Regression Models

Mahbub Latif, PhD

May 2025

Plan

- Comparison between two independent populations
  - Independent two-sample t-test
  - Simple linear regression model

Problem 33 (page 342)

- Twenty-five men between the ages of 25 and 30 were randomly selected to participate in Framingham Heart Study
- Of these, 11 were smokers, and 14 were not, and systolic blood pressure was measured from each of the 25 selected subjects.
- Use these data to test the hypothesis that the mean blood pressures of smokers and nonsmokers are the same.

Smokers	Non-smokers
124	130
134	122
136	128
125	129
133	118
127	122
135	116
131	127
133	135
125	120
118	122
	120
	115
	123

Problem 33 (page 342)

- Let μ_1 and μ_2 be the mean SBP for smokers and non-smokers, respectively, and we want to test

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2$$

- Assume SBPs follow normal distributions with a common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$, i.e.

$$Y_s \sim N(\mu_1, \sigma^2) \;\; and \;\; Y_{ns} \sim N(\mu_2, \sigma^2)$$

- The point estimates of the parameters

$$\hat{\mu}_s = \frac{1}{n_s}\sum_i Y_{s,i} \quad and \quad S_s^2 = \frac{1}{n_s - 1}\sum_i (Y_{s,i} - \hat{\mu}_s)^2$$

$$\hat{\mu}_{ns} = \frac{1}{n_{ns}}\sum_i Y_{ns,i} \quad and \quad S_{ns}^2 = \frac{1}{n_{ns} - 1}\sum_i (Y_{ns,i} - \hat{\mu}_{ns})^2$$

Problem 33 (page 342)

- Test statistic

$$T = \frac{\hat{\mu}_s - \hat{\mu}_{ns}}{S_p\sqrt{\frac{1}{n_s} + \frac{1}{n_{ns}}}} \sim t_{n_s+n_{ns}-2}$$

  - The pooled variance

$$S_p^2 = \frac{(n_s-1)S_s^2 + (n_{ns}-1)S_{ns}^2}{n_s + n_{ns} - 2}$$

- Reject H_0 at α level of significance if

$$|T| > t_{\alpha/2,\ n_s+n_{ns}-2}$$

- p-value of the test

$$p = 2P(T > |t|)$$

Estimates of model parameters

$$\hat{\mu}_s = \frac{1}{n_s}\sum_i Y_{s,i} = 123.36$$

$$\hat{\mu}_{ns} = \frac{1}{n_{ns}}\sum_i Y_{ns,i} = 129.18$$

$$S_s^2 = \frac{1}{n_s - 1}\sum_i (Y_{s,i} - \hat{\mu}_s)^2 = 5.73^2$$

$$S_{ns}^2 = \frac{1}{n_{ns} - 1}\sum_i (Y_{ns,i} - \hat{\mu}_{ns})^2 = 5.72^2$$

$$\begin{aligned} S_p^2 &= \frac{(n_s - 1)S_s^2 + (n_{ns} - 1)S_{ns}^2}{n_s + n_{ns} - 2} \\ &= \frac{(14 - 1)(5.73^2) + (11 - 1)(5.72^2)}{14 + 11 - 2} \\ &= 5.73^2 \end{aligned}$$

Problem 33 (page 342)

- Test statistic

$$T = \frac{\hat{\mu}_s - \hat{\mu}_{ns}}{S_p\sqrt{\frac{1}{n_s} + \frac{1}{n_{ns}}}} = \frac{123.36 - 129.18}{5.73\sqrt{\frac{1}{14} + \frac{1}{11}}} = -2.52$$

- Reject H_0 at 5% level of significance since $|T| = |-2.52| > t_{.025,23} = 2.07$
  - So, there is a significant difference between smokers and non-smokers in terms of mean SBP at a 5% level of significance
- p-value of the test

$$p = 2P(T > |-2.52|) = 0.019$$

Problem 33 (page 342)

- 95% confidence interval

$$
\begin{aligned}
(\hat{\mu}_s - \hat{\mu}_{ns}) \pm z_{.025}\, SE(\hat{\mu}_s - \hat{\mu}_{ns}) &= (\hat{\mu}_s - \hat{\mu}_{ns}) \pm t_{.025, n_s+n_{ns}-2}\; S_p \sqrt{1/n_s + 1/n_{ns}} \\
&= (123.36 - 129.18) \pm (2.07)(5.73) \\
&= -5.82 \pm 4.77 \\
&= (-10.59, -1.05)
\end{aligned}
$$

Representations of the SBP data

Smokers	Non-smokers
124	130
134	122
136	128
125	129
133	118
127	122
135	116
131	127
133	135
125	120
118	122
	120
	115
	123

id	sbp	smoke	x
1	124	yes	1
2	134	yes	1
3	136	yes	1
4	125	yes	1
5	133	yes	1
6	127	yes	1
7	135	yes	1
8	131	yes	1
9	133	yes	1
10	125	yes	1
11	118	yes	1
12	130	no	0
13	122	no	0

id	sbp	smoke	x
14	128	no	0
15	129	no	0
16	118	no	0
17	122	no	0
18	116	no	0
19	127	no	0
20	135	no	0
21	120	no	0
22	122	no	0
23	120	no	0
24	115	no	0
25	123	no	0

Linear regression models

- Define
  - $Y_i \rightarrow$ as the SBP measurement of the ith subject $(i = 1, \ldots, n)$ and

$$x_i = \begin{cases} 1 & \text{if the } ith \text{ subject is a smoker} \\ 0 & \text{otherwise} \end{cases}$$

  - $n \rightarrow$ total number of subjects (smokers + non-smokers)

- Mean SBP for two groups (smokers and non-smokers)

$$\begin{aligned} \mu_i = E(Y \mid x_i) &= \beta_0 + \beta_1 x_i \\ &= \begin{cases} \beta_0 & \text{if } x_i = 0 \text{ (for non-smokers)} \\ \beta_0 + \beta_1 & \text{if } x_i = 1 \text{ (for smokers)} \end{cases} \end{aligned}$$

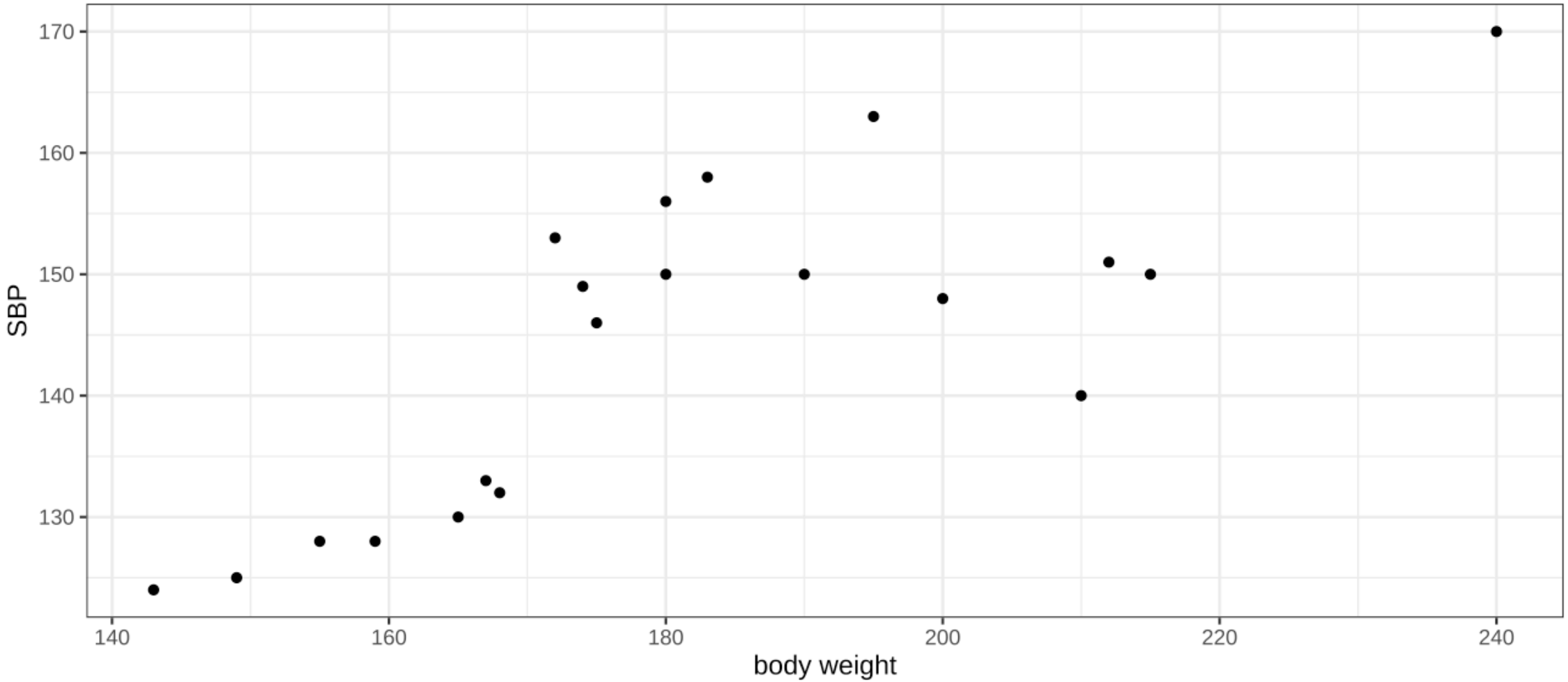
Exercise 31 (page 427)

- The data show body weight (in pounds) and systolic blood pressure (SBP) of 20 randomly selected males aged 25 to 30.
- The objective is to examine the effect of body weight on mean SBP

id	wt	sbp
1	165	130
2	167	133
3	180	150
4	155	128
5	212	151
6	175	146
7	190	150
8	210	140
9	200	148
10	149	125

id	wt	sbp
11	172	153
12	159	128
13	168	132
14	174	149
15	183	158
16	215	150
17	195	163
18	180	156
19	143	124
20	240	170

Scatter plot of body weight and SBP



Linear regression model

- Let Y_i be SBP measurement of the ith subject and x_i is the corresponding body weight $(i = 1, \ldots, n)$
- The regression model corresponding to the mean SBP of the ith subject

$$\mu_i = E(Y \mid x_i) = \beta_0 + \beta_1 x_i$$

  - $\beta_0 \rightarrow$ mean SBP of a subject with weight is zero, i.e., $x_i = 0$
  - $\beta_1 \rightarrow$ change in mean SBP for a 1-unit increase in body weight, i.e.,

$$\beta_1 = E(Y \mid x_i = x + 1) - E(Y \mid x_i = x)$$

Linear regression models

- Regression models study the effect of one variable (independent variable) on another variable (dependent variable)
    - Smoking status is an independent variable (also known as predictor or explanatory variable)
    - SBP is a dependent variable (also known as response or outcome)
- More than one independent variable can also be considered in a regression model
    - Y is used to denote the dependent variable
    - $X_1, \ldots, X_p$ are used to denote independent variables

Linear regression models

- Data: n independent pair of observations

$$\{(y_i, x_i), i = 1, \ldots, n\}$$

  - Subjects are different because of their values of independent variables

- Mean function

$$\mu_i = E(Y \mid x_i) = \beta_0 + \beta_1 x_i$$

  - $\beta_0 \rightarrow$ intercept
  - $\beta_1 \rightarrow$ slope

- Variance function

$$\sigma_i^2 = Var(Y \mid x_i) = \sigma^2$$

  - Constant variance, i.e., variance does not depend on independent variable

Linear regression models

- The mean and variance function of a response Y of a subject with the predictor x

$$\mu_i = E(Y \mid x_i) = \beta_0 + \beta_1 x_i \;\; and \;\; V(Y \mid x_i) = \sigma^2$$

- The regression model can also be expressed as

$$Y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- The random error term ϵ has the following properties
  - ϵ's are independent
  - $E(\epsilon \mid x_i) = 0$ and $Var(\epsilon \mid x_i) = \sigma^2$
  - $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \;\; \Rightarrow \;\; Y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$

Estimation of regression model parameters

Estimation of regression model parameters

- There are two methods of estimating parameters $(\beta_0, \beta_1, \sigma^2)$ of the linear regression model
  - Least square method of estimation and maximum likelihood method of estimation
- Least square estimates of model parameters correspond to the smallest error sum of squares

$$ESS = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- Maximum likelihood estimates correspond to the maximum of the likelihood function

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(Y_i - \beta_0 - \beta_1 x_i)^2}$$

Least square method of estimation

- Least square estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, correspond to the minimum of the error sum of squares

$$(\hat{\beta}_0, \hat{\beta}_0)' = \underset{(\beta_0,\beta_1)'\in\Theta}{\arg\min}\, ESS = \underset{(\beta_0,\beta_1)'\in\Theta}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_i\right)^2$$

- The estimates are the solutions of the following score equations

$$\left.\frac{\partial\, ESS}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0,\beta_1=\hat{\beta}_1} = 0$$

$$\left.\frac{\partial\, ESS}{\partial \beta_1}\right|_{\beta_0=\hat{\beta}_0,\beta_1=\hat{\beta}_1} = 0$$

Least square method of estimation

- Error sum of squares

$$ESS = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_i\right)^2$$

- The partial derivatives

$$\frac{\partial\, ESS}{\partial \beta_0} = 2 \sum_{i} \left(Y_i - \beta_0 - \beta_1 x_i\right)(-1)$$

$$\frac{\partial\, ESS}{\partial \beta_1} = 2 \sum_{i} \left(Y_i - \beta_0 - \beta_1 x_i\right)(-x_i)$$

Least square method of estimation

- The estimate of β_0

$$\left.\frac{\partial ESS}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0,\beta_1=\hat{\beta}_1} = 0$$

$$2\sum_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1) = 0 \;\Rightarrow\; \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \bar{x}$$

The estimate of β_1

$$\left.\frac{\partial\, ESS}{\partial \beta_1}\right|_{\beta_0=\hat{\beta}_0,\beta_1=\hat{\beta}_1} = 0$$

$$2\sum_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-x_i) = 0$$

$$\sum_i \left[(Y_i - \overline{Y})x_i - \hat{\beta}_1(x_i - \bar{x})x_i\right] = 0$$

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_i (Y_i - \overline{Y})x_i}{\sum_i (x_i - \bar{x})x_i} \\ &= \frac{\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{\sum_i Y_i^2 - n\overline{Y}\,\bar{x}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}\end{aligned}$$

Linear regression models

- The model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- The estimates of the parameters

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \bar{x} \;\; and \;\; \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

- The fitted model

$$\widehat{Y} = \hat{\mu} = \widehat{E}(Y \mid x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Regression model for estimating the effect of smoking on SBP

id	sbp	smoke	x
1	124	yes	1
2	134	yes	1
3	136	yes	1
4	125	yes	1
5	133	yes	1
6	127	yes	1
7	135	yes	1
8	131	yes	1
9	133	yes	1
10	125	yes	1
11	118	yes	1
12	130	no	0
13	122	no	0

id	sbp	smoke	x
14	128	no	0
15	129	no	0
16	118	no	0
17	122	no	0
18	116	no	0
19	127	no	0
20	135	no	0
21	120	no	0
22	122	no	0
23	120	no	0
24	115	no	0
25	123	no	0

Regression model for estimating the effect of smoking on SBP

n	sx	sy	sxx	sxy	xbar	ybar
25	11	3148	11	1421	0.44	125.92

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \\ &= \frac{1421 - (25)(0.44)(125.92)}{11 - (25)(0.44^2)} \\ &= 5.825\end{aligned}$$

Regression model for estimating the effect of smoking on SBP

n	sx	sy	sxx	sxy	xbar	ybar
25	11	3148	11	1421	0.44	125.92

$$\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$= \frac{1421 - (25)(0.44)(125.92)}{11 - (25)(0.44^2)}$$

$$= 5.825$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\bar{x}$$

$$= 125.92 - (5.825)(0.44)$$

$$= 123.357$$

Regression model for estimating the effect of smoking on SBP

- The fitted model for SBP on smoking status (x)

$$\widehat{Y} = \widehat{E}(Y \mid x) = 123.357 + 5.825\, x$$

  - Mean SBP for non-smokers

$$\widehat{E}(Y \mid x = 0) = \hat{\beta}_0 = 123.357$$

  - Mean SBP for smokers

$$\widehat{E}(Y \mid x = 1) = \hat{\beta}_0 + \hat{\beta}_1 = 129.182$$

Regression model for estimating the effect of body weight on SBP

id	wt	sbp
1	165	130
2	167	133
3	180	150
4	155	128
5	212	151
6	175	146
7	190	150
8	210	140
9	200	148
10	149	125

id	wt	sbp
11	172	153
12	159	128
13	168	132
14	174	149
15	183	158
16	215	150
17	195	163
18	180	156
19	143	124
20	240	170

Regression model for estimating the effect of body weight on SBP

n	sx	sy	sxx	sxy	xbar	ybar
20	3632	2884	671062	528519	181.6	144.2

$$\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$= \frac{528519 - (20)(181.6)(144.2)}{671062 - (20)(181.6^2)}$$

$$= 0.416$$

Regression model for estimating the effect of body weight on SBP

n	sx	sy	sxx	sxy	xbar	ybar
20	3632	2884	671062	528519	181.6	144.2

$$\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$= \frac{528519 - (20)(181.6)(144.2)}{671062 - (20)(181.6^2)}$$

$$= 0.416$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\bar{x}$$

$$= 144.2 - (0.416)(181.6)$$

$$= 68.584$$

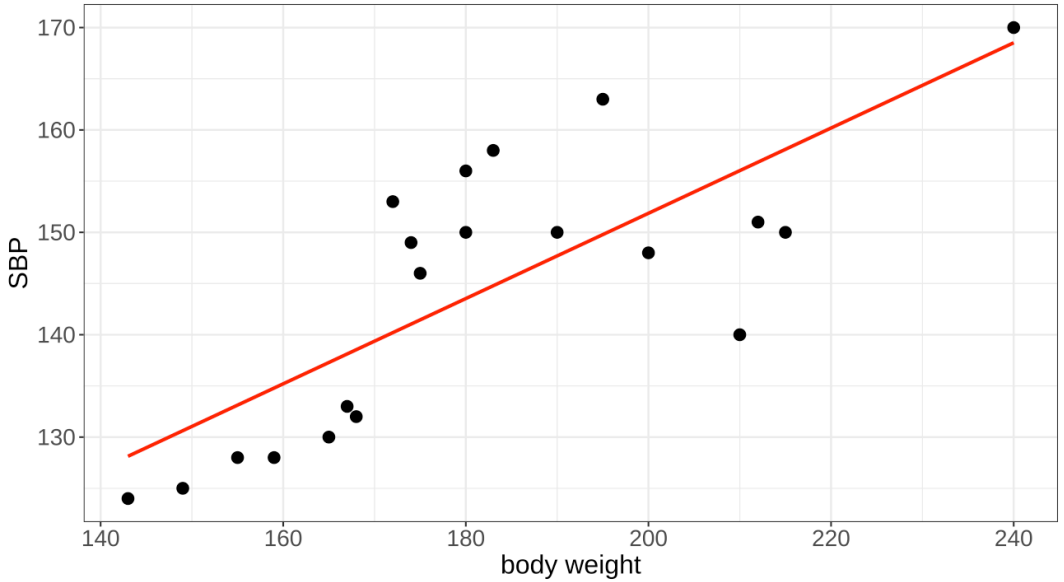
-

Regression model for estimating the effect of body weight on SBP

- The fitted model for SBP on body weight (x)

$$\widehat{Y} = \widehat{E}(Y \mid x) = 68.584 + 0.416\, x$$

  - For 1-pound increase in body weight, mean SBP increases by 0.416 unit



- Estimated SBP for a subject with a body weight of 225 pounds

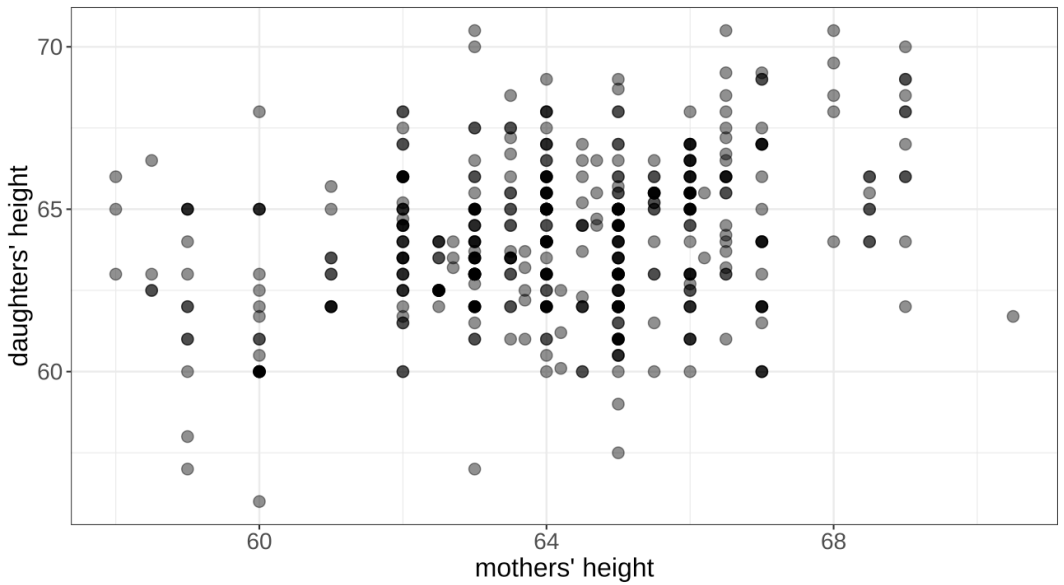
$$\widehat{E}(Y \mid x = 225) = 162.3$$

Mothers' height and daughters' height

Effect of mothers' height on her daughter's height

- A sample of 8 observations of Galton data, which has 934 observations in total

Mother Ht	Daughter Ht
66.5	67.5
64.5	71.0
62.0	70.0
69.0	73.0
63.0	66.0
66.5	71.0
62.0	66.0
67.0	73.2

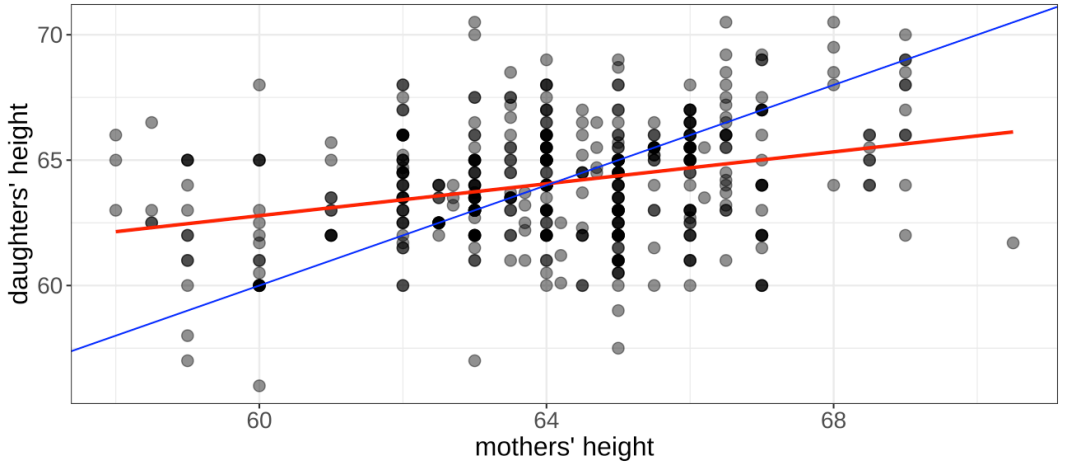


Effect of mothers' height on her daughter's height

- The fitted model for Daughters' height on Mothers' height (x)

$$\widehat{Y} = \widehat{E}(Y \mid x) = 46.587 + 0.315\, x$$

  - For a 1-inch increase in mothers' height, mean daughters' height increases by 0.315 inch



- Estimate average daughters' height for a mother of five feet and 4 inches tall

Sampling distributions of regression estimators

Sampling distributions of regression estimators

- Response of the ith subject Y_i follows a normal distribution

$$Y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

- The estimator of β_1 can be expressed of a linear combination of Y's

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(Y_i - \overline{Y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x}) Y_i}{S_{xx}} = \sum_i w_i Y_i$$

  - $w_i = \frac{(x_i - \bar{x})}{S_{xx}} \quad \Rightarrow \quad \sum_i w_i = 0$

Sampling distribution of estimators

- Response of the ith subject Y_i follows a normal distribution

$$Y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

- Since $\hat{\beta}_1$ is a linear combination of normally distributed random variables Y's, so $\hat{\beta}_1$ follows a normal distribution
  - The mean and variance of $\hat{\beta}_1$ are

$$E(\hat{\beta}_1) = E\Big[\sum_i w_i Y_i\Big] = \sum_i w_i E(Y_i) = \sum_i w_i \left(\beta_0 + \beta_1 x_i\right) = \beta_1$$

$$Var(\hat{\beta}_1) = Var\Big[\sum_i w_i Y_i\Big] = \sum_i w_i^2 Var(Y_i) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{Sxx}$$

Sampling distribution of estimators

- It can be shown that
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ is an unbiased estimator of β_0, i.e. $E(\hat{\beta}_0) = \beta_0$
  - The sampling distribution of $\hat{\beta}_0$ also follow a normal distribution with

$$
\begin{aligned}
E(\hat{\beta}_0) &= \beta_0 \\
Var(\hat{\beta}_0) &= Var\left(\bar{Y} - \hat{\beta}_1 \bar{x}\right) = Var \sum_i \left(\frac{1}{n} - w_i \bar{x}\right) Y_i \\
&= \sigma^2 \sum_i \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)^2 = \sigma^2 \sum_i \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) = \frac{\sigma^2 \sum_i x_i^2}{n S_{xx}}
\end{aligned}
$$

Estimation of error variance

- Linear regression model

$$Y_i = E(Y \mid x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Random error term

$$\epsilon_i = Y_i - E(Y \mid x_i) = Y_i - \beta_0 - \beta_1 x_i$$

- Fitted model

$$\widehat{Y}_i = \widehat{E}(Y \mid x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Residual $(\hat{\epsilon})$ is an estimate of random error (ϵ) and it should behave as a random sample from $N(0, \sigma^2)$ if model assumptions are satisfied

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Estimation of error variance

- It can be shown that

$$\begin{aligned}
\sum_i (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 \\
&= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \overline{Y})^2 \\
&= \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 + \sum_i (\hat{Y}_i - \overline{Y})^2 \\
SS_T &= SS_R + SS_M \\
\mathrm{df}: (n-1) &= (n-2) + 1
\end{aligned}$$

Estimation of error variance

$$
\begin{aligned}
SS_R &= \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_i (Y_i - \overline{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)^2 \\
&= \sum_i (Y_i - \overline{Y})^2 + \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_i (Y_i - \overline{Y})(x_i - \bar{x}) \\
&= S_{yy} + \frac{S_{xy}^2}{S_{xx}} - \frac{2S_{xy}^2}{S_{xx}} \\
&= \frac{S_{yy}S_{xx} - S_{xy}^2}{S_{xx}}
\end{aligned}
$$

Estimation of error variance

- Residual sum of squares (ResSS)

$$SS_R = \sum_i \hat{\epsilon}^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- It can be shown that

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - 2 \;\Rightarrow\; \sigma^2 = \frac{E\left[SS_R\right]}{n-2}$$

- The estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i \hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = MS_R$$

Estimation of error variance

- It can be shown that

$$\frac{SS_R}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

- We have already shown

$$\boxed{\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma^2/S_{xx}\right) \;\; and \;\; \hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \sum_i x_i^2/nS_{xx}\right)}$$

Effect of weight on SBP

SBP	Weight	fitted	residual
130	165	137.288	−7.288
133	167	138.121	−5.121
150	180	143.534	6.466
128	155	133.124	−5.124
151	212	156.858	−5.858
146	175	141.452	4.548
150	190	147.698	2.302
140	210	156.025	−16.025
148	200	151.861	−3.861
125	149	130.626	−5.626

- Fitted model

$$\hat{\mu} = 68.584 + 0.416\, Weight$$

- Residual sum of squares

$$SS_R = \sum \hat{\epsilon}_i^2 = 1416.963$$

- Error df

$$n - 2 = 18$$

- Estimate of error variance

$$\hat{\sigma}^2 = 78.72$$

Example 9.3a

- The data relate the moisture of a wet mix of a certain product (x) to the density of the finished product (Y)
- Consider a simple linear regression model for Y on x and obtain the estimates of regression parameters.

x	y
5	7.4
6	9.3
7	10.6
10	15.4
12	18.1
15	22.2
18	24.1
20	24.8

Statistical inference about the regression parameters

Statistical inference about β_1

- The null hypothesis $H_0 : \beta_1 = 0$
- The test statistic

$$T = \frac{\hat{\beta}_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

$$= \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$$

- Reject H_0 at α level of significance if

$$|T| > t_{n-2,\alpha/2}$$

- The p-value of the test

$$p = 2P(T > |t|)$$

- 95% confidence interval of β_1

$$\hat{\beta}_1 \pm se(\hat{\beta}_1)\, t_{n-2,\alpha/2}$$

$$= \hat{\beta}_1 \pm (\hat{\sigma}/\sqrt{S_{xx}})\, t_{n-2,\alpha/2}$$

Effect of Weight on SBP

- Fitted model

$$\hat{\mu} = 68.584 + 0.416\, Weight$$

- SE of $\hat{\beta}_1$

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{78.72}{11490.8}} = 0.083$$

- Test statistic for $H_0 : \beta_1 = 0$

$$T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0.416}{0.082769} = 5.031$$

- 95% CI for β_1

$$\begin{aligned}
&\hat{\beta}_1 \pm t_{18,.975}\, se(\hat{\beta}_1) \\
&= 0.416 \pm 2.101(0.083) \\
&= (0.242, 0.59)
\end{aligned}$$

Statistical inference about $\mu_{x_0} = \beta_0 + \beta_1 x_0$

- The estimator $\hat{\mu}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ and $E(\hat{\mu}_{x_0}) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$

$$
\begin{aligned}
Var(\hat{\mu}_{x_0}) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= Var(\hat{\beta}_0) + x_0^2 Var(\hat{\beta}_1) + 2x_0 Cov(\hat{\beta}_0, \hat{\beta}_1) \\
&= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] + x_0^2 \frac{\sigma^2}{S_{xx}} - 2x_0 \frac{\bar{x}\sigma^2}{S_{xx}} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right]
\end{aligned}
$$

Statistical inference about $\mu_{x_0} = \beta_0 + \beta_1 x_0$

- The estimator $\hat{\mu}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ follows a normal distribution with mean $E(\hat{\mu}_{x_0})$ and variance $Var(\hat{\mu}_{x_0})$
- The $(1-\alpha)100\%$ confidence interval of $\mu_{x_0} = \beta_0 + \beta_1 x_0$

$$\hat{\mu}_{x_0} \pm t_{n-2,\alpha/2}\, se(\hat{\mu}_{x_0}) = (\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right]}$$

Coeefficient of determination

- Total sum of squares $S_{yy} = \sum_i (Y_i - \bar{Y})^2$
- Residual sum of squares $SS_R = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- The term $S_{yy} - SS_R$ is known as regression sum of squares
- The coefficient of determination is defined as

$$R^2 = \frac{S_{yy} - SS_R}{S_{yy}} = 1 - \frac{SS_R}{S_{yy}}$$

  - R^2 takes a value between 0 and 1
  - $100R^2 \rightarrow$ proportion of total variation in the response explained by the model

Effect of Weight on SPB

- Sum of squares

$$S_{yy} = \sum (y_i - \bar{y})^2 = 3409.2 \;\; and \;\; SS_R = 1416.963$$

- Coefficient of determination

$$R^2 = \frac{S_{yy} - SS_R}{S_{yy}} = 1 - \frac{SS_R}{S_{yy}} = 1 - \frac{1416.963}{3409.2} = 0.584$$

Model diagnostics (Analysis of residuals)

- The random error term ϵ has the following properties
  - ϵ's are independent
  - $E(\epsilon \mid x_i) = 0$ and $Var(\epsilon \mid x_i) = \sigma^2$
  - $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad \Rightarrow \quad Y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$
- Different plots of residuals $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ are used to assess the model assumptions
  - A scatter plot of fitted values and residuals is used examine whether (i) Y's are independent, (ii) Y's have a connstant variance, and (iii) ϵ has a mean zero
  - A Q-Q plot is used to check whether residuals follow a normal distribution

Multiple Linear Regression Models

Simple linear regression model in matrix notation

- Data $\{(Y_i, x_i), i = 1, \ldots, n\}$

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

- In matrix notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \ and \ \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Simple linear regression model in matrix notation

- The model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Assumptions

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \; and \; V(\boldsymbol{\epsilon}) = I\sigma^2$$

  - I is an identity matrix
  - $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} \; and \; V(\boldsymbol{Y}) = I\sigma^2$

Simple linear regression model in matrix notation

- Error sum of squares

$$\begin{aligned} ESS &= \sum_i \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\beta - \beta'\boldsymbol{X}'\boldsymbol{Y} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta \end{aligned}$$

- Estimates of regression parameters

$$\begin{aligned} \frac{d\,ESS}{d\beta} = -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\beta \;&\Rightarrow\; \boldsymbol{X}'\boldsymbol{X}\hat{\beta} = \boldsymbol{X}'\boldsymbol{Y} \\ &\Rightarrow\; \hat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \end{aligned}$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \frac{1}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \begin{bmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{bmatrix}$$

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{nS_{xx}} \left[\sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i \right] \\
&= \frac{\bar{Y}}{S_{xx}} \sum x_i^2 - \frac{\bar{x}}{S_{xx}} \sum x_i Y_i \\
&= \frac{\bar{Y}}{S_{xx}} \left[\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2 \right] - \frac{\bar{x}}{S_{xx}} \left[\sum x_i Y_i - n\bar{x}\bar{Y} + n\bar{x}\bar{Y} \right] \\
&= \bar{Y} + \frac{n\bar{x}^2\bar{Y}}{S_{xx}} - \hat{\beta}_1\bar{x} - \frac{n\bar{x}^2\bar{Y}}{S_{xx}} \\
&= \bar{Y} - \hat{\beta}_1\bar{x}
\end{aligned}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \begin{bmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{bmatrix}$$

$$\begin{aligned} \hat{\beta}_1 &= \frac{1}{nS_{xx}} \Big[-\sum x_i \sum Y_i + n \sum x_i Y_i \Big] \\ &= \frac{1}{S_{xx}} \Big[-n\bar{x}\bar{Y} + \sum x_i Y_i \Big] \\ &= \frac{S_{xy}}{S_{xx}} \end{aligned}$$

Properties of $\hat{\beta}$

- $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- $E[\hat{\beta}] = E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta = \beta$
- $V[\hat{\beta}] = V\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

Multiple linear regression model

- A linear regression model with more than one predictors is known as multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

- In matrix notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \\ 1 & x_{1n} & x_{2n} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \ and \ \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Multiple linear regression model

- Estimate of regression parameters

$$\hat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- Expectation of estimated regression parameters

$$E[\hat{\beta}] = E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'E[\boldsymbol{Y}] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\beta = \beta$$

- Variance of estimated regression parameters

$$V[\hat{\beta}] = V\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'V\left(\boldsymbol{Y}\right)\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Multiple linear regression model

- Interpretations of model parameters

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$
$$E(Y \mid x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

  - $\beta_0 = E(Y \mid x_{1i} = 0, x_{2i} = 0)$ $\rightarrow$ mean response when both the predictors set at zero
  - $\beta_1 = E(Y \mid x_{1i} = x + 1, x_{2i} = c) - E(Y \mid x_{1i} = x, x_{2i} = c)$ $\rightarrow$ for one unit increase in x_1, the change in mean response when x_2 is fixed
  - $\beta_2 = E(Y \mid x_{1i} = c, x_{2i} = x + 1) - E(Y \mid x_{1i} = c, x_{2i} = x)$ $\rightarrow$ for one unit increase in x_2, the change in mean response when x_1 is fixed

Example 9.10a

- Estimate the effects of population size (x_1, in 1000) and divorce rate (x_2, per 100K) on suicide rate (y, per 100K) from a data set on 10 locations

y	x1	x2
11.6	679	30.4
16.1	1420	34.1
9.3	1349	17.2
9.1	296	26.8
8.4	6975	29.1
7.7	323	18.7
11.3	4200	32.6
8.4	633	32.5

Example 9.10a

- The model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

- Estimate of parameters

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 3.5074 \\ -0.0002 \\ 0.2609 \end{bmatrix}$$

term	estimate	std.error	statistic	p.value
(Intercept)	3.5074	4.3581	0.8048	0.457
x1	−0.0002	0.0004	−0.5775	0.589
x2	0.2609	0.1589	1.6427	0.161

Exercise (Pages 415-439)

- 1-6, 11, 12, 18, 19-23, 25, 26, 31, 43