merge some groups
   - Which ones? Machine Learning: regression tree, random forest
2. Interpolate between data points
   - Assume a relationship with a simple mathematical functional form

Split distance into bins



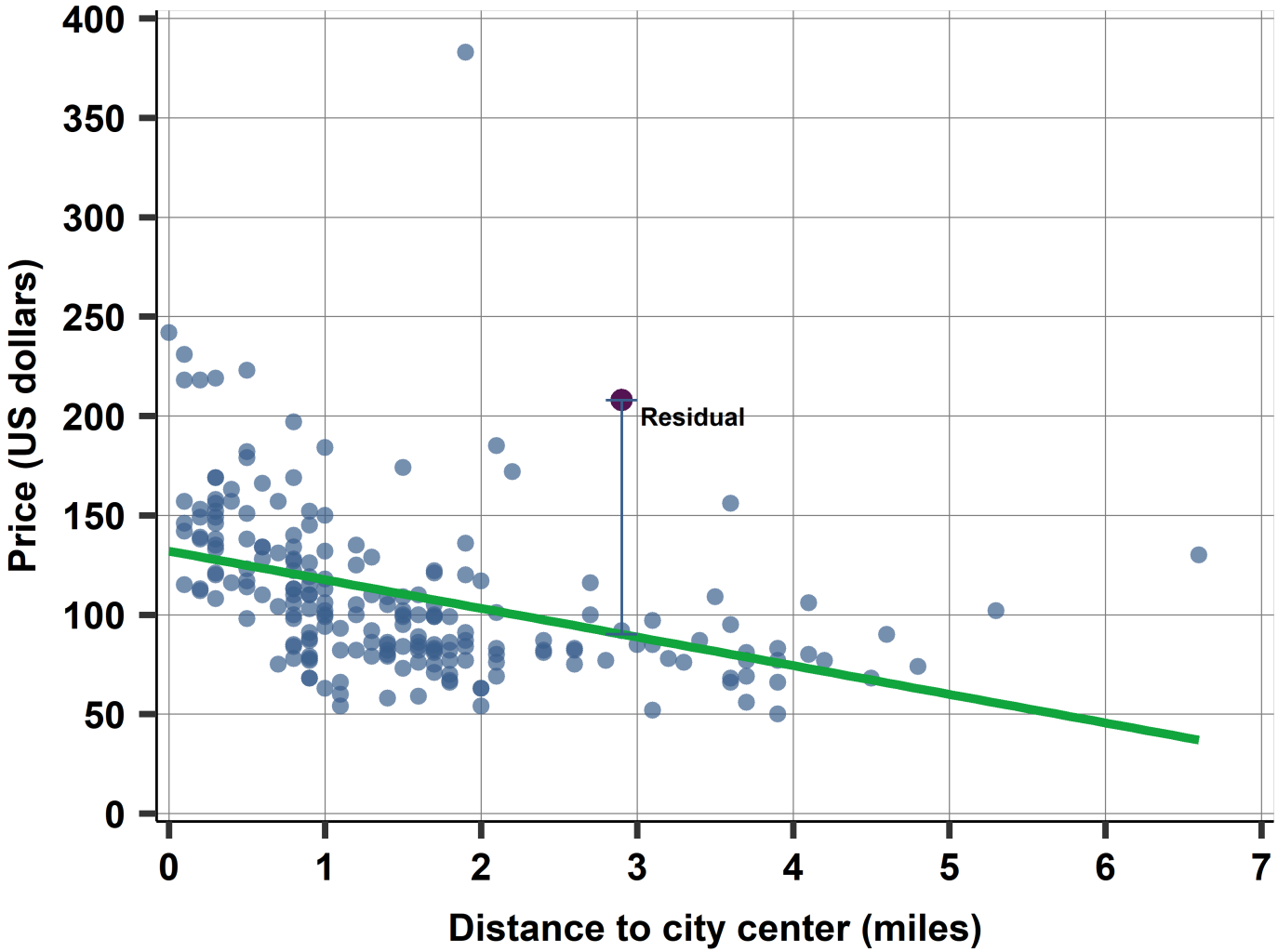
Békés and Kézdi (2021, Chapter 07)

Fit a line on the scatter plot



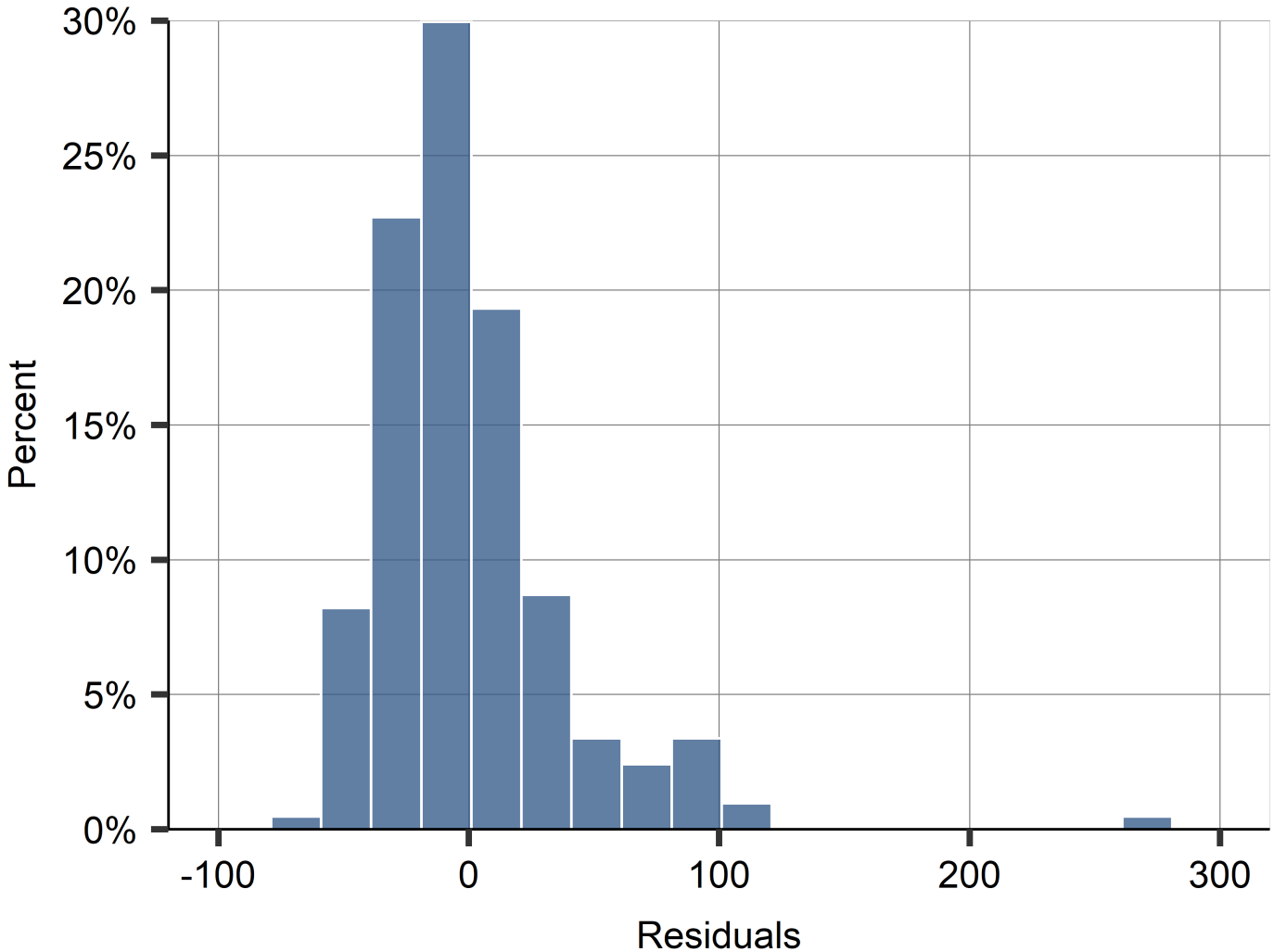
Békés and Kézdi (2021, Chapter 07)

Try to minimize residuals



Békés and Kézdi (2021, Chapter 07)

Unbiased but dispersed residuals



Békés and Kézdi (2021, Chapter 07)

Linear regression

Assume linear relationship:

$$E(Y|X) = a + bX$$

Need to estimate a and b.

But can use even for X for which we have no data (yet), like rating=3.7.

Interpret coefficients

$$E(Y|X) = a + bX$$

a mean value of Y when $X = 0$:

$$a + b \cdot 0 = a$$

b difference in mean value of Y when X *increases by 1 unit*:

$$[a + b(x + 1)] - [a + bx] = b(x + 1) - bx = b$$

Interpret regression of hotel price on distance

```
. regress price distance
```

Source	SS	df	MS		
				Number of obs	= 428
				F(1, 426)	= 19.51
Model	156858.055	1	156858.055	Prob > F	= 0.0000
Residual	3424389.35	426	8038.47266	R-squared	= 0.0438
				Adj R-squared	= 0.0416
Total	3581247.41	427	8386.99627	Root MSE	= 89.658

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	-12.01145	2.719123	-4.42	0.000	-17.35602	-6.666883
_cons	151.2924	6.255225	24.19	0.000	138.9974	163.5873

What is the expected price 2.2 miles from the center?

$$E(\text{price}|\text{distance}) = 151.29 - 12.01 \cdot \text{distance}$$

Can hold multiple things fixed

Multiple linear regression

$$E(Y|X,Z) = a + bX + cZ$$

Interpret coefficients

$$E(Y|X,Z) = a + bX + cZ$$

a mean value of Y when **both** $X = 0$ and $Z = 0$:

$$a + b \cdot 0 + c \cdot 0 = a$$

b difference in mean value of Y when X increases by 1 unit, **holding** Z **fixed**:

$$[a + b(x+1) + cz] - [a + bx + cz] = b(x+1) - bx = b$$

c difference in mean value of Y when Z increases by 1 unit, **holding** X **fixed**:

$$[a + bx + c(z+1)] - [a + bx + cz] = c(z+1) - cz = c$$

Hotel prices depend both on distance and rating

```
. regress price distance rating
```

Source	SS	df	MS		
				Number of obs	= 393
				F(2, 390)	= 13.97
Model	211966.86	2	105983.43	Prob > F	= 0.0000
Residual	2959529.45	390	7588.53705	R-squared	= 0.0668
				Adj R-squared	= 0.0620
Total	3171496.31	392	8090.55181	Root MSE	= 87.112

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	-9.027104	2.764378	-3.27	0.001	-14.46205	-3.592156
rating	26.42263	7.786677	3.39	0.001	11.11351	41.73174
_cons	38.14768	32.52452	1.17	0.242	-25.79765	102.093

What is the expected price 2.2 miles from the center with a 3.7 rating?

$$E(\text{price}|\text{distance}) = 38.14 - 9.03 \cdot \text{distance} + 26.42 \cdot \text{rating}$$

Multiplicative models

Multiplicative models

So far we studied

$$E(Y|X) = a + bX,$$

but often we want

$$E(Y|X) = aX^b.$$

Why? Because we may be interested in percentage or proportional changes, not changes by one unit.

$$\frac{E(Y|X = 2x)}{E(Y|X = x)} = \frac{a(2x)^b}{ax^b} = 2^b$$

Interpreting coefficients

Suppose

$$E(\text{wage}|\text{hours}) = 100 \times \text{hours}^{0.5}.$$

What is the effect of doubling working hours?

$$\frac{100 \times (2h)^{0.5}}{100 \times h^{0.5}} = 2^{0.5} \approx 1.41,$$

wages increase by 41 percent.

Categorical variables

Categorical variables

What X is *categorical*? Say, taking the values `male` and `female`.

One-hot encoding ("dummy variables") to the rescue.

$$\text{FEMALE}_i = \begin{cases} 1 & \text{if } X_i = \text{female} \\ 0 & \text{if } X_i = \text{male} \end{cases}$$

Similarly,

$$\text{MALE}_i = \begin{cases} 0 & \text{if } X_i = \text{female} \\ 1 & \text{if } X_i = \text{male} \end{cases}$$

One-hot encoded variables in a regression

Suppose

$$E(\text{WAGE}_i|X_i) = \begin{cases} 80 & \text{if } X_i = \text{female} \\ 100 & \text{if } X_i = \text{male} \end{cases}$$

This can be written as

$$E(\text{WAGE}_i|X_i) = 100 - 20 \times \text{FEMALE}_i,$$

which is a *linear regression*!

Huh?!

$$E(\text{WAGE}_i|X_i) = 100 - 20 \times \text{FEMALE}_i$$

$$= \begin{cases} 100 - 20 \times 1 & \text{if } X_i = \text{female} \\ 100 - 20 \times 0 & \text{if } X_i = \text{male} \end{cases}$$

$$= \begin{cases} 80 & \text{if } X_i = \text{female} \\ 100 & \text{if } X_i = \text{male} \end{cases}$$

Interpreting the effect of gender

$$E(\text{WAGE}_i|X_i) = 100 - 20 \times \text{FEMALE}_i,$$

is equal to

$$E(\text{WAGE}_i|X_i) = 80 + 20 \times \text{MALE}_i$$

Females earn 20 dollars *less than males* = Males earn 20 dollars *more than females*.

Question: why not enter both MALE and FEMALE?

Multiple categories

Suppose X captures the area of law.

$$E(\text{WAGE}_i|X_i) = \begin{cases} 80 & \text{if } X_i = \text{tax} \\ 100 & \text{if } X_i = \text{M\&A} \\ 95 & \text{if } X_i = \text{IP} \end{cases}$$

How to one-hot encode this? We can encode two of them, say `tax` and `IP`.

$$E(\text{WAGE}_i|X_i) = 100 - 20 \times \text{tax}_i - 5 \times \text{IP}_i$$

Questions: Why not include all three? Will expected wage take the value of 75?