BOOSTEDML

Articles on Statistics and Machine Learning for Healthcare

Search ... Sea

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

LINEAR AND GENERALIZED LINEAR MODELS

Linear Regression Summary(lm): Interpretting

POSTED ON JUNE 1, 2019 BY ALEX

Categories

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

Contents [hide]

Introduction to Linear Regression Summ Printouts

In this post we describe how to interpret the summary of a linear regression model in summary(lm). We discuss interpretation of the residual quantiles and summary statist standard errors and t statistics , along with the p-values of the latter, the residual star and the F-test. Let's first load the Boston housing dataset and fit a naive model. We w about assumptions, which are described in other posts.

```
library(mlbench)
data(BostonHousing)
model<-lm(log(medv) ~ crim + rm + tax + lstat , data = Bo
summary(model)

Call:
lm(formula = log(medv) ~ crim + rm + tax + lstat, data = BostonHousi

Residuals:
     Min       1Q   Median       3Q      Max
-0.72730 -0.13031 -0.01628  0.11215  0.92987

Coefficients:
```

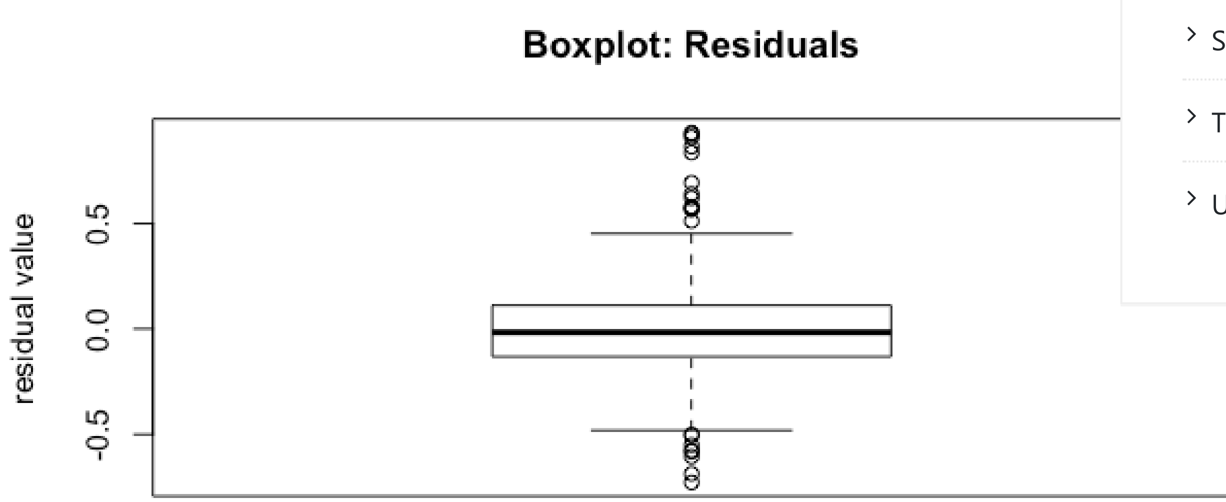
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.646e+00  1.256e-01  21.056  < 2e-16 ***
crim        -8.432e-03  1.406e-03  -5.998 3.82e-09 ***
rm           1.428e-01  1.738e-02   8.219 1.77e-15 ***
tax         -2.562e-04  7.599e-05  -3.372 0.000804 ***
lstat       -2.954e-02  1.987e-03 -14.867  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2158 on 501 degrees of freedom
Multiple R-squared:  0.7236,    Adjusted R-squared:  0.72
F-statistic: 327.9 on 4 and 501 DF,  p-value: < 2.2e-16
```

Residual Summary Statistics

The first info printed by the linear regression summary after the formula is the residu statistics. One of the assumptions for hypothesis testing is that the errors follow a Ga distribution. As a consequence the residuals should as well. The residual summary sta information about the symmetry of the residual distribution. The median should be cl the mean of the residuals is 0, and symmetric distributions have median=mean. Furth and 1Q should be close to each other in magnitude. They would be equal under a syr mean distribution. The max and min should also have similar magnitude. However, in holding may indicate an outlier rather than a symmetry violation.

We can investigate this further with a boxplot of the residuals.

```
boxplot(model[['residuals']],main='Boxplot: Residuals',yla
```



Categories
general
hypothesis test
linear and generalized lin models
Longitudinal D Analysis
Missing Data
Neural Networks+Dee Learning
Numerical Line Algebra
optimization
Survival
Time series
Uncategorized

We see that the median is close to 0. Further, the 25 and 75 percentile look approxim same distance from 0, and the non-outlier min and max also look about the same dis All of this is good as it suggests correct model specification.

Coefficients

The second thing printed by the linear regression summary call is information about t coefficients. This includes their estimates, standard errors, t statistics, and p-values.

Estimates

The intercept tells us that when all the features are at 0, the expected response is the Note that for an arguably better interpretation, you should consider [centering your fe](#) changes the interpretation. Now, when features are at their mean values, the expecte the intercept. For the other features, the estimates give us the expected change in the due to a unit change in the feature.

Standard Error

The standard error is the standard error of our estimate, which allows us to construct confidence intervals for the estimate of that particular feature. If $s.e.(\hat{\beta}_i)$ is the stand $\hat{\beta}_i$ is the estimated coefficient for feature i, then a 95% confidence interval is given by $\hat{\beta}_i \pm 1.96 \cdot s.e.(\hat{\beta}_i)$. Note that this requires two things for this confidence interval to

- your model assumptions hold
- you have enough data/samples to invoke the central limit theorem, as you be approximately Gaussian.

That is, assuming all model assumptions are satisfied, we can say that with 95% confi is not probability) the true parameter β_i lies in $[\hat{\beta}_i - 1.96 \cdot s.e.(\hat{\beta}_i), \hat{\beta}_i + 1.96 \cdot s.e$ on this, we can construct confidence intervals

```
confint(model)
                     2.5 %        97.5 %
(Intercept)  2.3987332457  2.8924423620
crim        -0.0111943622 -0.0056703707
rm           0.1086963289  0.1769912871
tax         -0.0004055169 -0.0001069386
lstat       -0.0334396331 -0.0256328293
```

Here we can see that the entire confidence interval for number of rooms has a large effect size relative to the other covariates.

Categories

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

t-value

The t-statistic is

$$\frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)} \quad (1)$$

which tells us about how far our estimated parameter is from a hypothesized 0 value, standard deviation of the estimate. Assuming that $\hat{\beta}_i$ is Gaussian, under the null hypo $\beta_i = 0$, this will be t distributed with $n-p-1$ degrees of freedom, where n is the observations and p is the number of parameters we need to estimate.

Pr(>|t|)

This is the p-value for the individual coefficient. Under the t distribution with $n-p$ - freedom, this tells us the probability of observing a value at least as extreme as our $\hat{\beta}$ probability is sufficiently low, we can reject the null hypothesis that this coefficient is note that when we care about looking at *all* of the coefficients, we are actually doing hypothesis tests, and need to correct for that. In this case we are making five hypothe for each feature and one for the coefficient. Instead of using the standard p-value of use the Bonferroni correction and divide by the number of hypothesis tests, and thus value threshold to 0.01.

Assessing Fit and Overall Significance

The linear regression summary printout then gives the residual standard error, the R statistic and test. These tell us about how good a fit the model is and whether any of coefficients are significant.

Residual Standard Error

The residual standard error is given by $\hat{\sigma} = \sqrt{\frac{\sum \hat{\epsilon}_i^2}{n-p}}$. It gives the standard deviation c residuals, and tells us about how large the prediction error is *in-sample* or *on the trai* We'd like this to be significantly different from the variability in the marginal response otherwise it's not clear that the model explains much.

Multiple and Adjusted R^2

Intuitively R^2 tells us what proportion of the variance is explained by our model, and is given by

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$
$$= 1 - \frac{\sum_i \hat{\epsilon}_i^2}{} \quad (2)$$

Categories

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

$$\sum_i (y_i - \bar{y})^2$$

both R^2 and the residual standard standard deviation tells us about how well our mc data. The adjusted R^2 deals with an increase in R^2 spuriously due to adding features fitting noise in the data. It is given by

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

thus as the number of features p increases, the required R^2 needed will increase as v maintain the same adjusted R^2.

F-Statistic and F-test

In addition to looking at whether individual features have a significant effect, we may whether *at least one* feature has a significant effect. That is, we would like to test the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

that all coefficients are 0 against the alternative hypothesis

$$H_1 : \exists i : 1 \leq i \leq p-1 : \beta_i \neq 0$$

Under the null hypothesis the F statistic will be F distributed with $(p-1, n-p)$ deg freedom. The probability of our observed data under the null hypothesis is then the p use the F-test alone without looking at the t-tests, then we do not need a Bonferroni while if we do look at the t-tests, we need one.

« Previous

Categories

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

Related Posts

[Linear Regression: Comparing Models Between Two Groups with linearHypothesis](#)
September 21, 2019

[Why You Should Center Your Features in Linear Regression](#)
August 31, 2019

Leave a Reply

Your email address will not be published. Required fields are marked*

Comment*

Name*

Email*

Website

Post Comment

This site uses Akismet to reduce spam. [Learn how your comment data is processed](#).

Categories

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

Categories

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

Categories

- general
- hypothesis test
- linear and generalized lin models
- Longitudinal D Analysis
- Missing Data
- Neural Networks+Dee Learning
- Numerical Line Algebra
- optimization
- Survival
- Time series
- Uncategorized

Categories