Chapter 6: Statistical inference for regression

Statistical inference, concisely, is process of drawing conclusions about a population from a sample. If certain distributional conditions are met by the data, then the scope of inference is extended, sometime greatly so. Statistical inference for regression exploits an often-met and usually forgiving set of conditions.

The discussion begins with a single explanatory variable, and simple linear regression. 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 \(E(Y|x)\) that depends on the value of \(x\).\(^1\) For every particular value of \(x\), we envision distribution of responses that is normally distributed with conditional mean \(E(Y|x)\) and a single standard deviation \(\sigma\).

In contrast, the model allows for a different mean or expected value of \(Y\) for every different value of \(x\):

\[
E(Y|x) = \beta_0 + \beta_1 x.
\]

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 \(E(Y|x_0) = \beta_0 + \beta_1 x_0\) and with standard deviation \(\sigma\).

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

\[
Y = E(Y|x) + \varepsilon,
\]

\[
E(Y|x) = \beta_0 + \beta_1 x,
\]

\[
\varepsilon \sim N(0, \sigma).
\]

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 \(Y \sim N(E(Y|x), \sigma)\).

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

\(^1\)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.
the data, a sample-oriented specification of the linear regression model is often set up as

\[ Y_i = E(Y_i|x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i, \]

where

\[ \epsilon_i \overset{iid}{\sim} N(0, \sigma), i = 1, \ldots, n. \]

The sample size is \( n \) and the abbreviation iid stands for independent and identically distributed. 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 \( E(Y_i|x_i) = \beta_0 = \mu \).

**Estimation of the least squares model parameters**

The least squares estimator of \( \beta \) is obtained by minimizing the objective function

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

with respect to \( \beta \). Minimization is accomplished by differentiating \( \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \) 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

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

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

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).

**Estimation of \( \sigma^2 \)**

A least squares estimator is not used for estimating \( \sigma^2 \). Instead, the estimator is a method of moments estimator. First, \( \sigma^2 \) is the conditional variance

\[
\sigma^2 = \text{var}(Y|x_1, \ldots, x_k) = E\{ (Y - E[Y|x_1, \ldots, x_k])^2 \}.
\]

\(^2\)The proof of uniqueness is called the Gauss-Markov Theorem.
The method of moments estimator uses a sample of independent observations. It is derived by first replacing the expectation operator with an average taken over the sample. Then, population quantities are replaced with sample quantities yielding

\[ \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{E}(Y_i|x_i)]^2 \]

However, this estimator is biased downwards; the method of moments estimators is corrected for bias. It is

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

The denominator \((n-k)\) is the degrees of freedom associated with the estimator of \(\sigma^2\). 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 \(E(Y|x)\).

The estimator \(\hat{\sigma}^2\) has a \(\chi^2_{n-k}\) distribution if the linear regression model is correct and it is properly scaled. Specifically,

\[ \frac{n-k}{\sigma^2} \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sigma^2} \sim \chi^2_{n-k}. \]

**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 (\(\hat{\beta}_1\)) 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).

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 \(\beta_1\).\(^3\) Consequently, the least squares estimators of \(\beta_0\) and \(\beta_1\) are unbiased (once again, 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.\(^4\) In this case, there are two statistics of interest: \(\hat{\beta}_0\) and \(\hat{\beta}_1\). A density plot of the bootstrap estimates provides

\(^3\)Therefore, the least squares estimators are unbiased.

\(^4\)The process is called bootstrapping and it has a wide variety of applications.
an empirical sampling distribution\(^5\) 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 earthquake data were used for this example. These data consist of measurements recorded at seismometer locations for 23 large earthquakes in western North America between 1940 and 1980. \(^6\) Both acceleration (the response variable) and distance to epicenter (explanatory variable) were transformed to the natural logarithm scale.

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.

**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.\(^7\) The normal theory standard errors are

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

$$\hat{\sigma}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \overline{x})^2}}.$$  

\(^5\)An empirical sampling distribution is derived purely from the data and without a model.


\(^7\)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 the only random variable.
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 a 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.

| Parameter | 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.

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

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

If the null hypothesis is true, then $T$ has a $t$-distribution with d.f. $= n - 2$ degrees of freedom; in shorthand, $T \sim t_{\text{d.f.}}$. 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 $H_a : \beta > \beta_{\text{null}}$, then p-value $= P(T > t| H_0 : \beta = \beta_{\text{null}})$.

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

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

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 $H_a : \beta_1 < 0$ is 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 give 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 figure above shows the weekly numbers of recorded meningococcal infections in France from 1985 to 1995. Presumably, after the vaccine has been introduced, the number of case 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 Meningococci 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 $E(Y_0|x_0)$ represents the underlying potential for meningococci infections in month $x_0$, and $y_0$ represents one particular realization of that potential.

### Table 3: Coefficients and standard errors obtained from the linear regression of weekly numbers of meningococcal infections against month (number of past December 1994).

| Parameter | Estimate | Std. Error | $t$-statistic | $P(T > |t|)$ |
|-----------|----------|------------|---------------|-------------|
| $\beta_0$ | 38.82    | 1.918      | 20.24         | < .0001     |
| $\beta_1$ | -.1187   | .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 confidence interval consists of a set of values for $\beta_0$ that are consistent with the data; all values not contained in the interval are considered to be inconsistent with the data. It is said that we are $100(1-\alpha)\%$ confident that the parameter is contained in the interval, where $1-\alpha$ is the confidence level. In repeated sampling, $100(1-\alpha)\%$ of all confidence intervals will contain the parameter. The probability that a particular interval, say $[35.03, 42.61]$ contains $\beta_0$ is 0 or 1, as neither the computed interval nor the parameter is a random variable.

A 95% confidence interval for $\beta_0$ is

$$\hat{\beta}_0 \pm t^*_{n-p, \hat{\sigma}(\hat{\beta}_0)} = 38.82 \pm 1.9755 \times 1.918 = [35.03, 42.61]$$

where $k = \text{number of model parameters} = 2 \Rightarrow df = 156 - 2 = 154 \Rightarrow t^* = -1.9755$. The estimator of the expected number of cases in December 1994 rate is $\hat{\beta}_0 = 38.82$ and the 95% confidence interval is $[35.03, 42.61]$.

---

8In the US, meningococcal infection rates increase when schools are in session.
To expand, the model supposes that there is an underlying expected rate $E(Y|\text{month} = 0)$ 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 $[-0.160, -0.077]$ infections/month.

**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 $\hat{\sigma} = 1.918$ on 154 degrees of freedom. For any other value $x_0$, we use the following formula to determine the estimated standard deviation of the estimated mean $\hat{E}(Y|x_0)$

$$\hat{\sigma}(\hat{E}[Y|\hat{x}_0]) = \hat{\sigma} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{n - 1}s^2_x}$$

where $s^2_x$ is the sample variance of the explanatory variable. The degrees of freedom associated with this statistic is $n - k = n - 2$. 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 $E(Y|x)$ across the range of $x$. Given $x_0$, a $100(1 - \alpha)$% confidence interval for $E(Y|x_0)$ is

$$\hat{E}(Y|x_0) \pm t_{n-2}^{\star} \hat{\sigma}(\hat{E}[Y|x_0])$$

where $t_{n-2}^{\star}$ 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. The figure below shows 95% confidence intervals for $E(Y|x)$ in red. The curved lines provides infinitely many confidence intervals visually; as most people interpret the intervals containing

---

9The command `summary(lm(y x))` command in R presents the estimate on the line labeled Residual standard error.

10These are also referred to as confidence bands, but it’s best to reserve that term for another set of intervals explained momentarily.
the true line rather than one or a few specific \( E(Y|x) \)’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 \( (x, E(Y|x)) \) satisfying \( E(Y|x) = \beta_0 + \beta_1 x \). 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

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

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.

**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 \( Y_0 = E(Y_0|x_0) + \varepsilon_0 \). The expectation of \( Y_0 \) is \( E(Y_0|x_0) = \beta_0 + \beta_1 x_0 \). The expectation is the best predictor of \( Y_0 \), and so the usual predictor in this situation is to use the estimated expectation \( \hat{E}(Y_0|x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \), otherwise known as the fitted value. The fitted values are graphed as the straight line in the figure.

Along with the prediction, and an estimate of prediction error should be presented. A preferred alternative presents 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 \( E(Y_0|x_0) \) since it must reflect the uncertainty or variability of the random variable \( y_0 \) about \( E(Y_0|x_0) \). 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.

Properties of the least squares estimators

The properties can be summarized tersely as

$$\hat{\beta}_0 \sim N \left( \beta_0, \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \right)$$

(1)

and

$$\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \right).$$

(2)

So the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables and their sampling distributions are normal. Further, the centers of the sampling distributions are the parameters being estimated, and so the estimators are unbiased.

The term $\sigma^2(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$ is the variance of the sampling distribution of $\hat{\beta}_0$. The precision of an estimator is reflected by and measured by the variance of its sampling distribution; the smaller the standard deviation, the more precise the estimator, and the closer an estimate will be (on average) to the true parameter value.

The standard deviations

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

(3)

and

$$\sigma(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

(4)

also reflect the precision. Both formulas show that variability among the explanatory variable values increases the precision of the estimators because $\sum (x_i - \bar{x})^2$ is in the denominator of the standard deviations.

Since $\sigma$, the standard deviation of the residuals $\varepsilon$ about the conditional mean must be estimated using the model residuals, the standard deviations of the estimators shown above must also be estimated. As discussed above, the estimators are referred to as the standard errors and are obtained by replacing $\sigma$ by its usual estimator $\hat{\sigma} = (n - k)^{-1} \sqrt{\sum e_i^2}$. The standard errors are

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

(5)

and

$$\hat{\sigma}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

(6)