Hypothesis Testing on Regression Coefficients

Dr. Marko Mitic

When you fit a linear model to a collection of (x,y) points where x and y are both continuous quantitative variables, it can be important to figure out whether any of your regression coefficients are significant. In this document, we will realize hypothesis testing on coefficient obtained from our regression model.

To test whether a particular independent variable is a significant predictor, you want to make sure the regression coefficient (or slope) associated is not zero. This is because if the slope is zero, no matter how much x changes there won't be any change in y. In order words, independent variables (e.g. the predictors) have no influence on dependent variable (e.g. the response), which is obviously wrong.

Before we start our analysis, we need to check initial assumptions:

- Observations are independent
- Residuals (error terms) associated with each observation are independent
- Relationship between the variables is linear
- The values for y vary normally around the mean of y
- Homoscedasticity.

If your assumptions check out, we can start building a linear model. For this example, the data is obtained from Prof. Nicole Radziwill github profile.The data refers to the Shenandoah Regional Airport (SHD) weather data for 2013.

```
# install.packages("RCurl")
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.2.1
```

```
## Loading required package: bitops
```

```
url = "https://raw.githubusercontent.com/NicoleRadziwill/Data-for-R-Examples/master/kshd-2013.txt"
wx.data = getURL(url,ssl.verifypeer=FALSE)
wx = read.table(text=wx.data, header=TRUE)
```

We'll make linear model on data collected in the summer. After checking the dataset, we can do this with:

```
summer.wx = wx[151:243,2:9]
head(summer.wx)
```

```
##     TEMP DEWP   STP VISIB WDSP MXSPD  MAX  MIN
## 151 72.7 57.0 978.8   9.8  4.1  13.0 88.0 55.6
## 152 75.0 59.7 974.8  10.0  4.9  11.1 87.1 61.2
## 153 74.1 60.4 971.2  10.0  6.0  13.0 82.8 66.9
## 154 71.7 59.9 969.9  10.0  3.7  13.0 80.1 65.7
## 155 65.7 47.6 975.2   9.8  3.4  12.0 75.2 55.6
## 156 66.3 50.5 977.5  10.0  2.7   9.9 77.4 55.6
```

We will build two models for linear regression. Model 1 which predict maximum temperature (MAX) from daily average visibility (VISIB), and Model 2 to predict MAX variable using all of the others available in the dataset. This is easily done as follows:

```r
LMmodel1 = lm(MAX ~ VISIB, data = summer.wx)
summary(LMmodel1)
```

```
## 
## Call:
## lm(formula = MAX ~ VISIB, data = summer.wx)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.5316  -2.4915   0.0095   3.2730  10.8525 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  76.9353     3.3970  22.648   <2e-16 ***
## VISIB         0.6355     0.3846   1.652    0.102    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.686 on 91 degrees of freedom
## Multiple R-squared:  0.02913,    Adjusted R-squared:  0.01846 
## F-statistic:  2.73 on 1 and 91 DF,  p-value: 0.1019
```

```r
LMmodel2 = lm(MAX~., data = summer.wx)
summary(LMmodel2)
```

```
## 
## Call:
## lm(formula = MAX ~ ., data = summer.wx)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -5.090 -1.170 -0.188  1.024  6.773 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 22.09360   51.38221   0.430   0.6683    
## TEMP         1.53806    0.11607  13.251  < 2e-16 ***
## DEWP        -0.10264    0.10966  -0.936   0.3519    
## STP         -0.01252    0.05262  -0.238   0.8125    
## VISIB       -0.42212    0.24639  -1.713   0.0903 .  
## WDSP        -0.51962    0.24699  -2.104   0.0384 *  
## MXSPD        0.12900    0.09869   1.307   0.1947    
## MIN         -0.43231    0.09016  -4.795 6.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.175 on 85 degrees of freedom
## Multiple R-squared:  0.8046, Adjusted R-squared:  0.7885 
## F-statistic: 49.99 on 7 and 85 DF,  p-value: < 2.2e-16
```

We will do a test for the slope associated with `VISIB` in the LMmodel1, as well as a test for `WDSP` in the LMmodel2. The null hypothesis is that the slope is actually equals to 0. Alternative hypothesis is the two-tailed version, meaning that the slope is nonzero.

The Test Statistics (for t-test) is calculated using the known formula:

```
## Warning: package 'png' was built under R version 3.2.1
```

$$t = \frac{\hat{\beta}_i}{s_e / \sqrt{SS_{xx}}}$$

where:

$$s_e = \sqrt{\frac{\sum Residuals^2}{n-2}}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2$$

One can also notice that `summary` command has the t-value calculated using simplified formula. We will use this value in our calculations.

P-value can be easily calculated using obtained values:

```
(1-pt(1.652,df=91))*2 #for model 1
```

```
## [1] 0.1019812
```

```
pt(-2.104,df=91)*2 #for model 2
```

```
## [1] 0.03813657
```

So our P-Value for the simple regression case, where we are exploring whether VISIB is a significant predictor, is approximately 0.102. For the multiple regression case, where we are exploring whether WDSP is a significant predictor, the P-Value is approximately 0.038.

Based on these values, we REJECT our null hypothesis for the multiple regression case, but not for the simple regression case. Although WDSP is a significant predictor in the multiple regression model, VISIB is not an effective predictor in the simple regression model.

As for the intercept in our simple linear regression model (`LMmodel1`), we observe from the `summary` command that it is approximatelly 22.68. Substituting that in our formula for p-value, we obtain:

```
(1-pt(22.648,df=91))*2
```

```
## [1] 0
```

The null hypothesis is testing our intercept regarding value zero. In other words, the default test in R sets this null hypothesis value to zero, and chooses a two-tailed test, so that we can tell whether the real intercept is zero or not.

To conclude, the p-value is close to zero (much more smaller than 0.05 threshold value), so we REJECT the null hypothesis that intercept of our model is equal to zero.