2. Regression Review

2.1 The Regression Model

The general form of the regression model

$$y_t = f(\mathbf{x}_t, \beta) + \varepsilon_t$$

where

$$\mathbf{x}_t = (x_{t1}, \cdots, x_{tp})', \beta = (\beta_1, \ldots, \beta_m)'.$$

- ε_t is a random variable, $\mathrm{E}\varepsilon_t = 0, \mathrm{Var}(\varepsilon_t) = \sigma^2$, and the error ε_t are uncorrelated.

- Normal distribution assumption for error ε_t implies independence among error

Examples:

1. $y_t = \beta_0 + \varepsilon_t$ (constant mean model)

2. $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ (simple linear regression model)

3. $y_t = \beta_0 \exp(\beta_1 x_t) + \varepsilon_t$ (exponential growth model)

4. $y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \varepsilon_t$ (quadratic model)

5. $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t$ (linear model with two independent variables)

6. $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_{11} x_{t1}^2 + \beta_{22} x_{t2}^2 + \beta_{12} x_{t1} x_{t2} + \varepsilon_t$ (quadratic model with two independent variables)

Linear Models: see Example 1, 2, 4, 5, 6

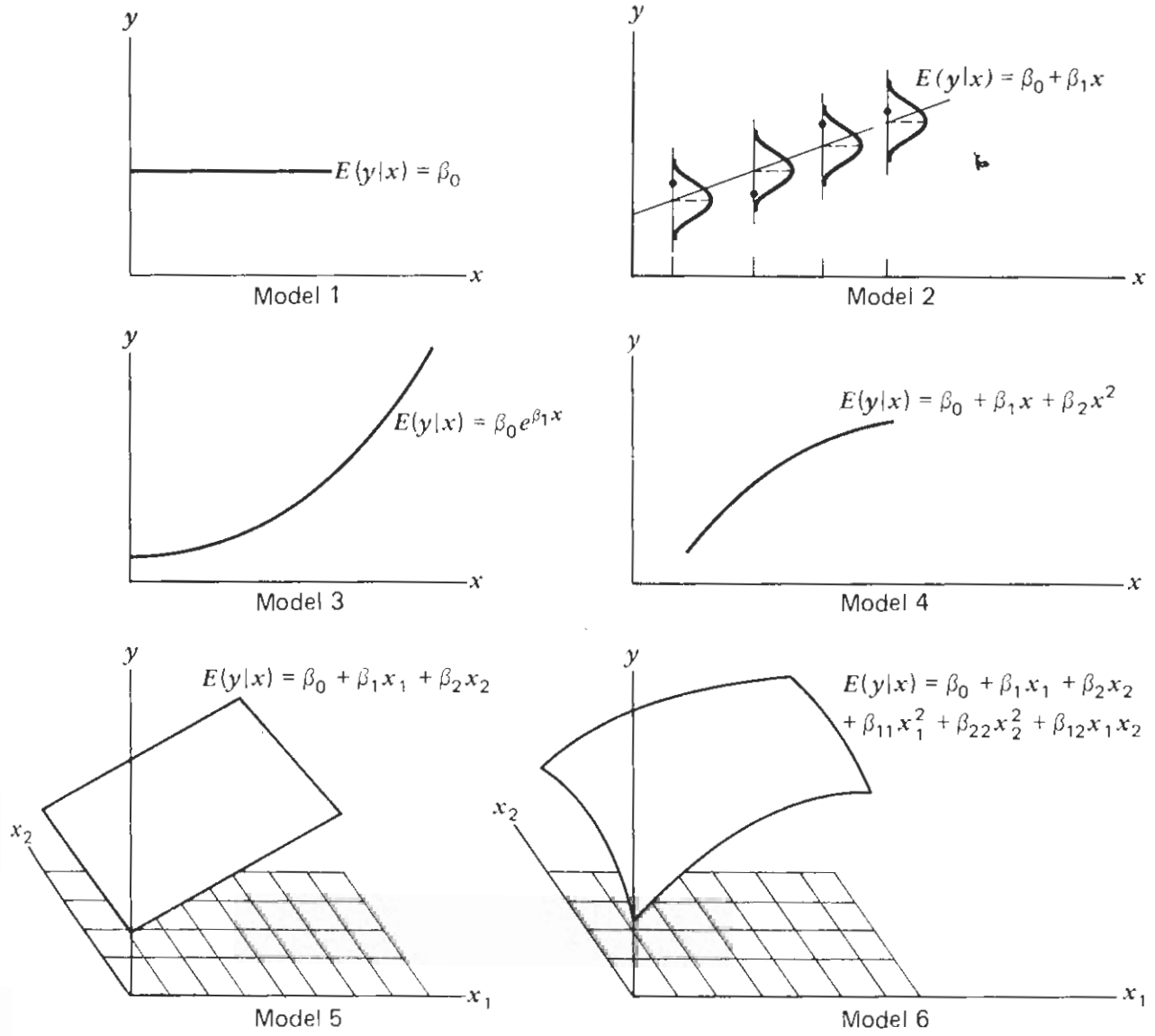


Figure 2.1. Graphical representation of the regression models 1–6. The dots in model 2 represent possible realizations.

2.2 Prediction from regression models with known coefficients

- prediction y_k^{pred} for $y_k = f(\mathbf{x}_k, \beta) + \varepsilon_k$
- forecast error $e_k = y_k^{\text{pred}} - y_k$
- The expected value of the squared forecast error

$$\mathrm{E}(e_k^2) = \mathrm{E}(y_k^{\text{pred}} - y_k)^2 = \sigma^2 + \mathrm{E}[f(\mathbf{x}_k, \beta) - y_k^{\text{pred}}]^2$$

- Minimum mean square error(MMSE)

$$y_k^{\text{pred}} = f(\mathbf{x}_k, \beta)$$

- $100(1-\alpha)$ percent prediction interval for the future value y_k

$$[f(\mathbf{x}_k, \beta) - \mu_{\alpha/2}\sigma; f(\mathbf{x}_k, \beta) + \mu_{\alpha/2}\sigma]$$

where $\mu_{\alpha/2}$ is the $100(1-\alpha/2)$ percentage point of the $\mathcal{N}(0,1)$ distribution.

2.3 Least squares estimates of unknown coefficients

The parameter estimates that minimize the sum of the squared deviations

$$S(\beta) = \sum_{t=1}^{n} [y_t - f(\mathbf{x}_t; \beta)]^2$$

are called the least squares estimates and are denoted by $\hat{\beta}$.

Examples

1. Simple linear regression model through the origin

$$y_t = \beta x_t + \varepsilon_t.$$

2. Simple linear regression model

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Example: Running Performance

For illustration, consider the data list in Table. In this example we are interested in predicting the racing performance of trained female distance runners. The measured variables include the running performance in 10-km road race; body composition such as height, weight, skinfold sum, and relative body fat; and maximal aerobic power. The subjects for study were 14 trained and competition-experienced female runners who had placed among the top 20 women in a 10-km road race with more than 300 female entrants. The laboratory data were collected during the week following the competitive run.
For the moment we are interested only in the relationship between the racing performance(running time) and maximal aerobic power(volume of maximal oxygen uptake; V_{O_2}.)

Table 2.1. Physical and Performance Characteristics of 14 Female Runners[a]

X_1	X_2	X_3	X_4	X_5	Y
163	53.6	76.4	17.9	61.32	39.37
167	56.4	62.1	15.2	55.29	39.80
166	58.1	65.0	17.0	52.83	40.03
157	43.1	44.9	12.6	57.94	41.32
150	44.8	59.7	13.9	53.31	42.03
151	39.5	59.3	19.2	51.32	42.37
162	52.1	98.7	19.6	52.18	43.93
168	58.8	73.1	19.6	52.37	44.90
152	44.3	59.2	17.4	57.91	44.90
161	47.4	51.5	14.4	53.93	45.12
161	47.8	61.4	7.9	47.88	45.60
165	49.1	62.5	10.5	47.41	46.03
157	50.4	60.3	12.6	47.17	47.83
154	46.4	76.7	19.6	51.05	48.55

[a] X_1, height; X_2, weight; X_3, skinfold sum; X_4, relative body fat; X_5, V_{O_2}; Y, running time.

Source: Conley et al. (1981).

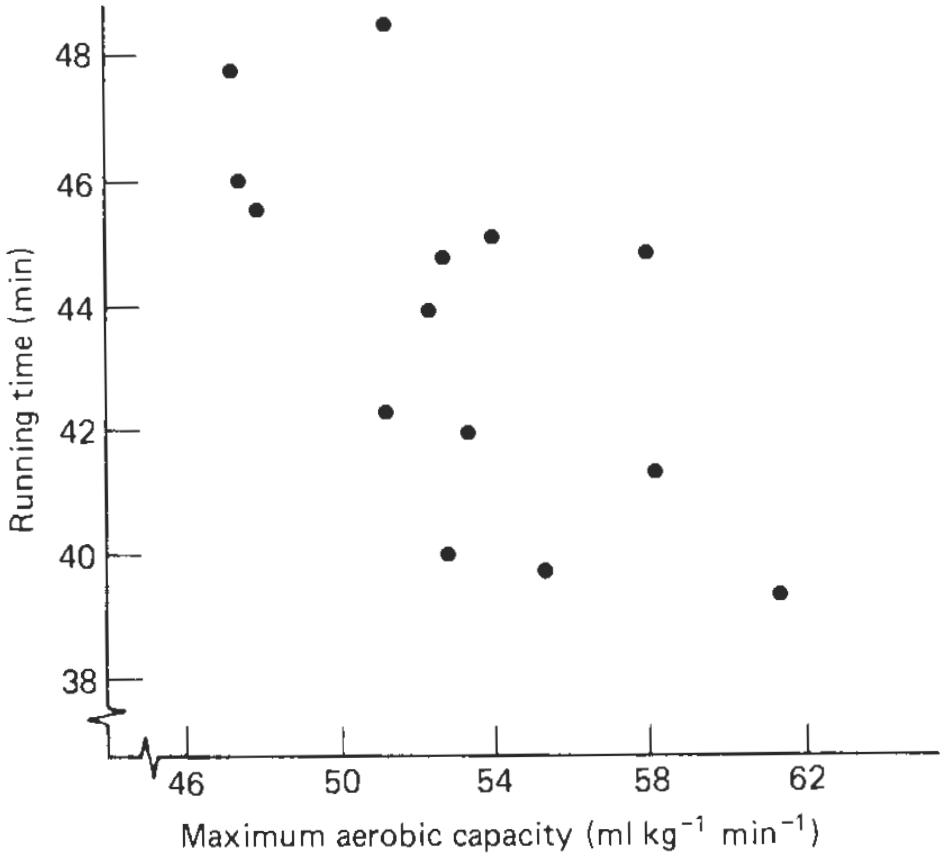


Figure 2.2. Scatter plot of running time against maximal aerobic capacity.

Estimation in the General Linear Regression Model

Linear Regression models can be written as

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_p x_{tp} + \varepsilon_t$$

or

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t$$

where $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tp})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ To simplify the calculation of the least squares estimates for the general linear regression model, we introduce matrix notation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathrm{E}(\boldsymbol{\varepsilon}) = 0, \mathrm{Var}((\boldsymbol{\varepsilon})) = \sigma^2\mathbf{I}$. Then

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{Var}(\mathbf{y}) = \mathrm{E}[\mathbf{y} - \mathrm{E}(\mathbf{y})][\mathbf{y} - \mathrm{E}(\mathbf{y})]' = \sigma^2\mathbf{I}$$

In matrix notation the least squares criterion can be expressed as minimizing

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{n}(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Then the solution is then given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Two special cases,

1. Simple linear regression through the origin:

$$y_t = \beta x_t + \varepsilon_t, \quad \text{for} \quad 1 \leq t \leq n.$$

2. Simple linear regression model

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad \text{for} \quad 1 \leq t \leq n.$$

2.4 Properties of Least Squares Estimation

1.
$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

2. If it is assumed that errors in linear model are normally distribution, then

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$$

3. The difference between the observed and fitted values are called residuals and are given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

Then we have

$$\mathbf{X}'\mathbf{e} = \mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = 0.$$

4. The variation of the observations y_t around their mean $\bar{y}$, *total sum of squares* (SSTO)

$$\text{SSTO} = \sum (y_t - \hat{y})^2 = \sum y_t^2 - n\bar{y}^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2.$$

The variation of fitted value $\hat{y}_t$ around the mean $\bar{y}$, *sum of squares due to regression* (SSR)

$$\text{SSR} = \sum (\hat{y}_t - \bar{y})^2 = \sum \hat{y}_t^2 - n\bar{y}^2 = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2.$$

sum of squares due to error (or residual) (SSE)

$$\text{SSE} = \sum (y_t - \hat{y}_t)^2 = \mathbf{e}'\mathbf{e}.$$

Then

$$\text{SSTO} = \text{SSR} + \text{SSE}$$

and define the *coefficient of determination* R^2 as the ratio of the sum of squares due to regression and the total sum of squares:

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}$$

5. The variance σ^2 is usually unknown. However it can be estimated by the mean square error

$$s^2 = \frac{\text{SSE}}{n-p-1}.$$

Hence the estimated covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

The estimated standard error of $\hat{\beta}_i$ is given by $s_{\hat{\beta}_i} = s\sqrt{c_{ii}}$, where $\sqrt{c_{ii}}$ is the corresponding diagonal element in $(\mathbf{X}'\mathbf{X})^{-1}$.

6. The least square estimate $\hat{\boldsymbol{\beta}}$ and s^2 are independent.

2.5 Confidence Intervals and Hypothesis Testing

Consider the quantity

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}, i = 0, 1, \ldots, p$$

where $\hat{\beta}_i$ is the least square estimate of β_i, and $s_{\hat{\beta}_i} = s\sqrt{c_{ii}}$ is its standard error. This quantity has a *t-distribution with $n - p - 1$ degrees of freedom*.

Confidence intervals

A $100(1 - \alpha)$ confidence interval for the unknown parameter β_i is given by

$$[\hat{\beta}_i - t_{\alpha/2}(n - p - 1)s\sqrt{c_{ii}}, \hat{\beta}_i + t_{\alpha/2}(n - p - 1)s\sqrt{c_{ii}}]$$

where $t_{\alpha/2}(n - p - 1)$ is the $100(1 - \alpha/2)$ percentage point of a t distribution with $n - p - 1$ degree of freedom

Hypothesis Tests for individual Coefficients

$$H_0 : \beta_i = \beta_{i0} \quad \text{vs} \quad H_1 : \beta_i \neq \beta_{i0}$$

The test statistic is given by

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{s\sqrt{c_{ii}}}$$

- If $|t| > t_{\alpha/2(n-p-1)}$, we reject H_0 in favor of H_1 at significance level α.
- If $|t| \leq t_{\alpha/2(n-p-1)}$, there is not enough evidence for rejecting the null hypothesis H_0 in favor of H_1. Thus loosely speaking, we "accept" H_0 at significance level α.
- Special case when $\beta_{i0} = 0$.

2.6 Prediction from Regression Models with Estimated Coefficients

The MMSE prediction of a future y_k from regression model $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_p x_{tp} + \varepsilon_t$ is given by

$$y_k^{\text{pred}} = \beta_0 + \beta_1 x_{k1} + \cdots + \beta_p x_{kp} = \mathbf{x}_k' \boldsymbol{\beta}$$

Replace $\boldsymbol{\beta}$ by their least estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$$\hat{y}_k^{\text{pred}} = \hat{\beta}_0 + \hat{\beta}_1 x_{k1} + \cdots + \hat{\beta}_p x_{kp} = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$$

These predictions have the following properties

1. Unbiased
$$\mathrm{E}(y_k - \hat{y}^{\text{pred}}) = 0$$

2. $\hat{y}^{\text{pred}}$ is the *minimum mean square error forecast among all linear unbias forecast*

3. The variance of the forecast error $y_k - \hat{y}_k^{\text{pred}}$ is given by
$$\mathrm{Var}(y_k - \hat{y}_k^{\text{pred}}) = \mathrm{Var}\{\varepsilon_k + \mathbf{x}_k'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\} = \sigma^2[1 + \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k]$$

4. The estimated variance of the forecast error is
$$\widehat{\mathrm{Var}}(y_k - \hat{y}_k^{\text{pred}}) = s^2[1 + \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k]$$

   A $100(1-\alpha)$ percent prediction interval for the future y_k is then given by its upper and lower limits
$$\hat{y}_k^{\text{pred}} \pm t_{\alpha/2}(n-p-1)s[1 + \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k]^{\frac{1}{2}}$$

Example: Running Performance

- The simple linear regression model

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

  where y_t is the time to run the 10-km race, and x_t is the maximal aerobic capacity V_{O_2}.

- The least square estimate of $\hat{\beta}_0 = 68.494$ and $\beta_1 = -0.468$.

- The fitting values $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$, the residuals $e_t = y_t - \hat{y}_t$, and the entries in ANOVA table are easily calculated.

Source	SS	df	MS	F
Regression	48.845	1	48.845	9.249
Error	63.374	12	5.281	
Total	112.219	13		

- The coefficient of determination is given by $R^2 = 48.845/112.219 = .435$.

- $$\begin{aligned}\widehat{\mathrm{Var}}(\hat{\beta}) &= s^2(\mathbf{X}'\mathbf{X})^{-1} = s^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_t-\bar{x})^2} & \frac{-\bar{x}}{\sum(x_t-\bar{x})^2} \\ \frac{-\bar{x}}{\sum(x_t-\bar{x})^2} & \frac{1}{\sum(x_t-\bar{x})^2} \end{bmatrix} \\ &= \begin{bmatrix} 66.8511 & -1.2544 \\ -1.2544 & 0.0237 \end{bmatrix}\end{aligned}$$

- The standard error $s_{\hat{\beta}_i}$ and t statistics $t_{\hat{\beta}_i} = \hat{\beta}_i / s_{\hat{\beta}_i}$ are given in the following table

	$\hat{\beta}_i$	$s_{\hat{\beta}_i}$	$t_{\hat{\beta}_i}$
Constant	68.494	8.176	8.38
V_{O_2}	-0.468	0.154	-3.04

- $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 < 0$
  The critical value for this one-sided test with significance level $\alpha = 0.05$ is $-t(12) = -1.78$.
  Since the t statistic is smaller than this critical value, we reject the null hypothesis. In the other words, there is evidence for a significance inverse relationship between running time and aerobic capacity.

- Prediction intervals
  For example, let us predict the running time for a female athlete with maximal aerobic capacity $x_k = 55$. Then

$$\hat{y}^{\text{pred}} = \hat{\beta}_0 + \hat{\beta}_1 x_k = 42.75$$

  and its standard error by

$$s\left(1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_t - \bar{x})^2}\right)^{1/2} = 2.40$$

  Thus a 95 percent prediction interval for the running time is given by

$$42.75 \pm t_{0.025}(12)2.40 = 42.75 \pm 5.23 \quad \text{or} \quad (37.52, 47.98)$$

To investigate whether other variable (height X_2, weight X_3, skinfold sum X_4, relative body fat X_5 help explain the variation in the data, fit the model

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_5 x_{t5} + \varepsilon_t$$

Then we have

	$\hat{\beta}_i$	$s_{\hat{\beta}_i}$	$t_{\hat{\beta}_i}$
Constant	80.921	32.556	2.49
V_{O_2}	-0.479	0.203	-2.36
Height	-0.064	0.262	-0.24
Weight	-0.085	0.285	-0.30
Skinfold sum	0.027	0.079	0.35
Relative body fat	0.047	0.292	0.16

and the ANOVA table

Source	SS	df	MS	F
Regression	57.989	5	11.598	1.71
Error	54.230	8	6.779	
Total	112.219	13		

Case Study I: Gas Mileage Data Consider predicting the gas mileage of an automobile as a function of its size and other engine characteristics. In table 2.2. we have list data on 38 automobile (1978-1979 models). These data, which were originally taken from Consumer Reports. The variables include gas mileage in miles per gallon (MPG), number of cylinders, cubic engine displacement, horse power, weight, acceleration, and engine type [straight(1), V(0)].

The gas consumption (in gallons) should be proportional to the effort(=force $\times$ distance) it takes to move the car. Furthermore, since force is proportional to weight, we except that gas consumption per unit distance is proportional to force and thus proportional to weight . Thus a model of the form

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

where $y = \text{MPG}^{-1} = GPM$ and x =weight.

To check whether a quadratic term $(\text{weight})^2$ is need, then fit the model

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \varepsilon_t$$

and make a hypothesis test

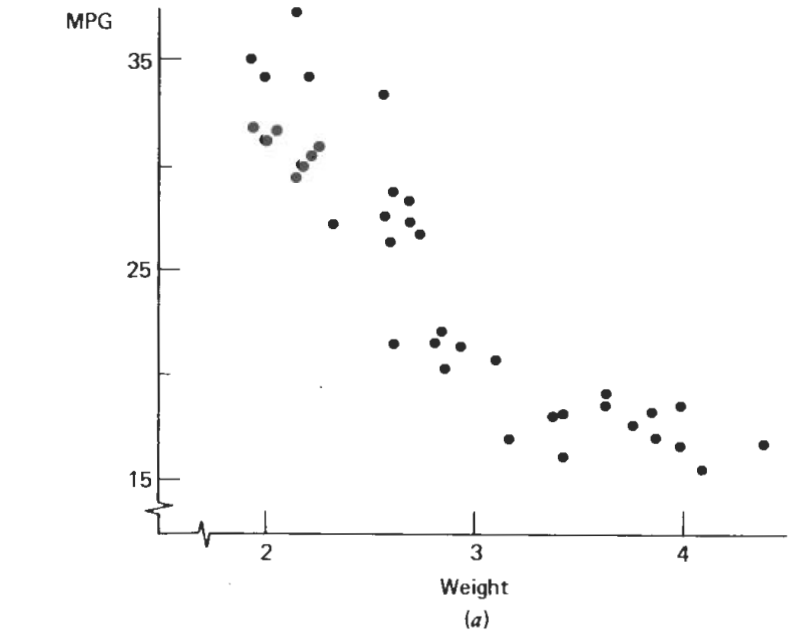
$$H_0 : \beta_2 = 0 \quad \text{againsit} \quad H_1 : \beta_2 \neq 0$$

Table 2.2. Gas Mileage Data on 38 Automobiles (1978–1979 Models)[a]

MPG	X_1	X_2	X_3	X_4	X_5	X_6
16.9	8	350	155	4.360	14.9	1
15.5	8	351	142	4.054	14.3	1
19.2	8	267	125	3.605	15.0	1
18.5	8	360	150	3.940	13.0	1
30.0	4	98	68	2.155	16.5	0
27.5	4	134	95	2.560	14.2	0
27.2	4	119	97	2.300	14.7	0
30.9	4	105	75	2.230	14.5	0
20.3	5	131	103	2.830	15.9	0
17.0	6	163	125	3.140	13.6	0
21.6	4	121	115	2.795	15.7	0
16.2	6	163	133	3.410	15.8	0
20.6	6	231	105	3.380	15.8	0
20.8	6	200	85	3.070	16.7	0
18.6	6	225	110	3.620	18.7	0
18.1	6	258	120	3.410	15.1	0
17.0	8	305	130	3.840	15.4	1
17.6	8	302	129	3.725	13.4	1
16.5	8	351	138	3.955	13.2	1
18.2	8	318	135	3.830	15.2	1
26.5	4	140	88	2.585	14.4	0
21.9	6	171	109	2.910	16.6	1
34.1	4	86	65	1.975	15.2	0
35.1	4	98	80	1.915	14.4	0
27.4	4	121	80	2.670	15.0	0
31.5	4	89	71	1.990	14.9	0
29.5	4	98	68	2.135	16.6	0
28.4	4	151	90	2.670	16.0	0
28.8	6	173	115	2.595	11.3	1
26.8	6	173	115	2.700	12.9	1
33.5	4	151	90	2.556	13.2	0
34.2	4	105	70	2.200	13.2	0
31.8	4	85	65	2.020	19.2	0
37.3	4	91	69	2.130	14.7	0
30.5	4	97	78	2.190	14.1	0
22.0	6	146	97	2.815	14.5	0
21.5	4	121	110	2.600	12.8	0
31.9	4	89	71	1.925	14.0	0

[a]MPG, miles per gallon; X_1, number of cylinders; X_2, engine displacement in cubic inches; X_3, horsepower; X_4, weight in 1000 lb; X_5, acceleration in sec; X_6, engine type [straight(1), V(0)].

Source: Henderson and Velleman (1981).



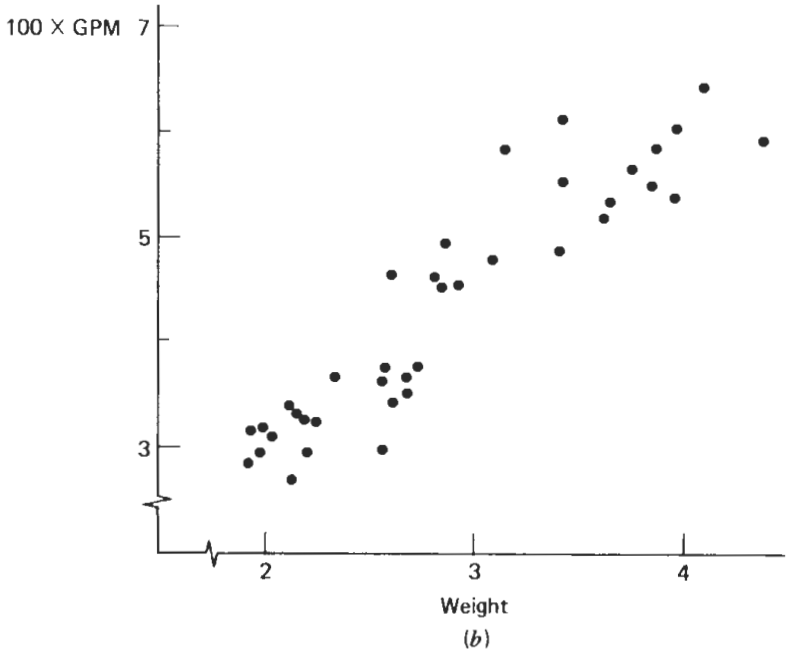


Figure 2.3. Scatter plots for the gas mileage example. (*a*) Plot of MPG against weight. (*b*) Plot of GPM = MPG^{-1} against weight.

2.7 Model Selection Techniques

(I) Adjusted coefficient of fitness

$$R_a^2 = 1 - \frac{\mathsf{SSE}/(n-p-1)}{\mathsf{SSTO}/(n-1)} = 1 - \frac{s^2}{\mathsf{SSTO}/(n-1)}.$$

Compared with R^2, we have extra penalty here for introducing more independent variables. Intention: maximize R_a^2 (equivalently minimizing s^2).

(II) Consider

$$C_p = \frac{\mathsf{SSE}_p}{s^2} - (n-2p).$$

Since $\mathrm{E}(\mathsf{SSE}_p) \approx (n-p)\sigma^2$, (large n), $C_p \sim p$. So choose smallest p such that $C_p \sim p$.

(III) Backward Elimination, Forward Selection, and Stepwise Regression.

Example 2 (Case study I continued)

Results from Fitting All possible Regression

p	Variables Included						R_a^2
1				X_4			.8540
2		X_2		X_4			.8866
3	X_1		X_3			X_6	.9071
4	X_1		X_3		X_5	X_6	.9231
5	X_1	X_2	X_3	X_4		X_6	.9235
6	X_1	X_2	X_3	X_4	X_5	X_6	.9267

Correlations Among the Predictor Variables

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.00					
X_2	.94	1.00				
X_3	.86	.87	1.00			
X_4	.92	.95	.92	1.00		
X_5	-.13	-.14	-.25	-.03	1.00	
X_6	.83	.77	.72	.67	-.31	1.00

t Statistics from Forward Selection and Backward Elimination Procedures

Step	Constant	X_1	X_2	X_3	X_4	X_5	X_6	R_a^2
			(a) Forward Selection					
1	-.02				14.75			.8540
2	-2.78		-3.377		8.39			.8866
3	-3.37	1.88	-3.95		8.08			.8943
			(b) Backward Elimination					
1	-3.92	3.62	-1.82	3.50	1.74	1.56	-3.60	.9267
2	-4.29	3.58	-2.54	3.11	2.97		-3.46	.9235

Selected Models

$$\text{GPM}_t = \beta_0 + \beta_1(\text{weight}_t) + \beta_2(\text{displacement}_t) + \varepsilon_t$$

or the even simpler model

$$\text{GPM}_t = \beta_0 + \beta_1(\text{weight}_t) + \varepsilon_t$$

2.8 Multi-colinearity in Predictor Variables

This is referring to the situation when the columns of $\mathbf{X}$ are almost linearly dependent. This sort of situation is common for observational data rather than artificially designed matrix. We say, in this case, that $\mathbf{X}'\mathbf{X}$ is ill conditioned. This could cause large variance of $\hat{\beta}_i$. So techniques like *Ridge Regression* will be applied, i.e. with restriction $\boldsymbol{\beta}'\boldsymbol{\beta} \leq r^2$.

Example

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t, \ t = 1, 2, \ldots, n.$$

The least square estimate of $\beta_i^* = \frac{s_i}{s_y}\beta_i, i = 1, 2$, where $s_i\sqrt{n-1} = [\sum(x_{it} - \bar{x}_i)^2]$, $s_y\sqrt{n-1} = [\sum(y_t - \bar{y})^2]$, has such property

$$\mathrm{Var}(\beta_1^*) = \mathrm{Var}(\beta_2^*) = \sigma_*^2 \frac{1}{1 - r_{12}^2}$$

and

$$\mathrm{Corr}(\beta_1^*, \beta_2^*) = -r_{12}$$

where $r_{12} = [\sum(x_{t1} - \bar{x}_1)(x_{t2} - \bar{x}_2)]/[\sum(x_{t1} - \bar{x}_1)^2 \sum(x_{t2} - \bar{x}_2)^2]^{\frac{1}{2}}$.

2.9 General Principles for Modelling

- Principle of Parsimony (As Simples as Possible)

$$\hat{\hat{\mathbf{y}}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \Longrightarrow \mathrm{Var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\hat{\mathbf{y}}_t) = \frac{\sigma^2}{n}\mathrm{Trace}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \frac{p\sigma^2}{n}.$$

Hence the average forecast error has variance

$$\sigma^2(1+\frac{p}{n}).$$

So extra independent variables would increase p and the forecast error.

- Plot Residuals
  Ideally, the residuals would be in the form of *white noise*, otherwise some modifications might be necessary, such as extra linear/quadratic terms. Particularly for the **funnel Shape** (like a trumpet), the logarithm transformation might be proper.

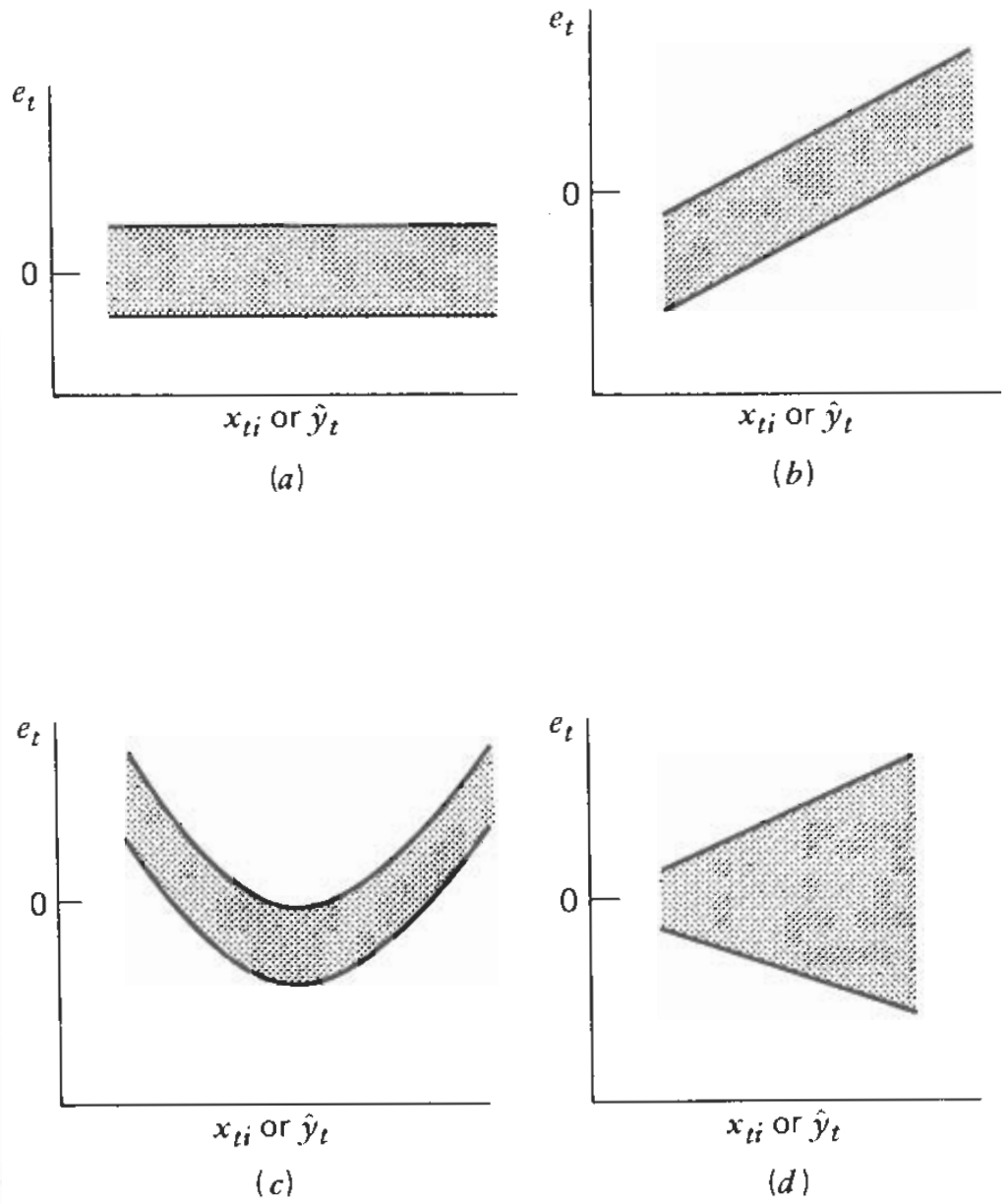


Figure 2.6. Various satisfactory and unsatisfactory residual plots. (*a*) Satisfactory residual plot. (*b*) Incorrect model form (a constant or a linear term should have been included). (*c*) Incorrect model form (a quadratic term should have been included). (*d*) Nonconstant variance.