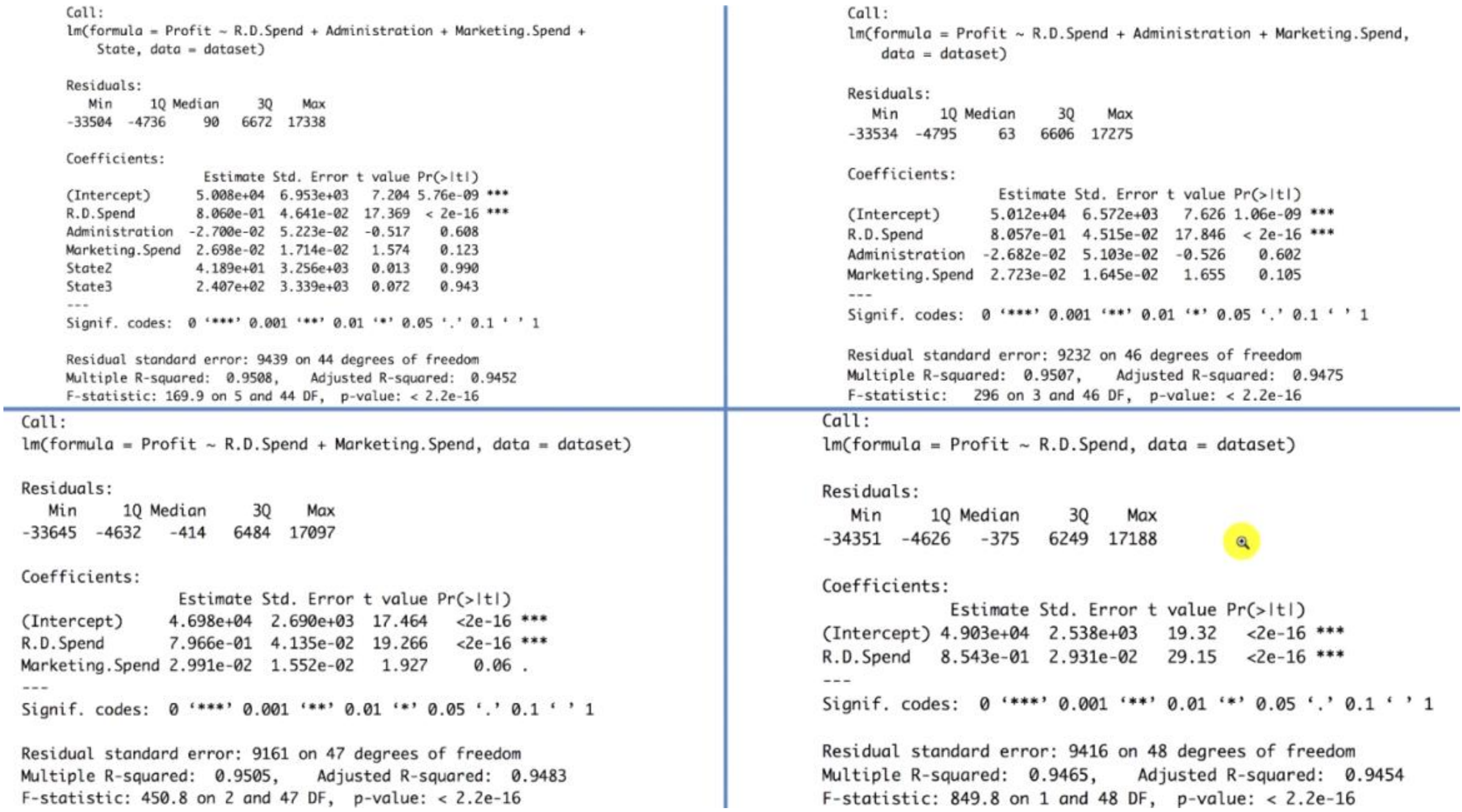
Evaluating Regression Models

Saturday, January 21, 2017 5:19 PM



```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
    State, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.008e+04  6.953e+03   7.204 5.76e-09 ***
R.D.Spend       8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration -2.700e-02  5.223e-02  -0.517    0.608
Marketing.Spend 2.698e-02  1.714e-02   1.574    0.123
State2          4.189e+01  3.256e+03   0.013    0.990
State3          2.407e+02  3.339e+03   0.072    0.943
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33534  -4795     63   6606  17275

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.012e+04  6.572e+03   7.626 1.06e-09 ***
R.D.Spend       8.057e-01  4.515e-02  17.846  < 2e-16 ***
Administration -2.682e-02  5.103e-02  -0.526    0.602
Marketing.Spend 2.723e-02  1.645e-02   1.655    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9232 on 46 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9475
F-statistic:   296 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-34351  -4626   -375   6249  17188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9416 on 48 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9454
F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

The 4 quadrants consists of outputs of 4 different models which were built using different number of variables. This is the result of Backward Elimination algorithm run on the original set of features mentioned in Quadrant 1. We slowly eliminate features which have a P-value greater than 0.05.

During this procedure, in Quadrant 3, we have a model which contains a feature having a P-value of 0.06. This is greater than threshold and hence we had removed it last time round.

Now we are going to employ other stats namely R-squared and adjusted R-squared in order to determine the optimal feature set.

R-squared keeps growing as long as we add variables. So this can't be used as an attribute for determining the optimal feature set owing to its bias.

We use adjusted R-squared instead owing to it having a penalization factor.

Method to select optimal feature set.

- Build a model using all features
- Determine the P-value of all features of the model and eliminate feature having the highest P-value
- Build another model excluding this feature
- Keep doing this procedure until the Adjusted R-squared value decreases
- When it does, use the previous model having the highest adjusted R-squared value obtained so far

In this example, the model in Quadrant 3 is the one selected as the best model to represent the data and it is the one used when making predictions