Chapter 2

Simple Linear Regression

Simple linear regression is typically used to model the relationship between two variables (a single predictor X and a response Y) as a straight line so that given a specific value of X, we can predict the value of Y. Mathematically, the regression of Y on X is

$$\mathrm{E}[Y|X = x],$$

which is the expected value of Y at a specific value x of X. For example, if X = Age of propellant and Y = Shear strength, then the regression of Y on X represents the average shear strength on a given propellant age. In simple linear regression model,

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

2.1 Description of the Data and Model

Data

Y (Response variable)	X (Predictor)
y_1	x_1
y_2	x_2
$\vdots$	$\vdots$
y_n	x_n

That is, we have n pairs of data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$.

Simple linear regression model

The simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where the unknown parameters β_0 and β_1 determine the intercept and slope of a specific straight line. The slope β_1 is the change in the mean response produced by a unit change in X. The intercept β_0 is the mean response when $x = 0$ if the range of data on x includes 0. If the range of x does not include zero, then β_0 has no practical interpretation.

The regression model in a data setting is written as

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

where the intercept β_0 and the slope β_1 are unknown constants and ϵ_i are random errors. The errors are assumed to have mean zero and unknown variance σ^2. Additionally we assume that the errors are uncorrelated.

Notes:

1. There will almost certainly be a variation in Y due to random phenomenon that cannot be controlled or explained. These unexplained variation is called random error.

2. The assumption $\text{Var}(\epsilon) = \sigma^2$ means $\text{Var}(Y|X = x) = \sigma^2$.

2.2 Estimation of the Parameters

How to estimate β_0, β_1 and σ^2 from given data (a random sample of n pairs from the population)?

2.2.1 Estimation of β_0 and β_1

How to estimate β_0 and β_1?

This can be achieved by finding the straight line as close as possible to the observations. For this, we will use the least-squares criterion.

Least-squares estimation

The estimates of β_0 and β_1, say $\hat{\beta}_0$ and $\hat{\beta}_1$, are defined by

$$(\hat{\beta}_0, \hat{\beta}_1) \equiv \text{argmin}_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by finding b_0 and b_1 such that minimizes

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Thus, $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\hat{\beta}_0 = b_0, \ \hat{\beta}_1 = b_1,$$

where b_0 and b_1 must satisfy

$$\begin{cases} \frac{\partial S(b_0,b_1)}{\partial b_0} = 0 \\ \frac{\partial S(b_0,b_1)}{\partial b_1} = 0 \end{cases} \quad \Longrightarrow \quad \begin{cases} \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0 \\ \sum_{i=1}^n x_i(y_i - b_0 - b_1 x_i) = 0. \end{cases}$$

These equations can be written as

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases}$$

This last two equations are called the **normal equations**. Consequently, the estimates of the intercept and slope are given by

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

and

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$ and $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$, $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$.

Notations

- Fitted value or Predicted value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n.$$

- Residual:

$$\begin{aligned} e_i &= y_i - \hat{y}_i, \quad i = 1, \ldots, n, \\ &= \text{difference between the observed value and the fitted value} \end{aligned}$$

Then, we notice that

1. $\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$
2. $\sum_{i=1}^n x_i e_i = 0$
3. $\sum_{i=1}^n \hat{y}_i e_i = 0$
4. The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$ of the data.

2.2.2 Estimation of σ^2

How to estimate σ^2 from the given data?

For this, we will use the residuals to get an unbiased estimate of σ^2. First, observe that

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i),\ i = 1, \ldots, n.$$

However, β_0 and β_1 are unknown and so all we can do is to estimate these errors by replacing β_0 and β_1 by their respective least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ giving the residuals

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i),\ i = 1, \ldots, n.$$

Next, we can observe that

$$\mathrm{E}\left[SS_{Res}\right] = (n-2)\sigma^2,$$

where the residual sum of squares $SS_{Res} = \sum_{i=1}^{n} e_i^2$. Then, it can be shown that

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} := MS_{Res}$$

is an unbiased estimate of σ^2. The quantity MS_{Res} is called the residual mean square.

Note: The divisor in $\hat{\sigma}^2$ is $n-2$ since we have estimated two parameters, namely β_0 and β_1.

2.2.3 Example

Consider the following data. Fit a simple linear regression model.

x	Y
0	12.22
2	14.96
4	21.23
6	21.15
8	23.67
10	28.21

2.2.4 Example

Let us revisit the rocket propellant data (Ex 2.1 in textbook).

SAS Program:

```
option ls=90 ps=75;
title 'Rocket Propellant Data';
proc plot data=rocket;
plot Strength*Age;
run;
proc reg data=rocket;
model Strength = Age;
plot Strength*Age;  /* The least-squares fitted line */
run; quit;
```

Output:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2627.82236	44.18391	59.47	<.0001
Age	1	-37.15359	2.88911	-12.86	<.0001

Output Statistics

Obs	Dependent Variable	Predicted Value	Residual
1	2159	2052	106.7583
2	1678	1745	-67.2746
3	2316	2331	-14.5936
4	2061	1996	65.0887
5	2208	2423	-215.9776
6	1708	1922	-213.6041
7	1785	1736	48.5638
8	2575	2535	40.0616
9	2358	2349	8.7296
10	2257	2219	37.5671
11	2165	2145	20.3743

12	2400	2488	-88.9464
13	1780	1699	80.8174
14	2337	2266	71.1752
15	1765	1810	-45.1434
16	2054	1959	94.4423
17	2414	2405	9.4992
18	2201	2163	37.0975
19	2654	2554	100.6848
20	1754	1829	-75.3202

Sum of Residuals	0
Sum of Squared Residuals	166255
Predicted Residual SS (PRESS)	205944

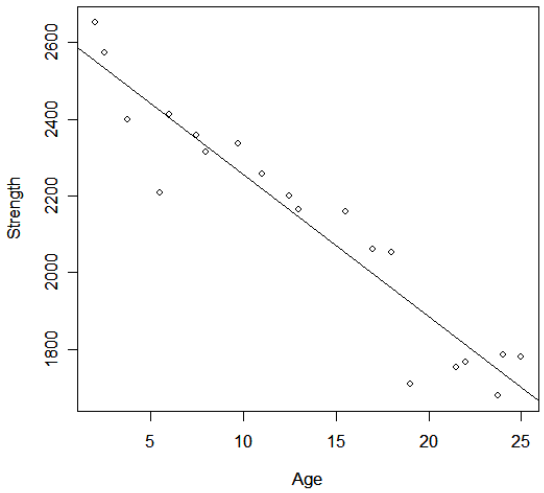

Figure 2.1: A plot of the rocket propellant data with the least-squares regression line.

Figure 2.1 shows a scatter plot of the propellant data with the least-squares regression line. The least-squares fit is

$$\hat{Y} = 2627.8224 - 37.1536X$$

or

$$\widehat{\text{Strength}} = 2627.8224 - 37.1536 * \text{Age}.$$

We interpret the slope -37.1536 as the average weekly decrease in propellant shear strength due to the age of the propellant. The intercept 2627.8224 represents the shear strength in a batch propellant immediately following manufacture.

The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = \frac{166255}{18} = 9236.389.$$

Remember that this estimate of σ^2 is model dependent.

2.2.5 Properties of estimates

Recall that the least-squares estimate of β_1 is given by

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and the least-squares estimate of β_0 is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Since $\sum_{i=1}^n (x_i - \bar{x}) = 0$, we find that

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

Thus, we can rewrite $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n c_i \epsilon_i,$$

where $c_i = (x_i - \bar{x})/S_{xx}$ for $i = 1, \ldots, n$. Similarly, we can rewrite $\hat{\beta}_0$ as

$$\hat{\beta}_0 = \beta_0 + \sum_{i=1}^n k_i \epsilon_i,$$

where $k_i = \left(\frac{1}{n} - c_i \bar{x}\right) = \left(\frac{1}{n} - \frac{x_i - \bar{x}}{S_{xx}} \bar{x}\right)$ for $i = 1, \ldots, n$. Then, it can be shown that

- $\mathrm{E}[\hat{\beta}_1] = \beta_1,\ \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}},\ \mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0,$
- $\mathrm{E}[\hat{\beta}_0] = \beta_0,\ \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right),$
- The least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is the best linear unbiased estimators (BLUE) of β_0 and β_1 (Gauss-Markov theorem).
- $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$.

2.2.6 Estimation by maximum likelihood

The method of least squares can be used to estimate the parameters in a linear regression model regardless of the form of the distribution of the error ϵ. Least squares produces best linear unbiased estimators of the intercept and slope. If the form of the distribution of the errors is known, the method of maximum likelihood can be used to estimate parameters.

Consider the data $(y_i, x_i), i = 1, \ldots, n$. Assume that the errors $\epsilon_1, \ldots, \epsilon_n$ are independently and normally distributed with mean 0 and variance σ^2. Then, the observations $y_1, \ldots, y_n$ are independently and normally distributed with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. For the simple linear regression model with normal errors, the likelihood function is

$$L(\beta_0, \beta_1, \sigma^2 | x_1, \ldots, x_n, y_1, \ldots, y_n) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - \frac{n}{2}\log(2\pi\sigma^2)\right].$$

Thus, the maximum likelihood estimates of $\beta_0, \beta_1, \sigma^2$ are given by

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}, \quad \tilde{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n}.$$

Notice that the maximum-likelihood estimates of the intercept and slope are identical to the least-squares estimates. Also, $\tilde{\sigma}^2$ is a biased estimate of σ^2; i.e., $\mathrm{E}[\tilde{\sigma}^2] \neq \sigma^2$. In fact, $\mathrm{E}[\tilde{\sigma}^2] = \frac{n-2}{n}\sigma^2$.

2.3 Inference About the Slope and Intercept

Recall that the assumptions for the simple linear regression model are

A1. Y is related to X by the simple linear regression model

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

A2. The errors $\epsilon_1, \ldots, \epsilon_n$ have a common variance σ^2; i.e., $\mathrm{Var}(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$.

A3. The errors $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated; i.e., $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

Additional assumption to make inferences about the regression model

A4. The errors are independently and normally distributed with a mean of 0 and variance σ^2; i.e.,

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n.$$

Methods for checking these assumptions will be considered in Chapter 5. We wish to perform such checks before model fitting but this is usually impossible so that we are force to conduct such checks after model fitting.

2.3.1 Inference about the slope

Under the assumption (A4) it can be shown that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Standardizing the above equation gives

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1).$$

When σ^2 is unknown (as is usually the case), σ^2 is replaced by $\hat{\sigma}^2 = MS_{Res}$. Additionally, it can be shown that under the assumption (A4)

- $\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$,
- $\hat{\beta}_1$ and MS_{Res} are independent.

Thus, we have the following distribution result under the assumption

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}} \sim t_{n-2}.$$

Hypothesis test on the slope

Suppose that we wish to test the hypothesis that the slope equals a constant, say β_{10}. The appropriate hypotheses are

$$H_0 : \beta_1 = \beta_{10} \quad \text{versus} \quad H_1 : \beta_1 \neq \beta_{10}.$$

When H_0 is true, the test statistic

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1).$$

If σ^2 was known then we could use Z_0 to test hypotheses. When σ^2 is unknown, σ^2 is replaced by $\hat{\sigma}^2 = MS_{Res}$. Thus, the test statistic is

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \sim t_{n-2}$$

when H_0 is true. We reject the null hypothesis at level α if $|t_0| > t_{\alpha/2,n-2}$ or $P - value < \alpha$. Here $t_{\alpha/2,n-2}$ is the $100(1-\alpha/2)th$ quantile of the t-distribution with $n-2$ degrees of freedom.

Note: The denominator of the test statistic, t_0, is called the estimated standard error of the slope. That is,

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}}.$$

Confidence interval (CI) for the slope

We have shown that

$$\frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}.$$

A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\left(\hat{\beta}_1 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1), \hat{\beta}_1 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1)\right).$$

Note: The larger the confidence interval coefficient $(1-\alpha)$ is, the wider the CI.

2.3.2 Inference about the intercept

Under the assumption (A4) it can be shown that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right).$$

Standardizing the above equation gives

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1).$$

When σ^2 is unknown replacing σ^2 by MS_{Res} results in

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

Hypothesis test on the intercept

We are interested in testing

$$H_0 : \beta_0 = \beta_{00} \quad \text{versus} \quad H_1 : \beta_0 \neq \beta_{00}.$$

If σ^2 was known then we could use

$$Z_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

to test hypotheses for β_0. When σ^2 is unknown, the test statistic

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$$

when H_0 is true. The null hypothesis is rejected at level α if $|t_0| > t_{\alpha/2,n-2}$ or $P-value < \alpha$. The estimated standard error of the intercept is

$$\text{se}(\hat{\beta}_0) = \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

Confidence interval for the intercept

We have shown that

$$\frac{\hat{\beta}_0 - \beta_0}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}.$$

A $100(1-\alpha)\%$ confidence interval for β_0 is given by

$$\left(\hat{\beta}_0 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0), \hat{\beta}_0 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0)\right).$$

2.3.3 Example

Revisit the artificial data.

2.3.4 Example

Revisit the rocket propellant data. We will test for significance of the slope and intercept in the rocket propellant regression model and we will construct the interval estimation of the slope and intercept.

SAS Program:

```
proc reg data=rocket;
model Strength = Age / clb;  /* CI for regression coefficients */
run; quit;
```

Output:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	2627.82236	44.18391	59.47	<.0001	2534.99540	2720.64931
Age	1	-37.15359	2.88911	-12.86	<.0001	-43.22338	-31.08380

Since $S_{xx} = 1106.559$ and $MS_{Res} = \hat{\sigma}^2 = 9236.389$, the standard error of the slope is

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = 2.8891.$$

Thus, the test statistic for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is

$$t_0 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{-37.1536}{2.8891} = -12.86.$$

If we choose $\alpha = .05$, $t_{.025,18} = 2.101$ from Table A.3. Also, we observe that $P - value = P(|t_{n-2}| > |t_0| = 12.86) < .0001$ from SAS Output. Thus, we reject $H_0 : \beta_1 = 0$ at significance level .05. We conclude that there is a statistically significant linear relationship between shear strength and the age of propellant.
We now wish to estimate the slope using a confidence interval. The 95% CI for β_1 is

$$\hat{\beta}_1 \pm t_{.025,18}\text{se}(\hat{\beta}_1) = -37.1536 \pm (2.101)(2.8891) = (-43.2234, -31.0838).$$

2.4 Confidence Interval for the Mean Response

A major use of a regression model is to estimate the mean response at a given value of the predictor. Let us recall that the mean response at $X = x_0$ is given by

$$\mu_0 = \text{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0.$$

An estimate of $\text{E}[Y|X = x_0]$ is

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Under the assumptions stated previously, it can be shown that

1. $\mathrm{E}[\hat{\mu}_0] = \beta_0 + \beta_1 x_0 = \mathrm{E}[Y|X = x_0]$
2. $\mathrm{Var}(\hat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)$
3. $\hat{\mu}_0 \sim N\left(\mu_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$.

Replacing σ^2 by MS_{Res} results in

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

Thus, a $100(1-\alpha)\%$ confidence interval for $\mu_0 = \mathrm{E}[Y|X = x_0]$ is given by

$$\hat{\mu}_0 \pm t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Note that the width of the CI increases as $|x_0 - \bar{x}|$ increases.

Example

Revisit the artificial data.

Example

Revisit the rocket propellant data. Suppose that we wish to estimate the mean shear strength of the propellant bond in a rocket motor made from a batch of sustainer propellant that is 10 weeks old.

SAS Program:

```
/* If you wish to estimate the mean response and/or predict Y at particular
 point whose corresponding response value is not observed, then you must add
 new data to the given data set. */
  data new;
  input Age Strength;
  datalines;
  10 .
  run;
```

```
data rocket2;
set rocket new;
run;
proc print data=rocket2;
run;
proc reg data=rocket2;
/* CI for mean response and PI for individual observations */
model Strength = Age / clm cli;
/* Confidence and prediction bands */
plot Strength*Age / conf pred;
run; quit;
```

Output:

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	2159	2052	22.3597	2005	2099	106.7583
2	1678	1745	36.9114	1668	1823	-67.2746
3	2316	2331	26.4924	2275	2386	-14.5936
4	2061	1996	23.9220	1946	2046	65.0887
5	2208	2423	31.2701	2358	2489	-215.9776
6	1708	1922	26.9647	1865	1979	-213.6041
7	1785	1736	37.5010	1657	1815	48.5638
8	2575	2535	38.0356	2455	2615	40.0616
9	2358	2349	27.3623	2292	2407	8.7296
10	2257	2219	22.5479	2172	2267	37.5671
11	2165	2145	21.5155	2100	2190	20.3743
12	2400	2488	35.1152	2415	2562	-88.9464
13	1780	1699	39.9031	1615	1783	80.8174
14	2337	2266	23.8903	2215	2316	71.1752
15	1765	1810	32.9326	1741	1880	-45.1434
16	2054	1959	25.3245	1906	2012	94.4423
17	2414	2405	30.2370	2341	2468	9.4992
18	2201	2163	21.6340	2118	2209	37.0975
19	2654	2554	39.2360	2471	2636	100.6848
20	1754	1829	31.8519	1762	1896	-75.3202
21	.	2256	23.5837	2207	2306	.

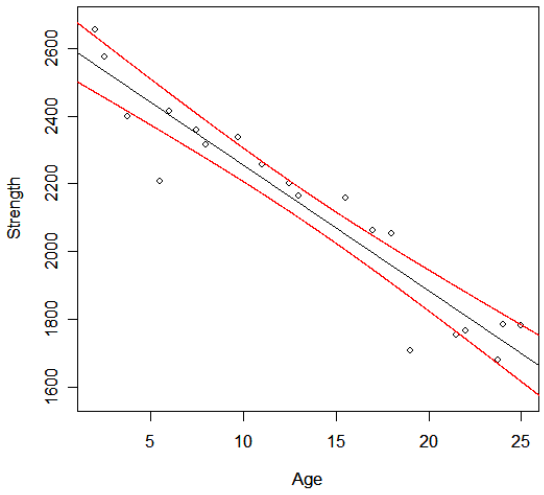

Figure 2.2: The 95% confidence intervals for the rocket propellant data.

SAS Output contains the 95% confidence intervals for $\mathrm{E}[Y|X = x_0]$ for several other values of x_0. These confidence intervals are illustrated graphically in Figure 2.2. In particular, when $x_0 = 10$, $\hat{\mu}_0 = 2627.8224 - (37.1536)(10) = 2256.286$ and hence the 95% CI for $\mu_0 = \mathrm{E}[Y|X = 10]$ is

$$\hat{\mu}_0 \pm t_{.025,18}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} = 2256.286 \pm (2.101)(23.5837) = (2206.739, 2305.834).$$

Extrapolation

The prediction beyond the range of the data (known as **extrapolation**) is often less meaningful. The linear regression model is not going to perform well over the range of the data because of model error. The greater the extrapolation, the higher is the chance of model error impacting the results. See Figure 1.5 in textbook. Read pp. 32–33 for the details.

2.5 Prediction Interval for New Observation

An important application of the regression model is prediction of new observations Y at a given value of X. If we denote a given value of X by x_0, then

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

is the point estimate of the new observation $Y_0 (= \beta_0 + \beta_1 x_0 + \epsilon_0)$. Under the assumptions A1–A4 and the additional assumption that $\epsilon_0 \sim N(0, \sigma^2)$ and ϵ_0 is independent of $\epsilon_1, \ldots, \epsilon_n$, it can be shown that

$$\psi = Y_0 - \hat{y}_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right).$$

Standardizing this and replacing σ^2 by MS_{Res} gives

$$\frac{Y_0 - \hat{y}_0}{\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

Thus, a $100(1-\alpha)\%$ prediction interval for Y_0 is given by

$$\hat{y}_0 \pm t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Note: A *confidence interval* is always reported for a *parameter* (e.g., $\mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0$) and a *prediction interval* is reported for the value of a *random variable* (e.g., Y_0).

Example

Revisit the artificial data.

Example

Suppose that we wish to estimate the "actual" shear strength of the propellant bond in a rocket motor made from a batch of sustainer propellant that is 10 weeks old.

Output:

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	2159	2052	22.3597	1845	2259	106.7583
2	1678	1745	36.9114	1529	1962	-67.2746
3	2316	2331	26.4924	2121	2540	-14.5936
4	2061	1996	23.9220	1788	2204	65.0887
5	2208	2423	31.2701	2211	2636	-215.9776

6	1708	1922	26.9647	1712	2132	-213.6041
7	1785	1736	37.5010	1519	1953	48.5638
8	2575	2535	38.0356	2318	2752	40.0616
9	2358	2349	27.3623	2139	2559	8.7296
10	2257	2219	22.5479	2012	2427	37.5671
11	2165	2145	21.5155	1938	2352	20.3743
12	2400	2488	35.1152	2274	2703	-88.9464
13	1780	1699	39.9031	1480	1918	80.8174
14	2337	2266	23.8903	2058	2474	71.1752
15	1765	1810	32.9326	1597	2024	-45.1434
16	2054	1959	25.3245	1750	2168	94.4423
17	2414	2405	30.2370	2193	2617	9.4992
18	2201	2163	21.6340	1956	2370	37.0975
19	2654	2554	39.2360	2335	2772	100.6848
20	1754	1829	31.8519	1616	2042	-75.3202
21	.	2256	23.5837	2048	2464	.

The 95% prediction interval (PI) for $Y_0 = \beta_0 + 10\beta_1 + \epsilon_0$ is

$$\hat{y}_0 \pm t_{.025,18}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} = 2256.286 \pm (2.101)(98.9538) = (2048.385, 2464.188).$$

Notice that SAS Output provides $\text{se}(\hat{\mu}_0)$ and we can compute $\text{se}(Y_0 - \hat{y}_0)$ from $\text{se}(\hat{\mu}_0)$. Therefore, a new motor made from a batch of 10-week-old sustainer propellant could reasonably be expected to have a propellant shear strength between 2048.385 psi and 2464.188 psi.

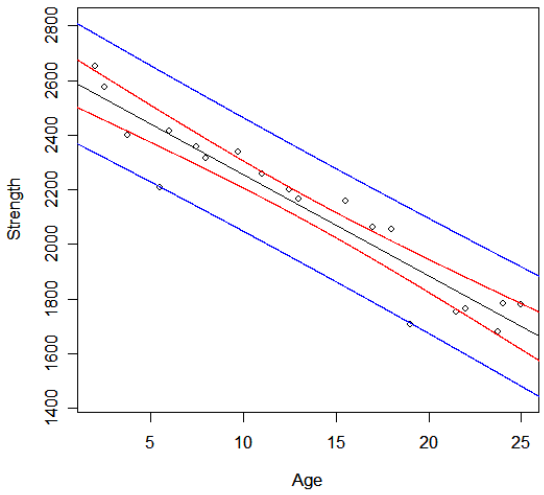

Figure 2.3: The 95% confidence and prediction intervals for the rocket propellant data.

Note that the prediction intervals are considerably wider than the confidence intervals.

2.6 Confidence Interval for σ^2

If the errors are normally and independently distributed,

$$\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}.$$

Thus, a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}}, \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}} \right).$$

Example

For the rocket propellant data, $MS_{Res} = 9236.38$, $\chi^2_{.025,18} = 31.53$ and $\chi^2_{.975,18} = 8.23$. Consequently, the 95% confidence interval for σ^2 is $(5273.52, 20199.26)$.

2.7 Analysis of Variance

There is a linear relationship between Y and X if $Y = \beta_0 + \beta_1 X + \epsilon$ and $\beta_1 \neq 0$. To test whether there is a linear relationship between Y and X we have to test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

As we saw in Section 2.3, we can perform this test using the following t-statistic

$$t_0 = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)}.$$

Note that $t_0 \sim t_{n-2}$ under H_0.

An alternative way to test the significance of the regression coefficient is an analysis-of-variance approach. For this, we need to know the following terminology.

- Total corrected sum of squares of the observations (SS_T):

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

- Residual sum of squares (SS_{Res}):

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- Regression sum of squares (SS_R):

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

  Note that the regression sum of squares is the sum of squares explained by the regression model. For the simple linear regression model it can be shown that $SS_R = \hat{\beta}_1^2 S_{xx}$.

Then, it can be shown that

$$SS_T = SS_R + SS_{Res}.$$

$$\left(\text{Total variability} = \begin{array}{c} \text{Variability explained} \\ \text{by the regression model} \end{array} + \begin{array}{c} \text{Unexplained variability} \\ \text{by the regression model} \end{array} \right)$$

To test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ we can use the test statistics

$$F_0 = \frac{SS_R/1}{SS_{Res}/(n-2)}.$$

Under the assumption that $\epsilon_1, \ldots, \epsilon_n$ are independently and normally distributed with mean 0 and variance σ^2, it can be shown that

- $\frac{SS_R}{\sigma^2} \sim \chi_1^2$ under H_0,
- $\frac{SS_{Res}}{\sigma^2} \sim \chi_{n-2}^2$,
- SS_R and SS_{Res} are independent.

Also, note that $\mathrm{E}[SS_R] = \sigma^2 + \beta_1^2 S_{xx}$. Thus, we have the following distribution result

$$F_0 = \frac{SS_R/1}{SS_{Res}/(n-2)} \sim F_{1,n-2}$$

when H_0 is true. If there is a linear relationship between Y and X then SS_{Res} should be "small" and SS_R should be "close" to SS_T; i.e., if F_0 is large enough then the null hypothesis is rejected.
How large is "large"?
The null hypothesis is rejected at level α if $F_0 > F_{\alpha,1,n-2}$ or $P - value < \alpha$.

Notes:

1. It can be shown that in the case of simple linear regression

$$t_0 = \frac{\hat{\beta}_1 - 0}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2}$$

   and

$$F_0 = \frac{SS_R/1}{SS_{Res}/(n-2)} \sim F_{1,n-2}$$

   are related via $t_0^2 = F_0$.

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean squares (MS)	F
Regression	1	SS_R	$MS_R = SS_R/1$	$F_0 = \frac{MS_R}{MS_{Res}}$
Residual	$n-2$	SS_{Res}	$MS_{Res} = SS_{Res}/(n-2)$	
Total	$n-1$	SS_T		

Table 2.1: Analysis of variance for testing significance of regression

2. The coefficient of determination of the regression line is defined as

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T},$$

which is the proportion of variability in Y explained by the regression model.

For the rocket propellant data, $R^2 = .9018$ that is, 90.18% of the variability in strength is explained by the regression model.

Notes that

- $0 \leq R^2 \leq 1$.
- $R^2 \approx 1$ implies that most of the variation in the observation is explained by the regression model.
- It is possible to make R^2 large by adding enough terms to the model.
- R^2 is equal to the square of the correlation between Y and X.

Example

We will test for significance of the slope of the regression model for the rocket propellant data using an analysis-of-variance (AOV) approach.

SAS Output:

```
                          Analysis of Variance

                                   Sum of          Mean
Source                     DF       Squares        Square   F Value   Pr > F

Model                       1       1527483       1527483    165.38   <.0001
Error                      18        166255    9236.38100
Corrected Total            19       1693738

          Root MSE             96.10609    R-Square     0.9018
          Dependent Mean     2131.35750    Adj R-Sq     0.8964
          Coeff Var             4.50915
```

The analysis of variance is summarized in the above SAS Output. The computed value of F_0 is 165.38 and the critical value at $\alpha = .05$ is $F_{.05,1,18} = 4.41$ from Table A.4. The $P - value = P(F_{1,18} > 165.38) < .0001$. Consequently, we reject $H_0 : \beta_1 = 0$. This agrees to t-test for the slope. Observe that $F_0 = 165.38 = t_0^2 = (12.86)^2$.

2.8 Centering of the Regression Model

There is an alternate form of the simple linear regression model that is occasionally useful. Recall the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n.$$

Observe that this model can be rewritten as

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i,\ i = 1, \ldots, n$$

with $\beta_0^* = \beta_0 + \beta_1\bar{x}$. Note that the intercept β_0^* for the model with centered predictor is the mean of the distribution of the response when the predictor is at its centered value. It is easy to show that the least-squares estimate of β_0^* is $\hat{\beta}_0^* = \bar{y}$ and the estimate of the slope is unaffected by centering.

There are two reasons to center. One is technical. The numerical routines that fit the model are often more accurate when variables are centered. Some computer programs automatically center variables and transform the model back to the original variables, all without the user's knowledge. The second reason is practical. Recall that the intercept β_0 for the model with uncentered predictor is the mean of the distribution of the response when the predictor is 0. However, if the range of the predictor does not include zero, then the intercept β_0 has no practical interpretation. The coefficients from a centered model are often easier to interpret.

2.9 Regression Through the Origin

A no-intercept regression model often seems appropriate in analyzing data from chemical and other manufacturing processes.

Example: The yield of a chemical process is zero when the process operating temperature is zero.

The no-intercept model is

$$Y = \beta_1 X + \epsilon.$$

Given n observations $(y_i, x_i), i = 1, \ldots, n$, the least-squares estimate of the slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Consequently, the least-squares regression line is

$$\hat{Y} = \hat{\beta}_1 X.$$

The estimate of σ^2 is

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

- A $100(1-\alpha)\%$ CI for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-1}\text{se}(\hat{\beta}_1)$$

  with $\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2}}$.

- A $100(1-\alpha)\%$ CI for $\text{E}[Y|X = x_0]$ is

$$\hat{\mu}_0 \pm t_{\alpha/2,n-1}\sqrt{\frac{x_0^2\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2}}.$$

- A $100(1-\alpha)\%$ PI for a new observation $Y_0 = \beta_1 x_0 + \epsilon_0$ is

$$\hat{Y}_0 \pm t_{\alpha/2,n-1}\sqrt{\hat{\sigma}^2\left(1 + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)}.$$

The no-intercept model analogue for R^2 is

$$R_0^2 = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2}$$

and this statistic R_0^2 indicates the proportion of variability around the origin (zero) accounted for by regression. Generally R^2 is not a good comparative statistic for the intercept and no-intercept models. We occasionally find that R_0^2 is larger than R^2 even though the residual mean square (which is a reasonable measure of the overall quality of the fit) for the intercept model is smaller than the residual mean square for the no-intercept model.

Note: You may decide whether or not to fit the no-intercept model via scatter plot, prior knowledge, comparison of MS_{Res} or hypothesis test.

Example

See the shelf-stocking data in Ex 2.8 in textbook.

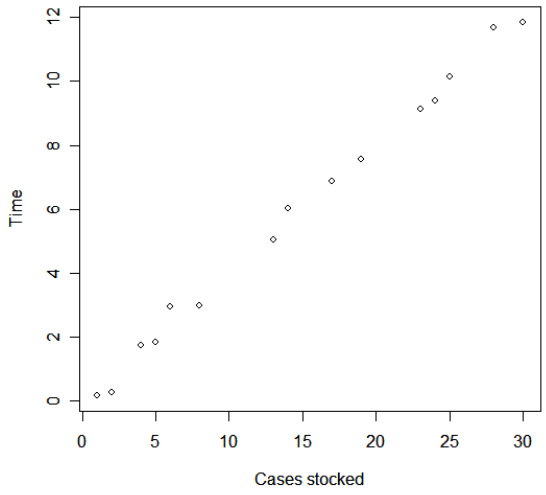

Figure 2.4: Scatter plot of shelf-stocking data.

Figure 2.4 suggests that a straight line passing through the origin could be used to express the relationship between time and the number of cases stocked. Furthermore, since if the number of cases = 0 then shelf stocking time = 0, this model seems intuitively reasonable.

SAS program:

```
option ls=90 ps=75;
title 'Shelf-Stocking Data';
proc reg data=stock;
model Time = Cases / noint;  /* no-intercept model */
plot Time*Cases / conf pred;
run; quit;
proc reg data=stock;  /* intercept model */
model Time = Cases;
run; quit;
```

Output:

- Intercept model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F

Model	1	228.31757	228.31757	2452.13	<.0001
Error	13	1.21043	0.09311		
Corrected Total	14	229.52800			

Root MSE	0.30514	R-Square	0.9947
Dependent Mean	5.85000	Adj R-Sq	0.9943
Coeff Var	5.21605		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.09376	0.14358	-0.65	0.5251
Cases	1	0.40711	0.00822	49.52	<.0001

The t statistic for testing $H_0 : \beta_0 = 0$ is $t_0 = -.65$, which is not significant, implying that the no-intercept model may provide a superior fit.

- No-intercept model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	741.61537	741.61537	8305.23	<.0001
Error	14	1.25013	0.08929		
Uncorrected Total	15	742.86550			

Root MSE	0.29882	R-Square	0.9983
Dependent Mean	5.85000	Adj R-Sq	0.9982
Coeff Var	5.10808		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Cases	1	0.40262	0.00442	91.13	<.0001

Also, we conclude that the no-intercept model is superior since MS_{Res} for the no-intercept model is smaller than MS_{Res} for the intercept model. As noted previously, the R^2 statistics for these two models are not directly comparable.

The least-squares fitted line is

$$\widehat{\text{Time}} = .4026 * \text{Cases Stocked}$$

which is shown in Figure 2.5. The t statistic for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is $t_0 = 91.13$, for which the $P - value$ is $< .0001$. These summary statistics do not reveal any startling inadequacy in the no-intercept model.

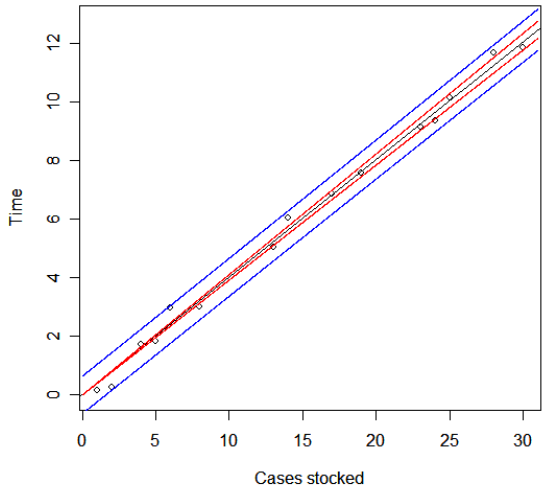

Figure 2.5: Scatter plot of shelf-stocking data with the least-squares regression line and the confidence and prediction intervals.

Figure 2.5 also shows the 95% CI for the mean response $\mathrm{E}[Y|X = x_0] = \beta_1 x_0$ and the 95% PI for a new observation $Y_0 = \beta_1 x_0 + \epsilon_0$. Notice that the length of the confidence interval at $x_0 = 0$ is zero.