北京师范大学 香港浸会大学 联合国际学院

BEIJING NORMAL UNIVERSITY · HONG KONG BAPTIST UNIVERSITY
UNITED INTERNATIONAL COLLEGE

STAT2013 Regression Analysis

Semester 1, 2018-2019

Room: T3-401-R1

Instructor: Ye, Huajun Terry

TA：Jenna Otto

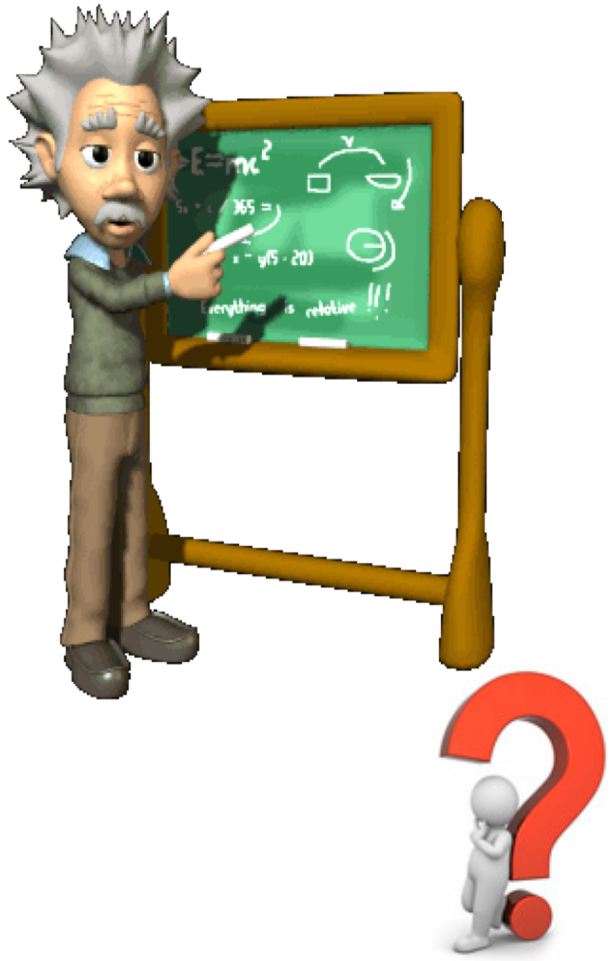
Email:hjye@uic.edu.hk

Jotto@uic.edu.hk (TA)

Chapter 7
Advanced Topics on Regression Models

In this Chapter, we will know some important concepts.

- Nonlinear regression
- Multicolinearity and ridge regressions
- AIC and BIC
- Regression with a binary response (Logistic Regression)



7.1 Nonlinear regression

The linear regression models provide a sufficiently large and complex range of models to suit the needs of many analysts. Yet linear regression cannot be expected to be appropriate for all problems. So we need to study nonlinear regression models:

$$y = f(x_1, \ldots, x_p; \theta_1, \ldots, \theta_q) + \varepsilon$$
$$\equiv f(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

where

y	:Response (dependent variable)
$x_1, \ldots, x_p$	:Regressors
$\theta_1, \ldots, \theta_q$	:Parameters (unknown)
ε	:Random error, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$
f	:Known function

We want to estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ and σ^2 based on a set of observations $\{(y_i, x_{ip}, \ldots x_{ip}), i = 1, \ldots, n\}$

The model

$$y = f(\boldsymbol{X}, \theta) + \varepsilon$$

is called *nonlinear model* if f is a nonlinear function of θ; otherwise, called *linear model*.

Example 7.1.1

$$y = \beta_0 + \beta_1 \log x + \varepsilon \qquad \leftarrow \textit{Nonlinear model in } x$$

Let $x^* = \log x$, then

$$y = \beta_0 + \beta_1 x^* + \varepsilon \qquad \leftarrow \textit{Linear model}$$

Example 7.1.2

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \varepsilon$$

$\uparrow$ *Nonlinear model in* x_1 *and* x_2

Let $x_3^* = x_1^2, x_4^* = x_1 x_2$ and $x_5^* = x_2^2$, then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_3^* + \beta_{12} x_4^* + \beta_{22} x_5^* + \varepsilon$$

$\uparrow$ *Linear model*

Example 7.1.3

$$y = \alpha e^{\beta x} + \varepsilon \qquad \leftarrow \textit{Nonlinear model in } \beta \textit{ and } x$$

$$\hat{y} = \alpha e^{\beta x}$$

$$\underbrace{\log \hat{y}}_{y^*} = \underbrace{\log \alpha}_{\beta_0} + \beta x \qquad \leftarrow \textit{Linear model}$$

This model can be treated as linear or nonlinear model.

Example 7.1.4

$$y = \alpha + \beta_1 x_1^{r_1} + \beta_2 x_2^{r_2} + \varepsilon \qquad \leftarrow \textit{Nonlinear model in } r_1 \textit{ and } r_2$$

Example 7.1.5

$$y = \frac{\alpha}{a + \exp\{-\beta_1 x_1 - \beta_2 x_2\}} + \varepsilon \qquad \leftarrow \textit{Nonlinear model in } \alpha,\ \beta_0,\ \beta_1,\ \beta_2$$

Linearization Method

When $f(\boldsymbol{X};\theta)$ is a nonlinear function of $\boldsymbol{X}$, some transformation of $x_1,\ldots,x_p$ are helpful.

Example 7.1.6

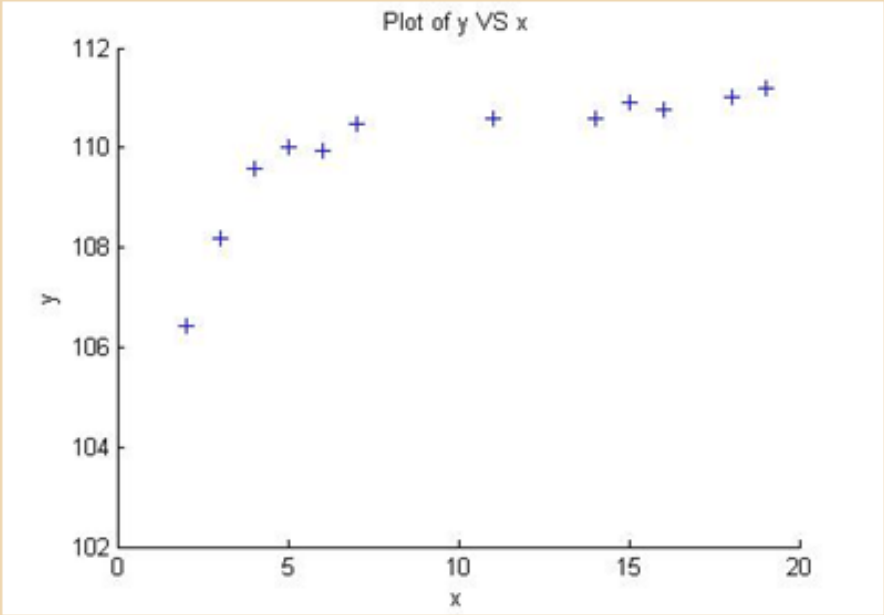
The volume of the steel tank varies by steel corrosion.

x: the number of using times of the steel tank
y: the volume of the tank

x	2	3	4	5	6	7
y	106.42	108.20	109.58	110.00	109.93	110.49
x	11	14	15	16	18	19
y	110.59	110.60	110.90	110.76	111.00	111.20

Example 7.1.6 (*cont*)

Figure 7.1a



The plot above suggests to fit the data by the model:

$$E(y) = \frac{x}{\beta_1 + \beta_0 x} = \frac{1}{\beta_0 + \beta_1 / x} \qquad (1)$$

Example 7.1.6 (*cont*)

Let $y^* = 1/y$ and $x^* = 1/x$. The model becomes:

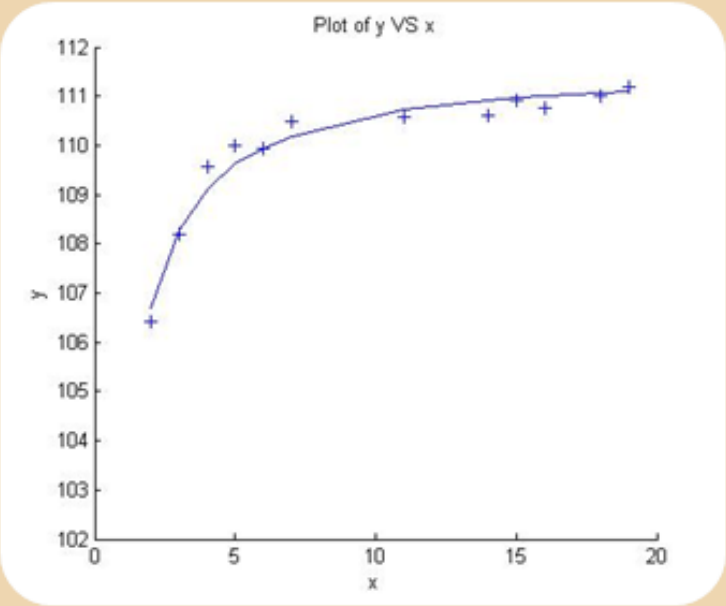
$$E(y^*) = \beta_0 + \beta_1 x^* \qquad (2)$$

By the least squares method, we have:

$$y = \frac{x}{0.00083 + 0.00896x}$$

If we plot the model above in figure 7.1a, we have:

Figure 7.1b



Example 7.1.6 (*cont*)

Remarks:

- The above method is not the real LSE method
- The $R^{*2} = 0.9660$ is based on model (2). For model (1):

  R^2 is different from R^{*2} .

  It can be easily calculated as:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0.7341$$

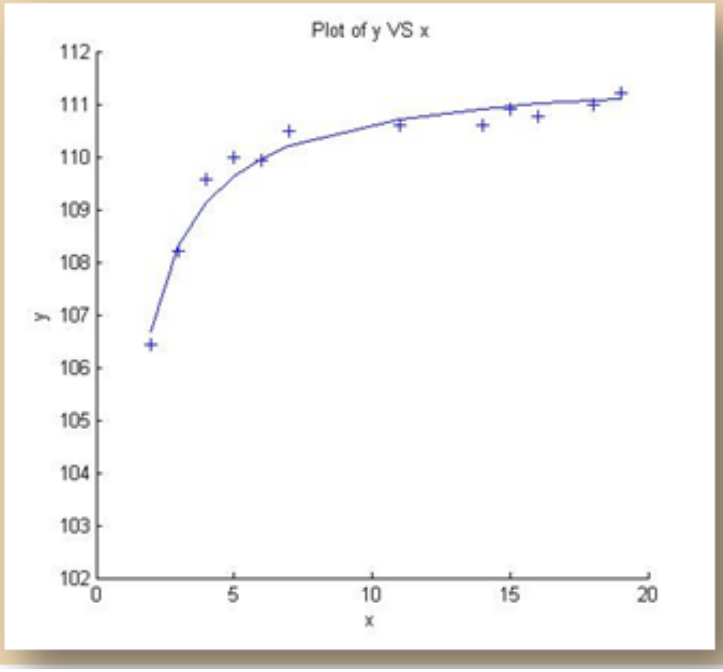
$$R^2 = \frac{SST - RSS}{SST} = 0.9650 \qquad \hat{\sigma} = \sqrt{\frac{RSS}{n-2}} = \sqrt{0.07341} \cong 0.2709$$

Example 7.1.6 (*cont*)

This data set can be also fitted by other models, for example,

$$\hat{y} = \beta_0 e^{\beta_1 / x} = 111.632 e^{-0.09066 / x}$$

Figure 7.1c



- Cox Models

➢ Cox's remarkable achievement was that his methods allowed to first estimate β in

$$h_{\mathbf{x}}(y) = h_0(y)e^{\beta^T \mathbf{x}}$$

➢ The baseline hazard in data that are possibly right-censored.

➢ Importantly, the baseline hazard does not need to be first estimated before the regression coefficients are estimated.

Kernel Smoothing and Spline Smoothing

- The aim of a regression analysis is to produce a reasonable analysis to the unknown response function m, where for n data points (X_i, Y_i), the relationship can be modeled as

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \cdots, n$$

- Unlike parametric approach where the function m is fully described by a finite set of parameters, nonparametric modeling accommodate a very flexible form of the regression curve.

Kernel Smoothing

- A reasonable approximation to the regression curve $m(x)$ will be the mean of response variables near a point x. This *local averaging procedure* can be defined as

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} W_{ni}(x) Y_i$$

Every smoothing method to be described is of the form above.

- The amount of averaging is controlled by a *smoothing parameter*. The choice of smoothing parameter is related to the balances between *bias* and *variance*.

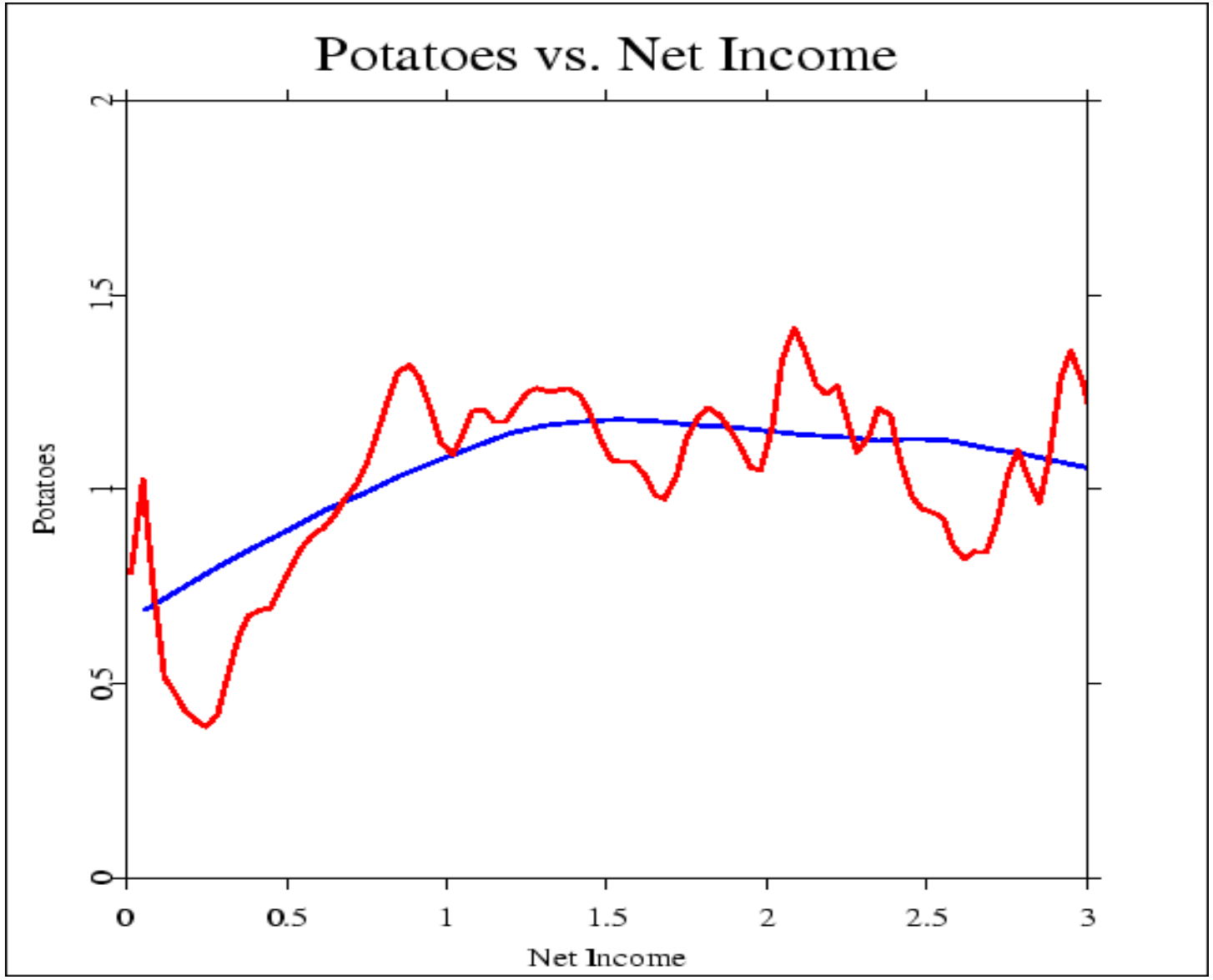


Figure 7.1d. Expenditure of potatoes as a function of net income. h = 0.1, 1.0, n = 7125, year = 1973.

Spline Smoothing

- Spline smoothing quantifies the competition between
  - the aim to produce a good fit to the data
  - the aim to produce a curve without too much rapid local variation.
- The regression curve $\hat{m}_\lambda(x)$ is obtained by minimizing the penalized sum of squares

$$S_\lambda(m) = \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 + \lambda \int_a^b \{m''(x)\}^2 dx$$

where m is twice-differentiable function on [a,b], and λ represents the rate of exchange between residual error and roughness of the curve m.

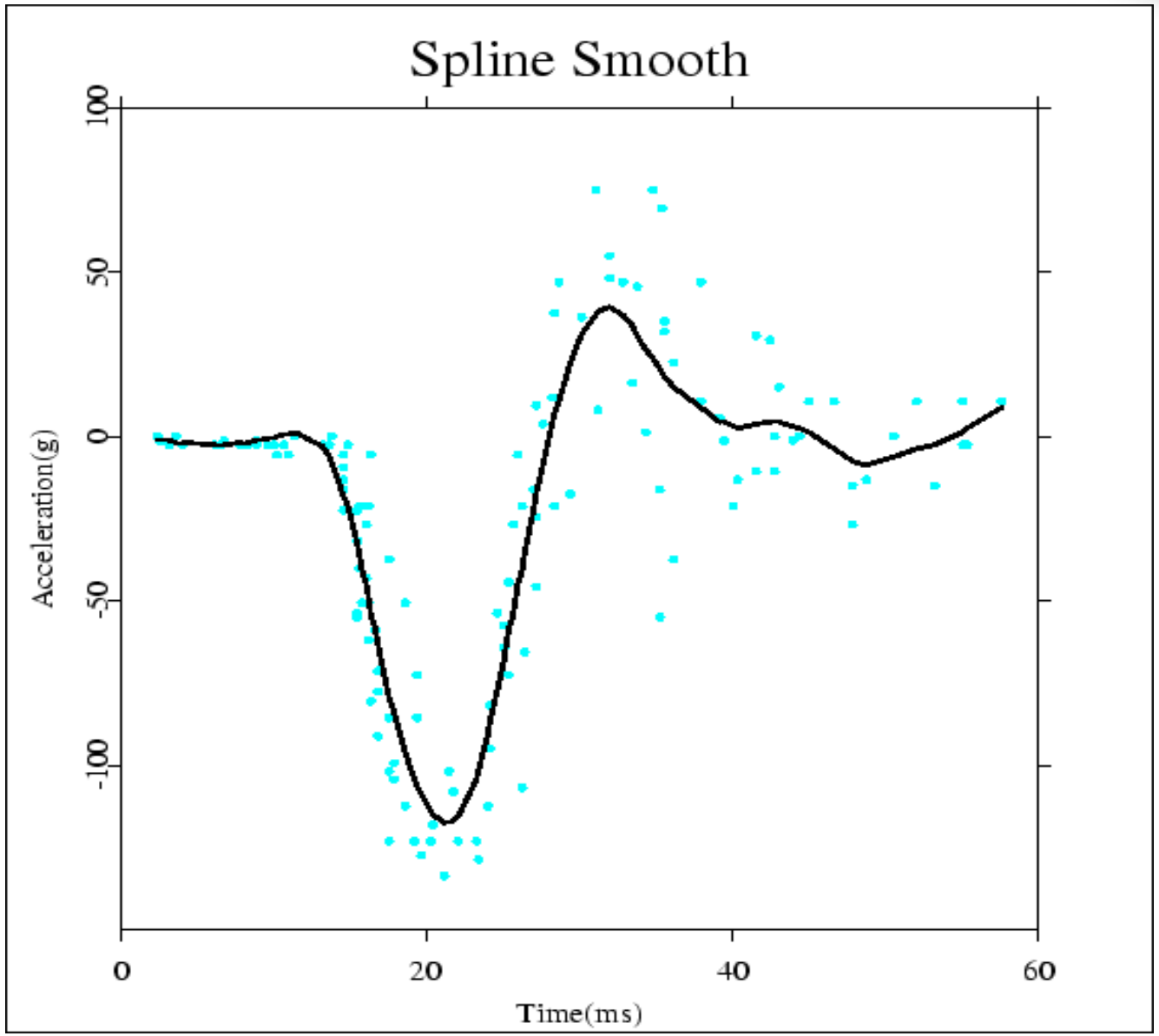


Figure 7.1e. A spline smooth of the Motorcycle data set.

7.2 Multicolinearity and Ridge Regression

Strong multicollinearity can result in large variances and covariances for the least squares estimates of the coefficients. $\mathbf{C} = (\mathbf{X'X})^{-1}$ and

$$C_{jj} = \frac{1}{1 - R_j^2}$$

Strong multicollinearity between x_j and any other regressor variable will cause R_j^2 to be large, and thus C_{jj} to be large.

In other words, the variance of the least squares estimate of the coefficient will be very large.

Variance Inflation Factors (VIF)

- Variance inflation factors are very useful in determining if multicollinearity is present.

$$VIF_j = C_{jj} = (1 - R_j^2)^{-1}$$

- VIFs > 5 to 10 are considered significant. The regressors that have high VIFs probably have poorly estimated regression coefficients

The **variance inflation factor (VIF)** is calculated from **two steps:**

1. Run an OLS regression that has X_i as a function of all the other explanatory variables in the equation—For i = 1, this equation would be:

$$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots + \alpha_K X_K + v$$

where v is a classical stochastic error term

2. Calculate the variance inflation factor for $\hat{\beta}_i$:

$$\text{VIF}(\hat{\beta}_i) = \frac{1}{(1 - R_i^2)}$$

where R_i^2 is the unadjusted R^2 from **step one**

Ridge Regression

Least squares estimation gives an unbiased estimate,

$$E(\hat{\beta}) = \beta$$

with minimum variance – but this variance may still be very large, resulting in unstable estimates of the coefficients.

Alternative: Find an estimate that is biased but with smaller variance than the unbiased estimator

Ridge Estimator $\hat{\beta}_R$

$$\hat{\beta}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$
$$= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$
$$= \mathbf{Z}_k\hat{\boldsymbol{\beta}}$$

k is a "biasing parameter" usually between 0 and 1.

The effect of k on the MSE

Recall: $MSE(\hat{\beta}^*) = Var(\hat{\beta}^*) + (bias)^2$

Now, $MSE(\hat{\beta}_R^*) = Var(\hat{\beta}_R^*) + (bias)^2$

$$= \sigma^2 \Sigma \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \boldsymbol{\beta}' (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2} \boldsymbol{\beta}$$

As $k \uparrow$, Var $\downarrow$, and bias $\uparrow$

Choose k such that the reduction in variance $>$ increase in bias.

$$SS_{\text{Res}} = (y - x\hat{\beta}_R)'(y - x\hat{\beta}_R)$$

Ridge Trace

Plots k against the coefficient estimates. If multicollinearity is severe, the ridge trace will show it. Choose k such that $\hat{\beta}_R$ is stable and hope the MSE is acceptable

Ridge regression is a good alternative if the model user wants to have all regressors in the model.

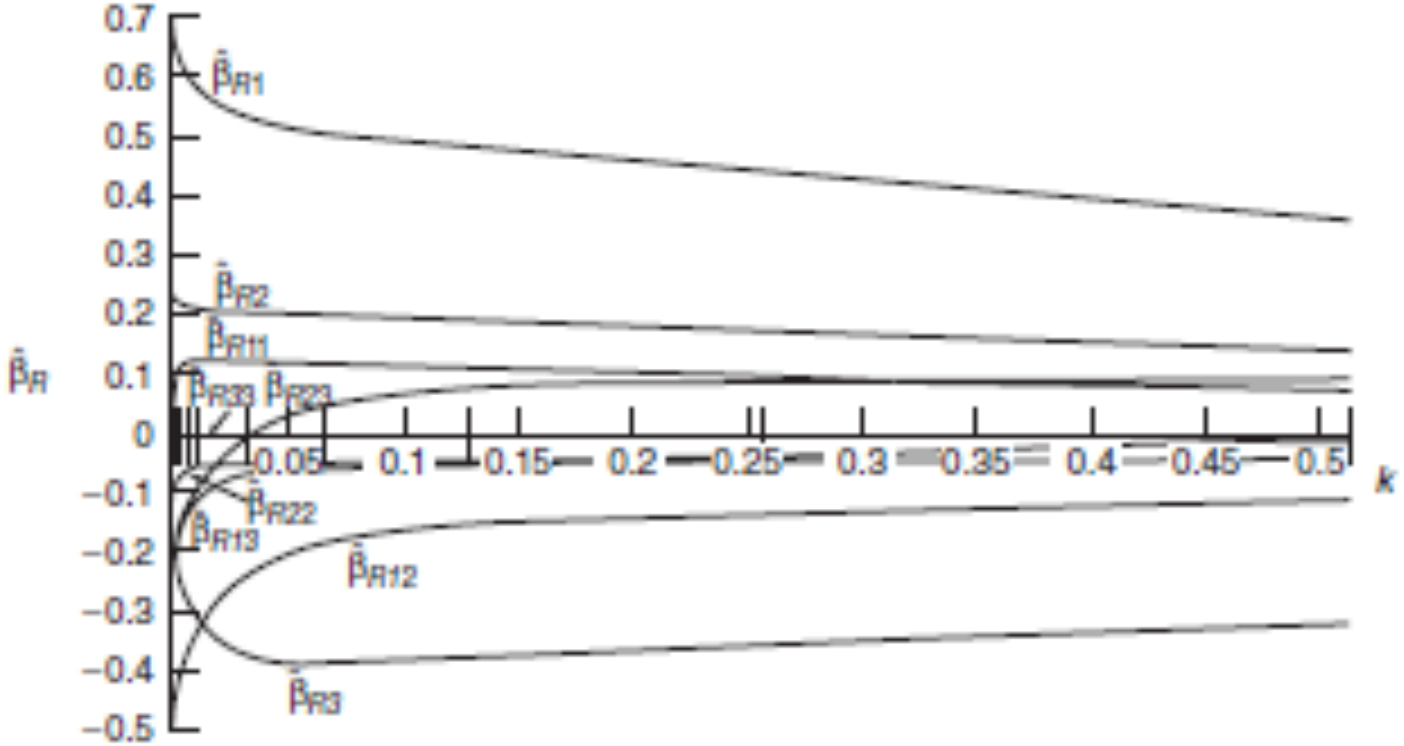


Figure 9.5 Ridge trace for acetylene data using nine regressors.

More About Ridge Regression

- Methods for choosing k
- Relationship to other estimators
- Ridge regression and variable selection
- Generalized ridge regression (a procedure with a biasing parameter k for each regressor

7.3 AIC and BIC

Akaike Information Criteria (AIC)

$$AIC = -2\log L + 2p$$

Akaike, Hirotugu (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* **19** (6): 716–723..

Bayesian Information Criteria (BIC/SIC)

$$BIC = -2\log L + p\ln(n)$$

Schwarz, Gideon E. (1978). "Estimating the dimension of a model". *Annals of Statistics* **6** (2): 461–464.

n = sample size

p = the number of free parameters to be estimated

L = the maximized value of the likelihood function for the estimated model

AIC versus BIC

$$2p \;\; vs. \;\; p\ln(\mathrm{n})$$

- BIC and AIC are similar
- Different penalty for number of parameters
- The BIC penalizes free parameters more strongly than does the AIC.
- Implications: BIC tends to choose smaller models
- **The larger the n, the more likely that AIC and BIC will disagree on model selection**

7.4 Regression with a binary response

Logistic Regression: Model

Y: Binary outcome variable (I.e. a categorical variable with only two levels).

Example: Smoking Status (Yes/No); Application Status (Admitted/Denied);

Assumption: *Y* has a binomial distribution

X: Explanatory variable

$\pi(x)$*:* the probability of "success" when *X* takes value *x*.
$\pi(x)$ is the parameter for the binomial distribution.

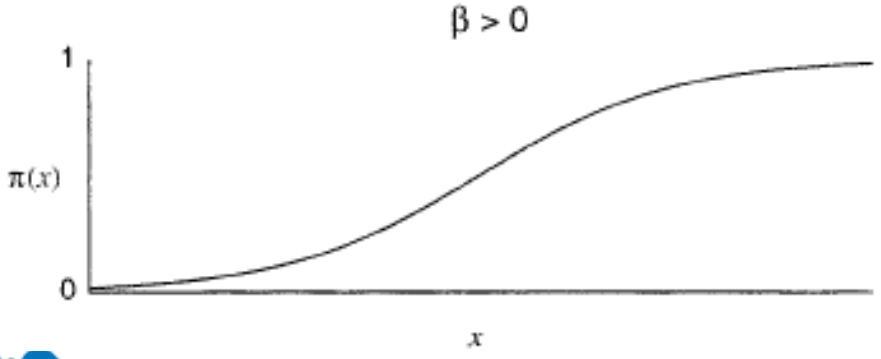
Logistic Regression: Model

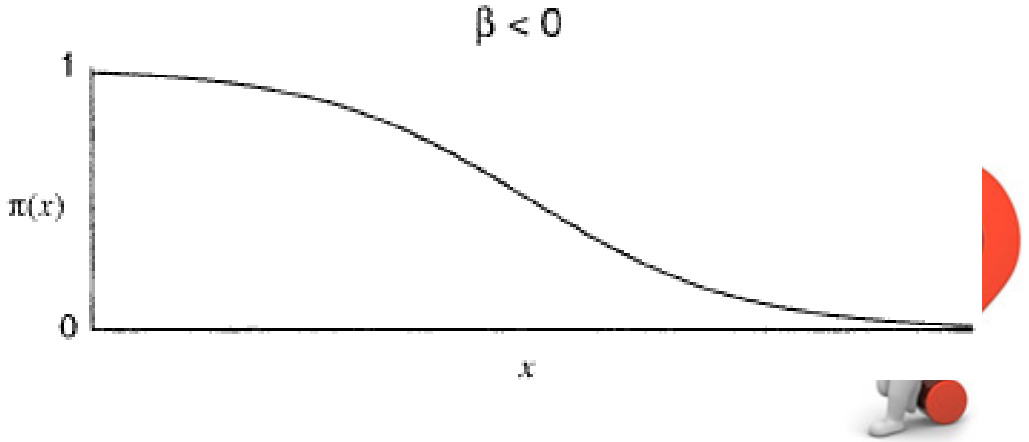
Supposed there is a single explanatory variable X, which is quantitative.

For a binary response variable Y, **π(x)** denotes the "success" probability at value x. The **logistic regression model** is as follows:

$$\log it(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

The formula implies that **π(x)** increases or decreases as an S-shaped function of x.





Logistic Regression: Model

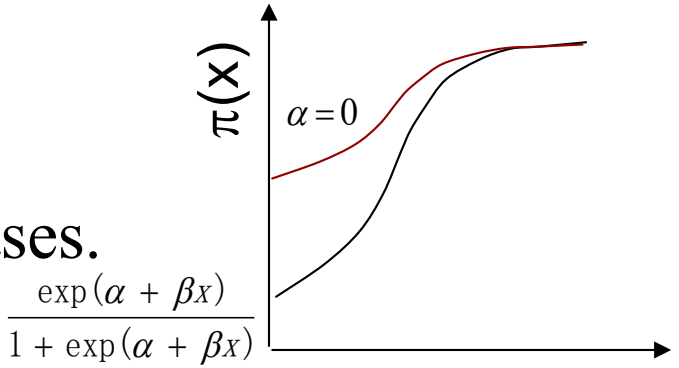
Model: $$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

The parameter β determines the rate of increase or decrease of the S-shaped curve for $\pi(x)$.

$\beta > 0$: ascending

$\beta < 0$: descending

And the rate of change increases as $|\beta|$ increases.



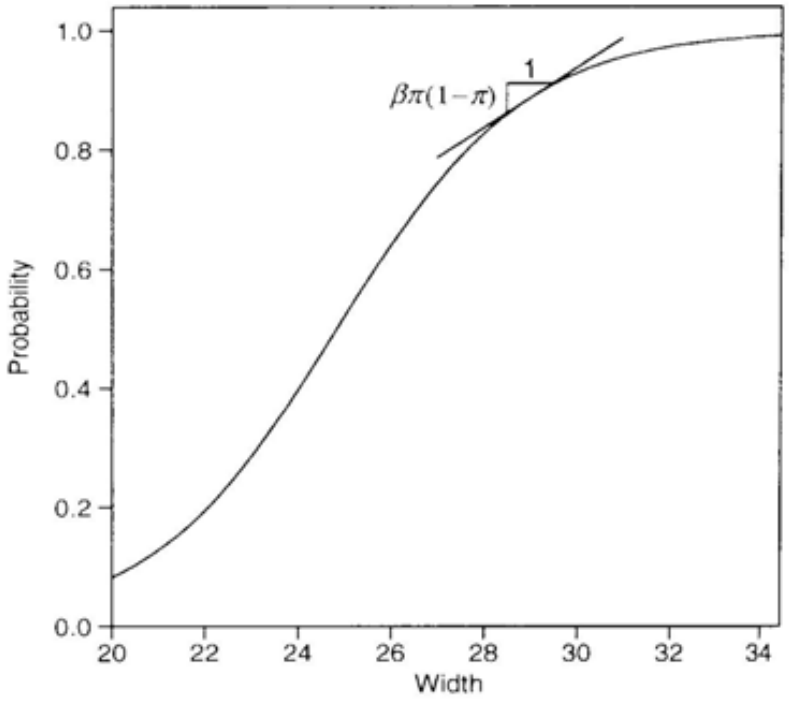
When $\beta = 0$, $\pi(x) = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$.

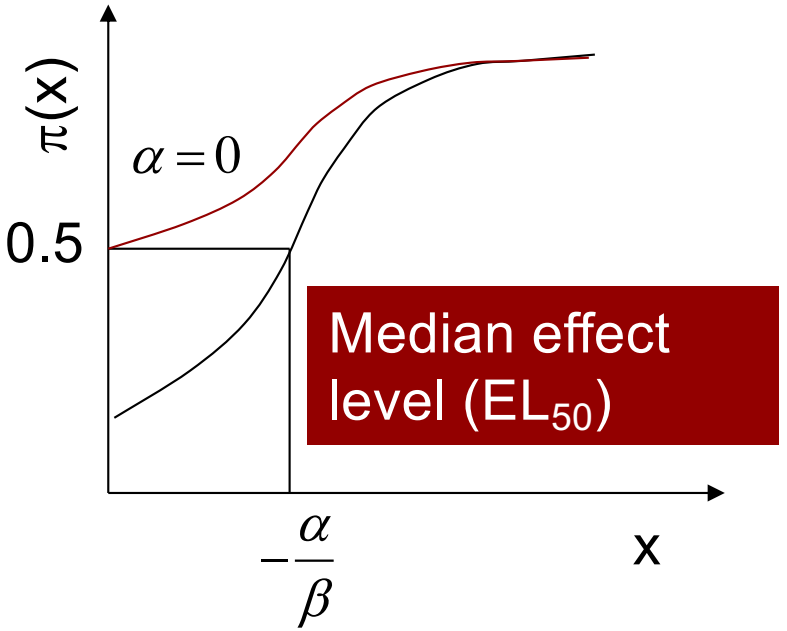
Then $\pi(x)$ is identical at all x, so the binary response Y is independent of X.

32

Logistic Regression: Model

$$\frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$





A straight line drawn tangent to the curve at a particular x value, such as the left figure. For logistic regression parameter β that line has slope equal to

$$\pi'(x) = \beta\pi(x)(1-\pi(x))$$

The slope approaches 0 as the probability approaches 1.0 or 0.

When $\pi(x) = 0.5$, the slope is steepest, the x values relating to the logistic regression parameters is $x = -\frac{\alpha}{\beta}$

We call this x value the median effective level and denote it by EL_{50} . It represents the level at which each outcome has a 50% chance.