Why linear models?

Roger Levy

Massachusetts Institute of Technology

March 10, 2025

Why linear models?

For modeling conditional distributions $P(Y|\mathrm{X})$, we get to linear regression from the following assumptions:

1. The expected response $E[Y|\mathrm{X}]$ is a linear combination of the predictors:

$$\widehat{y} = \beta \cdot \mathrm{x}$$

Why linear models?

For modeling conditional distributions $P(Y|\mathrm{X})$, we get to linear regression from the following assumptions:

1. The expected response $E[Y|\mathrm{X}]$ is a linear combination of the predictors:

$$\widehat{y} = \beta \cdot \mathrm{x}$$

2. Residual error, or noise, is distributed normally with mean zero around the expected response:

$$y \sim \mathcal{N}(\widehat{y}, \sigma^2)$$

Why linear models?

For modeling conditional distributions $P(Y|\mathrm{X})$, we get to linear regression from the following assumptions:

1. The expected response $E[Y|\mathrm{X}]$ is a linear combination of the predictors:

$$\widehat{y} = \beta \cdot \mathrm{x}$$

2. Residual error, or noise, is distributed normally with mean zero around the expected response:

$$y \sim \mathcal{N}(\widehat{y}, \sigma^2)$$

Why linear models?

For modeling conditional distributions $P(Y|\mathrm{X})$, we get to linear regression from the following assumptions:

1. The expected response $E[Y|\mathrm{X}]$ is a linear combination of the predictors:

$$\widehat{y} = \beta \cdot \mathrm{x}$$

2. Residual error, or noise, is distributed normally with mean zero around the expected response:

$$y \sim \mathcal{N}(\widehat{y}, \sigma^2)$$

But, why linear regression?

Linear effect of predictors

$$\boxed{\widehat{y} = \beta \cdot \mathsf{x}}$$

XXX

Normally distributed residual error

$$\boxed{y \sim \mathcal{N}(\widehat{y}, \sigma^2)}$$

Three major reasons why this is a good assumption to use:

Normally distributed residual error

$$\boxed{y \sim \mathcal{N}(\widehat{y}, \sigma^2)}$$

Three major reasons why this is a good assumption to use:

- It has good computational properties

Normally distributed residual error

$$\boxed{y \sim \mathcal{N}(\widehat{y}, \sigma^2)}$$

Three major reasons why this is a good assumption to use:

- It has good computational properties
- There are *a priori* mathematical reasons to expect residual error to often look Gaussian(-ish)

Normally distributed residual error

$$\boxed{y \sim \mathcal{N}(\widehat{y}, \sigma^2)}$$

Three major reasons why this is a good assumption to use:

- It has good computational properties
- There are *a priori* mathematical reasons to expect residual error to often look Gaussian(-ish)
- It is robust to mild-to-moderate violations

Good computational properties of Gaussian-error assumption

$$P(y|\widehat{y}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\widehat{y})^2}{2\sigma^2}\right]$$

- Regardless of the value of σ^2, this quantity is minimized when $\sum_i (y_i - \widehat{y}_i)^2$ is minimized across the dataset

Good computational properties of Gaussian-error assumption

$$P(y|\widehat{y}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\widehat{y})^2}{2\sigma^2}\right]$$

- Regardless of the value of σ^2, this quantity is minimized when $\sum_i (y_i - \widehat{y}_i)^2$ is minimized across the dataset
- Therefore, maximizing data likelihood is equivalent to **minimizing the sum of squared residual errors**

Good computational properties of Gaussian-error assumption

$$P(y|\widehat{y}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\widehat{y})^2}{2\sigma^2}\right]$$

- Regardless of the value of σ^2, this quantity is minimized when $\sum_i (y_i - \widehat{y}_i)^2$ is minimized across the dataset
- Therefore, maximizing data likelihood is equivalent to **minimizing the sum of squared residual errors**
  - This gives rise to the name commonly used for linear regression, ordinary least squares

The least-squares fit for linear regression

Let $\mathrm{RSS}(\beta) = \sum_i (y_i - \widehat{y}_i)^2$ and use matrix notation $\mathrm{y} = \mathrm{X}\beta + \epsilon$:

$$\mathrm{RSS}(\beta) = (\mathrm{y} - \mathrm{X}\beta)^\top (\mathrm{y} - \mathrm{X}\beta)$$

This is minimized when its gradient with respect to β is 0, i.e., $\nabla_\beta(\mathrm{RSS}(\beta)) = 0$. Using vector calculus we derive:

$$\begin{aligned}
0 &= \nabla_\beta \left((\mathrm{y} - \mathrm{X}\beta)^\top (\mathrm{y} - \mathrm{X}\beta) \right) \\
&= \left(\nabla_\beta (y - X\beta) \right)^\top (y - X\beta) + \left(\nabla_\beta (y - X\beta) \right) (y - X\beta)^\top \\
&= -\mathrm{X}^\top (y - X\beta) - \mathrm{X}(y - X\beta)^\top \\
&= -2 \left(\mathrm{X}^\top \mathrm{y} - \mathrm{X}^\top \mathrm{X}\beta \right)
\end{aligned}$$

If $\mathrm{X}^\top \mathrm{X}$ is invertible, then this equation is solved when:

$$\beta = \left(\mathrm{X}^\top \mathrm{X} \right)^{-1} \mathrm{X}^\top \mathrm{y}$$

This is thus the OLS, and equivalently the ML, estimate $\widehat{\beta}$.

Why to expect Gaussian(-ish) noise around $\widehat{y}$

- In regression modeling with a continuous response, a frequently plausible working assumption is that the residual error involves the **additive influence** of **large numbers** of **separate** (unconditionally independent) **stochastic factors** that are **not captured by the predictors included in the regression model**.

Why to expect Gaussian(-ish) noise around $\widehat{y}$

- In regression modeling with a continuous response, a frequently plausible working assumption is that the residual error involves the **additive influence** of **large numbers** of **separate** (unconditionally independent) **stochastic factors** that are **not captured by the predictors included in the regression model**.
- In this setting, we can invoke the Central Limit Theorem, stated here somewhat informally:

> The mean of n independently distributed random variables approaches a normal distribution as $n \to \infty$.

How bad is it if residual error is not truly Gaussian?

Even if residual error is not truly Gaussian, OLS linear regression still has desirable properties:

- So long as errors are IID, OLS is the Best Linear Unbiased Estimator (BLUE) of the regression coefficients—this is the Gauss–Markov theorem

How bad is it if residual error is not truly Gaussian?

Even if residual error is not truly Gaussian, OLS linear regression still has desirable properties:

- So long as errors are IID, OLS is the Best Linear Unbiased Estimator (BLUE) of the regression coefficients—this is the Gauss–Markov theorem
    - ("Best" here means lowest-variance)

How bad is it if residual error is not truly Gaussian?

Even if residual error is not truly Gaussian, OLS linear regression still has desirable properties:

- So long as errors are IID, OLS is the Best Linear Unbiased Estimator (BLUE) of the regression coefficients—this is the Gauss–Markov theorem
  - (“Best” here means lowest-variance)

How bad is it if residual error is not truly Gaussian?

Even if residual error is not truly Gaussian, OLS linear regression still has desirable properties:

- So long as errors are IID, OLS is the Best Linear Unbiased Estimator (BLUE) of the regression coefficients—this is the Gauss–Markov theorem
  - (“Best” here means lowest-variance)

However, non-normal residual error affects correctness of inferences regarding coefficient estimates, beyond simple unbiasedness. We now turn to a large family of techniques that manages the resulting issues.

References I