Regression Analysis

Last Updated:

September 20, 2023

Contents

1 Introduction **2**

1.1 The Method of Least Squares 3

1.1.1 Properties of Fitted Regression Line 6

1.2 Inference in Regression 8

1.3 Analysis of Variance Tables (ANOVA) 8

1.3.1 Hypothesis Testing 9

Chapter 1

Introduction

The primary goal in regression is to a devlop a model that relates a set of explanatory variables $X_1, \ldots, X_p$ to a response variable Y, then test the model and use it for inference and predicition.

Given a set of n pairs of data Y_i and X_i, we attempt to fit a straight line to these points, using a simple regression model

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

Where ϵ_i represents an unobserved random error term, β_0 is the intercept and β_1 is the slope of the line. β_0 and β_1 are parameters that need to be estimated from observed data. The model can also be expressed in terms of $(X_i - \bar{X})$.

$$Y_i = (\beta_0 + \beta_1 \bar{X}) + \beta_1 (X_i - \bar{X}) + \epsilon_i$$

Where $\bar{X}$ is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

This proposed model is linear in the parameters β_0, β_1, and would still be referred to as linear if we had X_i^2 instead of X_i. This model also makes the assumption that the random error terms ϵ_i are uncorrelated, have mean 0, and variance σ^2. Under these assumptions, we have

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

$$\mathrm{Var}(Y_i) = \sigma^2$$

,

1.1 The Method of Least Squares

The method of least squares is the most popular approach to fitting a regression model. Let Q be the sum of the error terms squared

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_i X_i)^2$$

Then we want to minimize Q with respect to the parameters β_1, β_2,

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

The linearity assumption gives us two linear equations with unknown solutions b_0, and b_1

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i$$

with

$$k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Note that k_i has important properites, such as

$$\sum_{i=1}^{n} k_i = 0, \ \sum_{i=1}^{n} k_i X_i = 1, \ \sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

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

$$\begin{aligned}
\sum_{i=1}^{n} k_i X_i &= \frac{\sum_{i=1}^{n} (X_i - \bar{X}) X_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n} (X_i^2 - X_i \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n} (X_i^2 - 2X_i \bar{X} + \bar{X}^2)} \\
&= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2} = 1
\end{aligned}$$

$$
\begin{aligned}
\sum_{i=1}^{n} k_i^2 &= \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right)^2 \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^4} \\
&= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned}
$$

The equation for the fitted line is then

$$\hat{Y} = b_0 + b_1 X$$

Or alternatively using $X - \bar{X}$,

$$\hat{Y} = (b_0 + b_1\bar{X}) + b_1(X - \bar{X})$$

Theorem 1.1.1 (Gauss Markov Theorem)**.** *The least square estimators b_0, b_1 are unbiased and have minimum variance among all unbiased linear estimators.*

Proof. Consider an unbiased linear estimator

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$$

$\hat{\beta}_1$ must satisfy $E(\hat{\beta}_1) = \beta_1$.

$$
\begin{aligned}
\beta_1 &= E(\hat{\beta}_1) \\
&= E\left(\sum_{i=1}^{n} c_i Y_i \right) \\
&= \sum_{i=1}^{n} c_i E(Y_i) \\
&= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 X_i) \\
&= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i
\end{aligned}
$$

Therefore, $\sum\limits_{i=1}^{n} c_i = 0$, and $\sum\limits_{i=1}^{n} c_i X_i = 1$. We can also see that the variance is

$$\operatorname{Var}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \operatorname{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2$$

Now, set $c_i = k_i + d_i$ where k_i is as defined previously above and d_i are arbitrary constants. We want to show that the variance is minimized, so

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \sum_{i=1}^{n} c_i^2 \,\mathrm{Var}(Y_i)\\ &= \sigma^2 \sum_{i=1}^{n} c_i^2\\ &= \sigma^2 \sum_{i=1}^{n} (k_i + d_i)^2\\ &= \sigma^2 \left(\sum_{i=1}^{n} k_i^2 + 2\sum_{i=1}^{n} k_i d_i + \sum_{i=1}^{n} d_i^2 \right)\end{aligned}$$

Note that the variance of b_1 is

$$\mathrm{Var}(b_1) = \mathrm{Var}\left(\sum_{i=1}^{n} k_i Y_i\right) = \sigma^2 \sum_{i=1}^{n} k_i^2$$

Now notice that there is a relationship between the variance of $\hat{\beta}_1$ and b_1, namely that the variance of $\hat{\beta}_1$ is the same as b_1 plus an additonal constants but these constants are indeed 0.

$$\begin{aligned}\sum_{i=1}^{n} k_i d_i &= \sum_{i=1}^{n} k_i (c_i - k_i)\\ &= \sum_{i=1}^{n} k_i c_i - \sum_{i=1}^{n} k_i^2\\ &= \sum_{i=1}^{n} c_i \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\\ &= \frac{\sum_{i=1}^{n} c_i X_i - \sum_{i=1}^{n} c_i \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\end{aligned}$$

We know that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i X_i = 1$, so this becomes

$$\sum_{i=1}^{n} k_i d_i = \frac{1-0}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0$$

Therefore,

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 \right)$$

Clearly the variance is minimized when $d_i = 0$ for all i, thus

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \mathrm{Var}(b_1)$$

Thus the least squares estimator b_1 has minimum variance along all unbiased estimators. □

We may write

$$\hat{Y} = b_0 + b_1 X$$

for the estimated or fitted line, and

$$e_i = Y_i - \hat{Y}_i$$

for the estimated i^{th} residual. The estimate for the variance σ^2 is then

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

The estimate of the variance σ^2 is also known as the mean square error (MSE).

1.1.1 Properties of Fitted Regression Line

(i) $\sum\limits_{i=1}^n e_i = 0$. Recall that $\hat{Y} = b_0 + b_1 X = (b_0 + b_1 \bar{X}) + b_1(X - \bar{X})$, and

$$\bar{Y} = b_0 + b_1 \bar{X}$$

So $\hat{Y} = \bar{Y} + b_1(X - \bar{X})$, then

$$\begin{aligned}
\sum_{i=1}^n e_i &= \sum_{i=1}^n (Y_i - \hat{Y}_i) \\
&= \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i \\
&= \sum_{i=1}^n Y_i - \sum_{i=1}^n (\bar{Y} + b_1(X_i - \bar{X})) \\
&= n\bar{Y} - n\bar{Y} + b_1 \sum_{i=1}^n (X_1 - \bar{X}) \\
&= n\bar{Y} - n\bar{Y} + b_1(n\bar{X} - n\bar{X}) = 0
\end{aligned}$$

(ii) $\sum\limits_{i=1}^n Y_i = \sum\limits_{i=1}^n \hat{Y}_i$. This follows from the previous property since

$$\sum_{i=1}^n e_i = \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i = 0 \implies \sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$$

(iii) $\sum_{i=1}^{n} X_i e_i = 0$. This can be shown from the definition

$$\begin{aligned}\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i (Y_i - \hat{Y}_i) \\ &= \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) \\ &= \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 \\ &= b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i \\ &= 0\end{aligned}$$

This is signficant because it tells us that the dot product between the vector of explanatory variables $\vec{X} = (X_1, \ldots, X_i)^T$ is orthogonal to the vector of error terms $\vec{e} = (e_1, \ldots, e_n)^T$, and from the previous property we get that

$$\vec{e} \cdot 1_n = \sum_{i=1}^{n} e_i = 0$$

Hence the vectors $\{1_n, X - \bar{X} 1_n\}$ are linearly independent and form a basis of the estimation space.

(iv) By applying the Pythagorean Theorem to the previous property we get

$$\begin{aligned}||Y||^2 &= ||\hat{Y}||^2 + ||Y - \hat{Y}||^2 \\ \sum_{i=1}^{n} Y_i^2 &= \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} \bar{Y}^2 + b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} e_i^2 \\ \implies \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 &= b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} e_i^2 \\ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 &= b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2\end{aligned}$$

This shows us the the total sum of squares is equal to the regression sum of squares plus the error sum of squares.

(v) The point $(\bar{X}, \bar{Y})$ is on the fitted line.

(vi) Under the normality assumption, $e_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The method of maximum likelihood leads to the method of least squares.

$$L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right)$$

So maximizing $L(\beta_0, \beta_1, \sigma^2)$ is equivalent to minimizing $\sum \epsilon_i^2$.

Note. The following 2 equations (known as the first and second normal equations) are important, the derivation is omitted.

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i \implies \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i = n\beta_0$$

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

1.2 Inference in Regression

Suppose that we have the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_1$, where $\{\epsilon_i\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Then

(a) $\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$ where $s^2(b_1) = \frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

(b) $\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$ where

$$s^2(b_0) = MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

(c) MSE is an unbiased estimate of σ^2 and is independent of b_0,b_1. Furthermore

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

Proof. Proof will be shown when we generalize this using matrices in later sections. □

1.3 Analysis of Variance Tables (ANOVA)

We create a table to summarize our results from statistical analysis.

Source	SS	DF	MS=SS/df	E(MS)
Regression	$SSR = b_1^2 \sum(X_i - \bar{X})^2$	$p - 1$	MSR	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n - p$	MSE	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n - 1$		

Each of the sums of squares is a quadratic form where the rank of the corresponding matrix is the degrees of freedom indicated. Chochran's theorem applies and we conclude that the quadratic forms are independent and have chi-sqaured distributions. Note that

$$\frac{SSR}{\sigma^2} = \frac{b_1^2 \sum (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{p-1}$$

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

Then, the ratio between 2 chi-sqaured distributions divided by their degrees of freedom has a F-distribution with their respective degrees of freedom.

$$F = \frac{SSR/\sigma^2(p-1)}{SSE/\sigma^2(n-p)} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{p-1,n-p}$$

The degrees of freedom are determined by the amount of data required to calculate each expression.

- $\sum\limits_{i=1}^{n} (Y_i - \bar{Y})^2$ has $n-1$ degrees of freedom since we have 1 constraint on the data that
$$\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0$$
- $b_1^2 \sum\limits_{i=1}^{n} (X_i - \bar{X})^2$ has one degree of freedom because it is a function of b_1
- $\sum\limits_{i=1}^{n} (Y - i - \hat{Y}_i)^2$ has $n-2$ degrees of freedom because it is a function of 2 parameters

1.3.1 Hypothesis Testing

The ANOVA table indactes how one can test the null hypothesis

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

The null Hypothesis is that the slope of the line is equal to 0. Under the null hypothesis, the expected mean square for regression and the expected mean square error are seperate independent