Theory of Regression Analysis with Applications

T Padma Ragaleena

National Institute of Science Education and Research
Bhubaneswar

20 November 2019

Multiple-linear regression model

Regression model

Response	Regressor 1	Regressor 2	$\cdots$	Regressor k
y	x_1	x_2	$\cdots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
.	.	.		.
.	.	.		.
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

- $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$
- We also assume : $cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$
- Y is a random vector ; all x_i's are not random and they are known with negligible error
- We assume the existence of at least an approximate linear relationship between response variables and other regressors.

Why are we assuming a distribution for ϵ?

- to get the p-values and confidence intervals for quantities of interest (hypothesis testing)

Why did we choose normal distribution?

- It describes random errors in real world processes reasonably well
- There is well developed mathematical theory behind normal distribution

Are non-normal distributions useful?

- In financial models, errors are assumed to come from a heavy tailed distribution , normal distribution is not suitable here.

Least Square Estimates

How do we estimate β

- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ measures the amount of deviation of the predicted value from the true value.
- One way to get a "good estimate" for β is to minimize the SSE.
- So we minimize $S(\beta) = (y - X\beta)'(y - X\beta)$ with respect to β and call the minimizing vector as the Least Square Estimate(LSE) for the model. It is denoted by $\hat{\beta}$.
- In order to find the β which minimizes $S(\beta)$, we use the following property of Hilbert spaces:

Closest point theorem

Let M be a closed convex subset of a Hilbert space H, $x \notin M$ then $\exists ! y_0 \in M$ such that $||x - y_0|| \leq ||x - m||$ for all $m \in M$. Also, $y_0 - x \in M^{\perp}$

- Using this theorem, we get :

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y = \text{Least Square Estimate}}$$

Least square estimates

- In Hilbert spaces, y_0 is called the projection of x on to the subspace M. Similarly, $H = X(X'X)^{-1}X'$ is called projection matrix because $\hat{y} = Hy$
- For Hilbert spaces, we know that the projection map defined as $P(x) = y_0$ is idempotent. Here also, H is idempotent i.e. $H^2 = H$

Properties of least square estimates(LSE)

- LSE is an unbiased estimate for β
- $\hat{\beta}$ is a maximum likelihood estimator for β.
- Least square estimators are Best Linear Unbiased Estimators - BLUE (Gauss-Markov theorem)

Gauss-Markov theorem

Let $Y = X\beta + \epsilon$ be a regression model such that each ϵ_i follows a distribution with mean 0 , variance σ^2 and $cov(\epsilon_i, \epsilon_j) = 0$. Then the LSE are Best Linear Unbiased Estimators.

- Observe that no normality is assumed for the errors
- $\hat{\beta}$ is best $\implies$ $Var(a'\hat{\beta}) \leq Var(a'\tilde{\beta})$ for all $a \in \mathbb{R}^p$; $\tilde{\beta} =$ any other linear unbiased estimate

Coefficient of determination

- $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ i.e. SST = SSRes + SSR
- SST measures the total variation of y_i's around $\bar{y}$
- SSRes measures the variation that could not be explained by the model.
- SSR is the variation that can be explained by the model.
- Then $R^2 = 1 - \frac{SSRes}{SST} \in [0, 1]$ gives a proportion of variation in y_i that could be explained by the model.

Coefficient of determination

Consider the data containing temperature (x-variable) and the log of the light intensity(y-variable) of 47 stars in the star cluster CYG OB1

```
data("CYGOB1")
model1 <- lm(CYGOB1$logli   CYGOB1$logst , data = CYGOB1)
summary(model1)$r.squared
0.04427374
```

Regerssion captures only 4.4% variation . This is not a good model.

Tests of significance

- $H_0 : \beta_j = 0$ for all j against H_1 : at least one $\beta_j \neq 0$ tests if there exists any linear relationship between response and predictors.

Test Statistic: Under the null hypothesis

$$\frac{\frac{SSR}{k}}{\frac{SSRes}{\sigma^2}} \sim F_{k,n-p}$$

Under a level of significance α, we have enough evidence to reject H_0 in favour of H_1 if

$$|F^*| \geq F_{\frac{\alpha}{2};k,n-p}$$

or reject the null hypothesis in favour of H_1 if

$$\text{p-value} \leq \alpha$$

Tests of significance

- Once we know that previous null hypothesis is rejected, then our next aim would be to know which coefficients β_j are non-zero.
- $H_0 : \beta_j = 0$ against $\beta_j \neq 0$

Test Statistic: Under the null hypothesis:

$$\frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}} \sim t_{n-k-1}$$

Under a level of significance α, we have enough evidence to reject H_0 in favour of H_1 if

$$|t^*| \geq t_{\frac{\alpha}{2};n-k-1}$$

or reject the null hypothesis in favour of H_1 if

$$\text{p-value} \leq \alpha$$

Tests of significance

- A more general hypothesis would be to test the r linearly independent hypothesis i.e. $H_0 : \hat{\beta}_0 a_{i0} + \hat{\beta}_1 a_{i1} + ... + \hat{\beta}_k a_{ik} = b_i$ for all $i = 1, 2, ..., r$.
- In other words, the hypothesis we want to test is $H_0 : A\hat{\beta} = \tilde{b}$ where T is a known linear transformation.

Test statistic: Under the null hypothesis:

$$\boxed{\frac{(A\hat{\beta}-\tilde{b})'(A(X'X)^{-1}A')^{-1}(A\hat{\beta}-\tilde{b})}{r\hat{\sigma}^2} \sim F_{r,n-p}}$$

Under a level of significance α, we have enough evidence to reject H_0 in favour of H_1 if

$$\boxed{|F^*| \geq F_{\frac{\alpha}{2};r,n-p}}$$

or reject the null hypothesis in favour of H_1 if

$$\boxed{\text{p-value} \leq \alpha}$$

Regression Diagnostics

- Our aim is to check if our model follows the regression assumptions. A few remedies are suggested if the assumptions are not being followed.
- The validity of these assumption is needed for the results to be meaningful. If these assumptions are violated, the results can be incorrect or misleading.
- So such underlying assumptions have to be verified before attempting to do regression modeling.

Residuals

- Residuals $e_i = y_i - \hat{y}_i$ can be thought of as a realization of the error terms. Thus any departure from assumptions on errors, should show up in the residuals.
- We can show that $e = (I - H)\epsilon$. Hence $Var(e) = \sigma^2(I - H)$.
- Even though errors ϵ_i are assumed to be uncorrelated and independent, the residuals e_i's are correlated and hence dependent.

Normality assumption

- **Q-Q plot** is a graphical tool that is used to assess normality.
- It plots the theoretical quantiles (horizontal axis) against the sample quantiles (vertical axis))
- Using the residual values (e_i), an empirical distribution is constructed using which we get sample quantiles.
- If X is a discrete random variable, then ξ_p is called the p^{th} quantile of a random variable X if
  $P(X \leq \xi_p) \geq p$ and $P(X \geq \xi_p) \geq 1 - p$
- If X is a continuous random variable, then p^{th} quatile is the unique ξ_p such that $P(X \leq \xi_p) = p$

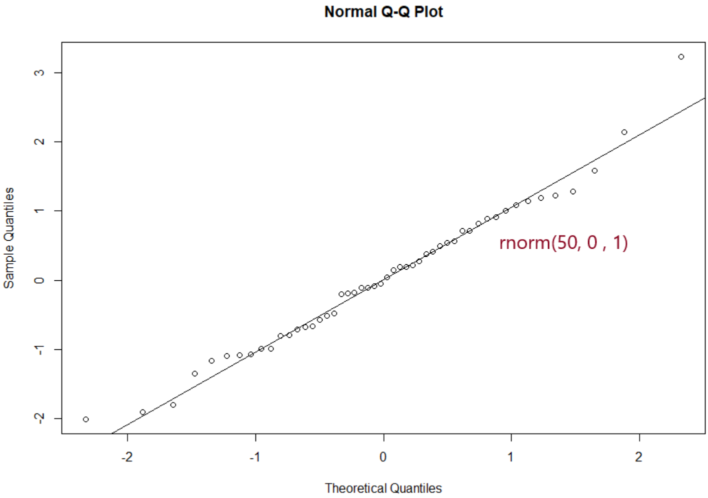
Q-Q plot

- Here we want to check if the residuals e_i are coming from a normal distribution.
- Considering the residual values we have, we can estimate the cdf from which these points have come from as:

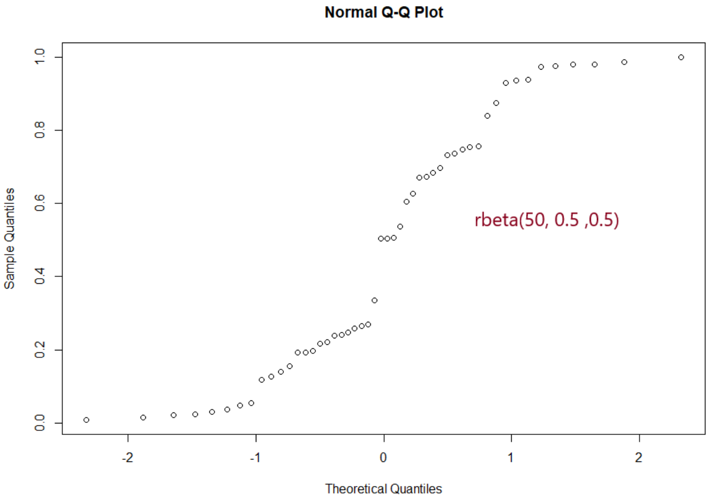
$$\boxed{\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(e_i \leq x)}$$

- If $e_{(1)} \leq e_{(2)} \leq ... \leq e_{(n)}$, then $e(i)$ will be the $\frac{i}{n}^{th}$ quantile.
- Plot $\hat{F}^{-1}(\frac{i}{n}) = \xi_{\frac{1}{n}}$ against $\Phi^{-1}(\frac{i}{n})$
- If the normality assumption is followed then the plot has to be an approximate $y = x$ line.

Normal Q-Q plot



Non-normal Q-Q plot



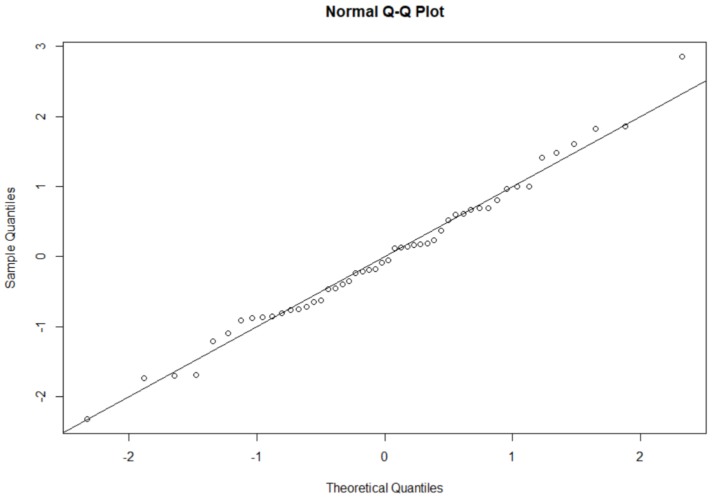
Data Example

Consider the "LifeCycleSavings" data set in R. This is a model proposed by Franco Modigliani to estimate savings ratio of a country.

```
g <- lm(sr   pop15 + pop75 + dpi + ddpi , data = LifeCycleSavings)
```

- “sr” is the savings ratio
- “pop15” is percentage of people under 15
- “pop75” is percentage of people over 75
- “dpi” is per capita disposable income
- “ddpi” is the percentage growth of dpi

Q-Q plot


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
3

Kolmogorov-Smirnov Test

- We should not rely on graphical tools to draw conclusions. A formal test to check for normality assumption is Kolomogorov-Smirnov test.
- If $X_1, X_2...X_n$ are assumed to come from a known continuous distribution P. Then we want to test the null hypothesis $H0$: The samples come from P against $H1$: they do not come from P.
- Let F_{exp} be the cdf associated with the null hypothesis and the empirical distribution function F_{obs} is given by : $F_{obs}(x) = \frac{1}{\text{total no. of obs}} \sum_{i=1}^{n} I(X_i \leq x)$.
- The test statistic is:

$$\boxed{D = sup|F_{exp}(x) - F_{obs(x)}|, \text{ sup over all } x}$$

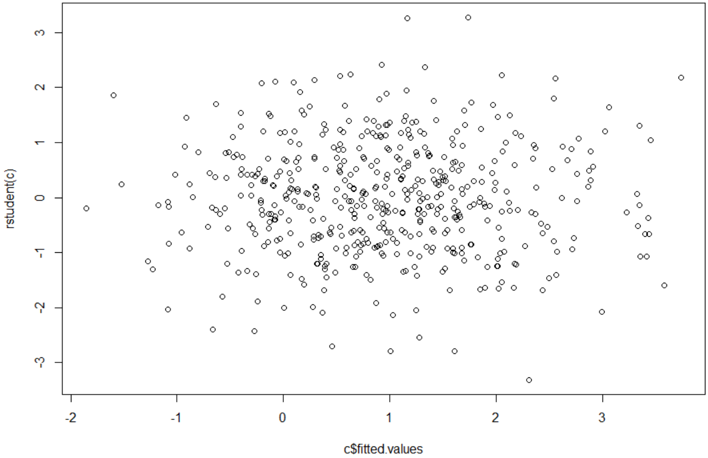
Kolomogorov-Smirnov test

```
One-sample Kolmogorov-Smirnov test
data: as.numeric(rstudent(g))
D = 0.067991, p-value = 0.9628
alternative hypothesis: two-sided
```

A very high p-value indicates that it is very likely that the normality assumption is being followed.
Other tests like Anderson Darling test, Shiapiro-Wilk test also exist to check the normality assumption.

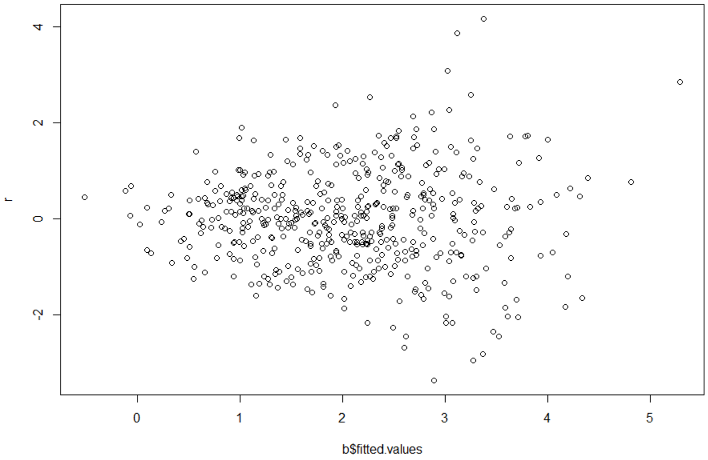
Constant Variance assumption

Fitted values vs Residuals for data from standard normal distribution



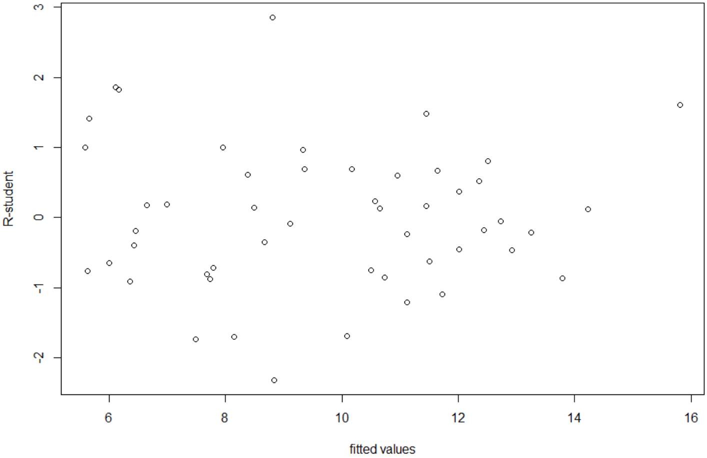
Constant Variance assumption

Fitted values vs Residuals for data from normal distribution with non-constant variance



Constant Variance assumption

Fitted values vs Residuals for LifeCycleSavings dataset

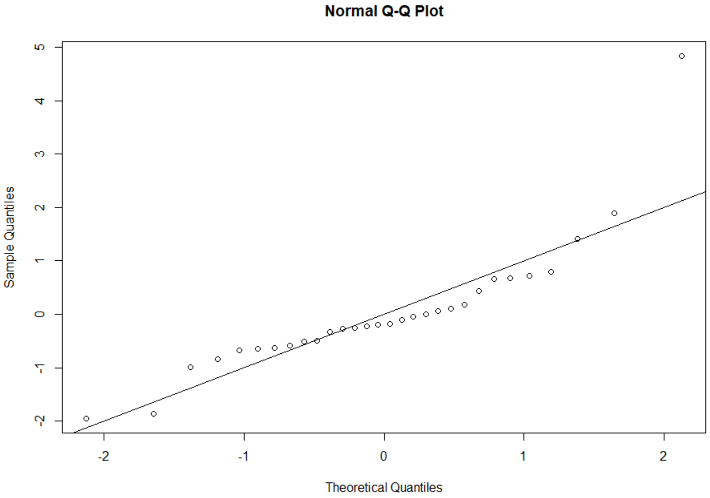


Linearity Assumption

- We can check the linearity assumption using the lack of fit test. In order to apply this test we should make sure that all the other assumptions are followed and only linearity is being questioned.
- Requirement : Take more than one observation for response given response x. $x_i \implies y_{i1}, ..., y_{i,n_i}$
- $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\hat{y}_i-\bar{y}_i)^2$
- SSRes = SSPE + SSLOF
- If the true regression function is linear: $\frac{SSLOF(n-m)}{SSPE(m-2)} \sim F_{m-2,n-m}$

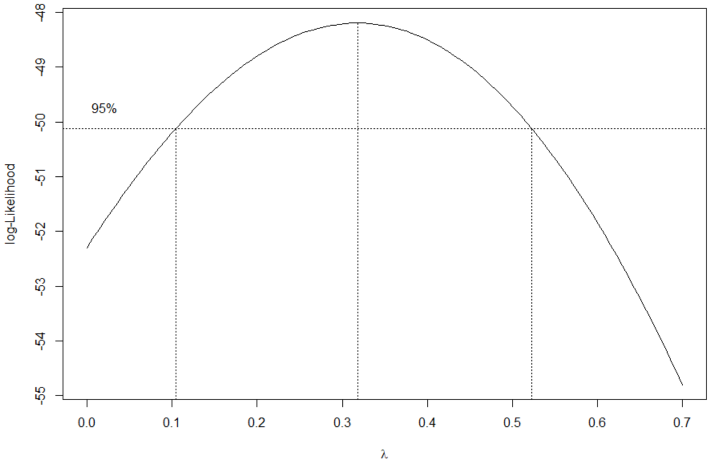
Box-Cox transformation

To correct the normality assumption if it isn't being followed. Consider the "gala" data from R



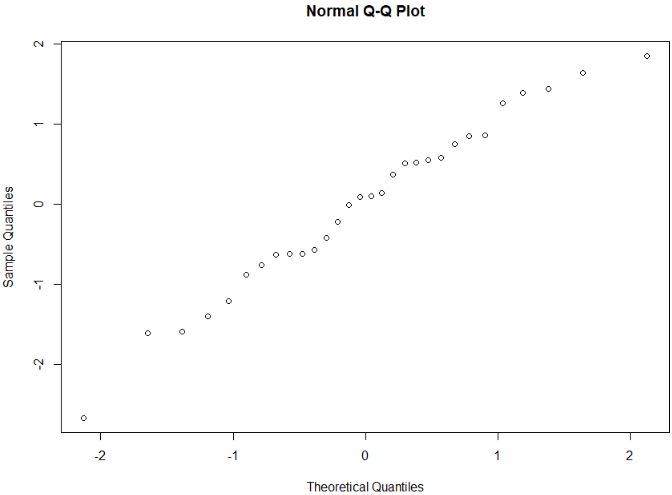
Box-Cox transformation

Find λ that maximizes likelihood



Box-Cox transformation

After applying a cube root transformation



Box-Cox method

```
One-sample Kolmogorov-Smirnov test
data: as.numeric(rstudent(gfit3))
D = 0.093249, p-value = 0.935
alternative hypothesis: two-sided
```

Hence our transformation on response was very useful. High p-value indicates strong evidence for normality.

Variance stabilizing transformations

One of the common reasons for the violation of constant variance is for the response variable to follow a distribution in which variance is a function of mean i.e. when $\sigma^2 = \omega(\mu)$.

AIM: We wish to find a function f such that $Var(f(Y))$ is roughly constant i.e. we "transform" the response variable.

$$f(Y) \approx f(\mu) + (Y - \mu)f'(\mu) \Rightarrow [f(Y) - f(\mu)]^2 \approx (Y - \mu)^2[f'(\mu)]^2$$

$$\text{hence, } V(f(Y)) \approx V(Y) \times [f'(\mu)]^2$$

Multicollinearity

The problem of multi-collinearity is said to exist when two or more regressor variables are strongly correlated. Or in other words, the columns of X exhibit near linear dependencies, then the problem of **multicollinearity** is said to exist. In case of perfect multicollinearity X will not be invertible.
We cannot find the least square estimates when multicollinearity exists

Least Square Estimates

How do we estimate β

- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ measures the amount of deviation of the predicted value from the true value.
- One way to get a "good estimate" for β is to minimize the SSE.
- So we minimize $S(\beta) = (y - X\beta)'(y - X\beta)$ with respect to β and call the minimizing vector as the Least Square Estimate(LSE) for the model. It is denoted by $\hat{\beta}$.
- In order to find the β which minimizes $S(\beta)$, we use the following property of Hilbert spaces:

Closest point theorem

Let M be a closed convex subset of a Hilbert space H, $x \notin M$ then $\exists! y_0 \in M$ such that $||x - y_0|| \leq ||x - m||$ for all $m \in M$. Also, $y_0 - x \in M^{\perp}$

- Using this theorem, we get :

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y = \text{Least Square Estimate}}$$

Problem with multi-collinearity

We can show that $Var(\hat{\beta}_j) = c_{jj}\sigma^2$ where c_{jj} is the j^{th} diagonal element of $(X'X)^{-1}$
In this case, $C_{jj} = \frac{1}{1-R_j^2}$ where R_j is the coefficient of determination when we regress x_j on the remaining p-variables.
If multicollinearity exists, $Var(\hat{\beta}_j) \longrightarrow \infty$ as $R_j^2 \longrightarrow 1$
This would mean that our estimates would be unreliable.

VariationInflation Factors(VIF)

VIF exists for each of the predictors in a multiple regression model. VIF for the j^{th} predictor is given by $VIF_j = \frac{1}{1-R_j^2}$.

Rule of thumb: If $VIF > 4$, it warrents further investigation. $VIF > 10$ indicates serious multicollinearity.

The followinf are the VIF values when BP is regressoed with respect to BSA and weight.

```
data$Weight data$BSA
4.276401 4.276401
```

We see some evidence of multicollinearity but we need to we need more evidence to confirm.

Ridge Regression

The ridge coefficients minimize a penalized residual sum of squares,
$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{k} \beta_k^2)^2$ is minimized w.r.t. β
This is equivalent to minimizing $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2$ given that $\sum_{i=1}^{k} \beta_k^2 < s$

$$\boxed{\beta_{ridge} = (X'X + \lambda I)^{-1} X'Y}$$

Data example

Consider the “meatspec” data in R from faraway package.

```
modified HKB estimator is 2.363535e-08
modified L-W estimator is 0.907997
smallest value of GCV at 3.25e-08
```

So the value of λ obtained is $3.25e-08$

Ridge trace