Data Science Bootcamp, 9th January 2016

Linear Regression

James McInerney

Columbia University
Data Science Institute

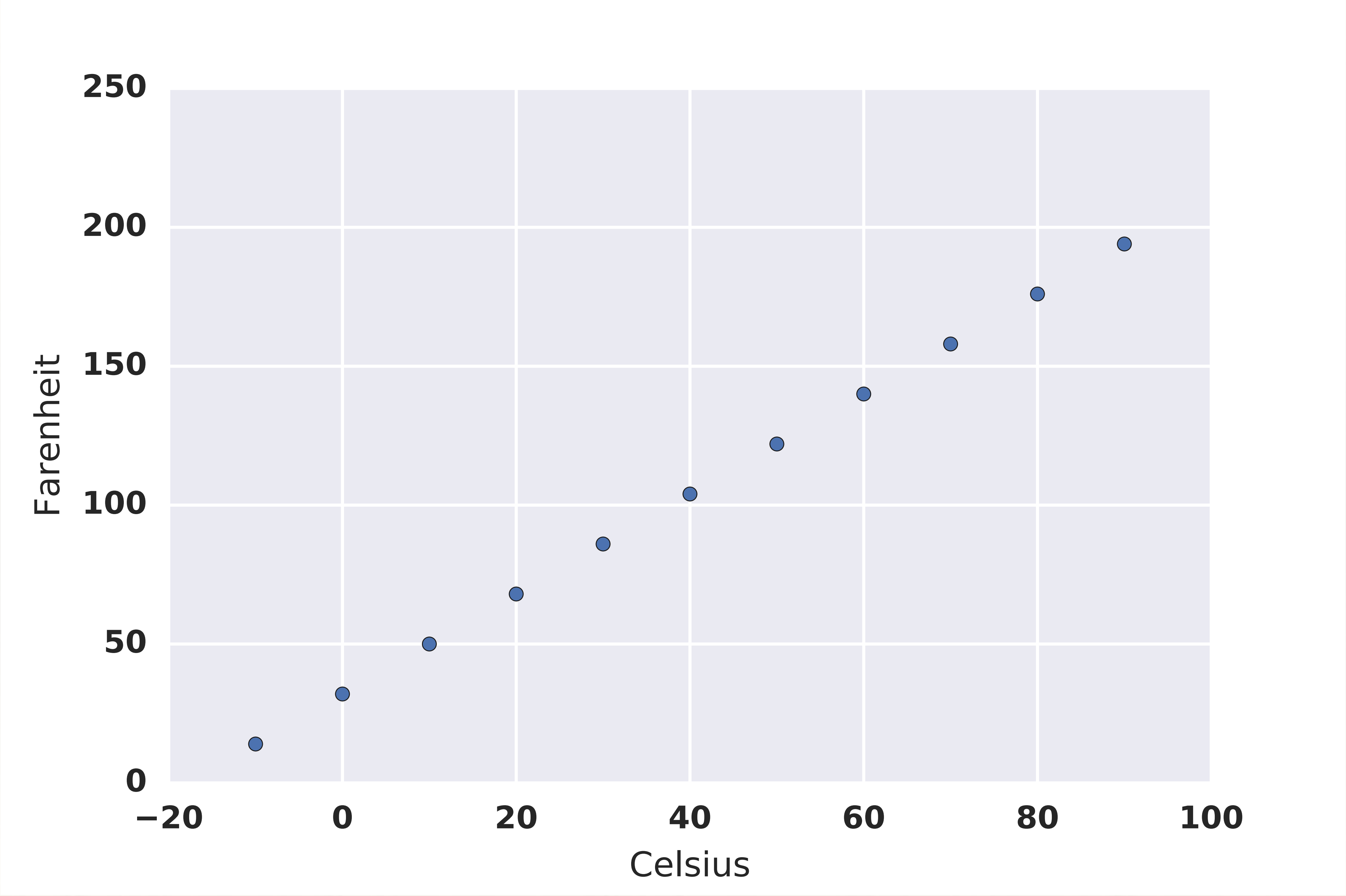
Regression Problem

- Goal: estimate the relationship between variables.
- Dependent variable: the value to be predicted.
- Independent variable(s): the inputs.
- Useful for prediction (of dependent variable), summarization (especially in higher dimensions), and gaining insight into a domain.

Examples

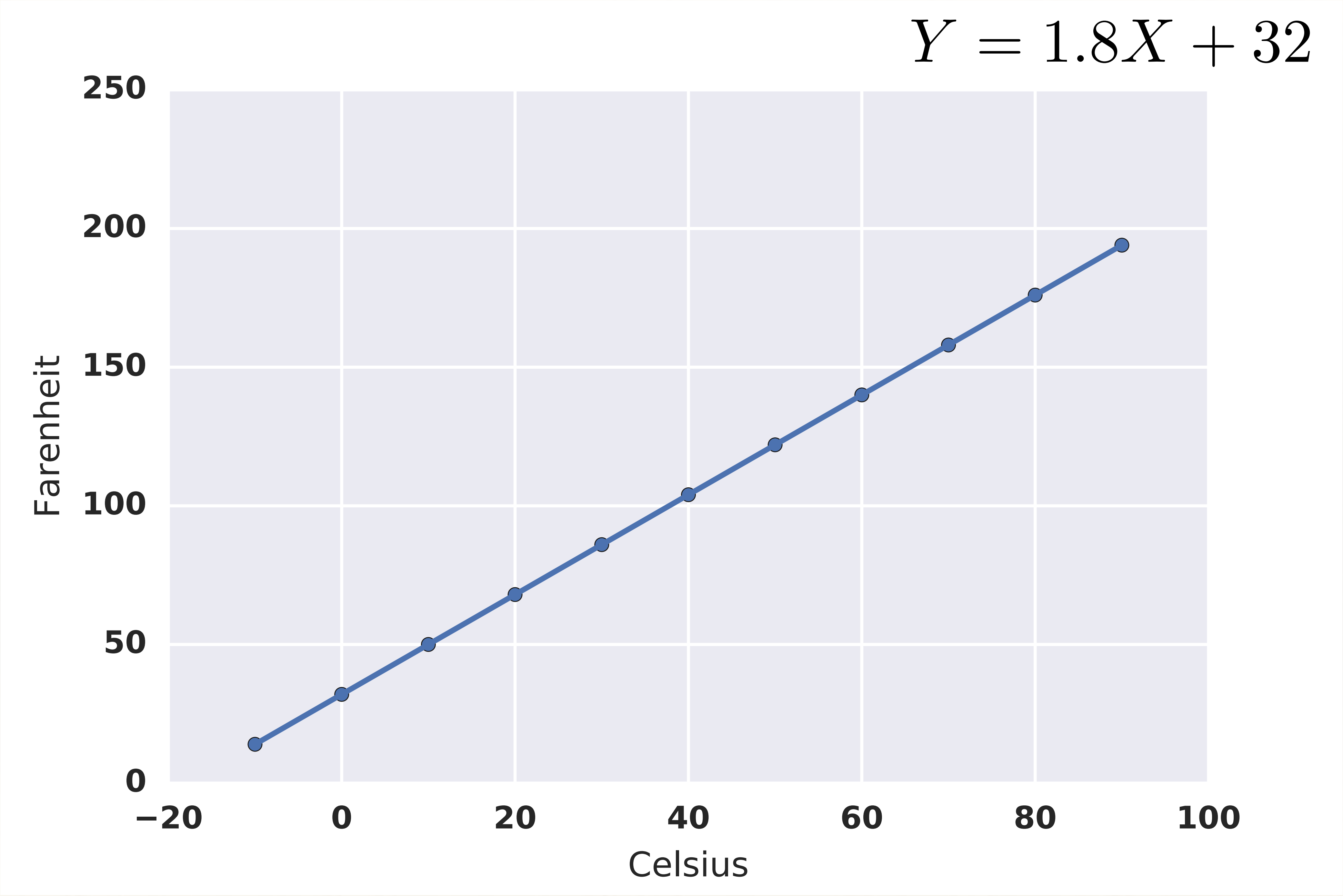
Problem	Dependent Variable	Independent Variable(s)
forecast sales estimates	sales	time
analyzing effect of medication	hours of effect	dosage (mg)
property valuation	house price	number of rooms, district, proximity to amenities

Deterministic Relationship

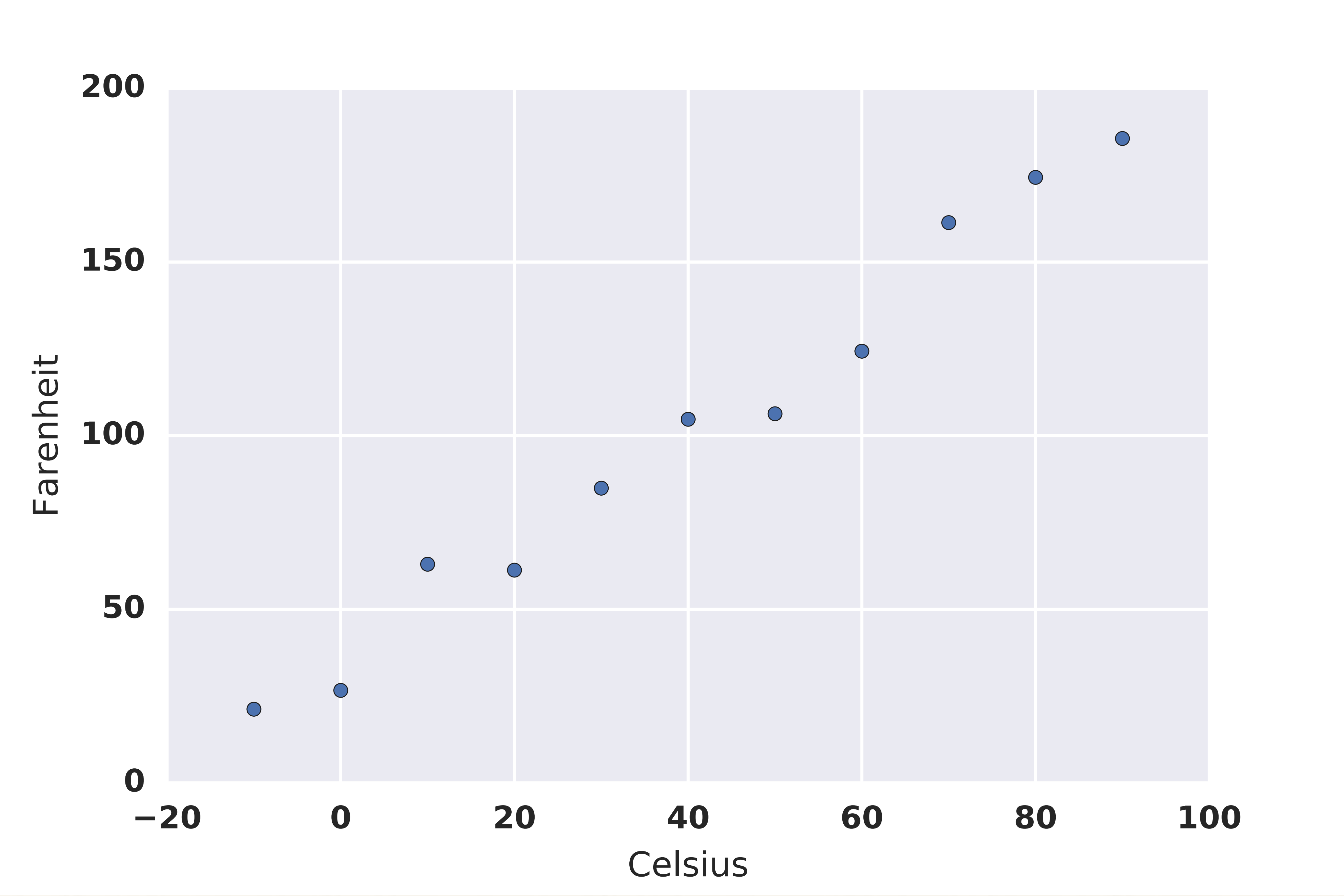


COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

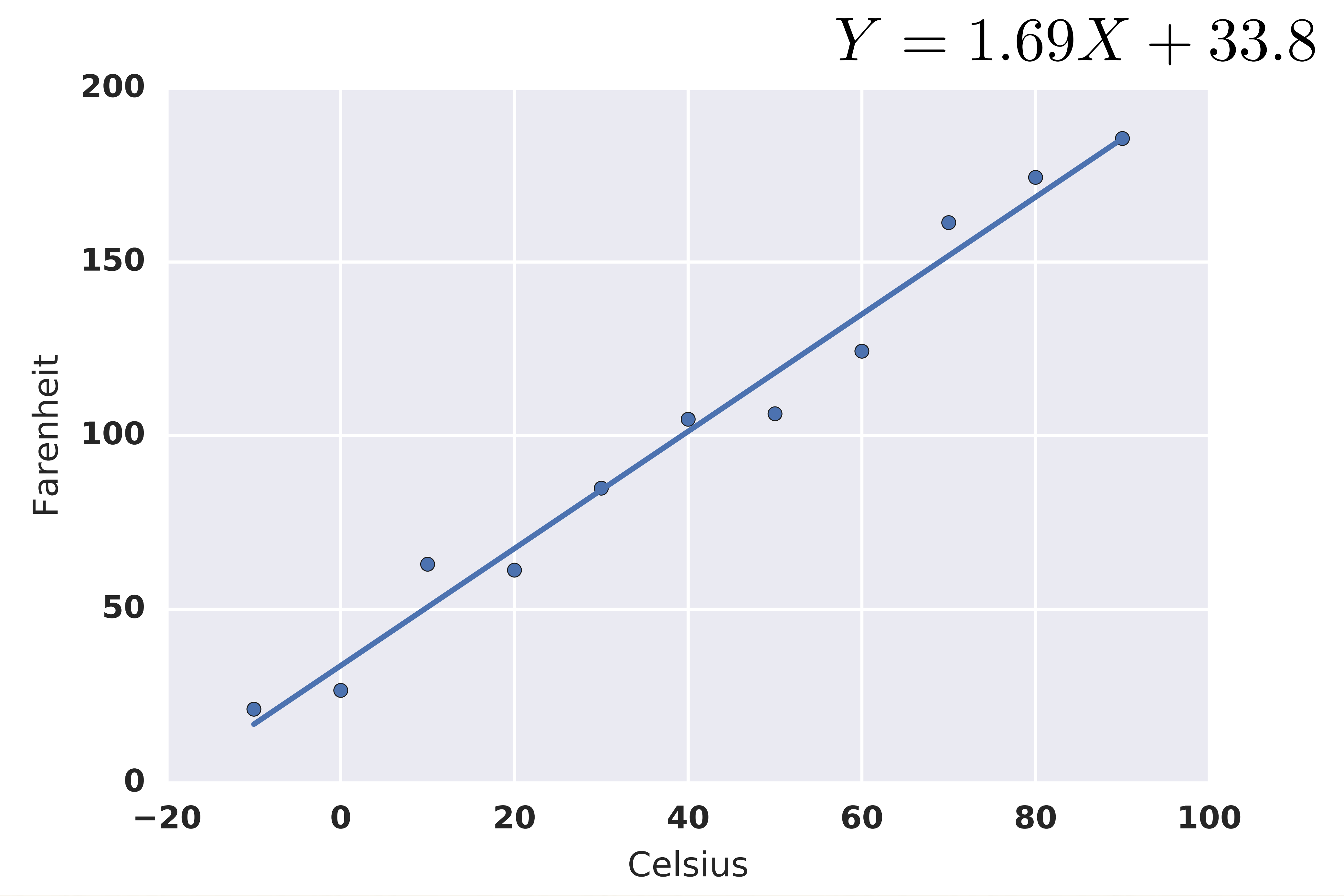
Deterministic Relationship



Stochastic Relationship



Stochastic Relationship



Regression and Linear Regression

- Independent variable $\mathbf{X}$
- Dependent variable Y
- Unknown coefficients β

Regression and Linear Regression

- Independent variable $\mathbf{X}$
- Dependent variable Y
- Unknown coefficients β

- Stochastic relationship $\mathbb{E}[Y|\mathbf{X}] = f(\mathbf{X}, \beta)$

Regression and Linear Regression

- Independent variable $\mathbf{X}$
- Dependent variable Y
- Unknown coefficients β

- Stochastic relationship $\mathbb{E}[Y|\mathbf{X}] = f(\mathbf{X}, \beta)$

- In **linear regression** $f(\mathbf{X}, \beta) := \mathbf{X}^\top \beta = \beta^{(0)} + \sum_{d=1}^{D} \beta^{(d)} X^{(d)}$

  where $X^{(0)} := 1$

Regression Error

- If we want an equality instead of an average, we need the concept of statistical error:

$$
\begin{aligned}
Y &= f(\mathbf{X}, \beta) + \epsilon \\
&= \mathbf{X}^\top \beta + \epsilon \quad \text{(linear regression)}
\end{aligned}
$$

Regression Error

- If we want an equality instead of an average, we need the concept of statistical error:

$$\begin{aligned} Y &= f(\mathbf{X}, \beta) + \epsilon \\ &= \mathbf{X}^\top \beta + \epsilon \quad \text{(linear regression)} \end{aligned}$$

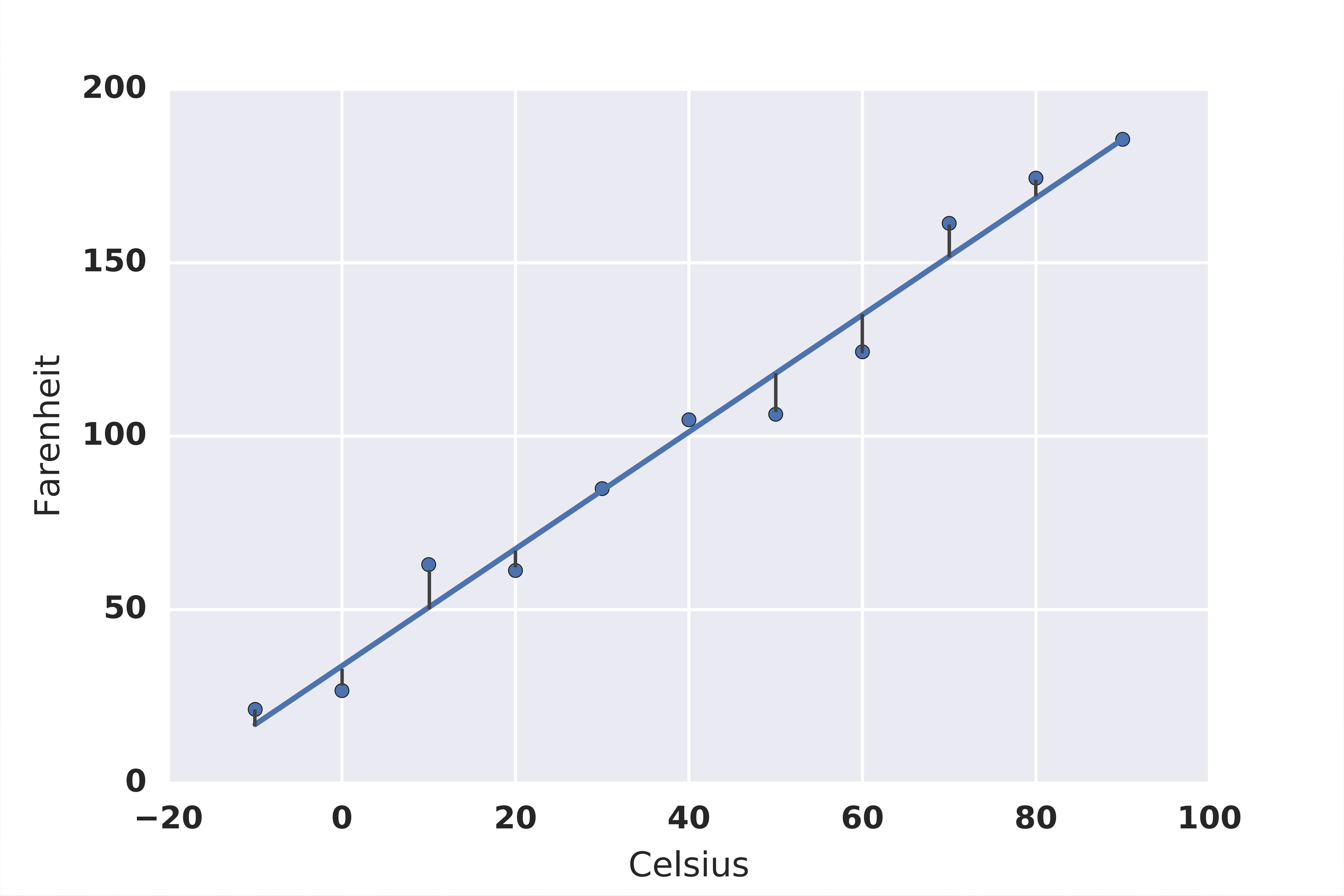
- Least-squares regression minimizes $\displaystyle \sum_{n=1}^{N} \epsilon_n^2$

$$= \sum_{n=1} (Y_n - \mathbf{X}_n^\top \beta)^2$$

What is the Error?

- Uncertainty from randomness in the data generating process.
- Epistemic uncertainty: "chance is a fool's name for fate" (Fred Astaire in the movie *Gay Divorcee*).
- —> in linear regression, and statistical models in general, the sum of all these errors has certain characteristics (e.g., central limit theorem).

Regression Error



Multiple Linear Regression

- Linear regression is not limited to being linear in $\mathbf{X}$.
- We are free to add various **basis functions** that allow us to capture non-linear relationships between the input and output while still using an inner product.
- For example, use first and second-order polynomials:

$$\mathbb{E}[Y|\hat{\mathbf{X}}] = \hat{\mathbf{X}}^\top \beta$$

where $\hat{\mathbf{X}} := \{\mathbf{X} \;\; \mathbf{X}^2\}$

Why is Linear Regression so Popular?

- Easy to interpret.
- Limited degrees of freedom helps avoid overfitting.
- Unique solution (<==> objective function is convex).
- Fast.

Why Other Approaches are Needed

- Some problems require a model with more degrees of freedom or break the linearity assumptions;
  - —> can be addressed to some extent with clever feature engineering, but this is not plug and play.
- When the assumptions of the next slide don’t hold.

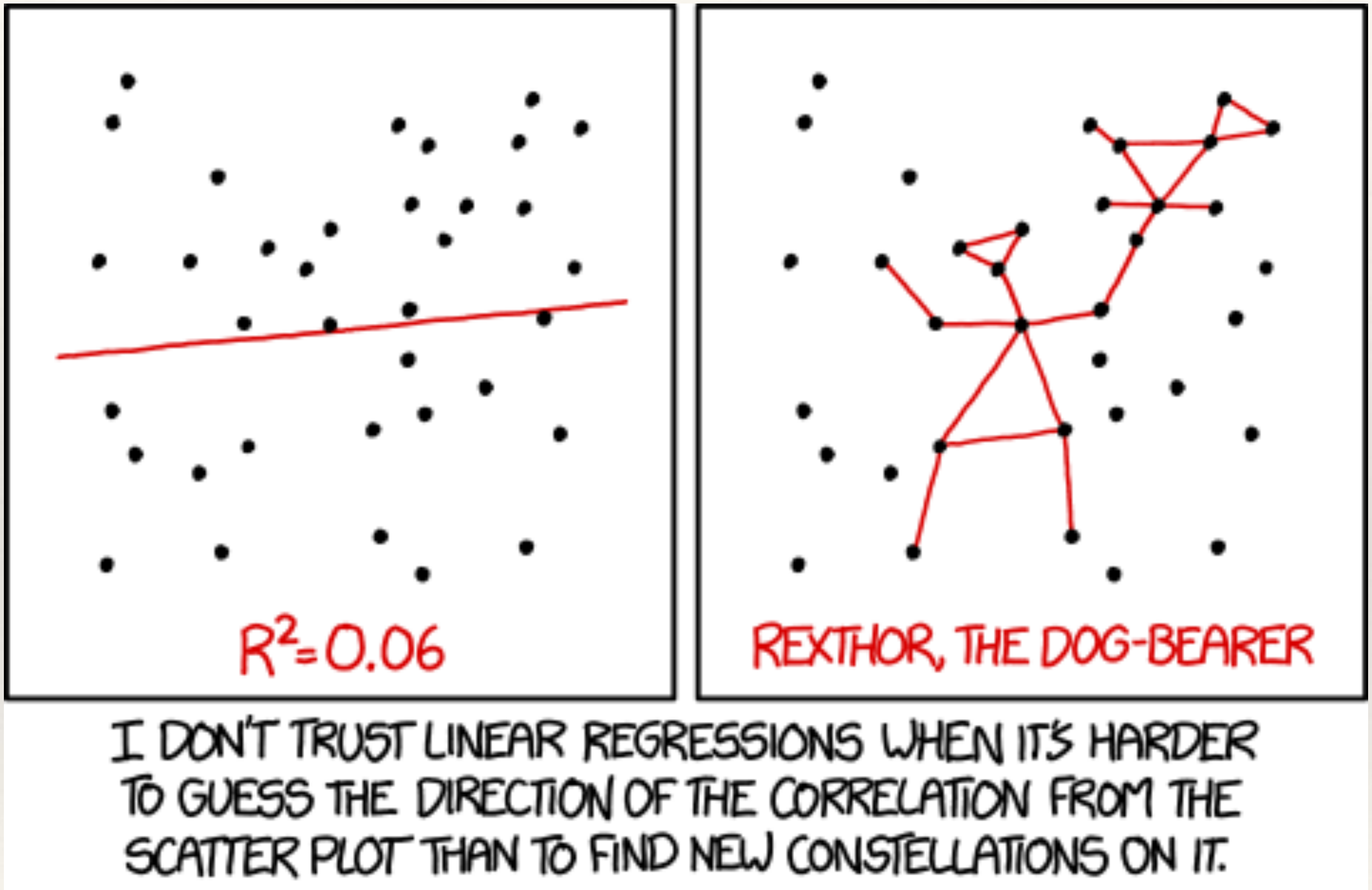
Key Assumptions

- Expected value of dependent variable is an affine function of the independent variables.
- Errors/residuals are independently and identically distributed (i.i.d.) from a normal distribution:
  - independent errors;
  - error drawn from the same distribution;
  - with the same mean and variance.
- No statistical error in the independent variables.
- Independent variables are linearly independent.

Today's Lab Session

- Fit linear regression to data simulated from a cubic function using *ordinary least squares* regression.
- Calculate coefficient of determination R^2 and plot residuals to look for model mismatch.
- Extend the set of *basis functions* to get a better fit with least squares.

Coefficient of Determination



http://xkcd.com/1725/

Conclusions

- Data usually exhibits stochastic relationships; we use probabilistic models to characterize this.
- Linear regression is a simple linear model of the data that is nonetheless unusually effective.