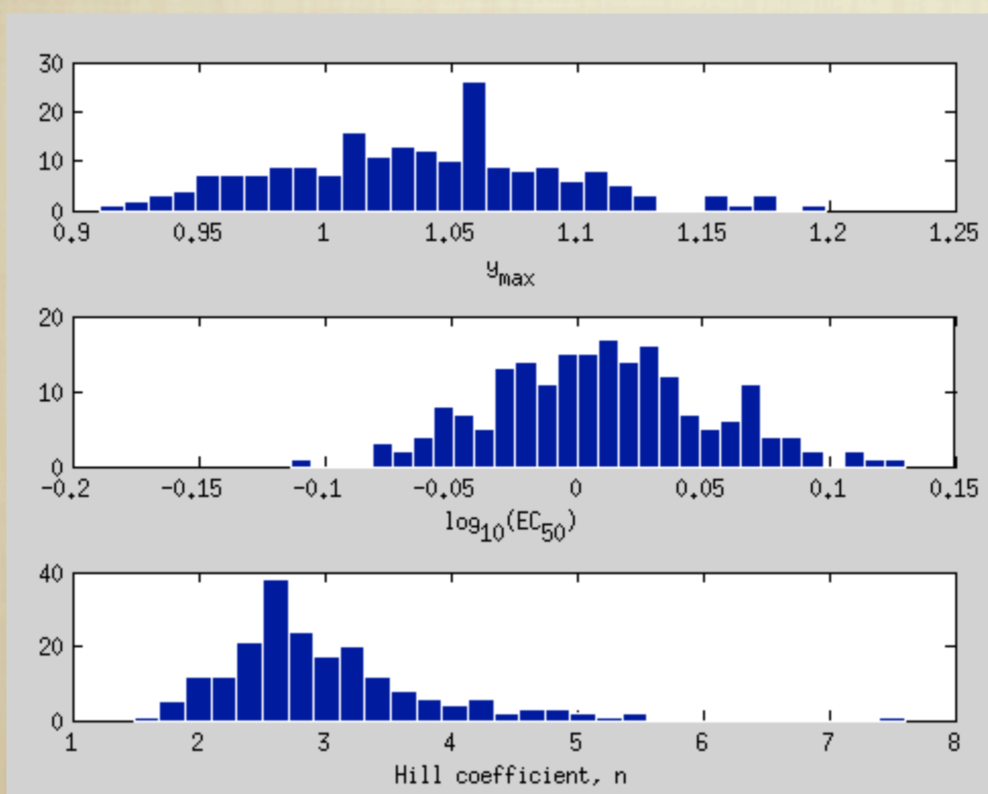
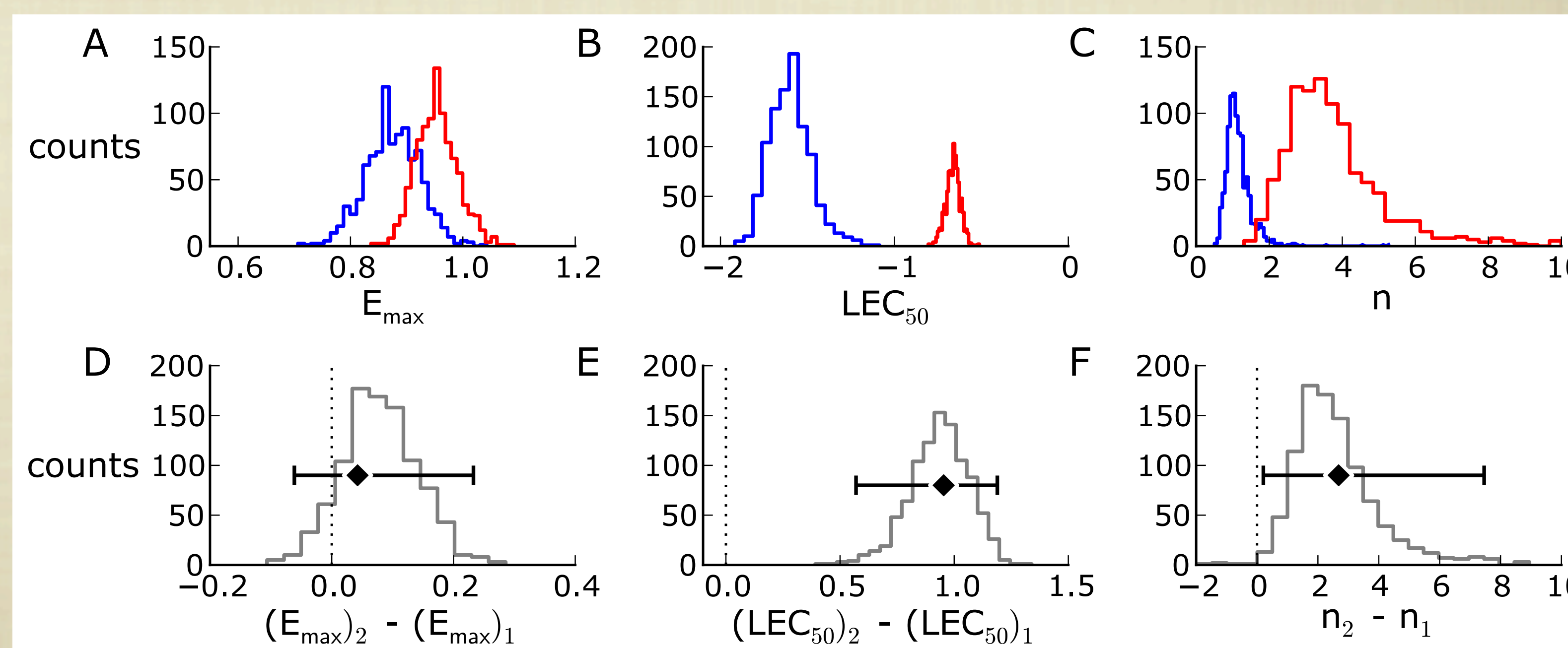
Experimental data analysis
Lecture 4: Model Selection

Dodo Das

Review of lecture 3

- Asymptotic confidence intervals (nlparci)
- Bootstrap confidence intervals (bootstrp)
- Bootstrap hypothesis testing





How to select the best model?

Case study: Fitting data with models of varying complexity.

- Simulate data from the following model:

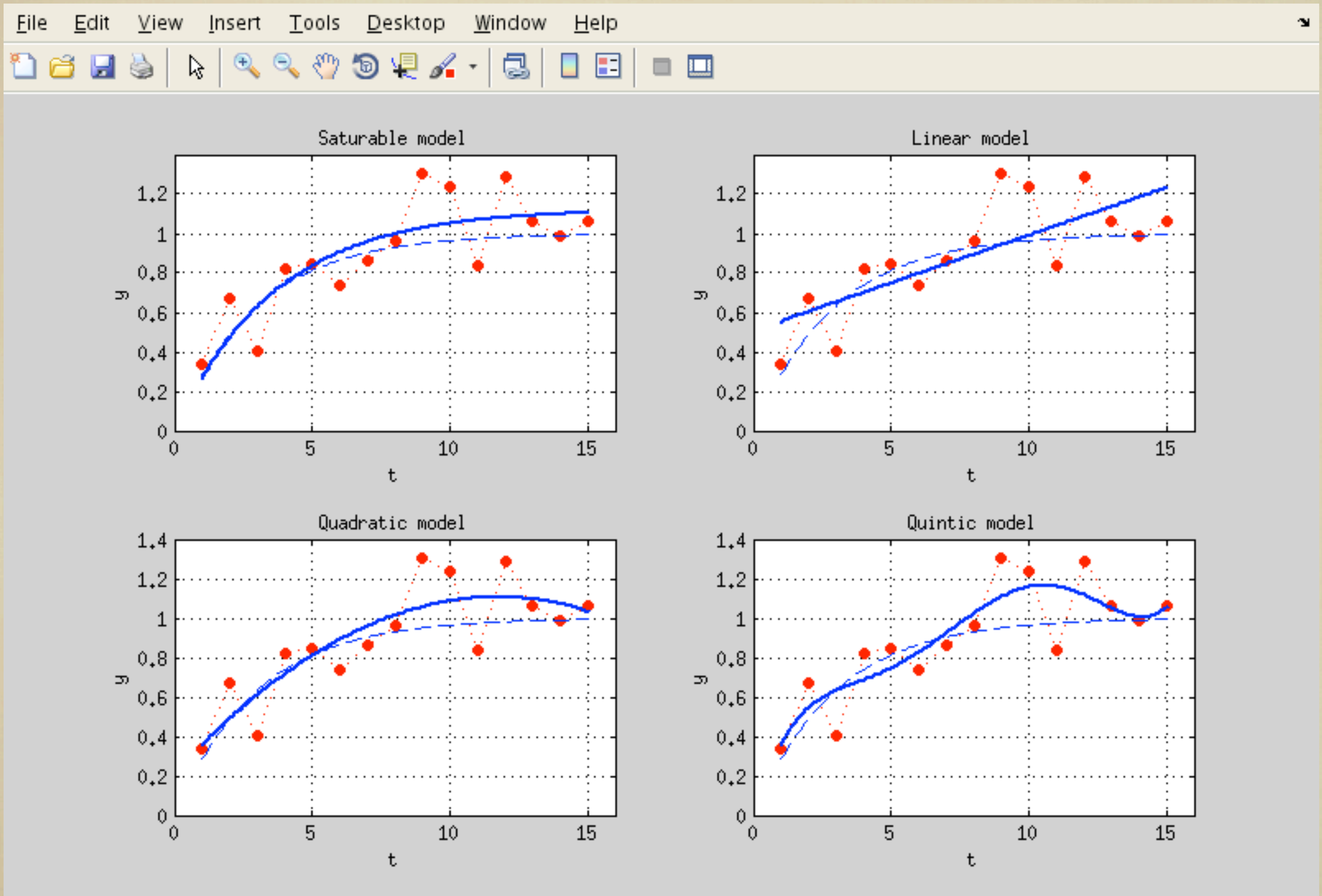
$$\boxed{y_{\text{true}} = y_{\max}\left(1 - e^{-t/\tau}\right)}$$

- Fit with 4 different models:

$$y_{\text{sat}} = y_{\text{true}} = y_{\max}\left(1 - e^{-t/\tau}\right) \qquad y_{\text{lin}} = c_0 + c_1 t$$

$$y_{\text{quad}} = c_0 + c_1 t + c_2 t^2 \qquad y_{\text{quint}} = c_0 + c_1 t + c_2 t^2 + \cdots c_5 t^5$$

Single dataset fits



How to rank models?

- The model with the highest number of parameters will typically give the best fit.
- But introducing more parameters increases model complexity, and the requires us to estimate values of all the additional parameters from limited data.
- Increasing the number of parameters also makes the model more susceptible to noise.

Quantifying model complexity: Defining Risk

- Squared error at a point:

$$L\big(f(x), \widehat{f}_n(x)\big) = \big(f(x) - \widehat{f}_n(x)\big)^2.$$

- Average over many experiments: mean squared error (or risk)

$$\mathrm{MSE} = R(f(x), \widehat{f}_n(x)) = \mathbb{E}\big(L(f(x), \widehat{f}_n(x))\big).$$

$$R\big(f(x), \widehat{f}_n(x)\big) = \mathsf{bias}_x^2 + \mathbb{V}_x$$

where

$$\mathsf{bias}_x = \mathbb{E}(\widehat{f}_n(x)) - f(x)$$

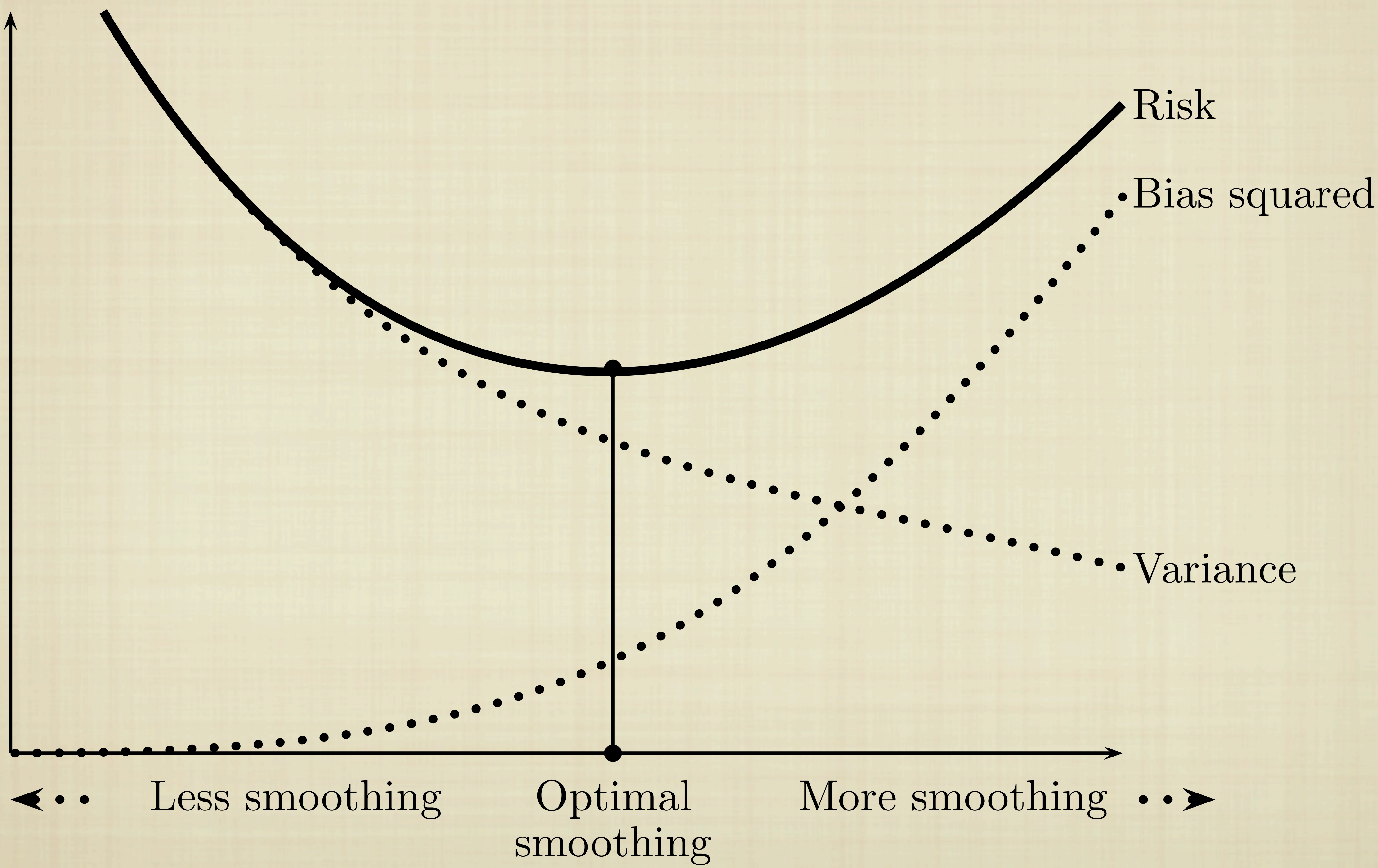
is the bias of $\widehat{f}_n(x)$ and

$$\mathbb{V}_x = \mathbb{V}(\widehat{f}_n(x))$$

is the variance of $\widehat{f}_n(x)$. In words:

Bias-Variance tradeoff

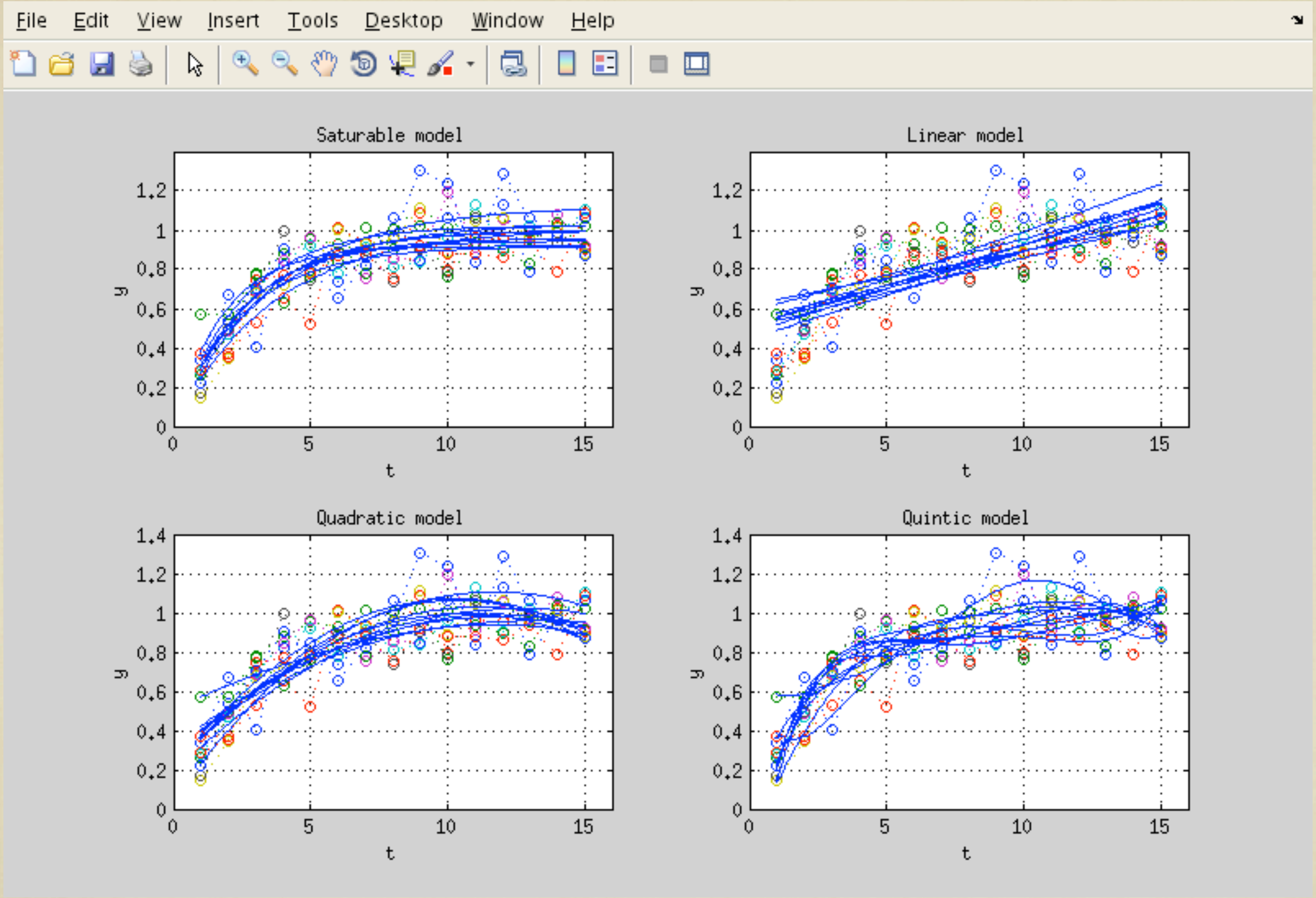
$$\text{RISK} = \text{MSE} = \text{BIAS}^2 + \text{VARIANCE}.$$



More complexity

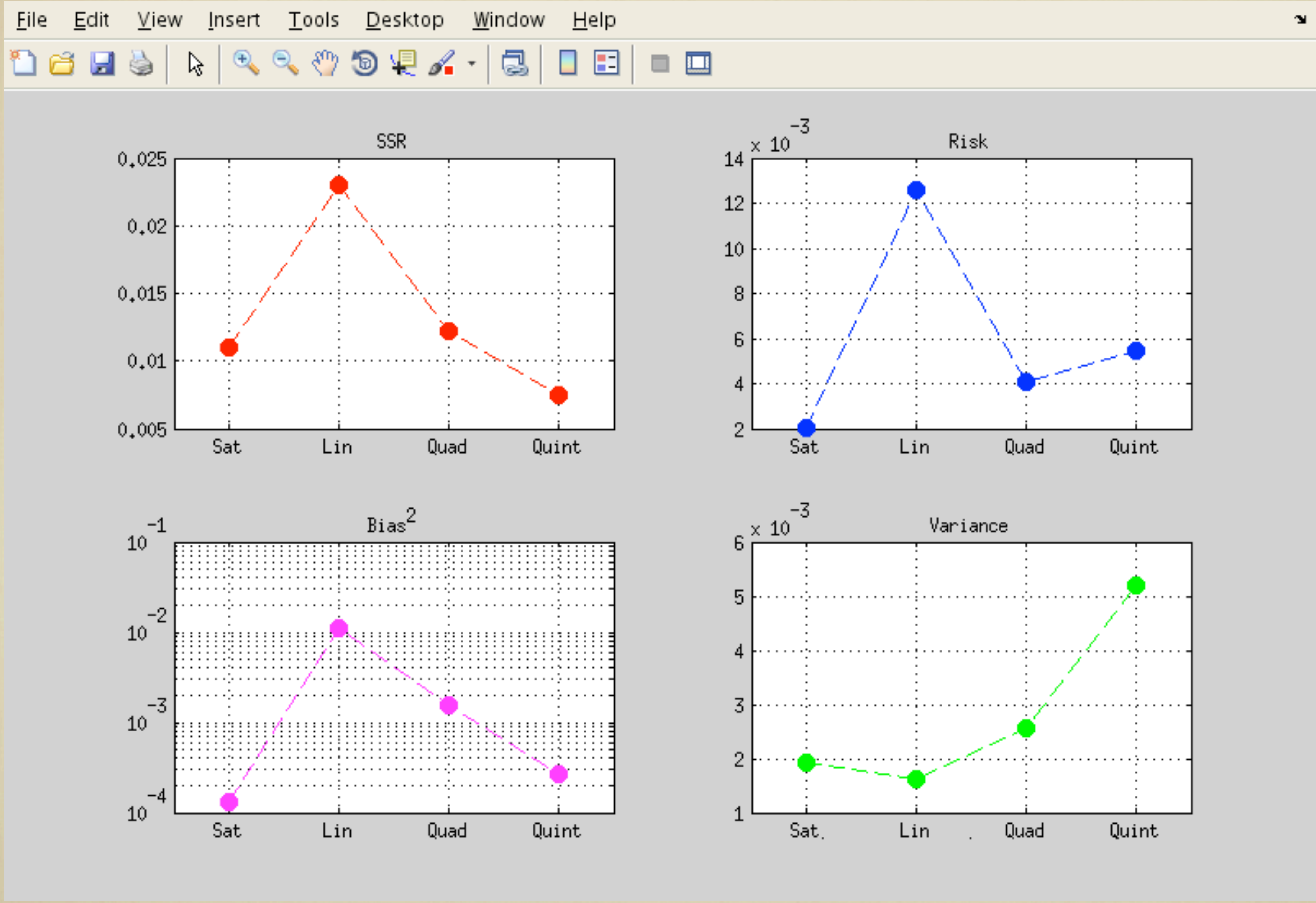
Less complexity

Bias-variance tradeoff for the simulated data



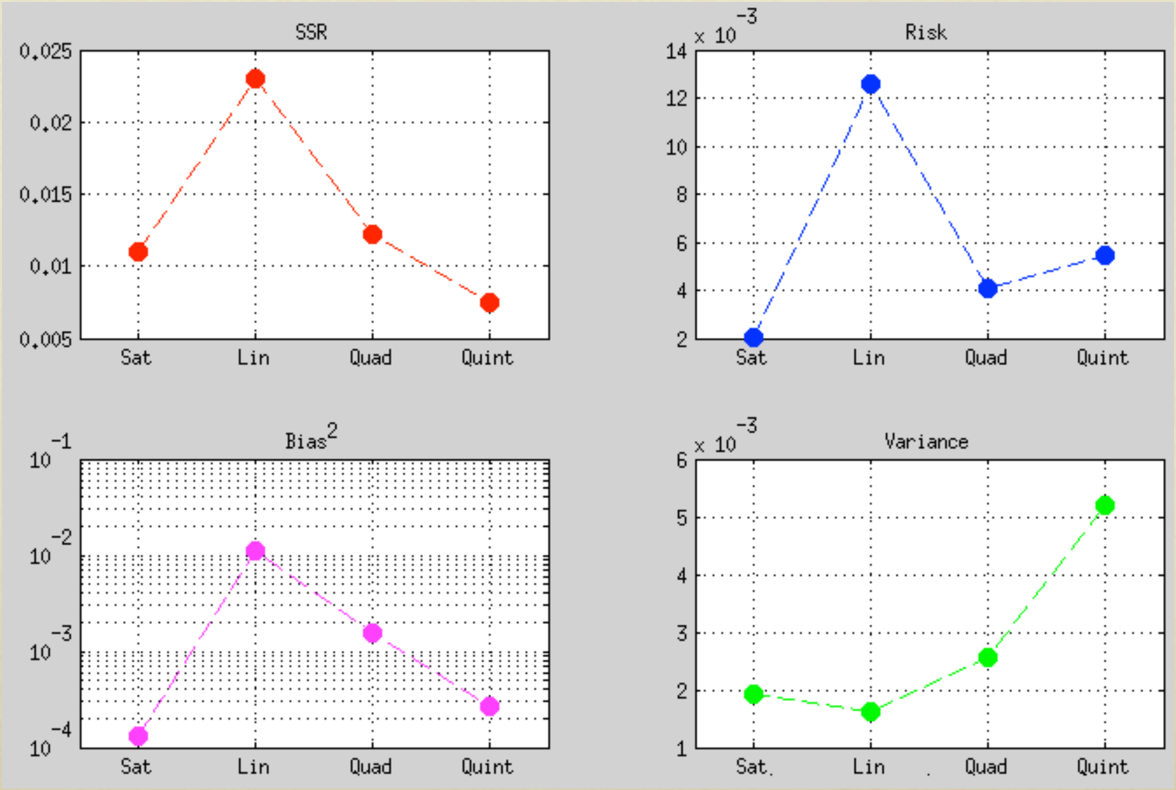
- Simulate many independent replicates, and fit them individually.

Bias-variance tradeoff for the simulated data



The guiding principle: 'Parsimony'

- Even though the 5th order polynomial fits the model the best (lowest SSR), it has high variance.
- In contrast, the linear model has the lowest variance, but a high bias.
- The goal is to pick the model that does the best job with the least number of parameters.



Methods for choosing an optimal model

- In general, we don't know the 'true' model, so we can't calculate the risk. Therefore, we need some way to rank model.

- The general idea is to give models preference if they reduce SSR, but penalize them if they have too many parameters.

- Two main techniques.

1. **Akaike's information criterion (AIC)**: Based on an information theory-based approximation for risk

2. **F-test**: Based on asymptotic statistical theory of the distribution of normal errors

Akaike's information criterion

- $AIC = -2\ln(L) + 2k + 2k(k+1)/(n-k-1)$
- For least squares regression: $AIC = SSR + 2k + 2k(k+1)/(n-k-1)$
- n = number of data points, k = number of parameters
- Lower AIC is better.
- For the i-th model among candidate models

$\Delta_i = (AIC)_i - (AIC)_{min}$

$w_i = \exp(-\Delta_i/2)/\Sigma \exp(-\Delta_i/2)$

F-test

- Can only use for nested models (eg: the linear, quadratic and quintic polynomial models). Cannot compare the saturable model with any of these using F-test.

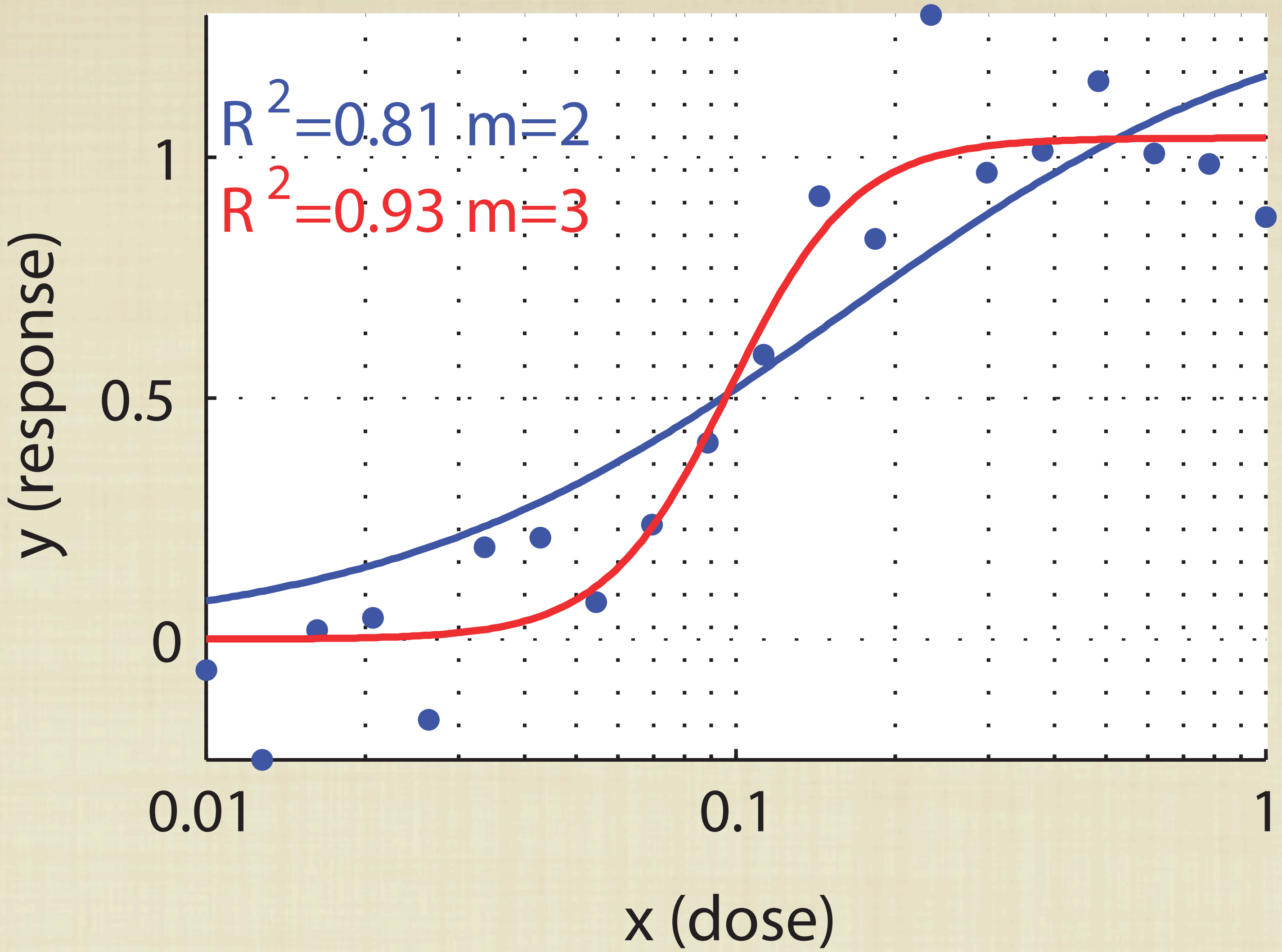
- Compute the F-statistic

$$F = \frac{\mathrm{SSR}_0 - \mathrm{SSR}_1}{\mathrm{SSR}_1} \cdot \frac{N - m_1}{m_1 - m_0}$$

- Null hypothesis: Simple model is sufficient.

- Look up table for associated *p*-value. If *p<0.05*, reject null hypothesis at 5% significance level.

Example: Application of F test



Number of data points, N	20	
Number of model parameters	$m_0 = 2$	$m_1 = 3$
Sum of squared residuals	$\mathrm{SSR}_0 = 0.881$	$\mathrm{SSR}_1 = 0.318$
F-score	30.1	
p-value	7.45×10^{-6}	
Is $p < 0.05$?	Yes	
Reject null hypothesis		