Applied Statistics Cheatsheet

Statistical Inference

An **inference** is a conclusion that patterns in the data are present in some broader context. A statistical inference is an inference justified by a probability model linking the data to the broader context.

- **Observational Study**: The group status of the subjects is established beyond the control of the investigator.
- **Randomized Experiment**: the investigator controls the assignment of experimental units to groups and uses a chance mechanism (like the flip of a coin) to make the assignment

Causal Inference

Statistical inferences of cause-and-effect relationships can be drawn from randomized experiments, but not from observational studies.

Counfounding Variables

A confounding variable is **related both to group membership and to the outcome**. Its presence makes it hard to establish the outcome as being a direct consequence of group membership

Inference to populations

Inferences to populations can be drawn from random sampling studies, but not otherwise.

Random sampling ensures that all subpopulations are represented in the sample in roughly the same mix as in the overall population. Again,

Simple Random Sample

A simple random sample of size n from a population is a subset of the population consisting of n members selected in such a way that every subset of size n is afforded the same chance of being selected.

Simple Linear Regression

$\mu\{Y|X\} = \beta_0 + \beta_1 X$

Model Assumption

1. Linearity
2. Normality: $Y|X \sim Normal$
3. Constant Variance: $\sigma(Y|X) = \sigma$
4. Independence

Least Square Method

Minimize $Q = \sum(Y_i - b_0 - b_1 X_i)^2 = \sum(Y_i - \hat{Y}_i)^2$

$$b_1 = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y}))}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b1\bar{X}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

Sampling Distribution

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \\ &= \frac{\mathrm{Var}\left(\sum_{i=1}^n Y_i (X_i - \bar{X})\right)}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \\ &= \frac{\sum_{i=1}^n \sigma^2 (X_i - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

- $\sigma^2_{\hat{\beta}_1} = \mathrm{Var}(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n (X_i - \bar{X})^2$
- $\sigma^2_{\hat{\beta}_0} = \mathrm{Var}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)\sigma^2$

$$SD(b_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)\sigma_x^2}}$$

$$SD(b_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\sigma_x^2}}$$

$$\frac{b_1 - \beta_1}{SE(B_1)} \sim t(n-2) \quad \frac{b_0 - \beta_0}{SE(B_0)} \sim t(n-2)$$

Matrix Form

$Y = X\beta + \epsilon$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ \epsilon_n \end{bmatrix}$$

$$\Psi = (Y - X\beta)^T (Y - X\beta)$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Confidence Intervals

$$SD(\mu(Y|X_0)) = \hat{\sigma}\sqrt{\tfrac{1}{n} + \tfrac{(X_0 - \bar{X})^2}{(n-1)\sigma_x^2}}$$

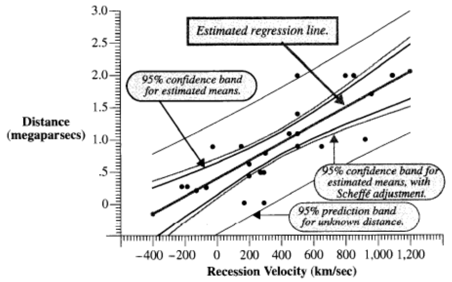
standardized $\mu(Y|X_0) \sim t(n-2)$

Prediction Interval

$$SD(Y|X_0) = \hat{\sigma}\sqrt{1 + \tfrac{1}{n} + \tfrac{(X_0 - \bar{X})^2}{(n-1)\sigma_x^2}}$$

standardized $Y|X_0 \sim t(n-2)$

Display 7.11 The 95% confidence band on the population regression line, the 95% confidence interval band for single mean estimates, and a 95% prediction interval band for the Big Bang example



R Squared

Total sum of squares(**SST**) $= \sum_{i=1}^n (Y_I - \bar{Y})^2$

Regression sum of squares(**SSR**) $= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$

Residual sum of squares(**SSE**) $= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

$$SST = SSR + SSE$$

$$R^2 = \left(\frac{SST - SSE}{SST}\right)\% = \left(\frac{SSR}{SST}\right)\%$$

$$MSE = \frac{SSE}{n-2}$$

Extra-Sums-of-Squares F-test

$H_0 : \beta_1 = 0$

$$F - stat = \frac{\left[\frac{\text{extra sums of squares}}{\#\beta \text{ being tested}}\right]}{\hat{\sigma}^2 \text{ from full model}} = \frac{\frac{SSR_{full} - SSR_{null}}{1}}{MSE}$$

Multiple Regression

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p} = \frac{SSE}{n-p}$$

$SD(b_j) = \sigma\sqrt{c_{ij}}$

standardized $b_j \sim t(n-p)$

c_{ij} is j^{th} diagonal element of $(X^T X)^{-1}$

Linear Combination Of Coefficients

$H_0 : c_0\beta_0 + c_1\beta_1 + ... + c_p\beta_p = 0$

$H_A : c_0\beta_0 + c_1\beta_1 + ... + c_p\beta_p \neq 0$

$est = c_0 b_0 + c_1 b_1 + ... + c_p b_p$

$$\begin{aligned} Var(est) &= c_0^2 Var(b_0)^2 + .. + c_p^2 Var(b_0) \\ &+ 2c_0c_1Cov(b_0, b_1) + ... + c_{p-1}c_pCov(b_{p-1}, b_p) \\ &= \hat{\sigma}^2 C(X^T X)^{-1} \end{aligned}$$

Extra-Sums-of-Squares F-test

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$

$$F - stat = \frac{\left[\frac{\text{extra sums of squares}}{\text{\# of } \beta\text{'s being tested}}\right]}{\hat{\sigma}^2 \text{ from full model}} = \frac{\frac{SSR_{full} - SSR_{reduce}}{df_{reduce} - df_f ull}}{MSE}$$

$$MSE = \frac{SSE}{n-p-1}$$

Adjusted R^2

Only for model comparison, not for model assessment.

$$\text{Adjusted } R^2 = \frac{\frac{SST}{n-1} - \frac{SSE}{n-p}}{\frac{SST}{n-1}}$$

Leverage

Measure the distance between explanatory values and the mean of explanatory values.

$H = X(X^TX)^{-1}X$

For ith observation: $h_i = H_{ii} = \dfrac{\partial \hat{Y}_i}{\partial Y_i}$

$SD(residual_i) = \sigma\sqrt{1-h_i}$, $\bar{h_i} = p/n$

Cutoff: larger than $2p/n$ (p : the number of parameters)

Studentized Residual

$$studres_i = \frac{residual_i}{\hat{\sigma}\sqrt{1-h_i}}$$

Roughly normal distributed. (Check absolute residual lager than 2)

Cook's Distance

$$D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{p\hat{\sigma}^2} = \frac{1}{p}(studres_i)^2(\frac{h_i}{1-h_i})$$

$\hat{Y}_{j(i)}$ is the jth fitted value without case i in the dataset

Cutoff: Larger than 1 → influential

Model Diagnosis

1. Residual v.s. Fitted Value Plot:
   - Pattern?
   - Non-constant Variance?
   - Influential Overservations?
2. QQ-Plot: Normality
3. Cook's Distance and Leverage Plot

Weighted Regression

$var(Y_i|X) = \dfrac{\sigma^2}{w_i}$

$Q = \sum w_i(Y_i - \hat{Y}_i)^2$

$\hat{\beta} = (X^TWX)^{-1}X^TWY$

$$W = \begin{bmatrix} w_1 & 0 & . & 0 \\ 0 & w_2 & 0 & 0 \\ . & . & . & . \\ 0 & 0 & 0 & w_n \end{bmatrix}$$

Ridge and Lasso Regression

$|\beta_j|$: L1-norm

Lasso: $\displaystyle\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$

$(\beta_j)^2$: L2-norm

Ridge: $\displaystyle\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda\sum_{j=1}^{p}(\beta_j)^2$

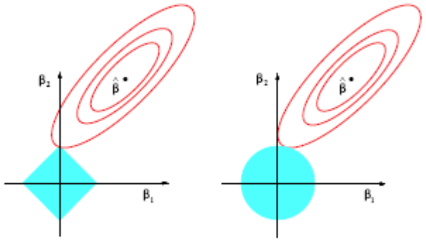
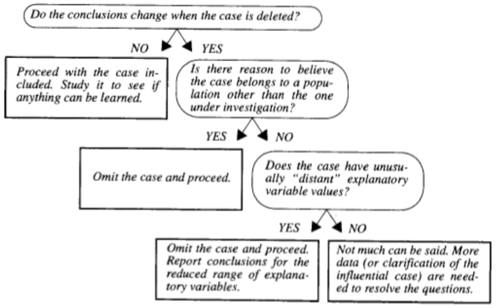


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Model Selection

Display 11.8 A strategy for dealing with suspected influential cases



Do the conclusions change when the case is deleted?

Proceed with the case included. Study it to see if anything can be learned.

Is there reason to believe the case belongs to a population other than the one under investigation?

Omit the case and proceed.

Does the case have unusually "distant" explanatory variable values?

Omit the case and proceed. Report conclusions for the reduced range of explanatory variables.

Not much can be said. More data (or clarification of the influential case) are needed to resolve the questions.

Strategies

Forward Selection

Start with the null model.

Backward Selection

Start with the full model.

Stepwise Selection

1. Start with null model.
2. Do on step of forward selection.
3. Do one step of backward elimination.
4. Repeat 2 and 3 until no explanatory variables can be added or removed.

Exhaustive Search Through All Subsets

Use the Cp statistics, R^2, Adjusted R^2, AIC and BIC.

Cp Statistic

The lower, the better.

$$Cp = p + (n-p)\frac{\hat{\sigma}^2 - \hat{\sigma}^2_{full}}{-\hat{\sigma}^2_{full}}$$

Akaike's Information Criteria(AIC)

The lower, the better.

$AIC = 2p + nlog(\hat{\sigma}^2) = 2p - 2log(L)$

Bayesian Information Criteria(BIC)

The lower, the better.

$BIC = p \cdot log(n) + n \cdot log(\hat{\sigma}^2) = p \cdot log(n) - 2log(L)$

Model Validation

For a new data set, define **mean square prediction error** as:

$$MSPE = \frac{\sum_{i=1}^{k=1}(Y_i - \hat{Y}_i)^2}{k}$$

Cross Validation

Serial Correlation

First-Order Autoregression Model {AR(1)}

$Y_t = \beta_0 + \beta_1X_1 + ... + \beta_kX_k + \epsilon_t$

$\epsilon_t = \alpha\epsilon_{t-1} + \psi_t$

$\psi_i \sim N(0, \sigma^2)$

Estimating α: Use the correlation coefficient between subsequent ordinary regression residuals.

Partial Auto Correlation Function(PACF)

A plot of the partial autocorrelations against lags.

cutoff: $[-\frac{2}{\sqrt{n}}, \frac{2}{\sqrt{n}}]$

Large-Sample Test

If one estimates the serial correlation coefficient from a series of n independent observations with constant variance, the estimate has an approx. normal distribution with mean 0 and standard deviation $\frac{1}{\sqrt{1}}$.

Variance Inflation Factor

$$\widehat{Var}(\hat{B}j) = \hat{\sigma^2}(X^TX)^{-1}_{j+1,j+1}$$
$$= \frac{\hat{\sigma^2}}{(n-1)\widehat{Var}(X_j)} \frac{1}{1-R_j^2}$$

$R_j^2 := R^2$ for the regression of X_j on the other covariates.

$$VIF := \frac{1}{1-R_j^2}$$

Cut-off rule of thumb:

$VIF(\hat{B}_j) > 5$ for high multicollinearity

Bootstrap

Assumption: **Independence** between samples.

Non-parametric Bootstrap

Repeated re-sampling with replacement.

The number of different bootstrap samples is $\binom{2n-1}{n}$ for sample size n.

Can obtain statistics(e.g. mean, standard deviation) of the estimator with only one set of samples.

Bootstrap Regression

Let X be the explanatory variable, Y be the responsive variable.

Case-based

1. Re-sample based on (X, Y) pairs.
2. Fit a regression model on the bootstrap sample.
3. Repeat 1 and 2 several times.

Problem: When X is an indicator variable.

Residual-based

1. Fit regression model on the original sample.
2. Re-sample the residuals from 1
3. Add bootstrap residuals to $hatY$ to form the new Y'.
4. Fit a regression model on Y' and X.
5. Repeat 1-3 several times.

Solve the problem of X being extremely skewed.

Bootstrap Confidence Interval

Useful when the distribution of estimator is skewed or not normal.

Use quantiles of the bootstrap estimations as the boundary of the confidence interval.

Parametric Bootstrap

1. A parametric model is fitted to the data (Often by maximum likelihood)
2. Samples of random numbers are drawn from this fitted model
3. Calculate the estimate/quantity of interest from these samples
4. Repeat 2 and 3 many times as for other bootstrap methods

Parametric bootstrap will be more accurate than non-parametric bootstrap if the parametric assumption is true, less accurate if false.

Natural Cubic Spline

splines package in R

1. Dividing the range of X into intervals.
2. Inside each interval, a cubic polynomial model is fitted.
3. At the interval split points(knots), the cubic polynomial are continuous and have continuous first and second derivatives.

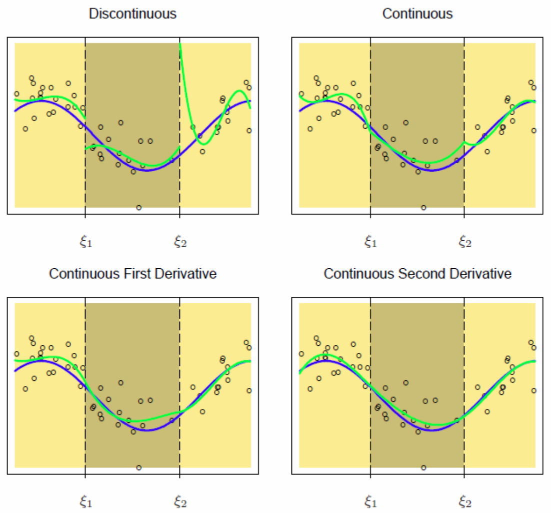


FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

R example:

$fit2 < -lm(ozone \sim ns(temperature, knots = c(70, 90)))$

$df = 2 + \#$ of knots

When to use natural cubic splines?

1. Smoothing.
2. To model confounding variables.
3. Higher order terms are required for X

Canonical Correlation Analysis CCA

CCA finds linear combinations in the two sets that have the largest possible correlations.

R command: **cancof**

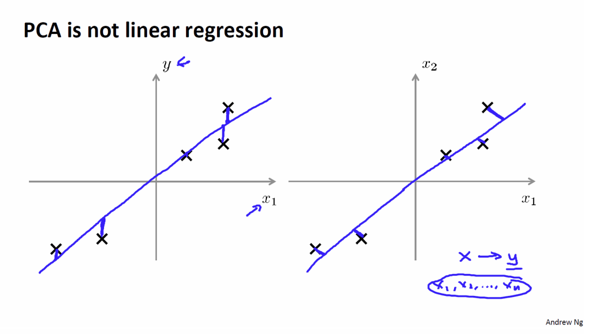
Bartlett's Chi-square Test

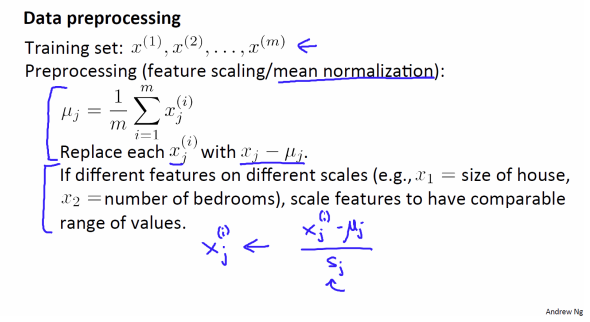
How many pairs of canonical variables are significant?

$$V = -\left((n-1) - \frac{p+q+1}{2}\right) ln(k)$$

n: number of obvservations
p: number of X variables minus number times test applied
q: number of Y variables minus number times test applied
k: $(1-r_t^2)(1-r_{t+1}^2)...(1-r_T^2)$
r_j^2: the squared correlation between the jth pair of canonical variables.
T: The totla number of canonical variables
t: number of times test applied
$V \sim \chi^2_{pq}$ under H_0 : the pair is significant.

Principle Component Analysis


PCA is not linear regression


Data preprocessing

Training set: $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ ←

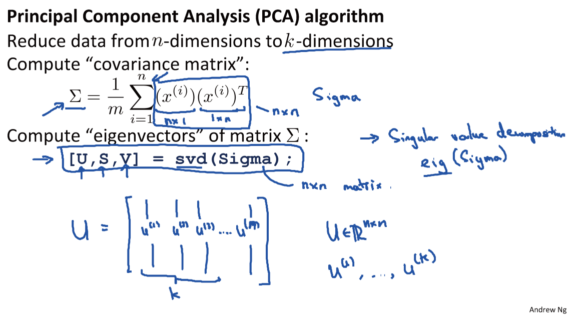
Preprocessing (feature scaling/$\underline{\text{mean normalization}}$):

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

Replace each $\underline{x_j^{(i)}}$ with $\underline{x_j - \mu_j}$.

If different features on different scales (e.g., $x_1 =$ size of house, $x_2 =$number of bedrooms), scale features to have comparable range of values.

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{s_j}$$



Principal Component Analysis (PCA) algorithm

Reduce data from n-dimensions to k-dimensions

Compute "covariance matrix":

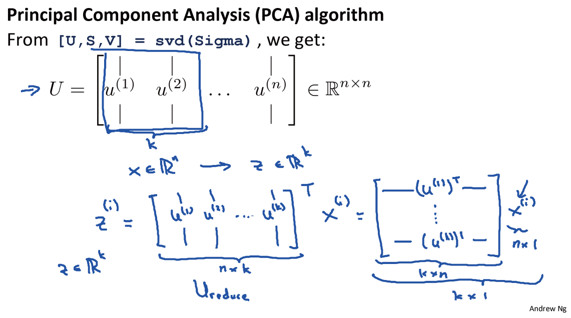
$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T$$

Compute "eigenvectors" of matrix Σ:

```
[U,S,V] = svd(Sigma);
```

Andrew Ng

Sigma
n×n
n×1
1×n
→ Singular value decomposition
eig(Sigma)
n×n matrix
$U = [u^{(1)}\ u^{(2)}\ u^{(3)} \dots u^{(n)}]$
k
$U \in \mathbb{R}^{n\times n}$
$u^{(1)}, \dots, u^{(k)}$



Principal Component Analysis (PCA) algorithm

From `[U,S,V] = svd(Sigma)`, we get:

$$U = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \dots & u^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n\times n}$$

Andrew Ng

k
$x \in \mathbb{R}^n \rightarrow z \in \mathbb{R}^k$
$z^{(i)} = [u^{(1)}\ u^{(2)} \dots u^{(k)}]^T x^{(i)} = \begin{bmatrix} —(u^{(1)})^T— \\ \vdots \\ —(u^{(k)})^T— \end{bmatrix} x^{(i)}$
n×k
U_{reduce}
k×n
n×1
k×1
$z \in \mathbb{R}^k$

$z = U_{redude} * X$

- Sensitive to scale: **Standardize before fitting!**
- Use z in regression: Solve multicollinearity and increase degrees of freedom.
- Benefits: **Low dimension** and **No correlation of X**.
- Drawback: Hard to interpretate.

Some True/False Questions

1. A multiple linear regression model should only include explanatory variables that have a normal distribution. **FALSE**
2. Adding an extra explanatory variable to a simple linear regression model **cannot** increase the significance, as measured by the t-test, for the explanatory variable that is already in the model. **FALSE**
3. The main reason logistic regression is preferred to multiple linear regression for a categorical response with two categories is that the logistic regression model allows for the non-constant variance of the response. **FALSE**
4. The multiple linear regression $\mu(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ will have the same R^2 value as the multiple linear regression $\mu(Y|Z) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$ where Z_1, Z_2, Z_3 are the first three principal component variables of X_1, X_2, X_3 **TRUE**
5. The bootstrap cannot be used for hypothesis testing. FALSE
6. It is possible for the first three principal component variables $(Z_1, ..., Z_3)$ from a principal components analysis of ten variables $(X_1, ..., X_10)$ to explain 100% of the total variation in the ten original variables $X_1, ..., X_10$. **TRUE**
7. The estimated mean response for the regression $\mu(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(X_1 \times X_2)$ corresponding to a particular set of explanatory variable values $X_1 = 3, X_2 = 2$ is 15. Based on this information we would estimate that there is more than a 50% chance that the response variable, given $X_1 = 3, X_2 = 2$, would take on a value greater than 17. **FALSE**
8. Modelling marital status (Single, Married, Divorced) as a categorical explanatory variable in a Poisson log-linear regression model will require three parameters to be estimated, not including the intercept and other variables included in the model. **FALSE**
9. A fitted linear regression model based on 500 observations returns $b_6 = 0.21, SE(b_5) = 0.06$. You are given two 90% confidence intervals (a) (0.09,0.33) and (b) (0.13,0.42) that have been computed based on the fitted regression model. One of the intervals was computed using the bootstrap and another using the standard linear regression theory. Interval (b) (0.13,0.42) is the confidence interval computed using the standard theory. **FALSE**
10. There are 64 possible logistic regression models that can be fitted in a situation there are seven explanatory variables and we are only interested in models that contain these seven variables; that is, we are not including interactions terms, terms for curvature etc. **FALSE**