STA 141A

Fall 2016

Lecture Note : Linear Regression

1 Simple linear regression (with one predictor)

1.1 Model assumptions and fitting procedure

- **Model :** X and Y are the *predictor* and *response* variables, respectively. Fit the model,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are uncorrelated, $E(\varepsilon_1) = 0$, $\mathrm{Var}(\varepsilon_1) = \sigma^2$. More generally, we can have $\mathrm{Vae}(\varepsilon_i) = \sigma_i^2$ where the value of σ_i^2 may depend on X_i. Here, we are treating X_i's as given, of fixed, rather than random. The source of randomness is through the randomness in $\{\varepsilon_i\}$. We may also write

$$\mathbb{E}(Y|X = X_i) = \beta_0 + \beta_1 X_i$$

and

$$\mathrm{Var}(Y|X = X_i) = \sigma_i^2.$$

- **Interpretation :** Look at the scatter plot of Y (vertical axis) versus X (horizontal axis). Consider narrow vertical strips around the different values of X:

  1. Means (measure of center) of the points falling in the vertical strips lie (approximately) on a straight line with **slope** β_1 and **intercept** β_0.
  2. Standard deviations (measure of spread) of the points falling in each vertical strip are (roughly) the same.

- **Estimation of β_0 and β_1 :** We employ the method of **least squares** to estimate β_0 and β_1. This means, we minimize the **sum of squared errors** : $Q(b_0, b_1) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$. This involves differentiating $Q(b_0, b_1)$ with respect to the *parameters* b_0 and b_1 and setting the derivatives to zero. This gives us the **normal equations**:

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^n X_i &= \sum_{i=1}^n Y_i & (2) \\ b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2 &= \sum_{i=1}^n X_i Y_i & (3) \end{aligned}$$

Solving these equations we have

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - n\overline{XY}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}, \qquad b_0 = \overline{Y} - b_1 \overline{X} \tag{4}$$

b_0 and b_1 are the *estimates* of β_0 and β_1, respectively, and are sometimes denoted as $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

- **Prediction :** The **fitted regression line** is given by the equation:

$$\widehat{Y} = b_0 + b_1 X \tag{5}$$

and is used to predict the value of Y given a value of X.

- **Residuals:** These are the quantities $e_i = Y_i - \widehat{Y}_i = Y_i - (b_0 + b_1 X_i)$, where $\widehat{Y}_i = b_0 + b_1 X_i$. Note that $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$. This means that e_i's estimate ε_i's. Some properties of the regression line and residuals are :

1. $\sum_i e_i = 0$.
2. $\sum_i e_i^2 \leq \sum_i (Y_i - u_0 - u_1 X_i)^2$ for any (u_0, u_1) (with equality when $(u_0, u_1) = (b_0, b_1)$).
3. $\sum_i Y_i = \sum_i \widehat{Y}_i$.
4. $\sum_i X_i e_i = 0$
5. $\sum_i \widehat{Y}_i e_i = 0$.
6. Regression line passes through the point $(\overline{X}, \overline{Y})$.
7. The slope b_1 of the regression line can be expressed as $b_1 = r_{XY} \frac{s_Y}{s_X}$, where r_{XY} is the correlation coefficient between X and Y and s_X and s_Y are the standard deviations of X and Y.

- **Error sum of squares**, denoted *SSE*, is given by

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - b_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- **Estimation of σ^2 :** It can be shown that $E(SSE) = (n-2)\sigma^2$. Therefore, σ^2 is estimated by the **mean squared error**, i.e., $MSE = \frac{SSE}{n-2}$. Note also that this justifies the statement that the **degree of freedom** of the errors is $n-2$ which is sample size (n) minus the number of regression coefficients (β_0 and β_1) being estimated.

1.2 Parameter estimation in simple linear regression

- **Estimates of the parameters :** We have the following estimates for β_0, β_1 and σ^2, respectively.

$$b_0 = \overline{Y} - b_1 \overline{X}, \quad b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}, \quad \widehat{\sigma}^2 = MSE = \frac{SSE}{n-2}, \tag{6}$$

where $SSE = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - b_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$.

- **Prediction :** The predicted valued of Y, given $X = X_h$ is $\widehat{Y}_h = b_0 + b_1 X_h = \overline{Y} + b_1 (X_h - \overline{X})$.

- **Expected values and variances :** Under the assumptions of the simple linear regression model, we have $E(b_0) = \beta_0$, $E(b_1) = \beta_1$ and $E(\widehat{\sigma}^2) = E(MSE) = \sigma^2$. In other words, the estimators b_0, b_1, $\widehat{\sigma}^2$ are *unbiased*. Also, $E(\widehat{Y}_h | X_h) = \beta_0 + \beta_1 X_h$.

  Assuming that $X_1, \ldots, X_n$ are *non-random*, the variances of b_0 and b_1 are given by:

$$\sigma^2(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_i (X_i - \overline{X})^2} \right], \quad \text{and} \quad \sigma^2(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}. \tag{7}$$

Replacing σ^2 by *MSE*, we obtain the estimates of the variances of b_0 and b_1, and these are denoted by

$$s^2(b_0) = MSE \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_i (X_i - \overline{X})^2}\right], \quad \text{and} \quad s^2(b_1) = \frac{MSE}{\sum_{i=1}^n (X_i - \overline{X})^2}, \tag{8}$$

respectively. Thus, $s(b_0)$ and $s(b_1)$ are the estimated **standard errors** of the estimators of β_0 and β_1, respectively.

Similarly, the variance and its estimate of $\widehat{Y}_h$ are

$$\sigma^2(\widehat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right], \qquad s^2(\widehat{Y}_h) = MSE \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right], \tag{9}$$

respectively.

1.3 Normal linear regression model

In model specified by (16), if the random variables $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$, then we have a *normal linear regression model*. This means that for each fixed value of X, the conditional distribution of Y given X is $N(\beta_0 + \beta_1 X, \sigma^2)$.

- **Maximum likelihood estimation :** Under this model, one can also obtain the estimates of b_0, b_1 and σ^2 by the method of **maximum likelihood**. This means that one treats the joint probability density function of $Y_1, \ldots, Y_n$ given $X_1, \ldots, X_n$

$$f(Y_1, \ldots, Y_n | X_1, \ldots, X_n; \beta_0, \beta_1, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2\right)$$

as a function, say $L(\beta_0, \beta_1, \sigma^2)$ of the parameters, and then maximizes this function w.r.t. the parameters by solving the equations:

$$\frac{\partial(\log L)}{\partial \beta_0} = 0, \qquad \frac{\partial(\log L)}{\partial \beta_1} = 0 \qquad \text{and} \qquad \frac{\partial(\log L)}{\partial \sigma^2} = 0 \tag{10}$$

to obtain the *maximum likelihood estimates* :

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = b_1, \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} = b_0 \\ \widehat{\sigma^2} &= \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = \frac{n-2}{n} MSE. \end{aligned} \tag{11}$$

- **Exact distribution :** Under the normality assumption, we can compute exact distribution of certain random variables that are very important for conducting tests of hypotheses for the different parameters. We have, SSE and (b_0, b_1) are independently distributed, and

$$SSE \sim \sigma^2 \chi^2_{n-2}, \qquad \frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2} \qquad \text{and} \qquad \frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}, \tag{12}$$

where χ^2_k and t_k denote the Chi-square and t-distribution, respectively, with k **degrees of freedom**.

1.4 Analysis of variance approach to regression

We divide the *total variability* in the observe data into two parts - one coming from the errors, the other coming from the predictor.

- **ANOVA decomposition :** The following decomposition

$$Y_i - \overline{Y} = (\widehat{Y}_i - \overline{Y}) + (Y_i - \widehat{Y}_i), \qquad i = 1, \ldots, n$$

represents the *deviation of observed response from mean response* in terms of the sum of the *deviation of fitted value from the mean* and the *residual*.

Taking sum of squares, and after some algebra we have

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2,$$

or

$$SSTO = SSR + SSE, \tag{13}$$

where $SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ and $SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$. (13) is referred to as the *ANOVA* decomposition to the variation in the response. Note that

$$SSR = b_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2.$$

- **Degrees of freedom :** The degrees of freedom of different terms in the decomposition (13) are

$$\text{d.f.}(SSTO) = n - 1, \qquad \text{d.f.}(SSR) = 1, \qquad \text{d.f.}(SSE) = n - 2.$$

So, d.f.$(SSTO)$ = d.f.(SSR)+ d.f.(SSE).

- **Expected value and distribution :** $E(SSE) = (n-2)\sigma^2$, and $E(SSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$. Also, under the *normal regression model*, and under $H_0 : \beta_1 = 0$,

$$SSR \sim \sigma^2 \chi_1^2, \qquad SSE \sim \sigma^2 \chi_{n-2}^2,$$

and these two are independent.

- **Mean squares :**

$$MSE = \frac{SSE}{\text{d.f.}(SSE)} = \frac{SSE}{n-2}, \qquad MSR = \frac{SSR}{\text{d.f.}(SSR)} = \frac{SSR}{1}.$$

Also, $E(MSE) = \sigma^2$, $E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$.

1.5 Descriptive measure of association between X and Y

Define the **coefficient of determination**:

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

Observe that $0 \leq R^2 \leq 1$, and the correlation coefficient, $\text{Corr}(X, Y)$ between X and Y is the (signed) square root of R^2. That is, $(\text{Corr}(X, Y))^2 = R^2$. Larger value of R^2 generally indicates higher degree of linear association between X and Y. Another (and considered better) measure of association is the **adjusted coefficient of determination** :

$$R^2_{ad} = 1 - \frac{MSE}{MSTO}.$$

R^2 is the proportion of variability in Y explained by its regression on X. Also, R^2 is unit free, i.e. does not depend on the units of measurements of the variables X and Y.

For the **housing price data**, $SSR = 352.91$, $SSTO = 556.08$, $n = 19$, and hence $SSE = 203.17$, d.f.$(SSE) = 17$, d.f.$(SSTO) = 18$. So, $R^2 = \frac{352.91}{556.08} = 0.635$ and $R^2_{ad} = 1 - \frac{11.95}{30.8} = 0.613$.

2 Inference in simple linear regression

- **Fact :** Under normal regression model (b_0, b_1) and SSE are independently distributed and

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}, \qquad \frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}, \qquad SSE \sim \sigma^2 \chi^2_{n-2}.$$

- **Confidence interval for β_0 and β_1:** $100(1-\alpha)\%$ (two-sided) confidence interval for β_i:

$$(b_i - t(1-\alpha/2; n-2)s(b_i), b_i + t(1-\alpha/2; n-2)s(b_i)),$$

  for $i = 0, 1$, where $t(1-\alpha/2; n-2)$ is the $1-\alpha/2$ *upper cut-off point* (or $(1-\alpha/2)$ quantile) of t_{n-2} distribution; i.e., $P(t_{n-2} > t(1-\alpha/2; n-2)) = \alpha/2$.

- **Hypothesis tests for β_0 and β_1 :** $H_0 : \beta_i = \beta_{i0}$ ($i = 0$ or 1).
  Test statistic : $T_i = \frac{b_i - \beta_{i0}}{s(b_i)}$.

  1. Alternative : $H_1 : \beta_i > \beta_{i0}$. Reject H_0 at level α if $\frac{b_i - \beta_{i0}}{s(b_i)} > t(1-\alpha; n-2)$. Or if, P-value $= P(t_{n-2} > T_i^{observed}) < \alpha$.
  2. Alternative : $H_1 : \beta_i < \beta_{i0}$. Reject H_0 at level α if $\frac{b_i - \beta_{i0}}{s(b_i)} < t(\alpha; n-2)$. Or if, P-value $= P(t_{n-2} < T_i^{observed}) < \alpha$.
  3. Alternative : $H_1 : \beta_i \neq \beta_{i0}$. Reject H_0 at level α if $|\frac{b_i - \beta_{i0}}{s(b_i)}| > t(1-\alpha/2; n-2)$. Or if, P-value $= P(|t_{n-2}| > |T_i^{observed}|) < \alpha$.

- **Inference for mean response at $X = X_h$:** Point estimate : $\widehat{Y}_h = b_0 + b_1 X_h$.
  Fact : $E(\widehat{Y}_h) = \beta_0 + \beta_1 X_h = E(Y_h)$, $\text{Var}(\widehat{Y}_h) = \sigma^2(\widehat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$. Estimated variance is

$$s^2(\widehat{Y}_h) = MSE \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right].$$

Distribution : $\frac{\widehat{Y}_h - E(Y_h)}{s(\widehat{Y}_h)} \sim t_{n-2}$.

Confidence interval : $100(1-\alpha)\%$ confidence interval for $E(Y_h)$ is

$$(\widehat{Y}_h - t(1-\alpha/2; n-2)s(\widehat{Y}_h), \widehat{Y}_h + t(1-\alpha/2; n-2)s(\widehat{Y}_h)).$$

- **Prediction of a new observation $Y_{h(new)}$ at $X = X_h$:** Prediction : $\widehat{Y}_{h(new)} = \widehat{Y}_h = b_0 + b_1 X_h$.

  Error in prediction : $Y_{h(new)} - \widehat{Y}_{h(new)} = Y_{h(new)} - \widehat{Y}_h$.

  Fact : $\sigma^2(Y_{h(new)} - \widehat{Y}_h) = \sigma^2(Y_{h(new)}) + \sigma^2(\widehat{Y}_h) = \sigma^2 + \sigma^2(\widehat{Y}_h) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$.

  Estimate of $\sigma^2(Y_{h(new)} - \widehat{Y}_h)$ is $s^2(Y_{h(new)} - \widehat{Y}_h) = MSE \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$.

  Distribution : $\frac{Y_{h(new)} - \widehat{Y}_h}{s(Y_{h(new)} - \widehat{Y}_h)} \sim t_{n-2}$.

  Prediction interval : $100(1-\alpha)\%$ prediction interval for $Y_{h(new)}$ is

$$(\widehat{Y}_h - t(1-\alpha/2; n-2)s(Y_{h(new)} - \widehat{Y}_h), \widehat{Y}_h + t(1-\alpha/2; n-2)s(Y_{h(new)} - \widehat{Y}_h)).$$

- **Confidence band for the regression line :** At $X = X_h$ the $100(1-\alpha)\%$ confidence band for the regression line is given by

$$\widehat{Y}_h \pm w_\alpha s(\widehat{Y}_h), \qquad \text{where } w_\alpha = \sqrt{2F(1-\alpha; 2, n-2)}.$$

  Here $F(1-\alpha; 2, n-2)$ is the $1-\alpha$ *upper cut-off point* (or, $(1-\alpha)$ quantile) for the $F_{2,n-2}$ distribution (F distribution with d.f. $(2, n-2)$).

2.1 Test of linear hypotheses

We are interested in testing for the dependence on the predictor variable from a different viewpoint. We call the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{14}$$

the **full model**. We want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Under $H_0 : \beta_1 = 0$, we have the **reduced model**:

$$Y_i = \beta_0 + \varepsilon_i. \tag{15}$$

Under the full model, $SSE_{full} = \sum_i (Y_i - \widehat{Y}_i)^2 = SSE$. Under the reduced model $SSE_{red} = \sum_i (Y_i - \overline{Y})^2 = SSTO$.

General structure of test statistic : Observe that d.f.$(SSE_{full}) = n-2$, d.f.$(SSE_{red}) = n-1$ and $SSE_{red} - SSE_{full} = SSR$.

$$F^* = \frac{\frac{SSE_{red} - SSE_{full}}{\text{d.f.}(SSE_{red}) - \text{d.f.}(SSE_{full})}}{\frac{SSE_{full}}{\text{d.f.}(SSE_{full})}} = \frac{\frac{SSR}{\text{d.f.}(SSR)}}{\frac{SSE}{\text{d.f.}(SSE)}} = \frac{MSR}{MSE}.$$

Under normal error model, and under $H_0 : \beta_1 = 0$, F^* has the F distribution with (paired) degrees of freedom (d.f.(SSE_{red})− d.f.(SSE_{full}), d.f.(SSE_{full})).

3 Simple linear regression : An example

Housing price data : Here Y = selling price of houses (in \$1000), and X = size of house (100 square feet). The summary statistics are given below:

$$n = 19, \quad \overline{X} = 15.719, \quad \overline{Y} = 75.211,$$

$$\sum_i (X_i - \overline{X})^2 = 40.805, \quad \sum_i (Y_i - \overline{Y})^2 = 556.078, \quad \sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = 120.001.$$

- **Estimates of β_1 and β_0 :**

$$b_1 = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} = \frac{120.001}{40.805} = 2.941,$$

and

$$b_0 = \overline{Y} - b_1 \overline{X} = 75.211 - (2.941)(15.719) = 28.981.$$

- **Fit and prediction :** The fitted regression line : $Y = 28.981 + 2.941X$. When $X = 18.5 = X_h$, the predicted value, that is an estimate of the mean selling price (in \$1000) when size of the house is 1850 sq. ft., is $\widehat{Y}_h = 28.981 + (2.941)(18.5) = 83.39$.

- **MSE :** The degrees of freedom (df) $= n - 2 = 17$. $SSE = \sum_i (Y_i - \overline{Y})^2 - b_1^2 \sum_i (X_i - \overline{X})^2 = 203.17$. So,

$$MSE = \frac{SSE}{n-2} = \frac{203.17}{17} = 11.95.$$

- **Standard error estimates :**

$$s^2(b_0) = MSE \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_i (X_i - \overline{X})^2} \right] = 73.00, \qquad s(b_0) = \sqrt{s^2(b_0)} = 8.544.$$

$$s^2(b_1) = \frac{MSE}{\sum_i (X_i - \overline{X})^2} = 0.2929, \qquad s(b_1) = \sqrt{s^2(b_1)} = 0.5412.$$

- **Confidence intervals :** We assume that the errors are normal to find confidence intervals for the parameters β_0 and β_1. We use the fact that $\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$ and $\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$ where t_{n-2} denotes the t-distribution with $n-2$ *degrees of freedom*. Since $t(0.975; 17) = 2.1098$, it follows that 95% two-sided confidence interval for β_1 is

$$2.941 \pm (2.1098)(0.5412) = (1.80, 4.08).$$

Since $t(0.95; 17) = 1.740$, the 90% two-sided confidence interval for β_0 is

$$28.981 \pm (1.740)(8.544) = (14.12, 43.84).$$

4 Regression diagnostics

Fitting the regression line is not a reaffirmation of its appropriateness for describing the relationship between X and Y. Indeed, there are several ways that the model assumptions could be violated. Also, the presence of some *extreme values* could heavily influence the results, even when a linear relationship may be valid. These aspects are usually determined by following various graphical diagnostic procedures.

4.1 Diagnostics for predictor

Diagnostic information about the predictor variable X, e.g. whether there are any outlying value, the range and concentration of X, are useful information that can provide clues to the appropriateness of the regression model assumptions.

- For moderate-sized data, we can use the **stem-and-leaf plot**, or the **dot plot** to gather information about the range and concentration of the data, as well as the possible extreme values of X. Similar information can be extracted from a summary plot like the **box plot**.
- If the data are observed over time, then both the predictor Y and the response X may show some pattern over time. A useful way of gathering this information is through the **sequence plot** (X values plotted against time).

To illustrate the effect of an extreme X value, consider the following example:

$$n = 16,\ \sum_{i=1}^{n-1} X_i = 90,\ \sum_{i=1}^{n-1} Y_i = 330,\ \sum_{i=1}^{n-1} X_i^2 = 1000,\ \sum_{i=1}^{n-1} X_i Y_i = 2400,\ X_n = 22,\ Y_n = 22.$$

Let $\overline{X}_{n-1} = \sum_{i=1}^{n-1} X_i/(n-1) = 6$, and $\overline{Y}_{n-1} = \sum_{i=1}^{n-1} Y_i/(n-1) = 22$. Also,

$$\sum_{i=1}^{n-1}(X_i - \overline{X}_{n-1})(Y_i - \overline{Y}_{n-1}) = \sum_{i=1}^{n-1} X_i Y_i - (n-1)\overline{X}_{n-1}\overline{Y}_{n-1} = 2400 - 15 \times 6 \times 22 = 420.$$

$$\sum_{i=1}^{n-1}(X_i - \overline{X}_{n-1})^2 = \sum_{i=1}^{n-1} X_i^2 - (n-1)(\overline{X})^2 = 1000 - 15 \times 6^2 = 460.$$

Hence, denoting by $b_1^{(n-1)}$ the least squares estimate of β_1 computed from the first $n-1$ observations, we have $b_1^{(n-1)} = \frac{\sum_{i=1}^{n-1}(X_i - \overline{X}_{n-1})(Y_i - \overline{Y}_{n-1})}{\sum_{i=1}^{n-1}(X_i - \overline{X}_{n-1})^2} = \frac{420}{460} = 0.913$. For the whole data set, $\overline{X} = (\sum_{i=1}^{n-1} X_i + X_n)/n = 7$. $\overline{Y} = (\sum_{i=1}^{n-1} Y_i + Y_n)/n = 22$.

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y} = (2400 + 22 \times 22) - 16 \times 7 \times 22 = 420.$$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2 = (1000 + 20^2) - 16 \times 7^2 = 616.$$

So, from the full data, estimate for β_1 is $b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{420}{616} = 0.6818$. Note that, in this example, standard deviation of X estimated from the first $n-1$ observations is $s_X^{(n-1)} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(X_i - \overline{X}_{n-1})^2} = 5.73$. And observe that $X_n > \overline{X}_{n-1} + 2s_X^{(n-1)}$.

It can be shown that if we use $b_0^{(n-1)}$ to denote the least squares estimate of β_0, from the first $n-1$ observations, and $e_n^{(n-1)} = Y_n - b_0^{(n-1)} - b_1^{(n-1)} X_n$, then

$$b_1 = b_1^{(n-1)} + \frac{(1 - \frac{1}{n})(X_n - \overline{X}_{n-1})e_n^{(n-1)}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = b_1^{(n-1)} + \frac{(X_n - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} e_n^{(n-1)}.$$

4.2 Diagnostics for residuals

Residuals $e_i = Y_i - \widehat{Y}_i$ convey information about the appropriateness of the model. In particular, possible departures from model assumptions are often reflected in the plot of residuals against either predictor(s) or fitted values, or in the distribution of the residuals.
Some important properties :

- **Mean :** We have seen that $\sum_i e_i = 0$ and hence $= \overline{e} = \frac{1}{n}\sum_i e_i = 0$.
- **Variance :** $\text{Var}(e) = s^2 = \frac{1}{n-2}\sum_i (e_i - \overline{e})^2 = \frac{1}{n-2}\sum_i e_i^2 = MSE$.
- **Correlations :** $\sum_i X_i e_i = 0$, $\sum_i \widehat{Y}_i e_i = 0$ and $\overline{e} = 0$ imply that $\text{Corr}(X, e) = 0$ and $\text{Corr}(\widehat{Y}, e) = 0$.
- **Nonindependence :** The residuals e_i are not independent even if the model errors ε_i are. This is because the e_i's satisfy two constraints: $\sum_i e_i = 0$ and $\sum_i X_i e_i = 0$. However, when n is large, the residuals are *almost independent* if the model assumptions hold.
- **Semi-studentized residuals :** Standardize the residuals by dividing through by $\sqrt{MSE}$ to get the semi-studentized residuals:

$$e_i^* = \frac{e_i - \overline{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}.$$

Model departures that can be studied by residual plots

1. The regression function is not linear.
2. The error terms do not have a constant variance.
3. The error terms are not independent.
4. The model fits all but one or a few outliers.
5. The error terms are not normally distributed.
6. One or several predictor variables have been omitted from the model.

4.3 Diagnostic plots

- **Nonlinearity of the regression function :** If the plot of residuals versus predictors show discernible, nonlinear pattern, that is an indication of possible nonlinearity of the regression function.

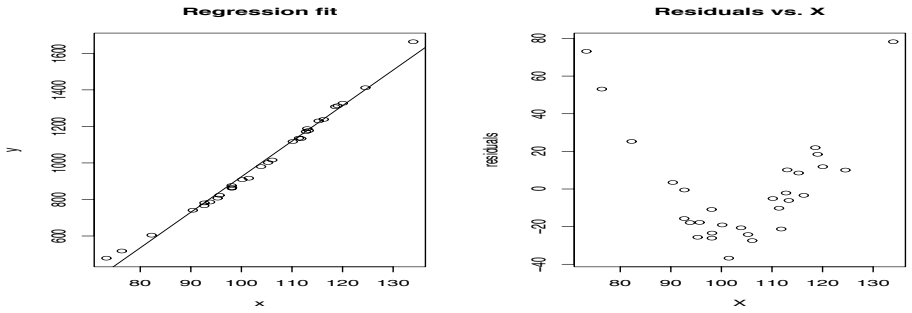
  Example : True model : $Y = 5 - X + 0.1 * X^2 + \varepsilon$ with $\varepsilon \sim N(0, (10)^2)$. We simulate 30 observations with X following a $N(100, (16)^2)$ distribution. The data summary is given below.

$$\overline{X} = 104.13, \overline{Y} = 1004.79, \sum_i X_i^2 = 330962.9, \sum_i Y_i^2 = 32466188, \sum_i X_i Y_i = 3249512.$$

  The linear model : $Y = \beta_0 + \beta_1 X + \epsilon$ was fitted to this data. The following table gives the summary.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	-1021.3803	40.0648	-25.49	$< 2 \times 10^{-16}$
Slope	19.4587	0.3814	51.01	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 28.78$, $R^2 = 0.9894$, $R^2_{ad} = 0.989$.



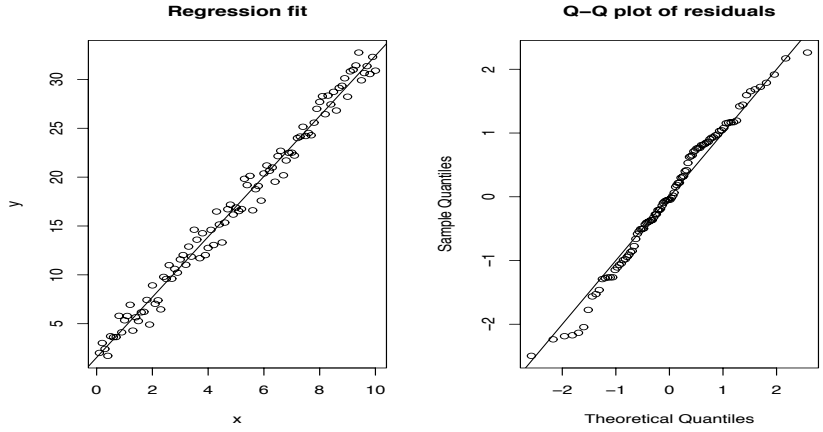
- **Presence of outliers :** If some of the semi-studentized residuals have "too large" absolute values (say $|e_i^*| > 3$ for some i) then the corresponding observation can be taken to be an outlier (in Y).

- **Nonnormality of errors :** This can be studied graphically by using the **normal probability plot**, or **Q-Q** (standing for quantile-quantile) plot. In this plot the ordered residuals (or observed quantiles) of the residuals are plotted against the expected quantiles assuming that e_i's are approximately normal and independent with mean 0 and variance $= MSE$. This results in plotting the k-th largest e_i against $\sqrt{MSE} \cdot z[(k - 0.375)/(n + 0.25)]$, where $z(q)$ is the q-th quantile of $N(0, 1)$ distribution, where $0 < q < 1$. If the errors are normally distributed then the points on the plot should almost along the diagonal line. Departures from that could indicate **skewness** or **heavier-tailed** distributions.

  (a) True model : $Y = 2 + 3X + \varepsilon$, where $\varepsilon \sim N(0, 1)$. 100 observations, with $X_i = \frac{i}{10}$, $i = 1, \ldots, 100$.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	1.5413	0.2196	7.02	2.92×10^{-10}
Slope	3.08907	0.03775	81.84	$< 2 \times 10^{-16}$

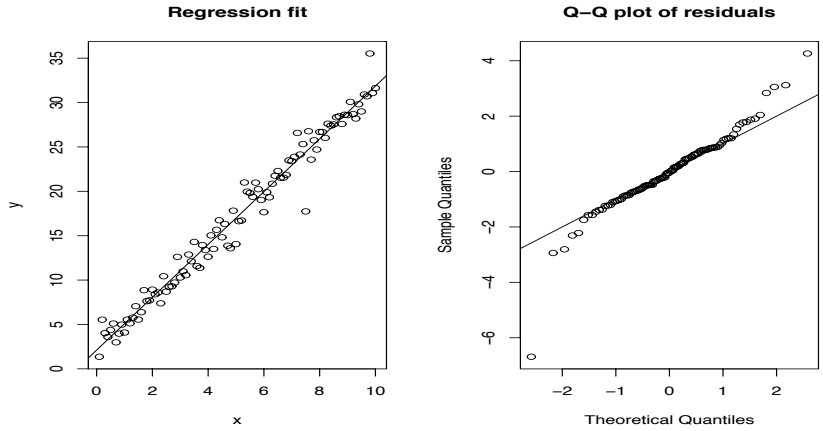
  $\sqrt{MSE} = 1.09$, $R^2 = 0.9856$.



(b) True model : $Y = 2 + 3X + \varepsilon$, where $\varepsilon \sim t_5$. 100 observations, with $X_i = \frac{i}{10}$, $i = 1, \ldots, 100$.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	2.11144	0.28279	7.467	3.42×10^{-11}
Slope	2.97458	0.04862	61.185	$< 2 \times 10^{-16}$

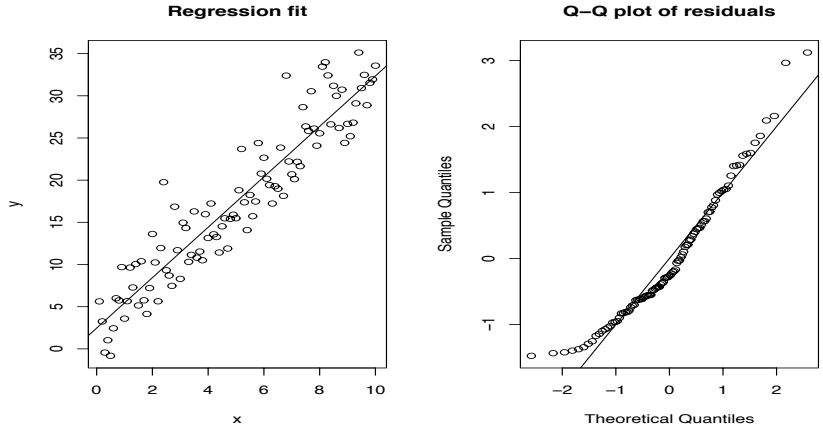
$\sqrt{MSE} = 1.403$, $R^2 = 0.9745$.



(c) True model : $Y = 2 + 3X + \varepsilon$ where $\varepsilon \sim (\chi^2_5 - 5)$. 100 observations, with $X_i = \frac{i}{10}$, $i = 1, \ldots, 100$.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	2.4615	0.6533	3.768	0.000281
Slope	2.9894	0.1123	26.617	$< 2 \times 10^{-16}$

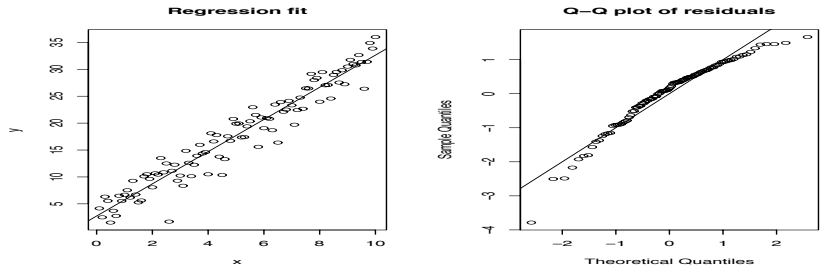
$\sqrt{MSE} = 3.242$, $R^2 = 0.8785$.



(d) True model : $Y = 2 + 3X + \varepsilon$ where $\varepsilon \sim (5 - \chi^2_5)$. 100 observations, with $X_i = \frac{i}{10}$, $i = 1, \ldots, 100$.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	2.7402	0.4694	5.838	6.87×10^{-8}
Slope	2.9896	0.0807	37.048	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 2.329$, $R^2 = 0.9334$.



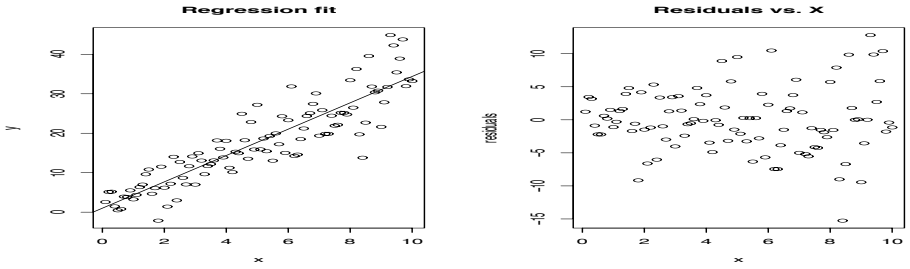
- **Heteroscedasticity or unequal variance :** The variance of the error ε_i may sometimes depend on the value of X_i. This is often reflected in the plot of residuals versus X through an unequal spread of the residuals along the X-axis.

  One possibility is that the variance either increases or decreases with increasing value of X. This is often true for financial data, where the volume of transactions usually has a role in the uncertainty of the market. Another possibility is that the data may come from different strata with different variabilities. E.g. different measuring instruments, with different precisions, may have been used.

  (a) True model : $Y = 2 + 3X + \sigma(X)\varepsilon$ where $\varepsilon \sim N(0, 1)$. 100 observations. $\log \sigma^2(X) = 1 + 0.1 * X$. $X_i = \frac{i}{10}$ for $i = 1, \ldots, 100$.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	1.0074	0.9729	1.035	0.303
Slope	3.3382	0.1673	19.958	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 4.828$, $R^2 = 0.8026$.



5 Multiple linear regression

A response variable Y is linearly related to p different explanatory variables $X^{(1)}, \ldots, X^{(p-1)}$ (where $p \geq 2$). The regression model is given by

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \cdots + \beta_p X_i^{(p-1)} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{16}$$

where ε_i have mean zero, variance σ^2 and are uncorrelated. The equation (16) can be expressed in matrix notations as

$$Y = \mathbf{X}\beta + \varepsilon, \qquad \text{where} \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ \varepsilon_n \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_1^{(1)} & X_1^{(2)} & \cdots & X_1^{(p-1)} \\ 1 & X_2^{(1)} & X_2^{(2)} & \cdots & X_2^{(p-1)} \\ . & . & . & \cdots & . \\ 1 & X_n^{(1)} & X_n^{(2)} & \cdots & X_n^{(p-1)} \end{bmatrix}, \qquad \text{and} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ \beta_{p-1} \end{bmatrix}.$$

So $\mathbf{X}$ is an $n \times p$ matrix.

5.1 Estimation problem

Note that β is estimated by the *least squares* procedure. That is minimizing the sum of squared errors $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^{(1)} - \cdots - \beta_{p-1} X_i^{(p-1)})^2$. The latter quantity can be expressed in matrix notations as $\| Y - \mathbf{X}\beta \|^2$. Minimization with respect to the parameter β (a $p \times 1$ vector) gives rise

to the **normal equations** :

$$\begin{aligned}
b_0 n + b_1 \sum_i X_i^{(1)} + b_2 \sum_i X_i^{(2)} + \cdots + b_{p-1} \sum_i X_i^{(p-1)} &= \sum_i Y_i \\
b_0 \sum_i X_i^{(1)} + b_1 \sum_i (X_i^{(1)})^2 + b_2 \sum_i X_i^{(1)} X_i^{(2)} + \cdots + b_{p-1} \sum_i X_i^{(1)} X_i^{(p-1)} &= \sum_i X_i^{(1)} Y_i \\
\cdots \quad \cdots \quad \cdots \quad \cdots &= \cdot \\
b_0 \sum_i X_i^{(p-1)} + b_1 \sum_i X_i^{(p-1)} X_i^{(1)} + b_2 \sum_i X_i^{(p-1)} X_i^{(2)} + \cdots + b_{p-1} \sum_i (X_i^{(p-1)})^2 &= \sum_i X_i^{(p-1)} Y_i
\end{aligned}$$

Observe that we can express this system of p equations in p variables $b_0, b_1, \ldots, b_{p-1}$ as

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T Y, \tag{17}$$

where $\mathbf{b}$ is a $p \times 1$ vector with $\mathbf{b}^T = (b_0, b_1, \ldots, b_{p-1})$.

If the $p \times p$ matrix $\mathbf{X}^T \mathbf{X}$ is nonsingular (as we shall assume for the time being), then the solution to this system is given by

$$\widehat{\beta} = \mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y.$$

This is the least squares estimate of β.

5.2 Expected value and variance of random vectors

For an $m \times 1$ vector $\mathbf{Z}$, with coordinates $Z_1, \ldots, Z_m$, the expected value (or mean), and variance of $\mathbf{Z}$ are defined as

$$E(\mathbf{Z}) = E \begin{bmatrix} Z_1 \\ Z_2 \\ \cdot \\ Z_m \end{bmatrix} = \begin{bmatrix} E(Z_1) \\ E(Z_2) \\ \cdot \\ E(Z_m) \end{bmatrix} \quad \text{and} \quad \mathrm{Var}(\mathbf{Z}) = \begin{bmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdot & \mathrm{Cov}(Z_1, Z_m) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Var}(Z_2) & \cdot & \mathrm{Cov}(Z_2, Z_m) \\ \cdot & \cdot & \cdots & \cdot \\ \mathrm{Cov}(Z_m, Z_1) & \mathrm{Cov}(Z_m, Z_2) & \cdot & \mathrm{Var}(Z_m) \end{bmatrix}.$$

Observe that $\mathrm{Var}(\mathbf{Z})$ is an $m \times m$ matrix. Also, since $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cov}(Z_j, Z_i)$ for all $1 \leq i, j \leq m$, $\mathrm{Var}(\mathbf{Z})$ is a symmetric matrix. Moreover, it can be checked, using the relationship that $\mathrm{Cov}(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)E(Z_j)$, that $\mathrm{Var}(\mathbf{Z}) = E(\mathbf{Z}\mathbf{Z}^T) - (E(\mathbf{Z}))(E(\mathbf{Z}))^T$.

5.3 Expected value and variance of least squares estimate

Using the assumptions on the errors $\varepsilon_1, \ldots, \varepsilon_n$, it follows that $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$. Since, we are treating $X_i^{(k)}$'s as non-random quantities, it follows, using the *linearity of expectation* that

$$E(Y) = E(\mathbf{X}\beta + \varepsilon) = \mathbf{X}\beta + E(\varepsilon) = \mathbf{X}\beta \quad \text{and} \quad \mathrm{Var}(Y) = \mathrm{Var}(\mathbf{X}\beta + \varepsilon) = \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Using these we have the following results about the least squares estimate $\widehat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y$.

$$E(\widehat{\beta}) = E((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(Y) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta$$

and

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}) &= \mathrm{Var}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathrm{Var}(Y) ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T \\
&= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\sigma^2 I_n) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X})^{-1} \\
&= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.
\end{aligned}$$

Here we have used the following facts : (i) $\mathrm{Var}(A\mathbf{Z}) = A\mathrm{Var}(\mathbf{Z})A^T$; (ii) $(AB)^T = B^T A^T$ and (iii) $(A^T)^T = A$.

5.4 Fitted values and residuals

The fitted value for i-th observation is $\widehat{Y}_i = b_0 + b_1 X_i^{(1)} + \cdots + b_{p-1} X_i^{(p-1)}$, and the residual is $e_i = Y_i - \widehat{Y}_i$. Using matrix notations, the vector of fitted values, $\widehat{Y}$, can be expressed as

$$\widehat{Y} = \mathbf{X}\mathbf{b} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

The $n \times n$ matrix $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the **hat matrix** and is denoted by $\mathbf{H}$. Thus $\widehat{Y} = \mathbf{H}Y$.

The vector of residuals, to be denoted by $\mathbf{e}$ (with i-th coordinate e_i, for $i = 1, \ldots, n$) can therefore be expressed as

$$\mathbf{e} = Y - \widehat{Y} = Y - \mathbf{H}Y = (I_n - \mathbf{H})Y = (I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)Y.$$

- **Hat matrix :** Check that the matrix $\mathbf{H}$ has the property that $\mathbf{H}\mathbf{H} = \mathbf{H}$ and $(I_n - \mathbf{H})(I_n - \mathbf{H}) = (I_n - \mathbf{H})$. A square matrix A having the property that $AA = A$ is called an *idempotent matrix*. So both $\mathbf{H}$ and $I_n - \mathbf{H}$ are idempotent matrices. The important implication of the equation

  $$\widehat{Y} = \mathbf{H}Y$$

  is that the matrix $\mathbf{H}$ **projects** the response vector Y as a linear combination of the columns of the matrix $\mathbf{X}$ to obtain the vector of fitted values, $\widehat{Y}$. Similarly, the matrix $I_n - \mathbf{H}$ applied to Y gives the residual vector $\mathbf{e}$.

- **Properties of residuals :** Many of the properties of residual can be deduced by studying the properties of the matrix $\mathbf{H}$. Some of them are listed below.

  1. $\sum_i e_i = 0$ and $\sum_i X_i^{(j)} e_i = 0$, for $j = 1, \ldots, p-1$. These are results of the following:

     $$\mathbf{X}^T\mathbf{e} = \mathbf{X}^T(I_n - \mathbf{H})Y = \mathbf{X}^T Y - \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \mathbf{X}^T Y - \mathbf{X}^T Y = 0. \qquad (18)$$

     Also, note that $\widehat{Y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$, and hence

     $$\sum_i \widehat{Y}_i e_i = \widehat{Y}^T\mathbf{e} = Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} = 0,$$

     using (18).

5.5 ANOVA decomposition

The matrix viewpoint gives a coherent way of representing the different components of the analysis of variance of the response in regression. As before, we need to deal with the objects

$$SSTO = \sum_i (Y_i - \overline{Y})^2, \qquad SSE = \sum_i (Y_i - \widehat{Y}_i)^2 = \sum_i e_i^2, \qquad \text{and} \quad SSR = SSTO - SSE.$$

The degrees of freedom of SSR is $p-1$. The degrees of freedom of $SSTO$ is $n-1$ and d.f.(SSE) = d.f.($SSTO$) - d.f.(SSR) = $n - 1 - (p-1) = n - p$. Moreover,

$$\begin{aligned}
\overline{Y} &= \frac{1}{n}\sum_i Y_i = (\frac{1}{n})Y^T\mathbf{1} \\
SSTO &= \sum_i Y_i^2 - \frac{1}{n}(\sum_i Y_i)^2 = Y^T Y - (\frac{1}{n})Y^T\mathbf{J}Y \\
SSE &= \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(I - \mathbf{H})(I - \mathbf{H})\mathbf{Y} = \mathbf{Y}^T(I - \mathbf{H})\mathbf{Y} \\
SSE &= Y^T Y - \widehat{\beta}^T\mathbf{X}^T Y
\end{aligned}$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}^T$.

5.6 Inference in multiple linear regression

We can ask the same questions regarding estimation of various parameters as we did in the case of regression with one predictor variable.

- **Mean and standard error of estimates :** We already checked that (with $\mathbf{b} \equiv \widehat{\beta}$) $E(\mathbf{b}) = \beta$ and $\text{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. And hence the estimated variance-covariance matrix of $\mathbf{b}$ is $\widehat{\text{Var}}(\mathbf{b}) = MSE(\mathbf{X}^T\mathbf{X})^{-1}$. Denote by $s(b_j)$ the standard error of $b_j = \widehat{\beta}_j$. Then $s^2(b_j)$ is the $(j+1)$-th diagonal entry of the $p \times p$ matrix $\widehat{\text{Var}}(\mathbf{b})$.

  Note that
  $$\text{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \text{ so that } \widehat{\text{Var}}(\mathbf{b}) = \text{MSE } (\mathbf{X}^T\mathbf{X})^{-1}.$$

- **Hypothesis test for regression effect:** We can use the ANOVA decomposition to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ (no regression effect), against H_1 : not all β_j are equal to zero. The test statistic is
  $$F^* = \frac{\frac{SSR}{\text{d.f.}(SSR)}}{\frac{SSE}{\text{d.f.}(SSE)}} = \frac{SSR/(p-1)}{SSE/(n-p)}.$$

  Under H_0 and assumption of normal errors, F^* has $F_{p-1,n-p}$ distribution. So, reject H_0 in favor of H_1, at level α if $F^* > F(1-\alpha; p-1, n-p)$.

- **Hypothesis tests for individual parameters :** Under $H_0 : \beta_j = \beta_j^0$, for a given $j \in \{1, \ldots, p-1\}$,
  $$t^* = \frac{b_j - \beta_j^0}{s(b_j)} \sim t_{n-p}.$$

  So, if $H_1 : \beta_j \neq \beta_j^0$, then reject H_0 in favor of H_1 at level α if $|t^*| > t(1-\alpha/2; n-p)$.

- **Confidence intervals for individual parameters :** Based on the result above, 100(1-α)% two-sided confidence interval for β_j is given by
  $$b_j \pm t(1-\alpha/2; n-p)s(b_j).$$

- **Estimation of mean response :** Since
  $$E(Y|X_h) = \beta^T X_h, \quad \text{where} \quad X_h = \begin{bmatrix} 1 \\ X_h^{(1)} \\ \cdot \\ \cdot \\ X_h^{(p-1)} \end{bmatrix},$$

  an *unbiased* point estimate of $E(Y|X_h)$ is $\widehat{Y}_h = \mathbf{b}^T X_h = b_0 + b_1 X_h^{(1)} + \cdots + b_{p-1} X_h^{(p-1)}$. Using the Working-Hotelling procedure, an $100(1-\alpha)\%$ confidence region for the entire regression surface (that is, confidence region for $E(Y|X_h)$ for all possible values of X_h), is given by
  $$\widehat{Y}_h \pm \sqrt{pF(1-\alpha; p, n-p)}s(\widehat{Y}_h),$$

where $s(\widehat{Y}_h)$ is the estimated standard error of $\widehat{Y}_h$ and is given by

$$s^2(\widehat{Y}_h) = (MSE) \cdot X_h^T (\mathbf{X}^T \mathbf{X})^{-1} X_h.$$

The last formula can be deduced from the fact that

$$\mathrm{Var}(\widehat{Y}_h) = \mathrm{Var}(X_h^T \mathbf{b}) = X_h^T \mathrm{Var}(\mathbf{b}) X_h = \sigma^2 X_h^T (\mathbf{X}^T \mathbf{X})^{-1} X_h.$$

Also, using the fact that $(\widehat{Y}_h - X_h^T \beta)/s(\widehat{Y}_h) \sim t_{n-p}$, a pointwise, $100(1-\alpha)\%$ two-sided confidence interval for $E(Y|X_h) = X_h^T \beta$ is given by

$$\widehat{Y}_h \pm t(1-\alpha/2; n-p) s(\widehat{Y}_h).$$

- **Coefficient of multiple determination :** The quantity $R^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}$ is a measure of association between the response Y and the predictors $X^{(1)}, \ldots, X^{(p-1)}$. This has the interpretation that R^2 is the proportion of variability in the response explained by the predictors. Another interpretation is that R^2 is the maximum squared correlation between Y and any linear function of $X^{(1)}, \ldots, X^{(p-1)}$.

- **Adjusted R^2 :** If one increases number of predictor variables in the regression model, then R^2 increases. To take into account the number of predictors, the measure called *adjusted multiple R-squared*, or,

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right) \frac{SSE}{SSTO},$$

is used. Notice that $R_a^2 < R^2$, and when the number of observations is not too large, R_a^2 can be substantially smaller than R^2. Even though R_a^2 does not have as nice an interpretation as R^2, in multiple linear regression, this considered to be a better measure of association.

Appendix: Basic facts about vectors and matrices

- **Addition rule for matrices :** If $c_1, \ldots, c_k$ are scalers, and $A_1, \ldots, A_k$ are all $m \times n$ matrices, then $B = c_1 A_1 + c_2 A_2 + \cdots + c_k A_k$ is an $m \times n$ matrix with (i,j)-th entry of B : $B(i,j) = c_1 A_1(i,j) + c_2 A_2(i,j) + \cdots + c_k A_k(i,j)$, for all $i = 1, \ldots, m$; $j = 1, \ldots, n$. (Note sometimes we denote the entries of a matrix by A_{ij} and sometimes by $A(i,j)$. But always the first index is for the row and the second index is for the column).

- **Transpose of a matrix :** If A is an $m \times n$ matrix, then A^T (spelled A-transpose) is the $n \times m$ matrix B whose (i,j)-th entry $B_{ij} = A_{ji}$ for all $i = 1, \ldots, n$; $j = 1, \ldots, m$.

- **Inner product of vectors :** If $\mathbf{x}$ and $\mathbf{y}$ are two $m \times 1$ vectors, then the *inner product* (or, *dot product*) between $\mathbf{x}$ and $\mathbf{y}$ is given by : $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^m x_i y_i$. Note that $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$.

- **Multiplication of matrices :** If A is an $m \times n$ matrix and B is an $n \times p$ matrix then the product $AB = C$, say, is defined and it is an $m \times p$ matrix with (i,j)-th entry : $C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$ for all $i = 1, \ldots, m$; $j = 1, \ldots, p$. Note that for $m \times 1$ vectors $\mathbf{x}$ and $\mathbf{y}$, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x}$. In other words, the (i,j)-th entry of AB is the inner product of i-th row of A and j-th column of B.

- **Special matrices :**
  1. **Square matrix :** A matrix A is square if it is $m \times m$ (that is, number of rows = number of columns).
  2. **Symmetric matrix :** An $m \times m$ (square) matrix A is symmetric if $A = A^T$. That is, for all $1 \leq i, j \leq m$, $A_{ij} = A_{ji}$.
  3. **Diagonal matrix :** A $m \times m$ matrix with all the entries zero except (possibly) the entries on the diagonal (that is the (i, i)-th entry for all $i = 1, \ldots, m$) is called a diagonal matrix.
  4. **Identity matrix :** The $m \times m$ diagonal matrix with all diagonal entries equal to 1 is called the identity matrix and is denoted by I (or, I_m). It has the property that for any $m \times n$ matrix A and any $p \times m$ matrix B, $IA = A$ and $BI = B$.
  5. **One vector :** The $m \times 1$ vector with all entries equal to 1 is usually called the one vector (non-standard term) and is denoted by $\mathbf{1}$ (or, $\mathbf{1}_m$).
  6. **Ones matrix :** The $m \times m$ matrix with all entries equal to 1 is denote by J (or, J_m). Note that $J_m = \mathbf{1}_m \mathbf{1}_m^T$.
  7. **Zero vector :** The $m \times 1$ vector with all entries zero is called the zero vector and is denoted by $\mathbf{0}$ (or, $\mathbf{0}_m$).
- **Multiplication is not commutative :** If A and B are both $m \times m$ matrices then both AB and BA are defined and are $m \times m$ matrices. However, in general $AB \neq BA$. Notice that $I_m B = B I_m = B$, where I_m is the identity matrix.
- **Linear independence :** The $m \times 1$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$, ($k$ arbitrary) are said to be *linearly dependent*, if there exist constants $c_1, \ldots, c_m$, **not all zero**, such that

  $$c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \cdots + c_m \mathbf{x}_m = \mathbf{0}.$$

  If no such sequence of numbers $c_1, \ldots, c_m$ exists then the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are said to be *linearly independent*.
  1. **Relationship with dimension :** If $k > m$ then the $m \times 1$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are **always** linearly **dependent**.
  2. **Rank of a matrix :** For an $m \times n$ matrix A, the **rank** of A, written $\text{rank}(A)$ is the maximal number of linearly independent columns of A (treating each column as an $m \times 1$ vector). Also, $\text{rank}(A) \leq \min\{m, n\}$.
  3. **Nonsingular matrix :** If an $m \times m$ matrix A has *full rank*, that is, $\text{rank}(A) = m$, (which is equivalent to saying that all the columns of A are linearly independent), then the matrix A is called **nonsingular**
- **Inverse of a matrix :** If an $m \times m$ matrix A is *nonsingular*, then it has an *inverse*, that is a *unique* $m \times m$ matrix denoted by A^{-1} that satisfies the relationship : $A^{-1}A = I_m = AA^{-1}$.
  1. **Inverse of a 2×2 matrix :** Let a 2×2 matrix A be expressed as $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then A is nonsingular (and hence has an inverse) *if and only if* $ad - bc \neq 0$. If this is satisfied then the inverse is

     $$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

2. **Solution of a system of linear equations :** A system of m linear equations in m variables $b_1, \ldots, b_m$ can be expressed as

$$\begin{aligned} a_{11}b_1 + a_{12}b_2 + \cdots + a_{1m}b_m &= c_1 \\ a_{21}b_1 + a_{22}b_2 + \cdots + a_{2m}b_m &= c_2 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots &= \cdot \\ a_{m1}b_1 + a_{m2}b_2 + \cdots + a_{mm}b_m &= c_m \end{aligned}$$

Here the *coefficients* a_{ij} and the constants c_i are considered known. This system can be expressed in matrix form as $A\mathbf{b} = \mathbf{c}$, where A is the $m \times m$ matrix with (i, j)-th entry a_{ij} and $\mathbf{b}$ and $\mathbf{c}$ are $m \times 1$ vectors with i-th entries b_i and c_i, respectively, for $i = 1, \ldots, m$; $j = 1, \ldots, m$.

If the matrix A is nonsingular, then a *unique solution* exists for this system of equations and is given by $\mathbf{b} = A^{-1}\mathbf{c}$. To see this, note that since $A(A^{-1}\mathbf{c}) = (AA^{-1})\mathbf{c} = I\ \mathbf{c} = \mathbf{c}$, it shows that $\mathbf{A}^{-1}\mathbf{c}$ is a solution. On the other hand, if $\mathbf{b} = \mathbf{b}^*$ is a solution, then it satisfies $A\mathbf{b}^* = \mathbf{c}$. Hence $\mathbf{b}^* = I\ \mathbf{b}^* = (A^{-1}A)\mathbf{b}^* = A^{-1}(A\mathbf{b}^*) = A^{-1}\mathbf{c}$, which proves uniqueness.