# Inferences in Regression Analysis

Notes from "Linear Statistical Models" by Kutner et al

Devere Anthony Weaver

---

## 1 Overview

This set of notes is concerned with inferences concerning the regression parameters $\beta_0$ and $\beta_1$, considering both interval estimation of these parameters, and tests about them.

It also covers interval estimation of the mean of the probability distribution of $Y$ for a given $X$ and prediction intervals for a new observation of $Y$ given $X$.

The final coverage is of the analysis of variance approach to regression analysis, the general test approach, and descriptive measures of association.

Throughout this document, the *normal error regression model* is assumed

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.1}$$

**Note:** There will be **no** expanded derivations nor proofs in this document, I don't want to do all the extra typing. If derivation or proof is desired, refer to the corresponding sections of "Applied Linear Statistical Models" by Kutner et al.

---

## 2 Inferences Concerning $\beta_1$

With simple linear regression, we're often concerned about drawing inferences about the slope of a fitted linear regression model.

Often, the most interesting test is given by

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

The reason for interest in this test is that when the slope is equal to zero, this indicates that there is no linear association between $X$ and $Y$. In this case, the estimated mean for all values of $Y$ is just a horizontal line at the slope intercept; there is no relation of *any type* between $X$ and $Y$ since the probability distributions of $Y$ are then identical at all levels of $X$ (i.e. $X$ contains no information about $Y$).

### 2.1 Sampling Distribution of $b_1$

The sampling distribution of $b_1$ refers to the different values of $b_1$t would be obtained with repeated sampling when the levels of the independent variable $X$ are held constant from sample to sample.

For 1.1, the sampling distribution is *normal* with

$$E[b_1] = \beta_1$$

and

$$Var[b_1] = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

Note that $b_1$ is an unbiased estimator by the Gauss-Markov theorem. We can estimate the variance of the sampling distribution of $b_1$ via

$$s^2[b_1] = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{MSE}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

$s^2[b_1]$ is an unbiased estimator for the variance of its sampling distribution.

### 2.2 Sampling Distribution of $(b_1 - \beta_1)/s[b_1]$

Since $b_1$ is normally distributed, we know that the standardized statistic is a standard normal variable.

An important result in statistics is that $\frac{b_1 - \beta_1}{s[b_1]}$ is distributed as $t(n-2)$ for 1.1.

### 2.3 Confidence Interval for $\beta_1$

Since the standardized $b_1$ follows a $t$ distribution, we can make the following probability statement

$$P[t(\alpha/2; n-2)] \leq \frac{b_1 - \beta_1}{s[b_1]} \leq t(1-\alpha/2; n-2)] = 1 - \alpha$$

This leads to the $1-\alpha$ confidence limits for $\beta_1$ being

$$b_1 \pm t(1-\alpha/2; n-2)s[b_1] \tag{2.1}$$

Of course, this can be computed by hand, but practically, in an applied setting we'll always use software.

### 2.4 Tests Concerning $\beta_1$

Generally, we'll concern ourselves with two different types of tests. The first being mentioned above, the two-sided test, that tests if there is any relation between the two variables. The second being a one-sided test such as

$$H_0 : \beta_1 \leq 0$$

$$H_A : \beta_1 > 0$$

Both of these test use a $t$-test with a $t$ test statistic $\frac{b_1}{s[b_1]}$; however, again since this is practically done using software, we'll often just use the p-value that is reported with the output. The p-value in statistical software packages will agree with the decision rule from the $t$-test.

---

## 3 Inferences Concerning $\beta_0$

There are only infrequent occasions when we wish to make inferences concerning $\beta_0$. These will really only occur when the scope of the model includes $X = 0$.

As such, I won't really delve into the detail here, but you would perform a $t$-test as you would for the slope. Again, software will report the p-value but you'll have to determine if there is any actual meaning for this result.

---

## 4 Some Consideration on Making Inferences Concerning $\beta_0$ and $\beta_1$

### 4.1 Effect of Departures from Normality

If the probability distributions of $Y$ are not exactly normal but do not depart seriously, the sampling distributions of $b_0$ and $b_1$ will be approximately normal, and the use of the $t$ distribution will provide approximately the specified confidence coefficient or level of significance.

Even if the distributions of $Y$ are far from normal, the estimators will have the property of *asymptotic normality*.

### 4.2 Interpretation of Confidence Coefficient and Risks of Errors

Since the regression model 1.1 assumes the $X_i$ are known constants, the confidence coefficient and risks of errors are interpreted with respect to taking repeated samples in which the $X$ observations are kept at the same levels as in the observed sample.

For example, a confidence interval with confidence coefficient .95 is interpreted to mean that if many independent samples are taken where the levels of $X$ in the first sample are repeated in these other samples, 95 percent of the intervals will contain the true value of the parameter.

### 4.3 Interval Estimation of $E[Y_h]$

Recall, in regression analysis, one of the major goals usually is to estimate the mean for one or more probability distributions of $Y$.

Let $X_h$ denote the level of $X$ for which we wish to estimate the mean response. $X_h$ may be a value which occurred in the sample, or it may be some other value of the independent variable within the scope of the model.

The mean response when $X = X_h$ is denoted by $E[Y_h]$. The following gives us the point estimation for the expected value of $Y_h$

$$\hat{Y_h} = b_0 + b_1 X_h \tag{4.1}$$

So, 4.1 is used to estimate the mean response at the level of $X_h$.

### 4.4 Sampling Distribution of $\hat{Y_h}$

For regression model 1.1, the sampling distribution of $\hat{Y_h}$ is normal, with mean and variance

$$E[\hat{Y_h}] = E[\hat{Y_h}]$$

$$Var[\hat{Y_h}] = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

### 4.5 Sampling Distribution of $\hat{Y_h} - E[Y_h]/s[\hat{Y_h}]$

This standardized version of y-hat is distributed as $t(n-2)$ for regression model. Hence, all inferences concerning $E[Y_h]$ are carried out in the usual fashion with the $t$ distribution.

### 4.6 Confidence Interval for $E[Y_h]$

A confidence interval for $E[Y_h]$ is constructed in the standard fashion, making use of the $t$ distribution. The $1-\alpha$ confidence limits are

$$\hat{Y_h} \pm t(1-\alpha/2; n-2)s[\hat{Y_h}] \tag{4.2}$$

Again, confidence intervals for $Y_h$ will practically be computed with statistical software.

---

## 5 Prediction of New Observation

This section considers the prediction of a new observation $Y$ corresponding to a given level of $X$ of the independent variable. The new observation on $Y$ is viewed as the result of a new trial, independent of the trials on which the regression analysis is based.

Denote the level of $X$ for a new trial as $X_h$ and the new observation on $Y$ as $Y_{h(new)}$. We assume the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation.

A **very important** thing to note is that there is a distinction between *estimating the mean response* $E[Y_h]$ and the *prediction of a new response* $Y_{h(new)}$.

In the former case, we estimate the mean of the distribution of $Y$ at a given level of $X$ (the regression line). In the later case, we predict an *individual outcome* drawn from the distribution of $Y$ at a given level of $X$.

The basic idea of a prediction interval is to choose a range in the distribution of $Y$ wherein most of the observations will fall, and to declare that the next observation will fall in this range.

Since we can't be certain of the location of the distribution of $Y$, prediction limits for the new observation clearly must take account of two elements:

1. Variation in possible location of the distribution of $Y$.
2. Variation within the probability distribution of $Y$.

Prediction limits for a new observation $Y$ at a given level of $X_h$ are obtained by means of the following theorem:

**Theorem 1.**

$$\frac{\hat{Y_h} - Y}{sY_{h(new)}}$$

*is distributed at a* $t(n-2)$ *for regression model 1.1.*

The $1-\alpha$ prediction limits for a new observation are

$$\hat{Y_h} \pm t(1-\alpha/2; n-2)s[Y_{h(new)}] \tag{5.1}$$

Prediction intervals are wider than confidence intervals since we encounter variability in the estimation of the new observation from sample to sample as well as the variation within the probability distribution of $Y$.

Again, a confidence interval represents an inference on a parameter, and its an interval which is intended to cover the value of the parameter while a prediction interval is a statement about the value to be taken by a random variable.

---

# 6 Analysis of Variance (ANOVA)

It should be noted that there is a substantial amount of content regarding the basics of ANOVA and its calculations; however, I won't include any of this information on this document as ANOVA isn't the primary focus here and mathematically, ANOVA and simply linear regression are equivalent in many cases.

Refer to the textbook for more details on the computation of ANOVA, its interpretation, ANOVA tables, etc.

The most important part of the section is the subsection on the $F$ test and the $F$ distribution that can be extended to the general linear test.

---

# 7 The General Linear Test

The ANOVA test of $\beta_1$ is an example of a generate test for a linear statistical model. The general linear test approach involves three basic steps:

1. Fit the full model and obtain the error sum of squares *SSE(F)*.
2. Fit the reduced model under $H_0$ and obtain the error sum of squares *SSS(R)*.
3. Use the $F^*$ test statistic and decision rule.

Again, I've not included any of the computational aspects of this in the notes since in practice, we can use the output of statistical software as a test. Specifically in R, the $F$ statistic and test are automatically computed as part of the output of a fitted linear model object.

---

# 8 Descriptive Measures of Association Between $X$ and $Y$ in Regression Model

No single descriptive measure of the "degree of linear association" can capture the essential information as to whether a given regression relation is useful in any particular application. However, there are times when the degree of linear association is of interest in its own right and the following subsections introduction two descriptive measures that are frequently used in practice to describe the degree of linear association between $X$ and $Y$.

## 8.1 Coefficient of Determination

T

$$r^2 = 1 - \frac{SSE}{SSTO} \tag{8.1}$$

where

$$0 \leq r^2 \leq 1$$

We may interpret $r^2$ as the proportionate reduction of total variation associated with the use of the independent variable $X$. The larger $r^2$ is, the more the total variation of $Y$ is reduced by introducing the independent variable $X$.

## 8.2 Coefficient of Correlation

The square root of $r^2$ is called the coefficient of $r^2$

$$r = \pm\sqrt{r^2} \tag{8.2}$$