Multivariate OLS Estimation of GPA Data

Analysis of Influential Cases

Previously, we fit a bivariate linear model, predicting GPA as a function of ACT score

$$colGPA = \beta_0 + \beta_1 ACT + u$$

Here's a quick recap of the steps we took.

Basic setup and loading of data.

```
# We use the stargazer package to display nice regression tables.
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
load("gpa1.RData")
head(data)
```

We examined the colGPA and ACT variables individually, which we omit here.

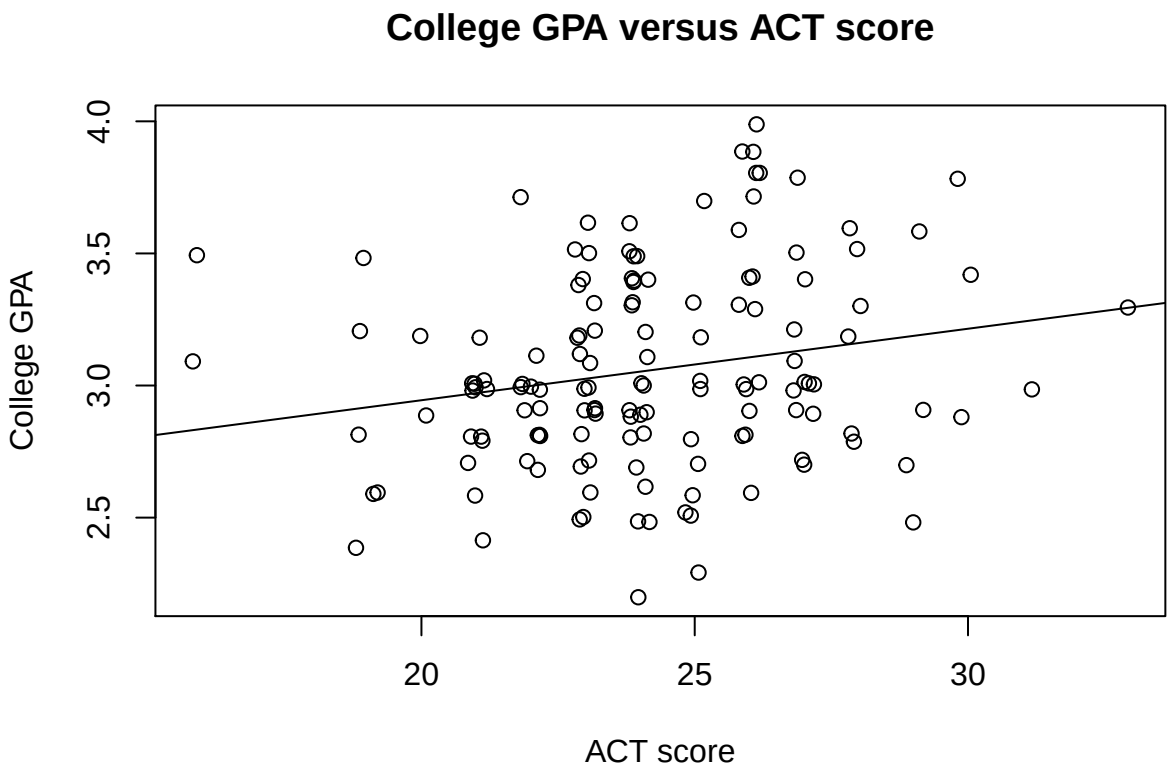
We then created a scatterplot of the two variables and fit a linear model.

```
# create the scatterplot
plot(jitter(data$ACT), jitter(data$colGPA), xlab = "ACT score", ylab = "College GPA", main = "College GP

# fit the linear model
(model1 = lm(colGPA ~ ACT, data = data))
```

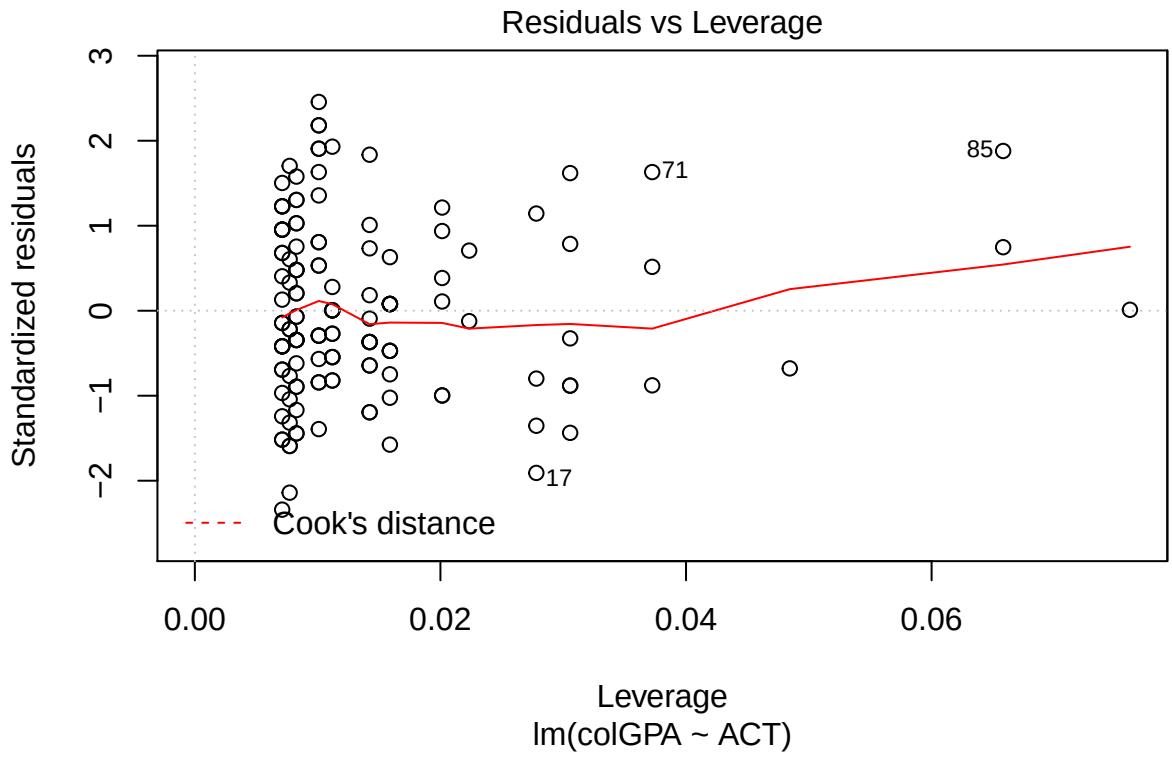
```
##
## Call:
## lm(formula = colGPA ~ ACT, data = data)
##
## Coefficients:
## (Intercept)          ACT
##     2.40298      0.02706
```

```
# Add regression line to scatterplot
abline(model1)
```



Next, we will want to examine our data to check for any unusually influential cases. We can use a residuals vs. leverage plot for this purpose.

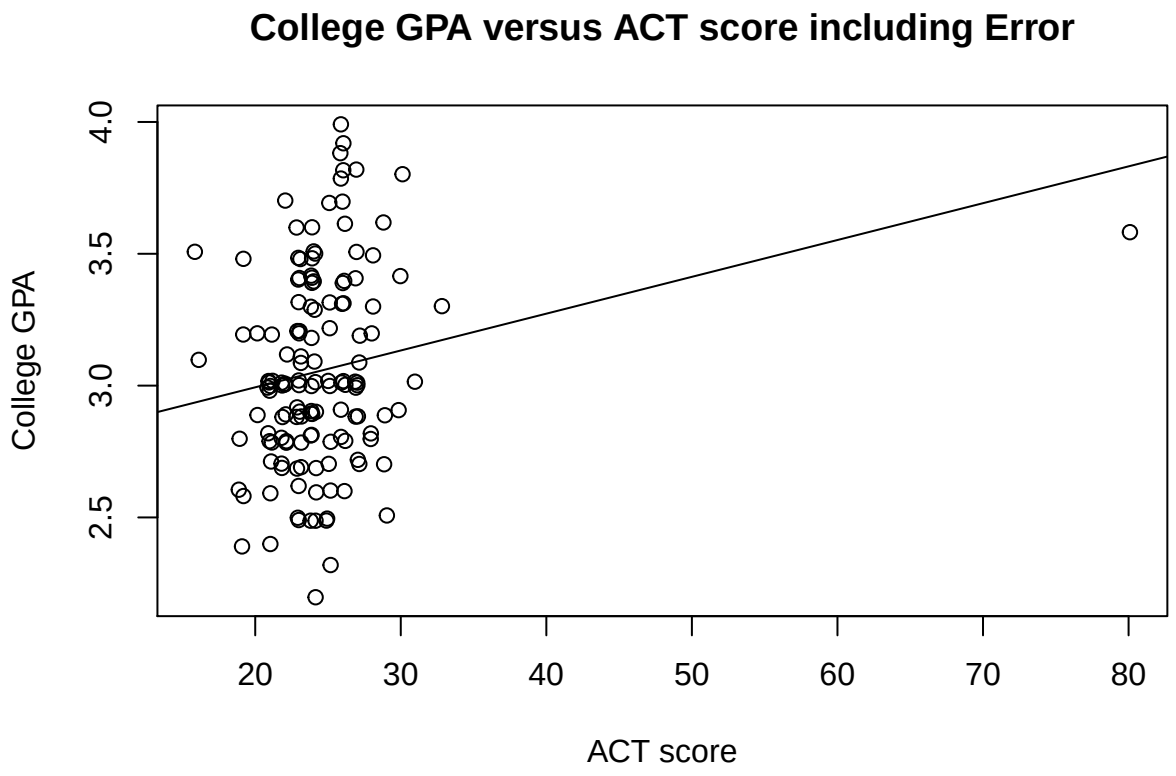
```
plot(model1, which = 5)
```



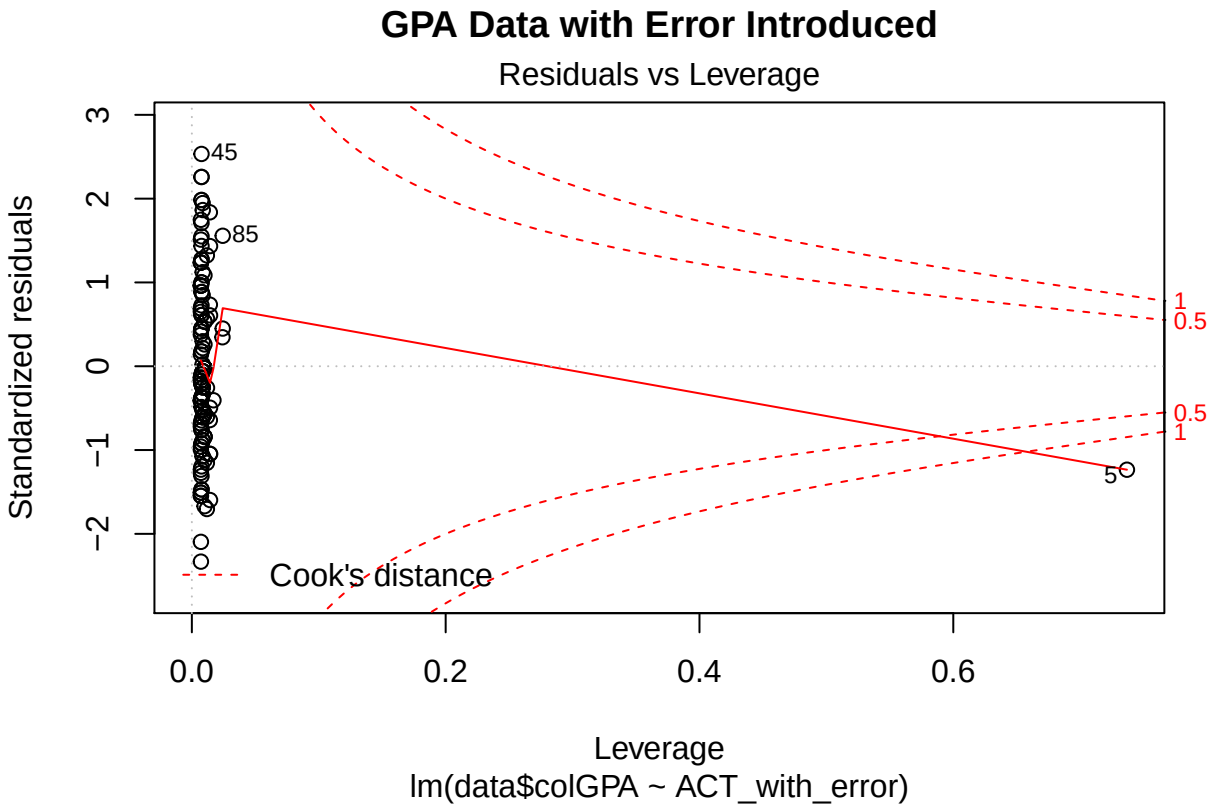
The following code shows what would happen if we introduced an error into the data set, resulting in a point with high influence.

```
ACT_with_error = data$ACT
ACT_with_error[5] = 80
model1_with_error = lm(data$colGPA ~ ACT_with_error)

# visualize the data with the error and the new ols line
plot(jitter(ACT_with_error), jitter(data$colGPA), xlab = "ACT score", ylab = "College GPA", main = "Col
# Add regression line to scatterplot
abline(model1_with_error)
```




```
plot(model1_with_error, which=5, main = "GPA Data with Error Introduced")
```



Notice that the point now stands out as having Cook's distance greater than 1.

Warning: when we find an influential case, we never automatically remove it from the data set.

Multivariate Linear Model Estimation

Here, we recreate the regression from the lecture and from Woodridge chapter 3. We predict colGPA from both ACT and high school GPA (hsGPA).

Our second model looks like this.

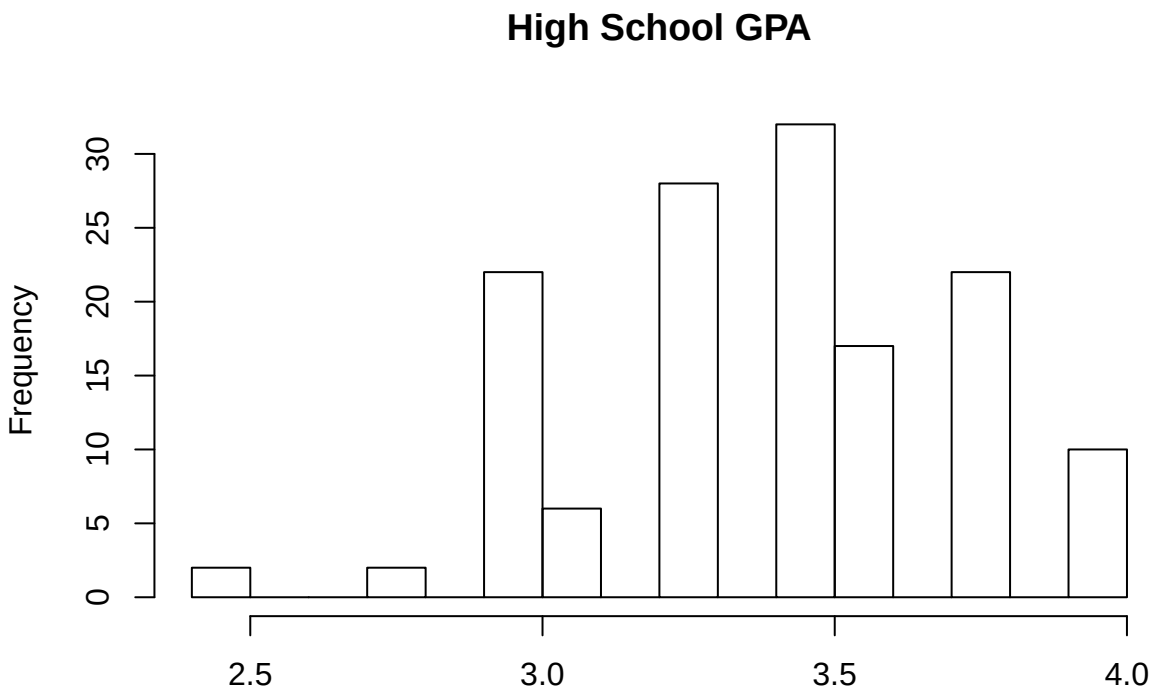
$$colGPA = \beta_0 + \beta_1 ACT + \beta_2 hsGPA + u$$

We first examine the high school GPA variable.

```
summary(data$hsGPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.400   3.200   3.400   3.402   3.600   4.000
```

```
hist(data$hsGPA, breaks = 20, main = "High School GPA", xlab = NULL)
```

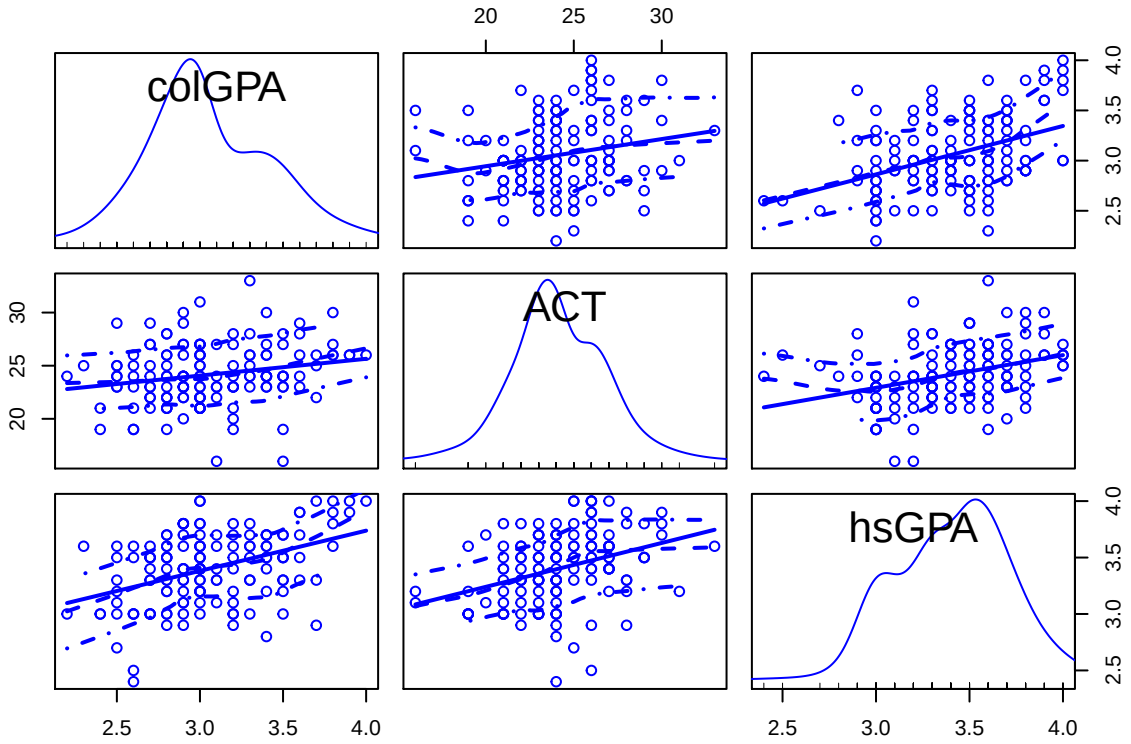



```
library(car)
```

```
## Loading required package: carData
```

```
scatterplotMatrix(data[,c("colGPA", "ACT", "hsGPA")], diagonal = "histogram")
```

```
## Warning in applyDefaults(diagonal, defaults = list(method =
## "adaptiveDensity"), : unnamed diag arguments, will be ignored
```



Next, we fit the linear model.

```
(model2 = lm(colGPA ~ ACT + hsGPA, data = data))
```

```
##
## Call:
## lm(formula = colGPA ~ ACT + hsGPA, data = data)
##
## Coefficients:
## (Intercept)          ACT        hsGPA
##    1.286328     0.009426     0.453456
```

```
model2$coefficients
```

```
## (Intercept)         ACT       hsGPA
## 1.286327767 0.009426012 0.453455885
```

Let's compare the R-squares for our two models.

```
summary(model1)$r.square
```

```
## [1] 0.04274732
```

```
summary(model2)$r.square
```

```
## [1] 0.1764216
```

Remember that R-squared can only go up when adding new variables.

For an assessment of model fit that penalizes extra variables, we can use the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC)

```r
AIC(model1)
```

```
## [1] 120.3534
```

```r
AIC(model2)
```

```
## [1] 101.1457
```

Presenting Regression Output

For a lot of reasons, we usually want to display the results of more than one linear model.

1. There can be good arguments for different specifications
2. We want to show that an effect is robust across models
3. We want to show that we're not cherry-picking a model that supports our argument

Because of these reasons, we often want to present the results of multiple models in a *regression table*. The stargazer package is a great way to create these tables.

```r
library(stargazer)
stargazer(model1, model2, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models Predicting College GPA",
          keep.stat = c("rsq", "n"),
          omit.table.layout = "n") # Omit more output related to errors
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Nov 23, 2018 - 3:36:38 PM

Table 1: Linear Models Predicting College GPA

	Dependent variable:	
	colGPA	
	(1)	(2)
ACT	0.027	0.009
hsGPA		0.453
Constant	2.403	1.286
Observations	141	141
R^2	0.043	0.176

The Stargazer Cheatsheet, by Jake Russ, is a great place to get started with stargazer. http://jakeruss.com/cheatsheets/stargazer.html