Assignment Project Exam Help

Linear Regression

https://powcoder.com

MAST90083 Computational Statistics and Data Mining

Dr Karim Seghouane

School of Mathematics & Statistics
The University of Melbourne

Add WeChat powcoder

Outline

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Statistical Models

Assignment Project Exam Help

- What is the simplest mathematical model that describes the relationship between two variables ?
- Straight line

https://powcoder.com

- Statistical models are fitted for a variety of reasons:
- Explanation and prediction: Uncover causes by studying the relationship between an interested variable (the response) and a set of variables called the explanatory variables& use the model for prediction

Add WeChat powcoder

- Examine and test scientific hypotheses

Linear Models

Assignment Project Exam Help

- Linear models have a long history in statistics, but even in today's computer era they are still important and widely used in supervised learning.

https://powcoder.com

- They are simple and provide a picture of how the inputs affect the output
- For prediction purposes they can sometimes outperform fancier nonlinear models, particularly in small sample cases, low signal-to-noise ratio or sparse data

Add WeChat powcoder

- We will study some of the key questions associated with the linear regression model

Linear regression

Given a vector of input variables $\mathbf{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ and a response variable $y \in \mathbb{R}$

$$y \approx f(\mathbf{x})$$

where

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- The linear model assumes that the dependence of y on $x_1, x_2, \ldots, x_p$ is linear. Or, it can be well approximated by the linear relationships.
- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.
- The β_j's are the unknown parameters that need to be determined.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Parameter Estimation

Assignment Project Exam Help

- We have at our disposal a set of training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ from which to estimate the parameters $\boldsymbol{\beta}$

https://powcoder.com

- The most popular method is least squares where β is obtained by minimizing the residual sum of squares

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Add WeChat powcoder

- **What is the statistical interpretation ?**

Statistical Interpretation

Assignment Project Exam Help

- From a statistical point of view, this represents the maximum likelihood estimation of $\boldsymbol{\beta}$ assuming

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \quad i = 1, \dots, N$$

https://powcoder.com

- and $\epsilon_1, \dots, \epsilon_N$ are independent random samples from $\mathcal{N}(0, \sigma^2)$, $\sigma > 0$ an unknown parameter so that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$
- Taking X as the $N \times (p+1)$ matrix with each row an input vector and 1 in the first position and $\mathbf{y}$ an $N \times 1$ vector of responses: $E(\mathbf{y}) = X\boldsymbol{\beta}$ and $\mathrm{Cov}(\mathbf{y}) = \sigma^2 \mathbf{I}_N$, so that

$$\mathbf{y} \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma^2 \mathbf{I}_N\right)$$

Add WeChat powcoder

Note 1

Matrix Form

Assignment Project Exam Help

- The residual sum of square takes the matrix form

$$\mathrm{RSS}(\beta) = (\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta)$$

https://powcoder.com

- Assuming that X has full column rank or $X^\top X$ is positive definite gives a unique solution

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top \mathbf{y}$$

Add WeChat powcoder

- and the fitted values at the training inputs are (Note 2)

$$\hat{\mathbf{y}} = X\hat{\beta} = X\left(X^\top X\right)^{-1} X^\top \mathbf{y} = H\mathbf{y}$$

Geometric Interpretation

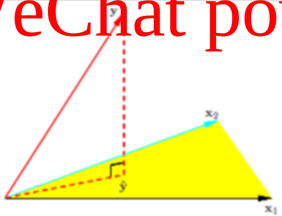
Assignment Project Exam Help

- The hat matrix H is square and satisfies: $H^2 = H$ and $H^\top = H$ (Note 3)
- H is the orthogonal projector onto $V = Sp(X)$ (column space of X or the subspace of $\mathcal{R}^N$ spanned by the column vectors of X)

https://powcoder.com

- and $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $Sp(X)$
- The residual vector $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to this subspace

Add WeChat powcoder



Statistical Properties

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- Assuming the model $\mathbf{y} \sim \mathcal{N}\left(X\beta, \sigma^2 \mathbf{I}_N\right)$ gives

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \left(X^\top X\right)^{-1}\right)$$

or

$$\hat{\mathbf{y}} \sim \mathcal{N}\left(X\beta, \sigma^2 H\right)$$

- where σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{N-p-1}(\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$$

- and $(N-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}$
- Furthermore $\hat{\sigma}^2$ and $\hat{\beta}$ are statistically independent. (**Why** ?)

Assessing the Accuracy of the coefficient estimates

Assignment Project Exam Help

- Approximate confidence set for the parameter vector β

$$C_{\beta} = \{\beta | \left(\hat{\beta} - \beta\right)^{\top} X^{\top} X \left(\hat{\beta} - \beta\right) \leq \hat{\sigma}^2 \chi^{(1-\alpha)}_{p+1}\}$$

- Test the null hypothesis of $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ using

https://powcoder.com

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

- and v_j is the jth diagonal element of $(X^{\top} X)^{-1}$. Under H_0, z_j is t_{N-p-1}.

Add WeChat powcoder

- Testing for a group of variables, H_0 : smaller model is correct

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} \sim F_{p_1 - p_0, N - p_1 - 1}$$

Note 4

Gauss-Markov Theorem

Assignment Project Exam Help

- LSE of β have the smallest variance among all linear unbiased estimates
- Assuming the estimation of $\theta = \mathbf{a}^\top \beta$, then its LSE is

https://powcoder.com

$$\hat{\theta} = \mathbf{a}^\top \hat{\beta} = \mathbf{a}^\top \left(X^\top X\right)^{-1} X^\top \mathbf{y} = \mathbf{c}_0^\top \mathbf{y}$$

- Then $E\left(\mathbf{a}^\top \hat{\beta}\right) = \mathbf{a}^\top \beta$ is unbiased and

Add WeChat powcoder

$$\mathrm{Var}\left(\hat{\theta}\right) \leq \mathrm{Var}\left(\mathbf{c}^\top \mathbf{y}\right)$$

- for any other linear estimator $\tilde{\theta} = \mathbf{c}^\top \mathbf{y}$

Reducing the MSE

Assignment Project Exam Help

- The LSE has the smallest MSE of all linear estimators with no bias.
- Biased estimator can generate smaller MSE

https://powcoder.com

- Shrinking a set of coefficients to zero may result in a biased estimate
- MSE is related to the prediction accuracy of a new response $y_0 = f(x_0) + \epsilon_0$ at input x_0

Add WeChat powcoder

$$E\left[y_0 - \tilde{f}(x_0)\right]^2 = \sigma^2 + E\left[x_0^\top \tilde{\beta} - f(x_0)\right]^2 = \sigma^2 + MSE\left[\tilde{f}(x_0)\right]$$

Note 5

Assessing the Overall Accuracy of the Model

Assignment Project Exam Help

- Quantify *how well the model fits the observations*
- Two quantities are used: the residual standard error (RSE) which measures the lack of fit

https://powcoder.com

$$RSE = \sqrt{\frac{\mathrm{RSS}\left(\hat{\beta}\right)}{N-p-1}} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-p-1}}$$

- R^2 Statistics

Add WeChat powcoder

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}}$$

- measure the amount of variability $\left(\mathrm{TSS} = \sum (y_i - \bar{y})^2\right)$ removed by the model

Other Considerations in the Regression Model

Assignment Project Exam Help

- Correlation of the error terms
- Interactions or collinearity
- Categorical predictors and their interpretation (two or more categories).
- Non-linear effects of predictors
- Outliers and high-leverage points
- Multiple outputs

https://powcoder.com

Add WeChat powcoder

Correlation of the error terms

Assignment Project Exam Help

- Use generalized LS

$$\mathrm{RSS}(\beta) = (\mathbf{y} - X\beta)^\top \Sigma^{-1} (\mathbf{y} - X\beta)$$

https://powcoder.com

- Similar to assuming

$$\mathbf{y} \sim \mathcal{N}(X\beta, \Sigma)$$

Add WeChat powcoder

- Still least square but using a different metric matrix Σ instead of $\mathbf{I}$

Note 6

Interactions or collinearity

Assignment Project Exam Help

- Variables closely related to one another which leads to linear dependence or collinearity among the columns of X.
- It is difficult to separate the individual effects of collinear variables on the response.

https://powcoder.com

$$\operatorname{Var}\left(\hat{\beta}\right) = \sigma^2 \left(X^\top X\right)^{-1}$$

Add WeChat powcoder

- Collinearity has considerable effect on the precision of $\hat{\beta} \mapsto$ large variances, wide confidence interval and low power of the tests
- It is important to identify and address potential collinearity problems

Detection of collinearity

Assignment Project Exam Help

- Look at the correlation matrix of the variables to detect pair of highly correlated variables
- Collinearity between three or more variables compute *variance inflation factor VIF* for each variable

https://powcoder.com

$$\mathrm{VIF}\left(\hat{\beta}_j\right) = \frac{1}{1 - R_j^2}$$

Add WeChat powcoder

- Geometrically $1 - R_j^2$ measures how close x_j is to the subspace spanned by X_{-j}

$$\mathrm{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{1 - R_j^2} \leq \lambda_{max}\lambda_{min}^{-1} = \kappa\left(X\right)^2$$

- Examine the eigenvalues and eigenvectors **Note 7**

Categorical predictors

- Also referred as categorical or discrete predictors or variables.
- Prediction task is called regression for quantitative output and classification for qualitative outputs
- Qualitative variables are represented by numerical codes

$$x_i = \begin{cases} 1 & \text{if the } ith \text{ experiment is a success} \\ 0 & \text{if the } ith \text{ experiment is a failure} \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } ith \text{ exp. is a success} \\ \beta_0 + \epsilon_i & \text{if the } ith \text{ exp. is a failure} \end{cases}$$

Note 8

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Non-linear effects of predictors

Assignment Project Exam Help

- The linear model assumes a linear relationship between the response and predictors
- The true relationship between the response and the predictors may be non-linear
- Polynomial regression is simple way to extend linear models to accommodate non-linear relationships
- In this case non-linearity is obtained by considering transformed versions of the predictors
- The parameters can be estimated using standard linear regression methods.

https://powcoder.com

Add WeChat powcoder

Note 9

Outliers

Assignment Project Exam Help

- Different reasons can lead to outliers. Example: incorrect recording of an observation
- The residual as an estimate of the error can be used to identify outliers by examination for extreme values

https://powcoder.com

- Better use the studentized residuals

$$E[\mathbf{e}] = E\left[\left(I_p - H\right)\mathbf{y}\right] = \left(I_p - H\right)\delta$$

Add WeChat powcoder

- If the diagonal of H is not close to 1 (small) then $\mathbf{e}$ reflects the presence of outliers.

Note 10

High-leverage points

Assignment Project Exam Help

- Observations with high leverage have an unusual value for x_i
- Difficult to identify when there are multiple predictors
- To quantify an observation's leverage use the leverage statistic

https://powcoder.com

$$h_i = \frac{1}{N} + (N-1)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

- $\mathbf{S}$ is the ample covariance matrix, $\mathbf{x}_i$ the ith row of X and $\bar{\mathbf{x}}$ the average row

Add WeChat powcoder

- The leverage statistic $\frac{1}{n} \leq h_i \leq 1$ and the average is $(p+1)/n$.
- If an observation has h_i greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage.

Note 11

Multiple outputs

Assignment Project Exam Help

- Multiple outputs $y_1, \dots, y_K$ need to be predicted from $x_1, \dots, x_p$ where a linear model is assumed for each output

https://powcoder.com

$$\mathbf{y}_k = \beta_{0k} + \sum_{j=1}^{p} \mathbf{x}_j \beta_{jk} + \epsilon_k = f_k(X) + \epsilon_k$$

- In matrix notation $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ where $\mathbf{Y}$ is $N \times K$, $\mathbf{X}$ is $N \times (p+1)$, $\mathbf{B}$ is $(p+1) \times K$ (matrix of parameters) and $\mathbf{E}$ is $N \times K$ matrix of errors.

Add WeChat powcoder

$$RSS(\mathbf{B}) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 = \mathrm{tr}\left[(\mathbf{Y} - \mathbf{XB})^\top (\mathbf{Y} - \mathbf{XB})\right]$$

Multiple outputs

Assignment Project Exam Help

- The least squares estimates have exactly the same form

$$\hat{\mathbf{B}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

https://powcoder.com

- In case of correlated errors $\epsilon \sim \mathcal{N}(0, \Sigma)$ the multivariate criterion becomes

Add WeChat powcoder

$$RSS(\mathbf{B}) = \sum_{i=1}^{N} \left(y_i - f\left(x_i\right)\right)^\top \Sigma^{-1} \left(y_i - f\left(x_i\right)\right)$$

Note 11

Why ?

Assignment Project Exam Help

- *The least squares estimates is in most cases not satisfied when a large number of potential explanatory variables are available*
- *Improving prediction accuracy:* LSE often has low bias but large variance, sacrifice a little bit of bias to reduce the variance of the predicted values and improve overall prediction accuracy
- *Interpretation:* Do all the predictors help to explain $\mathbf{y}$? determine a smaller subset with strongest effects and sacrifices the small details

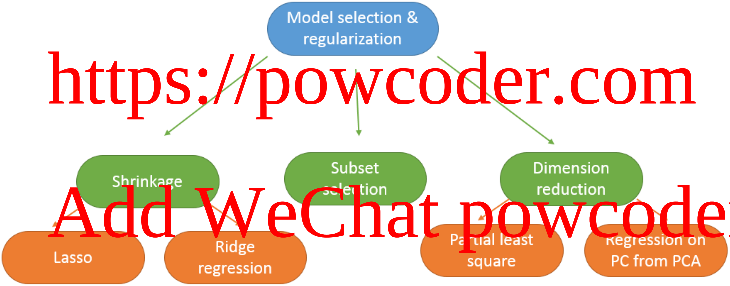
https://powcoder.com
Add WeChat powcoder

Linear Model Selection and Regularization

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder



Deciding on the Important Variables

Assignment Project Exam Help

Subset selection

- **All subsets** or **best subsets** regression (examine all potential combinations)
https://powcoder.com
- **Forward selection** - begin with intercept and iteratively add one variable.
- **Backward selection** - begin with the full model and iteratively remove one variable.
Add WeChat powcoder
- **What is best for cases where $p > n$?**

Best Subset

Assignment Project Exam Help

- Retain a subset of predictor and eliminate the rest
- LSE is used to obtain the coefficients of the retained variables
- For each $k \in \{0, 1, 2, \ldots, p\}$ find the subset k that gives the smallest residual
- The choice of k is obtained using a criterion and involves a tradeoff between bias & variance
- Different criteria $\leftarrow$ minimizes an estimate of the expected prediction error

https://powcoder.com

Add WeChat powcoder

Infeasible for large p

Note 12

Forward Selection

Assignment Project Exam Help

- Sequential addition of predictors $\rightarrow$ forward stepwise selection
- Starts with the intercept and sequentially add the predictor that most improve the fit

https://powcoder.com

- Add predictor producing the largest value of

$$F = \frac{\mathrm{RSS}\left(\hat{\boldsymbol{\beta}}_i\right) - \mathrm{RSS}\left(\hat{\boldsymbol{\beta}}_{i+1}\right)}{\mathrm{RSS}\left(\hat{\boldsymbol{\beta}}_{i+1}\right) / (N-k-2)}$$

Add WeChat powcoder

- Use 90th or 95th percentile of $F_{1,N-k-2}$ as F_e

Note 14

Backward Elimination

Assignment Project Exam Help

- Start with the full model and sequentially remove predictors
- Use F_d to choose the predictor to delete (smallest value of F)
- Stop when each predictor in the model produces $F > F_d$
- Can be used only when $N > p$

https://powcoder.com

$F_d \simeq F_e$

Add WeChat powcoder

Note 15

Alternative: Shrinkage Methods

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- Subset selection produces produce an interpretable model with possible lower prediction error than the full model
- The selection is discrete $\rightarrow$ often exhibits high variance
- Shrinkage methods are continuous and don't suffer as much from high variability
- We fit a model containing all p predictors using a technique that **constrains** or **regularizes** the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- Shrinking the coefficient estimates can significantly reduce their variance (not immediately obvious).

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Ridge Regression

Assignment Project Exam Help

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\underbrace{\lambda\sum_{j=1}^{p}\beta_j^2}_{\text{The shrinkage penalty.}}$$

https://powcoder.com

The tuning parameter.

- It serves to control the relative impact of the penalty on the regression coefficient estimates.
- Selecting a good value for λ is critical.
- cross-validation is used for this.

The shrinkage penalty.

- It is small when $\beta_1, \beta_2, \ldots, \beta_p$ are close to zero.
- So it has the effect of *shrinking* the estimates of β_j towards zero.

Add WeChat powcoder

Ridge Regression

Assignment Project Exam Help

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \underbrace{\sum_{j=1}^{p}\beta_j^2}_{\text{The shrinkage penalty.}}$$

https://powcoder.com

Add WeChat powcoder

Accuracy $\leftrightarrow$ Complexity

Ridge Regression

Assignment Project Exam Help

- Ridge regression shrinks the regression coefficients by constraining their size

https://powcoder.com

- This is the approach used in neural networks where it is known as weight decay
- The larger the value of λ, the greater the amount of shrinkage
- β_0 is left out of the penalty term

Add WeChat powcoder

Note 16

Ridge Regression

Assignment Project Exam Help

- Because we have now the addition of the penalty term, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant!

https://powcoder.com

- It is best to apply ridge regression after *standardizing the predictors*, using the formula

Add WeChat powcoder

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{ij})^2}}$$

Ridge Regression

Assignment Project Exam Help

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta) + \lambda\beta^\top\beta$$

$$\hat{\beta}^{ridge} = \left(X^\top X + \lambda \mathbf{I}\right)^{-1} X^\top \mathbf{y}$$

https://powcoder.com

- The ridge solution is a linear function of $\mathbf{y}$
- Avoid singularity when $X^\top X$ is not full by adding a positive constant to the diagonal of $X^\top X$.
- For orthogonal predictor $\hat{\beta}^{ridge} = \gamma\hat{\beta}$ where $0 \leq \gamma \leq 1$
- The effective degrees of freedom of the ridge regression fit is

Add WeChat powcoder

$$\mathbf{df}(\lambda) = \mathrm{tr}\left(X[X^\top X + \lambda \mathbf{I}]^{-1} X^\top\right)$$

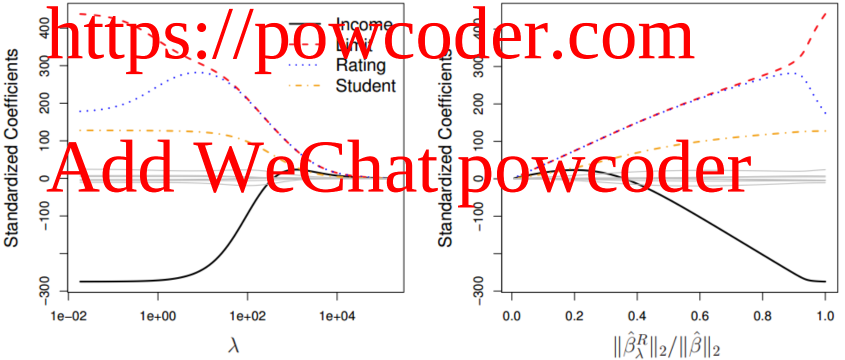
Note 17

Ridge regression - credit data example

Assignment Project Exam Help

balance ∼
age, *cards*, *education*, *income*, *limit*, *rating*, *gender*, *student*, *status*, *ethnicity*

https://powcoder.com



Add WeChat powcoder

Ridge regression vs. LS

Assignment Project Exam Help

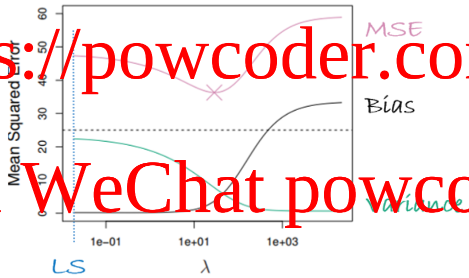
https://powcoder.com

Add WeChat powcoder

- Shrinkage methods can reduce the variance of the estimate $\hat{f}(x)$

  $\lambda \uparrow$ -> # of (influential) variables $\downarrow$

  -> Variance $\downarrow$ & Bias $\uparrow$



- Ridge regression (& Lasso) improves on LS!
- The MSE is reduced
- The variance is much smaller at the expense of a small increase in bias.

Lasso

Assignment Project Exam Help

- **Ridge regression** disadvantage: includes all p predictors (some of them with minor influence)

https://powcoder.com

- **Lasso**, in contrast, select subset.
- The lasso coefficients, $\hat{\beta}^L_\lambda$, minimize the quantity

Add WeChat powcoder

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

The Variable Selection Property of the Lasso

Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?

Assignment Project Exam Help

One can show that the lasso and ridge regression coefficient estimates solve the problems

https://powcoder.com

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

Add WeChat powcoder

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

respectively.

Some Remarks on Lasso

Assignment Project Exam Help

- Making s sufficiently small will cause some of the coefficients to be exactly zero $\to$ continuous subset selection
- If $s = \sum_{i=1}^{p} \|\hat{\beta}_j^{ls}\|$, then the lasso estimates are the $\hat{\beta}_j^{ls}$'s.
- s should be adaptively chosen to minimize an estimate of expected prediction error.

https://powcoder.com

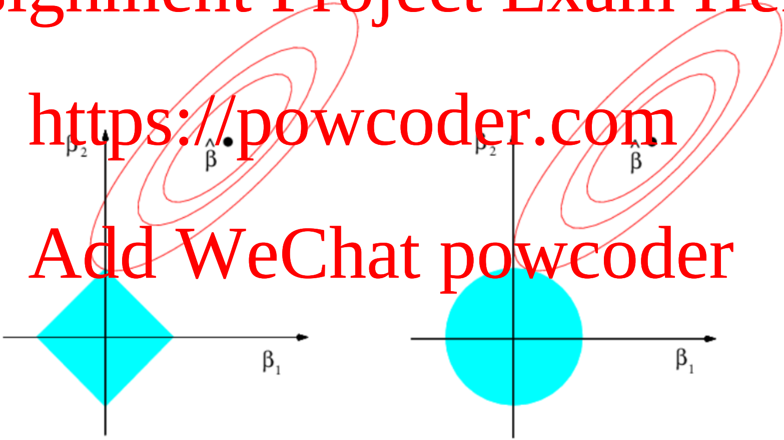
Add WeChat powcoder

The Variable Selection Property of the Lasso

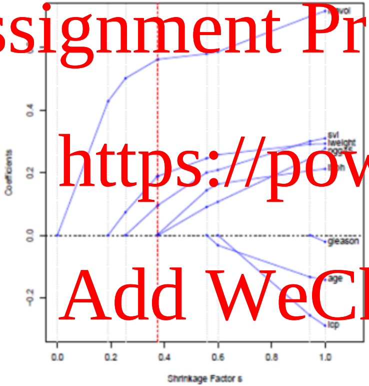
Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder



Profile of Lasso Coefficients



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- Profiles of lasso coefficients as the tuning parameter t is varied. The coefficients are plotted versus $t = s/\sum_{i=1}^{p} \|\hat{\beta}_j^{ls}\|$

Dimension Reduction Methods

Assignment Project Exam Help

- When there is a large number of predictors, often correlated, we can select what variables (dimensions) to use.
- But, why not transform the predictors (to a lower dimension) and then fit the least squares model using the transformed variables.

https://powcoder.com

- We will refer to these techniques as **dimension reduction** methods.

Add WeChat powcoder

- Use a small linear combinations $\mathbf{z}_m$, $m = 1, ..., M$ of $\mathbf{x}_j$
- The methods differ in how the linear combinations are obtained

Principal Components Regression

Assignment Project Exam Help

- The linear combinations $\mathbf{z}_m$ are the principal components
  $\mathbf{z}_m = X\mathbf{v}_m$

https://powcoder.com

$$\max_{\substack{\|\boldsymbol{\alpha}\|=1 \\ \mathbf{v}_\ell^\top \mathbf{S}\boldsymbol{\alpha}=0, \quad \ell=1,\dots,m-1}} \mathrm{Var}(X\boldsymbol{\alpha})$$

Add WeChat powcoder

- $\mathbf{v}_\ell^\top \mathbf{S}\alpha = 0$ ensure $\mathbf{z}_m = X\mathbf{v}_m$ is uncorrelated with all previous linear combinations $\mathbf{z}_\ell = X\mathbf{v}_\ell$, $\ell = 1, \dots, m-1$
- $\mathbf{y}$ is regressed on $\mathbf{z}_1, \dots, \mathbf{z}_M$ for $M \le p$

Principal Components Regression

Assignment Project Exam Help

- Since $\mathbf{z}_m$ are othrogonal, this regression is just a sum of univariate regressions

https://powcoder.com

$$\hat{\mathbf{y}}^{pcr} = \bar{\mathbf{y}} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

$$\hat{\theta}_m = < \mathbf{z}_m, \mathbf{y} > / < \mathbf{z}_m, \mathbf{z}_m >$$

Add WeChat powcoder

$$\hat{\beta}^{pcr} = \sum_{m=1}^{M} \hat{\theta}_m \mathbf{v}_m$$

- if $M = p$, $\hat{\mathbf{y}}^{pcr} = \hat{\mathbf{y}}^{LS}$ since the columns of $Z = UD$ span the column space of X
- PCR discards the $p - M$ smallest eigenvalue components.

Principal Components Regression

- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ *linear combinations* of our original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \qquad (1)$$

for some constants $\phi_{m1}, \ldots, \phi_{mp}$.

- We can then fit the linear regression model,

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n, \qquad (2)$$

using ordinary least squares.

- Note that in model (2), the regression coefficients are given by $\theta_0, \theta_1, \ldots, \theta_M$. If the constants $\phi_{m1}, \ldots, \phi_{mp}$ are chosen wisely, then such dimension reduction approaches can often outperform OLS regression.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Dimension Reduction Methods

Assignment Project Exam Help

- Notice that from definition (1):

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij},$$

https://powcoder.com

where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}. \tag{3}$$

Add WeChat powcoder

- Hence model (2) can be thought of as a special case of the original linear regression model.
- Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the form (3).
- Can win in the bias-variance tradeoff.

Principal Components

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

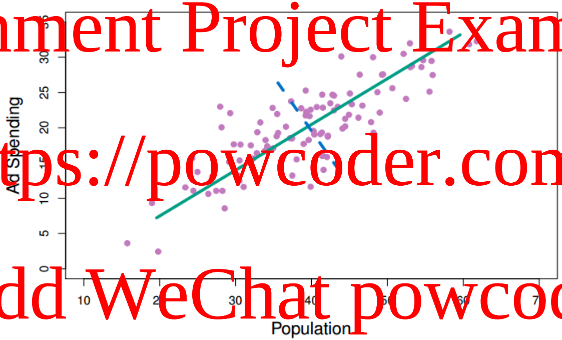
- Effectively, we change the orientation of the axes.
- The 1^{st} principal component is that (normalized) linear combination of the variables with the largest variance.
- The 2^{nd} principal component has the largest (remainder) variance, subject to being uncorrelated with the first.
- And so on...
- Many times we can explain most of the variation with only few principal components.
- More details in Chapter 10.2 of the book 'An introduction to statistical learning'

Principal Components

Assignment Project Exam Help

https://powcoder.com

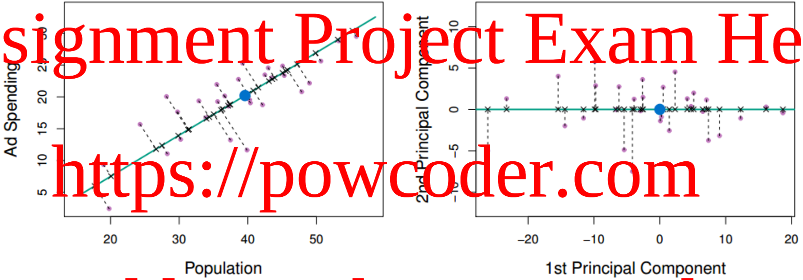
Add WeChat powcoder



The population size (`pop`*) and ad spending (*`ad`*) for* 100 *different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Principal Components

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder



A subset of the advertising data. Left: *The first principal component, chosen to minimize the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments.* Right: *The left-hand panel has been rotated so that the first principal component lies on the x-axis.*

Principal Components Regression

Assignment Project Exam Help

- Now we can run a regression analysis on only several principal components.
- We call it **Principal Components Regression** (**PCR**)

https://powcoder.com

- Note, these directions are identified in an *unsupervised* way, since the response Y is not used to help determine the principal component directions.

Add WeChat powcoder

- Consequently, PCR suffers from a potentially serious drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Partial Least Squares (PLS)

Assignment Project Exam Help

- It is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots, Z_M$ that are linear combinations of the original features.

https://powcoder.com

- Then fits a linear model via OLS using these M new features.
- Up to this point very much as PCR.
- PLS identifies these new features using the response Y (*supervised way*).

Add WeChat powcoder

- PLS approach attempts to find directions that help explain both the response and the predictors.

Partial Least Squares (PLS)

Assignment Project Exam Help

- Uses $\mathbf{y}$ to construct linear combinations of the inputs.
- The inputs are weighted by the strength of their univariate effect on $\mathbf{y}$

https://powcoder.com

- Regress $\mathbf{y}$ on $\mathbf{z}_m \to \hat{\theta}_m$ and orthogonalize with respect to $\mathbf{z}_m$
- Continue the process until $M < p$ directions are obtained
- PLS seeks directions that have high variance and have high correlation with the response

Add WeChat powcoder

$$\max_{\substack{\|\boldsymbol{\alpha}\|=1 \\ \mathbf{v}_\ell^\top \mathbf{S}\boldsymbol{\alpha}=0, \quad \ell=1,\dots,m-1}} \mathrm{Corr}^2\left(\mathbf{y}, X\boldsymbol{\alpha}\right) \mathrm{Var}\left(X\boldsymbol{\alpha}\right)$$

Partial Least Squares (PLS)

Assignment Project Exam Help

- The first component say $\mathbf{t} = X\boldsymbol{\alpha}_1$ maximizes

$$\boldsymbol{\alpha}_1 = \arg \max_{\|\boldsymbol{\alpha}\|=1} \operatorname{Cov}^2(\mathbf{y}, X\boldsymbol{\alpha})$$

https://powcoder.com

- Subsequent components $\mathbf{t}_2, \mathbf{t}_3, \ldots$ are chosen such that they maximizes the squared covariance to $\mathbf{y}$ and all components are mutually orthogonal
- Orthogonality is enforced by deflating the original variables X

Add WeChat powcoder

$$X_i = X - \mathcal{P}_{\mathbf{t}_1,\ldots,\mathbf{t}_{i-1}} X$$

- $\mathcal{P}_{\mathbf{t}_1,\ldots,\mathbf{t}_{i-1}}$ denotes the orthogonal projection onto the space spanned by $\mathbf{t}_1, \ldots, \mathbf{t}_{i-1}$
- $\hat{\mathbf{y}} = \mathcal{P}_{\mathbf{t}_1,\ldots,\mathbf{t}_m}\mathbf{y}$ instead of $\hat{\mathbf{y}} = X\hat{\beta} = X\left(X^\top X\right)^{-1} X^\top \mathbf{y}$

Illustrating the connection

The connection between these methods can be seen through the optimisation criterion they use to define projection directions

Assignment Project Exam Help

- PCR extracts components that explain the variance of the predictor space

https://powcoder.com

$$\max_{\substack{\|\boldsymbol{\alpha}\|=1 \\ \mathbf{v}_\ell^\top \mathbf{S}\boldsymbol{\alpha}=0, \quad \ell=1,\ldots,m-1}} \mathrm{Var}\left(X\boldsymbol{\alpha}\right)$$

- PLS extracts components that have a high covariance with

Add WeChat powcoder

$$\max_{\substack{\|\boldsymbol{\alpha}\|=1 \\ \mathbf{v}_\ell^\top \mathbf{S}\boldsymbol{\alpha}=0, \quad \ell=1,\ldots,m-1}} \mathrm{Corr}^2\left(\mathbf{y}, X\boldsymbol{\alpha}\right)\mathrm{Var}\left(X\boldsymbol{\alpha}\right)$$

- Both method are similar in there aim to extract m components from the predictor space X

Illustrating the connection

Assignment Project Exam Help

Both methods aims

- at expressing the solution in lower dimensional subspace $\beta = V\mathbf{z}$ where V is an $p \times m$ matrix of orthonormal columns

https://powcoder.com

- Using this basis for the subspace, an alternative approximate minimization problem is considered

Add WeChat powcoder

$$\min_{\beta} \|\mathbf{y} - X\beta\|_2 \approx \min_{\mathbf{z}} \|\mathbf{y} - XV\mathbf{z}\|$$

- In PCR V is directly obtained from X
- in PLS V depends on $\mathbf{y}$ in a complicated nonlinear way

Illustrating the connection

Assignment Project Exam Help

Considering

- The singular value decomposition $X = UDV^\top$ where U is $N \times p$, $D = \mathrm{diag}(d_1, ..., d_p)$ is $p \times p$ and V is $p \times p$
- The columns of U and V are orthogonal such that $U^\top U = I_p$ and $V^\top V = I_p$

https://powcoder.com

- The least squares solution takes the form

$$\hat{\beta} = \sum_{i=1}^{p} \frac{\mathbf{u}_i^\top \mathbf{y}}{d_i} \mathbf{v}_i = \sum_{i}^{p} \beta_i$$

Add WeChat powcoder

- The other estimator are shrinkage estimators and can be expressed as

$$\hat{\beta} = \sum_{i}^{p} f(d_i)\beta_i$$

Multiple Outcome Shrinkage

Assignment Project Exam Help

- When the output are not correlated $\rightarrow$ apply a univariate technique individually to each outcome or work with each column output individually
- Other approaches exploit correlations in the different responses $\rightarrow$ canonical correlation analysis

https://powcoder.com

- CCA find a sequence of linear combinations $\mathbf{X}v_m$ and $\mathbf{Y}u_m$ such that the correlations are maximized

$$\mathrm{Corr}^2\left(\mathbf{Y}u_m, \mathbf{X}v_m\right)$$

Add WeChat powcoder

- Reduced rank regression

$$\hat{\mathbf{B}}^{rr}(m) = \arg\min_{rank(\mathbf{B})=m} \sum_{i=1}^{N} \left(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i\right)^\top \Sigma^{-1} \left(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i\right)$$

Note 19

For more readings

Assignment Project Exam Help

- Summaries on LMS.
- Chapter 3, 5 & 14.5 from 'The elements of statistical learning' book.
- Chapters 3, 6 & 10.2 from 'An introduction to statistical learning' book.

https://powcoder.com

Add WeChat powcoder