Linear Regression-3

Prof. Asim Tewari
IIT Bombay

Multiple Linear Regression

Multiple Linear Regression assumes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The model can be expressed as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

with its coefficients being derived by minimizing RSS

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

$$\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Multiple Linear Regression

If X is a vector $\begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$

Data : n-data points

$(x_1^T, y_1), (x_2^T, y_2) \; \cdots \cdots \; (x_n^T, y_n)$

$x_1^T \uparrow \begin{bmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \\ \vdots \\ x_1^p \end{bmatrix} \qquad x_n^T \uparrow \begin{bmatrix} x_n^1 \\ x_n^2 \\ \vdots \\ x_n^p \end{bmatrix}$

Multiple Linear Regression

Data: $(x_1, y_1), \ldots (x_i, y_i) \ldots (x_n, y_n)$

$$\underset{p\times 1}{x_i} = (x_i^1, x_i^2 \ldots x_i^p)^T$$

$$\underset{(p+1)\times 1}{\beta} = (\beta_0, \beta_1, \beta_2 \ldots \beta_p)^T$$

$$\underset{(p+1)\times 1}{x_i} (1, x_i^1, x_i^2, x_i^3 \ldots x_i^p)$$

Multiple Linear Regression

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1} = \underbrace{\underbrace{\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}}_{n \times (p+1)} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{(p+1) \times 1}}_{n \times 1}$$

$$\underset{n \times 1}{Y} = \underset{n \times (p+1)}{X} \; \underset{(p+1) \times 1}{\beta}$$

$$x_1 = \begin{bmatrix} 1 \\ x_1^1 \\ x_1^2 \\ \vdots \\ x_1^p \end{bmatrix}$$

$p \rightarrow$ # of features

$n =$ # of samples

Multiple Linear Regression

$RSS = \|\underbrace{X\beta}_{\hat{Y}} - Y\|^2$, β^* is the value of β which min RSS

$\because (n > p+1)$

$RSS(\beta)$

Local min β^* would be sol of

$\nabla RSS(\beta^*) = 0$

Multiple Linear Regression

$$RSS(\beta) = \| X\beta - y \|^2$$

$$= \underbrace{\underbrace{(\underbrace{X\beta}_{n\times 1} - \underbrace{y}_{n\times 1})^T}_{1\times n} (\underbrace{X\beta}_{n\times 1} - \underbrace{y}_{n\times 1})}_{1\times 1}$$

$$= (X\beta)^T X\beta - (X\beta)^T y - y^T X\beta + y^T y$$

$$= \beta^T X^T X \beta - 2\,\beta^T X^T y + y^T y$$

$$\boxed{\nabla_x (a^T x) = a \quad \text{and} \quad \nabla_x (x^T A x) = (A + A^T) x}$$

Multiple Linear Regression

$$\boxed{\begin{array}{l} \nabla_x (a^T x) = a \\ \nabla_x (x^T A x) = (A + A^T) x \end{array}}$$

$$RSS = \beta^T \underbrace{X^T X} \beta - 2 \beta^T \underbrace{X^T y}_{a} + y^T y$$

$$\nabla_\beta RSS(\underset{(p+1)\times 1}{\underset{\uparrow}{\beta}}) = (X^T X + X^T X)\beta - 2 X^T y$$

$$\nabla_\beta RSS(\beta) = 2 X^T X \beta - 2 X^T y$$

Find $\nabla_\beta RSS(\beta^*) = 0$ }

Multiple Linear Regression

$$\nabla_\beta RSS(\beta) = 2X^TX\beta - 2X^Ty$$

$$\left.\text{Find } \nabla_\beta RSS(\beta^*) = 0 \right\}$$

$$2X^TX\beta^* - 2X^Ty = 0$$

$$X^TX\beta^* = X^Ty$$

$$\underbrace{(X^TX)^{-1}(X^TX)}_{I}\beta^* = (X^TX)^{-1}X^Ty$$

$$\boxed{\beta^* = (X^TX)^{-1}X^Ty}$$

Multiple Linear Regression

$\min \; RSS(\beta)$

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

$\hookrightarrow \min \quad (\text{global min?})$

$Y = X\beta$

$\nabla_\beta RSS(\beta) = 2X^T X\beta - 2X^T y$

$\nabla^2 RSS(\beta) = 2 X^T X$

$n \times (P+1)$

$\underbrace{(P+1) \times n}_{(P+1)\,(P+1)}$

Multiple Linear Regression

$$\underbrace{\nabla^2 RSS(\beta)}_{\text{Hessian of RSS}} = 2 \, \underbrace{\underset{(p+1)\times n}{X^T} \; \underset{n\times(p+1)}{X}}_{(p+1)\,(p+1)}$$

$$\forall \beta \qquad \beta^T (2X^TX)\beta = 2(X\beta)^T X\beta$$

$$= 2\,\|X\beta\|^2 \geq 0$$

β^* is now global minima.

Multiple Linear Regression

$$\boxed{\beta^* = (X^T X)^{-1} X^T y}$$

$$X = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2^1 & x_2^2 & \cdots & x_2^p \\ 1 & & & & \\ 1 & & & & \\ 1 & x_n^1 & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

$n \times (p+1)$

$$\overbrace{(X^T X)}^{(p+1)\times(p+1)}$$

$(p+1) \times n$ $n \times (p+1)$

Gauss-Markov Theorem

- The Gauss–Markov theorem states that if we have any other linear estimator $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$ that is unbiased for $\boldsymbol{\alpha}^T\beta$, that is, $E(c^Ty) = \boldsymbol{\alpha}^T\beta$, then

$$\mathrm{Var}(\alpha^T\hat{\beta}) \leq \mathrm{Var}(\mathbf{c}^T\mathbf{y}).$$

Gauss-Markov Theorem

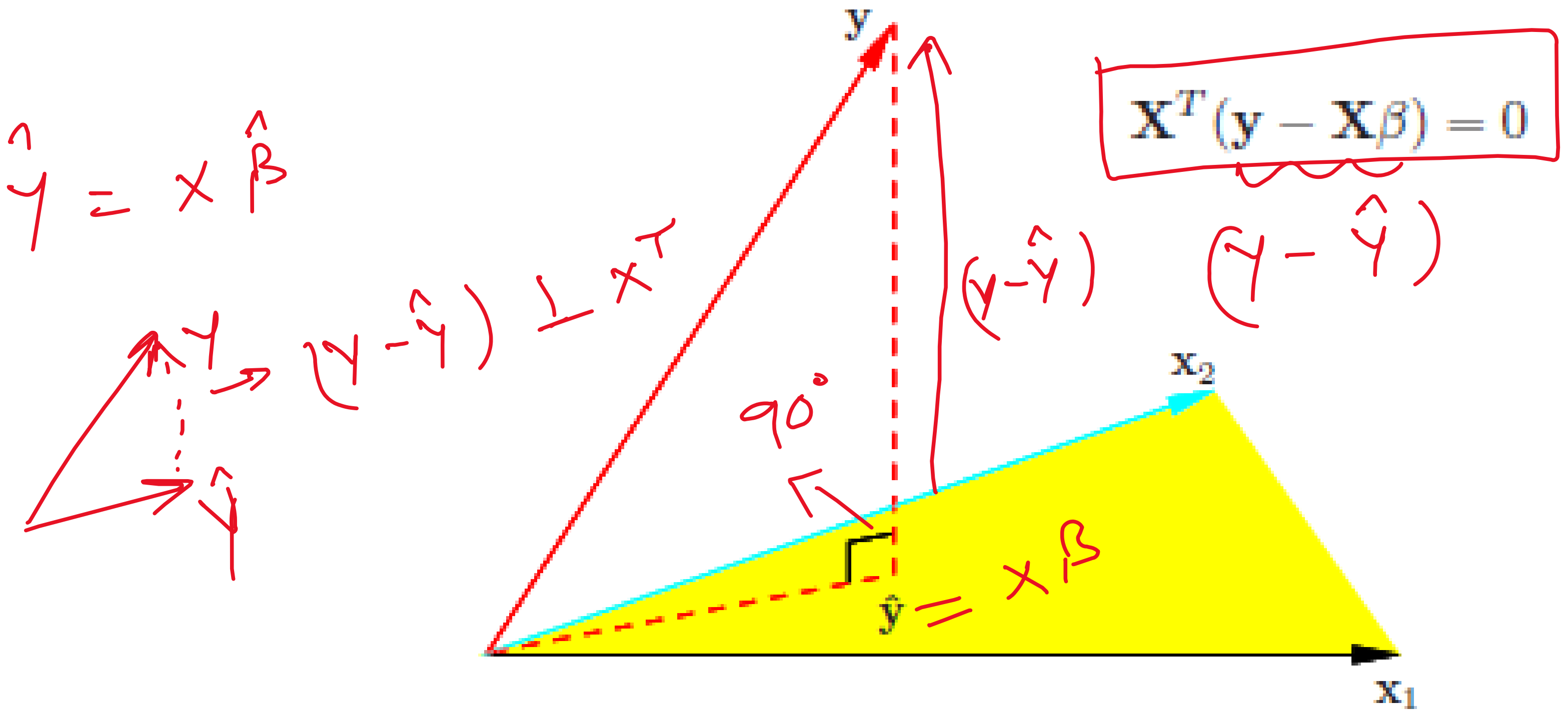
- The least squares estimate of $\boldsymbol{\alpha}^T\beta$ is

$$\hat{\theta} = a^T\hat{\beta} = a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

- Considering X to be fixed, this is a linear function $c_0^T y$ of the response vector y.
- If we assume that the linear model is correct, $\boldsymbol{\alpha}^T\beta$ is unbiased since

$$\begin{aligned} \mathrm{E}(a^T\hat{\beta}) &= \mathrm{E}(a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\ &= a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= a^T\beta. \end{aligned}$$

Linear Regression



The N-dimensional geometry of least squares regression with two predictors. The outcome vector y is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions

Multiple Linear Regression

In multiple linear regression, we usually are interested in answering a few important questions.

1. *Is at least one of the predictors $X1, X2, \ldots, Xp$ useful in predicting the response?*

2. *Do all the predictors help to explain Y , or is only a subset of the predictors useful?*

3. *How well does the model fit the data?*

4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

Extensions of Linear Models

- Removing the Additive Assumption

Introduce the interactive term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

- Where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$.
- Non-linear Relationships

$$\mathtt{mpg} = \beta_0 + \beta_1 \times \mathtt{horsepower} + \beta_2 \times \mathtt{horsepower}^2 + \epsilon$$

Nonlinear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots$$

$$\boxed{y = \beta_0 + \beta_1 x + \beta_2 x^2}$$

$$\equiv \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$x \qquad x^2$$

$$V \propto S$$

Basis function regression

$$y = \underline{\beta_0} + \underline{\beta_1} f_1(x) + \underline{\beta_2} f_2(x) + \cdots + \epsilon$$

e.g. 1 $\quad x \quad x^2$

2. $\quad f_1(x) = x^{1.5}, \quad f_2(x) = \left(1 + \frac{x}{3 + x^2}\right)$

Hypothesis testing in multi linear regression

$$y = f(x) + \epsilon \rightarrow f(x) \; ?$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

t statistic $\Rightarrow$ p-value

		p-value
Intercept	β_0	
x_1	β_1	
x_2	β_2	
	$\vdots$	

F-Statistic

$\checkmark \times \quad \checkmark \times \quad \checkmark \times$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{100} X_{100}$$

for 95% C $\Rightarrow$ 5 X_i $\checkmark$

t-statistic $\rightarrow$ p value for each variable X_i

$$\text{F-statistic} = \left\{ \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \right\}$$

$$\left\{ \begin{array}{lll} \text{If } F \approx 1 & H_0 & \checkmark \\ F \gg 1 & H_a & \checkmark \rightarrow n, p \end{array} \right.$$

F-Statistic

$$\overset{\checkmark \times}{} \quad \overset{\checkmark \times}{} \quad \overset{\checkmark \times}{}$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{100} X_{100} + \epsilon$$

for 95% C $\Rightarrow$ 5 X_i ✓

If H_0 is true and ϵ is $N(0, \sigma)$

then F-statistic follows a F-dist.

$$\left\{ \begin{array}{ll} \text{If } F \approx 1 & H_0 \checkmark \\ \quad F \gg 1 & H_a \checkmark \rightarrow n, p \end{array} \right.$$

Is There a Relationship Between the Response and Predictors?

We test the null hypothesis,

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

If this is true $(TSS - RSS)/p \rightarrow \sigma^2$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic,

If H_0 is true $F \approx 1$

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)},$$

Linear model is correct $\rightarrow \sigma^2$

If H_0 is not true $F >> 1$

Value of F-statistic close to 1 when null hypothesis true

Value of F-statistic greater than 1 when alternative hypothesis true

Hypothesis testing in multi linear regression

- F is very close to **one** we cannot reject the null hypothesis (thus, in a sense we accepted the **null hypothesis**)

- If F is **very large** we reject the null hypothesis (thus, in a sense we accepted the **alternate hypothesis**)

How large is large enough?

- This depends upon the values of n and p.
- If n is very large a small value above 1 is also a compelling evidence against the null hypothesis; however if n is a small then F has to be very large for us to reject the null hypothesis.

- When the null hypothesis is true and the error follows a Gaussian distribution, then it can be shown that F-statistic follows F-distribution

Hypothesis testing in multi linear regression

Why do we need F-statistic when t-statistic already exists?

(Given these individual p-values for each variable, why do we need to look at the overall F-statistic? After all, it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response.)

- However, the above logic is flawed, especially when the number of predictors p is large.

- For instance, consider an example in which $p = 100$ and $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ is true, so no variable is truly associated with the response. In this situation, about 5% of the p-values associated with each variable will be below 0.05 by chance. In other words, we expect to see approximately five small p-values even in the absence of any true association between the predictors and the response.

Multiple Linear Regression

- Predictors with Only Two Levels
- Define new variables

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

The model then takes the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Multiple Linear Regression

- Predictors with more than Two Levels
- Define new variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

The model then takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

Potential problems of Linear Regression

1. *Non-linearity of the response-predictor relationships.*
2. *Correlation of error terms.*
3. *Non-constant variance of error terms.*
4. *Outliers.*
5. *High-leverage points.*
6. *Collinearity.*

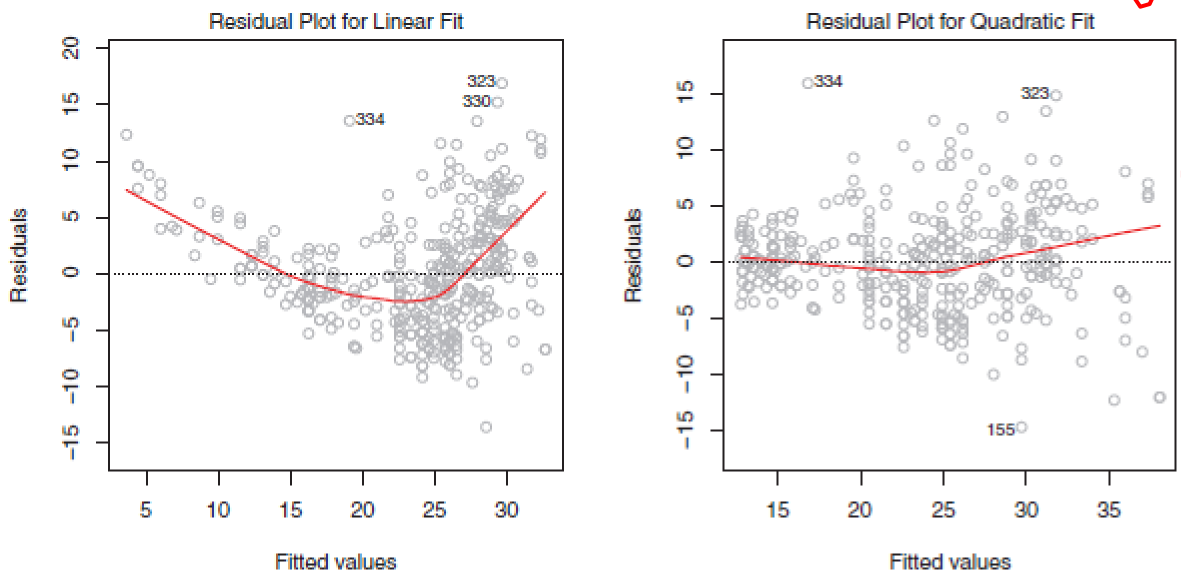
Potential problems of Linear Regression

1. *Non-linearity of the response-predictor relationships*

$y = A x^n + \epsilon$

Hyper Parameter

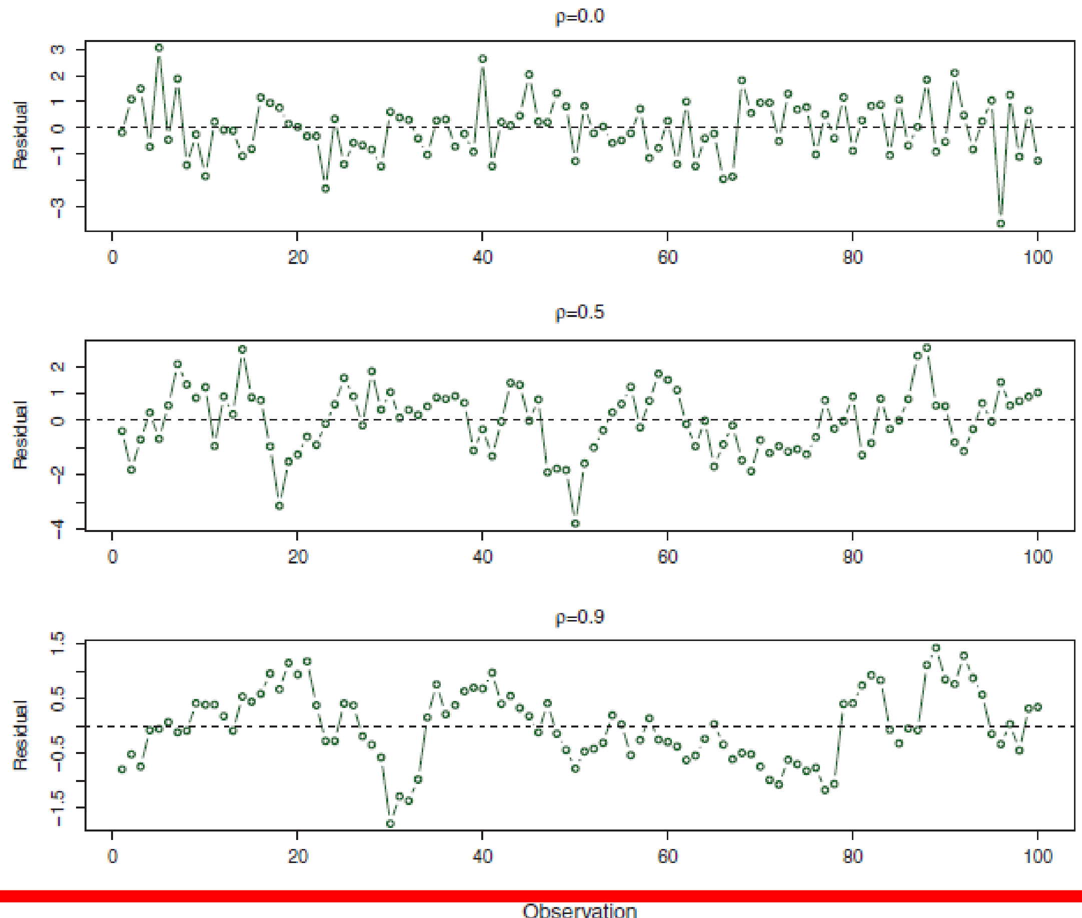
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$



Plots of residuals versus predicted (or fitted) values for the Auto *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* mpg *on* horsepower. *A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* mpg *on* horsepower *and* horsepower². *There is little pattern in the residuals.*

Potential problems of Linear Regression

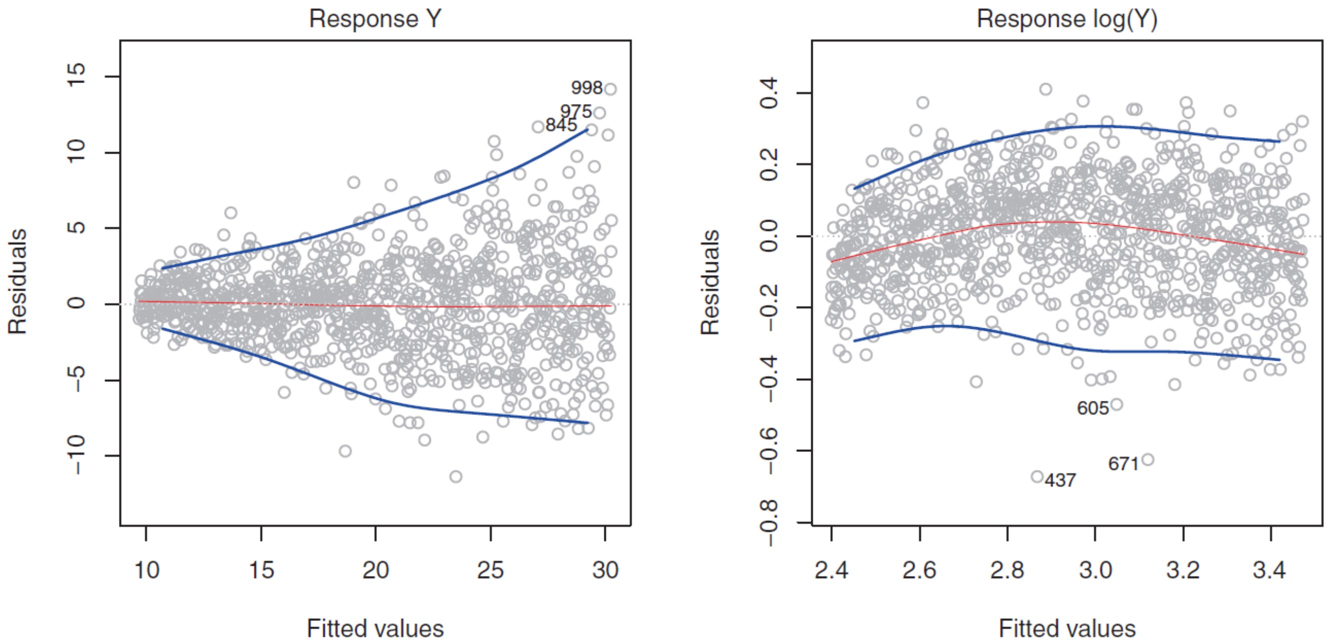
2. *Correlation of error terms*



Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

Potential problems of Linear Regression

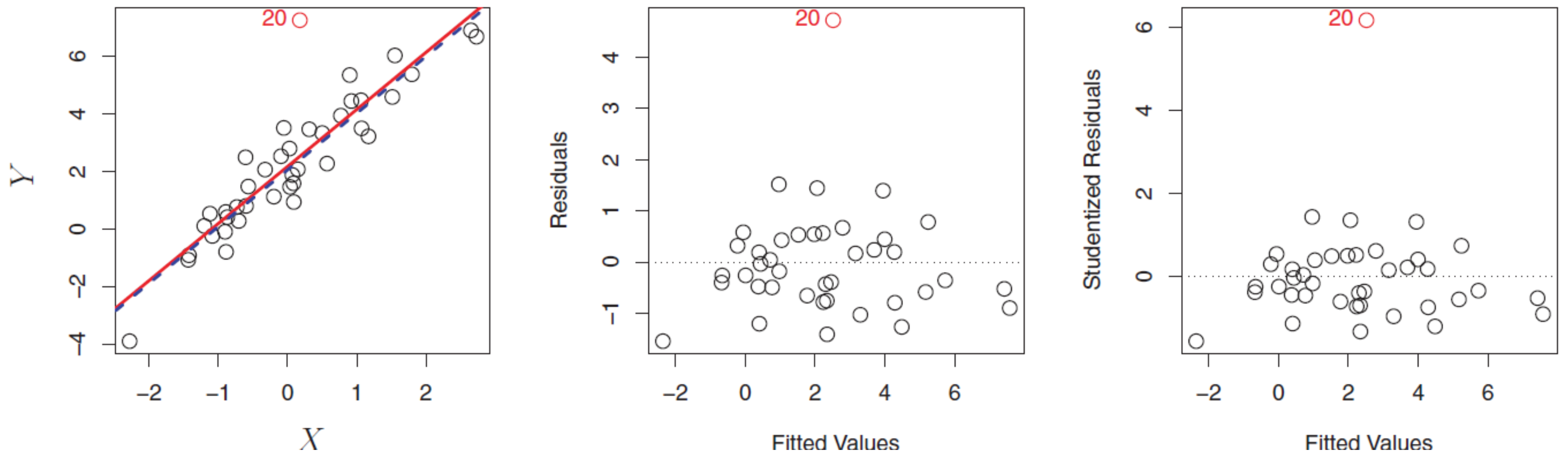
3. *Non-constant variance of error terms.*



Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity*

Potential problems of Linear Regression

4. *Outliers.*



Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of* 6*; typically we expect values between* −3 *and* 3*.*

Robust Regression

- The average quadratic error functional (RSS) is very sensitive to outliers
- Robust error functionals aim to reduce the influence of outliers.
- Linear regression with robust error functionals is called robust linear regression.

Robust Regression

- One example of a robust error functional is the Huber function where the errors are only squared if they are smaller than a threshold ε > 0, otherwise they have only a linear impact

$$E_H = \sum_{k=1}^{n} \begin{cases} \boxed{e_k^2} & \text{if } |e_k| < \epsilon \\ \boxed{2\epsilon \cdot |e_k| - \epsilon^2} & \text{otherwise} \end{cases}$$

e_k^2

Robust Regression

- Another example of a robust error functional is least trimmed squares which sorts the errors so that

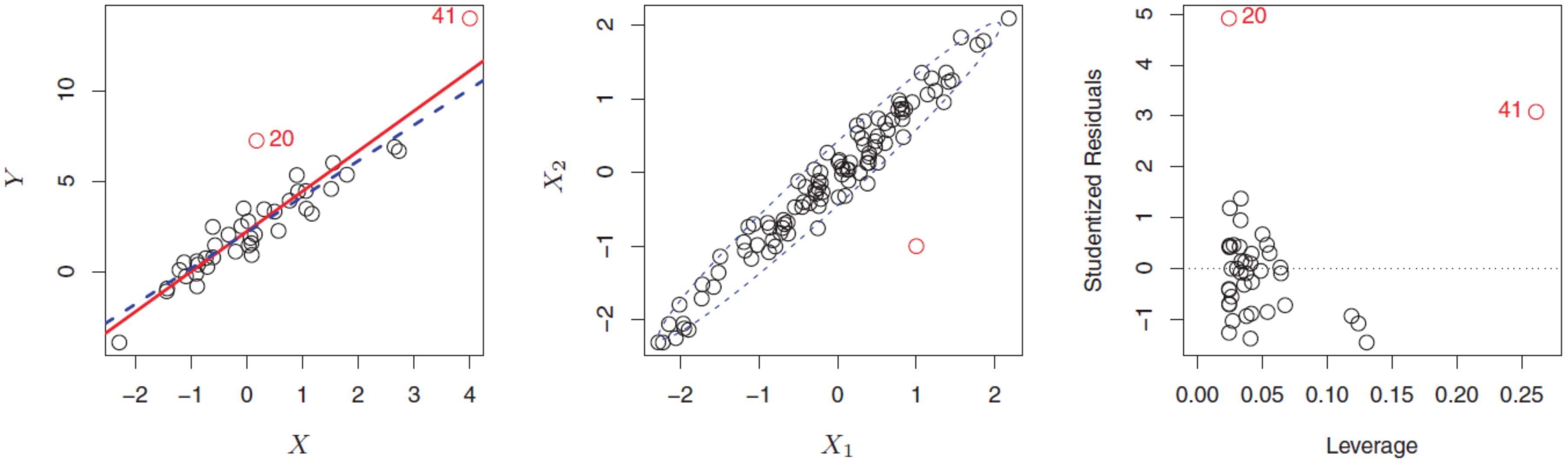
$$e'_1 \leq e'_2 \leq \ldots \leq e'_n$$

and only considers the m smallest errors, $1 \leq m \leq n.$

$$E_{LTS} = \sum_{k=1}^{m} e'^2_k$$

Potential problems of Linear Regression

5. *High-leverage points*



Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: *The red observation is not unusual in terms of its X*1 *value or its X*2 *value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual*

High-leverage points

- In order to quantify an observation's leverage, we compute the *leverage statistic*.
- A large value of this statistic indicates an observation with high leverage.
- For a simple linear regression,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

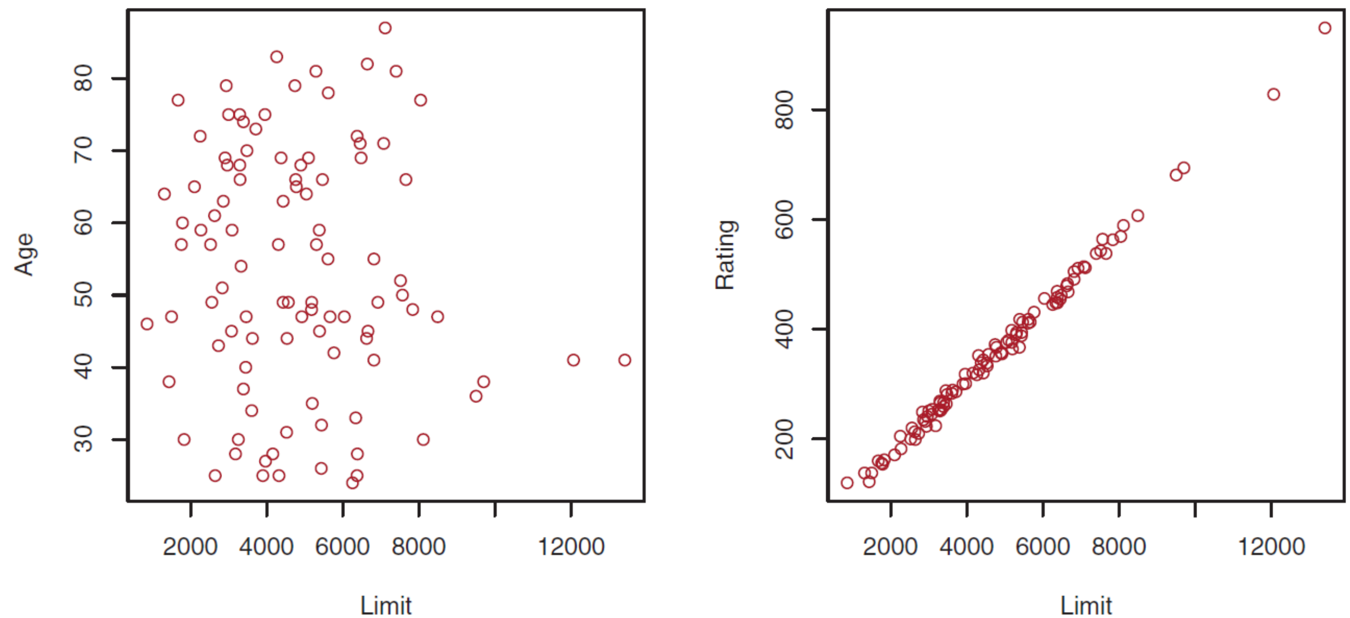
- The leverage statistic h_i is always between $1/n$ and 1.
- The average leverage for all the observations is always equal to $(p+1)/n$.

Potential problems of Linear Regression

$y = f(x) + \epsilon$

$$\boxed{\beta^* = (X^T X)^{-1} X^T y}$$

6. *Collinearity*



Scatterplots of the observations from the Credit *data set.* Left: *A plot of* age *versus* limit. *These two variables are not collinear.* Right: *A plot of* rating *versus* limit. *There is high collinearity.*

Singular Value Decomposition (SVD)

$$\begin{bmatrix} \Sigma_{ii} & 0 \\ 0 & 0 \end{bmatrix}$$

$$\underset{m \times n}{M} = \underset{m \times m}{U} \; \underset{m \times n}{\Sigma} \; \underset{n \times n}{V^*}$$

Both U and V^* are unitary matrix (Real or complex)

$\Rightarrow U^*U = I_{m \times m}$ and $V^*V = I_{n \times n}$

Σ is a diagonal rectangular matrix of non-negative real numbers.

Singular Value Decomposition (SVD)

Singular value decomposition is a generalization of the eigen- decomposition of a square Matrix to a non-square matrix.

$$M = U\Sigma V \wedge *$$

Where,

M is a real of complex m x n matrix

U is an m x m, real or complex unitary matrix (conjugate transpose, U^* is also its inverse)

Σ is an m x n rectangular diagonal matrix with non-negative real numbers

V^* is an n x n, real or complex unitary matrix

If $M$is real then U and $V$are real orthogonal matrices

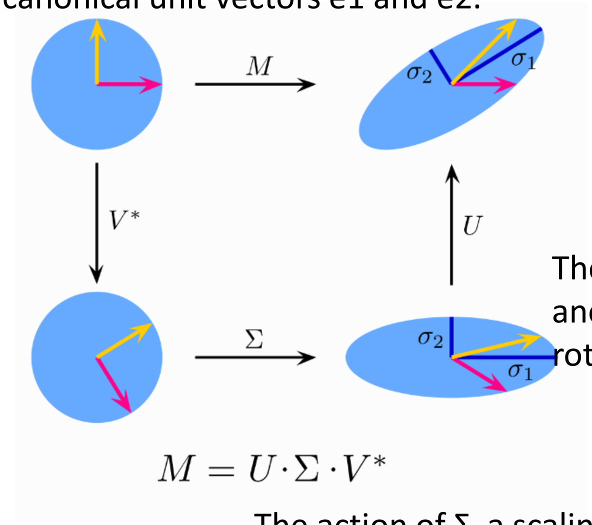
The diagonal values of Σ are know as the singular values. By convention they are written in descending order. In this case Σ (but not always $U$and V^*) is uniquely determined by M.

Singular Value Decomposition (SVD)

$$M = U\Sigma V^*$$

Illustration of the singular value decomposition UΣV* of a real 2×2 matrix M.

The action of M, indicated by its effect on the unit disc D and the two canonical unit vectors e1 and e2.



The action of V*, a rotation, on D, e1, and e2.

The action of U, another rotation.

The action of Σ, a scaling by the singular values σ1 horizontally and σ2 vertically.

Compact Singular Value Decomposition

Compact singular value decomposition is similar to SVD with a square diagonal matrix Σ_{d} .

$$M = U_c \Sigma_{\mathrm{d}} V_c^*$$

Where,

M is a real of complex m x n matrix

U_c is an m x r, semi-unitary matrix ($\mathrm{U}_{\mathrm{c}}^* U_c = I_{rxr}$)

Σ_{d} is an r x r square diagonal matrix with positive real numbers

V_c is an n x r , semi-unitary matrix ($\mathrm{V}_{\mathrm{c}}^* \mathrm{V}_c = I_{rxr}$)

$r \leq \min\{n, m\}$ is the rank of M matrix, and Σ_{d} has only the non-zero singular values of M.

Thus, $\Sigma = \begin{bmatrix} \Sigma_d & 0 \\ 0 & 0 \end{bmatrix}$

Singular Value Decomposition (SVD)

$$M_{m \times n} = U_{m \times m} \Sigma_{m \times n} V^*_{n \times n}$$

The diagonal values are called the singular values.

Compact Singular Value Decomposition

Σ [large → singular]

Semi-unitary Matrix $\Rightarrow U^*U = I$

$V^*V = I$

$$\underset{m\times n}{M} = \underset{m\times r}{U}\ \underset{r\times r}{\Sigma}\ \underset{r\times n}{V^*}$$

$V_{n\times r}$

$r \leq \min(m,n)$ is the rank of M

$\Sigma_{r\times r}$ is a non-zero diagonal matrix, it has non-zero singular value

Ridge Regression

$$y = f(x) + \epsilon$$

Random Error with mean $\Rightarrow$ Singular values will never be zero if for colinear points but will be close to zero.

Ridge Regression

If singular values are not zero then

$(X^T X)^{-1}$ will exist.

$$X = U \Sigma V^T$$

$$\Sigma = \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} \Sigma_d & 0 \\ 0 & 0 \end{bmatrix}$$

$X = U \Sigma V^T = U \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} V^T$ → diagonal matrix with singular values

Ridge Regression

$$(X^T X)^{-1} = \Big(V \begin{bmatrix} \Sigma_d & 0 \end{bmatrix} \underbrace{U^T U}_{I} \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix} V^T \Big)^{-1}$$

If singular values are very small

$$= \left(V \Sigma_d^2 V^T \right)^{-1}$$

⇒ large β

$$= (V^T)^{-1} (\Sigma_d^2)^{-1} (V)^{-1}$$

$$= V \Sigma_d^{-2} V^T$$

→ diagonal with $\frac{1}{\text{singular}^2}$

Ridge Regression

OLS regression

$$\min_{\beta} \| X\beta - Y \|_2^2 \Rightarrow (X^T X)^{-1} X^T Y$$

Ridge Regression

$$\min_{\beta} \| X\beta - Y \|_2^2 + \lambda \| \beta \|_2^2$$

+ve fitting parameter

Ridge Regression

$$RSS(\beta) = \| X\beta - y \|_2^2 + \lambda \| \beta \|_2^2$$

$$= \beta^T X^T X \beta - 2\beta^T X^T y$$
$$+ y^T y + \lambda \beta^T \beta$$

$$\nabla_\beta RSS(\beta) = 2X^T X\beta - 2X^T y + 2\lambda\beta$$

$$\nabla RSS(\beta) = 0$$

Ridge Regression

$$\nabla_\beta RSS(\beta) = 2X^T X\beta - 2X^T y + 2\lambda\beta$$

gy $\nabla RSS = 0$

$$\Rightarrow \quad 2X^T X\beta - 2X^T y + 2\lambda\beta = 0$$

$$(2X^T X + 2\lambda I)\beta = 2X^T y$$

$$\boxed{\beta = (X^T X + \lambda I)^{-1} X^T y}$$