Стандартизация (CPUE)

Промысловые данные и GLMs

- Использование GLMs в рыбохозяйственных исследованиях
  - Анализ биологических данных
  - Тестирование гипотез
  - Стандартизация индекса (улучшение индекса относительной численности путем удаления вариации прочих факторов которые могут изменятья во времени и пространстве)
  - Конверсия коэффициента уловистости между орудиями лова
  - Объединение CPUEs различных съемок

стандартизация

Maunder, M. N. and Punt, A. E. 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. 70:141-159.

Предположение

$$\frac{C}{E} = qN$$

Справедливо пока q константа, но в реальности q варьирует в пространстве и времени

CPUE как индекс численности

- Возможность использования CPUE как индекс численности зависит от способности удалить эффект всех факторов кроме численности которые оказывают эффект на *q*
- Исторически, Beverton and Holt (1957) стандартизировали CPUE выбирая стандартный тип судна и определяя относительную мощность других судов

$$RFP_i = \frac{C_i / E_i}{C_s / E_s} \qquad I_t = \frac{\sum_i C_{t,i}}{\sum_i (RFP_i \cdot E_{t,i})}$$

- Этот подход не позволяет учитывать эффект других факторов (месяц, район и проч) или оценить точность

- Now GLMs, GAMs, etc.

Как стандартизровать CPUE?

• каждая модель стандартизации будет иметь эффект года и другие факторы

• эффект года будет выражаться в оценках величины относительной численности

• удаляются эффекты других факторов

$$Y = Год + температура$$

•Если все другие эффекты – факторы то многофакторная ANOVA

• если другие эффекты непрерывные величины то ANCOVA

Hilborn and Walters (1992)
p. 126-128 (assuming log-normal, no zeros)

Catch rate (tons per hour) by vessel class

Year	Class I	Class II	Class III
1	0.63	0.85	1.28
2	0.46	0.65	1.09
3	0.35	0.66	1.01
4	0.43	0.48	0.84

$$U_{ti} = \beta_0 * Yr * Class * \varepsilon$$

$$\log_e(U_{ti}) = \log_e(\beta_0) + \log_e(Yr) + \log_e(Class) + \log_e(\varepsilon)$$

Estimated catch rate for Class 1 and Year 1

Hilborn and Walters (cont)

```
>data<-read.csv("P:/Rwork/book/HibornWalters 4.2 data.csv")
>data$year<-as.factor(data$year)                    #make year a factor
>data$class<-as.factor(data$class)                  #make class a factor


>modl<-glm(log(cpue)~year+class,data=data,family=gaussian)
>anova(modl,test="F")

Analysis of Deviance Table
Model: gaussian, link: identity
Response: log(cpue)
Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                    11    1.80717
year   3  0.35016         8    1.45701  8.3776 0.0144792 *
class  2  1.37341         6    0.08359 49.2890 0.0001889 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hilborn and Walters (cont)

>summary(modl)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.51678    0.08346  -6.192 0.000817 ***
year2       -0.24781    0.09638  -2.571 0.042259 *
year3       -0.35923    0.09638  -3.727 0.009766 **
year4       -0.45820    0.09638  -4.754 0.003145 **
class2       0.34739    0.08346   4.162 0.005930 **
class3       0.82525    0.08346   9.888 6.18e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.01393227)

    Null deviance: 1.807166  on 11  degrees of freedom
Residual deviance: 0.083594  on  6  degrees of freedom
AIC: -11.546

Number of Fisher Scoring iterations: 2
```

>yrcoef<-exp(dummy.coef(modl)[[2]]) #calculates all coeffs and back-transform

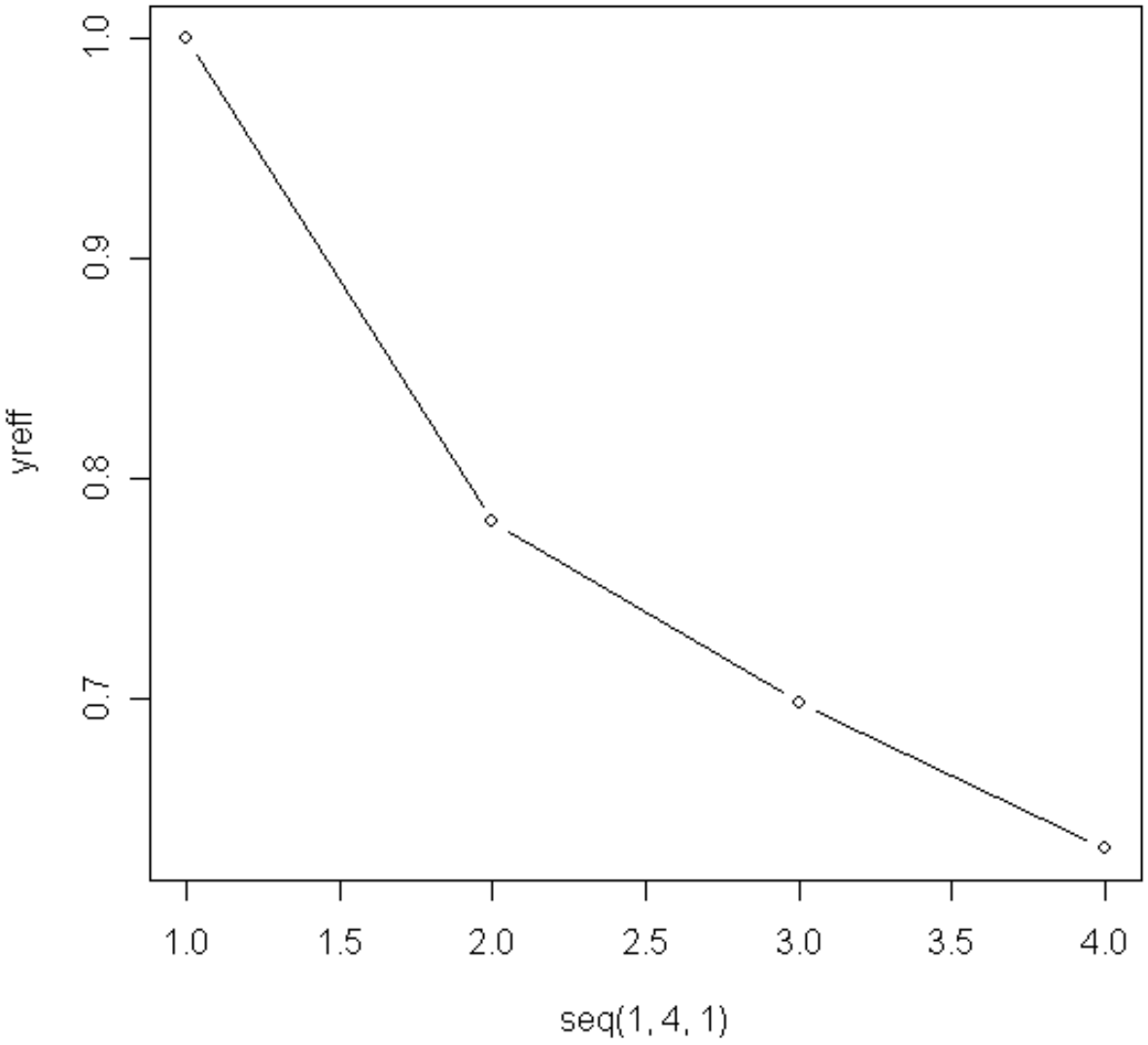
>yrcoef

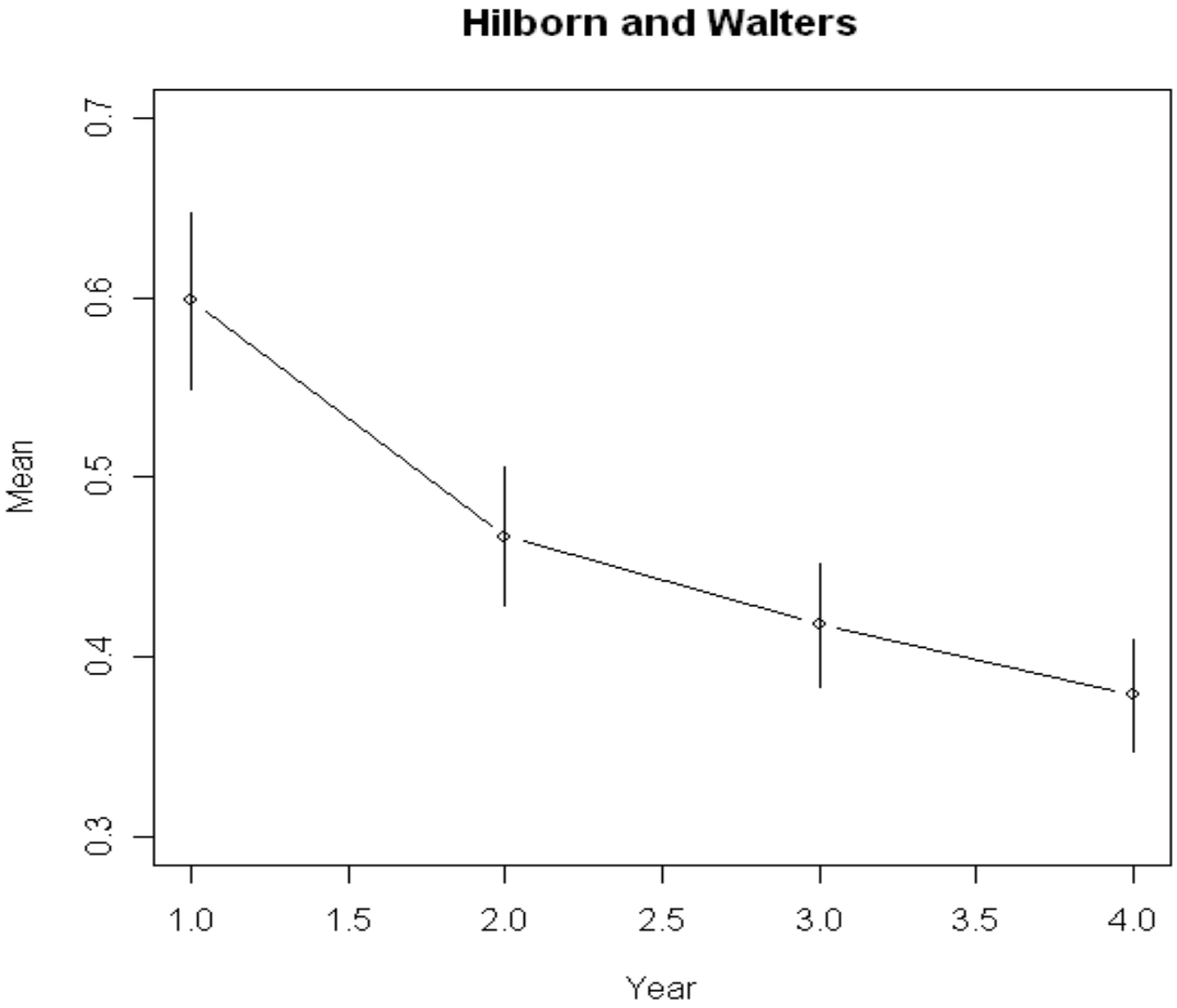
```
   1         2         3         4
1.0000000 0.7805057 0.6982131 0.6324213
```

#Year 2 = 78% of Reference Yr1

>plot(yrcoef~seq(1,4,1),type="b")

Year Effects




Hilborn and Walters
Mean
0.3
0.4
0.5
0.6
0.7
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Year

Пример стандартизации

- **Рассчитать индекс численности на основе данных полученных в результате траловой съемки**

- **Независимые переменные выбраны для модели на основе следующих предаоложений:**
  - **Температура оказывает эффект на перемещения рыб, поэтому учитывать этот фактор важно**
  - **Глубина также может влиять на распределение рыб**
  - **Еффект года важен и должен учитываться для описания изменений во времени**
  - **Другие переменные также протестированы**

- **Модель будет использована для расета временного ряда данных и анализа тренда**

Model

$$Abundance \sim \beta_0 + \beta_1 Year + \beta_2 Temperature + \beta_3 Depth + \ldots$$

Abundance = численность выловленной камбалы
Year = календарный год (1990 – 2012) – дискретная

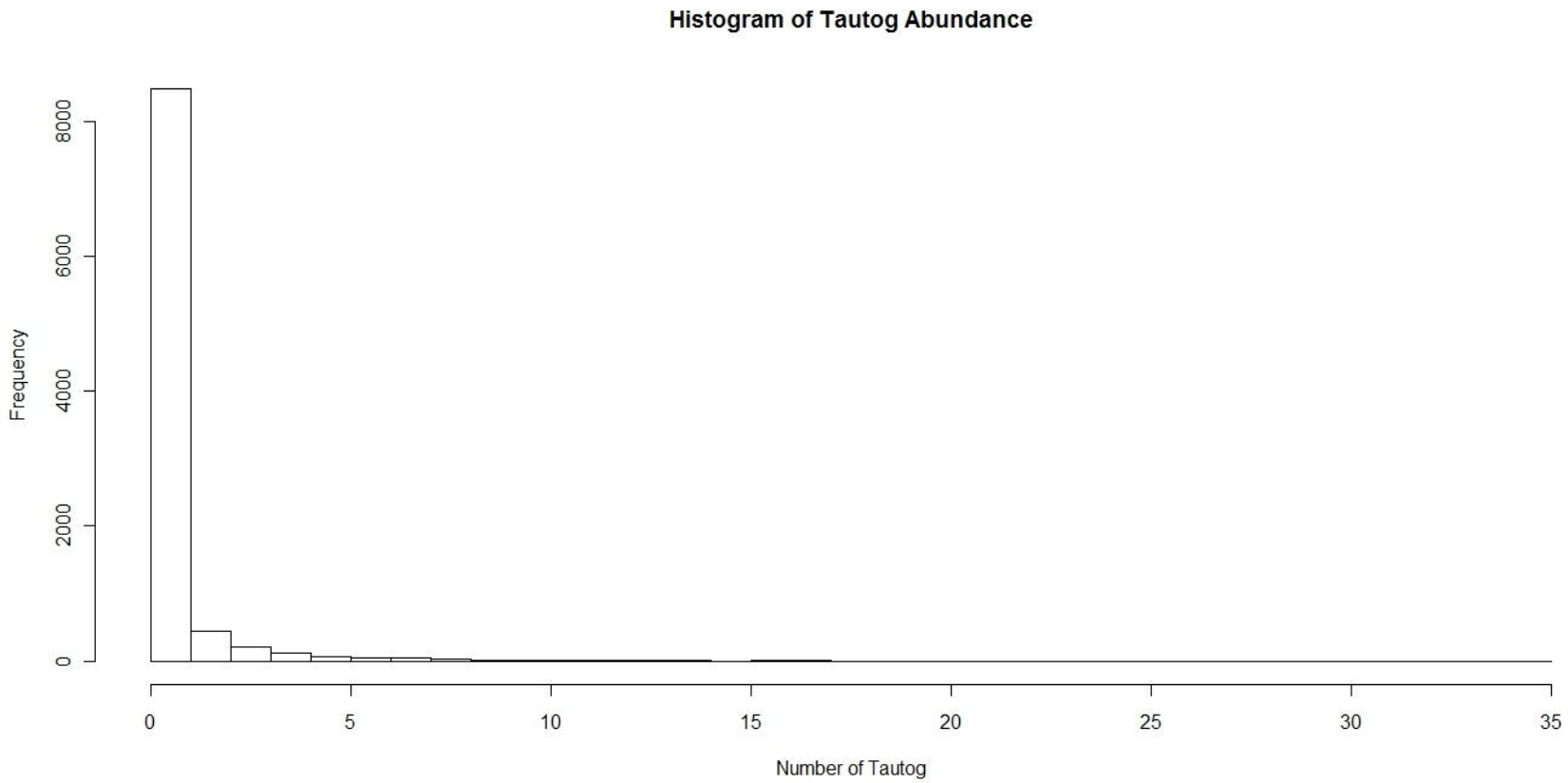
Temperature = ткмпература в С - непрерывная

Depth = глубина - непрерывная

Структура и анализ

- **На основе функциональной формы праспределения вылова следующие модели были выбраны для анализа:**
  - **Отрицательная биноминальная модель распредления**
  - **Модель оценена с использованием всех параметров, затем по одной переменной удаляли и переоценивали модель**
- **Окончательная модккль выбрана на основе AIC**
- **После диагностики окончательная модель модифицирована на основе результатов**

- **Сначала рассмотрим распределение вылова**



- **Данные распределены в соотв с отрицательнвм бином распределением**

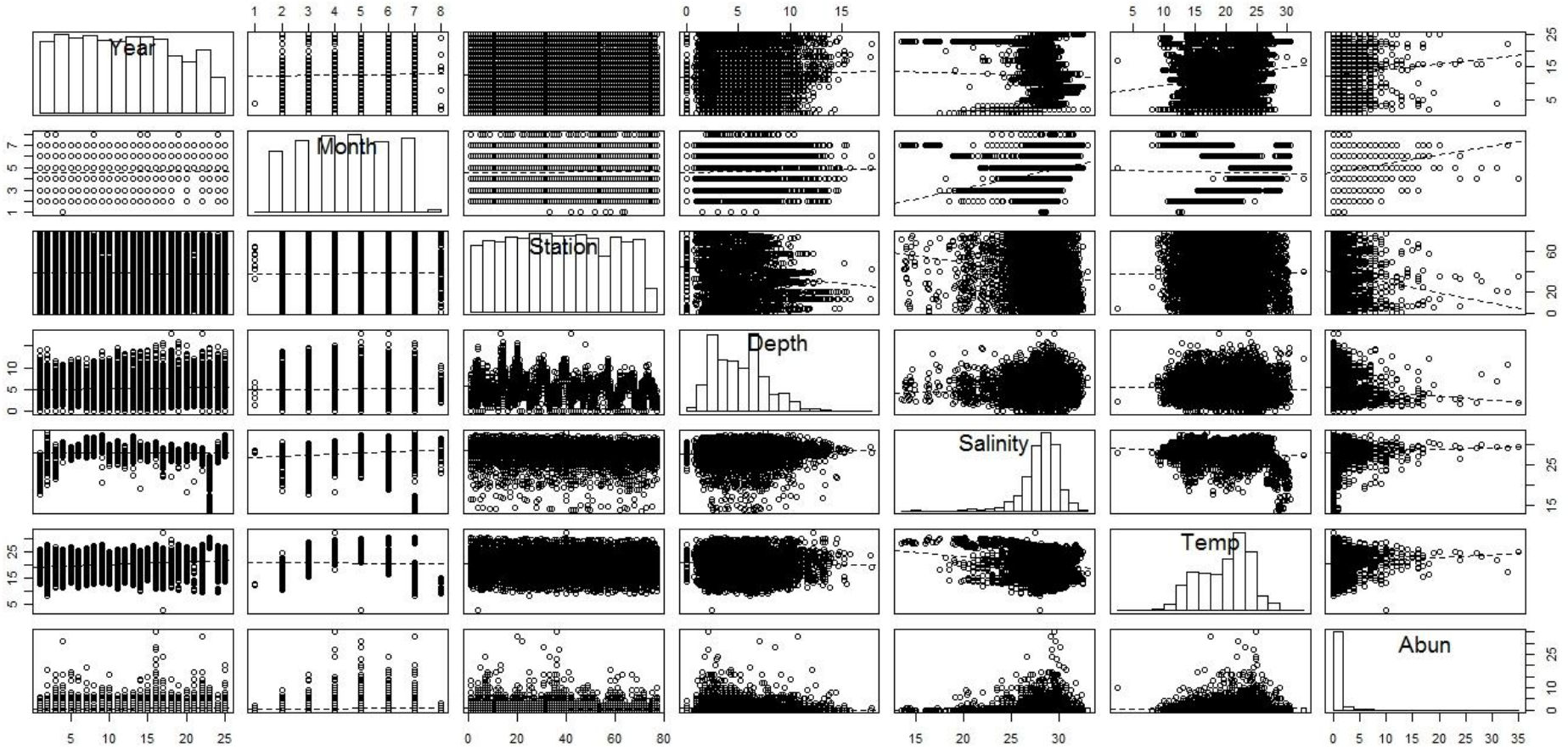
Выбор модели

Model	AIC
Year + Temp + Depth + Salinity + Station + Month	16,371
Year + Temp + Depth + Salinity + Station	16,463
Year + Temp + Depth + Salinity	17,051
Year + Temp + Depth	17,277
Year + Temp	17,298
Year	17,711

Multicollinearity

- Diagnostics of multicollinearity
  - Matrix of scatter plots and correlations
  - Adding or deleting a predictor variable changes the regression coefficients
  - Variance Inflation Factor (VIF): measure how much variances of estimated regression coefficients are inflated compared to when predictor variables are not correlated (VIF>10, problems)
  - For VIF, use *vif()* in package *car*

Output NY – Model Diagnostics Multicollinearity



Output NY – Model Diagnostics Multicollinearity

- **VIF test:**
  - **The VIF test indicates potential problems with multicollinearity**

Year	6.8
Temp	5.7
Depth	3.5
Salinity	4.4
Station	5.0
Month	7.3

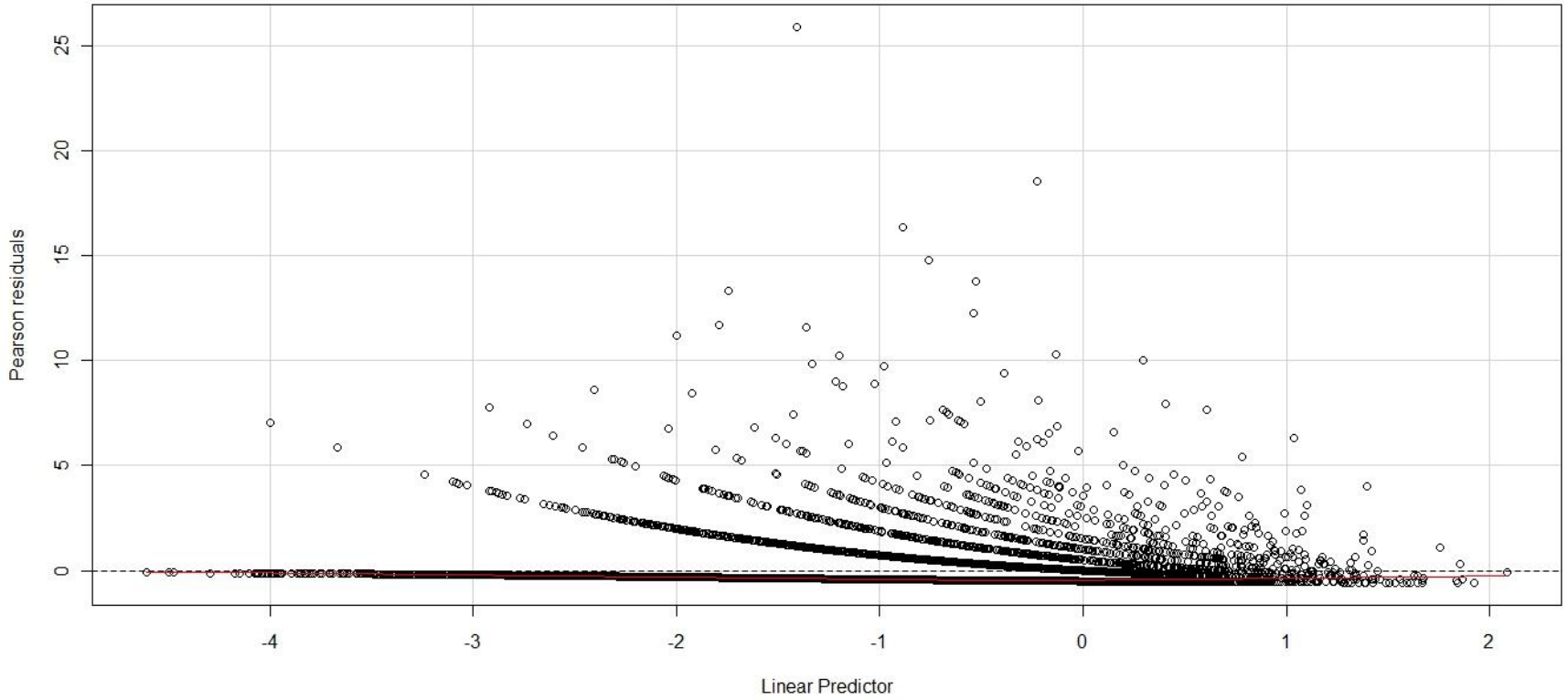
Output NY – Model Diagnostics Multicollinearity

- **VIF test:**
  - **Dropped Month (highest VIF value) and reran**
  - **All covariates under 5 so problem corrected**

Year	4.1
Temp	1.1
Depth	3.5
Salinity	3.2
Station	4.6

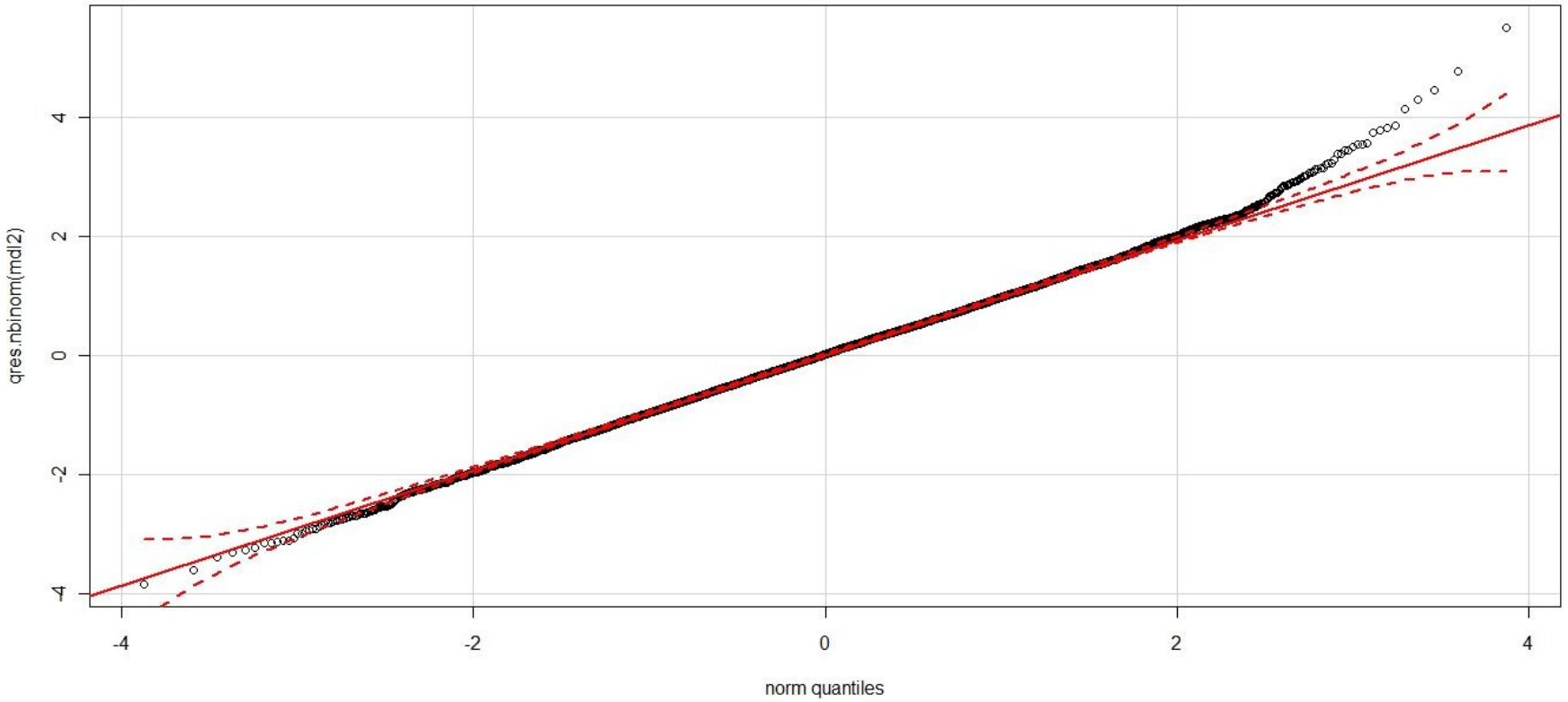
Output NY – Model Diagnostics Homoscedasticity

- **The next diagnostic was to test to make sure there were no issues with variance**



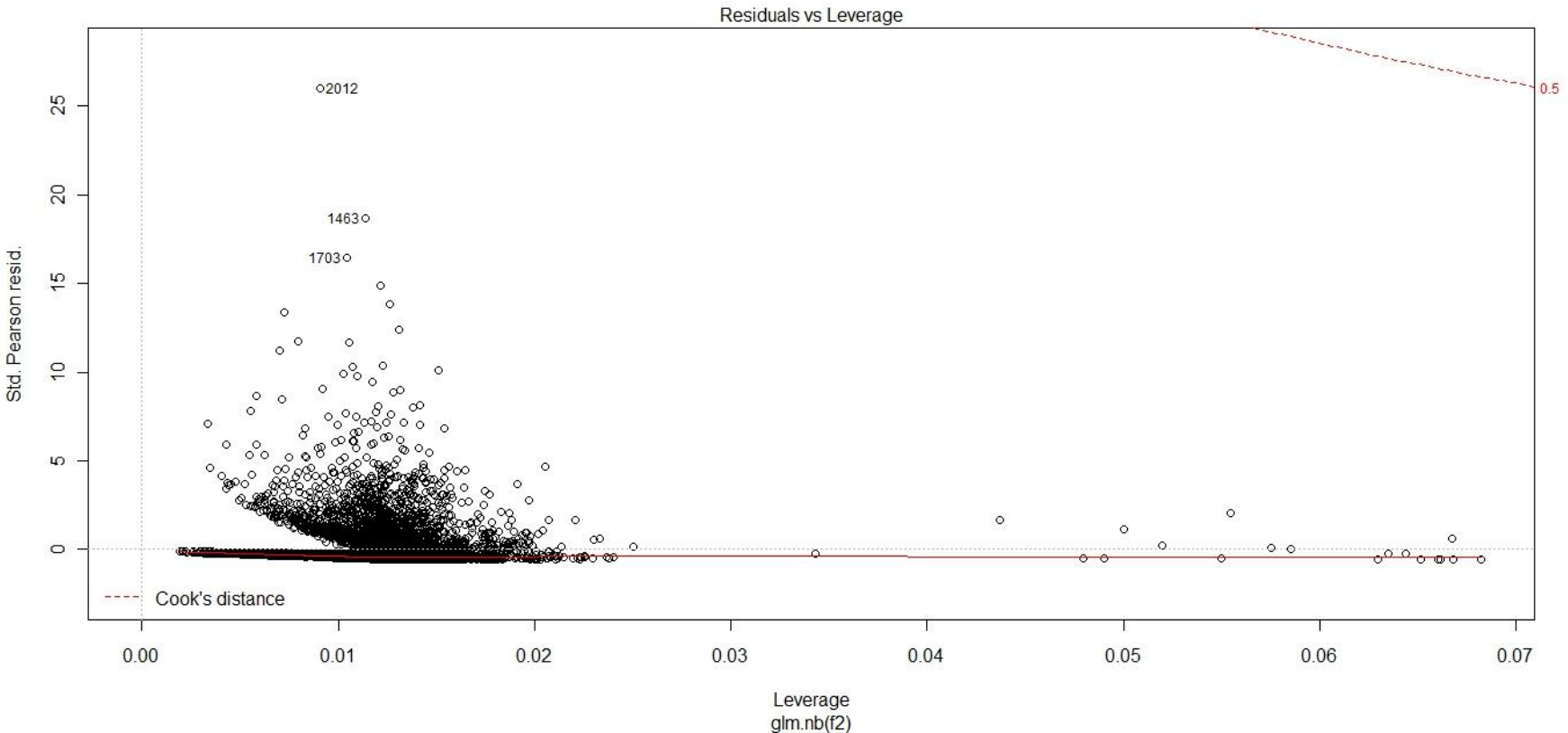
Output NY – Model Diagnostics
Data Distribution

- **The next diagnostic was to test the chosen data distribution**

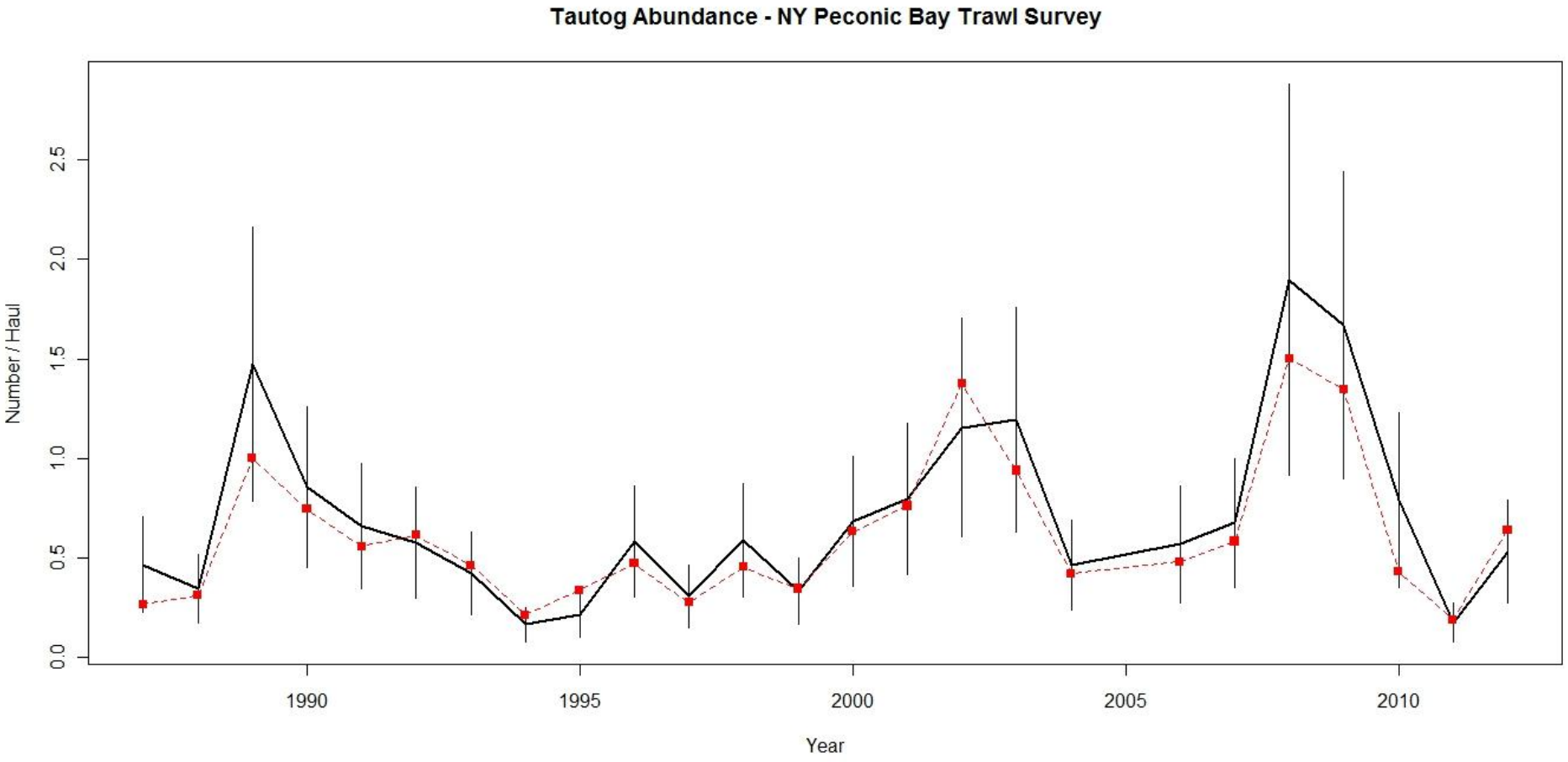


Output NY – Model Diagnostics Outliers

- **The next diagnostic was to test to make sure there were no major outliers**



Output NY – Results



Независимые от промысла данные

- Съемки
  - Количество выловленных рыб на единицу стандартного траления (много 0), предполагается что индекс связан с численностью
  - обычно GLMs не используются для стандартизации индекса относительной численности поскольку съемка планируется так чтобы минимизировать дисперсию (например стратифицированная случайная съемка)
  - GLMs может быть использована если имеется информация по дополнительным covariates
  - Covariates: глубина, температура, соленость, тип дна, района - страта

Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. 53: 577-588.

Промысловые данные

- промысловый CPUE
  - Обычно информация по общему улову и усилию (на уровне 1 дня или рейса)
  - Очень мало нулевых уловов для целенаправленных одновидовых рейсов
  - В случае многовидового промысла 0 уловы по отдельным видам
  - Covariates: год, месяц, усилие, характеристики судна (мощность, тоннаж), район промысла

Punt et al. 2000. Standardization of catch and effort data in a spatially-structured shark fishery. Fish. Res. 45:129-145.

CPUE в GLM

- C/E используется как Y
  - Hilborn and Walters предположили что C/E распределен log-normally

- C используется как Y и включает E как offset или как независимую переменную
  - C трансформируется или моделируется с соответствующей формой распеделения ошибки

- Если E стандартизировано (съемки) то услилие не используется

Проблемы стандартизации

- Если относительный вылов не пропорционален численности, GLM тренд не будет правильно отражатьчисленность

  например усилие сконцентрировано в одном районе, орудия лова которые могут перенасыщаться

- Любые из менения неучтенные GLM будут интерпретированы как изменения численности

Наиболее часто используемые распределения

- Лог нормальное (необходимо добавлять константу к каждой величине если есть нулевые значения из за логарифмирования)
- Гамма с log link (необходимо добавлять константу если есть нулевые значения)
- Пуассон (работает с 0) – необходимо округлять если используется C/E
- Отрицательное биномиальное (работает с 0) – необходимо округлять если используется C/E

Other Issues

Interactions

- Interactions among year and other variables may occur
- Common – year x month/week/day and area
- Significant terms raise some interesting questions about the interactions
- Interactions with year are problematic
- Makes interpretation of year effects difficult (may not be an index of relative abundance)

Remedial Measures for Year Interactions

- Ignore the interactions with year – don't include when model is estimated (most common in literature)
  - May lead to biased index of abundance if interaction is substantial
  - Try model with and without to see how year estimates change

- Interactions with year and month/week/day
  - Average estimates from year x month interaction over year
  - Use dummy dataset to get estimate then average across years

- Interactions with year and area
  - Use area size to calculate a weighted average like above
  - Recognize that these interactions imply different trends in abundance in different areas; suggests a spatially structure population

Zero catches

“Real” Zeros

- Due to clumped distribution of fish within its range and the sampling design

“False” Zeros

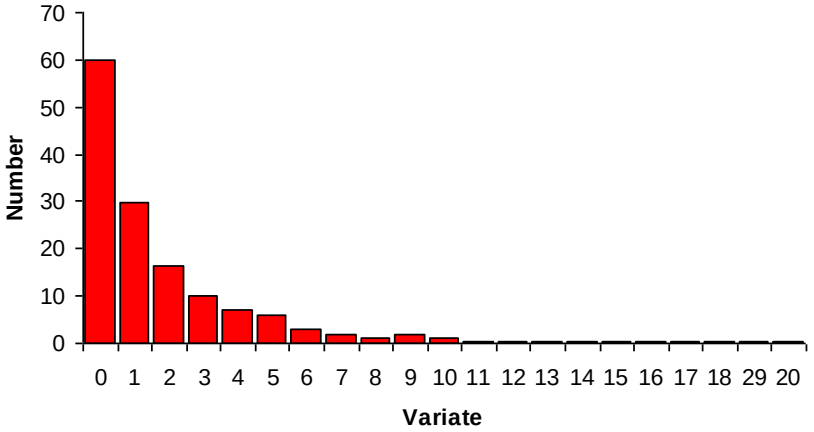
- Gear unknowingly malfunctions during tow
- Sampling occurs outside species’ distribution

Remedial Measures for Zero catches

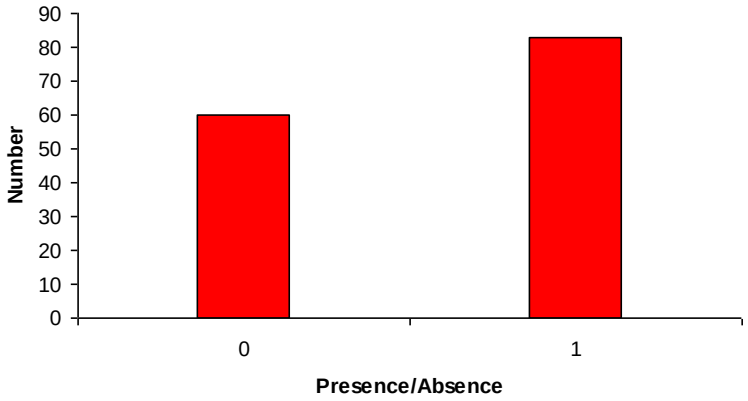
- Ignore them
  - Delete if “False” zeros
- Aggregate information in a cell
  - Results in loss of information
- Add a small constant to all data (common)
  - Choose based on appropriate transformation to normality
  - Remove constant when back-transformed
- Use distributions that have zeros
  - Poisson, Negative Binomial
  - If not best fit, try zero-inflated Poisson and Negative binomial distributions
- Delta (Conditional) Model
  - Model zeros and positive catches separately

Delta (Conditional) Models

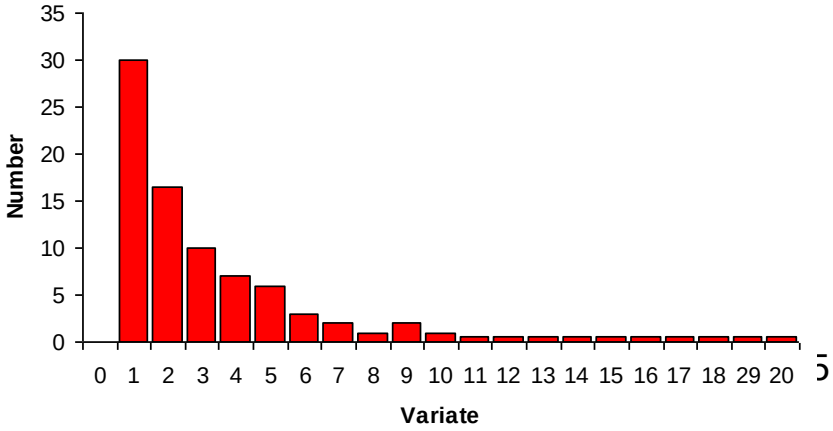
- If data aren't fit well to a single distribution
- Can decompose fish abundance data into two components and model each component separately
- Then recombine



Absence/Presence (binomial: 0/1)



Positive Values

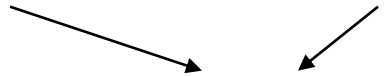


Decomposition of catch data into binary and positive components

No. per haul
0
4
6
1
0
3
4
6
0
0
2
Mean 2.360

Decomposed

Binary	Positive
0	
1	4
1	6
1	1
0	
1	3
1	4
1	6
0	
0	
1	2
Proportion 0.64	Mean 3.71



p*m= 2.36

Delta-X Models

- Delta refers to zeros and X can be any distribution for the positive catches

- Each component is modeled separately

Presence/Absence (0/1) – logistic regression

$$\log_e\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_n X_n + \varepsilon$$

Positive Values – glm assuming different error

$\log_e Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_i$ *(log-normal)*

Stefannson, G. 1996. Analysis of groundfish survey abundance: combining the GLM and delta approaches. ICES J. Mar. Sci. 53:577-588

Delta-X Models

- Get estimates of p and y for each year and combine to generate predicted unconditional mean

Estimate year means and probs as before

$$\hat{A}_i = \hat{p}_i \cdot \hat{\mu}_i$$

$$\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_{1i})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_{1i})}$$

$$\hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_{1i})$$

$$\mathrm{var}(\hat{A}_i) = \mathrm{var}(\hat{\mu}_i \cdot \hat{p}_i) = \hat{p}_i^2 \,\mathrm{var}(\hat{\mu}_i) + \hat{\mu}_i \,\mathrm{var}(\hat{p}_i) + 2 \cdot p_i \cdot \mu_i \cdot Cov(p, \mu)$$

See Appendix 1 in Lo et al. 1992 for complete example

Common Delta-X Models

- Presence/absence modeled as binary response using logistic GLM
- Positive response modeled as log-normal, Gamma (log link), Poisson (log link) , inverse Gaussian
- Other potential positive distributions – zero-truncated Poisson and Negative Binomial

Delta-X Models

- Interpretation of predicted mean index is dependent on the significant variables in both GLM models

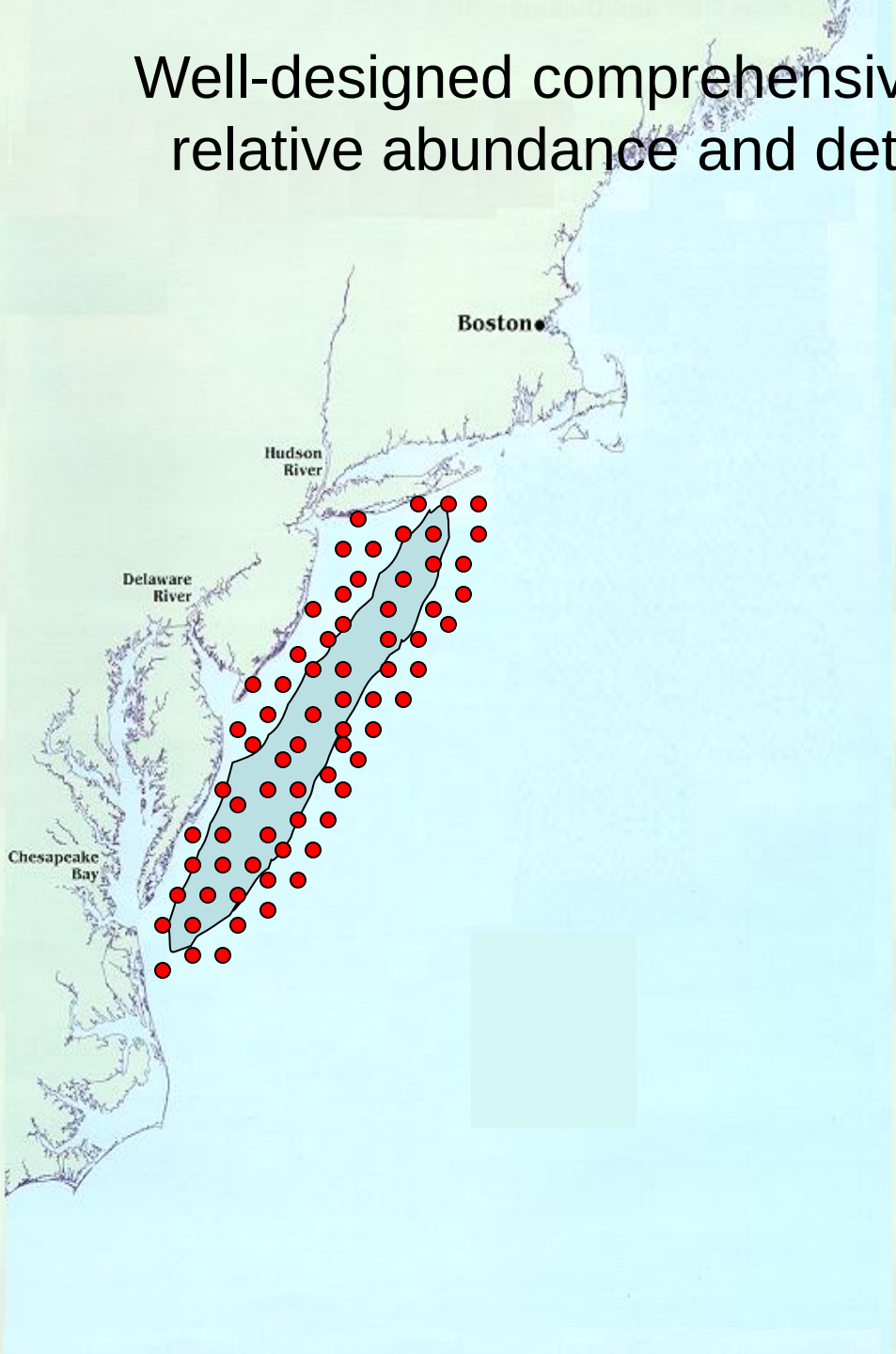
- For example, if logistic GLM finds p depends on salinity and the GLM for positive values finds positive catches depend on temperature, then the overall mean depends on salinity and temperature

Combining Survey Indices

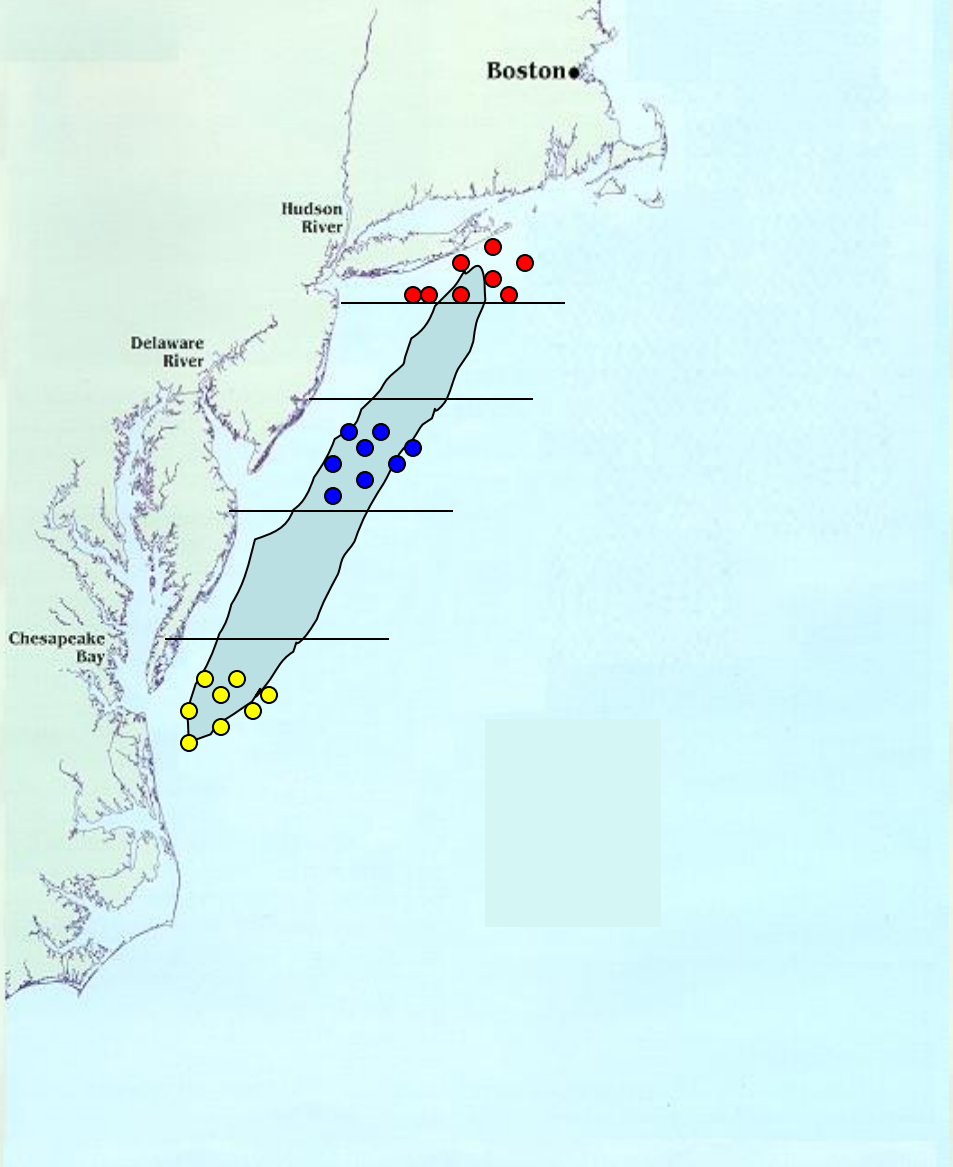

Range Expansion
(High Abundance)
Boston
Hudson River
Delaware River
Chesapeake Bay

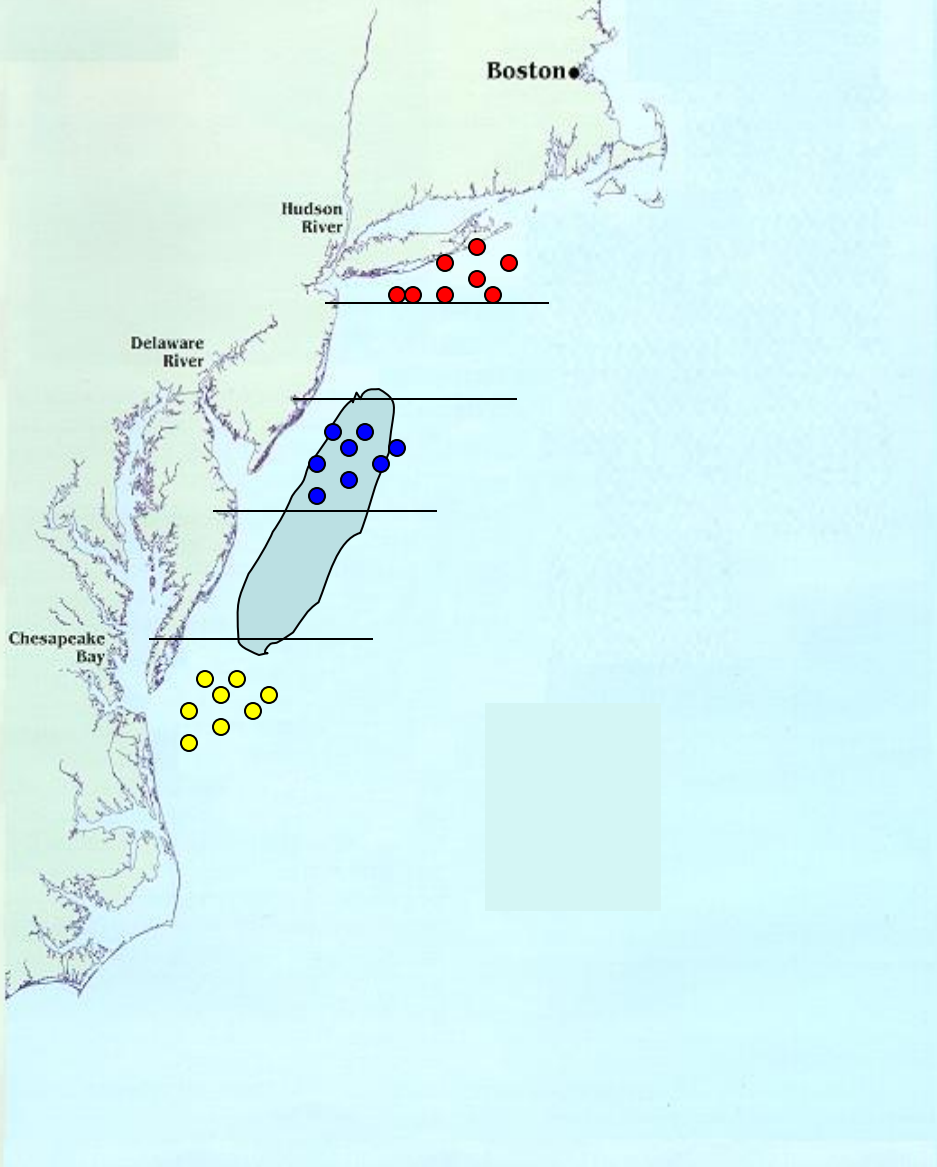

Range Contraction
(Low Abundance)
Boston
Hudson River
Delaware River
Chesapeake Bay

Well-designed comprehensive survey is best to estimate relative abundance and detect changes in abundance

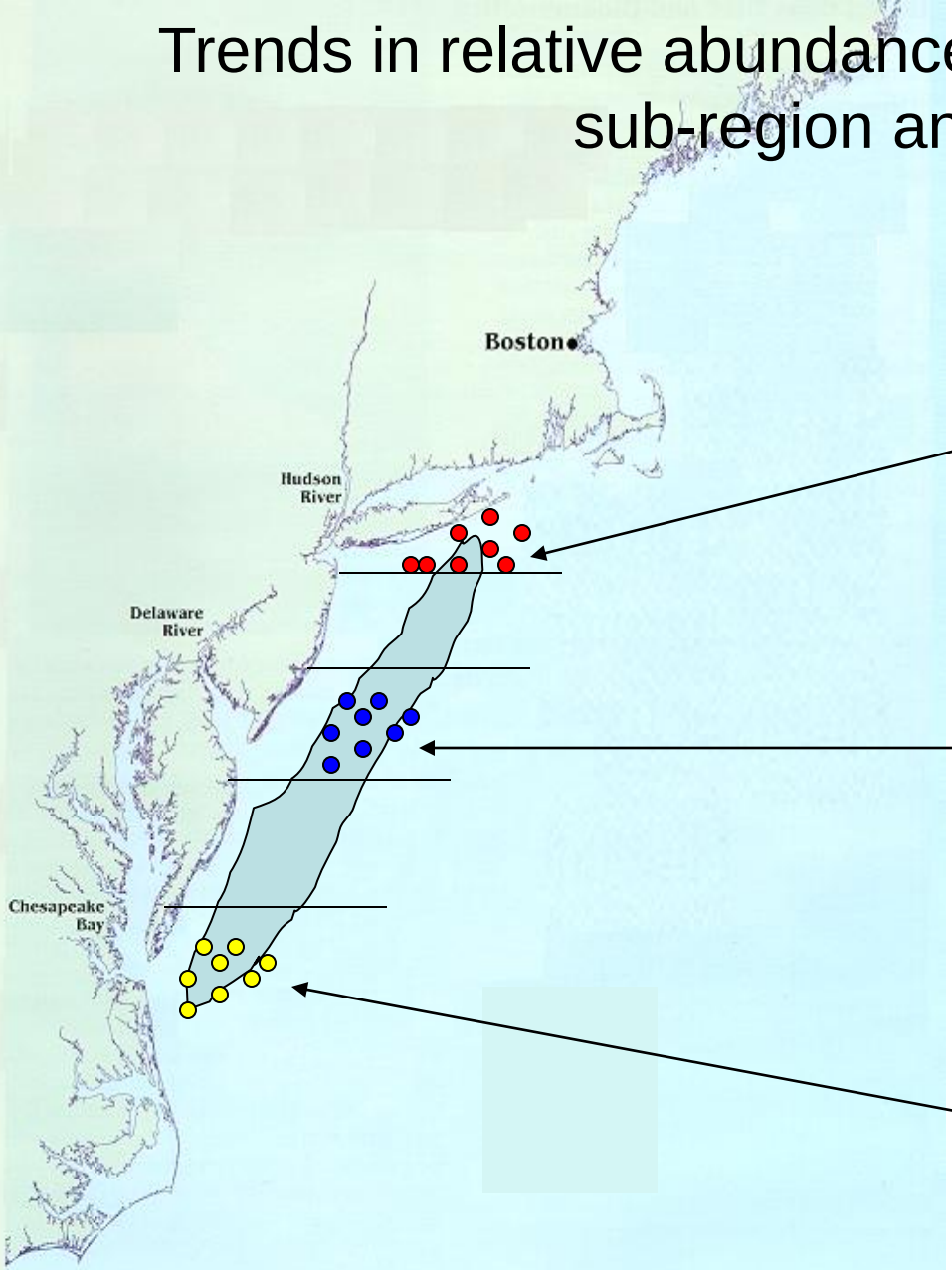


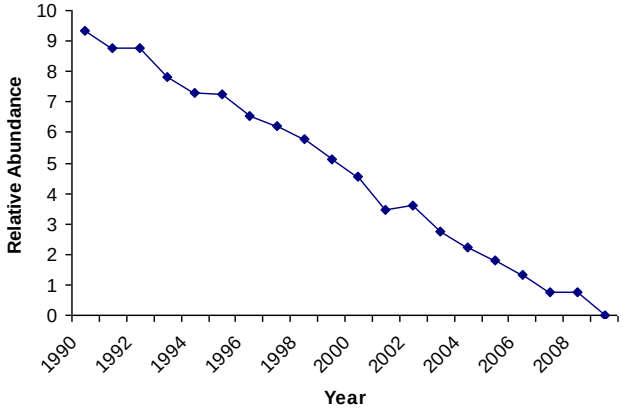
Because of jurisdictional boundaries, separate surveys may be conducted in sub-regions of a species range using different methodologies

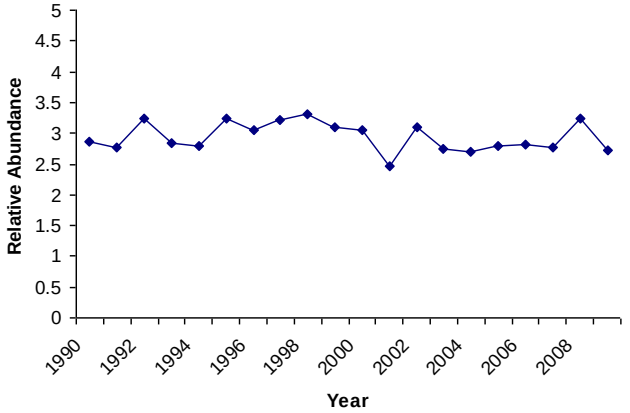


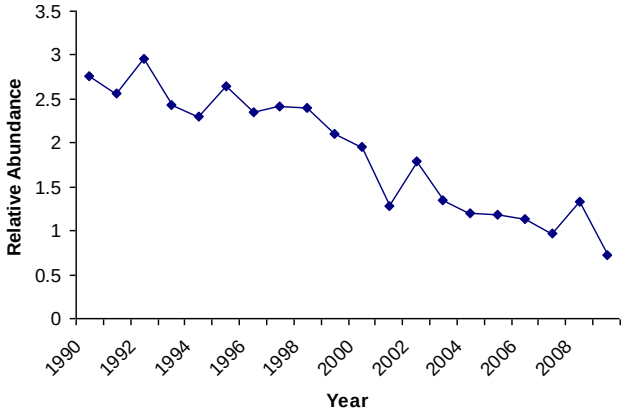


Trends in relative abundance will be vary depending on sub-region and gear used









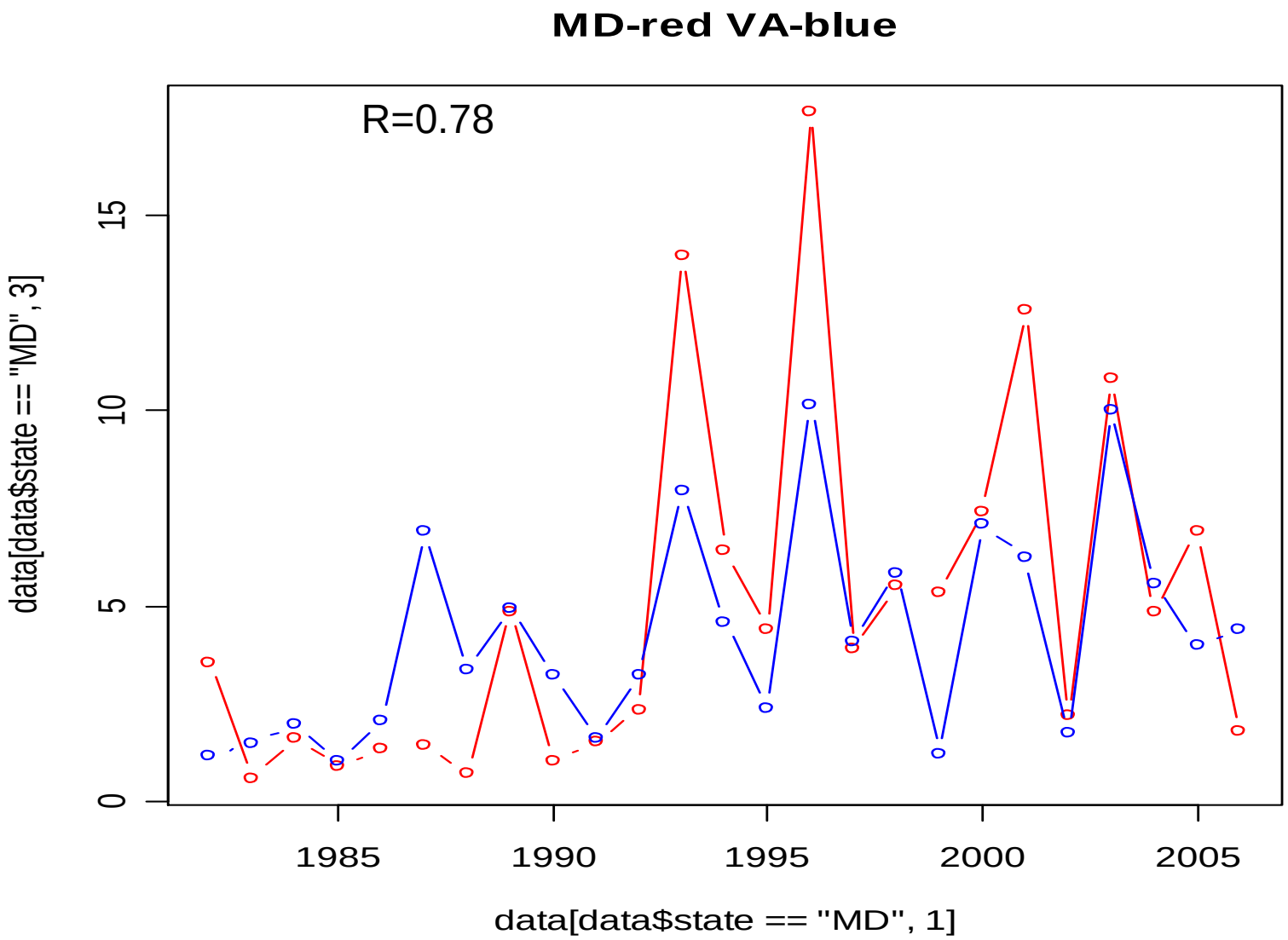
If using different mesh sizes, measuring different segments of the population

- How do you combine indices?
  - If all subregions were sampled (with similar gears and methodologies), equal to comprehensive survey - combine using area-weights in a weighted average (like stratified mean)
  - If few subregions (using same gear and methodologies), treat subregion as random effect and use GLM with correlated errors (Fabrizio et al., 2000)
  - If multiple surveys with different gears in same sub-region can use GLMs to standardize by including a gear effect (assuming same mesh sizes; Ault and Smith MS 2000)
  - If multiple surveys with different gears, different regions: gear may not be sampling same population (size-related difference)

Combining Survey Indices

- Multiple relative abundance indices may be available for single stock
- How to combine into one, overall index?
- Use GLM with gear factor – year effects are combined estimate
- Choice of distribution with depend on availability of data (raw versus means)

Combining MD and VA YOY SB Indices



```
data<-read.csv("P:/ASMFC/Glm Course/CPUE
    Standardization/Programs and Data/sbyoycb.csv")

modl<-glm(index~as.factor(year)+state,family=gaussian,data=data)

termplot(modl,rug=F,se=F,ask=T,partial.resid=T)
```