Multiple linear regression

Advanced statistical methods and models in experimental design

Bartosz Maćkiewicz

How to improve the quality of education? Guber (1999) study

This example that we will use originated in a paper by Guber (1999). There has been an ongoing debate in the U.S. about what we can do to improve the quality of primary and secondary education.

It is generally assumed that spending more money on education will lead to better-prepared students, but that is just an assumption. Guber addressed that question by collecting data for each of the 50 (U.S.) states. She recorded the amount spent on education, the pupil/teacher ratio, the average teacher's salary, the percentage of students in that state taking the SAT exams, and the combined SAT score.

SAT and ACT: a short crash course

The SAT and the ACT are two separate standardized tests that are used for university admissions. The SAT scores range from 200 to 800, while the ACT scores range from 1 to 36. The SAT has been characterized as mainly a test of ability, while the ACT has been characterized as more of a test of material covered in school. Most importantly, the SAT tends to be used by universities in the Northeast and the West and by the more prestigious schools. Students living elsewhere are probably more likely to take the ACT unless they are applying to schools on either coast, such as Harvard, Princeton, Berkeley, or Stanford.

Distribution of the variables (guber1999.csv)

Before we consider the regression solution itself, we need to look at the distribution of each variable or interest:

```
guber <- read.csv("guber1999.csv")
head(guber, 2)
```

```
##   X id   State Expend PTratio Salary PctSAT Verbal Math SATcombined
## 1 1  1 Alabama  4.405    17.2 31.144      8    491  538         1029
## 2 2  2  Alaska  8.963    17.6 47.951     47    445  489          934
##   ACTcombined LogPctSAT
## 1        20.2     2.079
## 2        21.0     3.850
```

- the amount spent on education (`Expend`),
- the pupil/teacher ratio (`PTratio`),
- average teacher's salary (`Salary`),
- the percentage of students in that state taking the SAT exams (`PctSAT` and `LogPctSAT`)
- the combined SAT score (`SATcombined`)

We can do that using histograms, Q-Q plots, and scatterplots.

Relation between expenditure and educational outcome

The most obvious thing to do with these data is to ask about the relationship between expenditure (`Expend`) and outcome (`SATcombined`, we will ignore ACT for now).

Relation between expenditure and educational outcome

```
cor.test(~ Expend + SATcombined, data = guber)
```

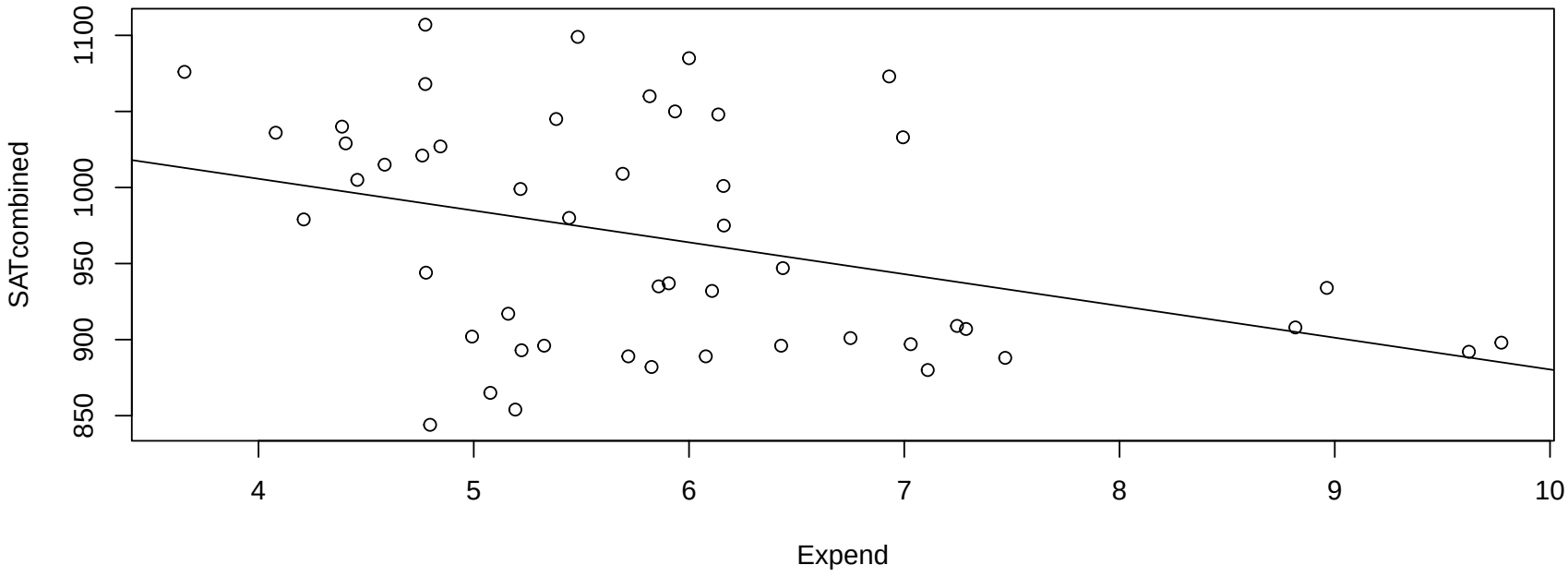
```
##
##  Pearson's product-moment correlation
##
## data:  Expend and SATcombined
## t = -2.8509, df = 48, p-value = 0.006408
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5957789 -0.1142957
## sample estimates:
##       cor
## -0.380537
```

Relation between expenditure and educational outcome

```
fit <- lm(SATcombined ~ Expend, data = guber); summary(fit)
```

```
##
## Call:
## lm(formula = SATcombined ~ Expend, data = guber)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -145.074  -46.821    4.087   40.034  128.489
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1089.294     44.390  24.539  < 2e-16 ***
## Expend       -20.892      7.328  -2.851  0.00641 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.91 on 48 degrees of freedom
## Multiple R-squared:  0.1448, Adjusted R-squared:  0.127
## F-statistic: 8.128 on 1 and 48 DF,  p-value: 0.006408
```

Relation between expenditure and educational outcome



We have found a negative correlation between SAT and Expend. The relationship is somewhat surprising because it would suggest that the more money we spend on educating our children the worse they do. The regression line is clearly decreasing and the correlation is $-.381$. Although that correlation is not terribly large, it is statistically significant ($p = .006$) and cannot just be ignored. Those students who come from wealthier schools tend to do worse. Why should this be?

Adding predictors

The question that now arises is what would happen if we used both variables (`Expend` and `LogPctSAT`) simultaneously as predictors of the SAT score. What this really means, though it may not be immediately obvious, is that we will look at the relationship between Expend and SAT *controlling for LogPctSAT*.

Interpretation:

When we say that we are controlling for `LogPctSAT`, we mean that we are looking at the relationship *while holding `LogPctSAT` constant*.

Imagine that we had many thousands of states instead of only 50 and we could pull out a collection of states that had exactly the same percentage of students taking the SAT (e.g. 60%) Then we could look at only the students from those states and compute the correlation and regression coefficient for predicting `SAT` from `Expend`. Then we could draw another sample of states, perhaps those with 40% of their students taking the exam, and do the same. Notice that we have calculated two correlations and two regression coefficients here, each with `PctSAT` held constant at a specific value (40% or 60%).

Obviously, we don't have thousands of states. However, that does not stop us from mathematically estimating what we would obtain if we could carry out the imaginary exercise that we just explained. And that is exactly what multiple regression is all about.

Multiple linear regression

```
fit <- lm(SATcombined ~ Expend + LogPctSAT, data = guber); summary(fit)
```

```
##
## Call:
## lm(formula = SATcombined ~ Expend + LogPctSAT, data = guber)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.515 -13.616  -2.572  16.541  54.901
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1147.100     16.701  68.684  < 2e-16 ***
## Expend        11.129      3.264   3.409  0.00135 **
## LogPctSAT    -78.203      4.471 -17.490  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.78 on 47 degrees of freedom
## Multiple R-squared:  0.8861, Adjusted R-squared:  0.8813
## F-statistic: 182.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Three ways of thinking about multiple regression

There are at least three ways of thinking about multiple regression. Each of them gives us a somewhat different view of what is going on. If one of them makes more sense to you than the others, you can focus on that approach.

- We can treat a regression coefficient as the coefficient we would get if we had a whole group of states that did not differ on any of the predictors except the one under consideration. In other words, all predictors but one are held constant, and we look at what varying that one predictor does.

Three ways of thinking about multiple regression

We can think of multiple regression as "adjusting" both of our variables for any variables we want to control. You can think of it as "removing" the influence of the controlling variable from the predictor and the outcome variable.

```
# Predicting SAT from LogPctSAT
modelsat <- lm(SATcombined ~ LogPctSAT, data = guber)
predictsat <- predict(modelsat)
# Residuals (= influence of LogPctSAT removed)
residsat <- guber$SATcombined - predictsat

# Predicting Expend from LogPctSAT
modelexpend <- lm(Expend ~ LogPctSAT, data = guber)
predictexpend <- predict(modelexpend)
# Residuals (= influence of LogPctSAT removed)
residexpend <- guber$Expend - predictexpend
```

Three ways of thinking about multiple regression

```r
model <- lm(residsat ~ residexpend)
summary(model)
```

```
##
## Call:
## lm(formula = residsat ~ residexpend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.515 -13.616  -2.572  16.541  54.901
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.526e-13  3.608e+00   0.000  1.00000
## residexpend  1.113e+01  3.230e+00   3.445  0.00119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.51 on 48 degrees of freedom
## Multiple R-squared:  0.1983, Adjusted R-squared:  0.1816
## F-statistic: 11.87 on 1 and 48 DF,  p-value: 0.001194
```

Three ways of thinking about multiple regression

- We can think of a regression coefficient in multiple regression as the same thing we would have in simple regression if we adjusted our two variables for any of the variables we want to control. In this example, it meant adjusting both `SAT` and `Expend` for `LogPctSAT` (by computing the difference between the obtained score for that variable and the score predicted from the variable to be controlled). The coefficient (slope) that we obtain is the same coefficient we find in the multiple regression solution.

Three ways of thinking about multiple regression

- We can think of the multiple correlation as the simple Pearson correlation between the criterion (call it Y) and another variable (call it $\hat{Y}$) that is the best linear combination of the predictor variables.

```
combination <- guber$Expend * 11.129 +
  guber$LogPctSAT * -78.203
cor(guber$SATcombined, combination) ** 2
```

```
## [1] 0.8861068
```

```
summary(fit)$r.squared
```

```
## [1] 0.8861068
```

Importance and standardized regression coefficients

Standardized regression coefficients could be used to compare the "importance" of individual predictors.

```
##
## Call:
## lm(formula = SATcombined ~ Expend + LogPctSAT, data = guber)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.515 -13.616  -2.572  16.541  54.901
##
## Coefficients:
##             Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept) 1147.0997           NA    16.7012  68.684  < 2e-16 ***
## Expend        11.1288       0.2027     3.2644   3.409  0.00135 **
## LogPctSAT    -78.2027      -1.0399     4.4712 -17.490  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.78 on 47 degrees of freedom
## Multiple R-squared:  0.8861, Adjusted R-squared:  0.8813
## F-statistic: 182.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Multicollinearity, VIF and Tolerance

When predictor variables are correlated among themselves we have what is called collinearity or multicollinearity. Collinearity has the effect of increasing the standard error of a regression coefficient, which increases the width of the confidence interval and decreases the t value for that coefficient. This is what is measured by the **VIF** (*variance inflation factor*). Moreover, when two predictors are highly correlated one has little to add over and above the other and only serves to increase the instability of the regression equation.

Tolerance is the reciprocal of the VIF and can be computed as $1 - R_j^2$, where R_j^2 is the multiple correlation between variable j and all other predictor variables.

So we want a low value of VIF and a high value of Tolerance.

$VIF > 4$ may indicate instability of the regression equation and $VIF > 10$ (strong multicollinearity) means that we need to remove at least one predictor from the model.

```
car::vif(fit)
```

```
##    Expend LogPctSAT
##  1.458884  1.458884
```

Power and Sample Size

A rule of thumb that has been kicking around for years is that we should have at least 10 observations for every predictor.

Harris (1985) points out, however, that he knows of no empirical evidence supporting this rule. It certainly fails in the extreme, because no one would be satisfied with 10 observations and 1 predictor. Harris advocates an alternative rule dealing not with the ratio of p to N, but with their difference. His rule is that N should exceed p by at least 50. Others have suggested the slightly more liberal $N \geq p + 40$.

Whereas these two rules relate directly to the reliability of a correlation coefficient, Cohen, Cohen, West, and Aiken (2003) approach the problem from the direction of statistical power analysis. They show that in the three-predictor case, to have power .80 for a population $R^2 = .20$ would require $N = 48$. With 5 predictors, a population $R^2 = .20$ would require 58 subjects for the same degree of power.

Power and sample size: example

How many observations do we need to have 80% probability of detecting true R^2 that is equal to 0.2 with three predictors?

We can answer this question using R and `pwr` library that contains `pwr.f2.test` function. This function relies on the formula for a statistical test for H_0: R^2:

$$F = \frac{(N - p - 1)R^2}{p(1 - R^2)}$$

and takes the following arguments:

- `u` - the number of degrees of freedom in the numerator (the number of predictors)
- `v` - the number of degrees of freedom in the denominator, so $N - p - 1$ (look at the formula for F)
- `f2` - effect size computed using the formula:

$$f^2 = \frac{R^2}{1 - R^2}$$

Comparing models

Often we have what are called nested models or hierarchical models in which the variables in one model represent a subset of the variables in a second model.

In the case of nested models, it is relatively easy to test whether one is significantly better than another. We can just compare their R^2 values or the sums of squares for regression.

```
fit1 <- lm(SATcombined ~ Expend, data = guber)
fit2 <- lm(SATcombined ~ Expend + LogPctSAT, data = guber)
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: SATcombined ~ Expend
## Model 2: SATcombined ~ Expend + LogPctSAT
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     48 234586
## 2     47  31242  1    203344 305.91 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```