Week-2

01 February 2024 10:58

ahw811@qmul.ac.uk.

Assessing the Simple Linear Regression Model (Statistical Modelling I)

Week 2, Lecture 2

Office Hours: Thursday 1:00 - 2:00pm.

Assessing the Simple Linear Regression Model

Outline

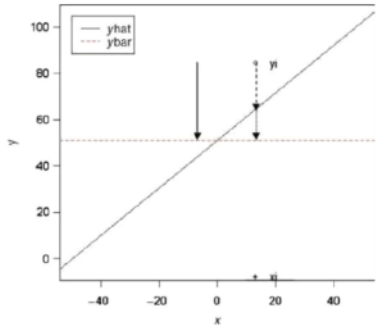
This Week Targets

1. Find your own data set that could be used for a simple linear regression model
2. Link it to one of the three things you said you would like to model earlier
3. Observations in (x, y) form with explanatory and response variables
4. Don't make it too large: 10- 50 observations
5. Save the data in Excel or csv file and upload that file to the submission point in the week 2 area of the module QM Plus site
6. Write down why you chose this data
7. There are no prizes for the data but we will use your data in the coming weeks and doing this now will make your first assessed coursework much much easier

Total Sum of Squares

The total variance of y_i around their mean $\overline{y}$ is the **Total Sum of Squares or** SS_T

$$SS_T = \sum_{i=0}^{n} (y_i - \overline{y})^2$$



Graphical depiction

a hypothetical situation with just a single data point (x_i, y_i) shown along with the least squares regression line and the mean of y based on all n data points.

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

Total Sum of Squares

In the simple linear regression model

Analysis of Variance Identity

Total sum of squares= Regression Sum of squares + Residual sum of squares

$$SS_T = SS_R + SS_E$$

Regression Sum of Squares

SS_R is the **Regression Sum of Squares**
Sometime called the Model Fit Sum of Squares

$$SS_R = \sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2$$

We will now show that $SS_T = SS_R + SS_E$.

Proof of $SS_T = SS_R + SS_E$

$$SS_T = \sum_{i=0}^{n} (y_i - \bar{y})^2$$

$$= \sum_{i=0}^{n} \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right)^2$$

$$= \sum_{i=0}^{n} \left[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 - 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \right]$$

$$= \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2$$

$$= SS_E + SS_R$$

$$-2\sum_{i=0}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = -2\sum_{i=0}^{n} (y_i - \hat{y}_i)\hat{y}_i + 2\bar{y}\sum_{i=0}^{n} (y_i - \hat{y}_i)$$

$$= -2\sum_{i=0}^{n} e_i \hat{y}_i + 2\bar{y}\sum_{i=0}^{n} e_i$$

$$\sum_{i=0}^{n} e_i = 0$$

$$\sum_{i=0}^{n} e_i \hat{y}_i = \sum_{i=0}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum e_i x_i$$

$$= 0$$

Total Sum of Squares (Summary)

Total Sum of Squares (Summary)

$\sum e_i x_i$

$\sum x_i (y_i - \hat{\beta}_1 - \hat{\beta}_1 x_i) = 0$ by first Normal Equation.

The total sum of squares SS_T is made up of:

1. The Regression Sum of Squares SS_R
   - The variability in the y_i around their mean $\overline{y}$
   - which is accounted for by the fitted model
2. The Residual Sum of Squares SS_E
   - The variability in the y_i
   - accounted for by the difference between observed and fitted values

This view can be presented in an **Analysis of Variance Table or ANOVA Table**

Sum of the residuals weighted by the corresponding value of the regressor variable always equal to zero.

The Anova Table

Source of variation	d.f.	SS	MS	VR
Regression	$v_R = 1$	SS_R	$MS_R = \frac{SS_R}{v_R}$	$F = \frac{MS_R}{MS_E}$
Residual	$v_E = \mathrm{n} - 2$	SS_E	$MS_E = \frac{SS_E}{v_E}$	
Total	$v_T = \mathrm{n} - 1$	SS_T		

The Anova Table

The variability in the y_i is accounted for by 4 quantities.
Each is a column of the Anova table:

- degree of freedom (d.f)
- sum of squares (SS)
- Mean Squares (MS)
- Variance Ration (VR)

We have already defined SS and will now consider the other three.

Mean Square

Mean Squares (MS) is a measure of the average variation for Residuals and for Regression found by dividing the relevant Sum of Squares (SS) by its degrees of freedom

$$MS_R = \frac{SS_R}{\nu_R} \text{ and } MS_E = \frac{SS_E}{\nu_E}$$

Variance Ratio

The Variance Ratio measures the variance explained by the model fit relative to that explained by the residuals. We usually denote this F.

$$F = \frac{MS_R}{MS_E}$$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

F Distribution often arises when we work with the Variance Ratios. We can also note that F Random variable is the Ration of 2 non-negative quantities so as such it will not take negative values. Areas or the percentiles for the F distribution can be found using R or an F table.

$\alpha = 0.05$

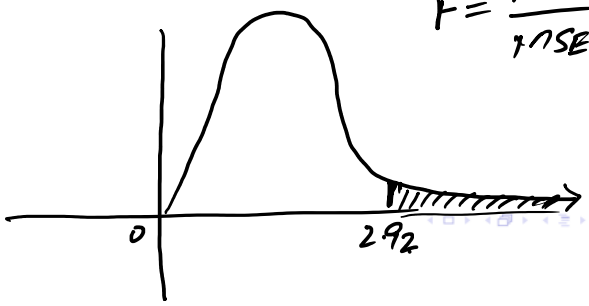
$\nu_1 = 3$

$\nu_2 = 30$

$F_{(3,30)}$

$F^{\nu_1,\nu_2}_{(3,30,0.05)} = 2.92$

$F = \frac{MSR}{MSE} \sim F_{(3,30,0.05)}$



Last term Prob & Stats II alert

If a random variable X follows a Chi-squared distribution on ν_1 degrees of freedom and variable Y follows a Chi-squared distribution on ν_2 degrees of freedom, then $\frac{X/\nu_1}{Y/\nu_2}$ follows a Fisher's F Distribution often simply called an F-Distribution with ν_1 and ν_2 degrees of freedom.

This is written as $\mathcal{F}_{\nu_1,\nu_2}$ or as $\mathcal{F}(\nu_1, \nu_2)$. The F-Distribution is skewed and depends on two parameters (ν_1, ν_2).

Using the Variance Ratio

The VR and the F-distribution are particularly useful in linear regression modelling for testing whether β_1 is statistically different from zero.

- If $\beta_1 = 0$ we could replace the linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with a simpler constant model $y_i = \beta_0 + \epsilon_i$.

We will see later in this course that if $\beta_1 = 0$ then the VR is such that

$$F = \frac{MS_R}{MS_E} \sim \mathcal{F}^1_{n-2}$$

To test the null hypothesis $\beta_1 = 0$

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0$$

Use the Variance Ratio as a test statistics. We reject H_0 at significance level α if

$$F > \mathcal{F}^1_{n-2}(\alpha)$$ critical point.

where $\mathcal{F}^1_{n-2}(\alpha)$ is such that $P[F > \mathcal{F}^1_{n-2}(\alpha)] = \alpha$.

F_{cal}

$\alpha = 0.05$
$\alpha = 0.1$ $\alpha = 0.5$.

Estimating σ^2

σ^2 is the variance of the residuals. In the Normal Simple Linear Regression model it is also the variance of the y_i. We can use the ANOVA table output to estimate σ^2.
Because the y_i are random variables, the Sums of Squares in the analysis of variance are also random variables.

- We will explore the properties of these SS_E random variables later in the course but for now we will use one of them $E(SS_E)$

We will see later that

$$E(SS_E) = (n-2)\sigma^2$$

Now from the ANOVA table the Mean Square for Residuals MS_E is

$$MS_E = \frac{SS_E}{\nu_E} = \frac{SS_E}{n-2}$$

Therefore $E(MS_E) = \sigma^2$. So MS_E is an unbiased estimator of σ^2.
This is interesting because MS_E is not the sample variance. We often denoted MS_E as S^2 and will use this notation in future.

Co-efficient of Determination R^2

Another quantity we get from the ANOVA table is R^2. The Coefficient of Determination, R^2 is the percentage of total variation in the y_i which is explained by the model fitted

$R^2 = 100\%$ all the observations fit exactly along the regression line

$R^2 = 0\%$ none of the variability in the data is explained by the model

$$\mathbf{R^2 = \frac{SS_R}{SS_T}100\% = (1 - \frac{SS_E}{SS_T})100\%}$$

Note that R^2 does not say whether there is any relationship between X and Y, just whether that relationship is linear.

Crude Residuals

From earlier we have the residuals or crude residuals are

$$e_i = y_i - \hat{y}_i \text{ and that } \sum_{i=0}^{n} e_i = 0$$

Studying these residuals, particularly through residual plots can be a useful way of assessing a linear regression model.

Before we are ready to do this we need to:

- Understand a bit more about the properties of the residuals
- Make some adjustments to the crude residuals

Expected value and Variance

$$\begin{aligned} E(e_i) &= E\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right) \\ &= E(y_i) - E\left((\hat{\beta}_0 + \hat{\beta}_1 x_i)\right) \\ &= (\beta_0 + \beta_1 x_i) - (\beta_0 + \beta_1 x_i) = 0 \end{aligned}$$

The mean of the i^{th} residual is zero

We will not derive $\text{Var}(e_i)$ from the first principles here

$$var(e_i) = \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

Notice though that $var(e_i)$ above is not the same as $\text{var}(\epsilon_i) = \sigma^2$. Here $\text{var}(e_i)$ changes with the x_i and therefore is different for each constant.

Covariance

For the residuals (again we won't derive from first principles)

$$cov(e_i, e_j) = -\sigma^2(\frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}})$$

Which again is different from $cov(\varepsilon_i, \varepsilon_j) = 0$

Adjustment to the residuals

Because the variance and covariance of the residuals in the fitted model (e_i) do not behave in the same way as the error term in the model specification ϵ_i It is sometimes better to work with standardised residuals which have

(i) variance closer to σ^2
(ii) covariances closer to zero

The standardised residuals are usually written d_i

Standardised residuals

The standardised residuals are given by

$$d_i = \frac{e_i}{(s^2(1-\nu_i))^{\frac{1}{2}}}$$

where

$$\nu_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}$$

Three useful Plots

d_i against x_i	• Check whether a linear model is appropriate • Check the Normal assumptions
d_i against $\hat{y}_i$	• Check for constant variance • Called homoscedasticity
QQ plot in R	• Good first indication of Normal residuals • Looking for a straight line

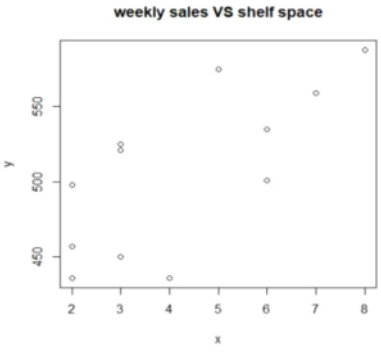
Exams Style Questions (2021)

Question 1 [25 marks].
A baker is interested to find the relationship between the width of the shelf-space for her brand of cookies (x, in feet) and monthly sales (y) of the product in a supermarket. Hence, she fits a model relating monthly sales y to the amount of shelf space x her cookies receive that month. That is, she is fitting the model in the following way

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

x (shelf space)	y (weekly sales)
3	535
2	425
6	575
5	639
3	450
8	630
4	435
2	498
6	534
3	530
2	457
7	559



Exams Style Questions (2021)

Using the R, we obtained the following output.

```
> mody <- lm(y ~ x)
> summary(mody)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-67.022 -31.346  -0.631  33.654  54.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  429.048     26.519  16.179 1.69e-08 ***
x             18.244      5.643   3.233  0.00898 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.2 on 10 degrees of freedom
Multiple R-squared:  0.511,     Adjusted R-squared:  0.4621
```

$\hat{y}_i = 429.048 + 18.244 x_i$

Exams Style Questions (2021)

```
F-statistic: 10.45 on 1 and 10 DF,  p-value: 0.008979

> anova(mody)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1  16059 16058.9  10.451 0.008979 **
Residuals 10  15366  1536.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MSE / MSR.

SSR. SSE

Use the R output above to answer the questions below.

(a) By looking at the summary output, write down the fitted model. [2]

(b) Write down the formula to compute the 95% confidence interval for β_1? [3]

(c) Compute the 95% confidence interval for β_1. [4]

(d) Fill in the blanks in the following table. [5]

Source of Variation	D F	Sum of Squares	Mean Square	F Value
Regression	1	SSR = 16059	MSR = 16058.9	?
Residual	12 2. = ?	SSE = 15366	MSE = $\frac{15366}{10}$ = ?	10.451

1536.6

Exams Style Questions (2021)

(e) (i) Write down the null hypothesis, that there is no effect on mean sales from increasing the amount of shelf space, versus a suitable alternative hypothesis. [4]

(ii) Compare the above value of F with the table value $F_{1,10,0.01}$. [3]

(iii) Comment on your findings. [4]

i) $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$

ii) $F = 10.451$ $F = 10.451$ $F_{1,10,0.01} = 10.044$

$F_{cal} > F_{1,10,0.01}$, Reject null Hypothesis $H_1: \beta_1 \neq 0$

There is significant effect on mean weekly sales when we increase Shelf Space.

Next Week Targets

1. Watch videos 7 and 8 posted at QMplus page
2. Try Questions of Exercise Sheet 1 and Exercise Sheet 2
3. Try Questions of Introduction to R
4. Attend tutorial session to discuss problems from (1), (2) and (3) above.