Chapter 2

Simple Linear Regression

2.1 Introduction

The term "regression" and the methods for investigating the relationships between two variables may date back to about 100 years ago. It was first introduced by Francis Galton in 1908, the renowned British biologist, when he was engaged in the study of heredity. One of his observations was that the children of tall parents to be taller than average but not as tall as their parents. This "regression toward mediocrity" gave these statistical methods their name. The term regression and its evolution primarily describe statistical relations between variables. In particular, the simple regression is the regression method to discuss the relationship between one dependent variable (y) and one independent variable (x). The following classical data set contains the information of parent's height and children's height.

Table 2.1 Parent's Height and Children's Height

Parent	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5
Children	65.8	66.7	67.2	67.6	68.2	68.9	69.5	69.9	72.2

The mean height is 68.44 for children and 68.5 for parents. The regression line for the data of parents and children can be described as

$$\text{child height} = 21.52 + 0.69 \text{ parent height}.$$

The simple linear regression model is typically stated in the form

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where y is the dependent variable, β_0 is the y intercept, β_1 is the slope of the simple linear regression line, x is the independent variable, and ε is the

random error. The dependent variable is also called response variable, and the independent variable is called explanatory or predictor variable. An explanatory variable explains causal changes in the response variables. A more general presentation of a regression model may be written as

$$y = E(y) + \epsilon,$$

where $E(y)$ is the mathematical expectation of the response variable. When $E(y)$ is a linear combination of exploratory variables $x_1, x_2, \cdots, x_k$ the regression is the linear regression. If $k = 1$ the regression is the simple linear regression. If $E(y)$ is a nonlinear function of $x_1, x_2, \cdots, x_k$ the regression is nonlinear. The classical assumptions on error term are $E(\varepsilon) = 0$ and a constant variance $\text{Var}(\varepsilon) = \sigma^2$. The typical experiment for the simple linear regression is that we observe n pairs of data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ from a scientific experiment, and model in terms of the n pairs of the data can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, \cdots, n,$$

with $E(\varepsilon_i) = 0$, a constant variance $\text{Var}(\varepsilon_i) = \sigma^2$, and all ε_i's are independent. Note that the actual value of σ^2 is usually unknown. The values of x_i's are measured "exactly", with no measurement error involved. After model is specified and data are collected, the next step is to find "good" estimates of β_0 and β_1 for the simple linear regression model that can best describe the data came from a scientific experiment. We will derive these estimates and discuss their statistical properties in the next section.

2.2 Least Squares Estimation

The least squares principle for the simple linear regression model is to find the estimates b_0 and b_1 such that the sum of the squared distance from actual response y_i and predicted response $\hat{y}_i = \beta_0 + \beta_1 x_i$ reaches the minimum among all possible choices of regression coefficients β_0 and β_1. i.e.,

$$(b_0, b_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

The motivation behind the least squares method is to find parameter estimates by choosing the regression line that is the most "closest" line to

all data points (x_i, y_i). Mathematically, the least squares estimates of the simple linear regression are given by solving the following system:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0 \tag{2.1}$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0 \tag{2.2}$$

Suppose that b_0 and b_1 are the solutions of the above system, we can describe the relationship between x and y by the regression line $\hat{y} = b_0 + b_1 x$ which is called the fitted regression line by convention. It is more convenient to solve for b_0 and b_1 using the centralized linear model:

$$y_i = \beta_0^* + \beta_1 (x_i - \bar{x}) + \varepsilon_i,$$

where $\beta_0 = \beta_0^* - \beta_1 \bar{x}$. We need to solve for

$$\frac{\partial}{\partial \beta_0^*} \sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \bar{x}))]^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \bar{x}))]^2 = 0$$

Taking the partial derivatives with respect to β_0 and β_1 we have

$$\sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \bar{x}))] = 0$$

$$\sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \bar{x}))](x_i - \bar{x}) = 0$$

Note that

$$\sum_{i=1}^{n} y_i = n\beta_0^* + \sum_{i=1}^{n} \beta_1 (x_i - \bar{x}) = n\beta_0^* \tag{2.3}$$

Therefore, we have $\beta_0^* = \dfrac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$. Substituting β_0^* by $\bar{y}$ in (2.3) we obtain

$$\sum_{i=1}^{n} [y_i - (\bar{y} + \beta_1 (x_i - \bar{x}))](x_i - \bar{x}) = 0.$$

Denote b_0 and b_1 be the solutions of the system (2.1) and (2.2). Now it is easy to see

$$b_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{2.4}$$

and

$$b_0 = b_0^* - b_1\bar{x} = \bar{y} - b_1\bar{x} \tag{2.5}$$

The fitted value of the simple linear regression is defined as $\hat{y}_i = b_0 + b_1 x_i$. The difference between y_i and the fitted value $\hat{y}_i$, $e_i = y_i - \hat{y}_i$, is referred to as the regression residual. Regression residuals play an important role in the regression diagnosis on which we will have extensive discussions later. Regression residuals can be computed from the observed responses y_i's and the fitted values $\hat{y}_i$'s, therefore, residuals are observable. It should be noted that the error term ε_i in the regression model is unobservable. Thus, regression error is unobservable and regression residual is observable. Regression error is the amount by which an observation differs from its expected value; the latter is based on the whole population from which the statistical unit was chosen randomly. The expected value, the average of the entire population, is typically unobservable.

Example 2.1. If the average height of 21-year-old male is 5 feet 9 inches, and one randomly chosen male is 5 feet 11 inches tall, then the "error" is 2 inches; if the randomly chosen man is 5 feet 7 inches tall, then the "error" is -2 inches. It is as if the measurement of man's height was an attempt to measure the population average, so that any difference between man's height and average would be a measurement error.

A residual, on the other hand, is an observable estimate of unobservable error. The simplest case involves a random sample of n men whose heights are measured. The sample average is used as an estimate of the population average. Then the difference between the height of each man in the sample and the unobservable population average is an error, and the difference between the height of each man in the sample and the observable sample average is a residual. Since residuals are observable we can use residual to estimate the unobservable model error. The detailed discussion will be provided later.

2.3 Statistical Properties of the Least Squares Estimation

In this section we discuss the statistical properties of the least squares estimates for the simple linear regression. We first discuss statistical properties without the distributional assumption on the error term, but we shall assume that $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2$, and ϵ_i's for $i = 1, 2, \cdots, n$ are independent.

Theorem 2.1. *The least squares estimator b_0 is an unbiased estimate of β_0.*

Proof.

$$Eb_0 = E(\bar{y} - b_1\bar{x}) = E\Big(\frac{1}{n}\sum_{i=1}^{n} y_i\Big) - Eb_1\bar{x} = \frac{1}{n}\sum_{i=1}^{n} Ey_i - \bar{x}Eb_1$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \beta_1\frac{1}{n}\sum_{i=1}^{n} x_i - \beta_1\bar{x} = \beta_0.$$

□

Theorem 2.2. *The least squares estimator b_1 is an unbiased estimate of β_1.*

Proof.

$$\begin{aligned}
E(b_1) &= E\Big(\frac{S_{xy}}{S_{xx}}\Big) \\
&= \frac{1}{S_{xx}}E\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})Ey_i \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i) \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\beta_1(x_i - \bar{x}) \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\beta_1 = \frac{S_{xx}}{S_{xx}}\beta_1 = \beta_1
\end{aligned}$$

□

Theorem 2.3. *$Var(b_1) = \dfrac{\sigma^2}{nS_{xx}}$.*

Proof.

$$\begin{aligned}
\mathrm{Var}(b_1) &= \mathrm{Var}\Big(\frac{S_{xy}}{S_{xx}}\Big) \\
&= \frac{1}{S_{xx}^2}\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\Big) \\
&= \frac{1}{S_{xx}^2}\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}y_i(x_i-\bar{x})\Big) \\
&= \frac{1}{S_{xx}^2}\frac{1}{n^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\mathrm{Var}(y_i) \\
&= \frac{1}{S_{xx}^2}\frac{1}{n^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2}{nS_{xx}}
\end{aligned}$$

□

Theorem 2.4. *The least squares estimator b_1 and $\bar{y}$ are uncorrelated. Under the normality assumption of y_i for $i = 1, 2, \cdots, n$, b_1 and $\bar{y}$ are normally distributed and independent.*

Proof.

$$\begin{aligned}
\mathrm{Cov}(b_1, \bar{y}) &= \mathrm{Cov}(\frac{S_{xy}}{S_{xx}},\ \ \bar{y}) \\
&= \frac{1}{S_{xx}}\ \mathrm{Cov}(S_{xy},\ \bar{y}) \\
&= \frac{1}{nS_{xx}}\ \mathrm{Cov}\Big(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),\ \ \bar{y}\Big) \\
&= \frac{1}{nS_{xx}}\ \mathrm{Cov}\Big(\sum_{i=1}^{n}(x_i-\bar{x})y_i,\ \ \bar{y}\Big) \\
&= \frac{1}{n^2S_{xx}}\ \mathrm{Cov}\Big(\sum_{i=1}^{n}(x_i-\bar{x})y_i,\ \ \sum_{i=1}^{n}y_i\Big) \\
&= \frac{1}{n^2S_{xx}}\ \sum_{i,j=1}^{n}(x_i-\bar{x})\ \mathrm{Cov}(y_i, y_j)
\end{aligned}$$

Note that $E\varepsilon_i = 0$ and ε_i's are independent we can write

$$\mathrm{Cov}(y_i, y_j) = E[\ (y_i - Ey_i)(y_j - Ey_j)\ ] = E(\varepsilon_i,\ \ \varepsilon_j) = \begin{cases} \sigma^2, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

Thus, we conclude that

$$\text{Cov}(b_1, \bar{y}) = \frac{1}{n^2 S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})\sigma^2 = 0.$$

Recall that zero correlation is equivalent to the independence between two normal variables. Thus, we conclude that b_0 and $\bar{y}$ are independent. □

Theorem 2.5. *Var*$(b_0) = \Big(\frac{1}{n} + \frac{\bar{x}^2}{nS_{xx}}\Big)\sigma^2$.

Proof.

$$\begin{aligned}
\text{Var}(b_0) &= \text{Var}(\bar{y} - b_1\bar{x}) \\
&= \text{Var}(\bar{y}) + (\bar{x})^2 \text{Var}(b_1) \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{nS_{xx}} \\
&= \Big(\frac{1}{n} + \frac{\bar{x}^2}{nS_{xx}}\Big)\sigma^2
\end{aligned}$$

□

The properties $1 - 5$, especially the variances of b_0 and b_1, are important when we would like to draw statistical inference on the intercept and slope of the simple linear regression.

Since the variances of least squares estimators b_0 and b_1 involve the variance of the error term in the simple regression model. This error variance is unknown to us. Therefore, we need to estimate it. Now we discuss how to estimate the variance of the error term in the simple linear regression model. Let y_i be the observed response variable, and $\hat{y}_i = b_0 + b_1 x_i$, the fitted value of the response. Both y_i and $\hat{y}_i$ are available to us. The true error σ_i in the model is not observable and we would like to estimate it. The quantity $y_i - \hat{y}_i$ is the empirical version of the error ε_i. This difference is regression residual which plays an important role in regression model diagnosis. We propose the following estimation of the error variance based on e_i:

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

Note that in the denominator is $n-2$. This makes s^2 an unbiased estimator of the error variance σ^2. The simple linear model has two parameters, therefore, $n - 2$ can be viewed as $n-$ number of parameters in simple

linear regression model. We will see in later chapters that it is true for all general linear models. In particular, in a multiple linear regression model with p parameters the denominator should be $n-p$ in order to construct an unbiased estimator of the error variance σ^2. Detailed discussion can be found in later chapters. The unbiasness of estimator s^2 for the simple linear regression can be shown in the following derivations.

$$y_i - \hat{y}_i = y_i - b_0 - b_1 x_i = y_i - (\bar{y} - b_1\bar{x}) - b_1 x_i = (y_i - \bar{y}) - b_1(x_i - \bar{x})$$

It follows that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \bar{y}) - b_1\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Note that $(y_i - \hat{y}_i)x_i = [(y_i - \bar{y}) - b_1(x_i - \bar{x})]x_i$, hence we have

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]x_i \\
&= \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})](x_i - \bar{x}) \\
&= \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - b_1\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= n(S_{xy} - b_1 S_{xx}) = n\Big(S_{xy} - \frac{S_{xy}}{S_{xx}}S_{xx}\Big) = 0
\end{aligned}$$

To show that s^2 is an unbiased estimate of the error variance, first we note that

$$(y_i - \hat{y}_i)^2 = [(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2,$$

therefore,

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2b_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}_i) + b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2nb_1 S_{xy} + nb_1^2 S_{xx} \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2n\frac{S_{xy}}{S_{xx}}S_{xy} + n\frac{S_{xy}^2}{S_{xx}^2}S_{xx} \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - n\frac{S_{xy}^2}{S_{xx}}
\end{aligned}$$

Since

$$(y_i - \bar{y})^2 = [\beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]^2$$

and

$$(y_i - \bar{y})^2 = \beta_1^2(x_i - \bar{x})^2 + (\varepsilon_i - \bar{\varepsilon})^2 + 2\beta_1(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}),$$

therefore,

$$E(y_i - \bar{y})^2 = \beta_1^2(x_i - \bar{x})^2 + E(\varepsilon_i - \bar{\varepsilon})^2 = \beta_1^2(x_i - \bar{x})^2 + \frac{n-1}{n}\sigma^2,$$

and

$$\sum_{i=1}^{n} E(y_i - \bar{y})^2 = n\beta_1^2 S_{xx} + \sum_{i=1}^{n} \frac{n-1}{n}\sigma^2 = n\beta_1^2 S_{xx} + (n-1)\sigma^2.$$

Furthermore, we have

$$\begin{aligned} E(S_{xy}) &= E\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\Big) \\ &= \frac{1}{n}E\sum_{i=1}^{n}(x_i - \bar{x})y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})Ey_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i) \\ &= \frac{1}{n}\beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i \\ &= \frac{1}{n}\beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 = \beta_1 S_{xx} \end{aligned}$$

and

$$\mathrm{Var}\Big(S_{xy}\Big) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})y_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\mathrm{Var}(y_i) = \frac{1}{n}S_{xx}\sigma^2$$

Thus, we can write

$$E(S_{xy}^2) = \mathrm{Var}(S_{xy}) + [E(S_{xy})]^2 = \frac{1}{n}S_{xx}\sigma^2 + \beta_1^2 S_{xx}^2$$

and

$$E\Big(\frac{nS_{xy}^2}{S_{xx}}\Big) = \sigma^2 + n\beta_1^2 S_{xx}.$$

Finally, $E(\hat{\sigma}^2)$ is given by:

$$E\sum_{i=1}^{n}(y_i-\hat{y})^2 = n\beta_1^2 S_{xx} + (n-1)\sigma^2 - n\beta_1^2 S_{xx} - \sigma^2 = (n-2)\sigma^2.$$

In other words, we prove that

$$E(s^2) = E\left(\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y})^2\right) = \sigma^2.$$

Thus, s^2, the estimation of the error variance, is an unbiased estimator of the error variance σ^2 in the simple linear regression. Another view of choosing $n-2$ is that in the simple linear regression model there are n observations and two restrictions on these observations:

(1) $\sum_{i=1}^{n}(y_i-\hat{y}) = 0,$

(2) $\sum_{i=1}^{n}(y_i-\hat{y})x_i = 0.$

Hence the error variance estimation has $n-2$ degrees of freedom which is also the number of total observations − total number of the parameters in the model. We will see similar feature in the multiple linear regression.

2.4 Maximum Likelihood Estimation

The maximum likelihood estimates of the simple linear regression can be developed if we assume that the dependent variable y_i has a normal distribution: $y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$. The likelihood function for $(y_1, y_2, \cdots, y_n)$ is given by

$$\mathrm{L} = \prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{(-1/2\sigma^2)\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}.$$

The estimators of β_0 and β_1 that maximize the likelihood function L are equivalent to the estimators that minimize the exponential part of the likelihood function, which yields the same estimators as the least squares estimators of the linear regression. Thus, under the normality assumption of the error term the MLEs of β_0 and β_1 and the least squares estimators of β_0 and β_1 are exactly the same.

After we obtain b_1 and b_0, the MLEs of the parameters β_0 and b_1, we can compute the fitted value $\hat{y}_i$, and the likelihood function in terms of the fitted values.

$$\mathrm{L} = \prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{(-1/2\sigma^2)\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

We then take the partial derivative with respect to σ^2 in the log likelihood function $\log(L)$ and set it to zero:

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$$

The MLE of σ^2 is $\hat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Note that it is a biased estimate of σ^2, since we know that $s^2 = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is an unbiased estimate of the error variance σ^2. $\dfrac{n}{n-2}\hat{\sigma}^2$ is an unbiased estimate of σ^2. Note also that the $\hat{\sigma}^2$ is an asymptotically unbiased estimate of σ^2, which coincides with the classical theory of MLE.

2.5 Confidence Interval on Regression Mean and Regression Prediction

Regression models are often constructed based on certain conditions that must be verified for the model to fit the data well, and to be able to predict the response for a given regressor as accurate as possible. One of the main objectives of regression analysis is to use the fitted regression model to make prediction. Regression prediction is the calculated response value from the fitted regression model at data point which is not used in the model fitting. Confidence interval of the regression prediction provides a way of assessing the quality of prediction. Often the following regression prediction confidence intervals are of interest:

- A confidence interval for a single pint on the regression line.
- A confidence interval for a single future value of y corresponding to a chosen value of x.
- A confidence region for the regression line as a whole.

If a particular value of predictor variable is of special importance, a confidence interval for the corresponding response variable y at particular regressor x may be of interest.

A confidence interval of interest can be used to evaluate the accuracy of a single future value of y at a chosen value of regressor x. Confidence interval estimator for a future value of y provides confidence interval for an estimated value y at x with a desirable confidence level $1-\alpha$.

It is of interest to compare the above two different kinds of confidence interval. The second kind has larger confidence interval which reflects the less accuracy resulting from the estimation of a single future value of y rather than the mean value computed for the first kind confidence interval.

When the entire regression line is of interest, a confidence region can provide simultaneous statements about estimates of y for a number of values of the predictor variable x. i.e., for a set of values of the regressor the $100(1-\alpha)$ percent of the corresponding response values will be in this interval.

To discuss the confidence interval for regression line we consider the fitted value of the regression line at $x = x_0$, which is $\hat{y}(x_0) = b_0 + b_1 x_0$ and the mean value at $x = x_0$ is $E(\hat{y}|x_0) = \beta_0 + \beta_1 x_0$. Note that b_1 is independent of $\bar{y}$ we have

$$\begin{aligned}
\mathrm{Var}(\hat{y}(x_0)) &= \mathrm{Var}(b_0 + b_1 x_0) \\
&= \mathrm{Var}(\bar{y} - b_1(x_0 - \bar{x})) \\
&= \mathrm{Var}(\bar{y}) + (x_0 - \bar{x})^2 \mathrm{Var}(b_1) \\
&= \frac{1}{n}\sigma^2 + (x_0 - \bar{x})^2 \frac{1}{S_{xx}}\sigma^2 \\
&= \sigma^2 \Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\Big]
\end{aligned}$$

Replacing σ by s, the standard error of the regression prediction at x_0 is given by

$$s_{\hat{y}(x_0)} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

If $\varepsilon \sim N(0, \sigma^2)$ the $(1-\alpha)100\%$ of confidence interval on $E(\hat{y}|x_0) = \beta_0 + \beta_1 x_0$ can be written as

$$\hat{y}(x_0) \pm t_{\alpha/2,n-2}\ s\ \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

We now discuss confidence interval on the regression prediction. Denoting the regression prediction at x_0 by y_0 and assuming that y_0 is independent of $\hat{y}(x_0)$, where $y(x_0) = b_0 + b_1 x_0$, and $E(y - \hat{y}(x_0)) = 0$, we have

$$\mathrm{Var}\big(y_0 - \hat{y}(x_0)\big) = \sigma^2 + \sigma^2\Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\Big] = \sigma^2\Big[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\Big].$$

Under the normality assumption of the error term

$$\frac{y_0 - \hat{y}(x_0)}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0, 1).$$

Substituting σ with s we have

$$\frac{y_0 - \hat{y}(x_0)}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

Thus the $(1 - \alpha)100\%$ confidence interval on regression prediction y_0 can be expressed as

$$\hat{y}(x_0) \pm t_{\alpha/2,n-2}\ s\ \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

2.6 Statistical Inference on Regression Parameters

We start with the discussions on the total variance of regression model which plays an important role in the regression analysis. In order to partition the total variance $\sum_{i=1}^{n}(y_i - \bar{y})^2$, we consider the fitted regression equation $\hat{y}_i = b_0 + b_1 x_i$, where $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 = S_{xy}/S_{xx}$. We can write

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}[(\bar{y} - b_1\bar{x}) + b_1 x_i] = \frac{1}{n}\sum_{i=1}^{n}[\bar{y} + b_1(x_i - \bar{x})] = \bar{y}.$$

For the regression response y_i, the total variance is $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$. Note that the product term is zero and the total variance can be partitioned into two parts:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i-\hat{y})^2+(\hat{y}_i-\bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y})^2 = \mathrm{SS}_{Reg}+\mathrm{SS}_{Res} \\
&= \text{Variance explained by regression} + \text{Variance unexplained}
\end{aligned}$$

It can be shown that the product term in the partition of variance is zero:

$$\begin{aligned}
&\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) \qquad \text{(use the fact that } \sum_{i=1}^{n}(y_i-\hat{y}_i)=0\text{)} \\
&= \sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i) = \sum_{i=1}^{n}\big[b_0+b_1(x_i-\bar{x})\big](y_i-\hat{y}) \\
&= b_1\sum_{i=1}^{n}x_i(y_i-\hat{y}_i) = b_1\sum_{i=1}^{n}x_i[y_i-b_0-b_1(x_i-\bar{x})] \\
&= b_1\sum_{i=1}^{n}x_i\big[(y_i-\bar{y})-b_1(x_i-\bar{x})\big] \\
&= b_1\Big[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) - b_1\sum_{i=1}^{n}(x_i-\bar{x})^2\Big] \\
&= b_1[S_{xy}-b_1S_{xx}] = b_1[S_{xy}-(S_{xy}/S_{xx})S_{xx}] = 0
\end{aligned}$$

The degrees of freedom for SS_{Reg} and SS_{Res} are displayed in Table 2.2.

Table 2.2 Degrees of Freedom in Partition of Total Variance

SS_{Total}	=	SS_{Reg}	+	SS_{Res}
n-1	=	1	+	n-2

To test the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ it is needed to assume that $\varepsilon_i \sim N(0,\sigma^2)$. Table 2.3 lists the distributions of SS_{Reg}, SS_{Res} and SS_{Total} under the hypothesis H_0. The test statistic is given by

$$F = \frac{SS_{Reg}}{SS_{Res}/n-2} \sim F_{1,\, n-2},$$

which is a one-sided, upper-tailed F test. Table 2.4 is a typical regression Analysis of Variance (ANOVA) table.

Table 2.3 Distributions of Partition of Total Variance

SS	df	Distribution
SS_{Reg}	1	$\sigma^2\chi^2_1$
SS_{Res}	n-2	$\sigma^2\chi^2_{n-2}$
SS_{Total}	n-1	$\sigma^2\chi^2_{n-1}$

Table 2.4 ANOVA Table 1

Source	SS	df	MS	F
Regression	SS_{Reg}	1	$SS_{Reg}/1$	$F = \frac{MS_{Reg}}{s^2}$
Residual	SS_{Res}	n-2	s^2	
Total	SS_{Total}	n-1		

To test for regression slope β_1, it is noted that b_1 follows the normal distribution

$$b_1 \sim N\Big(\beta_1, \frac{\sigma^2}{SS_{xx}}\Big)$$

and

$$\Big(\frac{b_1-\beta_1}{s}\Big)\sqrt{S_{xx}} \sim t_{n-2},$$

which can be used to test $H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$. Similar approach can be used to test for the regression intercept. Under the normality assumption of the error term

$$b_0 \sim N\Big[\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\Big].$$

Therefore, we can use the following t test statistic to test $H_0 : \beta_0 = \beta_{00}$ versus $H_1 : \beta_0 \neq \beta_{00}$.

$$t = \frac{b_0 - \beta_0}{s\sqrt{1/n + (\bar{x}^2/S_{xx})}} \sim t_{n-2}$$

It is straightforward to use the distributions of b_0 and b_1 to obtain the $(1-\alpha)100\%$ confidence intervals of β_0 and β_1:

$$b_0 \pm t_{\alpha/2,n-2}\ s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},$$

and

$$b_1 \pm t_{\alpha/2,n-2}\ s\sqrt{\frac{1}{S_{xx}}}.$$

Suppose that the regression line pass through $(0,\ \beta_0)$. i.e., the y intercept is a known constant β_0. The model is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with known constant β_0. Using the least squares principle we can estimate β_1:

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2}.$$

Correspondingly, the following test statistic can be used to test for $H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$. Under the normality assumption on ε_i

$$t = \frac{b_1 - \beta_{10}}{s}\sqrt{\sum_{i=1}^{n} x_i^2} \sim t_{n-1}$$

Note that we only have one parameter for the fixed y-intercept regression model and the t test statistic has $n-1$ degrees of freedom, which is different from the simple linear model with 2 parameters.

The quantity R^2, defined as below, is a measurement of regression fit:

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SS_{Res}}{SS_{Total}}$$

Note that $0 \leq \mathrm{R}^2 \leq 1$ and it represents the proportion of total variation explained by regression model.

Quantity $\text{CV} = \dfrac{s}{\bar{y}} \times 100$ is called the coefficient of variation, which is also a measurement of quality of fit and represents the spread of noise around the regression line. The values of R^2 and CV can be found from Table 2.7, an ANOVA table generated by SAS procedure REG.

We now discuss simultaneous inference on the simple linear regression. Note that so far we have discussed statistical inference on β_0 and β_1 individually. The individual test means that when we test $H_0 : \beta_0 = \beta_{00}$ we only test this H_0 regardless of the values of β_1. Likewise, when we test $H_0 : \beta_1 = \beta_{10}$ we only test H_0 regardless of the values of β_0. If we would like to test whether or not a regression line falls into certain region we need to test the multiple hypothesis: $H_0 : \beta_0 = \beta_{00}, \beta_1 = \beta_{10}$ simultaneously. This falls into the scope of multiple inference. For the multiple inference on β_0 and β_1 we notice that

$$\begin{aligned}
&\big(\, b_0 - \beta_0,\, b_1 - \beta_1 \,\big) \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} b_0 - \beta_0 \\ b_1 - \beta_1 \end{pmatrix} \\
&\sim 2s^2 F_{2,n-2}.
\end{aligned}$$

Thus, the $(1-\alpha)100\%$ confidence region of the β_0 and β_1 is given by

$$\begin{aligned}
&\big(\, b_0 - \beta_0,\, b_1 - \beta_1 \,\big) \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} b_0 - \beta_0 \\ b_1 - \beta_1 \end{pmatrix} \\
&\leq 2s^2 F_{\alpha,2,n-2},
\end{aligned}$$

where $F_{\alpha,2,n-2}$ is the upper tail of the αth percentage point of the F-distribution. Note that this confidence region is an ellipse.

2.7 Residual Analysis and Model Diagnosis

One way to check performance of a regression model is through regression residual, i.e., $e_i = y_i - \hat{y}_i$. For the simple linear regression a scatter plot of e_i against x_i provides a good graphic diagnosis for the regression model. An evenly distributed residuals around mean zero is an indication of a good regression model fit.

We now discuss the characteristics of regression residuals if a regression model is misspecified. Suppose that the correct model should take the quadratic form:

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2 x_i^2 + \varepsilon_i$$

with $E(\varepsilon_i) = 0$. Assume that the incorrectly specified linear regression model takes the following form:

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i^*.$$

Then $\varepsilon_i^* = \beta_2 x_i^2 + \varepsilon_i^*$ which is unknown to the analyst. Now, the mean of the error for the simple linear regression is not zero at all and it is a function of x_i. From the quadratic model we have

$$b_0 = \bar{y} = \beta_0 + \beta_2 \bar{x}^2 + \bar{\varepsilon}$$

and

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1(x_i - \bar{x}) + \beta_2 x_i^2 + \varepsilon_i)}{S_{xx}}$$

$$b_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i^2}{S_{xx}} + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{S_{xx}}.$$

It is easy to know that

$$E(b_0) = \beta_0 + \beta_2 \bar{x}^2$$

and

$$E(b_1) = \beta_1 + \beta_2 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i^2}{S_{xx}}.$$

Therefore, the estimators b_0 and b_1 are biased estimates of β_0 and β_1. Suppose that we fit the linear regression model and the fitted values are given by $\hat{y}_i = b_0 + b_1(x_i - \bar{x})$, the expected regression residual is given by

$$\begin{aligned} E(e_i) = E(y_i - \hat{y}_i) &= \left[\beta_0 + \beta_1(x_i - \bar{x}) + \beta_2 x_i^2\right] - \left[E(b_0) + E(b_1)(x_i - \bar{x})\right] \\ &= \left[\beta_0 + \beta_1(x_i - \bar{x}) + \beta_2 x_i^2\right] - \left[\beta_0 + \beta_2 \bar{x}^2\right] \\ &\quad - \left[\beta_1 + \beta_2 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i^2}{S_{xx}}\right](x_i - \bar{x}) \\ &= \beta_2 \left[(x_i^2 - \bar{x}^2) - \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i^2}{S_{xx}}\right] \end{aligned}$$

If $\beta_2 = 0$ then the fitted model is correct and $E(y_i - \hat{y}_i) = 0$. Otherwise, the expected value of residual takes the quadratic form of x_i's. As a result, the plot of residuals against x_i's should have a curvature of quadratic appearance.

Statistical inference on regression model is based on the normality assumption of the error term. The least squares estimators and the MLEs of the regression parameters are exactly identical only under the normality assumption of the error term. Now, question is how to check the normality of the error term? Consider the residual $y_i - \hat{y}_i$: we have $E(y_i - \hat{y}_i) = 0$ and

$$\begin{aligned}\mathrm{Var}(y_i - \hat{y}_i) &= Var(y_i) + \mathrm{Var}(\hat{y}_i) - 2\mathrm{Cov}(y_i, \hat{y}_i) \\ &= \sigma^2 + \sigma^2\Big[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\Big] - 2\mathrm{Cov}(y_i, \bar{y} + b_1(x_i - \bar{x}))\end{aligned}$$

We calculate the last term

$$\begin{aligned}&\mathrm{Cov}(y_i,\ \bar{y} + b_1(x_i - \bar{x})) = \mathrm{Cov}(y_i,\ \bar{y}) + (x_i - \bar{x})\mathrm{Cov}(y_i, b_1) \\ &= \frac{\sigma^2}{n} + (x_i - \bar{x})\mathrm{Cov}(y_i,\ S_{xy}/S_{xx}) \\ &= \frac{\sigma^2}{n} + (x_i - \bar{x})\frac{1}{S_{xx}}\mathrm{Cov}\Big(y_i,\ \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\Big) \\ &= \frac{\sigma^2}{n} + (x_i - \bar{x})\frac{1}{S_{xx}}\mathrm{Cov}\Big(y_i,\ \sum_{i=1}^{n}(x_i - \bar{x})y_i\Big) = \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\sigma^2\end{aligned}$$

Thus, the variance of the residual is given by

$$\mathrm{Var}(e_i) = Var(y_i - \hat{y}_i) = \sigma^2\Big[1 - \Big(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\Big)\Big],$$

which can be estimated by

$$s_{e_i} = s\Big[1 - \Big(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\Big)\Big].$$

If the error term in the simple linear regression is correctly specified, i.e., error is normally distributed, the standardized residuals should behave like the standard normal random variable. Therefore, the quantile of the standardized residuals in the simple linear regression will be similar to the quantile of the standardized normal random variable. Thus, the plot of the

quantile of the standardized residuals versus the normal quantile should follow a straight line in the first quadrant if the normality assumption on the error term is correct. It is usually called the normal plot and has been used as a useful tool for checking the normality of the error term in simple linear regression. Specifically, we can

(1) Plot ordered residual $\frac{y_i - \hat{y}_i}{s}$ against the normal quantile $Z\left(\frac{i-0.375}{n+0.25}\right)$

(2) Plot ordered standardized residual $\frac{y_i - \hat{y}_i}{s_{e_i}}$ against the normal quantile $Z\left(\frac{i-0.375}{n+0.25}\right)$

2.8 Example

The SAS procedure REG can be used to perform regression analysis. It is convenient and efficient. The REG procedure provides the most popular parameter estimation, residual analysis, regression diagnosis. We present the example of regression analysis of the density and stiffness data using SAS.

```
data example1;
input density  stiffness @@;
datalines;
 9.5  14814   8.4    17502   9.8  14007  11.0  19443   8.3   7573
 9.9  14191   8.6     9714   6.4   8076   7.0   5304   8.2  10728
17.4  43243  15.0    25319  15.2  28028  16.4  41792  16.7  49499
15.4  25312  15.0    26222  14.5  22148  14.8  26751  13.6  18036
25.6  96305  23.4   104170  24.4  72594  23.3  49512  19.5  32207
21.2  48218  22.8    70453  21.7  47661  19.8  38138  21.3  53045
;
proc reg data=example1 outest=out1 tableout;
model stiffness=density/all;
run;

ods rtf file="C:\Example1_out1.rtf";
proc print data=out1;
title "Parameter Estimates and CIs";
run;
ods rtf close;
```

```
*Trace ODS to find out the names of the output data sets;
ods trace on;
ods show;

ods rtf file="C:\Example1_out2.rtf";
proc reg data=Example1 alpha=0.05;
model stiffness=density;
ods select Reg.MODEL1.Fit.stiffness.ANOVA;
ods select Reg.MODEL1.Fit.stiffness.FitStatistics;
ods select Reg.MODEL1.Fit.stiffness.ParameterEstimates;
ods rtf close;

proc reg data=Example1;
model stiffness=density;
output out=out3 p=yhat r=yresid student=sresid;
run;

ods rtf file="C:\Example1_out3.rtf";
proc print data=out3;
title "Predicted Values and Residuals";
run;
ods rtf close;
```

The above SAS code generate the following output tables 2.5, 2.6, 2.7, 2.8, and 2.9.

Table 2.5 Confidence Intervals on Parameter Estimates

Obs	MODEL	TYPE	DEPVAR	RMSE	Intercept	density
1	Model1	Parms	stiffness	11622.44	-25433.74	3884.98
2	Model1	Stderr	stiffness	11622.44	6104.70	370.01
3	Model1	T	stiffness	11622.44	-4.17	10.50
4	Model1	P-value	stiffness	11622.44	0.00	0.00
5	Model1	L95B	stiffness	11622.44	-37938.66	3127.05
6	Model1	U95B	stiffness	11622.44	-12928.82	4642.91

Data Source: density and stiffness data

The following is an example of SAS program for computing the confidence band of regression mean, the confidence band for regression predic-

Table 2.6 ANOVA Table 2

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	1	14891739363	14891739363	110.24	<.0001
Error	28	3782270481	135081089		
Corrected Total	29	18674009844			

Data Source: density and stiffness data

Table 2.7 Regression Table

Root MSE	11622.00	R-Square	0.7975
Dependent Mean	34667.00	Adj R-Sq	0.7902
Coeff Var	33.53		

Data Source: density and stiffness data

Table 2.8 Parameter Estimates of Simple Linear Regression

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > $\|t\|$
Intercept	1	-25434.00	6104.70	-4.17	0.0003
density	1	3884.98	370.01	10.50	< .0001

Data Source: density and stiffness data

tion, and probability plot (QQ-plot and PP-plot).

```
data Example2;
input density  stiffness @@;
datalines;
9.5   14814  8.4  17502   9.8  14007  11    19443   8.3  7573
9.9   14191  8.6   9714   6.4   8076  7      5304   8.2  10728
17.4  43243  15   25319  15.2  28028  16.4  41792  16.7  49499
15.4  25312  15   26222  14.5  22148  14.8  26751  13.6  18036
25.6  96305  23.4 104170 24.4  72594  23.3  49512  19.5  32207
21.2  48218  22.8  70453 21.7  47661  19.8  38138  21.3  53045
;
across=1 cborder=red offset=(0,0)
shape=symbol(3,1) label=none value=(height=1);
symbol1 c=black value=- h=1;
symbol2 c=red;
```

Table 2.9 Table for Fitted Values and Residuals

Obs	density	stiffness	yhat	yresid
1	9.5	14814	11473.53	3340.47
2	8.4	17502	7200.06	10301.94
3	9.8	14007	12639.02	1367.98
4	11	19443	17300.99	2142.01
5	8.3	7573	6811.56	761.44
6	9.9	14191	13027.52	1163.48
7	8.6	9714	7977.05	1736.95
8	6.4	8076	-569.90	8645.90
9	7.0	5304	1761.09	3542.91
10	8.2	10728	6423.06	4304.94
11	17.4	43243	42164.84	1078.16
12	15.0	25319	32840.89	-7521.89
13	15.2	28028	33617.89	-5589.89
14	16.4	41792	38279.86	3512.14
15	16.7	49499	39445.35	10053.65
16	15.4	25312	34394.89	-9082.89
17	15.0	26222	32840.89	-6618.89
18	14.5	22148	30898.41	-8750.41
19	14.8	26751	32063.90	-5312.90
20	13.6	18036	27401.93	-9365.93
21	25.6	96305	74021.64	22283.36
22	23.4	104170	65474.69	38695.31
23	24.4	72594	69359.67	3234.33
24	23.3	49512	65086.19	-15574.19
25	19.5	32207	50323.28	-18116.28
26	21.2	48218	56927.74	-8709.74
27	22.8	70453	63143.70	7309.30
28	21.7	47661	58870.23	-11209.23
29	19.8	38138	51488.78	-13350.78
30	21.3	53045	57316.24	-4271.24

Data Source: density and stiffness data

```
symbol3 c=blue;
symbol4 c=blue;
proc reg data=Example2;
     model density=stiffness /noprint p r;
     output out=out p=pred r=resid LCL=lowpred
            UCL=uppred LCLM=lowreg UCLM=upreg;
run;
ods rtf file="C:\Example2.rtf";
ods graphics on;
title "PP Plot";
```

```
plot npp.*r./caxis=red ctext=blue nostat cframe=ligr;
run;
title "QQ Plot";
plot r.*nqq. /noline  mse
     caxis=red ctext=blue cframe=ligr;
run;

*Compute confidence band of regression mean;
plot  density*stiffness/conf caxis=red ctext=blue
      cframe=ligr legend=legend1;
      run;

*Compute confidence band of regression prediction;
plot density*stiffness/pred caxis=red ctext=blue
     cframe=ligr legend=legend1;
run;
ods graphics off;
ods rtf close;
quit;
```

The regression scatterplot, residual plot, 95% confidence bands for regression mean and prediction are presented in Fig. 2.1.

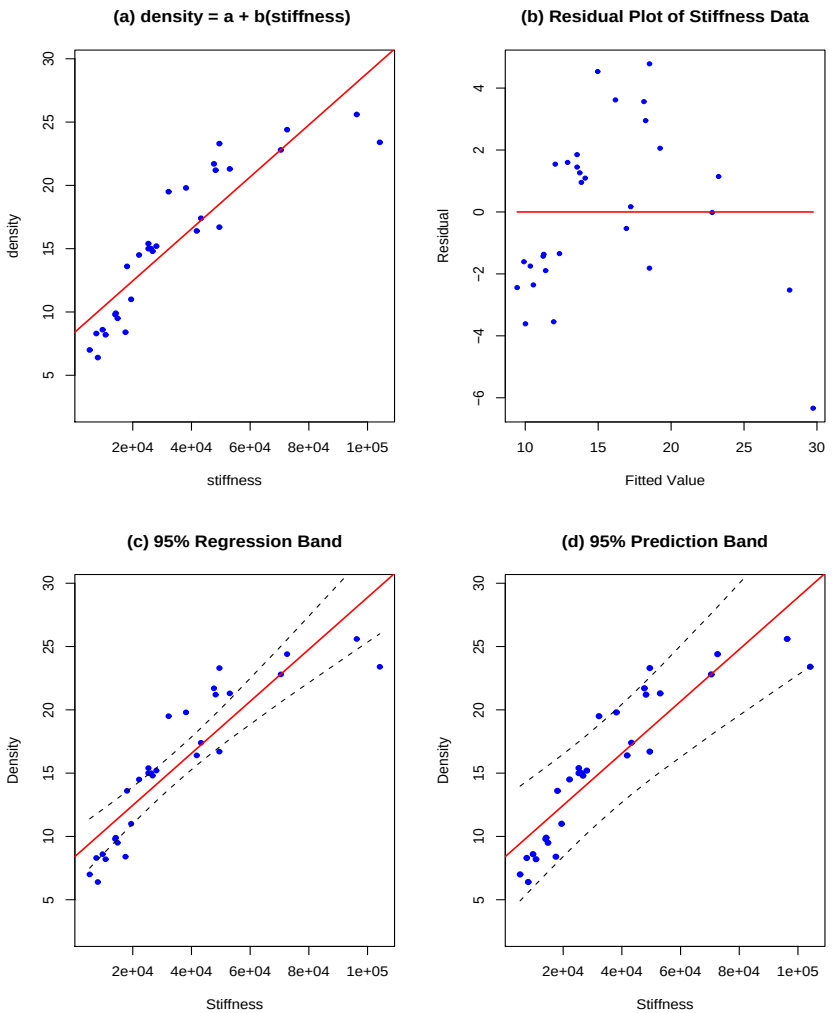

Fig. 2.1 (a) Regression Line and Scatter Plot. (b) Residual Plot, (c) 95% Confidence Band for Regression Mean. (d) 95% Confidence Band for Regression Prediction.

The Q-Q plot for regression model density=$\beta_0 + \beta_1$ stiffness is presented in Fig. 2.2.

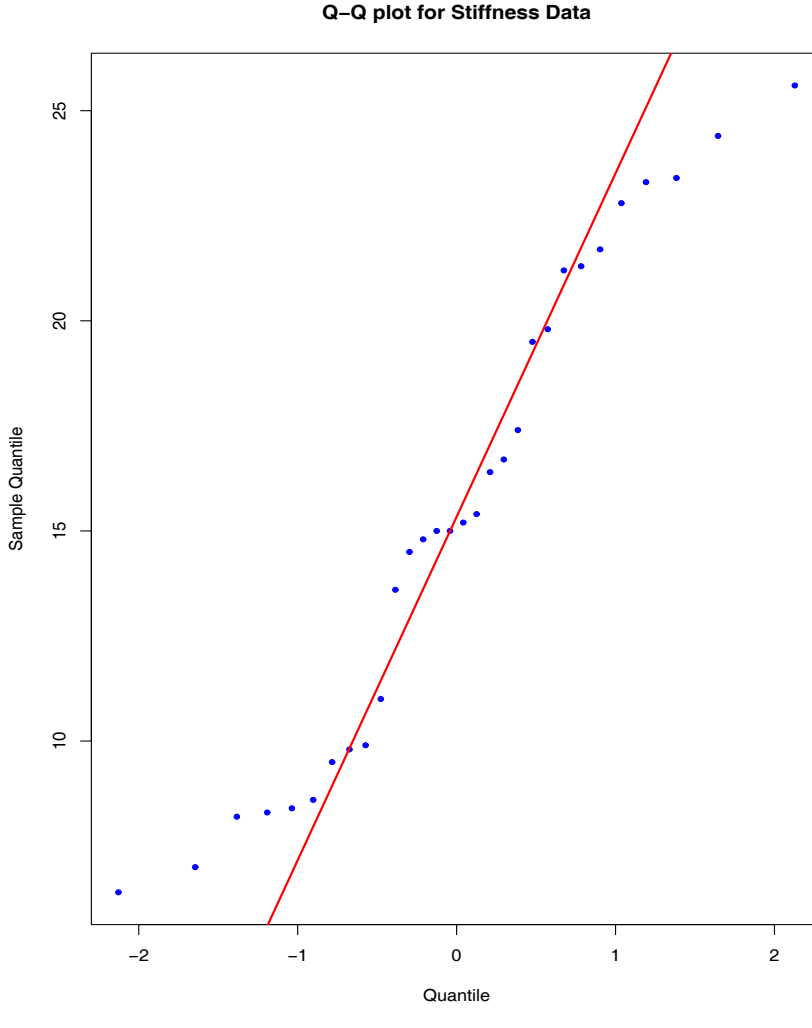

Fig. 2.2 Q-Q Plot for Regression Model density=$\beta_0 + \beta_1$ stiffness + ε.

Problems

1. Consider a set of data $(x_i,\ y_i)$, $i = 1, 2, \cdots, n$, and the following two regression models:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon, \qquad (i = 1, 2, \cdots, n), \quad \text{Model A}$$
$$y_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \varepsilon, \quad (i = 1, 2, \cdots, n), \quad \text{Model B}$$

   Suppose both models are fitted to the same data. Show that

$$SS_{Res,\ A} \geq SS_{Res,\ B}$$

   If more higher order terms are added into the above Model B, i.e.,

$$y_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \gamma_3 x_i^3 + \cdots + \gamma_k x_i^k + \varepsilon, \quad (i = 1, 2, \cdots, n),$$

   show that the inequality $SS_{Res,\ A} \geq SS_{Res,\ B}$ still holds.
2. Consider the zero intercept model given by

$$y_i = \beta_1 x_i + \varepsilon_i, \quad (i = 1, 2, \cdots, n)$$

   where the ε_i's are independent normal variables with constant variance σ^2. Show that the $100(1-\alpha)\%$ confidence interval on $E(y|x_0)$ is given by

$$b_1 x_0 + t_{\alpha/2,\ n-1} s \sqrt{\frac{x_0^2}{\sum_{i=1}^n x_i^2}}$$

   where $s = \sqrt{\sum_{i=1}^{n} (y_i - b_1 x_i)/(n-1)}$ and $b_1 = \dfrac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}$.
3. Derive and discuss the $(1-\alpha)100\%$ confidence interval on the slope β_1 for the simple linear model with zero intercept.
4. Consider the fixed zero intercept regression model

$$y_i = \beta_1 x_i + \varepsilon_i, \quad (i = 1, 2, \cdots, n)$$

   The appropriate estimator of σ^2 is given by

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

   Show that s^2 is an unbiased estimator of σ^2.

Table 2.10 Data for Two Parallel Regression Lines

x	y
x_1	y_1
$\vdots$	$\vdots$
x_{n_1}	y_{n_1}
x_{n_1+1}	y_{n_1+1}
$\vdots$	$\vdots$
$x_{n_1+n_2}$	$y_{n_1+n_2}$

5. Consider a situation in which the regression data set is divided into two parts as shown in Table 2.10.
The regression model is given by

$$y_i = \begin{cases} \beta_0^{(1)} + \beta_1 x_i + \varepsilon_i, & i = 1, 2, \cdots, n_1; \\ \\ \beta_0^{(2)} + \beta_1 x_i + \varepsilon_i, & i = n_1 + 1, \cdots, n_1 + n_2. \end{cases}$$

In other words, there are two regression lines with common slope. Using the centered regression model

$$y_i = \begin{cases} \beta_0^{(1*)} + \beta_1 (x_i - \bar{x}_1) + \varepsilon_i, & i = 1, 2, \cdots, n_1; \\ \\ \beta_0^{(2*)} + \beta_1 (x_i - \bar{x}_2) + \varepsilon_i, & i = n_1 + 1, \cdots, n_1 + n_2, \end{cases}$$

where $\bar{x}_1 = \sum_{i=1}^{n_1} x_i/n_1$ and $\bar{x}_2 = \sum_{i=n_1+1}^{n_1+n_2} x_i/n_2$. Show that the least squares estimate of β_1 is given by

$$b_1 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)y_i + \sum_{i=n_1+1}^{n_1+n_2}(x_i - \bar{x}_2)y_i}{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{i=n_1+1}^{n_1+n_2}(x_i - \bar{x}_2)^2}$$

6. Consider two simple linear models

$$Y_{1j} = \alpha_1 + \beta_1 x_{1j} + \varepsilon_{1j},\ j = 1, 2, \cdots, n_1$$

and

$$Y_{2j} = \alpha_2 + \beta_2 x_{2j} + \varepsilon_{2j},\ j = 1, 2, \cdots, n_2$$

Assume that $\beta_1 \neq \beta_2$ the above two simple linear models intersect. Let x_0 be the point on the x-axis at which the two linear models intersect. Also assume that ε_{ij} are independent normal variable with a variance σ^2. Show that

(a). $x_0 = \dfrac{\alpha_1 - \alpha_2}{\beta_1 - \beta_2}$

(b). Find the maximum likelihood estimates (MLE) of x_0 using the least squares estimators $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

(c). Show that the distribution of Z, where

$$Z = (\hat{\alpha}_1 - \hat{\alpha}_2) + x_0(\hat{\beta}_1 - \hat{\beta}_2),$$

is the normal distribution with mean 0 and variance $A^2\sigma^2$, where

$$A^2 = \frac{\sum x_{1j}^2 - 2x_0 \sum x_{1j} + x_0^2 n_1}{n_1 \sum (x_{1j} - \bar{x}_1)^2} + \frac{\sum x_{2j}^2 - 2x_0 \sum x_{2j} + x_0^2 n_2}{n_2 \sum (x_{2j} - \bar{x}_2)^2}.$$

(d). Show that $U = N\hat{\sigma}^2/\sigma^2$ is distributed as $\chi^2(N)$, where $N = n_1 + n_2 - 4$.

(e). Show that U and Z are independent.

(f). Show that $W = Z^2/A^2\hat{\sigma}^2$ has the F distribution with degrees of freedom 1 and N.

(g). Let $S_1^2 = \sum (x_{1j} - \bar{x}_1)^2$ and $S_2^2 = \sum (x_{2j} - \bar{x}_2)^2$, show that the solution of the following quadratic equation about x_0, $q(x_0) = ax_0^2 + 2bx_0 + c = 0$,

$$\Big[(\hat{\beta}_1 - \hat{\beta}_2)^2 - \Big(\frac{1}{S_1^2} + \frac{1}{S_2^2}\Big)\hat{\sigma}^2 F_{\alpha,1,N}\Big] x_0^2$$
$$+ 2\Big[(\hat{\alpha}_1 - \hat{\alpha}_2)(\hat{\beta}_1 - \hat{\beta}_2) + \Big(\frac{\bar{x}_1}{S_1^2} + \frac{\bar{x}_2}{S_2^2}\Big)\hat{\sigma}^2 F_{\alpha,1,N}\Big] x_0$$
$$+ \Big[(\hat{\alpha}_1 - \hat{\alpha}_2)^2 - \Big(\frac{\sum x_{1j}^2}{n_1 S_1^2} + \frac{\sum x_{2j}^2}{n_2 S_2^2}\Big)\hat{\sigma}^2 F_{\alpha,1,N}\Big] = 0.$$

Show that if $a \geq 0$ and $b^2 - ac \geq 0$, then $1 - \alpha$ confidence interval on x_0 is

$$\frac{-b - \sqrt{b^2 - ac}}{a} \leq x_0 \leq \frac{-b + \sqrt{b^2 - ac}}{a}.$$

7. Observations on the yield of a chemical reaction taken at various temperatures were recorded in Table 2.11:

(a). Fit a simple linear regression and estimate β_0 and β_1 using the least squares method.

(b). Compute 95% confidence intervals on $E(y|x)$ at 4 levels of temperatures in the data. Plot the upper and lower confidence intervals around the regression line.

Table 2.11 Chemical Reaction Data

temperature (C^0)	yield of chemical reaction (%)
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4
150	77.4

Data Source: Raymond H. Myers, Classical and Modern Regression Analysis With Applications, P77.

(c). Plot a 95% confidence band on the regression line. Plot on the same graph for part (b) and comment on it.

8. The study "Development of LIFETEST, a Dynamic Technique to Assess Individual Capability to Lift Material" was conducted in Virginia Polytechnic Institute and State University in 1982 to determine if certain static arm strength measures have influence on the "dynamic lift" characteristics of individual. 25 individuals were subjected to strength tests and then were asked to perform a weight-lifting test in which weight was dynamically lifted overhead. The data are in Table 2.12:

   (a). Find the linear regression line using the least squares method.
   (b). Define the joint hypothesis $H_0 : \beta_0 = 0,\ \beta_1 = 2.2$. Test this hypothesis problem using a 95% joint confidence region and β_0 and β_1 to draw your conclusion.
   (c). Calculate the studentized residuals for the regression model. Plot the studentized residuals against x and comment on the plot.

Table 2.12 Weight-lifting Test Data

Individual	Arm Strength (x)	Dynamic Lift (y)
1	17.3	71.4
2	19.5	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.1	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

Data Source: Raymond H. Myers, Classical and Modern Regression Analysis With Applications, P76.