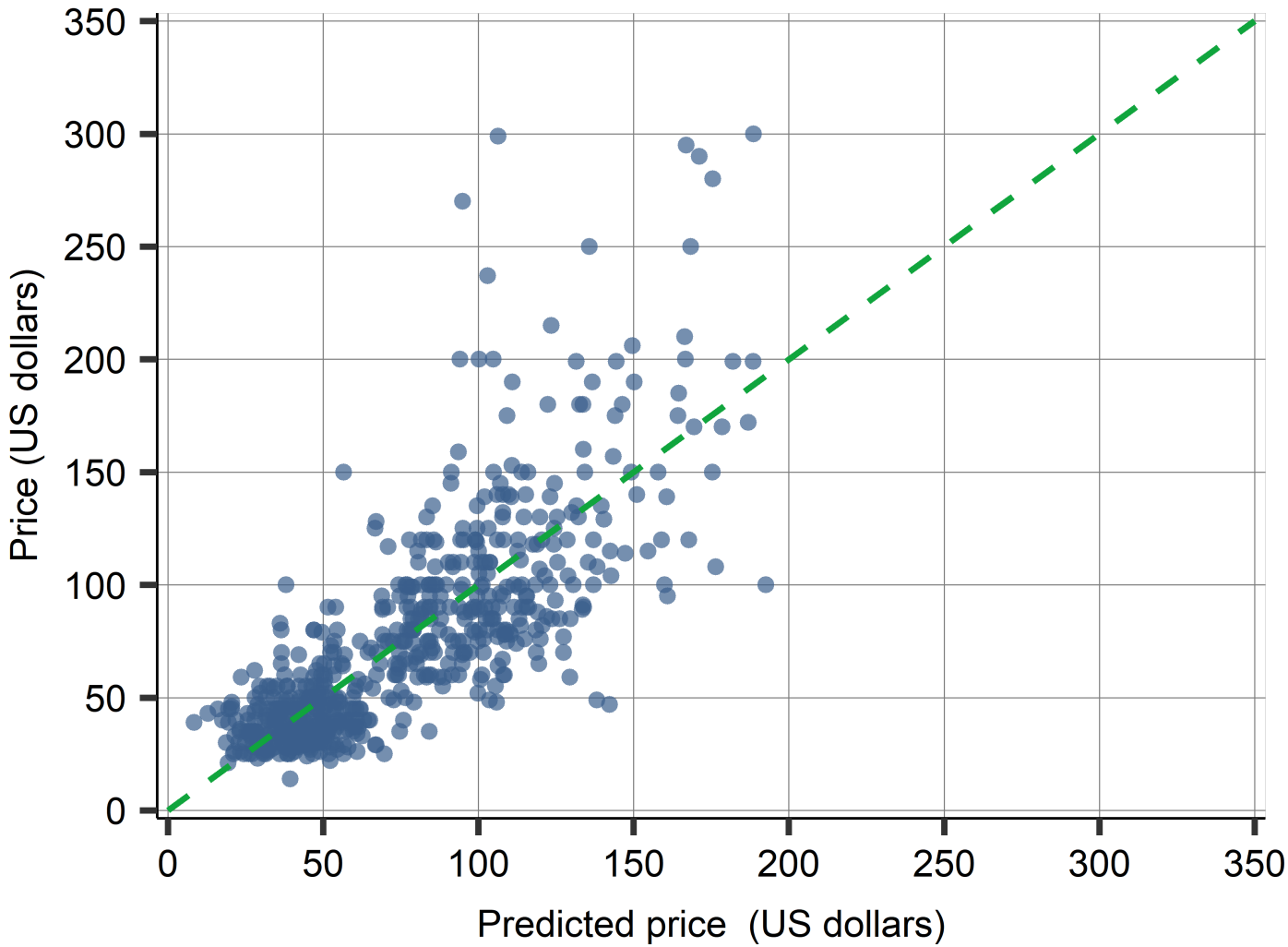
Goodness of fit

Goodness of fit

- Take a known variable Y_i and its predicted value $\hat{Y}_i$.
- Ideally, $\hat{Y}_i \approx Y_i$.
- If not, we need a measure of distance.
- Two cases:
  1. Y numeric
  2. Y categorical (not discussed)

AirBnB predicted prices for London, March 2017



Békés and Kézdi (2021, Chapter 14)

Root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Properties

- Only for quantitative variables.
- Standard deviation of prediction error (if unbiased).
- Only zero for perfect fit, positive otherwise.
- Symmetric in errors.
- Has same units as Y.
- Predictive error band:

$$Y_i \approx \hat{Y}_i \pm 2 \times \text{RMSE}$$

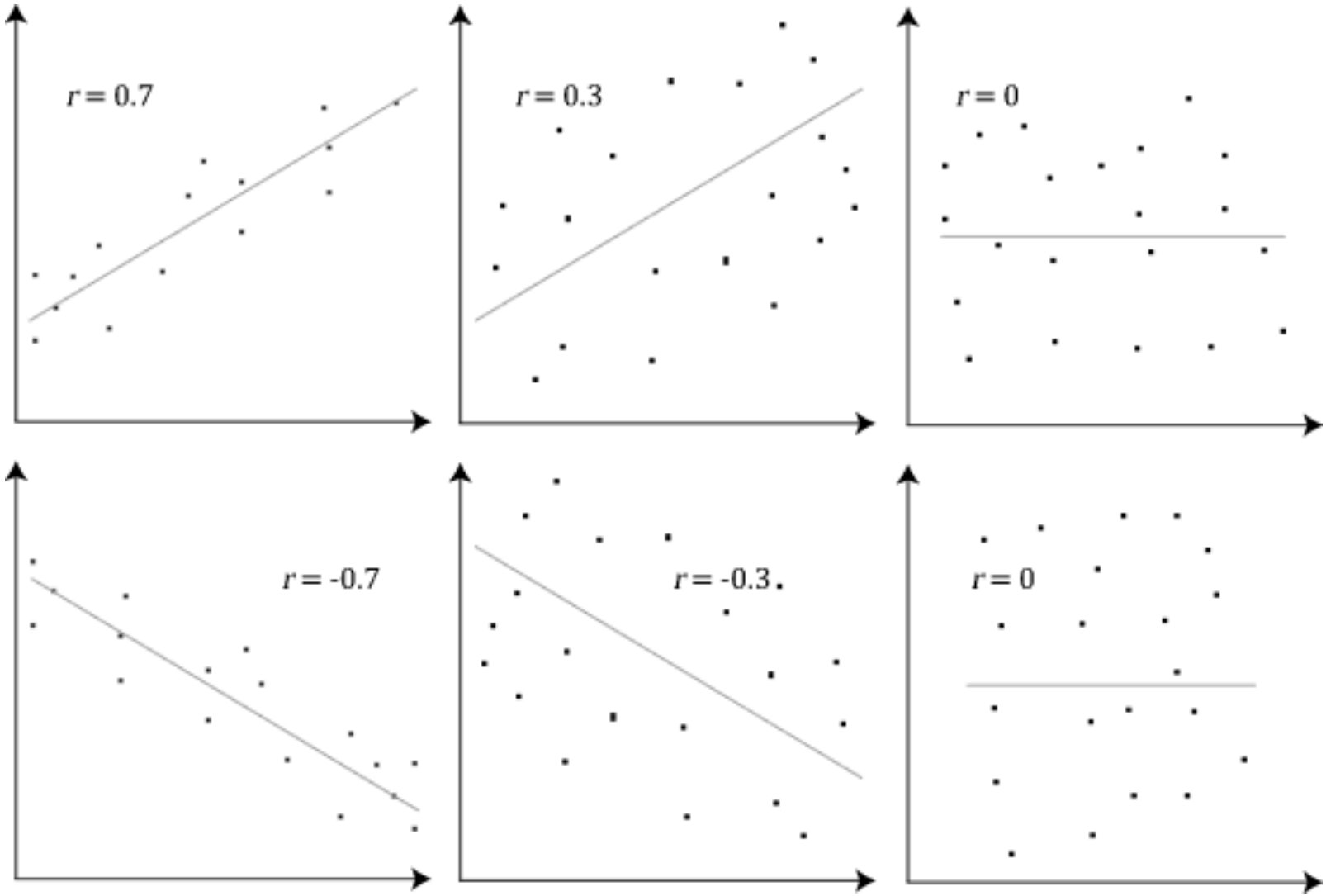
Correlation coefficient

$$\rho(Y, \hat{Y}) = \frac{\sum_{i=1}^{n}(Y_i - \mu_Y)(\hat{Y}_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(Y_i - \mu_Y)^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \mu_Y)^2}}$$

Properties

- Only for quantitative variables.
- Bounded between -1 and $+1$.
- Unitless.

Different degrees of Correlation



Laerd Statistics(2019)

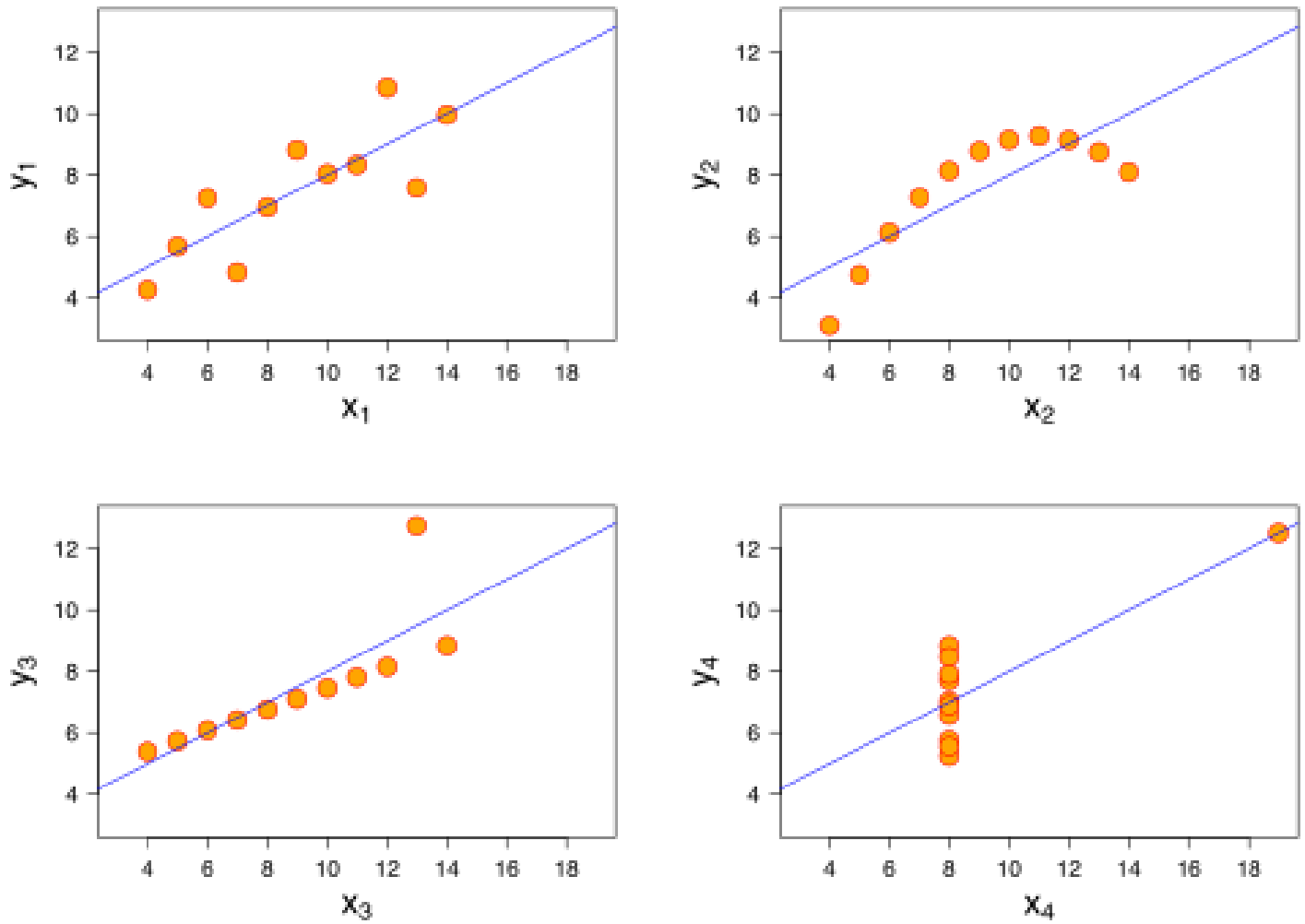
R-squared

$$R^2 = \rho(Y, \hat{Y})^2$$

Properties

- Only for quantitative variables.
- Bounded between 0 and 1.
- What fraction of the variation in the outcome does our prediction explain?
- Unitless.

Beware, all of these have $R^2 = 0.66$



Anscombe(1973)

Out-of-sample model evaluation

In-sample model evaluation

A trained model is a function of the training data. For example, an OLS estimator:

$$\hat{\beta}(X_{\text{train}}, Y_{\text{train}}) := \frac{\sum_{i \in \text{train}} (X_i - \bar{X}_{\text{train}})(Y_i - \bar{Y}_{\text{train}})}{\sum_{i \in \text{train}} (X_i - \bar{X}_{\text{train}})^2}$$

So far we evaluated the model on the same data it was trained on.

$$\hat{Y}_{\text{train}} = X_{\text{train}} \cdot \hat{\beta}(X_{\text{train}}, Y_{\text{train}})$$

The model is **trained to fit well** on this data.

Out-of-sample model evaluation

We need to evaluate the model on **data it has not seen**.

$$\hat{Y}_{\text{test}} = X_{\text{test}} \cdot \hat{\beta}(X_{\text{train}}, Y_{\text{train}})$$

All goodness of fit measures can be calculated on the test data.

Mean Squared Error

$$\text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} (Y_i - \hat{Y}_i)^2 = \underbrace{(\bar{Y}_{\text{test}} - \bar{\hat{Y}}_{\text{test}})^2}_{\text{bias}} + \underbrace{\frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} (\hat{Y}_i - \bar{\hat{Y}}_{\text{test}})^2}_{\text{variance}}$$

Typically larger.

R^2 for out-of-sample

$$R^2 = 1 - \frac{\sum_{i \in \text{test}} (Y_i - \hat{Y}_i)^2}{\sum_{i \in \text{test}} (Y_i - \bar{Y}_{\text{test}})^2}$$

Typically smaller.

May even be negative.