Regression

Nicholas Lambert

6/22/2021

Simple regression

Regression analysis is a supervised machine learning approach

Special case of the general linear model

$$outcome_i = (model) + error_i$$

Predict (estimate) value of one outcome (dependent or target) variable as:

- one predictor or feature (independent) variable: simple / univariate

$$Y_i = (b_0 + b_1 * X_{i1}) + \epsilon_i$$

- more predictors or features (independent) variables: multiple / multivar.

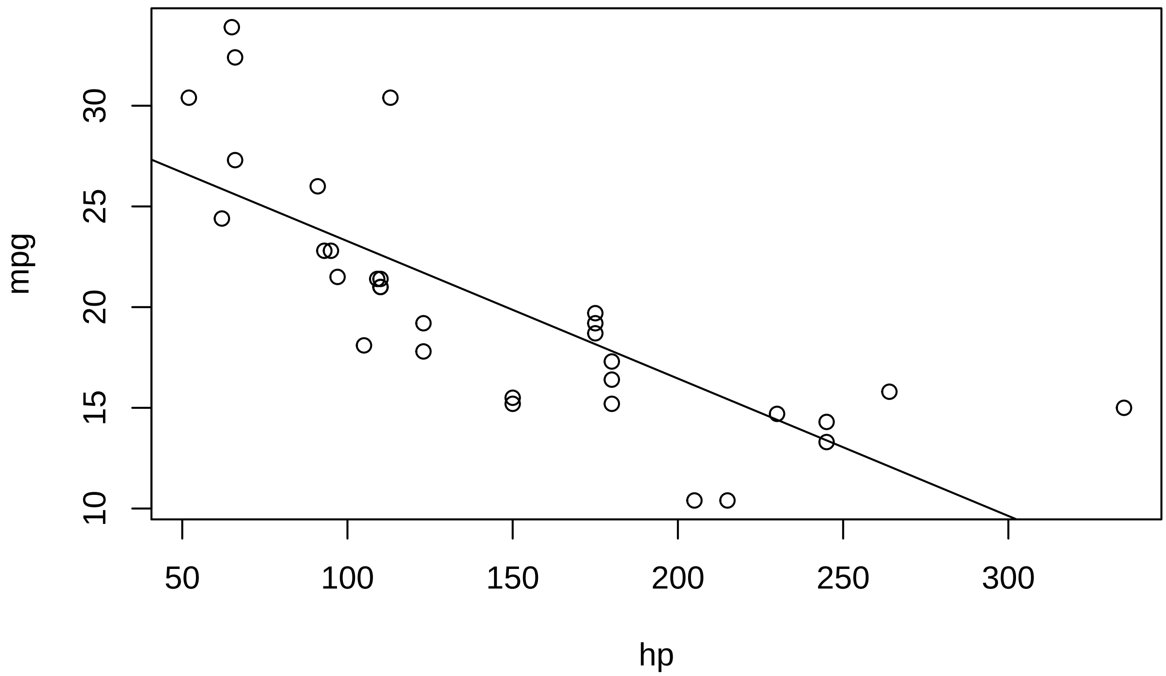
$$Y_i = (b_0 + b_1 * X_{i1} + b_2 * X_{i2} + \cdots + b_M * X_{iM}) + \epsilon_i$$

Example

Can we predict a a car's miles per gallon (mpg) from horsepower?

body massi=(b0+b1∗flipper lengthi)+ i

$$mpg_i = (b_0 + b_1 * horsepower) + \epsilon_i$$



Least squares

Least squares is the most commonly used approach to generate a regression model

The model fits a line:

- to minimise the squared values of the residuals (errors)
- that is squared difference between
- observed values

$$residual_i = observed_i - model_i$$

$$deviation = \sum_i (observed_i - model_i)^2$$

Assumptions

- **Linearity**
  - the relationship is actually linear
- **Normality of residuals**
  - standard residuals are normally distributed with mean 0
- **Homoscedasticity of residuals**
  - at each level of the predictor variable(s) the variance of the standard residuals should be the same (homo-scedasticity) rather than different (hetero-scedasticity)
- **Independence of residuals**
  - adjacent standard residuals are not correlated
- **When more than one predictor: no multicollinearity**
  - if two or more predictor variables are used in the model, each pair of variables not correlated

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421 < 0.0000000000000002 ***
## hp          -0.06823    0.01012  -6.742          0.000000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 0.0000001788
```

Overall fit

The output indicates:

- p-value: 0.0000001788: $p < .01$ the model is significant
  - derived by comparing F-statistic (45.46) to F distribution having specified degrees of freedom (1, 30)
  - Report as: $F(1, 30) = 45.46$
- Adjusted R-squared: 0.5892:
  - hoursepower can account for 58.92% variation in miles per gallon
- Coefficients
  - Intercept estimate 30.09886 is significant
  - horsepower (slope) estimate -0.06823 is significant

Outliers and influential cases

```
##                mpg model_stdres model_cook_dist
## Maserati Bora  15     2.357853        1.052231
```

One influential case (Cook's distance > 1) an no outliers (0 abs std res > 2.58)

Checking assumptions: normality

Shapiro-Wilk test for normality of standard residuals, robust models: should be not significant

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars_output$model_stdres
## W = 0.92058, p-value = 0.02156
```

Standard residuals are **not** normally distributed.

Checking assumptions: homoscedasticity

Breusch-Pagan test for homoscedasticity of standard residuals, robust models: should be not significant

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 0.049298, df = 1, p-value = 0.8243
```

Standard residuals are homoscedastic.

Checking assumptions: independence

Durbin-Watson test for the independence of residuals

- robust models: statistic should be close to 2 (advised between 1 and 3) and not significant

```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.1338, p-value = 0.00411
## alternative hypothesis: true autocorrelation is greater than 0
```

Standard residuals are **not** independent!

Note: the result depends on the order of the data.

Checking assumptions: multicollinearity

Checking the variance inflation factor (VIF)

- robust models should have no multicollinearity: largest VIF should be lower than 10 or the average VIF should not be greater than 1

Result

No, we cannot predict mpg from horsepower

- predictors are statistically significant[1], but
- model is not robust, as it doesn't satisfy most assumptions:
    - Standard residuals are NOT normally distributed
    - Standard residuals are NOT independent

$$mpg_i = (30.09886 - 0.06823 * horsepower) + \epsilon_i$$

Summary

Simple Regression:

- Regression
- Ordinary Least Squares
- Interpretation
- Checking assumptions

Comparing regression model

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
```

[1] test for multicollinearity does not apply in this example (univariate)

```
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421 < 0.0000000000000002 ***
## hp          -0.06823    0.01012  -6.742          0.000000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 0.0000001788
```

Is there a difference between:

Model1's R2= 0.5721 and Model2's R2= 0.6688?

Comparing R-squared

- $R2$
  - measure of correlation between:
  - values predicted by the model (fitted values)
  - observed values for outcome variable
- Adjusted R2
  - adjusts the R2 depending on
  - number of cases
  - number of predictor (independent) variables
  - "unnecessary" variables lower the value

The model with the highest adjusted R2 has the best fit.

Model difference with ANOVA

Can be used to test whether adjusted R2 are signif. different if models are hierarchical one uses all variables of the other plus some additional variables.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp + cyl
## Model 2: mpg ~ log10(hp) + cyl + wt
##   Res.Df    RSS Df Sum of Sq      F     Pr(>F)
## 1     29 291.98
## 2     28 155.08  1    136.89 24.716 0.00002998 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Still, neither model is robust.

Information criteria

- Akaike Information Criterion (**AIC**)
  - measure of model fit
- penalising model with more variables
  - not interpretable per-se, used to compare similar models
  - lower value, better fit
- Bayesian Information Criterion (**BIC**)

    - similar to AIC

```
## [1] 169.5618
```

```
## [1] 151.3149
```

Stepwise selection

Stepwise selection of predictor (independent) variables:

- iteratively adding and/or removing predictors
- to obtain best performing model

Three approaches

- forward: from no variable, iteratively add variables
- backward: from all variables, iteratively remove variables
    - both (a.k.a. step-wise):
    - from no variable
    - one step forward, add most promising variable
    - one step backward, remove any variable not improving

MASS::stepAIC

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9290 -1.5598 -0.5311  1.1850  5.8986
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 38.75179    1.78686  21.687 < 0.0000000000000002 ***
## wt          -3.16697    0.74058  -4.276             0.000199 ***
## cyl         -0.94162    0.55092  -1.709             0.098480 .
## hp          -0.01804    0.01188  -1.519             0.140015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.512 on 28 degrees of freedom
## Multiple R-squared:  0.8431, Adjusted R-squared:  0.8263
## F-statistic: 50.17 on 3 and 28 DF,  p-value: 0.00000000002184
```

```
## [1] 155.4766
```

Validation

Can the model be generalised?

- split data into:
    - training set: used to train the model
    - test set: used to test the model

Approaches:

- Validation
  - simple split: e.g. 80% traning, 20% test
- Cross-validation
  - leave-p-out: repeated split, leaving out p cases for test
  - leave-1-out
  - k-fold: repeated split, k equal size samples

caret::train

Use caret::train to cross-validate Model 3

```
## Linear Regression
##
## 32 samples
##  3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 25, 26, 27, 25, 25
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   2.280544  0.8762117  1.972755
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
##       RMSE  Rsquared       MAE Resample
## 1 3.248957 0.8572600 2.7480092    Fold1
## 2 1.947559 0.8815566 1.6988144    Fold2
## 3 1.013646 0.9804871 0.9150159    Fold3
## 4 1.624329 0.9313163 1.3054667    Fold4
## 5 3.568231 0.7304388 3.1964679    Fold5
```

Summary

Comparing regression models:

- Information criteria
- Model difference
- Stepwise selection Validation