STATS762 Regression for Data Science

Model assessment

May 8, 2019

Black cherry tree

	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9

This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees.

Girth	Tree diameter (rather than girth, actually) in inches
Height	Height in ft
Volume	Volume of timber in cubic ft

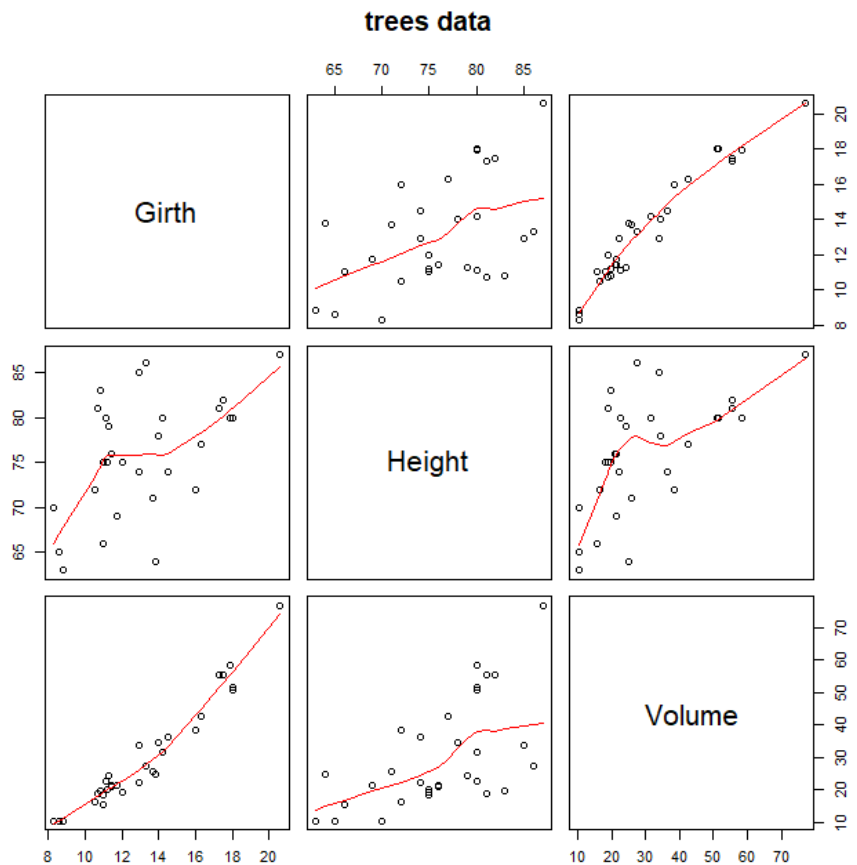
Black cherry tree

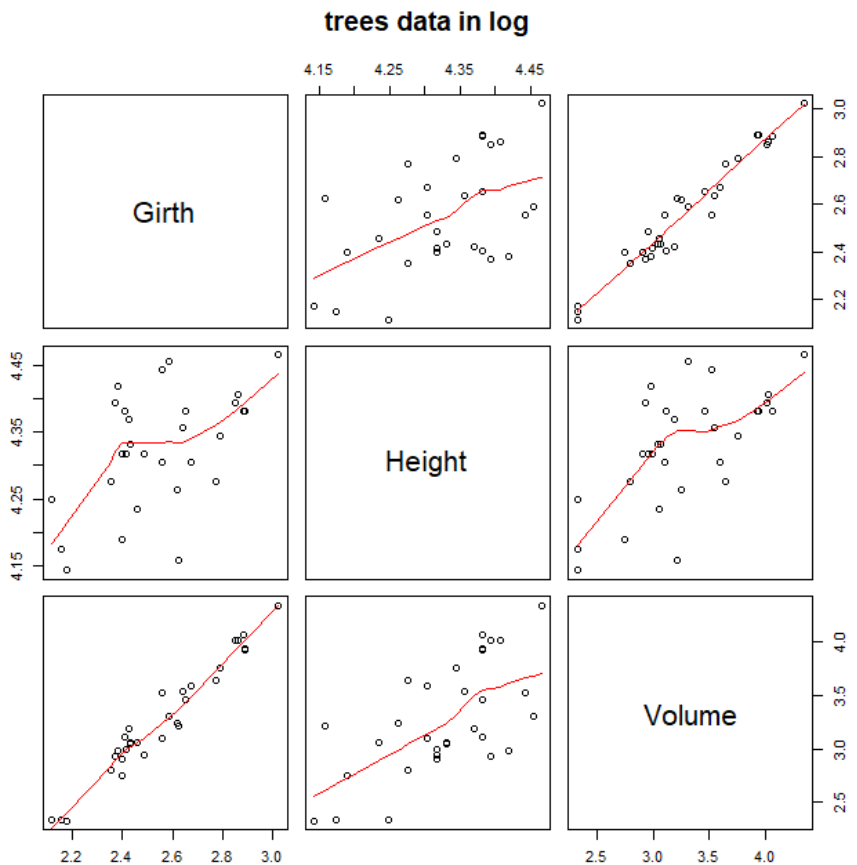
	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9

Can we predict the volume of a cherry tree?
E.g., if a particular tree has diameter 11 inches and height 85 feet, can we predict its volume?

Black cherry tree

Naive pair plots - (Left) the actual scale and (Right) log scale





Black cherry tree

We have three possible linear models of interest.

Model 1 $Volume = \beta_0 + \beta_1 Grith + \epsilon, \epsilon \sim N(0, \sigma^2)$

Model 2 $\log(Volume) = \beta_0 + \beta_1 \log(Grith) + \epsilon, \epsilon \sim N(0, \sigma^2)$

Model 3 $\log(Volume) = \beta_0 + \beta_1 \log(Grith^2 * Height) + \epsilon, \epsilon \sim N(0, \sigma^2)$

Here ϵ is a random error associated with observation and is normally distributed with mean of 0 and variance of σ^2.

Which model describes the observation best?

Let's choose the model resulting the smallest discrepancy between prediction and observation.

Black cherry tree

Statistical models for *Volumne*

Model 1 $$f = \frac{e^{-(Volume-(\beta_0+\beta_1 Grith))^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$$

Model 2 $$f = \frac{e^{-(\log(Volume)-(\beta_0+\beta_1 \log(Grith)))^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$$

Model 3 $$f = \frac{e^{-(\log(Volume)-(\beta_0+\beta_1 \log(Grith^2 Height)))^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$$

Black cherry tree

Given $D = \{Volume_i, Girth_i, Height_i\}_{i=1}^{31}$, the likelihood for $\theta = (\beta_0, \beta_1, \sigma^2)$ for the three models follow;

$$\text{Model 1} \quad L(\theta; D) = \prod_{i=1}^{31} \frac{e^{-(Volume_i - (\beta_0 + \beta_1 Grith_i))^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$$

$$\text{Model 2} \quad L(\theta; D) = \prod_{i=1}^{31} \frac{e^{-(\log(Volume_i) - (\beta_0 + \beta_1 \log(Grith_i)))^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$$

$$\text{Model 3} \quad L(\theta; D) = \prod_{i=1}^{31} \frac{e^{-(\log(Volume_i) - (\beta_0 + \beta_1 \log(Grith_i^2 Height_i)))^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$$

For each model, we estimate $(\beta_0, \beta_1, \sigma^2)$ by maximizing the likelihood. This is so called Maximum Likelihood Estimation (MLE).

Model selection - Criterion based procedures

For a true (unknown) model g, a model f parameterized by θ is picked. The Kullack-Leibler divergence from g to f_θ is given by

$$D_{KL}(g||f_\theta) = \int g(x) \log(g(x)/f_\theta(x)) dx$$

i.e., Information lost when f_θ is used to approximate g.

- Always non-negative, $D_{KL}(g||f_\theta) \geq 0$.
- If $g(x) = f_\theta(x)$ almost everywhere, $D_{KL}(g||f_\theta) = 0$.
- A large value of $D_{KL}(g||f_\theta)$ means that f_θ is very different to g.

We want to choose f_θ minimizing the KL divergence!

Model selection - Criterion based procedures

Akaike (1974) [1] showed that $-\log(L(\theta^*)) + k + c$ is an asymptotic bias corrected estimator for $\mathbb{E}[D_{KL}(g||f_\theta)]$

$\log(L(\theta^*))$ - log-likelihood at the maximum likelihood estimate θ^*

k - Number of parameters in the model f

c Constant depending on the true model g
i.e., It can be ignored for model comparison.

KL divergence estimator has two components; model fit and model size.

e.g., For linear regression model,
$-2\log(L(\theta^*)) = n\log(RSS/n) + \text{constant}$. Here the constant is same for a given data and can be ignored for model comparison.

[1] H. Akaike, "A new look at the statistical model identification", IEEE Trans. Autom. Control, vol. AC-19, no. 6, pp. 716-723, Dec. 1974.

Model selection - Criterion based procedures

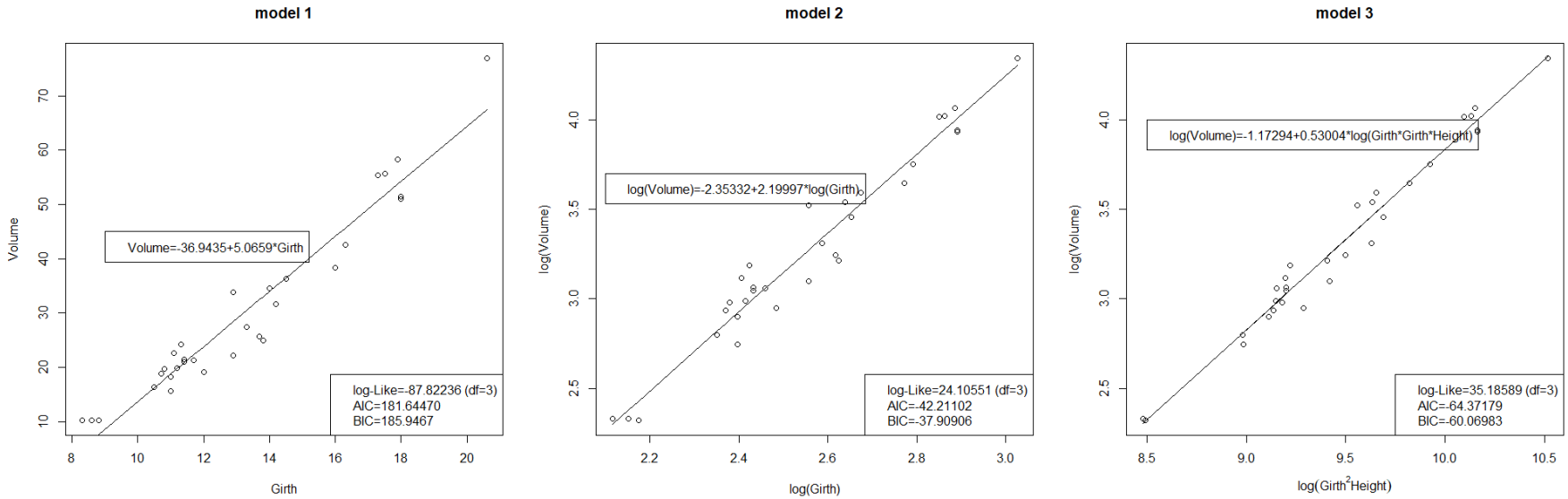
Information criterion is a method to select a model balancing between goodness of fit and simplicity of models.

- Akaike information criterion: $AIC = \underbrace{-2\log(L(\theta^*))}_{\text{goodness of fit}} + \underbrace{2k}_{\text{penalty}}$
- Bayesain information criterion: $BIC = \underbrace{-2\log(L(\theta^*))}_{\text{goodness of fit}} + \underbrace{k\log(n)}_{\text{penalty}}$

Here k is a number of parameters, n a number of observations and θ^* is the maximum likelihood estimate.

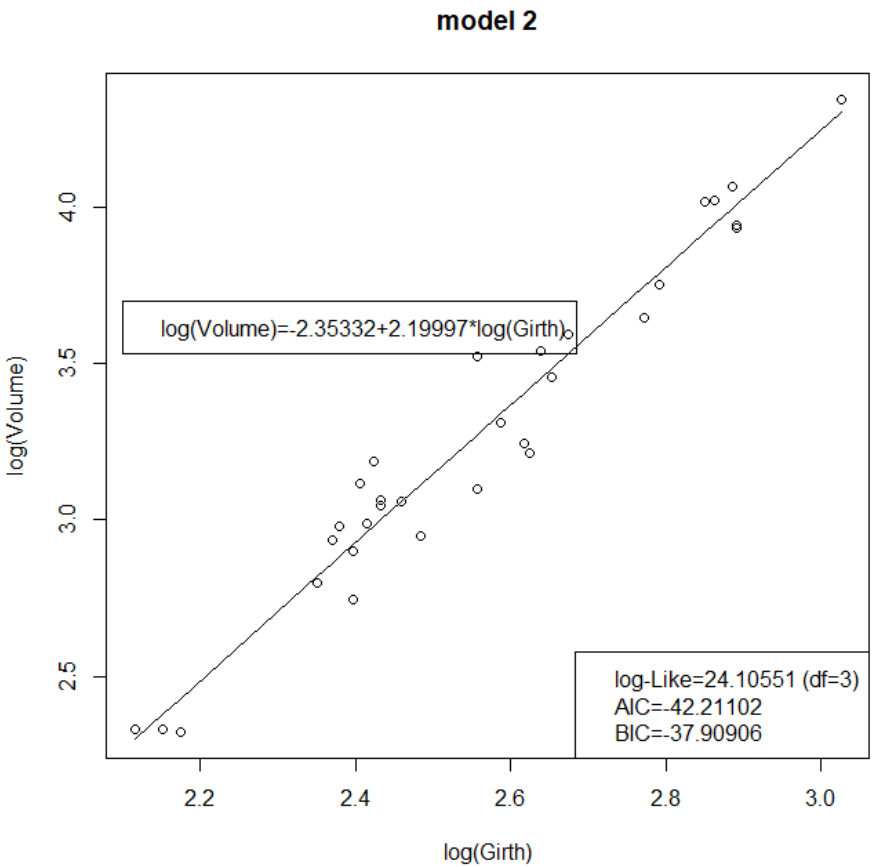
- Relative quality of statistical models for a given set of data.
- A model with the smallest value is preferred.
- BIC puts a heavier penalty (i.e., it will tend to prefer smaller models in comparison to AIC.)

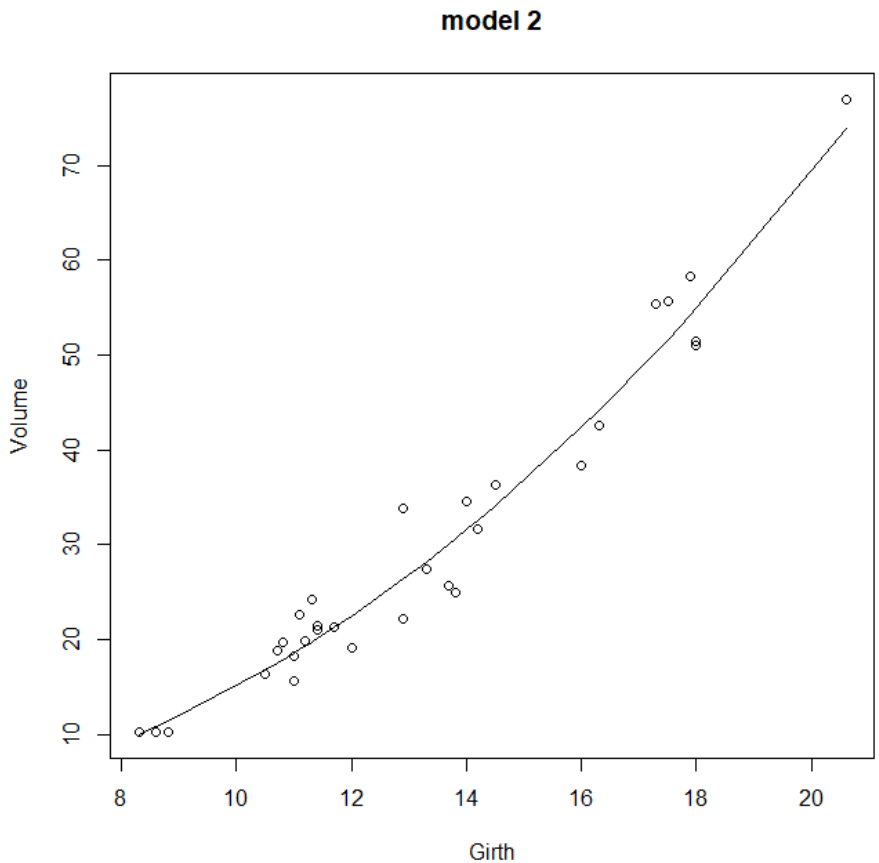
Black cherry tree



Among the three model, Model 3 gives the smallest AIC and BIC values. Hence Model 3 is chosen.
Transform the model back to the data scale for easy interpretation.

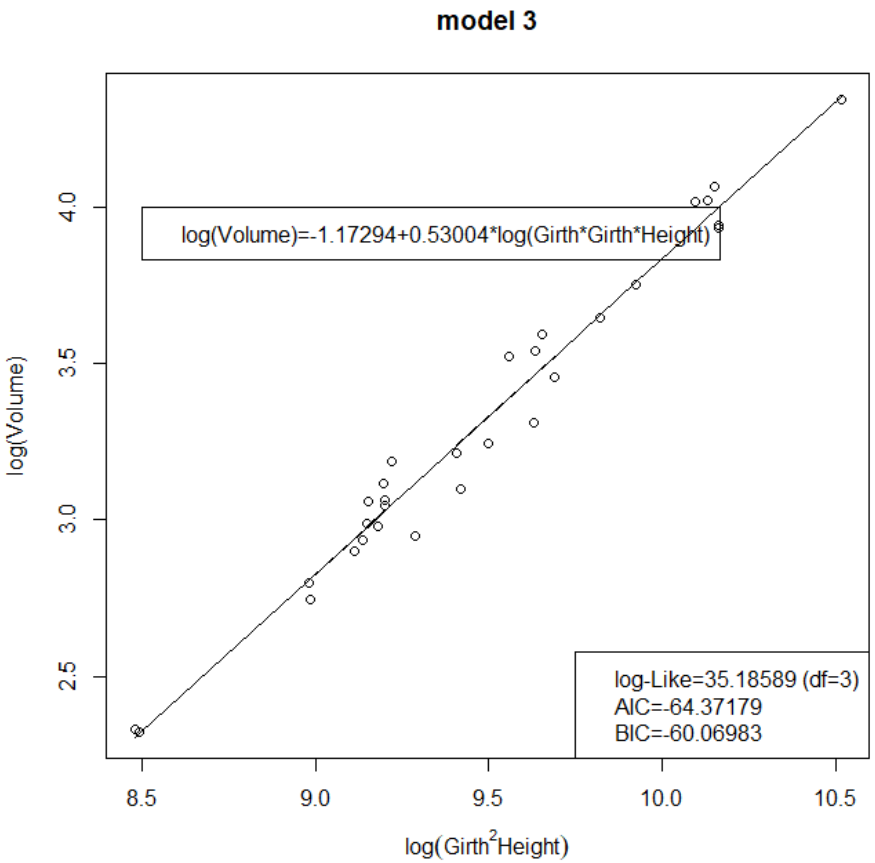
Black cherry tree

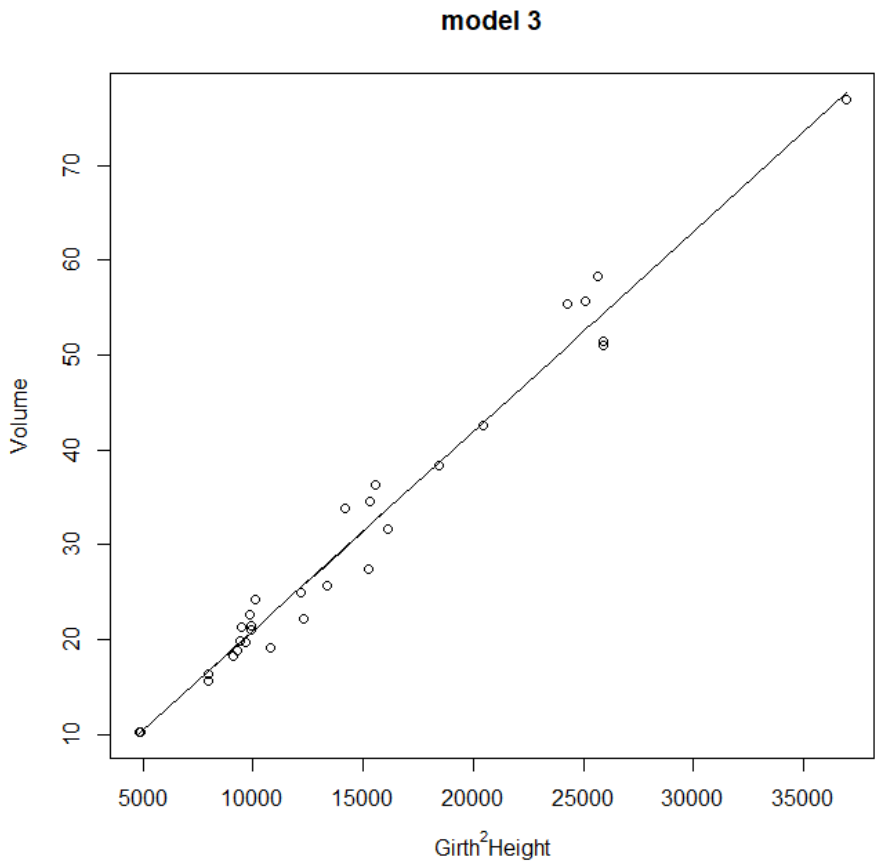




$$Volume = e^{-2.35332+2.19997 \log(Girth)}$$

Black cherry tree





$$Volume = e^{-1.17294+0.53004 \log(Girth^2 Height)}$$

Prediction

There is an old Danish proverb:

Prediction is difficult, especially when dealing with the future.

Or there is the following quote from Nostradamus:

For a long time, I have been making many predictions, far in advance, of events since come to pass, naming the particular locality. I acknowledge all to have been accomplished through divine power and inspiration.

Prediction is one of the two most common uses of a regression model - the other is explanation.

Prediction

Some examples of regression models used for prediction:

a) Predicting the volume of a cherry tree.

   - Ordinary regression used to predict volume of tree based on size information.

b) Predicting soil evaporation.

   - Ordinary regression used to predict rate of evaporation based on meteorological conditions.

c) Predicting the age of abalone.

   - Poisson regression used to predict age of abalone based on physical measurements.

d) Breast cancer diagnosis.

   - Logistic regression used to predict probability of breast cancer based on ?

Model selection and assessment

Model assessment Estimate the performance of a model by estimating its prediction error on new data.

Model selection Estimate the performance of different models and choose the best one.

Choice of assessment method relates to the final selection of model. What kind of assessment methods are available? Predictability of a model.

Loss function

A response variable Y, an input X and a prediction model $f(X)$ estimated from a training dataset $\mathcal{T}$.

The loss function $L(Y, f(X))$ measures error between Y and $f(X)$ and some popular choices follow;

Loss function quantifies how well a chosen model f describes the data and, a model with the smallest loss (smallest discrepancy between the prediction and data) is favoured.

Loss function

Some well known loss functions follow;

Continuous response, $Y \in \mathbb{R}$

- Squared error, $L(Y, f_\theta(X)) = (Y - f_\theta(X))^2$
- Absolute error, $L(Y, f_\theta(X)) = |Y - f_\theta(X)|$
- Log-likelihood loss, $L(Y, f_\theta(X)) = -2\log(p_{\theta(X)}(Y))$

Categorical response, $Y \in \{1, 2, ..., K\}$

- 0-1 loss, $L(Y, f_\theta(X)) = I(Y \neq_\theta f_\theta(X))$
- Log-likelihood loss, $L(Y, f_\theta(X)) = -2\sum_{k=1}^{K} I(Y = k)\log(p_k(X))$

Loss function expectation

Quantity of interest is a loss expectation

$$err = \mathbb{E}_f[L(Y, f(X))]$$

and it is approximated by the average loss over the n-training samples ($\boldsymbol{y} = (y_1, ..., y_n)$ and $\boldsymbol{x} = (x_i, ..., x_n)$)

$$\widehat{err} \approx \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$

Auto data

Gas mileage, horsepower, and other information for 392 vehicles.

mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
18	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
15	8	350.0	165	3693	11.5	70	1	buick skylark 320
18	8	318.0	150	3436	11.0	70	1	plymouth satellite
16	8	304.0	150	3433	12.0	70	1	amc rebel sst
17	8	302.0	140	3449	10.5	70	1	ford torino
15	8	429.0	198	4341	10.0	70	1	ford galaxie 500
14	8	454.0	220	4354	9.0	70	1	chevrolet impala
14	8	440.0	215	4312	8.5	70	1	plymouth fury iii
14	8	455.0	225	4425	10.0	70	1	pontiac catalina
15	8	390.0	190	3850	8.5	70	1	amc ambassador dpl

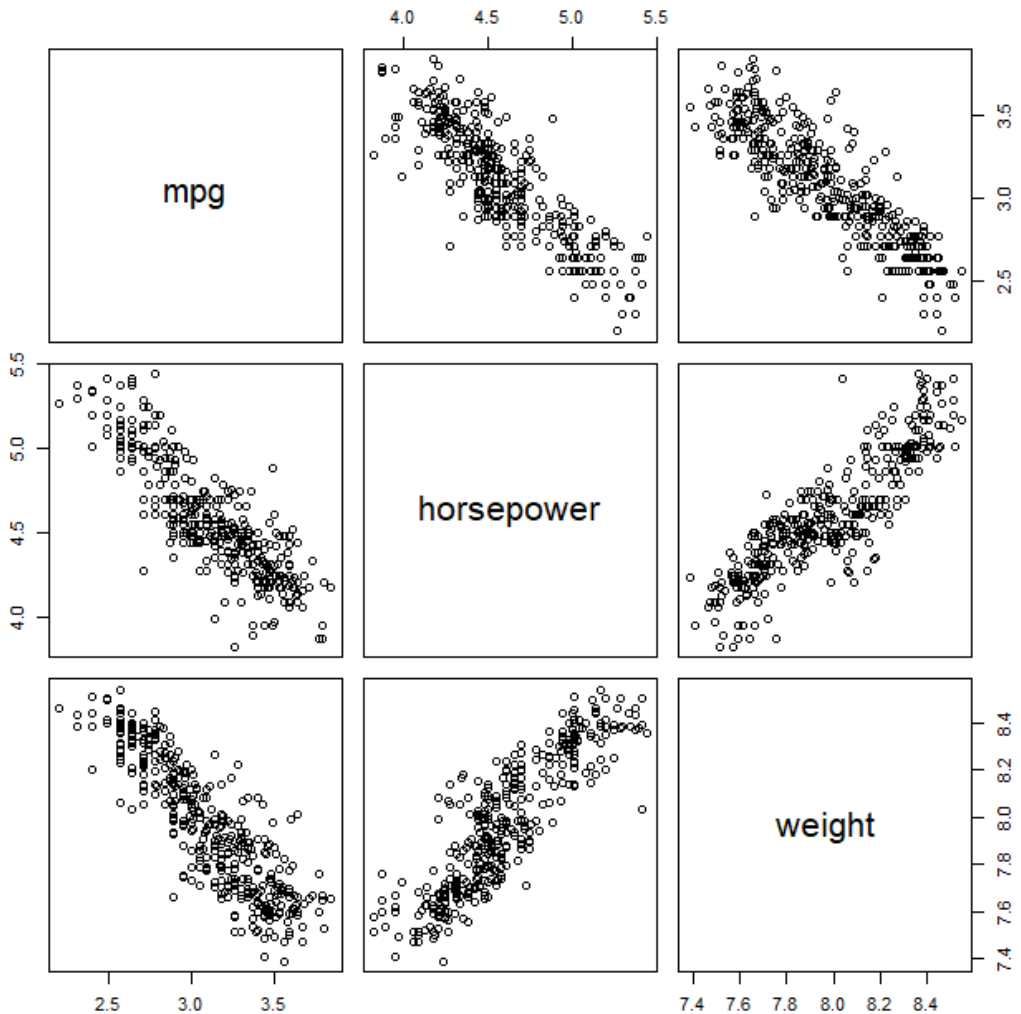
mpg	miles per gallon
cylinders	Number of cylinders between 4 and 8
displacement	Engine displacement (cu. inches)
horsepower	Engine horsepower
weight	Vehicle weight (lbs.)
acceleration	Time to accelerate from 0 to 60 mph (sec.)
year	Model year (modulo 100)
origin	Origin of car (1. American, 2. European, 3. Japanese)
name	Vehicle name

2

The goal was to create a model that relates the mpg to technical

Auto data

We consider two variables weight and horsepower and, a linear model in log-scale.



Auto data

```
glm(formula = log(mpg) ~ log(weight) + log(horsepower),
data = Auto)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.10954    0.28614  35.330  < 2e-16 ***
log(weight)     -0.67347    0.05698 -11.818  < 2e-16 ***
log(horsepower) -0.35985    0.04667  -7.711 1.05e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for gaussian family taken to be 0.0237138)

Null deviance: 45.2100  on 391  degrees of freedom
Residual deviance:  9.2247  on 389  degrees of freedom
AIC: -349.31
```

Question of interest

How do we evaluate how well this model predicts the mpg for new observations?

Cross validation

If we use the same data to fit and assess a model, we only know about the discrepancy of a model to one set of samples.

Ideally we would like to see how does the model performs when we have a new data in terms of accuracy of its predictions. Theories are judged by its predictive performance.

Cross validation is a model validation technique in which it assesses the predictive performance of the models for a new data (test data) outside a training data.
i.e., we use a training data to fit a model and a test data to validate a model.

K-fold cross validation

Given training data (Y_i, X_i), $i = 1, ..., n$ there is an predictor $f_\theta(x)$ to y. The K-fold cross validation is an iterative train-validation processes.

(1) Split the data into K roughly equal-sized parts (y_i^k, x_i^k), $i = 1, ..., n_k$ for $k = 1, ..., K$.

(2) For $k = 1, .., K$
(i) Train a model $\hat{f}^{-k}(x)$ on all parts except the k-th part.
(ii) Validate the model on the k-th part.

K-fold cross validation procedure

Given data (Y_i, X_i), $i = 1, ..., n$, an predictor model is $f_\theta(x)$.

(1) Split the data into K roughly equal-sized parts $S_k = \{y_i^k, x_i^k\}_{i=1}^{n_k}$, $k = 1, ..., K$.

(2) For $k = 1, .., K$

(i) Find the model $\hat{f}^{-k}(x)$ for the training set $\cup_{j \neq k} S_j$.
(ii) Find an error prediction on the test data S_k using the choice of loss function $L(y, \hat{f}^{-k}(x))$

$$\widehat{err}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} L(y_i^k, \hat{f}^{-k}(x_i^k))$$

(3) The cross-validation estimation of prediction error is

$$CV(\hat{f}) = \frac{1}{K} \sum_{k=1}^{K} \widehat{err}_k$$

Auto data

$$\log(mpg) = \log(horsepower) + \log(weight) + \epsilon\,, \epsilon \sim N(0, \sigma^2)$$

We fit a 8-folder cross-validation; 49 samples per partition.
Loss estimates over partitions;

	absolute	squared	-2log-like
partition 1	0.1186186	0.02421976	-0.8822753
partition 2	0.1133114	0.02073725	-1.0275694
partition 3	0.1211903	0.02537510	-0.8325743
partition 4	0.1176297	0.02306609	-0.9311738
partition 5	0.1397195	0.02883929	-0.6789373
partition 6	0.1178929	0.02424649	-0.8811472
partition 7	0.1212162	0.02478317	-0.8580983
partition 8	0.1092155	0.01884942	-1.1038791

The 8-folder cross-validation estimates of three prediction errors are

$$CV_1(\hat{f}) = 0.11984927\,, CV_2(\hat{f}) = 0.02376457\,, CV_3(\hat{f}) = -0.89945684$$

K-fold cross validation

How to decide the value of K?

- The value for K is chosen such that each train/test group of data samples is large enough to be statistically representative of the broader dataset.
- It is preferable to split the data sample into equal sized K groups.
- If a value for K is chosen that does not evenly split the data sample, then one group will contain a remainder of the examples.

Leave one out cross validation (LOOCV)

- Instead of taking a block out each time, we take a single datapoint out each time.
- Training set of $n - 1$ datapoints and each prediction is based on $n - 1$ data points.
- Equivalent to n-fold cross validation.

Leave one out cross validation (LOOCV)

Given $\{Y_i, X_i\}_{i=1}^n$, an predictor model is $f_\theta(x)$.
For $k = 1, ..., n$,

(1) The training set is all data points except (Y_k, X_k);
$S_k = \{Y_i, X_i\}_{i \neq k}$. Find the model $\hat{f}^{-i}(x)$ for the training set S_k.

(ii) Find an error prediction on the test data point X_k using the choice of loss function $L(y, \hat{f}^{-k}(x))$

$$\widehat{err}_k = L(y_i^k, \hat{f}^{-k}(x_i^k))$$

The cross-validation estimation of prediction error is

$$CV(\hat{f}) = \frac{1}{n} \sum_{k=1}^{n} \widehat{err}_k$$

Auto data

$$\log(mpg) = \log(horsepower) + \log(weight) + \epsilon\,, \epsilon \sim N(0, \sigma^2)$$

Loss estimates over partitions; (392 partitions)

```
                absolute      squared -2log-like
partition 1  0.02916213 0.0008504301  -1.870522
partition 2  0.03244227 0.0010525012  -1.861999
partition 3  0.06790444 0.0046110130  -1.711843
...
partition 390 0.16302087 0.026575804 -0.7824587
partition 391 0.09753578 0.009513228 -1.5048027
partition 392 0.23723417 0.056280051  0.4814673
```

The LOOCV estimates of three prediction errors are

$$CV_1(\hat{f}) = 0.12033510\,, CV_2(\hat{f}) = 0.02396192\,, CV_3(\hat{f}) = -0.88671228\,.$$

Cross validation

Useful to estimate a model's forecasting error but not a great tool for comparing models.

CV total error shows the patterns of predictions and does not necessarily show the difference between model.

The cross-validate loop has to be on the outside of anything that uses the data

- Can't run model selection, then cross-validate
- Can't choose the tuning parameter based on the data, then cross-validate

If you pick a model and then do cross-validation, youre just getting the error due to the parameter estimates: the one thing we can do analytically.

Stepwise model selection by AIC

With a penalty multiplier λ, we can specify the model size penalty in AIC

$$AIC(\lambda) = \lambda k - 2\log(\hat{L}).$$

i.e., If $\lambda = 2$, $AIC(\lambda)$ is a standard AIC. If $\lambda = \sqrt{n}$, $AIC(\lambda)$ is a BIC.

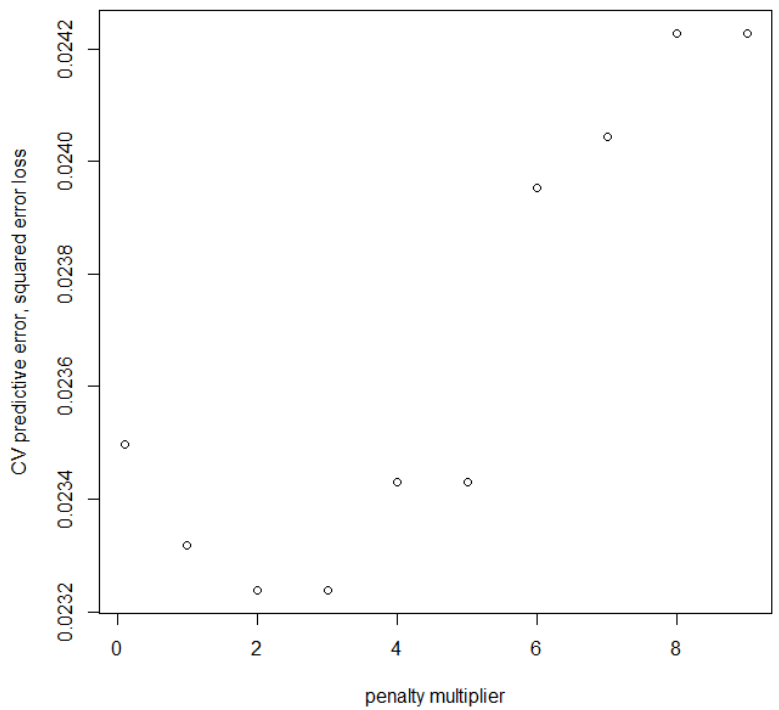
As λ gets larger, the penalty increases and a smaller model is likely to be chosen.
As λ gets smaller, the penalty decreases and a larger model is likely to be chosen.

Auto data

The full model is

$$\log(mpg) = \log(displacement) + \log(horsepower) + \log(weight) + \log(acceleration) + \log(horsepower) * \log(acceleration) + \epsilon\,.$$

and a model is chosen based on $AIC(\lambda)$. The 10-CV predictive errors of chosen models by $AIC(\lambda)$ follow;



Auto data

When $\lambda = 2, 3$, the 10-folder CV predictive squared error is minimized and the same linear combination of covariates is chosen for all three λ-values.

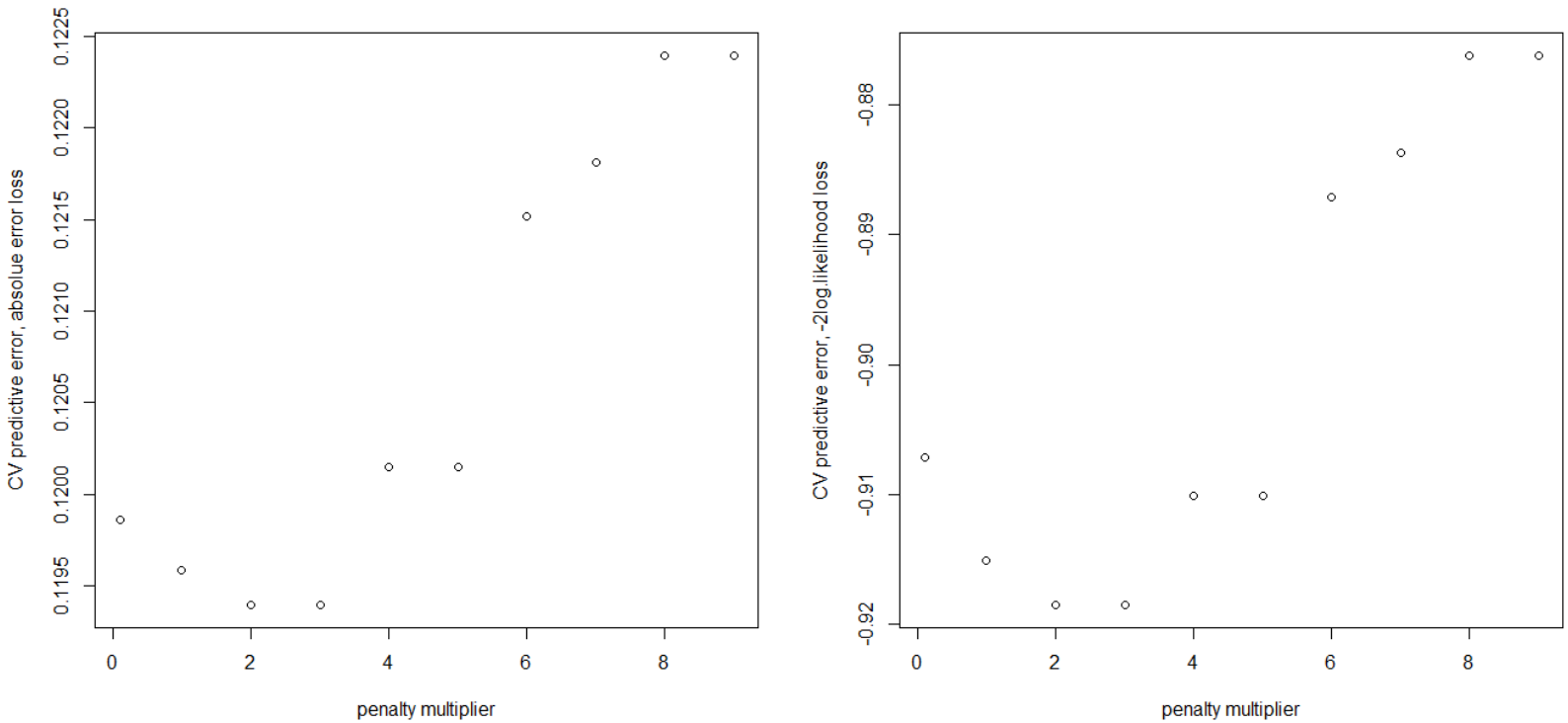
We now fit the model using all data

$$\log(\textit{mpg}) = 9.3480 - 0.1620 \log(\textit{displacement}) - 0.5156 \log(\textit{horsepower}) \\ -0.2915 \log(\textit{weight}) - 0.2689 \log(\textit{acceleration})$$

and the MSE is 0.02248063.

Auto data

Both absolute and log-likelihood loss yield the same result.



$\lambda = 2, 3$ minimizes the predictive loss; $\lambda = 2, 3$ maximizes the model predictability.

Bootstrap

Data $(Z_1, ..., Z_n)$ contains n independent samples from the true (but unknown) distribution f and the statistics of interest $S(z)$ has a probability distribution.

The estimator based on $(Z_1, ..., Z_n)$ is $\hat{S}(z)$ and it depends on f and n.

We would like to know this sampling distribution, but there are two difficulties;

(i) f is unknown.

(ii) Even if we know f, $S(z)$ may be such a complicated function of $(Z_1, ..., Z_n)$ that finding its distribution would exceed our analytic abilities.

Bootstrap

Important question;

How could we find the probability distribution of $S(z)$ without going through incredibly complicated analytic calculations?

Bootstrap

If we know f, we generate many datasets with n-samples from f and estimate S for each dataset.
B datasets and corresponding S estimates are

$$Z^{*1} = (Z_1^{*1}, ..., Z_n^{*1}) \sim g\,, S(Z^{*1}) = h(Z_1^{*1}, ..., Z_n^{*1})$$

$$Z^{*2} = (Z_1^{*2}, ..., Z_n^{*2}) \sim g\,, S(Z^{*2}) = h(Z_1^{*2}, ..., Z_n^{*2})$$

$$\vdots$$

$$Z^{*B} = (Z_1^{*B}, ..., Z_n^{*B}) \sim g\,, S(Z^{*B}) = h(Z_1^{*B}, ..., Z_n^{*B})$$

The empirical distribution of $(S(Z^{*1}), ..., S(Z^{*B}))$ approximate the distribution for $S(z)$.

Bootstrap

In real world, we don't know f and only have n-samples $(X_1, ..., X_n)$!

One idea is generating B datasets from the empirical distribution of $(X_1, ..., X_n)$.
The empirical distribution $\tilde{F}_n$ is

$$\tilde{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq y)$$

In other words, $\hat{f}_n$ puts mass $1/n$ at each X_i.

Bootstrap

Bootstrapping is a type of resampling where samples of the same size are repeatedly drawn, with replacement, from a single original sample.

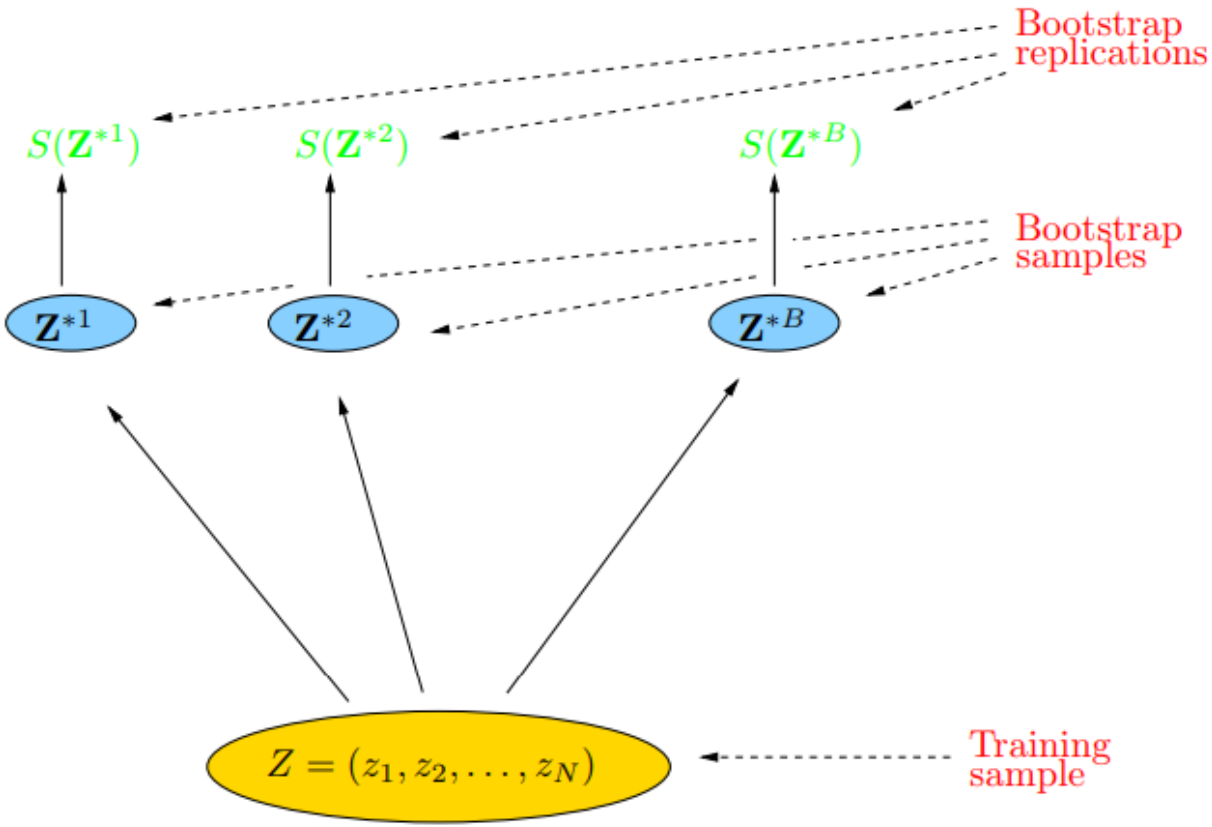
Efron (1979) [3] introduced the Bootstrap method to evaluate the accuracy of an estimator in the field of statistic inference.

For example, the accuracy of an estimator is defined by bias, variance, confidence intervals, prediction error or some other such measure.

Resampling technique used to estimate statistics on a population by sampling a dataset with replacement.

[3] Bootstrap methods: another look at the jackknife. Efron, B. Annals of Statistics. 7. 1-26 (1979)

Bootstrap



4

[4] The Element of Statistical Learning. Hastie, T., Tibshirani, R. and Friedman

Bootstrap

Data $(Z_1, ..., Z_n)$ contains n independent samples from the true (but unknown) distribution g and the statistic of interest is $S(z)$. The estimator based on $(Z_1, ..., Z_n)$ is $\hat{S}_n$ and it depends on g and n. The distribution of $\hat{S}_n$ is found by B bootstrap samples.

For $b = 1, ..., B$,

1. Bootstrap dataset Z^{*b} is a randomly draw dataset with replacement from $(Z_1, ..., Z_n)$, each sample the same size as the original training set n.
2. Estimate a statistic $S(Z^{*b})$ for Z^{*b}.

Estimate an aspect of the distribution of $S(Z)$ from bootstrap sampling.

Bootstrap

Some popular measures of the distribution;

- Sample mean: $\bar{S}^* = \sum_{b=1}^{B} S(Z^{*b})/B$
- Bias : $Bias = \bar{S}^* - \hat{S}_n$
- Variance: $Var(S^*) = \frac{1}{B}\sum_{b=1}^{B}(S(Z^{*b}) - \bar{S}^*)^2$
- $1 - \alpha$ Confidence interval:
  Sort sample statistics, $S(Z^{*1}) < ... < S(Z^{*B})$.
  The $1 - \alpha$ confidence interval (S_{lower}, S_{upper}) is defined by

$$\alpha/2 = \text{proportion of } (S(Z^*) < S_{lower})$$

$$1 - \alpha/2 = \text{proportion of } (S(Z^*) < S_{upper})$$

Bootstrap

The distribution of $S(Z^*) - \hat{S}_n$ is an approximation to the distribution of $\hat{S}_n - S(Z)$.

The variance of bootstrap estimator is $V(S^*) = \frac{1}{B}\sum_{b=1}^{B}(S(Z^{*b}) - \bar{S}^*)^2$

and by law of large numbers $V(S^*) \to V(\hat{S}_n)$.
The bootstrap is called to be consistent for $\hat{S}_n$ if, for all x,

$$P(a_n(\hat{S}_n - S) \leq x) - P(a_n(S^* - \hat{S}_n) \leq x) \to 0\,, \quad n \to \infty\,.$$

Consistency of the bootstrap implies (usually) that variance can be consistently estimated,

$$V(S^*)/V(\hat{S}_n) \to 1\,, \quad n \to \infty$$

The bias can also be estimated by $\mathbb{E}[S^*] - \hat{S}_n$, the difference between the bootstrap means and the observed values

$$\frac{\mathbb{E}[S^*] - S_n}{\mathbb{E}[\hat{S}_n] - S(z)} \to 1\,, \quad n \to \infty\,.$$

Black cherry tree

	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9

This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees.

Girth	Tree diameter (rather than girth, actually) in inches
Height	Height in ft
Volume	Volume of timber in cubic ft

Bootstrap - Black cherry tree

A simple linear regression model for $\log(Volume)$ is

$$\log(Volume) = -2.35332 + 2.1997 * \log(Girth)\,.$$

Estimate summary

```
                 Estimate Std. Error   t value     Pr(>|t|)
(Intercept)      -2.353325 0.23066284 -10.20244 4.180317e-11
log(cherry$Girth)  2.199970 0.08983455  24.48913 6.364197e-21
```

These estimates are based on 31 samples from the population. If we have more samples, would the distribution of estimates be different?

Bootstrap - Black cherry tree

1st bootstrap sample, Z^{*1}

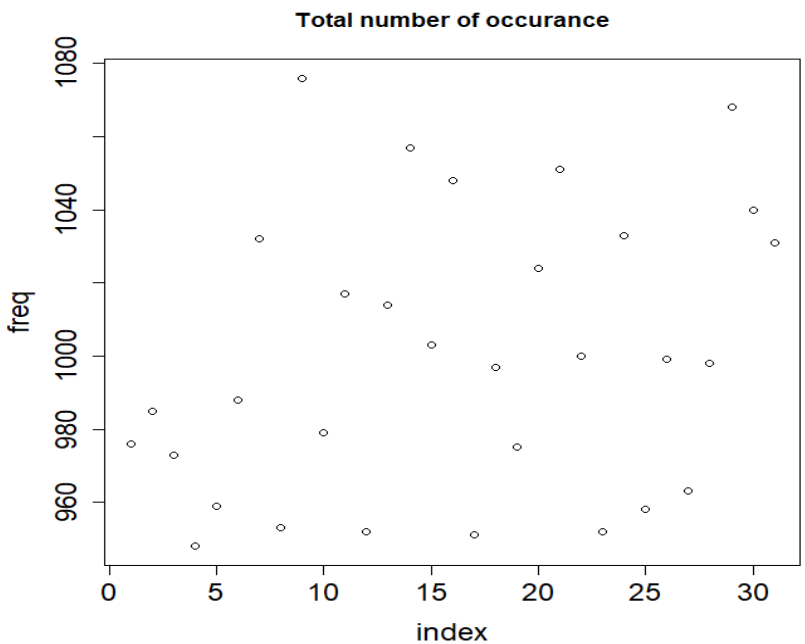
1 0 2 2 2 0 1 0 0 3 2 0 2 1 0 1 1 0 0 0 0 1 1 4 1 2 0 0 2 2 0

2nd bootstrap sample, Z^{*2}

1 0 0 1 0 2 1 1 2 0 4 1 0 1 4 1 1 2 0 0 1 2 0 1 2 0 0 2 0 0 1

3rd bootstrap sample, Z^{*3}

2 1 1 0 1 0 1 1 1 1 3 1 1 0 1 1 0 0 0 1 0 0 1 3 3 0 2 0 0 2 3



Some datapoints are appeared more than 1000 times and some are not.

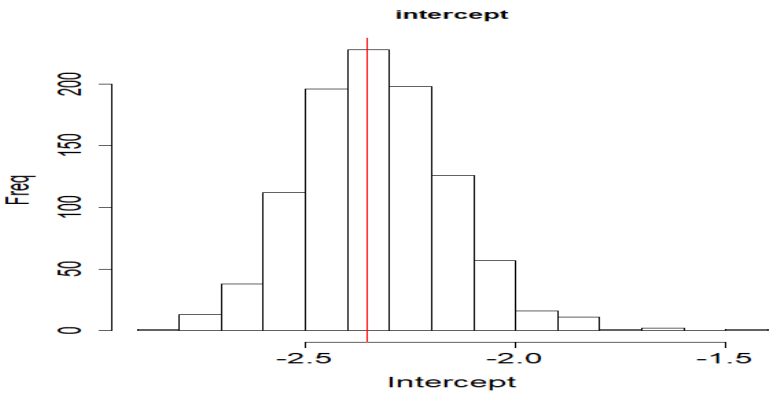
Bootstrap - Black cherry tree

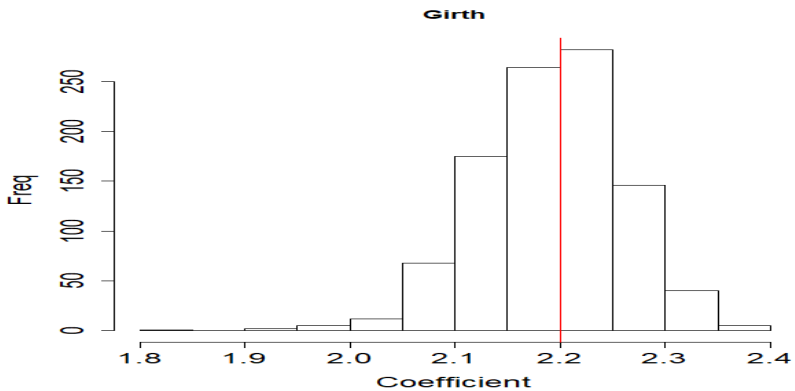
1st bootstrap sample, Z^{*1}
log(Volume) = -2.160316 + 2.123936*log(Girth)
2nd bootstrap sample, Z^{*2}
log(Volume) = -2.122209 + 20117111*log(Girth)
3rd bootstrap sample, Z^{*3}
log(Volume) = -2.240933 + 2.155957*log(Girth)
$\vdots$
1000th bootstrap sample, Z^{*3}
log(Volume) = -2.258098 + 2.174107*log(Girth)

Using 1000 S's (coefficients and intercept samples), we estimate an aspect of the distribution of $S(Z)$.

Bootstrap - Black cherry tree

Distribution of estimates based on 1000 bootstrap samples





Some key quantities of the distribution for $S(Z)$ are estimated;

	mean	bias	std	5%	95%
intercept	-6.0933006	-0.028637685	3.0003360	-11.1236345	-1.2880177
income	0.6059915	0.007258678	0.1655278	0.3393013	0.8825628
education	0.5382901	-0.007543831	0.1336451	0.3085326	0.7437392

Bootstrap - Black cherry tree

Number of bootstrap samples are changed from 10 to 5,000.

```
B=10
              mean        bias        std        5%       95%
intercept -2.276183  0.07714166 0.20574579 -2.551766 -1.992272
Girth      2.170509 -0.02946140 0.07961963  2.051785  2.272320

B=100
              mean          bias        std        5%       95%
intercept -2.355505 -0.0021798623 0.17971302 -2.656064 -2.093946
Girth      2.200516  0.0005460258 0.06841397  2.098669  2.318013

B=1000
              mean         bias        std        5%       95%
intercept -2.331896  0.021428797 0.17511167 -2.603781 -2.050586
Girth      2.191288 -0.008681432 0.06740208  2.078173  2.294095

B=5000
              mean         bias        std        5%       95%
intercept -2.344175  0.009149674 0.17012795 -2.609627 -2.058209
Girth      2.196345 -0.003624647 0.06499253  2.088327  2.298252
```

Bootstrap - Remark

- The basic bootstrap approach uses Monte Carlo sampling to generate an empirical estimate of the S's sampling distribution.
- Nonparametric distribution estimation.
- A larger number of samples of size n from a population, the empirical distribution based on samples gets closer to the population and the more accurate the relative frequency distribution of these estimates will be.

Reference

- Linear Models with R. Faraway, J. J. Chapman Hall/CRC Press (2015)
- The Elements of Statistical Learning. Hastie, T., Tibshirani, R., Friedman, J. Spring Series in Statistics. `http://web.stanford.edu/~hastie/ElemStatLearn/`
- An Introduction to the Bootstrap. Efron, B. and Tibshirani, R. (1993) Chapman and Hall, London.
- Bootstrap methods: another look at the jackknife. Efron, B. Annals of Statistics. 7. 1-26 (1979) `https://projecteuclid.org/euclid.aos/1176344552`