SAT study

Tyler Frankenberg

02/20/2022

Import packages

```
library(tidyverse)
```

Import data

```
url <- "https://raw.githubusercontent.com/curdferguson/data621/main/datasets/sat.txt"

sat <- read_tsv(url, skip = 1, col_names = c("state", "expend", "ratio", "salary", "takers", "ve
rbal", "math", "total"), show_col_types=FALSE)
```

Glimpse dataset structure and each column's summary statistics

```
sat %>% head(5)
```

```
## # A tibble: 5 x 8
##   state      expend ratio salary takers verbal  math total
##   <chr>       <dbl> <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 Alabama      4.40  17.2   31.1      8    491   538  1029
## 2 Alaska       8.96  17.6   48.0     47    445   489   934
## 3 Arizona      4.78  19.3   32.2     27    448   496   944
## 4 Arkansas     4.46  17.1   28.9      6    482   523  1005
## 5 California   4.99  24     41.1     45    417   485   902
```

```
sat[,1:4] %>% summary()
```

```
##     state               expend          ratio           salary     
##  Length:50          Min.   :3.656   Min.   :13.80   Min.   :25.99  
##  Class :character   1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98  
##  Mode  :character   Median :5.768   Median :16.60   Median :33.29  
##                     Mean   :5.905   Mean   :16.86   Mean   :34.83  
##                     3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55  
##                     Max.   :9.774   Max.   :24.30   Max.   :50.05
```

```
cat("\n")
```

```
sat[,5:8] %>% summary()
```

```
##      takers          verbal           math           total
##  Min.   : 4.00   Min.   :401.0   Min.   :443.0   Min.   : 844.0
##  1st Qu.: 9.00   1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :28.00   Median :448.0   Median :497.5   Median : 945.5
##  Mean   :35.24   Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:63.00   3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :81.00   Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

ANOVA

We construct a linear model with `expend`, `ratio`, and `salary` as predictors of the response variable `total`.

Is the effect of these predictors on the response statistically significant? We use an F-test for Analysis of Variance to test whether any of the predictors' coefficients is statistically different from zero.

Our F-statistic of 4.0662 is sufficiently greater than that of the null model, and our p-value of 0.01209 indicates this result would be the result of chance in only 0.12% of hypothetical samples.

We reject the null hypothesis that the coefficients of our predictors are statistically equivalent to zero, and take the effect of this model on the response as statistically significant at the 0.95 level.

```
sat_lm1 <- lm(total ~ expend + ratio + salary, data=sat)
sat_nullmod <- lm(total ~ 1, data=sat)

lm1_anova <- anova(sat_nullmod, sat_lm1)
lm1_anova
```

```
## Analysis of Variance Table
##
## Model 1: total ~ 1
## Model 2: total ~ expend + ratio + salary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     49 274308
## 2     46 216812  3     57496 4.0662 0.01209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Examine the effect of a new variable using ANOVA and T-test

We add the predictor `takers` to the model.

Is the additon of this predictor on the response statistically significant? We can test this in two ways; using a t-test for the specific variable and using an F-test to compare the effect of the first and second models as a whole. Then we can show that the results of these two methods are actually the same.

T-test

first, we can output the regresion summary of the new model and observe the value of the t-statistic and p-value for `takers`.

Our regression summary output gives a t-value of -12.559 and a p-value of 2.61e-16 for `takers`. This indicates the coefficient is about 12.5 times the size of its standard error, and that we'd expect this to be the result of chance in well fewer than 0.01% of hypothetical samples.

We conclude by this result that we can reject the null hypothesis at the 0.95% level of statistical significance, and assume the impact of `takers` to be significant.

ANOVA

Second, we can use an F-test for Analysis of Variance between the new model and previous model to test whether the additional impact of the coefficient for `takers` is statistically different from zero.

Our F-statistic of 157.74 is sufficiently greater than that of the model without `takers`, and our p-value of 2.607e-16 indicates this result would be the result of chance in well fewer than 0.01% of hypothetical samples.

We reject the null hypothesis that the difference in the coefficients of our predictors is statistically equivalent to zero, and take the effect of this model on the response as statistically significant at the 0.95 level.

Verify equivalence

Finally, we can verify that our results from these two tests are the same. We expect that our ANOVA F statistic should be approximately the square of our t-value for the added variable `takers`, and that their p-values would be equal.

As we see in our output below, the difference between the t-value squared and the F-statistic, as well as between the p-values, are each so small as to be functionally equivalent to zero.

```
# method 1 - regression summary output t-test
sat_lm2 <- lm(total ~ expend + ratio + salary + takers, data=sat)
lm2_sum <- summary(sat_lm2)
lm2_sum
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
# method 2 - ANOVA
lm2_anova <- anova(sat_lm1, sat_lm2)
lm2_anova
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio + salary + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 2.607e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# verification
(lm2_sum$coefficients["takers", "t value"])^2 - lm2_anova$`F`
```

```
## [1]           NA 2.273737e-13
```

```
(lm2_sum$coefficients["takers", "Pr(>|t|)"]) - lm2_anova$`Pr(>F)`
```

```
## [1]            NA -7.494179e-30
```