Chapter 7: Simple linear regression

The absolute movement of the ground and buildings during an earthquake is small even in major earthquakes. The damage that a building suffers depends not upon its displacement, but upon the acceleration. Understanding the characteristics of earthquake-caused ground acceleration is critical for the effective design and construction of earthquake resistant buildings. A first investigation into acceleration begins with examining the relationship between acceleration and distance to the epicenter of an earthquake.

Measurements recorded at seismometer locations for 23 large earthquakes in western North America between 1940 and 1980 are plotted in the figure.\(^1\)

The panels show the original data and the log-transformed data. Two statistical summaries of the data are also shown: a least squares regression line and a smooth\(^2\). In the left panel, the linear regression line does not summarize the relationship with any degree of reality whereas in the right panel, it is apparent that the linear model provides a reasonable though less than perfect summary of the relationship between the two variables. The utility of the linear regression equation is that it is simple and can be used to predict acceleration as a function of distance from epicenter in further analyses.

Last semester, linear regression models were introduced, but a critical topic was omitted: inferential methods. The context or framework for inference is a population of pairs \((x, y)\) where the *explanatory variable* \(x\) is fixed (non-random) and the *response variable* \(y\) is normally distributed around a mean \(\mu(y|x)\) that depends on the value of \(x\).\(^3\) For every particular value of \(x\), we envision of distribution of responses that is normally distributed with mean

\[
\mu(y|x) = E(y|x)
\]

---

\(^1\) Joyner and Boore (1981), Peak horizontal acceleration and velocity from strong-motion records including records from the 1979 Imperial Valley, California, earthquake, *Bulletin of the Seismological Society of America*, 71, 2011-2038.

\(^2\) A smooth represents the general trend between the two variables with relatively little assumptions or conditions imposed on the form of trend.

\(^3\) Often, perhaps usually, \(x\) is not fixed but random, as it is in the earthquake example. The failure to control \(x\) is not a deterrent to using linear regression and associated inferential methods.
and (again) a common standard deviation $\sigma$ for all values of $x$. The model assumes that there is a (possibly) different mean or expected value of $y$ for every different value of $x$. The means can be described by the simple linear regression model

$$\mu(y|x) = \beta_0 + \beta_1 x.$$  

We can also write $E(y|x) = \beta_0 + \beta_1 x$. There are two parameters: $\beta_0$ is the intercept and $\beta_1$ is the slope.

If a value of $x$ is selected, say $x_0$ and many pairs are observed with $x = x_0$, say

$$(x_0, y_1), (x_0, y_2), \ldots, (x_0, y_m)$$

then $y_1, y_2, \ldots, y_m$ will appear to be normally distributed about $\mu(y|x_0) = \beta_0 + \beta_1 x_0$ and with standard deviation $\sigma$. A helpful picture is shown in Ramsey and Schafer, p. 180. For example, ground acceleration might observed at regular intervals along the circumference of a circle with radius of 20 km (or $x_0 = 2.99 \text{ log-km}$) and centered on an earthquake epicenter. The average acceleration may be described by the linear model, but it is expected that the observed accelerations will vary about the mean according to the normal distribution.$^4$

If formal inference is to be conducted, then the linear regression model is

$$y = \mu(y|x) + \varepsilon,$$

$$\mu(y|x) = \beta_0 + \beta_1 x,$$

$$\varepsilon \sim N(0, \sigma).$$

The purpose of inference, of course, is to describe a population. According to this model, there may be uncountably many values of $y$ associated with a particular value $x$. Then, the subscript notation (e.g., $y_i$) is irrelevant at best. It’s convenient to think of the $y$’s associated with a particular value of $x$ as a subpopulation and specify the subpopulation distribution by writing $y \sim N(\mu(y|x), \sigma)$.

To use inferential methods with accuracy and confidence, it’s necessary to use sample data and verify that the model conditions are consistent with the data. To emphasize the role of the data, a sample-oriented specification of the linear regression model is often set up as

$$y_i = \mu(y_i|x_i) + \varepsilon_i$$

$$= \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma), i = 1, \ldots, n.$$  

$^4$Presumably, some of this variation may be accounted for by other variables such as soil type.
The sample size is \( n \) and the abbreviation iid stands for \textit{independent and identically distributed}. This last form of the model is commonly given in statistical texts (but not the \textit{Statistical Sleuth}). A number of inferential procedures are commonly used in formal linear regression analyses. The most important of these is a test of the hypothesis \( H_0 : \beta_1 = 0 \) versus \( H_a : \beta_1 \neq 0 \), since without convincing evidence that the slope is different from 0, there is no value to a fitted regression model. If the null hypothesis is adopted, then the adopted model is \( \mu(y|x) = \beta_0 = \mu \).

\textit{Interpolation and extrapolation}

If a fitted regression model is used to generate an estimate of \( \mu(y|x) \) for some value \( x_0 \) not in the data set but within the range of the observed explanatory variable\(^5\), the process is \textit{interpolation}. Interpolation is a very useful application of the linear model. If an estimate is computed for \( x_0 \) beyond the range of the explanatory variable, the process is \textit{extrapolation}, and fraught with risk. The figure above suggests that extrapolating for value of \( \log(x) \) larger than the maximum data set value is likely to be biased upwards. Relying on a figure to guide extrapolation is foolish however, since there is no assurance that linear trend apparent within the explanatory variable range continues above or below the maximum and minimum of \( x \).

\textit{Estimation of the least squares model parameters}

This discussion assumes an understanding of multivariable calculus and is optional reading. The objective is to find an optimal estimator of the parameter vector \( \beta = (\beta_0 \beta_1)^T \). The objective function is the sum of the squared prediction errors, or residuals given by

\[
\text{SSRes} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.
\]

The least squares estimator of \( \beta \) is obtained by minimizing the objective function with respect to \( \beta \). Minimization is accomplished by differentiating \( \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \) with respect to \( \beta \), setting the vector of partial derivatives equal to the zero vector, and solving the resulting system of equations for \( \beta \). The solution is \( \hat{\beta} = (\hat{\beta}_0 \hat{\beta}_1)^T \) where

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.
\]

\(^5\)\text{I.e., } \min(x_1, \ldots, x_n) \leq x_0 \leq \max(x_1, \ldots, x_n).
Problem 25, page 200 provides some assistance if you care to verify that the estimators given above truly yield the least possible sum of squared residuals. More can be said about the estimator: it is unique in the sense that every possible modification of the estimator will yield a worse estimator (in terms of SSRes).\textsuperscript{6}

**Estimation of $\sigma$**

Because the model specifies $y_i = \mu(y_i|x_i) + \varepsilon_i$, the realized value of $i$th residual is

\[
\hat{\varepsilon}_i = y_i - \hat{\mu}(y_i|x_i)
= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.
\]

The model specifies $\varepsilon_i \sim N(0, \sigma)$, and $\overline{\varepsilon}_i = 0$ is a property of least squares regression, so the estimator of $\sigma^2$ is

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n - 2} = \frac{\sum_{i=1}^{n} [y_i - \hat{\mu}(y_i|x_i)]^2}{n - 2}
\]

where $\hat{\mu}(y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The denominator $(n - 2)$ is the degrees of freedom associated with the estimator of $\sigma^2$. As with analysis of variance, the degrees of freedom is the difference between the number of observations and the number of estimated terms appearing to the right of the minus sign. In general, the degrees of freedom are the difference between the number of observations and the number of parameters in the linear regression model used to model $\mu(y|x)$\textsuperscript{7}.

The fitted values and a residual plot ($\hat{\varepsilon}_i$ against $\hat{\mu}(y_i|x_i)$) for the earthquake data are shown below. The residual plot indicates that relationship between $y$ and $x$ is not linear and instead contains a curvilinear component. A normal quantile plot of residuals indicates that the residuals are not normal distribution.

**Sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$**

The sampling distribution of $\hat{\beta}_1$ must be identified and used to draw formal inferences about $\beta_1$ (similarly for $\beta_0$). The sampling distribution of a statistic (e.g., $\hat{\beta}_1$) describes the distribution of the realized values of the statistic if the population is repeatedly sampled (that is, drawing many times samples of size $n$ and computing the value of the statistic from each sample).

\textsuperscript{6}The proof of uniqueness is called the Gauss-Markov Theorem.

\textsuperscript{7}The same denominator was used in the one-way analysis for variance: specifically, the degrees of freedom associated with $\hat{\sigma}_{\text{full}}$ is $n - I$, where $I$ is the number of sub-population means specified by the full model.
If the linear regression model described above is correct, then the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal distributions and the centers (specifically, the means) of the distributions are $\beta_0$ and $\beta_1$. Consequently, the least squares estimators of $\beta_0$ and $\beta_1$ are unbiased (if the linear regression model is correct).

It’s possible to construct an estimate of the sampling distribution of an estimator by repeated sampling the data with replacement and computing statistic of interest. In this case, there are two statistics of interest: $\hat{\beta}_0$ and $\hat{\beta}_1$. A smoothed histogram of the bootstrap estimates provides an empirical sampling distribution against which the normal theory sampling distribution may be compared. The figures below show the normal theory sampling distributions and bootstrap sampling distributions (based on 10,000 bootstrap samples).

The normal theory sampling distributions are similar to the empirical sampling distributions. Both distributions are estimates of the true unknown sampling distribution. There is some, but not a great deal, of assurance that the normal model is reasonably accurate.

---

8 Therefore, the least squares estimators are unbiased.
9 The process is called bootstrapping and it has a wide variety of applications.
10 An empirical sampling distribution is derived purely from the data and without a model.
Standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

The spread of the sampling distribution of a statistic is estimated by a standard error. The standard errors are

\[
\hat{\sigma}(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{x}{\sum(x_i - \bar{x})^2}}
\]

\[
\hat{\sigma}(\hat{\beta}_1) = \hat{\sigma} \sqrt{\sum(x_i - \bar{x})^2}.
\]

Note that the standard errors depend on the variability in the explanatory variable $x$, and the greater the variability in $x$, the smaller the standard errors are. A smaller standard error results in a more sensitive test of the hypothesis $H_0 : \beta = 0$ and an narrower confidence interval. The R function call `summary(lm(y~x))` produced Table 1 of summary statistics for the earthquake data.

Table 1: Linear regression coefficients and standard errors obtained from the Earthquake data set.

|                   | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------------|----------|------------|---------|----------|
| Intercept         | 0.18461  | 0.15410    | 1.198   | 0.232    |
| $x$               | -0.82466 | 0.04559    | -18.087 | < 2e-16  |

Table 2 is more appropriate for presentation to an audience.

---

11 The main reason that the explanatory variable $x$ is assumed to be fixed (and not a random variable) is that the sampling distributions of $\beta_0$ and $\beta_1$ are simple if only $y$ is random variable.
Table 2: Linear regression coefficients and standard errors obtained from the Earthquake data set. The p-value for a test of $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ is given in the column headed by $P(T > |t|)$.

| Parameter | Estimate | Std. Error | t-statistic | $P(T > |t|)$ |
|-----------|----------|------------|-------------|-------------|
| $\beta_0$ | .185     | .154       | 1.198       | .232        |
| $\beta_1$ | -.825    | .0456      | -18.09      | < .0001     |

For comparison, the bootstrap standard errors are $\hat{\sigma}_{\text{boot}}(\hat{\beta}_0) = .181$ compared to $\hat{\sigma}(\hat{\beta}_0) = .154$, and $\hat{\sigma}_{\text{boot}}(\hat{\beta}_1) = .0552$ compared to $\hat{\sigma}(\hat{\beta}_1) = .0456$. These differences are reflected in the plots of the empirical and theoretical sampling distributions.

**Tests involving $\beta_0$ and $\beta_1$**

Usually, the objective is to formally test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ since $\beta_1 = 0$ implies that there is no association between $y$ and $x$. Occasionally, there is interest in testing $H_a : \beta_0 = 0$ versus $H_a : \beta_0 \neq 0$ since $\beta_0 = 0$ implies that the linear model passes through the origin. Since the tests involving $\beta_0$ and $\beta_1$ are very similar, let $\beta$ denote either parameter and let $\beta_{\text{null}}$ the value of $\beta$ specified by the null hypothesis. The null hypothesis is

$$H_0 : \beta = \beta_{\text{null}}.$$ 

Three alternatives are possible:

1. $H_a : \beta > \beta_{\text{null}}$.
2. $H_a : \beta < \beta_{\text{null}}$.
3. $H_a : \beta \neq \beta_{\text{null}}$.

The test statistic is a $t$-ratio

$$T = \frac{\hat{\beta} - \beta_{\text{null}}}{\hat{\sigma}(\hat{\beta})},$$

where

$$\hat{\sigma}(\hat{\beta}) = \begin{cases} 
\hat{\sigma} \sqrt{\frac{1}{n} + \frac{x}{\sum(x_i - \bar{x})^2}} & \text{if } \beta = \beta_0, \\
\hat{\sigma} & \text{if } \beta = \beta_1.
\end{cases}$$

If the null hypothesis is true, then $T$ has a $t$-distribution with d.f. = $n - 2$ degrees of freedom; in shorthand, $T \sim t_{n-2}$. If $n$ is sufficiently large (e.g., $n \geq 80$), then $T$ is approximately $N(0, 1)$ in distribution.

Suppose that the observed value of $T$ is $t$. P-values are computed according to
1. If $H_a : \beta > \beta_{null}$, then p-value = $P(T > t|H_0 : \beta = \beta_{null})$.

2. If $H_a : \beta < \beta_{null}$, then p-value = $P(T < t|H_0 : \beta = \beta_{null})$.

3. If $H_a : \beta \neq \beta_{null}$, then p-value = $2P(T > |t||H_0 : \beta = \beta_{null})$.

For the earthquake data analysis, hypotheses involving $\beta_0$ and $\beta_1$ are not particularly interesting. There’s no question that damage and hence acceleration diminishes with distance from the epicenter of an earthquake. The alternative $H_a : \beta_1 < 0$ are understood to be true; the questions of interest involve the magnitude of $\beta_1$ the form of the relationship, and the extent to which other factors (e.g., soil type, depth of the epicenter) affect the relationship. I cannot develop a meaningful hypothesis for the intercept since I cannot identify any value that is sensible. The test statistics and p-values are given in Table 2, however.

Meningococcal infections are caused by a type of bacteria called Neisseria meningitidis. A very small number of people exposed to this organism develop a serious illness from it, such as meningitis (swelling of the tissues around the brain). Meningococci are spread by direct contact with secretions from the nose and throat of an infected person. In 1978, the first meningococcal vaccines were licensed in the United States. The data to the right show the weekly numbers of recorded meningococcal infections in France from 1985 to 1995. Presumably, after the vaccine has been introduced, the number of cases should decline. The data appear to support this premise.

Table 3 provides a summary of the fitted regression model. Before inference is discussed, the population should be described. The term population, however, is inaccurate. Rather, inference is aimed at the process generating the observed counts. The process is the spread of Meningococi infections in the French population, and the observed counts depend on the number of contagious individuals, susceptible individuals, weather and other conditions that affect the transmission of the bacteria and development of the disease in infected individuals. In this context $\mu(y_0|x_0)$ represents the underlying potential for meningococci infections in month $x_0$, and $y_0$ represents one particular realization of that potential.

---

12 In the US, meningococcal infection rates increase when schools are in session.
Table 3: Coefficients and standard errors obtained from the linear regression of weekly numbers of meningococcal infections against month (number past December 1994).

| Parameter | Estimate | Std. Error | t-statistic | P(T > |t|) |
|-----------|----------|------------|-------------|----------|
| $\beta_0$ | 38.82    | 1.918      | 20.24       | < .0001  |
| $\beta_1$ | -1.1187  | 0.0212     | 5.599       | < .0001  |

A test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 < 0$ is informative since $\beta_1 < 0$ implies that the meningococcal infection rate declined over the time span of interest. A test of $H_0 : \beta_0 = 0$ is not interesting because $\beta_0 = 0$ implies that the meningococcal infection rate in December of 1994 (month 0) was zero (obviously false to anyone with some knowledge of the disease). However, a confidence interval for $\beta_0$ is informative since it brackets all values for the December 1994 rate that are consistent with the data. A 95% confidence interval for $\beta_0$ is

$$\hat{\beta}_0 \pm t^* p \hat{\sigma}(\hat{\beta}_0) = 38.82 \pm 1.975 \times 1.918$$

$$= [35.03, 42.61]$$

where $p = \text{number of model parameters} = 2 \Rightarrow \text{df} = 156 - 2 = 154 \Rightarrow t^* = -1.975$. The estimator of the expected number of cases in December 1994 is $\hat{\beta}_0 = 38.82$ and the 95% confidence interval is [35.03, 42.61].

To expand, the model supposes that there is an underlying expected rate $\mu(y|\text{month} = 0)$ that is estimated by $\hat{\beta}_0$. The estimator uses all of the data and the assumed model to produce the estimate $\hat{\beta}_0 = 38.82$ and it’s standard error $\hat{\sigma}(\hat{\beta}_0) = 1.918$. Two alternative estimates are the mean over all 156 months and it’s standard error ($\bar{y} = 29.5$ and $\hat{\sigma}(\bar{y}) = 1.04$), and the nearest chronological observation and it’s estimated standard deviation ($y_1 = 51$ and $\hat{\sigma}(y) = s = 13.03$). Since the evidence of a declining trend over time is convincing, I accept that the estimated intercept is a better estimator of the December 2004 rate.

Similarly, the slope estimate is important, as is a confidence interval for the slope, since the slope estimates the monthly change in the expected number of cases. The estimate is $\hat{\beta}_1 = -0.119$ infections/month and the 95% confidence interval is [−.160, −.077] infections/month. 95% confidence intervals can be computed in R using the function call `confint(lm(y~x))`.

Describing the distribution of the response given a particular value of the explanatory variable

The discussion above described the distribution at a particular value of the explanatory variable ($x = 0$). Specifically, the distribution is (by assumption) normal and the estimated mean is $\hat{\beta}_0 = 38.82$. The estimated standard deviation of the response variable about this
mean is reported to be $\hat{\sigma} = 11.92$ on 154 degrees of freedom.\textsuperscript{13} For any other value $x_0$, we use the following formula to determine the estimated standard deviation of the estimated mean $\hat{\mu}(y|x_0)$

\[
\hat{\sigma}(\hat{\mu}[y|x_0]) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}
\]

where $s_x^2$ is the sample variance of the explanatory variable. The degrees of freedom associated with this statistic is $n - p = n - 2$. Notice that this standard error is smallest when $x_0 = \bar{x}$ and increases as the distance between $x_0$ and $\bar{x}$ increases. The usual application of the standard error is to compute $100(1-\alpha)\%$ confidence intervals for $\mu(y|x)$ across the range of $x$. Given $x_0$, a $100(1-\alpha)\%$ confidence interval for $\mu(y|x_0)$ is

\[
\hat{\mu}(y|x_0) \pm t'_{n-2}\hat{\sigma}(\hat{\mu}[y|x_0])
\]

where $t'_{n-2}$ is the $\alpha/2$ quantile from the $t$ distribution with $n - 2$ degrees of freedom.

A set of intervals with connected endpoints are referred to as a set of confidence intervals.\textsuperscript{14} The figure below shows 95% confidence intervals for $\mu(y|x)$ in red. The curved lines provides infinitely many confidence intervals visually; as most people interpret the intervals as containing the true line rather than one or a few specific $\mu(y|x)$’s, it’s preferable to construct Scheffe confidence bands that have the property that we are $100(1-\alpha)\%$ confident that the bands contain all the pairs $(x, \mu[y|x])$ satisfying $\mu(y|x) = \beta_0 + \beta_1x$. The only change necessary to obtain the Scheffe confidence bands is to replace the $t$-critical value with a term using the $1 - \alpha$ quantile from the $F_{2,n-2}$ distribution. The $100(1-\alpha)\%$ confidence bands are obtained by computing

\[
\hat{\mu}(y|x_0) \pm \sqrt{2F^*_{2,n-2}}\hat{\sigma}(\hat{\mu}[y|x_0]).
\]

The figure shows that there is little practical difference between the individual confidence intervals (in red) and the Scheffe confidence bands (in blue). Very few of the data pairs are contained in the confidence intervals or bands, but that is not a matter of concern since the intervals are not meant to bracket the (random) response variables; rather they are meant to bracket their expectation.

\textsuperscript{13}The command \texttt{summary(lm(y x))} command in R presents the estimate on the line labeled Residual standard error.

\textsuperscript{14}These are also referred to as confidence bands, but Ramsey and Schafer reserve that term for another set of intervals explained momentarily.
**Prediction of a future response value**

An important problem is the prediction of a future response value given the explanatory variable value \( x_0 \). The future response value is a random variable defined (by the model) to be \( y_0 = \mu(y|x) + \varepsilon_0 \).

The expectation of \( y_0 \) is \( E(y_0|x_0) = \mu(y|x) = \beta_0 + \beta_1 x_0 \). The expectation is the best predictor of \( y_0 \), and so the usual predictor in this situation is to use the estimated expectation \( \hat{E}(y_0|x_0) = \hat{\mu}(y|x) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \).

These values are called the fitted values. The fitted values are graphed as the black line in the figure.

An estimate of prediction error should be presented with the prediction. A preferred alternative is to present a 100(1 - \( \alpha \))% prediction interval. A 100(1 - \( \alpha \))% prediction interval has the property of bracketing a realization (or future value) of \( y_0 \) associated with \( x_0 \) with probability 1 - \( \alpha \). A 100(1 - \( \alpha \))% prediction interval for \( y_0 \) is wider than a 100(1 - \( \alpha \))% confidence interval for \( \mu(y_0|x_0) \) since it must reflect the uncertainty or variability of the random variable \( y_0 \) about \( \mu(y_0|x_0) \). A 100(1 - \( \alpha \))% prediction interval for \( y_0 \) is

\[
\hat{y}_0 \pm t_{n-2}^* \hat{\sigma}(\hat{y}_0|x_0),
\]

where \( \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \) and the standard error of the prediction is

\[
\hat{\sigma}(\hat{y}_0) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}.
\]

As above, \( t_{n-2}^* \) is the \( \alpha/2 \) quantile from the \( t \) distribution with \( n - 2 \) degrees of freedom. A set of 95% prediction intervals are shown in green in the figure above.

A few other topics of possible interest are omitted from this discussion. Ramsey and Schafer discuss calibration (or inverse prediction) problems wherein the objective is to predict the value of \( x \) that yielded an observed value \( y \). Ramsey and Schafer’s discussion of causation and correlation should be read for review. Their half-page discussion of the value of replication in experimental design is useful.