# Linear Regression

Madhur

### Import Libraries

```r
library(tidyverse)
library(MASS)
library(ISLR2)
library(car)
```

## Simple Linear Regression

### Boston Dataset

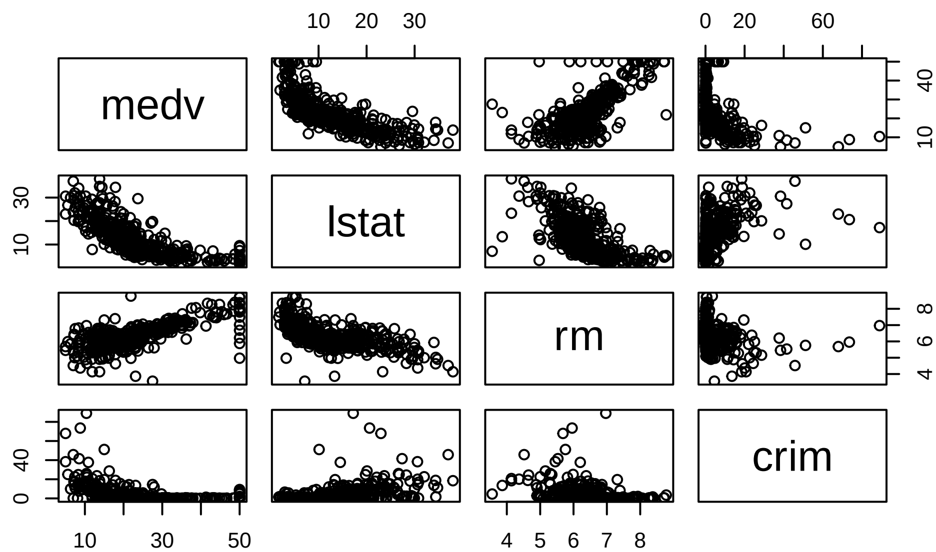
- 506 observations : 506 census tracts in Boston
- 12 predictors
- Target variable : medv = median house value

```r
glimpse(Boston)
```

```
Rows: 506
Columns: 13
$ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.08829,~
$ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5, 1~
$ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87, 7.~
$ chas    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.524,~
$ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.631,~
$ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9, 9~
$ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9505~
$ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4,~
$ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311, 31~
```

```
$ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2, 15~
$ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17.10~
$ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9, 15~
```

```r
pairs(Boston[, c("medv", "lstat", "rm", "crim")])
```




- **Simple Linear Regression**
  - predictor:lstat (lower status of the population %).
  - target : medv

```r
# simple linear regression with lstat predictor
attach(Boston)
lm.fit <- lm(medv ~ lstat, data = Boston)
summary(lm.fit)
```

```

Call:
lm(formula = medv ~ lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
names(lm.fit)
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

```r
coefficients(lm.fit)
```

```
(Intercept)       lstat
 34.5538409  -0.9500494
```

```r
confint(lm.fit)
```

```
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
```

- **Prediction Interval vs Confidence Interval**
  - Confidence interval : when lstat = 10, the confidence interval is (24.47, 25.63). This means we are 95% confident that the true average value of medv for lstat = 10 lies within this range.
  - For lstat = 10, the prediction interval is (12.83, 37.28). This range is wider because it accounts for the variability in individual data points, not just the variability in the estimated mean.
  - The predicted value (fit) for lstat = 10 is the same for both intervals: 25.05.

```r
predict(lm.fit, data.frame(lstat = (c(5,10,15,20))),
        interval = "confidence")
```

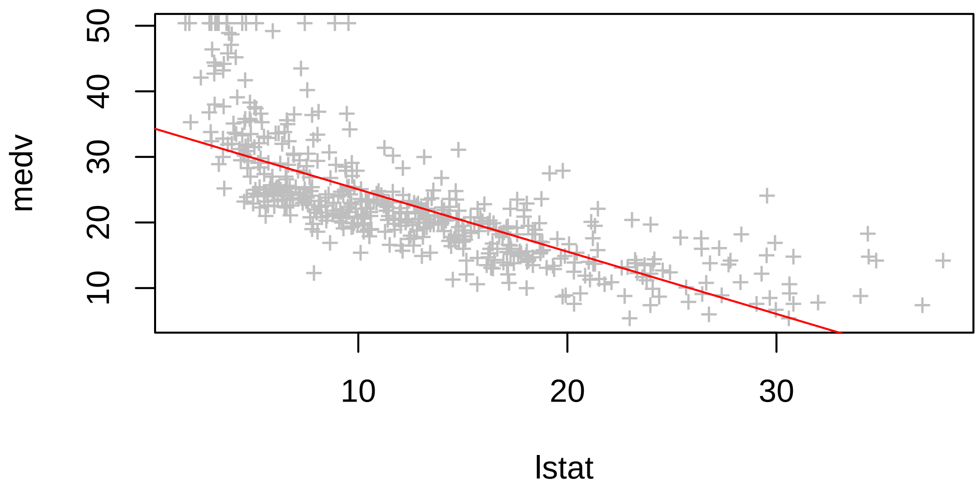
```
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
4 15.55285 14.77355 16.33216
```

```r
predict(lm.fit, data.frame(lstat = (c(5,10,15,20))),
        interval = "prediction")
```

```
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
4 15.55285  3.316021 27.78969
```

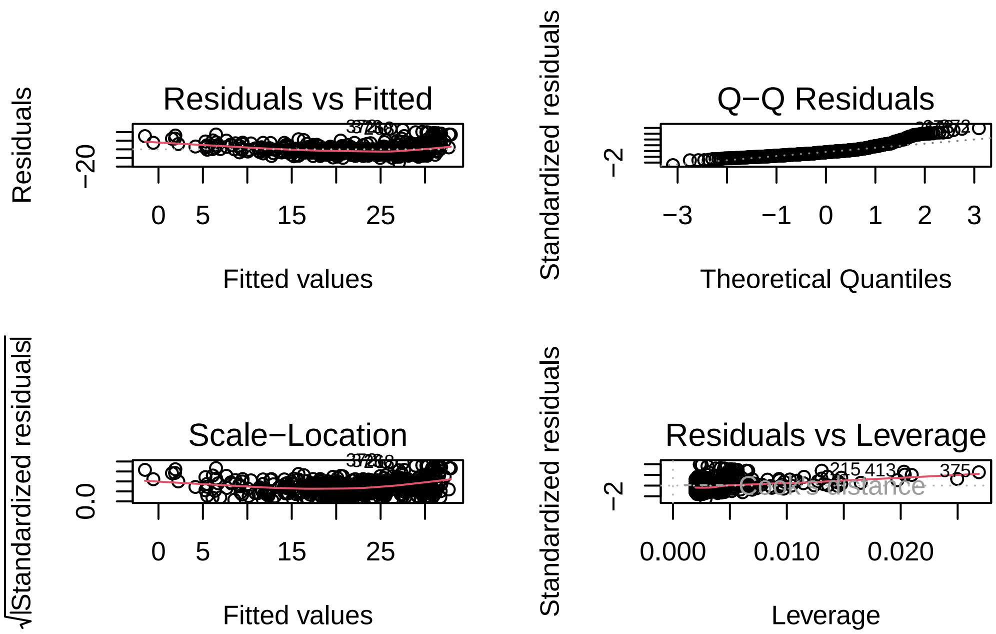
- **Plot**
  - plot(predictor, target variable)
  - add least square regression line `abline(model)`

```r
# plot - predictor and target varaible
plot(lstat, medv, col = "grey", pch = "+")
abline(lm.fit, col = "red")
```



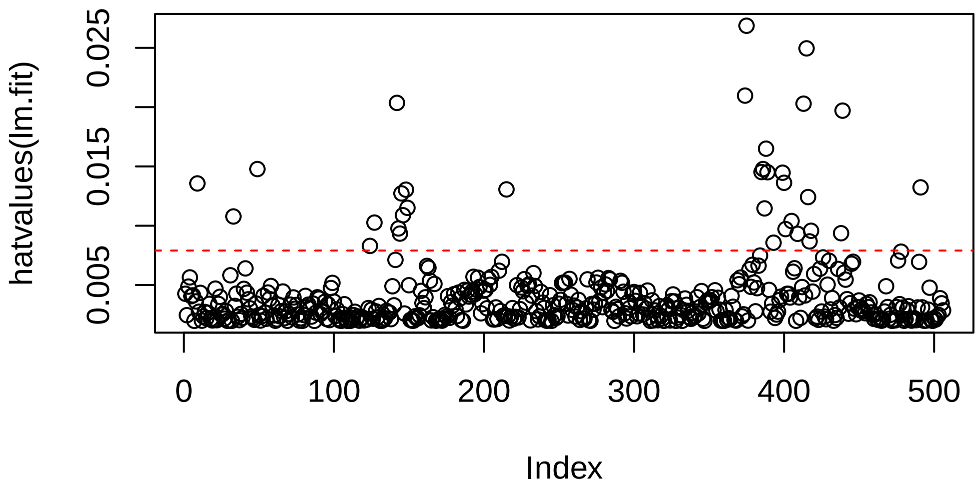

- **Diagnostic Plots**
  - Residual = Observed value − Predicted value
  - Fitted values mean prediction values
  - 1. Residuals vs. Fitted : Check for linearity and homoscedasticity (constant variance).
    - Random scatter : good model
    - Patterns (e.g., curvature) suggest non-linearity
    - Funnel-shaped patterns (widening or narrowing of residuals) suggest heteroscedasticity (non-constant variance).
  - 2. Normal Q-Q Plot : check whether the residuals are normally distributed.
  - 3. Scale-Location (Spread-Location) Plot : Purpose: To check for homoscedasticity (constant variance of residuals).
    - The points should show a horizontal line with random scatter.
    - A clear trend (e.g., an upward or downward slope) suggests heteroscedasticity.
  - 4. Residuals vs. Leverage
    - Identify influential data points.
    - Points with high leverage (far to the right or left) and large residuals are influential and could

```r
par(mfrow = c(2, 2))
plot(lm.fit)
```




- Leverage statistics can be computed for any number of predictors using the hatvalues() function. - influential data points.

```r
plot(hatvalues(lm.fit))
abline(h = 2 * (length(coef(lm.fit)) / nrow(Boston)), col = 'red', lty =2 )
```




```
which.max(hatvalues(lm.fit))
```

```
375
375
```

- The `which.max()` function : identifies the index of the largest element of a vector.
- It tells us which observation has the largest leverage statistic.

### Multiple Linear Regression

- predictor :
  - lstat : lower status of the population (percent).
  - age : proportion of owner-occupied units built prior to 1940.
- target : medv : median value of owner-occupied homes in $1000s.

```
mlm.fit <- lm(medv ~ lstat + age, data = Boston)
summary(lm.fit)
```

```
Call:
```

```
lm(formula = medv ~ lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.168  -3.990  -1.318   2.034  24.500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,	Adjusted R-squared:  0.5432 
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

- MLR with all 12 predictors
  - RSE
  - R2

```r
lm.fit <- lm(medv ~ ., data = Boston)
summary(lm.fit)
```

```
Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.1304  -2.7673  -0.5814   1.9414  26.2526 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  41.617270   4.936039   8.431 3.79e-16 ***
crim         -0.121389   0.033000  -3.678 0.000261 ***
zn            0.046963   0.013879   3.384 0.000772 ***
indus         0.013468   0.062145   0.217 0.828520    
chas          2.839993   0.870007   3.264 0.001173 ** 
nox         -18.758022   3.851355  -4.870 1.50e-06 ***
rm            3.658119   0.420246   8.705  < 2e-16 ***
age           0.003611   0.013329   0.271 0.786595    
```

```
dis          -1.490754   0.201623  -7.394 6.17e-13 ***
rad           0.289405   0.066908   4.325 1.84e-05 ***
tax          -0.012682   0.003801  -3.337 0.000912 ***
ptratio      -0.937533   0.132206  -7.091 4.63e-12 ***
lstat        -0.552019   0.050659 -10.897  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.798 on 493 degrees of freedom
Multiple R-squared:  0.7343,    Adjusted R-squared:  0.7278
F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

- **R2** measures the proportion of variance in the dependent variable (response) that is explained by the independent variables (predictors) in the model.
  - Multiple R-squared measures the proportion of the variance in the response variable that is explained by the predictors in the model. -This means 73.43% of the variance in the dependent variable is explained by the independent variables in the model.
  - **Adjusted R2** : it penalizes adding predictors that do not significantly improve the model's performance.
  - Adjusted R2 is slightly lower than  2(0.7278 compared to 0.7343) because it adjusts for the model's complexity. -If the difference between R2 and Adjusted R2 is large, it may indicate that unnecessary predictors are included in the model.
- **RSE** : RSE is a measure of the average deviation of the observed values from the fitted regression line, expressed in the same units as the response variable.
  - If RSE $= 4.7$ for a model predicting housing prices in `$1000s`, the predictions are, on average, \$4700 off from the actual values.
- **VIF** : A high VIF indicates that a predictor is highly collinear with other predictors, which can make regression coefficients unstable.
  - VIF $< 5$: Generally acceptable.
  - VIF $> 10$: Strong multicollinearity that requires attention.

```r
vif(lm.fit)
```

```
    crim       zn    indus     chas      nox       rm      age      dis 
1.767486 2.298459 3.987181 1.071168 4.369093 1.912532 3.088232 3.954037 
     rad      tax  ptratio    lstat 
7.445301 9.002158 1.797060 2.870777 
```

- Since `age` has high p-value remove it from the model

```r
mlm.fit <- lm(medv ~ . -age , data = Boston)
summary(mlm.fit)
```

```
Call:
lm(formula = medv ~ . - age, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.1851  -2.7330  -0.6116   1.8555  26.3838 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  41.525128   4.919684   8.441 3.52e-16 ***
crim         -0.121426   0.032969  -3.683 0.000256 ***
zn            0.046512   0.013766   3.379 0.000785 ***
indus         0.013451   0.062086   0.217 0.828577    
chas          2.852773   0.867912   3.287 0.001085 ** 
nox         -18.485070   3.713714  -4.978 8.91e-07 ***
rm            3.681070   0.411230   8.951  < 2e-16 ***
dis          -1.506777   0.192570  -7.825 3.12e-14 ***
rad           0.287940   0.066627   4.322 1.87e-05 ***
tax          -0.012653   0.003796  -3.333 0.000923 ***
ptratio      -0.934649   0.131653  -7.099 4.39e-12 ***
lstat        -0.547409   0.047669 -11.483  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.794 on 494 degrees of freedom
Multiple R-squared:  0.7343,    Adjusted R-squared:  0.7284 
F-statistic: 124.1 on 11 and 494 DF,  p-value: < 2.2e-16
```

#### Interaction Terms

```r
imlr.fit <- lm(medv ~ lstat*age, data = Boston)
summary(imlr.fit)
```

```
Call:
```

```
lm(formula = medv ~ lstat * age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.806  -4.045  -1.333   2.085  27.552 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
age         -0.0007209  0.0198792  -0.036   0.9711    
lstat:age    0.0041560  0.0018518   2.244   0.0252 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared:  0.5557,    Adjusted R-squared:  0.5531 
F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

### Non-linear Transformations of the Predictors

- Predictors : lstat and lstat^2
- Use ANOVA to quantify if quadratic is better fit than linear.

```r
qlm.fit <- lm(medv ~ lstat + I(lstat^2), data = Boston)
summary(qlm.fit)
```

```
Call:
lm(formula = medv ~ lstat + I(lstat^2), data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2834  -3.8313  -0.5295   2.3095  25.4148 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,	Adjusted R-squared:  0.6393 
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

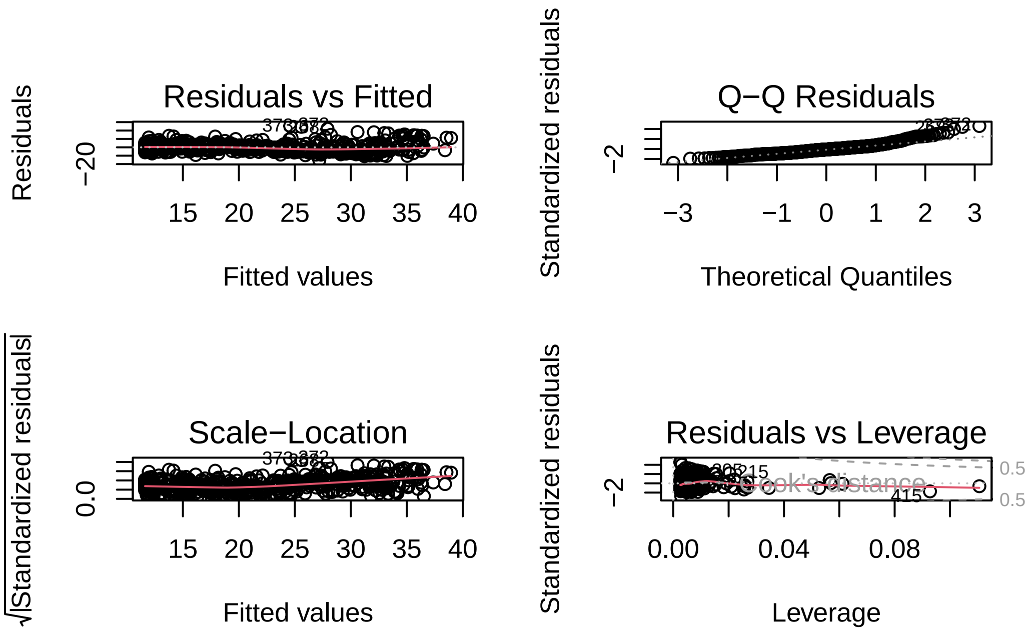
- **ANOVA** : From the result we can see
  - Model 1: medv ~ lstat
  - Model 2: medv ~ lstat + I(lstat^2)
  - NULL Hypothesis : both model same. Alternate Hypothesis : Model 2 better
  - p-value almost 0 : Alternative hypothesis is true
  - We could have guessed it as there was non linear relationship (From Diagnostic Plot)

```r
lm.fit <- lm(medv ~ lstat)
anova(lm.fit, qlm.fit)
```

```
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)    
1    504 19472                                 
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow = c(2, 2))
plot(qlm.fit)
```




```r
plm.fit <- lm(medv ~ poly(lstat, 5), data = Boston)
summary(plm.fit)
```

```
Call:
lm(formula = medv ~ poly(lstat, 5), data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5433  -3.1039  -0.7052   2.0844  27.1153 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       22.5328     0.2318  97.197  < 2e-16 ***
poly(lstat, 5)1 -152.4595     5.2148 -29.236  < 2e-16 ***
poly(lstat, 5)2   64.2272     5.2148  12.316  < 2e-16 ***
poly(lstat, 5)3  -27.0511     5.2148  -5.187 3.10e-07 ***
poly(lstat, 5)4   25.4517     5.2148   4.881 1.42e-06 ***
poly(lstat, 5)5  -19.2524     5.2148  -3.692 0.000247 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.215 on 500 degrees of freedom
Multiple R-squared:  0.6817,	Adjusted R-squared:  0.6785 
F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16
```

- Log transformation of model

```r
summary(lm(medv ~ log(rm), data = Boston))
```

```
Call:
lm(formula = medv ~ log(rm), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.487  -2.875  -0.104   2.837  39.816 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -76.488      5.028  -15.21   <2e-16 ***
log(rm)       54.055      2.739   19.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.915 on 504 degrees of freedom
Multiple R-squared:  0.4358,	Adjusted R-squared:  0.4347 
F-statistic: 389.3 on 1 and 504 DF,  p-value: < 2.2e-16
```