

Inference in Regression Analysis

Dr. Frank Wood

Today: Normal Error Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

1. Y_i value of the response variable in the i^{th} trial
2. β_0 and β_1 are parameters
3. X_i is a known constant, the value of the predictor variable in the i^{th} trial
4. $\epsilon_i \sim_{iid} N(0, \sigma^2)$
5. $i = 1, \dots, n$

Inferences concerning β_1

Tests concerning β_1 (the slope) are often of interest, particularly

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

the null hypothesis model

$$Y_i = \beta_0 + (0)X_i + \epsilon_i$$

implies that there is no relationship between Y and X

Review : Hypothesis Testing

1. Elements of a statistical test
 - 1.1 Null hypothesis, H_0
 - 1.2 Alternative hypothesis, H_a
 - 1.3 Test statistic
 - 1.4 Rejection region

Review : Hypothesis Testing - Errors

1. Errors

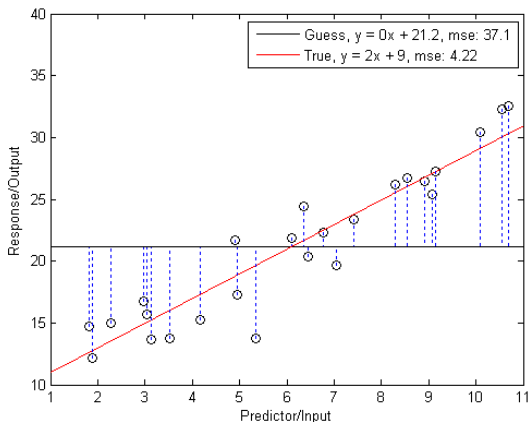
- 1.1 A type I error is made if H_0 is rejected when H_0 is true. The probability of a type I error is denoted by α . The value of α is called the level of the test.
- 1.2 A type II error is made if H_0 is accepted when H_a is true. The probability of a type II error is denoted by β .

P-value

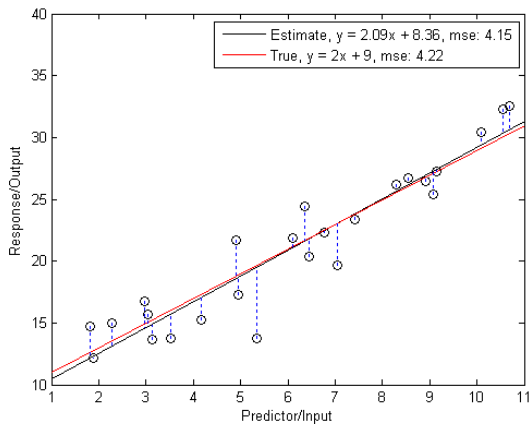
The p-value, or attained significance level, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.

Null Hypothesis

If $\beta_1 = 0$ then with 95% confidence the b_1 would fall in some range around zero



Alternative Hypothesis : Least Squares Fit



Testing This Hypothesis

1. Only have a finite sample
2. Different finite set of samples (from the same population / source) will (almost always) produce different estimates of β_0 and β_1 (b_0, b_1) given the same estimation procedure
3. b_0 and b_1 are random variables whose sampling distributions can be statistically characterized
4. Hypothesis tests can be constructed using these distributions.

Example : Sampling Dist. Of b_1

1. The point estimator for b_1 is

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

2. The sampling distribution for b_1 is the distribution over b_1 that occurs when the predictor variables X_i are held fixed and the observed outputs are repeatedly sampled

Sampling Dist. Of b_1 In Normal Regr. Model

1. For a normal error regression model the sampling distribution of b_1 is normal, with mean and variance given by

$$\begin{aligned}\mathbb{E}(b_1) &= \beta_1 \\ \text{Var}(b_1) &= \frac{\sigma^2}{\sum(X_i - \bar{X})^2}\end{aligned}$$

2. To show this we need to go through a number of algebraic steps.

First step

To show

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X})Y_i$$

we observe

$$\begin{aligned}\sum (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum (X_i - \bar{X})Y_i - \sum (X_i - \bar{X})\bar{Y} \\ &= \sum (X_i - \bar{X})Y_i - \bar{Y} \sum (X_i - \bar{X}) \\ &= \sum (X_i - \bar{X})Y_i - \bar{Y} \sum (X_i) + \bar{Y}n \frac{\sum X_i}{n} \\ &= \sum (X_i - \bar{X})Y_i\end{aligned}$$

Slope as linear combination of outputs

b_1 can be expressed as a linear combination of the Y_i 's

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ &= \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} \\ &= \sum k_i Y_i \end{aligned}$$

where

$$k_i = \frac{\sum(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$$

Properties of the k_i 's

It can be shown that

$$\begin{aligned}\sum k_i &= 0 \\ \sum k_i X_i &= 1 \\ \sum k_i^2 &= \frac{1}{\sum (X_i - \bar{X})^2}\end{aligned}$$

(possible homework). We will use these properties to prove various properties of the sampling distributions of b_1 and b_0 .

Normality of b'_1 s Sampling Distribution

1. Useful fact:

- 1.1 A linear combination of independent normal random variables is normally distributed
- 1.2 More formally: when Y_1, \dots, Y_n are independent normal random variables, the linear combination $a_1 Y_1 + a_2 Y_2 + \dots + a_n Y_n$ is normally distributed, with mean $\sum a_i \mathbb{E}(Y_i)$ and variance $\sum a_i^2 \text{Var}(Y_i)$

Normality of b_1 's Sampling Distribution

Since b_1 is a linear combination of the Y_i 's and each Y_i is an independent normal random variable, then b_1 is distributed normally as well

$$b_1 = \sum k_i Y_i, \quad k_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

b_1 is an unbiased estimator

This can be seen using two of the properties

$$\begin{aligned} E(b_1) &= E\left(\sum k_i Y_i\right) = \sum k_i E(Y_i) = \sum k_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \\ &= \beta_0(0) + \beta_1(1) \\ &= \beta_1 \end{aligned}$$

Variance of b_1

Since the Y_i are independent random variables with variance σ^2 and the k_i 's are constants we get

$$\begin{aligned}V(b_1) &= V\left(\sum k_i Y_i\right) = \sum k_i^2 V(Y_i) \\&= \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2 \\&= \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}\end{aligned}$$

note that this assumes that we know σ^2 . Can we?

Estimated variance of b_1

1. When we don't know σ^2 then we have to replace it with the MSE estimate
2. Remember

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

plugging in we get

$$V(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$
$$\hat{V}(b_1) = \frac{s^2}{\sum(X_i - \bar{X})^2}$$

Digression : Gauss-Markov Theorem

In a regression model where $E(2i) = 0$ and variance $V(2i) = \sigma^2$ and $2i$ and $2j$ are uncorrelated for all i and j the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Remember

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$
$$b_0 = \bar{Y} - b_1\bar{X}$$

Proof

1. The theorem states that b_1 as minimum variance among all unbiased linear estimators of the form

$$\hat{\beta}_1 = \sum c_i Y_i$$

2. As this estimator must be unbiased we have

$$\begin{aligned} E(\hat{\beta}_1) &= \sum c_i E(Y_i) = \beta_1 \\ &= \sum c_i (\beta_0 + \beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1 \end{aligned}$$

Proof cont.

1. Given these constraints

$$\beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1$$

clearly it must be the case that $\sum c_i = 0$ and $\sum c_i X_i = 1$

2. The variance of this estimator is

$$V(\hat{\beta}_1) = \sum c_i^2 V(Y_i) = \sigma^2 \sum c_i^2$$

Proof cont.

Now define $c_i = k_i + d_i$ where the k_i are the constants we already defined and the d_i are arbitrary constants. Let's look at the variance of the estimator

$$\begin{aligned}V(\hat{\beta}_1) &= \sum c_i^2 V(Y_i) = \sigma^2 \sum (k_i + d_i)^2 \\ &= \sigma^2 \left(\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \right)\end{aligned}$$

Note we just demonstrated that

$$\sigma^2 \sum k_i^2 = V(b_1)$$

Proof cont.

Now by showing that $\sum k_i d_i = 0$ we're almost done

$$\begin{aligned}\sum k_i d_i &= \sum k_i (c_i - k_i) \\ &= \sum k_i (c_i - k_i) \\ &= \sum k_i c_i - \sum k_i^2 \\ &= \sum c_i \left(\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right) - \frac{1}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum c_i X_i - \bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} = 0\end{aligned}$$

Proof end

So we are left with

$$\begin{aligned}V(\hat{\beta}_1) &= \sigma^2(\sum k_i^2 + \sum d_i^2) \\ &= V(b_1) + \sigma^2(\sum d_i^2)\end{aligned}$$

which is minimized when the d_i 's = 0. This means that the least squares estimator b_1 has minimum variance among all unbiased linear estimators.

Sampling Distribution of $(b_1 - \beta_1)/S(b_1)$

1. b_1 is normally distributed so $(b_1 - \beta_1)/(\text{Var}(b_1)^{1/2})$ is a standard normal variable
2. We don't know $\text{Var}(b_1)$ so it must be estimated from data. We have already denoted it's estimate
3. Using this estimate we it can be shown that

$$\frac{b_1 - \beta_1}{\hat{S}(b_1)} \sim t(n - 2)$$

$$\hat{S}(b_1) = \sqrt{\hat{V}(b_1)}$$

Where does this come from?

1. We need to rely upon the following theorem
For the normal error regression model

$$\frac{SSE}{\sigma^2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2(n - 2)$$

and is independent of b_0 and b_1

2. Intuitively this follows the standard result for the sum of squared normal random variables
Here there are two linear constraints imposed by the regression parameter estimation that each reduce the number of degrees of freedom by one.

Another useful fact : t distribution

Let z and $\chi^2(\nu)$ be independent random variables (standard normal and χ^2 respectively). We then define a t random variable as follows:

$$t(\nu) = \frac{z}{\sqrt{\frac{\chi^2(\nu)}{\nu}}}$$

This version of the t distribution has one parameter, the degrees of freedom ν

Distribution of the studentized statistic

To derive the distribution of this statistic, first we do the following rewrite

$$\frac{b_1 - \beta_1}{\hat{S}(b_1)} = \frac{\frac{b_1 - \beta_1}{S(b_1)}}{\frac{\hat{S}(b_1)}{S(b_1)}}$$

$$\frac{\hat{S}(b_1)}{S(b_1)} = \sqrt{\frac{\hat{V}(b_1)}{V(b_1)}}$$

Studentized statistic cont.

And note the following

$$\frac{\hat{V}(b_1)}{V(b_1)} = \frac{\frac{MSE}{\sum(x_i - \bar{X})^2}}{\frac{\sigma^2}{\sum(x_i - \bar{X})^2}} = \frac{MSE}{\sigma^2} = \frac{SSE}{\sigma^2(n-2)}$$

where we know (by the given theorem) the distribution of the last term is χ^2 and indep. of b_1 and b_0

$$\frac{SSE}{\sigma^2(n-2)} \sim \frac{\chi^2(n-2)}{n-2}$$

Studentized statistic final

But by the given definition of the t distribution we have our result

$$\frac{b_1 - \beta_1}{\hat{S}(b_1)} \sim t(n - 2)$$

because putting everything together we can see that

$$\frac{b_1 - \beta_1}{\hat{S}(b_1)} \sim \frac{z}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}$$

Confidence Intervals and Hypothesis Tests

Now that we know the sampling distribution of b_1 (t with $n-2$ degrees of freedom) we can construct confidence intervals and hypothesis tests easily