Bootstrapping, Crossvalidation and Discrete Data

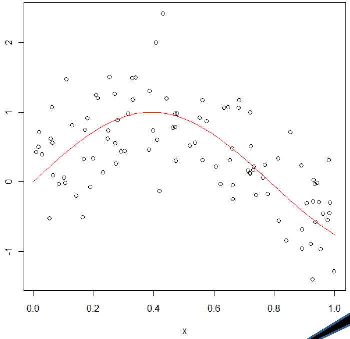
David B King, Ph.D.

May 12, 2015

Bootstrapping and Crossvalidation

Model Assessment, Bootstrapping and Crossvalidation

Bootstrapping and Crossvalidation



- x=runif(100)
- truey=sin(4*x)
- errors=rnorm(100,mean=0,sd=0.5)
- y=truey+errors
- plot(x,y)
- curve(sin(4*x),from=0,to=1,col="red")

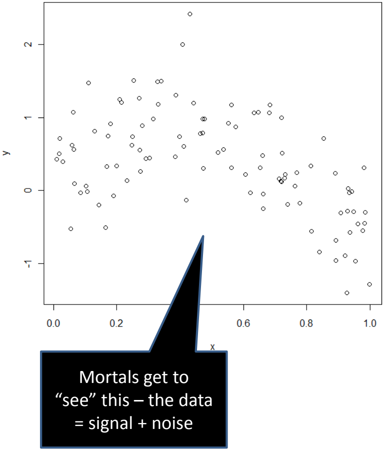
True Functional Relationship between X and Y

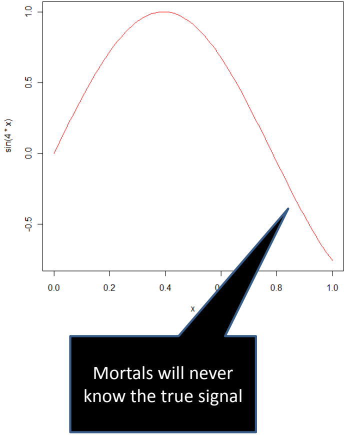
Y = f(X) + e

Experimental Errors

Game: Estimate the true relationship between X and Y

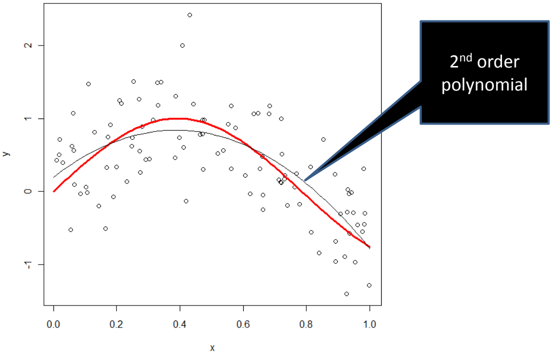
Bootstrapping and Crossvalidation





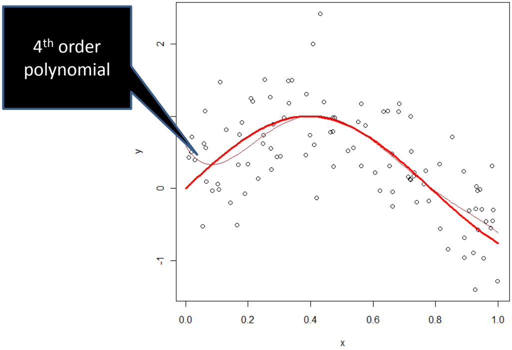
Bootstrapping and Crossvalidation

```
> fit1=lm(y~poly(x,degree=2),data=data)   # fit quadratic to data
> fit2=lm(y~poly(x,degree=3),data=data)   # fit 3rd order polynomial
> fit3=lm(y~poly(x,degree=4),data=data)
> fit4=lm(y~poly(x,degree=5),data=data)
```

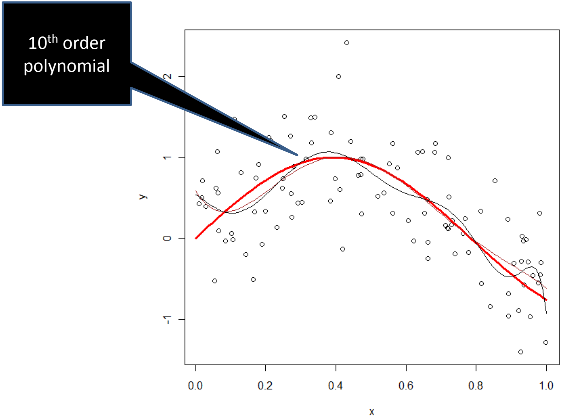


Bootstrapping and Crossvalidation

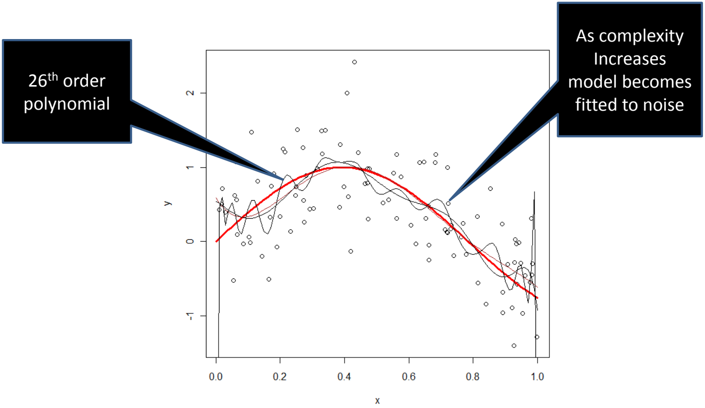
```
> fit1=lm(y~poly(x,degree=2),data=data)   # fit quadratic to data
> fit2=lm(y~poly(x,degree=3),data=data)   # fit 3rd order polynomial
> fit3=lm(y~poly(x,degree=4),data=data)
> fit4=lm(y~poly(x,degree=5),data=data)
```



Bootstrapping and Crossvalidation



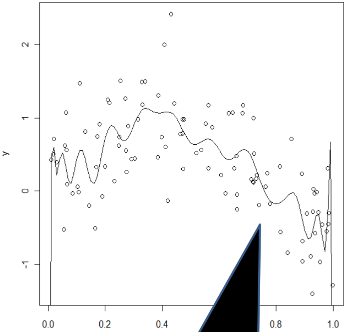
Bootstrapping and Crossvalidation



Bootstrapping and Crossvalidation

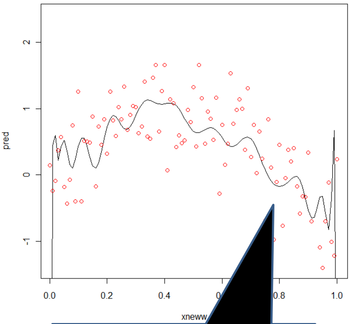
The Dangers of Over-fitting a Model

Training Data



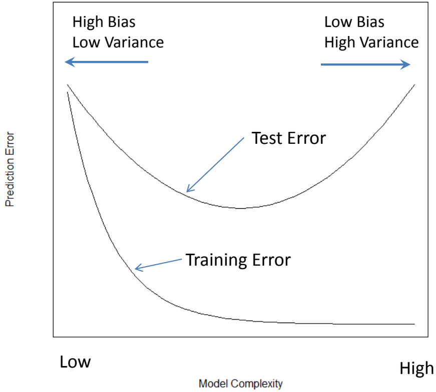
Because the complex Model was fit to every Every little quirky feature of the training data...

New Data



The complex model does a poor job of predicting new – as yet unseen – data coming from the process.

Bootstrapping and Crossvalidation



Training Err = MSE will always go down as model complexity increases

Bootstrapping and Crossvalidation

Ways of measuring error, using a "loss" function:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{Squared Error} \\ |Y - \hat{f}(X)| & \text{Absolute Error} \end{cases}$$

$$\text{Test Err} = Err = E\left[L(Y, \hat{f}(X)) \right]$$

Expectation taken over everything random, including X and Y as well as randomness
In the training sample which produced $\hat{f}(x)$

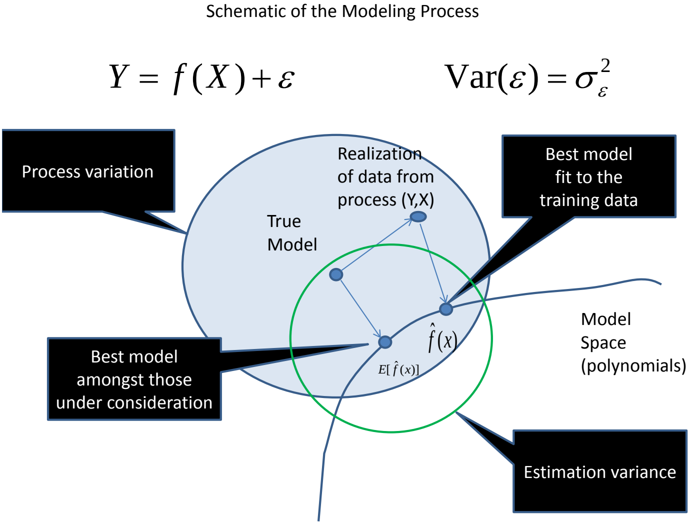
$$\text{Training Err} = err = \frac{1}{N} \sum_{i=1}^{N} \left[L(y_i, \hat{f}(x_i)) \right]$$

Empirical average of loss observed in the training data

$$\text{Training Err} < \text{Test Err}$$

Training error underestimates test error and does worse as model complexity grows

Bootstrapping and Crossvalidation



Bootstrapping and Crossvalidation

The Bias – Variance Decomposition

The Process: $$Y = f(X) + \varepsilon$$

Prediction error at X = x_0 =

$$
\begin{aligned}
Err(x_0) &= E\left[(Y - \hat{f}(x_0))^2 \mid X = x_0\right] \\
&= E\left[(Y - f(x_0))^2 \mid X = x_0\right] + \left[(f(x_0) - E(\hat{f}(x_0)))^2\right] + E\left[(\hat{f}(x_0) - E(\hat{f}(x_0)))^2 \mid X = x_0\right] \\
&= \sigma_\varepsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \text{Bias}^2 + Variance
\end{aligned}
$$

- First term is the variation of the target $f(x_0)$ around its true mean And never goes away no matter how well we estimate f.
- Second term measures deviation from the true function $f(x_0)$ to best estimable function $E[\hat{f}(x_0)]$
- Third term is the variation present in estimation

Model Complexity ↑ Bias ↓ Variance ↑

Bootstrapping and Crossvalidation

Extra-Sample Error, In-Sample Error, and Optimism

$$\text{Test Err} = Err = E\left[L(Y, \hat{f}(X))\right]$$

Test Err (or generalization error) is a kind of extra-sample error since the Test features occur at different X values than the samples in the training data.

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_y E_{Y^{New}}\left[L(Y_i^{New}, \hat{f}(X_i))\right]$$

In sample error measures the expected difference between N new responses Y_i^{New} At each of the training points x_i, i = 1, …, N

$$Op = Err_{in} - E_y[err]$$

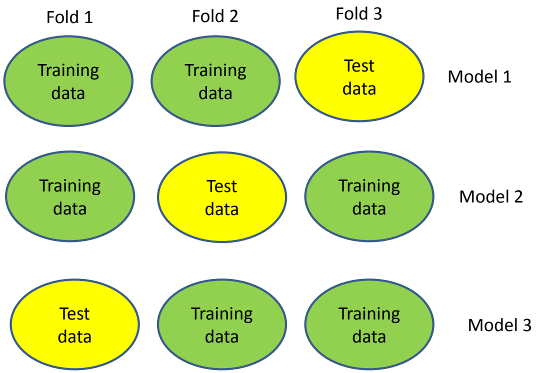
Optimism is the expected difference between In sample error and training error.

Cross-Validation and Bootstrapping measure Extra-Sample Error

Bootstrapping and Crossvalidation

K-Fold Cross Validation

Depiction of 3-fold cross validation:



Divide your data into K –folds, train a model on (K-1) folds and test model
On the remaining 1 fold

Bootstrapping and Crossvalidation

Big Idea in Cross Validation

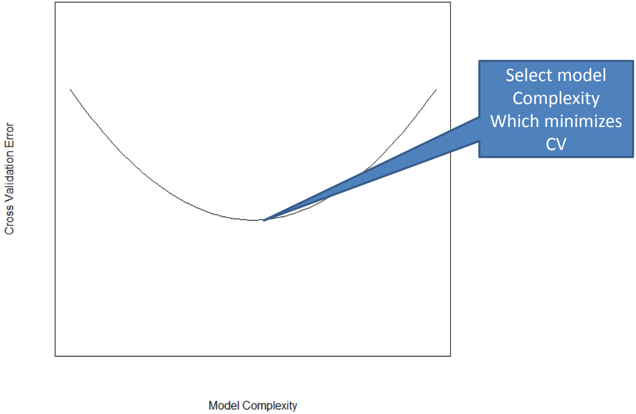
Because the Test data was not used to build each of the models,
The test data acts as NEW INDEPENDENT, YET TO BE OBSERVED DATA FROM THE PROCESS

Let $\hat{f}^{-K}(x)$ denote the fitted function with the Kth part of the data removed

$$CV = \frac{1}{KN}\sum_{i=1}^{K}\sum_{j=1}^{N}(y_i - \hat{f}^{-K}(x_i))^2$$

CV is an estimator of Extra-Sample Err

Bootstrapping and Crossvalidation



Bootstrapping and Crossvalidation

Write an R function which will compute CV

Bootstrapping and Crossvalidation

Bootstrapping

A Dataset:

Row	X	Y
1	A	F
2	B	G
3	C	H
4	D	I
5	E	J

Sample the rows 1:5 WITH REPLACEMENT randomly:

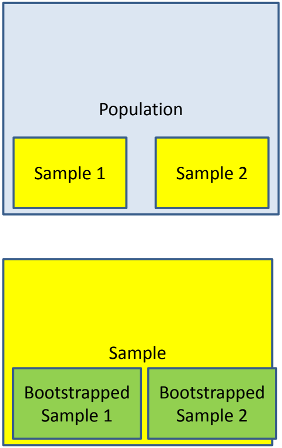
```
> sample(1:5,5,replace=TRUE)
[1] 5 1 2 2 1
```

The bootstrapped data is the dataset with the rows in the sample

Rows	X	Y
5	E	J
1	A	F
2	B	G
2	B	G
1	A	F

Bootstrapping and Crossvalidation

Idea behind Bootstrapping:



The process of repeatedly subsampling a sample should mimic the process of repeatedly drawing samples from a population.

Bootstrapping and Crossvalidation

Bootstrapping

Sample the rows 1:5 WITH REPLACEMENT randomly:

```
> sample(1:5,5,replace=TRUE)
[1] 5 1 2 2 1
```

The bootstrapped data is the dataset with the rows in the sample

Rows	X	Y
5	E	J
1	A	F
2	B	G
2	B	G
1	A	F

Rows	X	Y
3	C	H
4	D	I

OOB = out of bag

The rows which were selected are “in the bag”,
the rows not selected are “out of the bag”.

Bootstrapping and Crossvalidation

Uses for Bootstrapping

Bootstrapping can be used to measure the variance of any statistic S(Data)

$$\hat{V}(S(Data)) = \frac{1}{B-1}\sum_{i=1}^{B}(S(Data^{*i}) - \bar{S}^{*})^2$$

Bootstrapping can also be used to measure the bias of any statistic.

OOB data can be used to estimate extra-sample error.

If C_{-i} denotes the set of bootstrap samples which do not contain observation i

$$E\hat{rr} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{|C_{-i}|}\sum_{b \in C_{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

Jackknifing

From an historical standpoint leave one out crossvalidation, or n-fold crossvalidation was the first measure of out of sample error which was measured. Consider the general linear model

$$\mathbf{y}_{(i)} = \mathbf{X}_{(i)}\beta_{(i)} + \epsilon_{(i)}$$

formed by **deleting row** i from the model. Here $\mathbf{y}_{(i)}$ and $\epsilon_{(i)}$ are $(n-1)\times 1$ vectors, $\mathbf{X}_{(i)}$ is a $(n-1)\times p$ matrix and $\beta_{(i)}$ is $p\times 1$. What makes leave one out crossvalidation special is we know explicit formulas for how the prediction will change. One formula comes as a consequence of the Sherman-Morrison-Woodbury theorem we have

$$\begin{aligned}
(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1} &= (\mathbf{X}^T\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^T)^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}
\end{aligned}$$

Jackknifing

where $\mathbf{x}_i$ denotes the $p \times 1$ vector corresponding to the i^{th} row of $\mathbf{X}$. Now since $\mathbf{x}_i = \mathbf{X}^T \mathbf{e}_i$ with $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)^T$ denoting the i^{th} element of the standard basis in $\mathbb{R}^n$ it follows that

$$\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = \mathbf{e}_i^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e}_i = h_{ii}$$

hence

$$\begin{aligned}(\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} &= (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}}.\end{aligned}$$

It follows that

$$\begin{aligned}\hat{\beta}_{(i)} &= (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^T \mathbf{y}_{(i)} \\ &= \left[(\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}} \right] \left[\mathbf{X}^T \mathbf{y} - \mathbf{x}_i y_i \right]\end{aligned}$$

Jackknifing

Hence,

$$\begin{aligned}
\hat{\beta}_{(i)} &= \hat{\beta} + \left[\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}\right]\left[\mathbf{x}_i^T\hat{\beta} - y_i h_{ii}\right] - y_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \\
&= \hat{\beta} + \left[\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}\right]\left[\hat{y}_i - y_i h_{ii} - y_i(1-h_{ii})\right] \\
&= \hat{\beta} + \left[\frac{\hat{\epsilon}_i}{1-h_{ii}}\right]\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
\hat{\epsilon}_{i(i)} \equiv y_i - \hat{y}_{i(i)} &= y_i - \mathbf{x}_i^T\hat{\beta}_{(i)} \\
&= y_i - \mathbf{x}_i^T\hat{\beta} + \left[\frac{\hat{\epsilon}_i}{1-h_{ii}}\right]\left[\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\right] \\
&= \hat{\epsilon}_i + \left[\frac{h_{ii}\hat{\epsilon}_i}{1-h_{ii}}\right] = \frac{\hat{\epsilon}_i}{1-h_{ii}}.
\end{aligned}$$

PRESS

The PRESS residuals are defined by $\hat{\epsilon}_{i(i)} = \frac{\hat{\epsilon}_i}{1-h_{ii}}$ and the PRESS statistic is given by

$$PRESS = \sum_{i=1}^{n}(y_i - y_{i(i)})^2 = \sum_{i=1}^{n}(\hat{\epsilon}_{i(i)})^2 = \sum_{i=1}^{n}\left(\frac{\hat{\epsilon}_i}{1-h_{ii}}\right)^2.$$

A related statistic is given by

$$R^2_{PRESS} = 1 - \frac{PRESS}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Notice that since formulas are given for CV it saves us a lot of computation.

Scenario: We have:

- a sample of observations: $x_1, \ldots, x_n$
- a target parameter in mind: θ
  (e.g., σ, or $\log \sigma^2$, or e^{μ}, or σ/μ, ...)
- an estimate of θ, $\hat{\theta} \equiv y_{all}$
  e.g., s or $\log s^2$, or $e^{\bar{x}}$, or $0.7413\cdot$ F-spread/median ...

Questions:

1. What is a confidence interval for θ based on $\hat{\theta}$?
2. If $\hat{\theta}$ is biased for θ (i.e., if you repeatedly estimate θ using $\hat{\theta}$ and found that mean of $\hat{\theta}$'s is **not** θ), can we find an estimate of θ, say $\hat{\theta}^*$, which is less biased?

Answers:

1. Yes. Construct CIs using pseudo-values, or bootstrap estimates.
2. Yes. $\hat{\theta}_{JK}$ or $\hat{\theta}_B$.

We illustrate Jackknife approach first, then bootstrap.

Question 1: Confidence Interval

Q1: What is a confidnence interval for θ based on $\hat{\theta}$?

Ans: For "95% CI", expect answer is of the form

$$\hat{\theta} \pm t_{n-1}(0.975) \cdot [??]$$

where [??] is probably something like $\sqrt{var(\hat{\theta})}$ and $t_{n-1}(0.975)$ is 97.5%-point of Student's t, $n-1$ d.f.

- If $\hat{\theta} = \bar{x}$, then we use $\widehat{Var}(\hat{\theta}) = s^2/n$, because theory says $Var(\bar{x}) = \sigma^2/n$ and s^2 estimates σ^2:
- But if $\hat{\theta} = \log s^2$ or $e^{\bar{x}}$ or $0.7413 \cdot F - spread/median$?

Question 1: Confidence Interval

Situations where we *know* the answer:

1. Sample mean
2. LS coefficients (Gaussian error)

Cases where we *don't know* the answer: **Everything else**

1. Most nonlinear statistics (e.g., RRline)
2. Non-Gaussian error distributions
3. Long-tails

What can we do?

- If we had many samples, then we would have many $\hat{\theta}$'s, and then we could calculate a standard deviation of them, as a measure of the variability of the $\hat{\theta}$'s.
- Generally we do not have a lot of samples– we have only one, and only one $\hat{\theta}$ (y_{all}).
- We need to generate more $\hat{\theta}$'s.
- We do so by estimating θ on subsamples of original sample.
- Jackknife: subsample $= \{$all x's except one$\}$ (size $n-1$)
- Bootstrap: sample (with replacement) from the x's (size n)

Jackknife – notation

Jackknife notation:

- **Pseudo-values** $= ny_{all} - (n-1)y_{(-i)}$
- **Target parameter** $= \theta$
- **Estimate of target parameter** = statistic $S \equiv \hat{\theta}$
- $y_{all} \equiv \hat{\theta}_{all}$
- Mean of pseudo-values $\equiv \hat{\theta}_{JK}$ = jackknife estimate of θ

Jackknife – Example 1

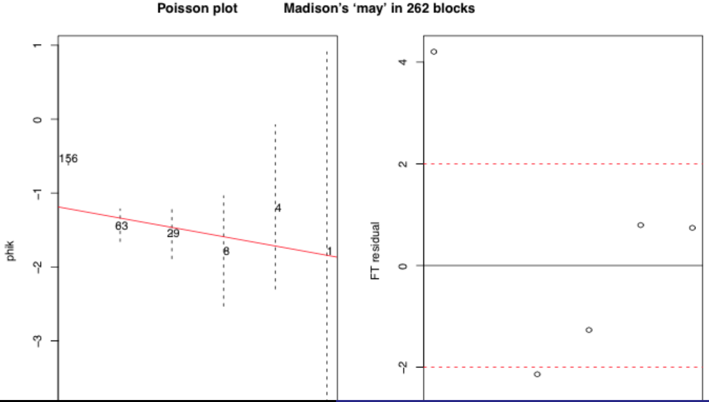
Example: **Sample mean, single batch** $x_1, \ldots, x_{20}$:
Confidence limits on θ using $S = y_{all} = \bar{x}$
($\bar{x} = 31.11826$, SE=SD/$\sqrt{20} = 4.09458$)

5.913934	17.225504	-0.525662	5.838253	16.726366
18.323374	20.972659	31.927836	20.437125	25.978228
26.039943	29.045681	42.689422	45.187752	47.038066
48.324157	53.771513	59.261140	51.370595	56.819336

```
 x    xx     xxxxx  xx x x      x  x xx xx x  x
  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
  0             20             40             60
```

Jackknife – Example 1

- $y_{all} = \text{mean}(y) = \sum_{i=1}^{20} x_i/20$
- $y_{(-i)} = \text{mean}(x \text{ without } x_i) = \sum_{k \neq i} x_k/19$
- Pseudo-values $= 20 \cdot y_{all} - 19 \cdot y_{(-i)}$:



Jackknife – Example 1

- average of pseudo-values $= 31.11826 = \bar{x}$
- Standard error of pseudo-values is $s/\sqrt{n} = 4.09458$

So, when statistic S is the sample mean,

- jackknife mean $=$ usual sample mean
- jackknife standard error $=$ usual standard error

i.e., results are expected.

Jackknife

Notes:

1. Mean of pseudo-values = "jackknife estimate of θ" ("*an estimate of much reduced bias*")
2. SE(pseudo-values) = "jackknife SE of $\hat{\theta}$"
3. **HW**: Show: SE(PVs) $= [(n-1)/\sqrt{n}] \cdot SD(y_{(-i)})$

Jackknife – procedure

Procedure: Statistic $S = \hat{\theta}$, $\hat{\theta}_{all} = y_{all}$

1. Calculate $\hat{\theta}_{all} = y_{all}$ using all data
2. Calculate same statistic without x_i : $y_{(-i)}$
3. Calculate pseudo-values $\hat{\theta}_i^* \equiv PV_i \equiv n y_{all} - (n-1) y_{(-i)}$
4. Calculate: $\hat{\theta}_{JK} = \sum \hat{\theta}_i^* / n$

$$\widehat{Var}(\hat{\theta}_{JK}) = (\text{SE}(\text{mean PVs}))^2 = \sum_{i=1}^{n} (\hat{\theta}_i^* - \hat{\theta}_{JK})^2 / [n(n-1)]$$

5. Calculate approximate 95% CI for θ using

$$\text{mean (PVs)} \pm t_{n-1}(0.975) \cdot \text{SE(mean PVs)}$$

$$= \hat{\theta}_{JK} \pm t_{n-1}(0.975) \cdot \sqrt{\widehat{Var}(\hat{\theta}_{JK})}$$

Jackknife – Example 2

Example: (Confidence interval on a standard deviation) A sample from a distribution produced the 11 values

$$0.1,\ 0.1,\ 0.1,\ 0.4,\ 0.5,\ 1.0,\ 1.1,\ 1.3,\ 1.9,\ 1.9,\ 4.7$$

There is no reason to suppose that the distribution is normal and some reason to suppose it is not.

- Sample standard deviation: y_{all}
- Leave-one-out: $y_{(-i)}$
- Pseudo-values: $\hat{\theta}_i^*$
- Average of pseudo-values: $\hat{\theta}_{JK}$
- SE($\hat{\theta}_{JK}$):
- 95% CI for σ:

Question 2: Reduce Bias

Q2: Is $\hat{\theta}_{JK}$ *"an estimate of much reduced bias"?*

Example: Use sample statistic s to estimate σ

- $E(s) \neq \sigma$; $E(s) = a(n) \cdot \sigma$ where $a(n) < 1$ (i.e., s slightly underestimates σ).
- $\lim_{n \to \infty} a(n) = 1$. So, as n gets large, the bias is negligible.
- But when $n = 5$, $E(s) = 0.8812\sigma$; i.e., s is about 12% too small.
- In general, $E(s) - \sigma \neq 0$, i.e., $E(s)$ is biased for σ.

We show that, in general, $\hat{\theta}_{JK}$ is less biased (or has no greater bias) for θ than $\hat{\theta} = y_{all}$ is.

Jackknife – Reduce Bias

- Suppose the bias has this form:

$$bias_n = E(\hat{\theta}) - \theta = a_1/n + a_2/n^2 + a_3/n^3 + \cdots$$

  i.e., bias is of order $1/n$.
- Then bias in $\hat{\theta}_{(-i)}$ ($n-1$ observations) is:

$$bias_n = E(\hat{\theta}_{(-i)}) - \theta = a_1/(n-1) + a_2/(n-1)^2 + a_3/(n-1)^3 + \cdots$$

- So bias in (average of leave-out-ones) $\equiv \hat{\theta}_L$ is

$$bias_n = E(\hat{\theta}_L) - \theta = (1/n) \cdot \sum_{i=1}^{n} E(\hat{\theta}_{(-i)} - \theta) = \sum_{k=1}^{\infty} a_k/(n-1)^k$$

Jackknife – Reduce Bias

- So bias in $\hat{\theta}_{JK}$ is

$$bias(\hat{\theta}_{JK}) = E(\hat{\theta}_{JK}) - \theta$$

$$= E[n(\hat{\theta}_{all} - \theta) - (n-1)(\hat{\theta}_L - \theta)]$$

$$= n(a_1/n + a_2/n^2 + \cdots) - (n-1)(a_1/(n-1) + a_2/(n-1)^2 + \cdots)$$

$$= (a_2/n - a_2/(n-1)) + a_3(1/n^2 - 1/(n-1)^2) + \cdots$$

$$= -a_2/(n(n-1)) - a_3(2n-1)/(n^2(n-1)^2) - \cdots$$

- Leading term is $a_2/(n(n-1)) \approx a_2/n^2$, of order $1/n^2$, less than order $1/n$: "*much reduced bias*".
- If $a_2 = a_3 = \cdots = 0$, then $\hat{\theta}_{JK}$ is **unbiased** for θ.

Jackknife – problem

- We can use the jackknife for any procedure — e.g., intercepts and slopes of RRline, effects from median polish, etc.
- Usually the jackknife over-estimates the standard error.
- But, more worrisome, sometimes it can under-estimate it.
- Efron (1979) asked himeself, “Why does the jackknife work?”
- In developing theory for it, he developed an alternative (And, in many ways, more intuitive) way of generating more $\hat{\theta}$'s
- So we next discuss Efron's bootstrap, then illustrate on
  — LS line and RR line
  — Median polish

About standard errors

Standard errors:

- SE=SD(statistic); e.g., SD(estimate)
- n observations $\rightarrow SD^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$
- For Gaussian$(0, \sigma^2)$: SD $\approx$ F-spread$/1.349$
- What happens when $n = 1$?

Jackknife and Bootstrap: Applications

RRline: We have only

- **one** intercept
- **one** slope

Median polish: We have only

- **one** M,
- **one** a_1, **one** a_2, ...
- **one** b_1, **one** b_2, ...

How to compute a SE when we have only one of each ??

Efron's bootstrap

- Ideally, we'd like to have another sample, so we can calculate another θ
- All we have is the sample at hand, which we hope is representative of the entire population. If the distribution of the quantity in the entire population is F, then we have only an $\hat{F}$, a sample from F
- We can use $\hat{F}$ to generate another sample, say $\hat{\hat{F}}$
- $\hat{\hat{F}}$ is obtained by take a random sample, with replacement, of the original x's.
- Note that some x_i will be duplicated, or triplicated, or ..., while other x_i's will not be represented at all.

Bootstrap – procedure

- Calculate $\hat{\theta}$ on this "bootstrap sample" from $\hat{F}$; denote it by $\hat{\theta}_b$, $b = 1, 2, \ldots, B$ (B can be very large).
- $\hat{\theta}_B = \sum_{b=1}^{B} \hat{\theta}_b / B \equiv$ "bootstrap estimate of θ"
- $\widehat{Var}(\hat{\theta}_B) \equiv \sum_{b=1}^{B} (\hat{\theta}_b - \hat{\theta}_B)^2 / (B - 1)$
- Approximate 95% CI: $\hat{\theta}_B \pm t_{n-1}(0.975) \cdot \sqrt{\widehat{Var}(\hat{\theta}_B)}$
- If $\hat{F}$ was not a good representation of true F, we are sorely out of luck

Bootstrap: How to obtain multiple samples/estimates?

We need to get more intercepts/slopes (RRline) or more Ms, a_is, b_js
Jackknife: Single sample
Bootstrap: More general

- Collect residuals in a pot
- Shuffle them around
- Put them back in "residual"
- Add back in the effects
- Now you have new set of data $\rightarrow$ recompute
- Get another set of estimates
- Repeat! Calculate standard errors

Jackknife and Bootstrap on LS and RRline

- $x = (1:20)$, $y =$ as before
- "True" line: $y = 1 + 3x + error$ (Gauusian, mean 0, SD=5)
- LS:$(\hat{a}_{LS}, \hat{b}_{LS}, RMS) = (0.1487, 2.9495, 5.412)$
- RR:$(\hat{a}_{RR}, \hat{b}_{RR}, |res|)$=(-1.06, 3.186, 91.9942)
- Jackknife estimate and jackknife SE:

	apv	bpv	rpv	Apv	Bpv	Rpv
mean	0.3389	2.9354	5.9777	-18.9427	5.7761	215.9494
SE	3.4719	0.2558	0.8624	1.6759	0.2102	13.7573

- LS theory: $\hat{SE} = RMS \cdot \sqrt{diag\{(X'X)^{-1}\}}$
  $SE(\hat{a}_{LS}) = 2.514$, $SE(\hat{b}_{LS})$= 0.210
- Jackknife SEs are generous when $\hat{\theta} = \hat{\theta}_{LS}$

Bootstrap: 3 approaches:

1. Sample indicies: repeat $B(200?)$ times:

```
ii <- sample(1:20, 20, replace=TRUE); xb <- x[ii];
yb <- y[ii]; lm(yb ~ xb); run.rrline(xb,yb)
```

2. Sample residuals, add back to original line:

```
res <- lm(y $\sim$ x)$res
for (j in 1:200) {
    b.res <- sample(res,20,replace=TRUE)
    yb <- 0.1487 + 2.9495*x + b.res
    b.coef <- lm(yb $\sim$ x)$coef
    [ save b.coef in file ]
    }
```

Depends critically on y being linear in x

Theory of Bootstrapping

- Statistics: Construct a statistic $T = f(\text{data})$ whose target of inference is an unknown parameter θ for a distribution function.
- Often, T is known to follow some distribution F.
- We can estimate F by use of the empirical CDF $\hat{F}$ via a function $t(\hat{F})$.
- Suppose, we want to calculate a $(1-2\alpha)$ confidence interval for θ.
- Often it is possible to show that $T \sim N(\theta+\beta, v)$ where v is the variance and β is the bias of T. I
- If both β and v are known then

$$P(T \leq t|F) \cong \Phi\left(\frac{t-(\theta+\beta)}{v^{1/2}}\right),$$

where $\Phi(\cdot)$ is the CDF of a standard normal.

Theory of Bootstrapping

If the α quantile of the standard normal distribution is $z_\alpha = \Phi^{-1}(\alpha)$, then an approximate $(1-2\alpha)$ confidence interval for θ has limits

$$(t - \beta - v^{1/2} z_{1-\alpha}, t - \beta - v^{1/2} z_\alpha),$$

where t is the assumed value of T as the above CI follows from

$$P(\beta + v^{1/2} z_\alpha \leq T - \theta \leq \beta + v^{1/2} z_{1-\alpha}) \cong 1 - 2\alpha.$$

The Two Flavors of Bootstrapping

- There are two different flavors of bootstrapping: **Parametric and Non-Parametric Bootstrapping**
- Parametric Bootstrapping: Suppose the CDF $F(\cdot|\theta)$ is known, then estimating the bias and variance of T can be accomplished quite easily.
- In this regard, suppose that $y_1, \ldots y_n$ are i.i.d. and that the CDF and PDF are $F(y|\theta)$ and $f(y|\theta)$, respectively.
- When we estimate θ with $\hat{\theta}$, it's substitution into the model gives the fitted model, with CDF $\hat{F}(y) := F(y|\hat{\theta})$ can be used to estimate the properties of T.
- We shall use Y^* to denote the random variable distributed according to the fitted model $\hat{F}$, and the notation E^* and Var^* will be used when these moments are calculated according to the fitted distribution.
- We may then generate many simulated data sets to estimate the bias and variance of T.

Parametric Bootstrapping

To illustrate the computation of the bias and variance of T from a single data set, we let $Y_1^*, \dots Y_n^*$ be a independently drawn sample from the fitted distribution $\hat{F}$. When the statistic of interest is calculated from such a simulated data set, we denote it by T^*. From R repetitions of the data simulation we obtain $T_1^*, \dots T_R^*$. The estimator of the bias $b(F) = \mathrm{E}[T|F] - \theta$ of T is then

$$\hat{B} = b(\hat{F}) = \mathrm{E}[T|\hat{F}] - t = \mathrm{E}^*[T^*] - t,$$

and this in turn is estimated by

$$\hat{B}_R = \frac{1}{R}\sum_{r=1}^{R} T_r^* - t = \bar{T}^* - t.$$

Note that in the simulation, t is the parameter value of the model, so that $T^* - t$ is the simulation analogue of $T - \theta$. The corresponding estimator of the variance of T is then

$$\hat{V}_R = \frac{1}{R-1}\sum_{r=1}^{R}(T_r^* - \bar{T}^*)^2,$$

and estimators for other moments can be made in similar fashion.

Non-Parametric Bootstrapping

- In many cases we have no parametric model, but we can assume $Y_1, \ldots Y_n$ are independently and identically distributed according to an unknown distribution function F. We use the **empirical CDF** $\hat{F}$ to estimate the unknown CDF F. To estimate the properties of T then, we utilize $\hat{F}$ just as we would in the parametric model when drawing simulated samples.
- Because the empirical CDF $\hat{F}$ puts equal probabilities on the original data values $y_1 \ldots, y_n$, each Y^* is independently sampled uniformly from these values.

Non-Parametric Bootstrapping

- Therefore each sample $Y_1^*, \ldots, Y_n^*$ is a random sample taken with replacement from the data.
- We may then repeat such sampling R times and estimate the bias of T by

$$\hat{B}_R = \frac{1}{R}\sum_{r=1}^{R} T_r^* - t = \bar{T}^* - t,$$

and the variance with

$$\hat{V}_R = \frac{1}{R-1}\sum_{r=1}^{R}(T_r^* - \bar{T}^*)^2,$$

which are the same formulae as in the parametric case, with the only exception being that in the non-parametric case the samples are drawn in different fashion.

Non-Parametric Bootstrapping

- Suppose that T estimates θ and we seek a confidence interval on θ with both left- and right-tail errors both equal to α. If the quantiles of $T - \theta$ are denoted a_p, we have
$$P(T - \theta \leq a_\alpha) = \alpha = P(T - \theta \geq a_{1-\alpha}).$$
Rewriting the events $T - \theta \leq a_\alpha$ and $T - \theta \geq a_{1-\alpha}$ as $\theta \geq T - a_\alpha$ and $\theta \leq T - a_{1-\alpha}$, respectively, we see that the $(1 - 2\alpha)$ equi-tailed confidence interval has limits
$$(\hat{\theta}_\alpha := t - a_\alpha, \hat{\theta}_{1-\alpha} := t - a_{1-\alpha}).$$
- This ideal solution to the confidence interval rarely applies because the distribution of $T - \theta$ is usually unknown. This leads us to several approximate methods, most of which are based off approximating the quantiles of $T - \theta$.

Normal Approximation Method

- The simplest approach is to apply a $N(\beta, v)$ approximation for $T - \theta$. This leads to approximate confidence limits given by

$$\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha} = t - \hat{B}_R \mp \hat{V}_R^{1/2} z_{1-\alpha},$$

  where $\hat{B}_R$ and $\hat{V}_R$ are calculated with the bias and variance formulas in the previous slide.
- Whether or not a normal approximation method is appropriate can be assessed through making a Q-Q plot of the simulated estimates $t_1^*, \ldots, t_R^*$. If such a plot suggests that the normal approximation is poor, then we can either try to improve the approximation in some way or replace it completely.

Normal Approximation Method

- If we start again at the general confidence interval formula, we can estimate the quantiles a_α and $a_{1-\alpha}$ by the corresponding quantiles of $T^* - t$. Assuming that the R simulations result in $t^*_{(1)} - t, \ldots, t^*_{(R)} - t$, ordered realizations of $T^* - t$, then the respective quantiles can be approximated by $a_\alpha = t^*_{((R+1)\alpha)} - t$ and $a_\alpha = t^*_{((R+1)(1-\alpha))} - t$, respectively.
- By substituting these values into the confidence interval for θ this results in the lower and upper confidence limits on θ given by

$$\hat{\theta}_\alpha = 2t - t^*_{((R+1)(1-\alpha))}, \quad \text{and} \quad \hat{\theta}_{1-\alpha} = 2t - t^*_{((R+1)\alpha)}.$$

- These are referred to as the basic bootstrap confidence limits for θ.

Studentized Bootstrap Method

- A modification of this is to use the form of the normal approximation confidence limit in (57), by replacing the $N(0,1)$ approximation for $Z = (T - \theta)/V^{1/2}$ by a bootstrap approximation.
- In this method, each simulated sample is used to calculate t^*, the variance estimate $\hat{V}^*$, and hence the bootstrap version $z^* = (t^* - t)/(\hat{V}^*)^{1/2}$ of Z.
- We note that in order to calculate $\hat{V}^*$ for each simulated sample for example, this method requires that a bootstrap be done for each simulated sample, or a bootstrap performed for each bootstrap if you will. Once the R simulated values of z^* are computed, they are ordered, and the p^{th} quantile of Z is estimated by the $(R+1)p^{th}$ ordered value of these.
- Then the confidence limits are replaced by

$$\hat{\theta}_\alpha = t - (\hat{V})^{1/2} z^*_{((R+1)(1-\alpha))}, \quad \text{and} \quad \hat{\theta}_{1-\alpha} = t - (\hat{V})^{1/2} z^*_{((R+1)\alpha)}.$$

- These are referred to as the studentized bootstrap confidence limits for θ.

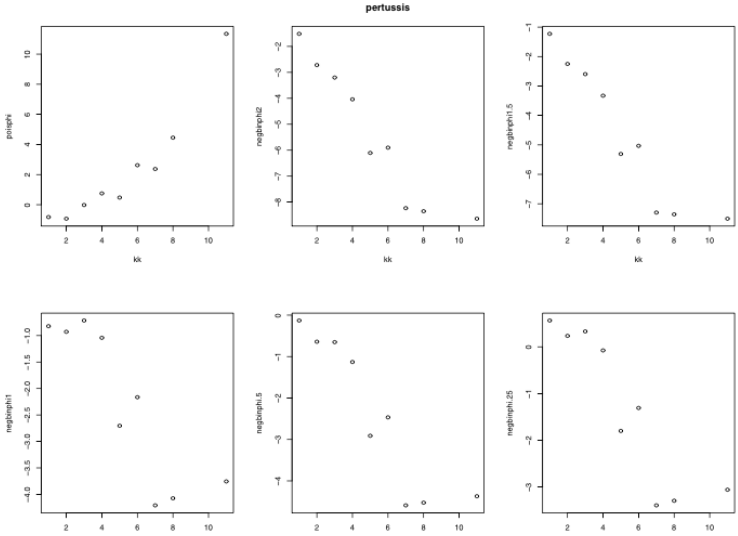
Example

3. Fit distribution to residuals and sample from it:
Ex: mean(residuals) = 0, SD = RMS = 5.412, N(0, sd=5.412):

```
for ( j in 1:200)  {
     b.res <- 5.412*rnorm(20)
     yb <- 0.1487 + 2.9495*x + b.res
     b.coef <- lm(yb $\sim$ x)$coef
     [ save b.coef in file ]
     }
```

Depends critically on y being linear in x **and** distribution of residuals in Gaussian

Example

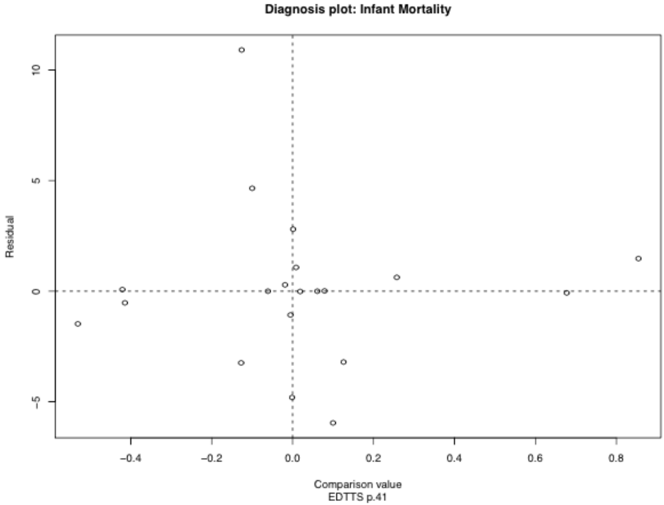

pertussis
poisphi
negbinphi2
negbinphi1.5
negbinphi1
negbinphi.5
negbinphi.25
kk

Example

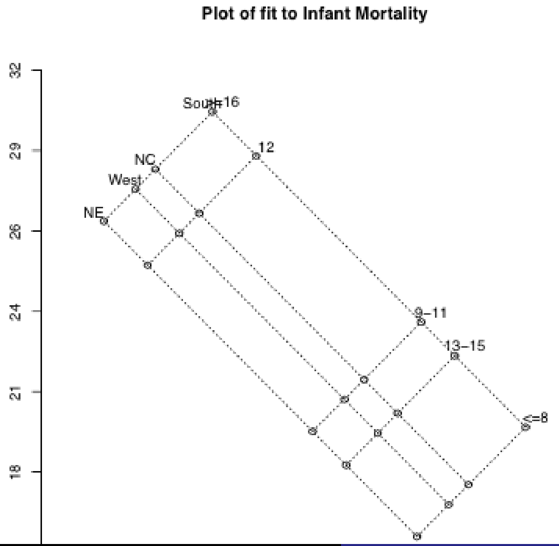
Median polish of infant mortality rates:

	Father's education					
Region	<=8	9-11	12	13-15	>=16	Row
NE	-1.475	0.075	0.012	-1.075	0.625	-1.475
NC	1.475	-0.075	-3.237	1.075	-0.525	2.375
South	10.900	4.650	-0.012	-4.800	0.000	-0.350
West	-3.200	-5.950	0.288	2.800	0.000	0.350
Col	7.475	5.925	-1.113	0.075	-3.625	20.775

Example


Diagnosis plot: Infant Mortality
Residual
10
5
0
−5
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Comparison value
EDTTS p.41

Example


Plot of fit to Infant Mortality
32
29
26
24
21
18
South
≥16
12
NC
West
NE
9–11
13–15
≤8

Example

	M	a1	a2	a3	a4
mean	21.211	-1.629	2.304	-0.515	0.237
SD	1.064	1.328	1.331	1.148	1.158

	b1	b2	b3	b4	b5
mean	6.979	5.442	-1.254	-0.508	-3.891
SD	1.941	1.917	1.403	0.943	1.840

EDTTS, Ch9

- "A Poissonness Plot": Count data (unbounded)
  #bufferflies/region, #accidents/month, #calls/week, ...
  (D.C. Hoaglin, *The American Statistician* 1980, 146–149)
- Binomial plot: #successes (failures) in fixed # of trails
  #women/#faculty; #rooms in use/total # rooms;
  #death/#people
  (NB: $n \uparrow$, $p \downarrow$ so $np = \lambda$ fixed $\Rightarrow$ binomial$(n, p) \sim$ Poisson(λ))
- Negative Binomial plot: # failures before nth success
- Logarithmic series plot: model for counts in ecology

Poisson Example 1: Data

"A Poissonness Plot"

Motivating data: London bombs hits, WWII

- City of London: 6 sq miles, divided into 1/4-sq miles blocks

$$24 \times 24 = 576 \text{ blocks}$$

- Count the #blocks with 0 bomb hits: 229
- Count the #blocks with 1 bomb hits: 211
- Etc: $k, n_k, \hat{p}_k \equiv n_k/576$:

k	0	1	2	3	4	5	6	7
n_k	229	211	93	35	7	0	0	1
$\hat{p}_k$	0.40	0.37	0.16	0.06	0.01	0.00	0.00	0.002

Models: $p_k = \mathrm{P}\{k \text{ hits}\} = e^{-\lambda}\lambda^k/k!$

Poisson Example 1: Poissonness plot

N = total #observations (here, $N = 576$)

$$E(n_k) = \eta_k = N \cdot p_k = N \cdot \mathrm{P}(X = k) = N \cdot e^{-\lambda}\lambda^k/k!$$

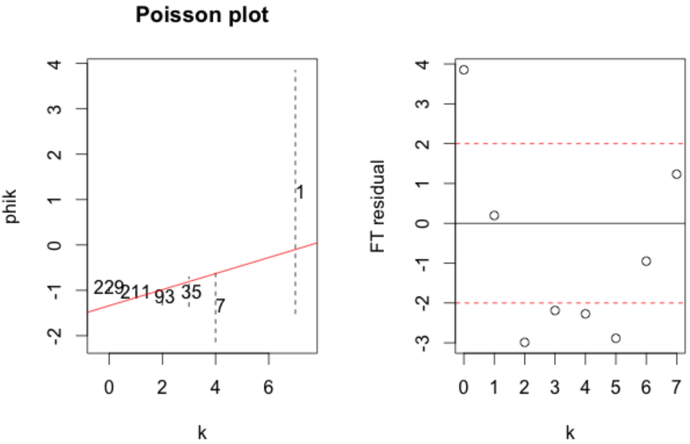
$$\Rightarrow \log(n_k) \approx \log(\eta_k) = \log(N) - \lambda + k\log(\lambda) - \log(k!)$$

$$\Rightarrow \phi_k \equiv \log(n_k) - \log(N) + \log(k!) \approx -\lambda + k\log(\lambda)$$

- Plot ——— (y-axis) vs ——— (x-axis)
- intercept =
- slope =
- often slope is better estimate... less variance
- $\log(n_k)$ is undefined when $n_k = 0$; do not plot

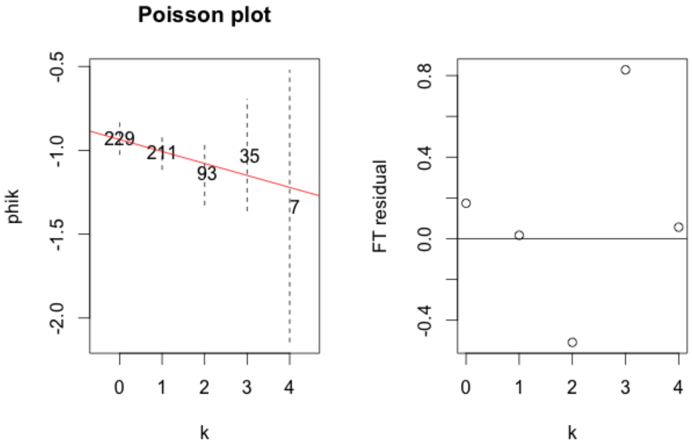
Poisson Example 1: Result (a)

R code: `poisplot(k,nk)`



Poisson Example 1: Result (b)

R code: `poisplot(k,nk,1:5)`



Poisson Example 1: Variability of estimates

London bomb data:

- $\lambda = 0.9357$ (intercept), $e^{-0.07142} = 0.9323$ (slope)
- average: $\hat{\lambda} = 0.934$
- If very different, weight slope estimate more heavily:

$$Var(int_{LS}) = \sum x_i^2 / den, Var(slope_{LS}) = n/den$$

  $(den = n(n-1)s_x^2)$, so weights $\propto 1/\sum x_i^2$, $1/n$
  (less drastically: $\sqrt{1/\sum x_i^2}, 1/\sqrt{n}$)
- For London Bomb data, $SE(int) \approx 2.45 SE(slope)$ (variances 30/50 and 5/50); weights 1:6 or 1:2.45

Poisson Example 1: FT residuals

Compare observed & expected via Freeman-Tukey residuals =

$$FT_k \equiv \sqrt{4 \cdot obs_k + 2} - \sqrt{4 \cdot exp_k + 1},$$

$$exp_k = 576 \cdot e^{-0.934} 0.934^k / k!$$

k	0	1	2	3	4	5	6	7
obs_k	229	211	93	35	7	0	0	1
exp_k	226.4	211.4	98.7	30.7	7.2	1.3	0.2	0.03
FT_k	0.19	-0.01	-0.56	0.78	0.03	-1.52	-0.35	1.40

Formula for Freeman-Tukey Residuals (theory)

Reference: M. F. Freeman, J. W. Tukey (1950), "Transformations related to the angular and the square root", *Annals of Mathematical Statistics* 21:607–611.

- Recall: $Z \sim N(0,1)$ (standard Gaussian) $\Rightarrow Z^2 \sim \chi^2_1$
- Freeman & Tukey (1950): $X \sim \mathit{Poisson}(\lambda)$ (mean, var=?)

$$\Rightarrow \sqrt{X} + \sqrt{X+1} \sim N(\sqrt{\lambda} + \sqrt{\lambda}, 1) \quad \text{(constant var)}$$

- $\sqrt{X} + \sqrt{X+1} \approx \sqrt{X+1/2} + \sqrt{X+1/2}$
  (added 0.5 to first X; subtracted 0.5 from second $X+1$)

Formula for Freeman-Tukey Residuals (theory)

- $\sqrt{X+1/2}+\sqrt{X+1/2}=\sqrt{4X+2}$
- $\sqrt{4X+2} \approx\sim N(\sqrt{4\lambda+2},1)$
- But not quite: Jensen's inequality: $E(g(X)) < g(E(X))$ if $g(\cdot)$ is concave (as is $g(x)=\sqrt{x}$)
- So slightly better: $\sqrt{4X+2} \approx\sim N(\sqrt{4\lambda+1},1)$

$$\Rightarrow FTres \equiv \sqrt{4X+2}-\sqrt{4\lambda+1} \approx\sim N(0,1)$$

- $(FTres)^2 \approx\sim \chi^2_1$, $sum \approx\sim \chi^2_{k-1}$ ($k = \#$ categories)

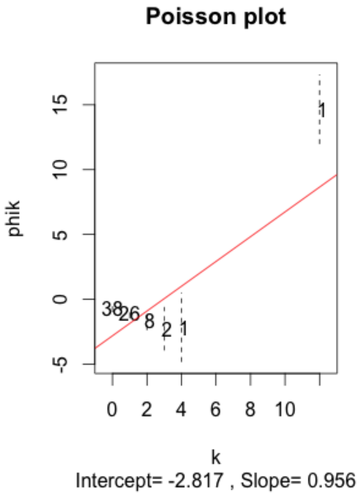
Poisson Example 2: Data

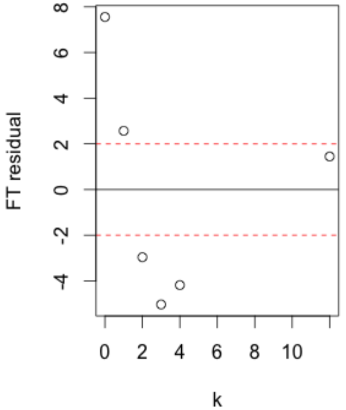
EDTTS, p.351: Incidents of international terrorism in U.S., Jan 1968 — Apr 1974 (Jenkins & Johnson 1975)

k	0	1	2	3	4		...	12
n_k	38	26	8	2	1	...0...		1

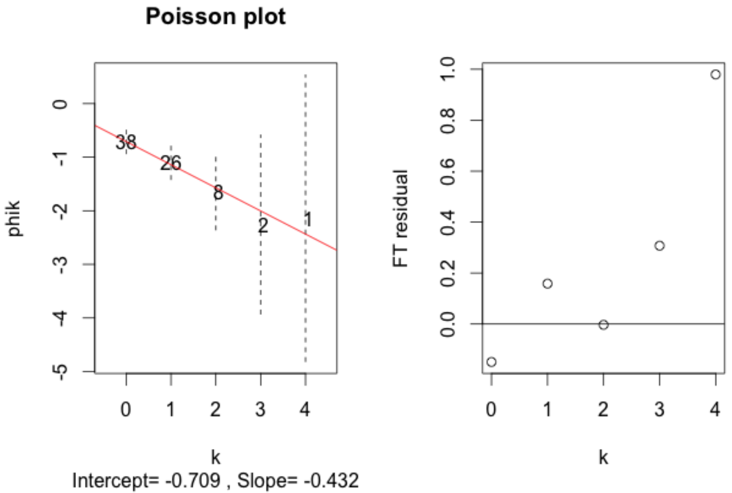
- $N = 75$ months (Jan'68 — Apr'74)
- Most months had 0 or 1 incident
- one month (July 1968) had 12 incidents
- 11 of the 12 July'68 incidents attributed to *El Poder Cubano* ("Cuban Power") anti-Castro group
- underlying assumption (individual Poisson occurences are independent) not satisfied for July'68 ($n_{12} = 1$); omit

Poisson Example 2: Result (a)





Poisson Example 2: Result (b)



Poisson Example 2: Estimate and Residuals

$\hat{\lambda} = 0.71$ (intercept) or $e^{-0.432} = 0.65 \Rightarrow \hat{\lambda} \approx 0.68$
Compare observed and expected:
Freeman-Tukey residuals $= \sqrt{4 \cdot obs_k + 2} - \sqrt{4 \cdot exp_k + 1}$
$exp_k = 76 \cdot e^{-0.68} 0.68^k / k!$

k	0	1	2	3	4	5	6	7
obs_k	38	26	8	2	1	0	0	0
exp_k	38.0	25.8	8.8	2.0	0.3	0.05	0.01	0.00
FT_k	0.04	0.08	-0.18	0.17	0.92	-0.09	-0.01	0.00

(For comparisons, all Freeman-Tukey residuals for $\hat{\lambda} = 0.66$, 0.67, 0.68, 0.69, 0.70 were very similar.)

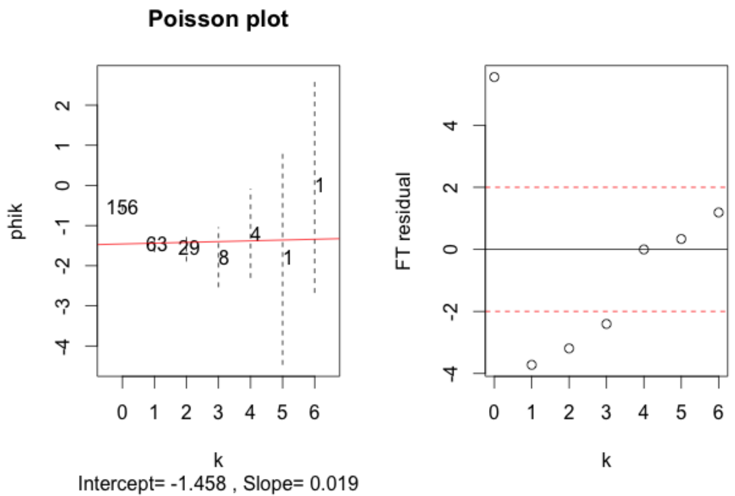
Poisson Example 3: Data

James Madison's *The Federalist* (Mosteller & Wallace 1964)

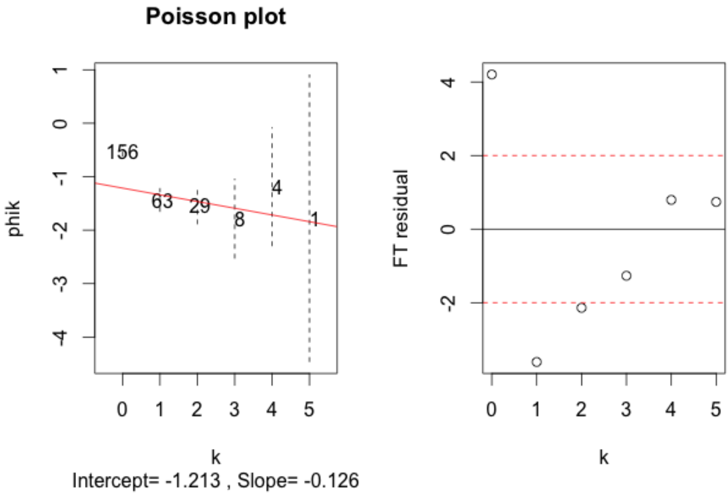
- Initial series of 77 essays
- Jay: 5; Hamilton: 43; Madison: 14; Hamilton & Madison: 3
- Disputed authorship: 12
- Mosteller & Wallace 1964: 12 disputed = Madison
- Conclusion based in part on use of certain words
- Frequency distribution of *may* in 262 blocks of text

k	0	1	2	3	4	5	6
n_k	156	63	29	8	4	1	1

Poisson Example 3: Result (a)



Poisson Example 3: Result (b)



Binomial Example: Data

Binomial plot

- Motivating example: 12 seats in First Class, B-737
- Stewardess: "Holy cow, it's all women!"
- KK: "Is that unusual?" — "Even half is unusual!"
- # of the 12 seats occupied by women on 100 B-737 ORD-DCA flights ($N = 100$)

k	0	1	2	3	4	5	6
n_k	1	3	4	23	25	19	18
k	7	8	9	10	11	12	
n_k	5	1	1	0	0	0	

Binomial Example: Binomial plot

- Possible values are 0, 1, 2, ..., 12
- Binomial distribution: Probability that k of the 12 seats are occupied by women:

$$\mathrm{P}\{X = k\} = C(n,k)p^k(1-p)^{n-k}, \quad C(n,k) = n!/[k!(n-k)!]$$

- Same approach as before: Compare expected counts η_k (from binomial distribution with $n = 12$ and estimated p) with observed n_k:

$$\eta_k = N \cdot \mathrm{P}\{X = k\} = N \cdot C(n,k)p^k(1-p)^{n-k} \approx n_k$$

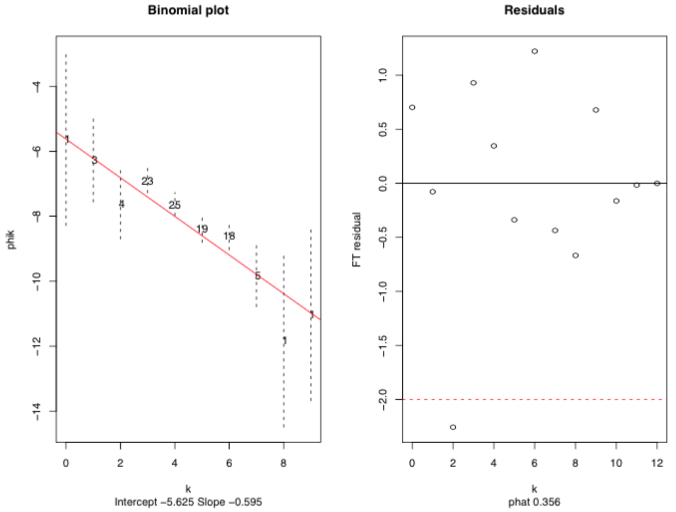
$$\Rightarrow n_k \approx N \cdot C(n,k) \cdot (p/(1-p))^k \cdot (1-p)^n$$

$$\Rightarrow \phi_k \equiv \log(n_k) - \log(N) - \log(C(n,k)) \approx$$

$$k \cdot \log(p/(1-p)) + n \log(1-p)$$

Slope = ______, Intercept = ______

Binomial Example: Result



Negative Binomial Plot

Negative Binomial (NB) distribution

- Motivating example: Madison frequency of the word "*may*"
- When Poisson doesn't quite fit, sometimes NB does
- "Over-dispersed Poisson" or # failures before n^{th} success

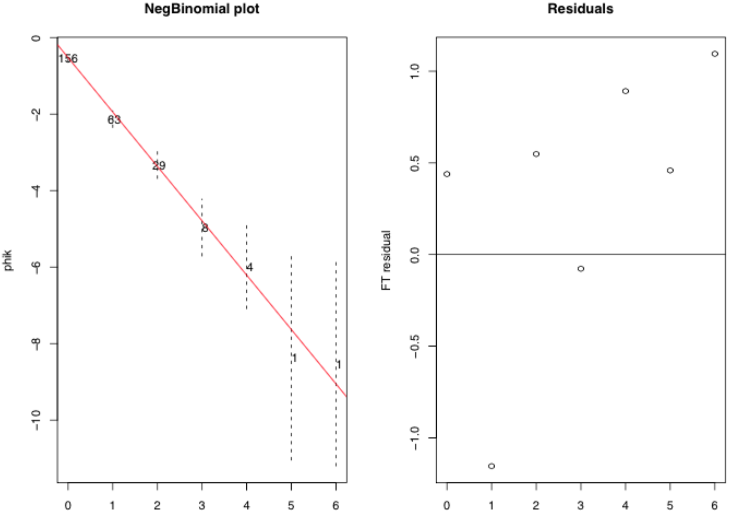
$$P(X = k) = \binom{n+k-1}{k} p^n (1-p)^k$$

- When $n = 1$ NB = geometric distribution $\propto (1-p)^k p$
- Take same approach as before:

$$\phi_k \equiv \log(n_k) - \log(N) - \log\left[\binom{n+k-1}{k}\right] \approx n\log(p) + k\cdot\log(1-p$$

Intercept =———, Slope=———

Negative Binomial Plot: Result



Negative Binomial Example

- Madison *may* frequencies: NB($n = 2$, $p = 0.76$)
- Interpretation: # words before 2nd occurrence of *may* $\approx$ negative binomial, probability of *may* $= 0.76$
- FT-residuals, $n = 2$, $p = 0.76$:

k	0	1	2	3	4	5	6
n_k	0.40	-1.14	0.59	-0.04	0.92	-0.97	-0.34

Logarithmic Series plot

- Model for ecological observations
- Sir Ronald Fisher: $butterfly species (EDTTS, p.385)

$$\mathrm{P}\{X = k\} = \alpha \cdot \theta^k / k, \quad 0 < \theta < 1, \quad \alpha = [-\log(1-\theta)]^{-1}$$

$$\Rightarrow \phi_k \equiv \log(k \cdot n_k / N) \approx -\log[-\log(1-\theta)] + k \cdot \log(\theta)$$

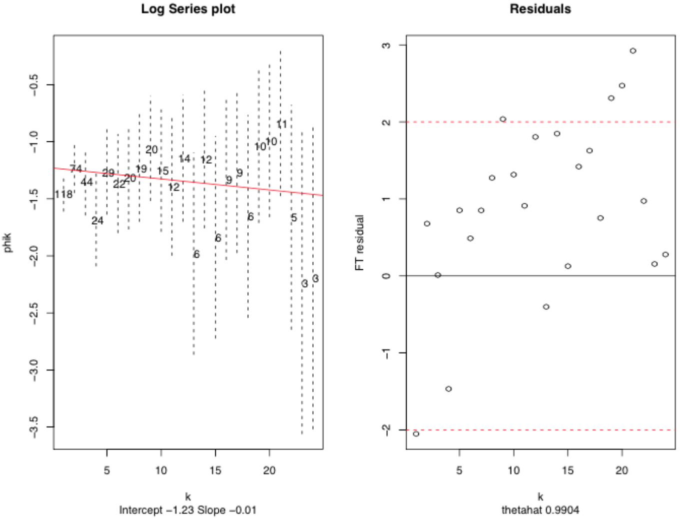
- Plot ———— (y-axis) vs ———— (x-axis)
- intercept =
- slope =

Logarithmic Series Example: Data

- Data: 501 species of butterflies (EDTTS, p.385)
- $n_k = \#$ of butterfly species for which k individuals were collected
- Ex: for 118 of the 501 species, only 1 individual was collected
- Ex: for 20 of the 501 species, 7 individuals were collected

k	1	2	3	4	5	6	7	8	9	10	11	12
n_k	118	74	44	24	29	22	20	19	20	15	12	14
k	13	14	15	16	17	18	19	20	21	22	23	24
n_k	6	12	6	9	9	6	10	10	11	5	3	3

Logarithmic Series Example: Result



A slight modification to n_k

A slight modification to n_k

- We would like approximately symmetric "confidence intervals" on our ϕ_k ordinates (y-axis)
- Randomness in ϕ_k is usually in $\log(n_k)$
- Small $n_k \Rightarrow \log(n_k)$ not symmetric, and $\log(\eta_k)$ not the center
- Could transform $\log(n_k)$ so it is more symmetric about transformed $\log(\eta_k)$ (not easy; changes interpretation)
- Easier: Modify n_k slightly to n_k^*:

$$n_k^* = 1/e = 0.36788 \text{ when } n_k = 1$$

$$n_k^* = n_k - 0.67 - 0.8 n_k / N \text{ when } n_k > 1$$

A slight modification to n_k

$$n_k^* = 1/e = 0.36788 \text{when } n_k = 1$$

$$n_k^* = n_k - 0.67 - 0.8 n_k/N \text{when } n_k > 1$$

Approximate CIs for $\log(n_k^*)$:

$$\log(n^*) \pm 1.96\sqrt{1-\hat{p}_k}/\sqrt{n_k - \sqrt{n}_k \cdot (0.47 + \hat{p}_k/4)}$$

where $\hat{p}_k = n_k/N$

See eqn(10), EDTTS p.365

As before, do not plot point if $n_k = 0$.

Alternative: Plot Frequency ratios

Plot Frequency ratios: Poisson

- Recall Poisson probabilities:
  $P(X = k) \equiv p_\lambda(k) = \exp(-\lambda)\lambda^k/k!$
- So $p_\lambda(k)/p_\lambda(k-1) = [e^{-\lambda}\lambda^k/k!]/[e^{-\lambda}\lambda^{k-1}/(k-1)!] = \lambda/k$

$$\Longrightarrow kp_\lambda(k)/p_\lambda(k-1) = \lambda \approx kn_k/n_{k-1}$$

- Plot kn_k/n_{k-1} vs k
- Slope $= 0$; Intercept $= \lambda$
- Madison *may* (EDTTS p.392)

Plot Frequency ratios: Other discrete distributions

Frequency ratio plots work also for:

- Binomial:

$$\frac{k \cdot b_k(p)}{b_{k-1}(p)} = \frac{k \cdot C(n,k)p^k(1-p)^{n-k}}{C(n,k-1)p^{k-1}(1-p)^{n-k+1}} = \frac{(n+1)p}{(1-p)} - k\frac{p}{(1-p)}$$

- Negative binomial: $k \cdot B_k(p)/B_{k-1}(p) =$

$$k \cdot C(n+k-1,k)p^n(1-p)^k / C(n+k-2,k-1)p^n(1-p)^{k-1}$$

$$= (n-1)(1-p) - (1-p) \cdot k$$

- Logarithmic series: $k \cdot L_k(\theta)/L_{k-1}(\theta) =$

$$k \cdot \theta^k / [-k \ln(1-\theta)] / \theta^{k-1} / [-(k-1)\ln(1-\theta)] = -\theta + \theta \cdot k$$

Problems with Frequency ratio plots

- Less resistant: One discrepant n_k affects .... ?
- What happens when $n_k = 0$, or, worse, $n_{k-1} = 0$?
- $Var(kn_k/n_{k-1})$ **not constant!**
- Large variability when n_{k-1} is small

Conclusion: Use approach based directly on probability distribution, not ratio of successive probabilities

R code

R code

- Uses n_k^* so CI for $\log(n_k^*)$ is roughly symmetric about $\log(\eta_k)$; i.e., CI for $\log(\eta_k) - \log(n_k^*)$ is roughly symmetric about 0
- For $\hat{p}_k = n_k/N$, approximately 95% CI for $\log(\eta_k)$:

$$\log(n_k^*) \pm 1.96 \cdot \sqrt{(1 - \hat{p}_k)/(n_k - (0.47 + 0.25 \cdot \hat{p}_k)\sqrt{n_k})}$$

R code

```
poisplot <- function(k,nk,which) {
  lenk <- length(k)
  if(missing(which))     which <- (1:lenk)
  k0 <- k[which]
  nk0 <- nk[which]
  k1 <- k0[nk[which] > 0]
  nk1 <- nk0[nk0 > 0]
  N <- sum(nk1)

  # Modification to n_k, used in CI
  nk2 <- nk1
  nk2[nk1==1] <- exp(-1)
  nk2[nk1 > 1] <- (nk1[nk1 > 1])*(1 - 0.8/N) - 0.67
  phik <- log((gamma(k1 + 1))*nk2/N)
  rr <- run.rrline(k1,phik)
```

R code

```
pkhat <- nk1/N
 cilim <- 1.96*sqrt((1-pkhat)/(nk1-(.47+.25*pkhat)*sqrt(nk1)))
rng <- range(c(phik-cilim,phik+cilim))

par(mfrow=c(1,2))

# Poissonness plot with confidence intervals
 plot(k1,phik,ylim=rng,xlim=range(k0)+c(-0.5,0.5),xlab="k",
 ylab="phik",type="n", main="Poisson plot", sub=
       paste( paste("Intercept=",format(round(rr$coef[6,1],3)))
              paste(", Slope=", format(round(rr$coef[6,2],3))))
text(k1,phik,format(nk1))
segments(k1,phik-cilim,k1,phik+cilim,lty=2)
abline(rr$coef[6,1],rr$coef[6,2],col=2)
```

R code

```
lamhat <- exp(rr$coef[6,2])
tmp <- ifelse(nk0 > 0, sqrt(2+4*nk0), 1)
exptd <- N*exp(-1*lamhat)*(lamhat^k0)/gamma(k0+1)
dk <- tmp - sqrt(4*exptd + 1)

# Residual plot
plot(k0,dk,xlab="k",ylab="FT residual")
abline(h=c(-2,0,2),lty=c(2,1,2),col=c(2,1,2))
list(k=k1,nk=nk1,nkstar=nk2,phik=phik,cilim=cilim,
     int=rr$coef[6,1],slope=rr$coef[6,2],res=dk,expected=exptd
```