Should you always check assumptions first?

In practice, it's mixed. Many experienced data scientists actually start with exploratory modeling using various algorithms (including linear regression) to get a baseline understanding, then dive deeper into assumptions for models they're seriously considering for production.

However, **checking assumptions is still valuable** because:

1. **Interpretability matters** - Linear regression coefficients are easily interpretable for stakeholders
2. **Computational efficiency** - Linear models are fast to train and predict
3. **Baseline establishment** - Simple models often perform surprisingly well and set a good benchmark
4. **Debugging complex models** - If a simple linear model fails badly, complex models might fail for the same fundamental reasons

Why not jump straight to complex models?

You absolutely can, and many practitioners do. But there are trade-offs:

Advantages of starting complex:

- Potentially better performance immediately
- Less time spent on assumption checking
- Modern tools make complex models easier to use

Disadvantages:

- **Black box problem** - Harder to explain to business stakeholders
- **Overfitting risk** - Complex models can memorize noise
- **Computational cost** - Training and inference are more expensive
- **Debugging difficulty** - When things go wrong, it's harder to understand why
- **Maintenance complexity** - More moving parts in production

Real-world practical approach:

Most successful data scientists use a **tiered approach**:

1. **Quick EDA** - Basic plots to understand data structure
2. **Simple baseline** - Linear regression, basic tree models
3. **Assumption checking** - Only for models you might actually deploy
4. **Complex models** - Gradient boosting, neural networks, etc.
5. **Model selection** - Balance performance, interpretability, and operational requirements

When assumptions really matter:

- **High-stakes decisions** (medical, financial, legal)
- **Regulatory environments** requiring explainable models
- **Limited data** where overfitting is a real concern
- **Production systems** where model drift monitoring is crucial

The key insight is that **model selection isn't just about accuracy** - it's about finding the right balance of performance, interpretability, maintainability, and business requirements for your specific use case.