Bentley University MA252 in R

Content extracted from the How to Data Website

This PDF generated on 30 November 2021

Contents

MA252 is an undergraduate statistics course at Bentley University that focuses on model building using regression. The description from the course catalog can be found here.

It covers simple linear regression, multiple linear regression, logistic linear regression, model building, transformations, and interactions.

Simple Linear Regression

Multiple Linear Regression

Model Building

Residual Analysis

Content last modified on 30 November 2021.

How to fit a linear model to two columns of data

Description

Let's say we have two columns of data, one for a single independent variable x and the other for a single dependent variable y. How can I find the best fit linear model that predicts y based on x?

In other words, what are the model coefficients β_0 and β_1 that give me the best linear model $\hat{y} = \beta_0 + \beta_1 x$ based on my data?

Related tasks:

- How to compute R-squared for a simple linear model
- How to fit a multiple linear regression model
- How to predict the response variable in a linear model

Solution in R

This solution uses fake example data. When using this code, replace our fake data with your real data.

```
# Here is the fake data you should replace with your real data.
xs <- c( 393, 453, 553, 679, 729, 748, 817 )
ys <- c(  24,  25,  27,  36,  55,  68,  84 )

# If you need the model coefficients stored in variables for later use, do:
model <- lm( ys ~ xs )
beta0 = model$coefficients[1]
beta1 = model$coefficients[2]

# If you just need to see the coefficients, do this alone:
lm( ys ~ xs )
```

```
Call:
lm(formula = ys ~ xs)

Coefficients:
(Intercept)           xs
   -37.3214       0.1327
```

The linear model in this example is approximately $y = 0.133x - 37.32$.

Content last modified on 28 May 2021.

See a problem? Tell us or edit the source.

How to compute a confidence interval for the expected value of a response variable

Description

If we have a simple linear regression model, $y = \beta_0 + \beta_1 x + \epsilon$, where ϵ is some random error, then given any x input, y can be veiwed as a random variable because of ϵ. Let's consider its expected value. How do we construct a confidence interval for that expected value, given a value for the predictor x?

Related tasks:

- How to compute a confidence interval for a mean difference (matched pairs) (on website)
- How to compute a confidence interval for a regression coefficient
- How to compute a confidence interval for a population mean (on website)
- How to compute a confidence interval for a single population variance (on website)
- How to compute a confidence interval for the difference between two means when both population variances are known (on website)
- How to compute a confidence interval for the difference between two means when population variances are unknown (on website)
- How to compute a confidence interval for the difference between two proportions (on website)
- How to compute a confidence interval for the population proportion (on website)
- How to compute a confidence interval for the ratio of two population variances (on website)

Solution in R

Let's assume that you already have a linear model. We construct an example one here from some fabricated data.

```
# Make the linear model
x <- c(34, 9, 78, 60, 22, 45, 83, 59, 25)
y <- c(126, 347, 298, 309, 450, 187, 266, 385, 400)
model <- lm(y ~ x)
```

Construct a data frame containing just one entry, the value of the independent variable for which you want to compute the confidence interval. That data frame can then be passed to R's `predict` function to get a confidence interval for the expected value of y.

```
# Use your chosen value of x below:
data <- data.frame(x=40)
# Compute the confidence interval for y:
predict(model, data, interval="confidence", level=0.95) # or choose a different confidence level; here we use 0
```

```
  fit      lwr     upr
1 313.7217 226.648 400.7954
```

Our 95% confidence interval is $[226.648, 400.7954]$. We can be 95% confident that the true average value of y, given that x is 40, is between 226.648 and 400.7954.

Content last modified on 09 September 2021.

See a problem? Tell us or edit the source.

How to compute R-squared for a simple linear model

Description

Let's say we have fit a linear model to two columns of data, one for a single independent variable x and the other for a single dependent variable y. How can we compute R^2 for that model, to measure its goodness of fit?

Related tasks:

- How to fit a linear model to two columns of data
- How to compute adjusted R-squared

Solution in R

We assume you have already fit a linear model to the data, as in the code below, which is explained fully in a separate task, how to fit a linear model to two columns of data.

```
xs <- c( 393, 453, 553, 679, 729, 748, 817 )
ys <- c(  24,  25,  27,  36,  55,  68,  84 )
model <- lm( ys ~ xs )
```

You can get a lot of information about your model from its summary.

```
summary( model )
```

```
Call:
lm(formula = ys ~ xs)

Residuals:
      1       2       3       4       5       6       7 
  9.163   2.199  -9.072 -16.795  -4.431   6.047  12.890 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) -37.32142   18.99544  -1.965  0.10664   
xs            0.13272    0.02959   4.485  0.00649 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.62 on 5 degrees of freedom
Multiple R-squared:  0.8009,	Adjusted R-squared:  0.7611 
F-statistic: 20.12 on 1 and 5 DF,  p-value: 0.006486
```

In particular, it contains the R^2 value.

```
summary( model )$r.squared
```

```
[1] 0.8009488
```

Content last modified on 01 June 2021.

See a problem? Tell us or edit the source.

How to predict the response variable in a linear model

Description

If we have a linear model and a value for each explanatory variable, how do we predict the corresponding value of the response variable?

Related tasks:

- How to fit a linear model to two columns of data
- How to fit a multiple linear regression model

Solution in R

Let's assume that you've already built a linear model. We do an example below with fake data, but you can use your own actual data. For more information on the following code, see how to fit a multiple linear regression linear model.

```
x1 <- c( 2,  7,  4,  3, 11, 18,   6, 15,   9,  12)
x2 <- c( 4,  6, 10,  1, 18, 11,   8, 20,  16,  13)
x3 <- c(11, 16, 20,  6, 14,  8,   5, 23,  13,  10)
y  <- c(24, 60, 32, 29, 90, 45, 130, 76, 100, 120)
model <- lm(y ~ x1 + x2 + x3)
```

Let's say we want to estimate y given that $x_1 = 5$, $x_2 = 12$, and $x_3 = 50$. We can use R's `predict()` function as shown below.

```
predict(model, newdata = data.frame(x1 = 5, x2 = 12, x3 = 50))
```

```
        1 
-91.71014 
```

For the given values of the explanatory variables, our predicted response variable is -91.71014.

Note that if you want to compute the predicted values for all the data on which the model was trained, simply call `predict(model)` with no new data, and it defaults to using the training data.

```
predict(model)
```

```
        1         2         3         4         5         6         7         8 
 47.57012  24.35988  42.21531  47.27614 110.86526  70.03098  95.12690  70.91291 
        9        10 
106.52987  91.11264 
```

Content last modified on 24 October 2021.

See a problem? Tell us or edit the source.

How to fit a multiple linear regression model

Description

Let's say we have several independent variables, $x_1, x_2, \ldots, x_k$, and a dependent variable y. How can I fit a linear model that uses these independent variables to best predict the dependent variable?

In other words, what are the model coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ that give me the best linear model $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x + \cdots + \beta_k x$ based on my data?

Related tasks:

- How to fit a linear model to two columns of data
- How to predict the response variable in a linear model

Solution in R

We're going to use fake data here for illustrative purposes. You can replace our fake data with your real data in the code below.

```r
# Replace this fake data with your real data
x1 <- c(2, 7, 4, 3, 11, 18, 6, 15, 9, 12)
x2 <- c(4, 6, 10, 1, 18, 11, 8, 20, 16, 13)
x3 <- c(11, 16, 20, 6, 14, 8, 5, 23, 13, 10)
y <- c(24, 60, 32, 29, 90, 45, 130, 76, 100, 120)

# If you'll need the model coefficients later, store them as variables like this:
model <- lm(y ~ x1 + x2 + x3)
beta0 <- model$coefficients[1]
beta1 <- model$coefficients[2]
beta2 <- model$coefficients[3]
beta3 <- model$coefficients[4]

# To see the model summary, which includes the coefficients and much more, do this:
summary(model)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-25.031 -20.218  -8.373  22.937  35.640

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.244     27.366   2.823   0.0302 *
x1            -2.701      2.855  -0.946   0.3806
x2             7.299      2.875   2.539   0.0441 *
x3            -4.861      2.187  -2.223   0.0679 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.13 on 6 degrees of freedom
Multiple R-squared:  0.5936,    Adjusted R-squared:  0.3904
F-statistic: 2.921 on 3 and 6 DF,  p-value: 0.1222
```

The coefficients and intercept appear on the left hand side of the output, about half way down, under the heading “Estimate.”

Thus the multiple linear regression model from the example data is $\hat{y} = 77.244 - 2.701x_1 + 7.299x_2 - 4.861x_3$.

Content last modified on 09 September 2021.

See a problem? Tell us or edit the source.

How to compute a confidence interval for a regression coefficient

Description

Say we have a linear regression model, either single variable or multivariate. How do we compute a confidence interval for the coefficient of one of the explanatory variables in the model?

Related tasks:

- How to compute a confidence interval for a mean difference (matched pairs) (on website)
- How to compute a confidence interval for a population mean (on website)
- How to compute a confidence interval for a single population variance (on website)
- How to compute a confidence interval for the difference between two means when both population variances are known (on website)
- How to compute a confidence interval for the difference between two means when population variances are unknown (on website)
- How to compute a confidence interval for the difference between two proportions (on website)
- How to compute a confidence interval for the expected value of a response variable
- How to compute a confidence interval for the population proportion (on website)
- How to compute a confidence interval for the ratio of two population variances (on website)

Solution in R

We'll assume that you have fit a single linear model to your data, as in the code below, which uses fake example data. You can replace it with your actual data.

```
x <- c(34, 9, 78, 60, 22, 45, 83, 59, 25)
y <- c(126, 347, 298, 309, 450, 187, 266, 385, 400)
model <- lm(y ~ x)
```

We can use R's `confint()` function to find the confidence interval for the model coefficients. You can change the `level` parameter to specify a different confidence level. Note that if you have a multiple regression model, it will make confidence intervals for all of the coefficient values.

```
confint(model, level = 0.95)   # or choose any confidence level; here we use 0.95
```

```
                2.5 %      97.5 %
(Intercept) 172.638075 535.526421
x            -4.491961   2.473935
```

The 95% confidence interval for the regression coefficient is $[-4.491961, 2.473935]$.

Content last modified on 09 September 2021.

See a problem? Tell us or edit the source.

How to compute adjusted R-squared

Description

If we have fit a multiple linear regression model, how can we compute the Adjusted R^2 for that model, to measure its goodness of fit?

Related tasks:

- How to compute R-squared for a simple linear model

Solution in R

We assume you have already fit a multiple linear regression model to the data, as in the code below. (If you're unfamiliar with how to do so, see how to fit a multiple linear regression model.) The data shown below is fake, and we assume you will replace it with your own real data if you use this code.

```
x1 <- c(2, 7, 4, 3, 11, 18, 6, 15, 9, 12)
x2 <- c(4, 6, 10, 1, 18, 11, 8, 20, 16, 13)
x3 <- c(11, 16, 20, 6, 14, 8, 5, 23, 13, 10)
y <- c(24, 60, 32, 29, 90, 45, 130, 76, 100, 120)
model <- lm(y ~ x1 + x2 + x3)
```

You can get a lot of information about your model from its summary.

```
summary(model)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.031 -20.218  -8.373  22.937  35.640 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   77.244     27.366   2.823   0.0302 *
x1            -2.701      2.855  -0.946   0.3806  
x2             7.299      2.875   2.539   0.0441 *
x3            -4.861      2.187  -2.223   0.0679 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.13 on 6 degrees of freedom
Multiple R-squared:  0.5936,	Adjusted R-squared:  0.3904 
F-statistic: 2.921 on 3 and 6 DF,  p-value: 0.1222
```

In particular, that printout contains the Adjusted R^2 value; it is the second value in the right-hand column, near the top.

You can also obtain it directly, as follows:

```r
summary(model)$adj.r.squared
```

```
[1] 0.3903924
```

In this case, the Adjusted R^2 is 0.3904.

Content last modified on 09 September 2021.

See a problem? Tell us or edit the source.

How to do a test of joint significance

Description

If we have a multivariate linear model, how do we test the joint significance of all the variables in the model? In other words, how do we test the overall significance of the regression model?

Solution in R

Let's assume that you already made your multiple regression model, similar to the one shown below. You can visit this task, , to see how to construct a multiple linear regression model.

Let's assume that you already made your multivariate linear model, similar to the one shown below. If you still need to create one, first see how to fit a multiple linear regression model.

We use example data here, but you would use your own data instead.

```
x1 <- c( 2,  7,  4,  3, 11, 18,   6, 15,   9,  12)
x2 <- c( 4,  6, 10,  1, 18, 11,   8, 20,  16,  13)
x3 <- c(11, 16, 20,  6, 14,  8,   5, 23,  13,  10)
y  <- c(24, 60, 32, 29, 90, 45, 130, 76, 100, 120)
model <- lm(y ~ x1 + x2 + x3)
```

Now we want to test whether the model is significant. We will use a null hypothesis that states that all of the model's coefficients are equal to zero, that is, they are not jointly significant in predicting y. We can write $H_0 : \beta_0 = \beta_1 = \beta 2 = \beta_3 = 0$.

We also choose a value $0 \leq \alpha \leq 1$ as our Type 1 error rate. Herer we'll use $\alpha = 0.05$.

The summary output for the model will give us both the F-statistic and the p-value.

```
summary(model)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.031 -20.218  -8.373  22.937  35.640 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   77.244     27.366   2.823   0.0302 *
x1            -2.701      2.855  -0.946   0.3806  
x2             7.299      2.875   2.539   0.0441 *
x3            -4.861      2.187  -2.223   0.0679 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.13 on 6 degrees of freedom
Multiple R-squared:  0.5936,	Adjusted R-squared:  0.3904 
F-statistic: 2.921 on 3 and 6 DF,  p-value: 0.1222
```

In the final line of the output, we can see that the F-statistic is 2.921. The corresponding p-value in the same line is 0.1222, which is greater than α, so we do not have sufficient evidence to reject the null hypothesis.

We cannot conclude that the independent variables in our model are jointly significant in predicting the response variable.

Content last modified on 05 October 2021.

See a problem? Tell us or edit the source.

How to compute covariance and correlation coefficients

Description

Covariance is a measure of how much two variables "change together." It is positive when the variables tend to increase or decrease together, and negative when they upward motion of one variable is correlated with downward motion of the other. Correlation normalizes covariance to the interval $[-1, 1]$.

Solution in R

How to Data does not yet contain a solution for this task in R.

How to compute the standard error of the estimate for a model

Description

One measure of the goodness of fit of a model is the standard error of its estimates. If the actual values are y_i and the estimates are $\hat{y}_i$, the definition of this quantity is as follows, for n data points.

$$\sigma_{\text{est}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

If we've fit a linear model, how do we compute the standard error of its estimates?

Solution in R

Let's assume that you already fit the linear model, as shown in the code below. This one uses a small amount of fake data, but it's just an example. See also how to fit a linear model to two columns of data.

```
x <- c(34, 9, 78, 60, 22, 45, 83, 59, 25)
y <- c(126, 347, 298, 309, 450, 187, 266, 385, 400)
model <- lm(y ~ x)
```

The standard error is shown as part of the model summary, reported by R's built-in `summary` function; see the third line from the bottom.

```
summary(model)
```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-193.776   -4.334   15.459   71.143  118.116 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  354.082     76.733   4.614  0.00244 **
x             -1.009      1.473  -0.685  0.51536   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 107.1 on 7 degrees of freedom
Multiple R-squared:  0.06283,	Adjusted R-squared:  -0.07106 
F-statistic: 0.4693 on 1 and 7 DF,  p-value: 0.5154
```

We can also extract just that one value using the code shown below.

```
summary(model)$sigma
```

```
[1] 107.1191
```

The standard error of the estimate is 107.119.

Content last modified on 14 September 2021.

See a problem? Tell us or edit the source.

How to do a hypothesis test of a coefficient's significance

Description

Let's say we have a linear model, either one variable or many. How do we conduct a test of significance for the coefficient of a single explanatory variable in the model? Similarly, how can we determine if an explanatory variable has a significant impact on the response variable?

Related tasks:

- How to compute a confidence interval for the difference between two proportions (on website)
- How to do a hypothesis test for a mean difference (matched pairs) (on website)
- How to do a hypothesis test for a population proportion (on website)
- How to do a hypothesis test for population variance (on website)
- How to do a hypothesis test for the difference between means when both population variances are known (on website)
- How to do a hypothesis test for the difference between two proportions (on website)
- How to do a hypothesis test for the mean with known standard deviation (on website)
- How to do a hypothesis test for the ratio of two population variances (on website)
- How to do a one-sided hypothesis test for two sample means (on website)
- How to do a two-sided hypothesis test for a sample mean (on website)
- How to do a two-sided hypothesis test for two sample means (on website)

Solution in R

We will use the fake data shown below with a single variable model. You can use a model created from your own actual data instead.

```
x <- c( 34,   9,  78,  60,  22,  45,  83,  59,  25)
y <- c(126, 347, 298, 309, 450, 187, 266, 385, 400)
model <- lm(y ~ x)
```

We can test whether a coefficient is zero by using that as our null hypothesis, $H_0 : \beta_i = 0$. We can use any value $0 \leq \alpha \leq 1$ as our Type 1 error rate; we will set α to be 0.05 here.

The answer to our hypothesis test can be obtained by looking at just the coefficients portion of the model summary:

```
summary(model)$coef
```

```
              Estimate   Std. Error t value    Pr(>|t|)
(Intercept) 354.082248 76.732772   4.6144853 0.002441995
x            -1.009013  1.472939  -0.6850334 0.515358250
```

The final column of output shows p-values for each β_i. The p-value associated with the x row is therefore for β_1, the coefficient on x. Because it is 0.515358250, which is greater than α, we cannot reject the null hypothesis, and we should continue to assume that $\beta_1 = 0$ and there is no significant relationship between the explanatory and response variable in this situation.

Content last modified on 24 October 2021.

See a problem? Tell us or edit the source.

How to do a Spearman rank correlation test

Description

When we want to determine whether there is a relationship between two variables, but our samples do not come from normally distributed populations, we can use the Spearman Rank Correlation Test. How do we conduct it?

Solution in R

We will use some fake data about height and weight measurements for people. You can replace it with your real data.

Our data should be stored in R vectors, as shown below.

```
heights <- c(60, 76, 57, 68, 70, 62, 63)
weights <- c(145, 178, 120, 143, 174, 130, 137)
```

Let's say we want to test the correlation between height (inches) and weight (pounds). Our null hypothesis would state that the Pearson correlation coefficient is equal to zero, or that there is no relationship between height and weight, $H_0 : \rho_s = 0$. We choose α, or the Type I error rate, to be 0.05 and carry out the Spearman Rank Correlation Test to get the test-statistic and p-value.

```
# Run the Spearman Rank Correlation Test to get the test-statistic and p-value
cor.test(heights, weights, alternative = "two.sided", method = "spearman")
```

```
    Spearman's rank correlation rho

data:  heights and weights
S = 12, p-value = 0.04802
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7857143
```

Our p-value is 0.04802, which is less than $\alpha = 0.05$, so we reject the null hypothesis. There does appear to be a relationship between height and weight.

(This p-value is different than the one computed in the solution using Python, because different approximation methods are used by the two software packages when the sample size is small.)

Note that for a right-tailed test, you can replace "two.sided" with "greater" and for a left-tailed test, you can replace "two.sided" with "less".

Content last modified on 05 October 2021.

See a problem? Tell us or edit the source.

How to compute the residuals of a linear model

Description

If a model has been fit to a dataset, the *residuals* are the differences between the actual data points and the results the model would predict. Given a linear model and a dataset, how can we compute those residuals?

Solution in R

Let's assume that you've already built a linear model similar to the one below. This one uses a small amount of fake data, but it's just an example. See also how to fit a linear model to two columns of data.

```
xs <- c( 393, 453, 553, 679, 729, 748, 817 )
ys <- c(  24,  25,  27,  36,  55,  68,  84 )
model <- lm(ys ~ xs)
```

We can extract the residuals of the model in either of two ways.

R has a built-in `residuals()` function for this purpose.

```
residuals(model)
```

```
         1          2          3          4          5          6          7 
  9.162630   2.199457  -9.072500 -16.795165  -4.431143   6.047185  12.889535 
```

The model itself has a `$residuals` attribute.

```
model$residuals
```

```
         1          2          3          4          5          6          7 
  9.162630   2.199457  -9.072500 -16.795165  -4.431143   6.047185  12.889535 
```

Content last modified on 14 September 2021.

See a problem? Tell us or edit the source.