## Model Fitting

- Non-linear regression
- Density (shape) estimation
- Parameter estimation of the assumed model
- Goodness of fit
Model Fitting
- Non-linear regression
- Density (shape) estimation
- Parameter estimation of the assumed model
- Goodness of fit

Model Selection
- Nested (In quasar spectrum, should one add a broad absorption line BAL component to a power law continuum? Are there 4 planets or 6 orbiting a star?)
- Non-nested (is the quasar emission process a mixture of blackbodies or a power law?).
- Model misspecification
Is the underlying nature of an X-ray stellar spectrum a non-thermal power law?
Is the underlying nature of an X-ray stellar spectrum a non-thermal power law?

Are the fluctuations in the cosmic microwave background best fit by Big Bang models with dark energy?
Is the underlying nature of an X-ray stellar spectrum a non-thermal power law?

Are the fluctuations in the cosmic microwave background best fit by Big Bang models with dark energy?

Are there interesting correlations among the properties of objects in any given class (e.g. the Fundamental Plane of elliptical galaxies), and what are the optimal analytical expressions of such correlations?
A good model should be

- Parsimonious (model simplicity)
- Conform fitted model to the data (goodness of fit)
- Easily generalizable.
- Not *under-fit* that excludes key variables or effects
- Not *over-fit* that is unnecessarily complex by including extraneous explanatory variables or effects.
- Under-fitting induces bias and over-fitting induces high variability.

A good model should balance the competing objectives of conformity to the data and parsimony.
$4Bn Chandra X-Ray observatory NASA 1999
1616 Bright Sources. Two weeks of observations in 2003
What is the underlying nature of a stellar spectrum?

Successful model for high signal-to-noise X-ray spectrum. Complicated thermal model with several temperatures and element abundances (17 parameters)
COUP source # 410 in Orion Nebula with 468 photons
Thermal model with absorption $A_V \sim 1$ mag
Fitting binned data using $\chi^2$
Model assuming a single-temperature thermal plasma with solar abundances of elements. The model has three free parameters denoted by a vector $\theta$.
- plasma temperature
- line-of-sight absorption
- normalization

The astrophysical model has been convolved with complicated functions representing the sensitivity of the telescope and detector.

The model is fitted by minimizing sum of squares (‘minimum chi-square’) with an iterative procedure.

$$\hat{\theta} = \arg \min_{\theta} \chi^2(\theta) = \arg \min_{\theta} \sum_{i=1}^{N} \left( \frac{y_i - M_i(\theta)}{\sigma_i} \right)^2.$$

*Chi-square minimization* is a misnomer. It is parameter estimation by *weighted least squares*. 
Limitations to $\chi^2$ ‘minimization’

- Fails when bins have too few data points.
- Binning is arbitrary. Binning involves loss of information.
Limitations to $\chi^2$ ‘minimization’

- Fails when bins have too few data points.
- Binning is arbitrary. Binning involves loss of information.
- Data points should be independent.
- Failure of independence assumption is common in astronomical data due to effects of the instrumental setup; e.g. it is typical to have $\geq 3$ pixels for each telescope point spread function (in an image) or spectrograph resolution element (in a spectrum). Thus adjacent pixels are not independent.

- Does not provide clear procedures for adjudicating between models with different numbers of parameters (e.g. one- vs. two-temperature models) or between different acceptable models (e.g. local minima in $\chi^2(\theta)$ space).

- Unsuitable to obtain confidence intervals on parameters when complex correlations between the estimators of parameters are present (e.g. non-parabolic shape near the minimum in $\chi^2(\theta)$ space).
Limitations to $\chi^2$ ‘minimization’

- Fails when bins have too few data points.
- Binning is arbitrary. Binning involves loss of information.
- Data points should be independent.
- Failure of independence assumption is common in astronomical data due to effects of the instrumental setup; e.g. it is typical to have $\geq 3$ pixels for each telescope point spread function (in an image) or spectrograph resolution element (in a spectrum). Thus adjacent pixels are not independent.
- Does not provide clear procedures for adjudicating between models with different numbers of parameters (e.g. one- vs. two-temperature models) or between different acceptable models (e.g. local minima in $\chi^2(\theta)$ space).
- Unsuitable to obtain confidence intervals on parameters when complex correlations between the estimators of parameters are present (e.g. non-parabolic shape near the minimum in $\chi^2(\theta)$ space).
Alternative approach to the model fitting based on EDF

Fitting to unbinned EDF
Correct model family, incorrect parameter value
Thermal model with absorption set at $A_V \sim 10$ mag
Misspecified model family!
Power law model with absorption set at $A_V \sim 1$ mag
Can the power law model be excluded with 99% confidence
1. Statistics based on EDF
2. Kolmogorov-Smirnov Statistic
3. Processes with estimated parameters
4. Bootstrap
5. Bootstrap for Time Series
6. Bootstrap outlier detection
7. Functional Model Fitting using Bootstrap
8. Confidence Limits Under Model Misspecification
Empirical Distribution Function

Cumulative Fraction Plot

[Graph showing a cumulative fraction plot with steps along the x-axis from 0.1 to 10, and corresponding cumulative fractions along the y-axis from 0.0 to 1.0]
Statistics based on EDF

Kolmogrov-Smirnov:  \[ D_n = \sup_x |F_n(x) - F(x)|, \]

\[ H(y) = P(D_n \leq y), \quad 1 - H(d_n(\alpha)) = \alpha \]

Cramér-von Mises:  \[ \int (F_n(x) - F(x))^2 \ dF(x) \]

Anderson - Darling:  \[ \int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \ dF(x) \]

is more sensitive at tails.
Statistics based on EDF

**Kolmogrov-Smirnov:** \[ D_n = \sup_x |F_n(x) - F(x)|, \]

\[ H(y) = P(D_n \leq y), \quad 1 - H(d_n(\alpha)) = \alpha \]

**Cramér-von Mises:** \[ \int (F_n(x) - F(x))^2 \, dF(x) \]

**Anderson - Darling:** \[ \int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x) \]

is more sensitive at tails.

- These statistics are distribution free if \( F \) is continuous & univariate.
Statistics based on EDF

Kolmogrov-Smirnov: \[ D_n = \sup_x |F_n(x) - F(x)|, \]

\[ H(y) = P(D_n \leq y), \quad 1 - H(d_n(\alpha)) = \alpha \]

Cramér-von Mises: \[ \int (F_n(x) - F(x))^2 \, dF(x) \]

Anderson - Darling: \[ \int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x) \]

is more sensitive at tails.

- These statistics are distribution free if \( F \) is continuous & univariate.
- No longer distribution free if either \( F \) is not univariate or parameters of \( F \) are estimated.
Kolmogrov-Smirnov: \( D_n = \sup_x |F_n(x) - F(x)|, \)

\[ H(y) = P(D_n \leq y), \quad 1 - H(d_n(\alpha)) = \alpha \]

Cramér-von Mises: \[ \int (F_n(x) - F(x))^2 \, dF(x) \]

Anderson - Darling: \[ \int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x) \]

is more sensitive at tails.

- These statistics are distribution free if \( F \) is continuous & univariate.
- No longer distribution free if either \( F \) is not univariate or parameters of \( F \) are estimated.

**EDF based fitting requires little or no probability distributional assumptions such as Gaussianity or Poisson structure.**
The KS statistic is used in ~500 astronomical papers/yr, but often incorrectly or with less efficiency than an alternative statistic.
The KS statistic is used in ~500 astronomical papers/yr, but often incorrectly or with less efficiency than an alternative statistic.

The 1-sample KS test (data vs. model comparison) is distribution-free only when the model is not derived from the dataset.

The KS test is distribution-free (i.e. probabilities can be used for hypothesis testing) only in 1-dimension.
The KS statistic is used in ∼500 astronomical papers/yr, but often incorrectly or with less efficiency than an alternative statistic.

The 1-sample KS test (data vs. model comparison) is distribution-free only when the model is not derived from the dataset.

The KS test is distribution-free (i.e. probabilities can be used for hypothesis testing) only in 1-dimension.

Probabilities need to be obtained from bootstrap resampling in these cases.
Misuse of Kolmogorov-Smirnov

The KS statistic is used in \( \sim 500 \) astronomical papers/yr, but often incorrectly or with less efficiency than an alternative statistic.

The 1-sample KS test (data vs. model comparison) is distribution-free only when the model is not derived from the dataset.

The KS test is distribution-free (i.e. probabilities can be used for hypothesis testing) only in 1-dimension.

Probabilities need to be obtained from bootstrap resampling in these cases.

**Numerical Recipe’s treatment of a 2-dim KS test is mathematically invalid.**

See the viral page

**Beware the Kolmogorov-Smirnov test!**

at http://asaip.psu.edu
KS probabilities are invalid when the model parameters are estimated from the data. Some astronomers use them incorrectly.

– Lillifors (1964)
Kolmogorov-Smirnov and Anderson-Darling Statistics

The KS statistic efficiently detects differences in global shapes, but not small scale effects or differences near the tails. The Anderson-Darling statistic (tail-weighted Cramer-von Mises statistic) is more sensitive.

\[
KS_n = \sqrt{n} \sup_x |F_n(x) - F(x)| \quad AD_n = n \int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)
\]
Example – Paul B. Simpson (1951)

\[ F(x, y) = ax^2y + (1 - a)y^2x, \quad 0 < x, y < 1 \]

\((X_1, Y_1) \sim F. \quad F_1 \text{ denotes the EDF of } (X_1, Y_1)\)

\[
P(|F_1(x, y) - F(x, y)| < .72, \text{ for all } x, y)\]

\[
> .065 \text{ if } a = 0, \quad (F(x, y) = y^2x)\]

\[
< .058 \text{ if } a = .5, \quad (F(x, y) = \frac{1}{2}xy(x + y))\]

Numerical Recipe’s treatment of a 2-dim KS test is mathematically invalid.
Processes with estimated parameters

\[ \{ F(\cdot; \theta) : \theta \in \Theta \} \] – a family of continuous distributions

\( \Theta \) is a open region in a \( p \)-dimensional space.

\( X_1, \ldots, X_n \) sample from \( F \)

Test \( F = F(\cdot; \theta) \) for some \( \theta = \theta_0 \)

Kolmogorov-Smirnov, Cramér-von Mises statistics, etc., when \( \theta \) is estimated from the data, are continuous functionals of the empirical process

\[ Y_n(x; \hat{\theta}_n) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n)) \]

\( \hat{\theta}_n = \theta_n(X_1, \ldots, X_n) \) is an estimator \( \theta \)

\( F_n \) – the EDF of \( X_1, \ldots, X_n \)

– The so-called ‘Bootstrap’ helps here.
Astronomers have often used *Monte Carlo methods* to simulate datasets from uniform or Gaussian populations. While helpful in some cases, this does not avoid the assumption of a simple underlying distribution.
Astronomers have often used *Monte Carlo methods* to simulate datasets from uniform or Gaussian populations. While helpful in some cases, this does not avoid the assumption of a simple underlying distribution.

Instead, what if we take the observed data as hypothetical ‘population’ and use Monte Carlo simulation on it. Can simulate many ‘datasets’ and, each of these can be analyzed in the same way to see how the estimates depend on plausible random variations in the data.

(No costly observations for ‘new/additional’ data).
Bootstrap (a resampling procedure) is a Monte Carlo method of simulating ‘datasets’ from an observed/given data, without any assumption on the underlying population.
What is Bootstrap?

- Bootstrap (a resampling procedure) is a Monte Carlo method of simulating ‘datasets’ from an observed/given data, without any assumption on the underlying population.

- Resampling the original data preserves (adaptively) whatever distributions are truly present, including selection effects such as truncation (flux limits or saturation).
Bootstrap (a resampling procedure) is a Monte Carlo method of simulating ‘datasets’ from an observed/given data, without any assumption on the underlying population.

Resampling the original data preserves (adaptively) whatever distributions are truly present, including selection effects such as truncation (flux limits or saturation).

Bootstrap helps evaluate statistical properties using data rather than an assumed Gaussian or power law or other distributions.
What is Bootstrap?

- Bootstrap (a resampling procedure) is a Monte Carlo method of simulating ‘datasets’ from an observed/given data, without any assumption on the underlying population.

- Resampling the original data preserves (adaptively) whatever distributions are truly present, including selection effects such as truncation (flux limits or saturation).

- Bootstrap helps evaluate statistical properties using data rather than an assumed Gaussian or power law or other distributions.

- Bootstrap procedures are supported by solid theoretical foundations.
Bootstrap Procedure

\[ X = (X_1, \ldots, X_n) - \text{a sample from } F \]
\[ X^* = (X^*_1, \ldots, X^*_n) - \text{a simple random sample from the data.} \]
\[ \hat{\theta} \quad \text{is an estimator of} \quad \theta \]
\[ \theta^* \quad \text{is based on} \quad X^*_i \]

**Examples:**

\[ \hat{\theta} = \bar{X}_n, \quad \theta^* = \bar{X}^*_n \]
\[ \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \quad \theta^* = \frac{1}{n} \sum_{i=1}^{n} (X^*_i - \bar{X}^*_n)^2 \]
\[ \theta^* - \hat{\theta} \quad \text{behaves like} \quad \hat{\theta} - \theta \]
Simple random sampling from data is equivalent to drawing a set of i.i.d. random variables from the empirical distribution. This is Nonparametric Bootstrap.
Simple random sampling from data is equivalent to drawing a set of i.i.d. random variables from the empirical distribution. This is **Nonparametric Bootstrap**.

**Parametric Bootstrap** if $X_1^*, \ldots, X_n^*$ are i.i.d. r.v. from $\hat{H}_n$, an estimator of $F$ based on data $(X_1, \ldots, X_n)$.
Nonparametric and Parametric Bootstrap

Simple random sampling from data is equivalent to drawing a set of i.i.d. random variables from the empirical distribution. This is Nonparametric Bootstrap.

Parametric Bootstrap if $X_1^*, \ldots, X_n^*$ are i.i.d. r.v. from $\hat{H}_n$, an estimator of $F$ based on data $(X_1, \ldots, X_n)$.

Example of Parametric Bootstrap:

$X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$

$X_1^*, \ldots, X_n^*$ i.i.d. $\sim N(\bar{X}_n, s_n^2)$; $s_n^2 = \frac{1}{n} \sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$N(\bar{X}_n, s_n^2)$ is a good estimator of the distribution $N(\mu, \sigma^2)$
\( \hat{\theta} \) is an estimator of \( \theta \) based on \( X_1, \ldots, X_n \).

\( \theta^* \) denotes the bootstrap estimator based on \( X_{1*}, \ldots, X_{n*} \).

\[ \text{Var}^*(\hat{\theta}) = E^* (\theta^* - E(\theta^*))^2 \]

In practice, generate \( N \) bootstrap samples of size \( n \). Compute \( \theta_{1*}, \ldots, \theta_{N*} \) for each of the \( N \) samples.

\[ \bar{\theta}^* = \frac{1}{N} \sum_{i=1}^{N} \theta_{i*} \]

\[ \text{Var}(\hat{\theta}) \approx \frac{1}{N} \sum_{i=1}^{N} (\theta_{i*} - \bar{\theta}^*)^2 \]
Bootstrap Distribution

Statistical inference requires sampling distribution $G_n$, given by $G_n(x) = P(\sqrt{n}(\bar{X} - \mu)/\sigma \leq x)$

<table>
<thead>
<tr>
<th>statistic</th>
<th>bootstrap version</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sqrt{n}(\bar{X} - \mu)/\sigma$</td>
<td>$\sqrt{n}(\bar{X}^* - \bar{X})/s_n$</td>
</tr>
<tr>
<td>$\sqrt{n}(\bar{X} - \mu)/s_n$</td>
<td>$\sqrt{n}(\bar{X}^* - \bar{X})/s_n^*$</td>
</tr>
</tbody>
</table>

where $s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ and $s_n^* = \frac{1}{n} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)^2$

For a given data, the bootstrap distribution $G_B$ is given by

$$G_B(x) = P^*(\sqrt{n}(\bar{X}^* - \bar{X})/s_n \leq x|X)$$

$G_B$ is completely known and $G_n \approx G_B$. 
If $G_n$ denotes the sampling distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$ then the corresponding bootstrap distribution $G_B$ is given by

$$G_B(x) = P^*(\sqrt{n}(\bar{X}^* - \bar{X})/s_n \leq x|X).$$

**Construction of Bootstrap Histogram**

$M = n^n$ bootstrap samples possible

$$X_1^{*(1)}, \ldots, X_n^{*(1)} \quad r_1 = \sqrt{n}(\bar{X}^{*(1)} - \bar{X})/s_n$$

$$X_1^{*(2)}, \ldots, X_n^{*(2)} \quad r_2 = \sqrt{n}(\bar{X}^{*(2)} - \bar{X})/s_n$$

\[\vdots\] \[\vdots\]

$$X_1^{*(M)}, \ldots, X_n^{*(M)} \quad r_M = \sqrt{n}(\bar{X}^{*(M)} - \bar{X})/s_n$$

Frequency table or histogram based on $r_1, \ldots, r_M$ gives $G_B$.

$$G_B(x) = \frac{1}{M} \#(r_i \leq x).$$
Confidence Interval for the mean

For $n = 10$ data points, $M = \text{ten billion}$

$N \sim n(\log n)^2$ bootstrap replications suffice
$N$ is much smaller than $n^n$.


Compute $\sqrt{n}(\bar{X}^*(j) - \bar{X})/s_n$ for $N$ bootstrap samples

Arrange them in increasing order

$r_1 < r_2 < \cdots < r_N \quad k = [0.05N], \quad m = [0.95N]$

90% Confidence Interval for $\mu$ is

$$
\bar{X} - r_m \frac{s_n}{\sqrt{n}} \leq \mu < \bar{X} - r_k \frac{s_n}{\sqrt{n}}
$$
Pearson’s correlation coefficient and its bootstrap version

\[
\hat{\rho} = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i Y_i - \bar{X} \bar{Y})}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2\right) \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2\right)}}
\]

\[
\rho^* = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i^* Y_i^* - \bar{X}_n^* \bar{Y}_n^*)}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^{n} (X_i^* - \bar{X}_n^*)^2\right) \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i^* - \bar{Y}_n^*)^2\right)}}
\]

**Smooth Functional Model**

\[
\hat{\rho} = H(\bar{Z}), \quad \text{where} \quad Z_i = (X_i Y_i, X_i^2, Y_i^2, X_i, Y_i)
\]

\[
H(a_1, a_2, a_3, a_4, a_5) = \frac{(a_1 - a_4 a_5)}{\sqrt{((a_2 - a_4^2)(a_3 - a_5^2))}}
\]

\[
\rho^* = H(\bar{Z}^*), \quad \text{where} \quad Z_i^* = (X_i^* Y_i^*, X_i^{*2}, Y_i^{*2}, X_i^*, Y_i^*)
\]
$H$ is a smooth function and $\mathbf{Z}_1$ is a random vector.

$\hat{\theta} = H(\bar{\mathbf{Z}})$ is an estimator of the parameter $\theta = H(\mathbb{E}(\mathbf{Z}_1))$

Division (normalization) of $\sqrt{n}(H(\bar{\mathbf{Z}}) - H(\mathbb{E}(\mathbf{Z}_1)))$ by its standard deviation makes them units free.

Studentization, if estimates of standard deviations are used. Under some regularity conditions Bootstrap distribution gives a very good approximation to the sampling distribution of such normalized statistics.

The theory works for both parametric and nonparametric Bootstrap.

- Babu and Singh (1984) Sankhyā
In practice

- Randomly generate $N \sim n(\log n)^2$ bootstrap samples
- Compute $t_n^{*}(j)$ for each bootstrap sample
- Arrange them in increasing order $u_1 < u_2 < \cdots < u_N$, $k = [0.05N]$, $m = [0.95N]$
- 90% Confidence Interval for the parameter $\theta$ is

$$\hat{\theta} - u_m \frac{\hat{\sigma}_n}{\sqrt{n}} \leq \theta < \hat{\theta} - u_k \frac{\hat{\sigma}_n}{\sqrt{n}}$$

This is called bootstrap PERCENTILE-$t$ confidence interval
When does bootstrap work well

- Sample Means
- Sample Variances
- Central and Non-central t-statistics
  (with possibly non-normal populations)
- Sample Coefficient of Variation
- Maximum Likelihood Estimators
- Least Squares Estimators
- Correlation Coefficients
- Regression Coefficients
- Smooth transforms of these statistics
When does Bootstrap fail

\[ \hat{\theta} = \max_{1 \leq i \leq n} X_i \]  
Non-smooth estimator

When does Bootstrap fail

\[ \hat{\theta} = \max_{1 \leq i \leq n} X_i \quad \text{Non-smooth estimator} \]


\[ \hat{\theta} = \bar{X} \quad \text{and} \quad E X_1^2 = \infty \quad \text{Heavy tails} \]

- Babu (1984) Sankhyā
When does Bootstrap fail

- $\hat{\theta} = \max_{1 \leq i \leq n} X_i$ Non-smooth estimator

- $\hat{\theta} = \bar{X}$ and $EX_1^2 = \infty$ Heavy tails
  - Babu (1984) Sankhyā

- $\hat{\theta} - \theta = H(\tilde{Z}) - H(E(Z_1))$ and $\partial H(E(Z_1)) = 0$
  Limit distribution is like linear combinations of Chi-squares. But here a modified version works.
  - Babu (1984) Sankhyā
Non-independent case

\( X_1, \ldots, X_n \) are identically distributed but not independent

- Straight forward bootstrap does not work in the dependent case. Variances of sums of random variables do not match.
- A clear knowledge of the dependent structure is needed to replicate resampling procedure.
- Classical bootstrap fails in the case of Time Series data.
- If the process is auto-regressive or moving-average one can replicate resampling procedure.
- In the general time-series case the *moving block bootstrap* is suggested.
$X_1, \cdots, X_n$ is a stationary sequence.

1. The sequence is split into overlapping blocks $B_1, \cdots, B_{n-b+1}$, of length $b$, where $B_j$ consists of $b$ consecutive observations starting from $X_j$, i.e., $B_j = \{X_j, X_{j+1}, \cdots, X_{j+b-1}\}$. Observation 1 to $b$ will be block 1, observation 2 to $b+1$ will be block 2 etc.

2. From these $n-b+1$ blocks, $n/b$ blocks will be drawn at random with replacement.

3. Align these $n/b$ blocks in the order they were picked.

This bootstrap procedure works with dependent data. By construction, the resampled data will not be stationary.

Varying randomly the block length can avoid this problem. However, the moving block bootstrap is still to be preferred.

Singh and Xie (2003, Sankhya) proposed a bootstrap density plot (histogram) of “mean – trimmed mean” for a suitable trimming number as a nonparametric graphical tool for detecting outlier(s) in a data set.

‘Bootlier’ plot is multimodal in the presence of outliers.

This method can be applied to data sets from a wide range of distributions, and it is quite effective in detecting outlying values in data sets with small portion of outliers.

**Strengths:**

- Its ability to incorporate heavy or short tailed data in outlier detections.
- Its effectiveness for outlier detection in multivariate settings where only few tools are available.
Density plots (histograms) of bootstrap sample mean (left), and bootstrap “mean − trimmed mean” (right). Original data are 20 standard normal observations with an outlier 6.

Linear Regression

\[ Y_i = \alpha + \beta X_i + \epsilon_i \]

\[ \mathbb{E}(\epsilon_i) = 0 \text{ and } \text{Var}(\epsilon_i) = \sigma_i^2 \]

Least squares estimators of \( \beta \) and \( \alpha \)

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
\]

\[
\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}
\]

\[
\text{Var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sigma_i^2}{L_n^2}
\]

\[
L_n = \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
Estimate the residuals \( e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i \)

Draw \( e_1^*, \ldots, e_n^* \) from \( \hat{e}_1, \ldots, \hat{e}_n \), where \( \hat{e}_i = e_i - \frac{1}{n}\sum_{j=1}^{n} e_j \).

Bootstrap estimators

\[
\beta^* = \hat{\beta} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(e_i^* - \bar{e}^*)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
\alpha^* = \hat{\alpha} + (\hat{\beta} - \beta^*)\bar{X} + \bar{e}^*
\]

\( V_B = E_B(\beta^* - \hat{\beta})^2 \approx \text{Var}(\hat{\beta}) \) efficient if \( \sigma_i = \sigma \)

\( V_B \) does not approximate the variance of \( \hat{\beta} \) under heteroscedasticity (i.e. unequal variances \( \sigma_i \))
Paired Bootstrap

Resample the pairs \((X_1, Y_1), \ldots, (X_n, Y_n)\)
\((\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Y}_n)\)

\[
\tilde{\beta} = \frac{\sum_{i=1}^{n}(\tilde{X}_i - \bar{\tilde{X}})(\tilde{Y}_i - \bar{\tilde{Y}})}{\sum_{i=1}^{n}(\tilde{X}_i - \bar{\tilde{X}})^2}, \quad \tilde{\alpha} = \bar{\tilde{Y}} - \tilde{\beta}\bar{\tilde{X}}
\]

Repeat the resampling \(N\) times and get

\[
\beta_{PB}^{(1)}, \ldots, \beta_{PB}^{(N)}
\]

\[
\frac{1}{N} \sum_{i=1}^{N}(\beta_{PB}^{(i)} - \hat{\beta})^2 \approx \text{Var}(\hat{\beta})
\]

even when not all \(\sigma_i\) are the same
Comparison

- **The Classical Bootstrap**
  - Efficient when $\sigma_i = \sigma$
  - But inconsistent when $\sigma_i$’s differ

- **The Paired Bootstrap**
  - Robust against heteroscedasticity
  - Works well even when $\sigma_i$ are all different


A handbook on ‘bootstrap’ for engineers to analyze complicated data with little or no model assumptions. Includes applications to radar and sonar signal processing.
We shall now get back to Goodness of Fit when parameters are estimated.
Parametric bootstrap

$X_1^*, \ldots, X_n^*$ sample generated from $F(.; \hat{\theta}_n)$

In Gaussian case $\hat{\theta}_n^* = (\bar{X}_n^*, s_n^2)$.

Both

$$\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n)|$$

and

$$\sqrt{n} \sup_x |F_n^*(x) - F(x; \hat{\theta}_n^*)|$$

have the same limiting distribution

In XSPEC package, the parametric bootstrap is command FAKEIT, which makes Monte Carlo simulation of specified spectral model.

Numerical Recipes describes a parametric bootstrap (random sampling of a specified pdf) as the ‘transformation method’ of generating random deviates.
Extreme daily precipitation over the Euro-Mediterranean area are modeled by a high-resolution Global Climate Model based on extreme value theory.

A modified Anderson-Darling statistic is used with Generalized Pareto family of distributions.

\[ F_{\mu,\sigma,\xi}(y) = \begin{cases} 1 - \left\{ 1 + \left( \frac{\xi(y - \mu)}{\sigma} \right) \right\}^{-1/\xi}, & \xi \neq 0, \ y \geq \mu \\ 1 - \exp\left( -\frac{(y - \mu)}{\sigma} \right), & \xi = 0 \ y \geq \mu, \end{cases} \]

where \( \sigma > 0, y \geq \mu \) when \( \xi > 0 \) and \( y \in [\mu, \mu - \sigma\xi] \), when \( \xi < 0 \). For modified Anderson-Darling statistic, both

\[ \int n \left( F_n(x) - F(x; \hat{\theta}_n) \right)^2 (1 - F(x; \hat{\theta}_n))^{-1} dF(x; \hat{\theta}_n) \]

and its bootstrap version

\[ \int n \left( F_n^*(x) - F(x; \hat{\theta}_n^*) \right)^2 (1 - F(x; \hat{\theta}_n^*))^{-1} dF(x; \hat{\theta}_n^*) \]

have the same limiting distribution, when \( \xi > 0 \).
Nonparametric bootstrap

$X_1^*, \ldots, X_n^*$ sample from $F_n$
i.e., a simple random sample from $X_1, \ldots, X_n$.

Bias correction

$$B_n(x) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n))$$

is needed.

Both

$$\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n)|$$

and

$$\sup_x |\sqrt{n} (F_n^*(x) - F(x; \hat{\theta}_n^*)) - B_n(x)|$$

have the same limiting distribution.

XSPEC does not provide a nonparametric bootstrap capability
Need for such bias corrections in special situations are well documented in the bootstrap literature.


Model misspecification

\( X_1, \ldots, X_n \) data from unknown \( H \).

\( H \) may or may not belong to the family \( \{ F(\cdot; \theta) : \theta \in \Theta \} \)

\( H \) is closest to \( F(\cdot, \theta_0) \)

Kullback-Leibler (information) divergence

\[
\int h(x) \log \left( \frac{h(x)}{f(x; \theta)} \right) d\nu(x) \geq 0
\]

\[
\int | \log h(x) | h(x) d\nu(x) < \infty
\]

\[
\int h(x) \log f(x; \theta_0) d\nu(x) = \max_{\theta \in \Theta} \int h(x) \log f(x; \theta) d\nu(x)
\]
For any $0 < \alpha < 1$,

$$P\left( \sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n) - (H(x) - F(x; \theta_0))| \leq C^*_\alpha \right) - \alpha \to 0$$

$C^*_\alpha$ is the $\alpha$-th quantile of

$$\sup_x |\sqrt{n} \left( F^*_n(x) - F(x; \hat{\theta}^*_n) \right) - \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n))|$$

This provide an estimate of the distance between the true distribution and the family of distributions under consideration.
Discussion so far

- K-S goodness of fit is often better than Chi-square test.
- K-S cannot handle heteroscedastic errors
- Anderson-Darling is better in handling the tail part of the distributions.
- K-S probabilities are incorrect if the model parameters are estimated from the same data.
- K-S does not work in more than one dimension.
- Bootstrap helps in the last two cases.

So far we considered model fitting part.

We shall now discuss model selection issues.
MLE and Model Selection

1. Model Selection Framework

2. Hypothesis testing for model selection: Nested models

3. Limitations

4. Penalized likelihood

5. Information Criteria based model selection
   - Akaike Information Criterion (AIC)
   - Bayesian Information Criterion (BIC)
Model Selection Framework (based on likelihoods)

- Observed data $D$
- $M_1, \ldots, M_k$ are models for $D$ under consideration
- Likelihood $f(D|\theta_j; M_j)$ and loglikelihood $\ell(\theta_j) = \log f(D|\theta_j; M_j)$ for model $M_j$.
  - $f(D|\theta_j; M_j)$ is the probability density function (in the continuous case) or probability mass function (in the discrete case) evaluated at data $D$.
  - $\theta_j$ is a $k_j$ dimensional parameter vector.
- Observed data $D$
- $M_1, \ldots, M_k$ are models for $D$ under consideration
- Likelihood $f(D|\theta_j; M_j)$ and loglikelihood $\ell(\theta_j) = \log f(D|\theta_j; M_j)$ for model $M_j$.
  - $f(D|\theta_j; M_j)$ is the probability density function (in the continuous case) or probability mass function (in the discrete case) evaluated at data $D$.
  - $\theta_j$ is a $k_j$ dimensional parameter vector.

**Example**

$D = (X_1, \ldots, X_n)$, $X_i$, i.i.d. $N(\mu, \sigma^2)$ r.v. Likelihood

$$f(D|\mu, \sigma^2) = (2\pi \sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right\}$$

Most of the methodology can be framed as a comparison between two models $M_1$ and $M_2$. 
The model $M_1$ is said to be nested in $M_2$, if some coordinates of $\theta_1$ are fixed, i.e. the parameter vector is partitioned as

- $\theta_2 = (\alpha, \gamma)$ and $\theta_1 = (\alpha, \gamma_0)$
- $\gamma_0$ is some known fixed constant vector.

Comparison of $M_1$ and $M_2$ can be viewed as a classical hypothesis testing problem of $H_0 : \gamma = \gamma_0$. 

Hypothesis testing is a criteria used for comparing two models. Classical testing methods are generally used for nested models.
The model $M_1$ is said to be nested in $M_2$, if some coordinates of $\theta_1$ are fixed, i.e. the parameter vector is partitioned as

- $\theta_2 = (\alpha, \gamma)$ and $\theta_1 = (\alpha, \gamma_0)$
- $\gamma_0$ is some known fixed constant vector.

Comparison of $M_1$ and $M_2$ can be viewed as a classical hypothesis testing problem of $H_0: \gamma = \gamma_0$.

**Example**

- $M_2$ Gaussian with mean $\mu$ and variance $\sigma^2$
- $M_1$ Gaussian with mean $0$ and variance $\sigma^2$

The model selection problem here can be framed in terms of statistical hypothesis testing $H_0: \mu = 0$, with free parameter $\sigma$.

Hypothesis testing is a criteria used for comparing two models. Classical testing methods are generally used for nested models.
Caution/Objections

- $M_1$ and $M_2$ are not treated symmetrically as the null hypothesis is $M_1$.

- Cannot accept $H_0$

- Can only reject or fail to reject $H_0$.

- Larger samples can detect the discrepancies and more likely to lead to rejection of the null hypothesis.
If $M_1$ is nested in $M_2$, then the largest likelihood achievable by $M_2$ will always be larger than that of $M_1$.

Adding a penalty on larger models would achieve a balance between over-fitting and under-fitting, leading to the so called Penalized Likelihood approach.

Information criteria based model selection procedures are penalized likelihood procedures.
Grounding in the concept of entropy, Akaike proposed an information criterion (AIC), now popularly known as Akaike Information Criterion, where both model estimation and selection could be simultaneously accomplished.
Grounding in the concept of entropy, Akaike proposed an information criterion (AIC), now popularly known as Akaike Information Criterion, where both model estimation and selection could be simultaneously accomplished.

AIC for model $M_j$ is $-2\ell(\hat{\theta}_j) + 2k_j$. The term $2\ell(\hat{\theta}_j)$ is known as the goodness of fit term, and $2k_j$ is known as the penalty.

The penalty term increase as the complexity of the model grows.
Grounding in the concept of entropy, Akaike proposed an information criterion (AIC), now popularly known as Akaike Information Criterion, where both model estimation and selection could be simultaneously accomplished.

AIC for model $M_j$ is $-2\ell(\hat{\theta}_j) + 2k_j$. The term $2\ell(\hat{\theta}_j)$ is known as the goodness of fit term, and $2k_j$ is known as the penalty.

The penalty term increase as the complexity of the model grows.

AIC is generally regarded as the first model selection criterion.

It continues to be the most widely known and used model selection tool among practitioners.
Advantages of AIC

- Does not require the assumption that one of the candidate models is the “true” or “correct” model.
- All the models are treated symmetrically, unlike hypothesis testing.
- Can be used to compare nested as well as non-nested models.
- Can be used to compare models based on different families of probability distributions.
Advantages of AIC

- Does not require the assumption that one of the candidate models is the “true” or “correct” model.
- All the models are treated symmetrically, unlike hypothesis testing.
- Can be used to compare nested as well as non-nested models.
- Can be used to compare models based on different families of probability distributions.

Disadvantages of AIC

- Large data are required especially in complex modeling frameworks.
- Leads to an *inconsistent model selection* if there exists a true model of finite order. That is, if $k_0$ is the correct number of parameters, and $\hat{k} = k_i \ (i = \arg \min_j (-2\ell(\hat{\theta}_j) + 2k_j))$, then $\lim_{n \to \infty} P(\hat{k} > k_0) > 0$. That is even if we have very large number of observations, $\hat{k}$ does not approach the true value.
Bayesian Information Criterion (BIC)

BIC is also known as the **Schwarz Bayesian Criterion**

$$-2\ell(\hat{\theta}_j) + k_j \log n$$

- BIC is consistent unlike AIC
- Like AIC, the models need not be nested to use BIC
- AIC penalizes free parameters less strongly than does the BIC
<table>
<thead>
<tr>
<th>Bayesian Information Criterion (BIC)</th>
</tr>
</thead>
</table>

BIC is also known as the **Schwarz Bayesian Criterion**

\[-2\ell(\hat{\theta}_j) + k_j \log n\]

- BIC is consistent unlike AIC
- Like AIC, the models need not be nested to use BIC
- AIC penalizes free parameters less strongly than does the BIC

- Conditions under which these two criteria are mathematically justified are often ignored in practice.
- Some practitioners apply them even in situations where they should not be applied.
Bayesian Information Criterion (BIC)

BIC is also known as the Schwarz Bayesian Criterion

$$-2 \ell(\hat{\theta}_j) + k_j \log n$$

- BIC is consistent unlike AIC
- Like AIC, the models need not be nested to use BIC
- AIC penalizes free parameters less strongly than does the BIC

- Conditions under which these two criteria are mathematically justified are often ignored in practice.
- Some practitioners apply them even in situations where they should not be applied.

Caution: Sometimes these criteria are multiplied by $-1$ so the goal changes to finding the maximizer.


