STAT430.Regression

Benjamin Kedem

June 2020

1 The General Problem of Linear Regression

Consider a set of non-random predictors or explanatory variables or *covariates* $x_1, ..., x_k$ and the corresponding random response variable y. The regression problem is to relate y to its covariates $x_1, ..., x_k$. That is, we wish to **regress** y on the corresponding covariates $x_1, ..., x_k$. This fundamental problem can be approached in a number of ways, one of which is referred to as **multiple linear regression**, where

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k$$

The problem is to estimate the β's. Many of the ideas of multiple regression are illustrated in terms of simple linear regression.

1.1 Simple Linear Regression

Suppose we decided to to fill a car gas tank with x gallons of gas (added to a constant quantity already in the tank) and measure the corresponding number of travel miles y. Clearly, x is a fixed (i.e. not random) quantity that we control. On the other hand, the number of travel miles y is random. The problem is how to relate a random variable y to a non-random variable x.

Consider the pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where $x_1, ..., x_n$ are non-random covariates and $y_1, ..., y_n$ are random observations.

Assume the model:

$$y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad i = 1, 2, ..., n$$

where:
$\epsilon_1, ..., \epsilon_2$ are uncorrelated: $Cov(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$.
$E(\epsilon_i) = 0$
$Var(\epsilon_i) = \sigma^2$ is the *same* for all i.
$E(y_i) = \alpha + \beta(x_i - \bar{x})$
$Var(y_i) = \sigma^2$.

Are the y_i iid? Are the x_i iid?

By centering the x_i we gain a mathematical simplification which helps in the description of the main ideas.

1.1.1 Least Squares Estimation

The parameters α, β are estimated by the method of least squares where the sum of squares

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \bar{x})]^2$$

is minimized w.r.t. α and β.

1. $\frac{\partial}{\partial \alpha} \sum_{i=1}^{n} \epsilon_i^2 = -2 \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \bar{x})] = 0$

2. $\frac{\partial}{\partial \beta} \sum_{i=1}^{n} \epsilon_i^2 = -2 \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \bar{x})](x_i - \bar{x}) = 0$

Fact: $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

Using the fact: The LSE's are given by

$\hat{\alpha} = \bar{y}$

$\hat{\beta} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

Regression line:

$$\hat{y} = \hat{\alpha} + \hat{\beta}(x - \bar{x})$$

In particular,

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$$

Statistical properties of LSE's:

$\hat{\alpha}$ and $\hat{\beta}$ are LINEAR and UNBIASED:
$E(\hat{\alpha}) = \alpha$, $E(\hat{\beta}) = \beta$

$$Var(\hat{\alpha}) = \frac{1}{n^2} \sum Var(y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$Var(\hat{\beta}) = \frac{1}{[\sum (x_i - \bar{x})^2]^2} \sum (x_i - \bar{x})^2 \sigma^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Does it makes sense to choose $x_i = 7$ for all i?

Theorem (Gauss-Markov): The variances of the LSE are the smallest in the class of all linear unbiased estimates.

Proof: We want to show $\hat{\beta}$ has the smallest variance in the class of all linear unbiased estimates. So let $b' = \sum c_i y_i$ another linear unbiased estimate. Then, for all β,

$$E(b') = \sum c_i E(y_i) = \sum c_i(\alpha + \beta(x_i - \bar{x})) = \alpha \sum c_i + \beta \sum (x_i - \bar{x})c_i = \beta$$

This is an IDENTITY in β!!! Therefore,

$$\sum c_i = 0, \qquad \sum (x_i - \bar{x})c_i = 1$$

and hence,

$$Var(b') = \sum c_i^2 Var(y_i) = \sigma^2 \sum c_i^2 = \sigma^2 \sum \left[\left(c_i - \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2}\right) + \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2}\right]^2$$

The middle term in 0. Therefore,

$$Var(b') = \sigma^2 \sum \left(c_i - \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2}\right)^2 + \frac{\sigma^2}{\sum (x_j - \bar{x})^2}$$

which is minimized for

$$c_i = \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2}$$

Therefore,

$$b' = \sum c_i y_i = \sum \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2} y_i = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_j - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_j - \bar{x})^2} = \hat{\beta}$$

Hence $\hat{\beta}$ has minimum variance.

We can show the same for $\hat{\alpha}$. Suppose $a' = \sum c_i y_i$ is linear unbiased for α. Then

$$E(a') = \sum c_i[\alpha + \beta(x_i - \bar{x})] = \alpha \sum c_i + \beta \sum (x_i - \bar{x})c_i = \alpha$$

This is an identity in α. Therefore

$$\sum c_i = 1, \quad \sum (x_i - \bar{x})c_i = 0$$

and

$$\begin{aligned} Var(a') = \sigma^2 \sum c_i^2 &= \sigma^2 \sum [(c_i - 1/n) + 1/n]^2 \\ &= \sigma^2 \sum \left[\left(c_i - \frac{1}{n}\right)^2 + \frac{2}{n}\left(c_i - \frac{1}{n}\right) + \frac{1}{n^2}\right] \\ &= \sigma^2 \left[\sum \left(c_i - \frac{1}{n}\right)^2 + \frac{2}{n}(1 - 1) + \frac{1}{n}\right] \end{aligned}$$

which is minimized for $c_i = 1/n$. Hence $a' = \bar{y} = \hat{\alpha}$ has minimum variance.

This result is stated as follows: **The LSE are Best Linear Unbiased Estimates or BLUE.**

1.1.2 Basic Decomposition of Total Sum of Squares

To judge the goodness of fit of the linear regression model it is useful to consider the residuals $y_i - \hat{y}_i,\ i = 1, ..., n$. Smaller residuals which do not appear to follow any particular pattern point to a reasonable fit.

Consider now the identity:

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

Then

$$\sum (y_i - \hat{y}_i)^2 = \sum [(y_i - \bar{y}) - (\hat{y}_i - \bar{y})]^2$$

Since the middle term is equal to $\sum (\hat{y}_i - \bar{y})^2$,

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 - 2 \sum (\hat{y}_i - \bar{y})^2 + \sum (\hat{y}_i - \bar{y})^2$$

By rearranging terms, we arrive at the basic decomposition of the total sum of squares

$$\sum (y_i - \bar{y})^2 (= SST) = \sum (y_i - \hat{y}_i)^2 (= SSE) + \sum (\hat{y}_i - \bar{y})^2 (= SSR)$$

From this we define a well known quantity:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SSR}{SST}$$

Obviously,

$$0 \leq R^2 \leq 1$$

The closer R^2 to 1, the better is the fit.

$$R^2 = \frac{SSR/n}{SST/n} = \frac{\mathit{Explained\ variance}}{\mathit{Total\ variance}}$$

R^2 is also known as the *coefficient of determination.*

We observe that in simple linear regression R^2 is equal to the the sample correlation between x and y squared:

$$\begin{aligned} r_{xy}^2 &= \frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} \times \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \\ &= \frac{\hat{\beta}^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = R^2 \end{aligned}$$

1.2 Hypothesis Testing

To test hypotheses we need distributions. For that we shall assume that:

The $\epsilon_i,\ i = 1, ..., n$, are independent $N(0, \sigma^2)$

With this assumption we can get the distributions of test statistics as well as confidence intervals.

1.2.1 Degrees of Freedom

Definition: The number of degrees of freedom of a sum of squares (SS) is the number of variables minus the number of linear relationships between them.

Example: Consider $\sum_{i=1}^n (x - \bar{x})^2$. This SS has n variables,

$$(x_1 - \bar{x}), ..., (x_n - \bar{x})$$

which satisfy the linear restriction $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Hence $df = n - 1$.

Example: Consider our centered regression model. We have,

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\bar{y} - \hat{\beta}(x_i - \bar{x}) - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

Only ONE variable is involved $\hat{\beta}$. Therefore the number of df of $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ is 1.

Example: Consider SSE.

$$1.\ \sum (y_i - \hat{y}_i) = \sum [(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})] = 0$$

That is, the sum of the residuals is 0. Therefore,

$$\begin{aligned} 2.\ \sum (y_i - \hat{y}_i)x_i &= \sum (y_i - \hat{y}_i)(x_i - \bar{x}) = \sum [(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})](x_i - \bar{x}) \\ &= \sum y_i(x_i - \bar{x}) - \bar{y} \sum (x_i - \bar{x}) - \hat{\beta} \sum (x_i - \bar{x})^2 \\ &= \sum y_i(x_i - \bar{x}) - \hat{\beta} \sum (x_i - \bar{x})^2 \\ &= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} \sum (x_j - \bar{x})^2 - \hat{\beta} \sum (x_i - \bar{x})^2 \\ &= \hat{\beta} \sum (x_j - \bar{x})^2 - \hat{\beta} \sum (x_i - \bar{x})^2 = 0 \end{aligned}$$

Therefore, $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ has $n - 2$ df.

This gives the basic ANOVA:

Source	SS	df
Model	SSR=$\sum(\hat{y}_i - \bar{y})^2$	1
Error	SSE=$\sum(y_i - \hat{y}_i)^2$	$n-2$
Corrected Total	SST=$\sum(y_i - \bar{y})^2$	$n-1$

From the normality of the ϵ_i we have.

Theorem: The following quantities are independent.

a. $\frac{n(\hat{\alpha}-\alpha)^2}{\sigma^2} \sim \chi^2_{(1)}$

b. $\frac{(\hat{\beta}-\beta)^2}{\sigma^2} \sum(x_i - \bar{x})^2 \sim \chi^2_{(1)}$

c. $\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{\sigma^2} \sim \chi^2_{(n-2)}$

Proof of b: Since the ϵ_i are iid $N(0, \sigma^2)$ the y_i are independent normal obsrvations. Since $\hat{\beta}$ is linear in the y_i's it is normal as well with mean β and variance (from above) $\sigma^2 / \sum(x_i - \bar{x})^2$. Therefore

$$\frac{(\hat{\beta}-\beta)}{\sigma}\sqrt{\sum(x_i-\bar{x})^2} \sim N(0,1)$$

But the square of $N(0,1)$ is a $\chi^2_{(1)}$, hence b.

We now define:

$$S^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

From c we see that S^2 is unbiased for σ^2:

$E(S^2) = \sigma^2$

Also: $Var(S^2) = \frac{2\sigma^4}{n-2}$

1.2.2 t and F Tests

From the theorem we see that:

$T = \frac{(\hat{\beta}-\beta_0)}{S}[\sum(x_i - \bar{x})^2]^{1/2} \sim t_{(n-2)}$

Hence, in testing $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$, H_0 is rejected when

$$|T| > t_{(n-2,\alpha/2)}$$

Squaring T we get

$$F = T^2 \sim F_{(1,n-2)}$$

Hence, under $H_0 : \beta = 0$,

$$F = \frac{\hat{\beta}^2 \sum(x_i - \bar{x})^2}{S^2} = \frac{\sum(\hat{y}_i - \bar{y})^2/1}{\sum(y_i - \hat{y}_i)^2/(n-2)} = \frac{SSR/1}{SSE/(n-2)} \sim F_{(1,n-2)}$$

and we reject for large values. THIS IS THE FIRST TEST PERFORMED BY SAS IN SIMPLE LINEAR REGRESSION.

1.2.3 Prediction Intervals

For a new covariate x_0 we want to compute a prediction interval for an unobserved y_0. Clearly, $E(y_0 - \hat{y}_0) = 0$.

Since y_0 is independent of $\hat{y}_0 = \hat{y}_0(y_1, ..., y_n)$,

$$Var(y_0 - \hat{y}_0) = Var(y_0) + Var(\hat{y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

From the definition of the t distribution,

$$\frac{y_0 - \hat{y}_0}{S\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]^{1/2}} \sim t_{(n-2)}$$

and a 95% prediction interval for y_0 is

$$\hat{y}_0 \pm t_{(0.025,n-2)} S \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]^{1/2}$$

SAS output provides plots of such prediction intervals. The prediction intervals are wider than confidence intervals for $E(y_0)$.

1.3 Simple Linear Regression in SAS

Remark: SAS uses the non-centered model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

All the above assumptions and results also hold for this model except that

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

From now on we shall use the SAS model as well.

Example: model Weight = Height

```
    Simple Linear Regression

 title ’Simple Linear Regression’;
   data Class;
      input Name $ Height Weight Age @@;
      datalines;
   Alfred  69.0 112.5 14  Alice  56.5  84.0 13  Barbara 65.3  98.0 13
   Carol   62.8 102.5 14  Henry  63.5 102.5 14  James   57.3  83.0 12
   Jane    59.8  84.5 12  Janet  62.5 112.5 15  Jeffrey 62.5  84.0 13
   John    59.0  99.5 12  Joyce  51.3  50.5 11  Judy    64.3  90.0 14
   Louise  56.3  77.0 12  Mary   66.5 112.0 15  Philip  72.0 150.0 16
   Robert  64.8 128.0 12  Ronald 67.0 133.0 15  Thomas  57.5  85.0 11
   William 66.5 112.0 15
   ;

   proc reg;
      model Weight = Height;
   run;

The REG Procedure

Model: MODEL1

Dependent Variable: Weight

Number of Observations Read 19
Number of Observations Used 19

Analysis of Variance
Source         DF Sum of Squares Mean Square F Value   Pr > F
Model           1 7193.24912      7193.24912   57.08   <.0001
Error          17 2142.48772       126.02869
Correct Total 18 9335.73684

Root MSE         11.22625    R-Square 0.7705
Dependent Mean  100.02632    Adj R-Sq 0.7570
Coeff Var      11.22330

Parameter Estimates
Variable DF     Param Est  SE        t Value Pr > |t|
Intercept 1    -143.02692 32.27459    -4.43  0.0004
Height    1       3.89903  0.51609     7.55    <.0001
```

Example: WEIGHT=HEIGHT

```
DATA SET1;
INPUT GENDER $ HEIGHT WEIGHT AGE;
DATALINES;
M 68 155 23
F 61 99 20
F 63 115 21
M 70 205 45
M 69 170 .
F 65 125 30
M 72 220 48
;
PROC REG DATA=SET1;
MODEL WEIGHT=HEIGHT; (Weight = a + b*Height + e)
RUN;

The REG Procedure

Model: MODEL1

Dependent Variable: WEIGHT

Number of Observations Read 7
Number of Observations Used 6
Number of Observations with Missing Values 1


Analysis of Variance:
Source         DF     SS           MS     F Value Pr > F
Model           1 8309.27052    8309.27052 61.47 0.0014
Error           4 540.72948     135.18237
Corrected Total 5 8850.00000
Check p-val=P(F(1,4)>61.47)=1-pf(61.47,1,4)=0.001429331

Root MSE      11.62680   R-Square 0.9389
Dependent Mean 165.00000   Adj R-Sq 0.9236
Coeff Var       7.04654

Parameter Estimates

Variable DF   Param Est       SE         t Value     Pr > |t|
Intercept   1   -670.03040 106.61338   -6.28         0.0033
HEIGHT    1     12.31003    1.57014     7.84         0.0014
```