# Analysis of Total Expenditures for Materials

*Michele Claibourn and Clay Ford*

*November 2018*

## Request from Donna Tolson

> I'd like you to run regressions to see where we land among the various peer groups, and if possible, to forecast what the model would predict our expenditures should be, compared to what they are.

Donna provided a 2017 spreadsheet with requested outcomes/dependent variables and inputs/independent variables highlighted. The initial analysis begins with the **collections expenditures** as the highest priority outcome; additional outcome variables can be assessed given more time.

In what follows, we estimate the amount spent on materials as a function of the number of faculty, the number of full time students, and the number of full time graduate students. Donna suggested number of faculty, number of full-time students, and number of PhDs awarded as predictors; model diagnostics revealed that the number of graduate students was a consistently stronger predictor of expenditures than number of PhDs awarded. As the predictive value of the model depends on the ability of the model to explain variation in expenditures, we chose to include the more strongly related of these two highly correlated features.

The same model is estimated on multiple samples of institutions: (1) The full set of ARL institutions (123 observations), (2) only US institutions (99), (3) SCHEV-identified peer institutions (27), (4) library-identified ARL peer institutions (30), (5) and a small sample of "scholarly infrastructure" peers (9). In each case the model is estimated on the sample excluding UVA. The model is then used to estimate expenditures for UVA, given the number of faculty, students, and graduate students; this predicted value is provided along with the actual UVA collection expenditures.

In addition, a histogram of the distribution of the residuals (predicted values minus actual values for all observations) are graphed and the inidividual residuals are plotted; UVA is identified by an orange dot.

## All ARL

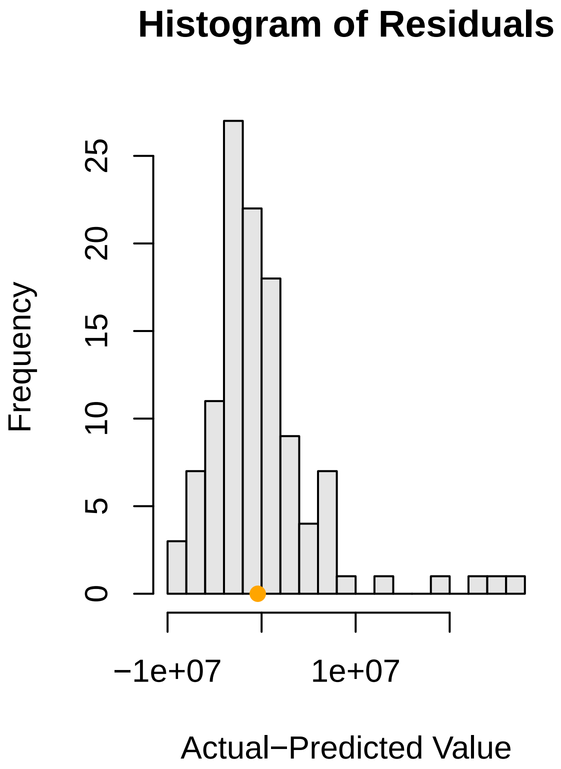
Based on the full set of ARL institutions, UVA's collection budget is about $400,000 less than predicted.

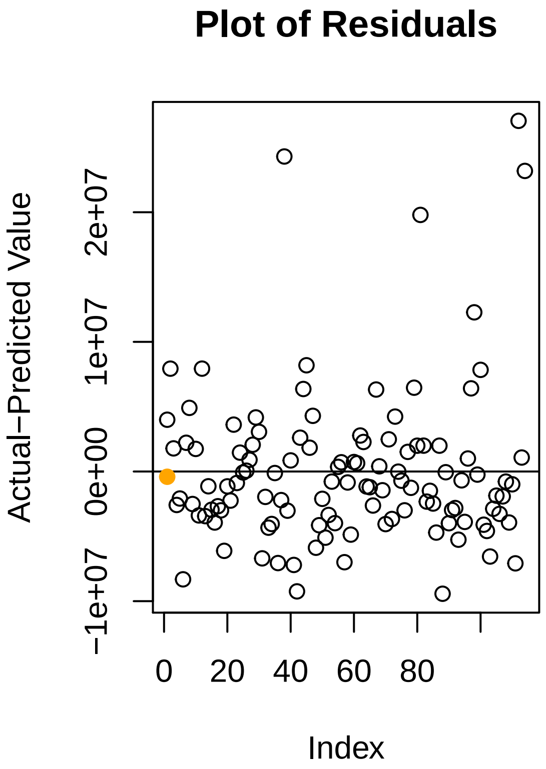
```
mod1 <- lm(explm ~ gradstu + fac + totstu,
           data = arl,
           subset = subset == "all_arl" & inam != "VIRGINIA")
pred1 <- predict(mod1, newdata = uva)
modelsum(summary(mod1)$adj.r.squared)
```

```
## Adjusted R^2: 0.353
```

```
compare(pred1)
```

```
## Actual:    $14,033,517
## Predicted: $14,438,536
## Diff:      $-405,019
```





For the full ARL sample, the model accounts for 36% of the variation in expenditures. Harvard, Princeton, Yale, and the Library of Congress are the notable outliers in this sample, evident in the highly positive residuals in the second graph.

## US Only

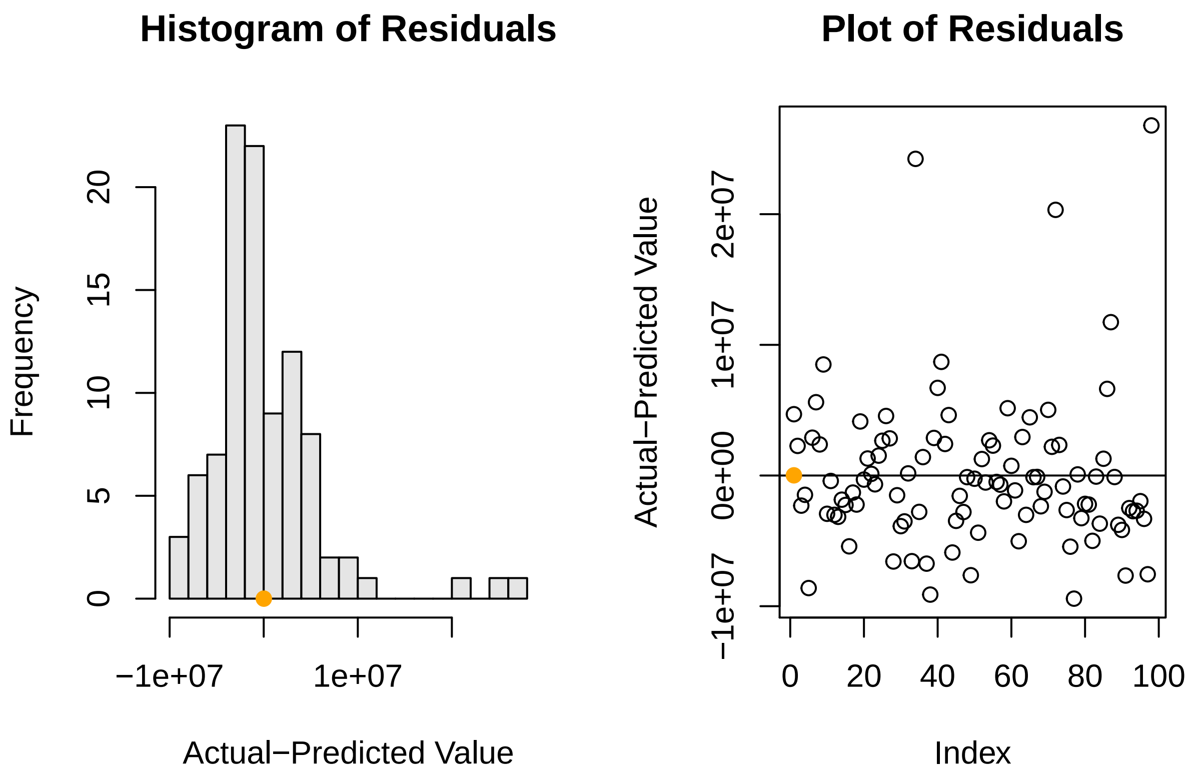
Restricting the estimation sample to only US insitutions, the collections expenditures for UVA are about $14,000 more than predicted.

```
mod2 <- lm(explm ~ gradstu + fac + totstu,
           data = arl,
           subset = subset == "us_only" & inam != "VIRGINIA")
pred2 <- predict(mod2, newdata = uva)
modelsum(summary(mod2)$adj.r.squared)
```

```
## Adjusted R^2: 0.401
```

```
compare(pred2)
```

```
## Actual:    $14,033,517
## Predicted: $14,019,059
## Diff:      $14,457.72
```



For the US-only sample, the model accounts for 40% of the variation in expenditures. Harvard, Princeton, and Yale are, again, the notable outliers evident in the second figure.

### SCHEV UVA

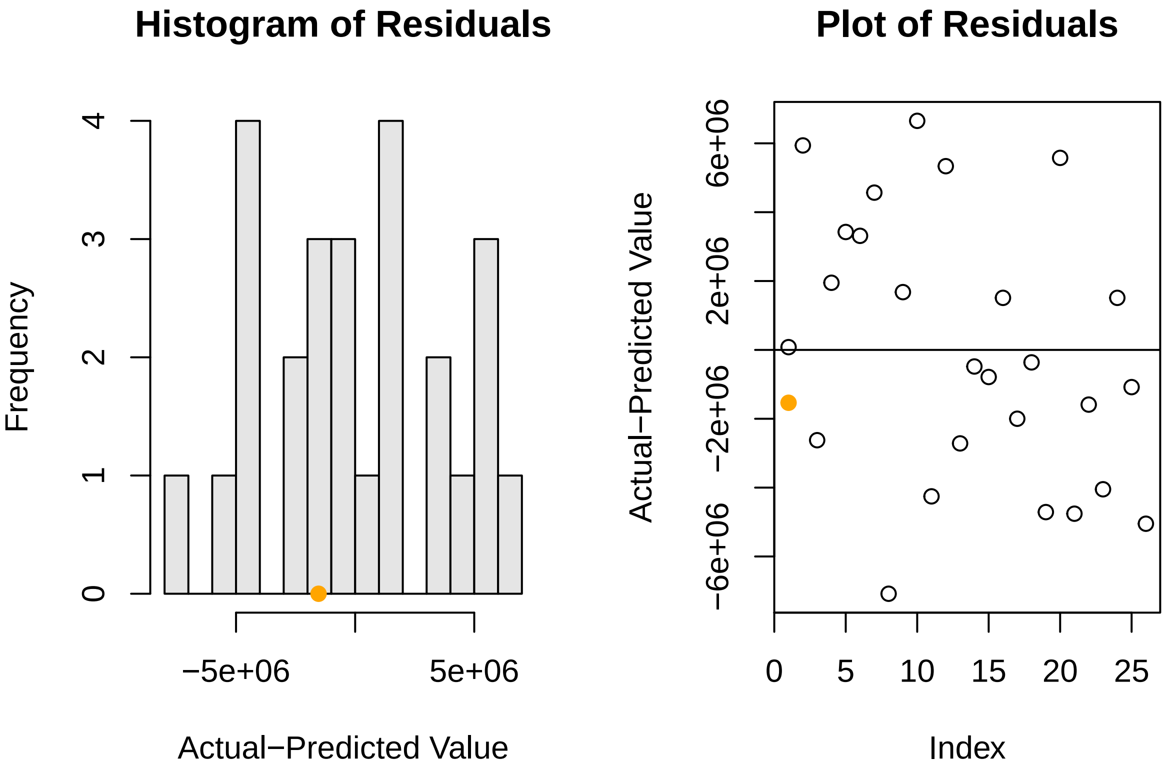
Restricting the sample still further to institutions identified by SCHEV as peer institutions for UVA, the UVA collections budget is about $1.5M less than predicted given UVA's size.

```r
mod3 <- lm(explm ~ gradstu + fac + totstu,
           data = arl,
           subset = subset == "schev_uva" & inam != "VIRGINIA")
pred3 <- predict(mod3, newdata = uva)
modelsum(summary(mod3)$adj.r.squared)
```

```
## Adjusted R^2: 0.344
```

```r
compare(pred3)
```

```
## Actual:    $14,033,517
## Predicted: $15,568,692
## Diff:      $-1,535,175
```



For the SCHEV peer sample, the model accounts for 34% of the variation in expenditures. No institutions register as outliers in this smaller sample. Importantly, only the number of graduate students is a significant predictor in the model estimated on the 26 SCHEV peer institutiions (all three predictors were significantly related to expenditures in the larger samples above).

### ARL peers

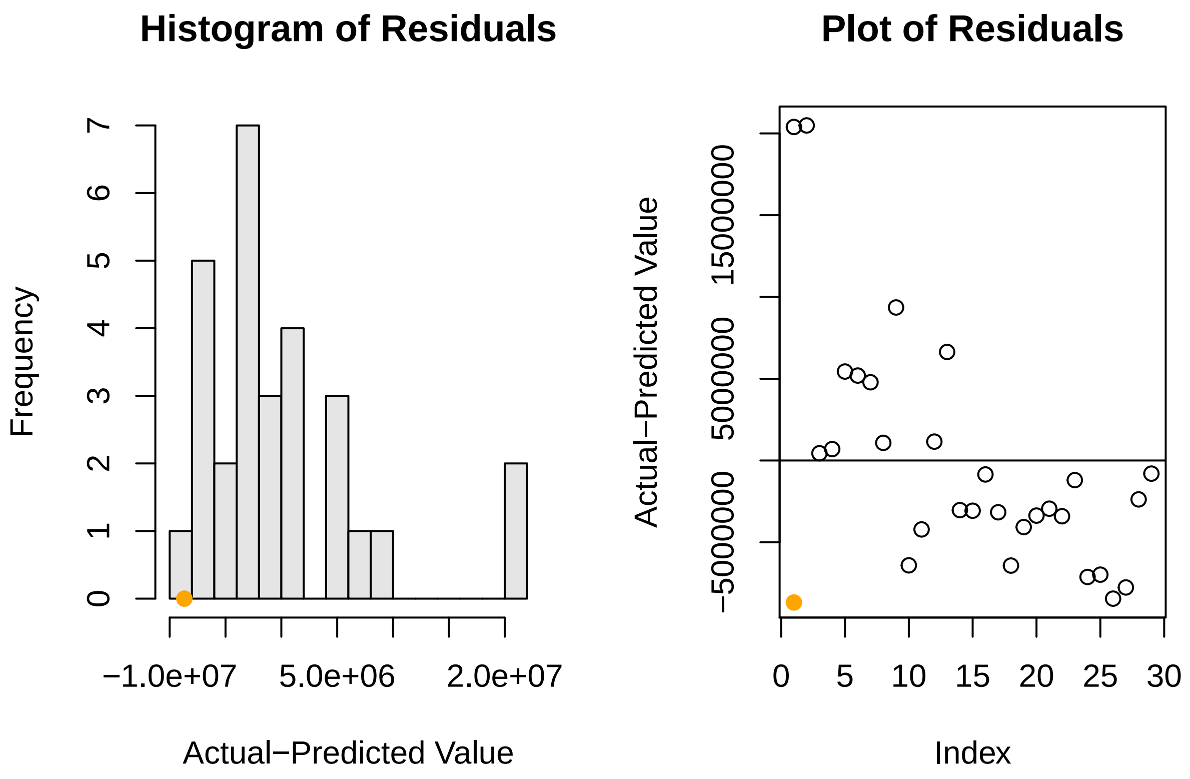
Using a similarly restricted sample, the 29 ARL peer institusions identified by the Library, UVA's collection budget is nearly $8.7M less than predicted.

```
mod4 <- lm(explm ~ gradstu + fac + totstu,
           data = arl,
           subset = subset == "arl_peers" & inam != "VIRGINIA")
modelsum(summary(mod4)$adj.r.squared)
```

```
## Adjusted R^2: 0.075
```

```
pred4 <- predict(mod4, newdata = uva)
compare(pred4)
```

```
## Actual:    $14,033,517
## Predicted: $22,716,176
## Diff:      $-8,682,659
```



For the SCHEV peer sample, the model accounts for only 8% of the variation in expenditures; that is, these predictors are not particularly related to materials expenditures. Here, again, Harvard and Yale stand out as notable outliers. In addition, the residual plot indicates non-constant variances. This model requires additional assessment and its predictions should be used with caution.

### Scholarly Infrastructure peers

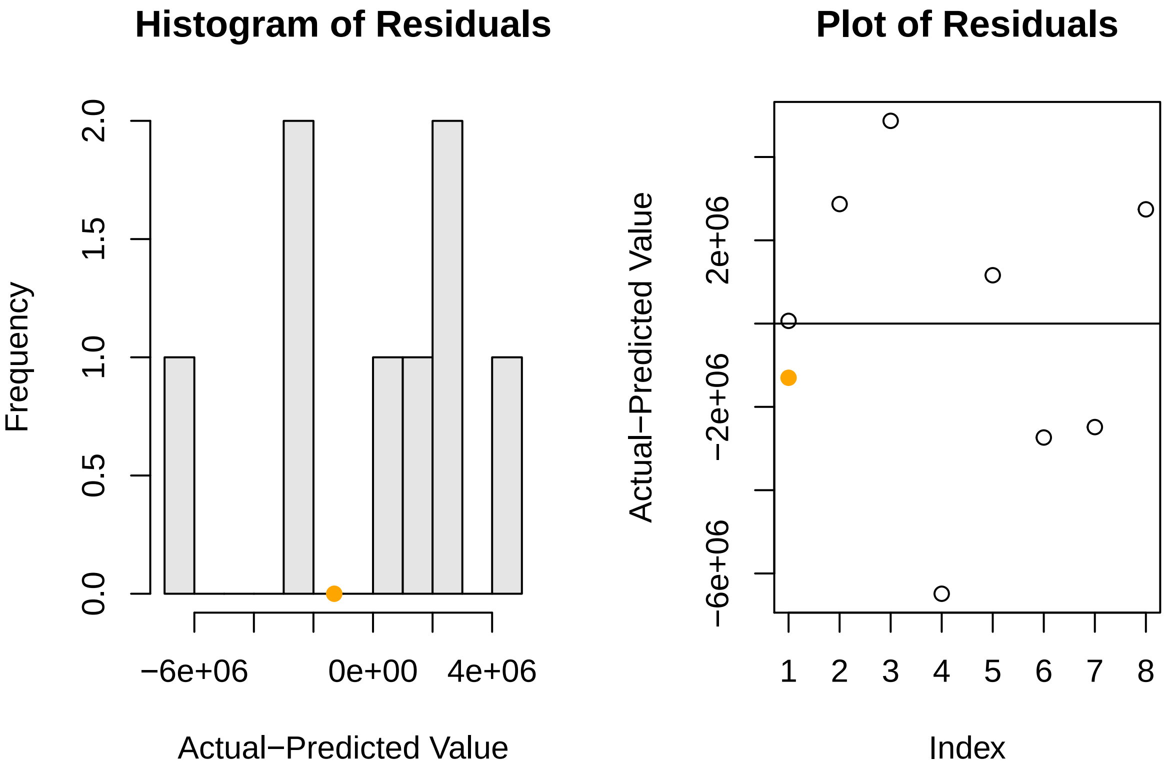
For the model for the scholarly infrastructure sample, UVA's collection expenditures are about $1.3M less than predicted.

```
mod5 <- lm(explm ~ gradstu + fac + totstu,
           data = arl,
           subset = subset == "scholarly_infrastructure_peers" & inam != "VIRGINIA")
modelsum(summary(mod5)$adj.r.squared)
```

```
## Adjusted R^2: 0.535
```

```
pred5 <- predict(mod5, newdata = uva)
compare(pred5)
```

```
## Actual:    $14,033,517
## Predicted: $15,334,861
## Diff:      $-1,301,344
```



This subsample is very small, with only 8 peers on which to estimate a model. Only number of graduate students bears any statistically discernible relationship to expenditures among this set. The model accounts for 54% of the variation in expenditures.

## Conclusions

Based on this model, the UVA Library collection budget is less than expected relative to ARL institutions for all subsamples except the "US-only" subsample. That said, this model incorporates a very limited set of predictors and may not represent the most appropriate model specification available. We considered multiple model specifications among the seven input variables initially identified by Donna; in a 2015 version of a similar analysis, we included additional variables as well (e.g., volumes, inclusion of a medical or law library). Such additional inputs were not considered here given time constraints.

With additional time, we could provide a fuller assessment of the regression model and subsamples, repeat the analysis for additional outcomes (total expenditures, total salary expenditures, staff salary expenditures), and incorporate data from multiple years to increase robustness or potentially examine change over time.