S631 HW6

Erik Parker

September 30, 2017

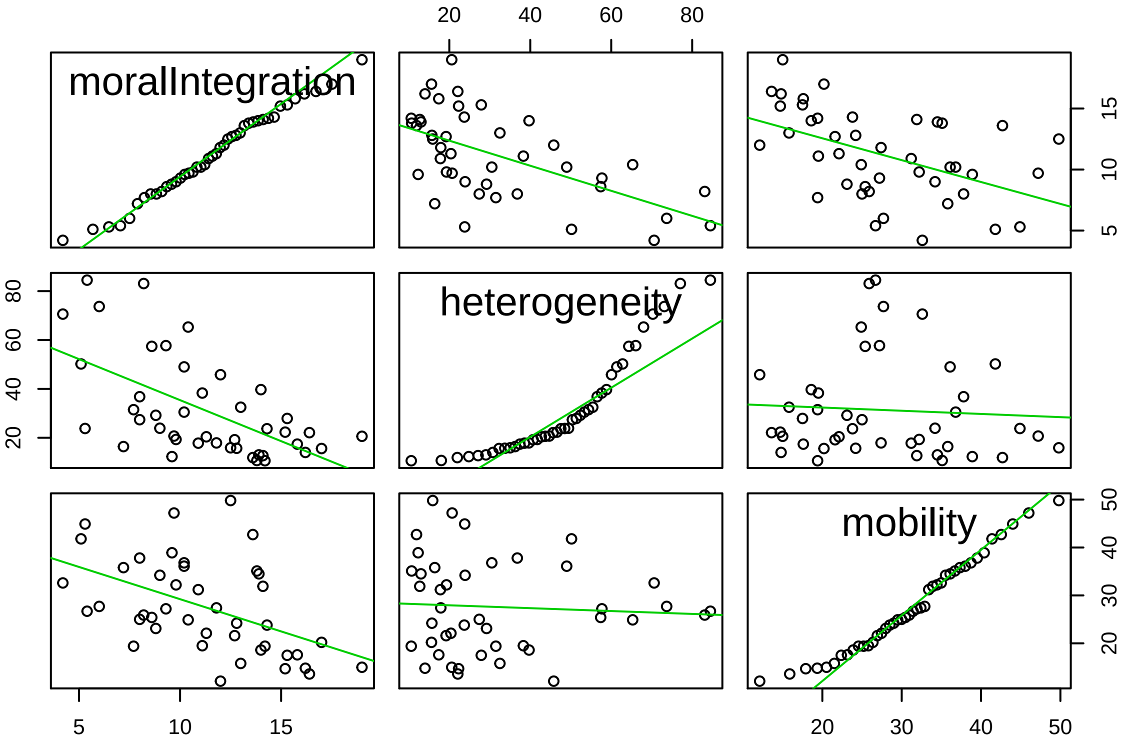
Our dataset contains information about moral integration of American Cities, and we are to study the influence of *mobility* and *heterogeneity* on *moralIntegration*.

a. Draw scatterplots, including the least-squared lines, showing the relationship of the response to each predictor. Are the least-squares lines reasonable summaries of the relationship between the response and each predictor?

```r
library(alr4)

cities <- read.table("Angell.txt", header = TRUE)

scatterplotMatrix(~moralIntegration + heterogeneity +
                    mobility, data = cities, smoother = FALSE, diagonal = "qqplot")
```



Based on the scatterplots generated above, it seems that the least-squared method provides a reasonable summary of the relationships between the response (*moralIntegration*) and each of the two predictors (*heterogeneity* and *mobility*). Both of these predictors seem to show reasonably linear, negative relationships with the response variable, and for the most part there appears to be no major violations of the assumption of homoscedasticity for these data. So overall, yes, the least-squares lines do appear to be reasonable summaries of the relationships between these responses and the predictor.

b. Compute the simle linear regression of *moralIntegration* on *heterogeneity* and interpret the coefficient estimates and the coefficitent of determination.

```
m1 <- lm(moralIntegration ~ heterogeneity, data = cities)

summary(m1)
```

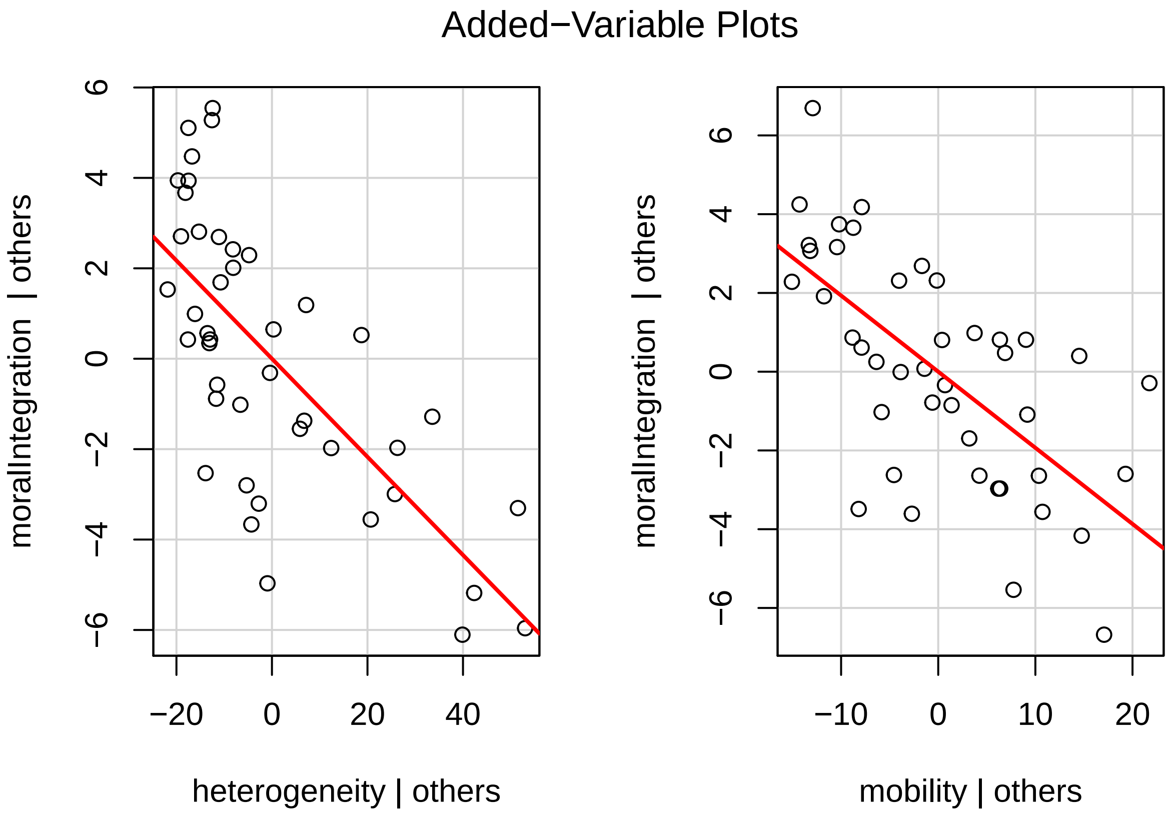
```
##
## Call:
## lm(formula = moralIntegration ~ heterogeneity, data = cities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6780 -2.6099  0.2493  2.2971  6.6931
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.42355    0.82507  17.482  < 2e-16 ***
## heterogeneity -0.10275    0.02212  -4.645 3.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.926 on 41 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3288
## F-statistic: 21.58 on 1 and 41 DF,  p-value: 3.486e-05
```

> From the summary of this model, we can see that our $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates are 14.424 and -0.103 respectively. As the range of values for *heterogeneity* (and all our regressors) does not contain 0 though, this value for $\hat{\beta}_0$ has not practical meaning here and only represents the intercept value for our line. This means that most important for interpretation here is the slope estimate, $\hat{\beta}_1$, which is telling us that for every increase of one unit in *heterogeneity* we see a decrease of 0.103 units of our response, *moralIntegration*. The next coefficient, the standard error, represents how precisely our model estimates the unknown value of our $\hat{\beta}_1$ coefficient by representing the square root of the estimated variance of $\hat{\beta}_1$. Here, the standard error for the slope is quite low at 0.0221, meaning that given these data the model was able to accurately predict a coefficient for *heterogeneity*. Next, the t- and p-values listed in the summary give us an idea of the statistical significance of our coefficients when testing the null hypothesis that they are not different from zero. The t-values are used to find a p-value which represents the probability of finding a coefficient at least as extreme as the one seen given that the null hypothesis is true and the *real* coefficient is actually equal to zero. Here the p-value for *heterogeneity* is quite low, meaning that there is a very small probability of seeing this value for $\hat{\beta}_1$ given that the true value of β_1 is zero, so we are safe to reject this null hypothesis.
>
> Finally, R^2 (the coefficient of determination) represents a ratio of the variation explained in the response variable by the regressor ($SSreg$) over the total amount of variation seen in the response (SYY). The value seen here for R^2 is not too high, at 0.345, meaning that there is still a fair amount of variation left unexplained by this model.

c. Compute the multiple regression using both predictors and interpret the coefficient estimates, corresponding added-variable plots, and the coefficient of determination.

```
m2 <- lm(moralIntegration ~ heterogeneity + mobility, data = cities)

avPlots(m2)
```



```r
summary(m2)
```

```
## 
## Call:
## lm(formula = moralIntegration ~ heterogeneity + mobility, data = cities)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -5.071 -1.194 -0.206  1.738  4.195 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   19.94076    1.19265  16.720  < 2e-16 ***
## heterogeneity -0.10856    0.01699  -6.389 1.34e-07 ***
## mobility      -0.19331    0.03543  -5.456 2.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.243 on 40 degrees of freedom
## Multiple R-squared:  0.6244, Adjusted R-squared:  0.6056 
## F-statistic: 33.25 on 2 and 40 DF,  p-value: 3.126e-09
```

Building off of the in depth descriptions in the previous problem: we see here that the $\hat{\beta}_1$ and $\hat{\beta}_2$ coefficient estimates for the slopes of *heterogeneity* and *mobility* respectively are both negative and of similar magnitude. Interpreting them, we see that for every one unit change in *heterogeneity* there is a decrease of the response variable by 0.109 units, when *mobility* is held constant, and for every one unit change in *mobility* we see a decrease in *moralIntegration* by 0.193 units when *heterogeneity* is the predictor which is held constant. The standard errors for both of these regressors are once again quite low, as are their p-values. The interpretation of these low standard errors is the same as the previous problem, but the interpretation of these low p-values has changed slightly as they now represent the same probabilities as before, but under the condition that are only valid when we think about them as a test of the additional influence of a regressor when it is added to a model that already contains the other regressor. So, the p-value of 1.34e-07 for *heterogeneity* means that there is a very low probability of$\hat{}_1$ equaling zero in this particular context when *heterogeneity* has been added to a model that already contains *mobility* as a predictor. This is an important distinction to make, as it can be seen that these p-values only make sense when considered in the context of this one particular model.

The value for R^2 here is quite a bit larger than the one from the previous model with *heterogeneity* alone, about double, meaning that the addition of *mobility* to this model (either alone, or with its interactions with *heterogeneity*) explains quite a bit more of the total variance seen in the response.

To get a different look at the variation explained by both *heterogeneity* and *mobility* alone, we can look at the added variable plots above. On the y-axis of both plots, we can see the variance in our response, *moralIntegrity*, not explained by the other regressor that is not seen on each plot's x-axis (So, on the left plot it is the variation in *moralIntegration* not explained by *mobility*). Then, in he left plot, the x-axis represents the variance not explained by the regression of *heterogeneity* on the other regressor, *mobility*. While the x-axis of the right plot represents the variance not explained by the regression of *mobility* on *heterogeneity*. So these plots essentially show us the relationship between the response and the two regressors when the other regressor is adjusted for, which is just a graphical representation of the numerical $\hat{\beta}$ values we obtained previously. The slopes of these two plots are both negative and large enough in relation to the units of the response variable, *moralIntegration*, suggesting that there is an influence of each of our two regressors on the response when they are considered alone. This conclusion regarding their influence is only explicitly testable through a hypothesis test though, which was addressed above and in the question below.

d. Explain the information contained in the output about the coefficient for *heterogeneity* in terms of the hypothesis test and conclusion. Also, obtain and interpret a 97% confidence interval for this coefficient.

```r
confint(m1, level = .97)
```

```
##                    1.5 %      98.5 %
## (Intercept)   12.5685788 16.27852516
## heterogeneity -0.1524859 -0.05301849
```

```r
confint(m2, level = .97)
```

```
##                    1.5 %      98.5 %
## (Intercept)   17.2569659 22.62454925
## heterogeneity -0.1467933 -0.07032628
## mobility      -0.2730376 -0.11358988
```

As stated above, the p-value given for *heterogeneity* in the first model is low enough for us to reject the null hypothesis that the slope of our model ($\hat{\beta}_1$) is equal to zero. *Heterogeneity*'s p-value given by the second model is also low enough to reject the null hypothesis that the estimated coefficient is equal to zero, but this time with the qualification that we are interpreting the associated probability when considering this $\hat{\beta}$ value as being obtained when *heterogeneity* is added to a model already containing the other regressor, *mobility* and that regressor is held constant.

When we obtain a 97% confidence interval for *heterogeneity* under the first regression model, we can say we are 97% confident that the interval ranging from -0.152 to -0.0530 will contain the true value of the coefficient (β_1). Under the second, multiple regression model, this changes slightly in that we can now say that we are 97% confident that the interval from -0.147 to -0.0703 will contain the true value of β_1, given that β_2 is present in the model. So, these two models show some slight differences in their 97% confidence interval ranges, but they tell the same general story that the true value for the *heterogeneity* slope coefficient is very likely negative and non-zero, allowing us to further reject the null hypothesis that $\beta_1 = 0$.

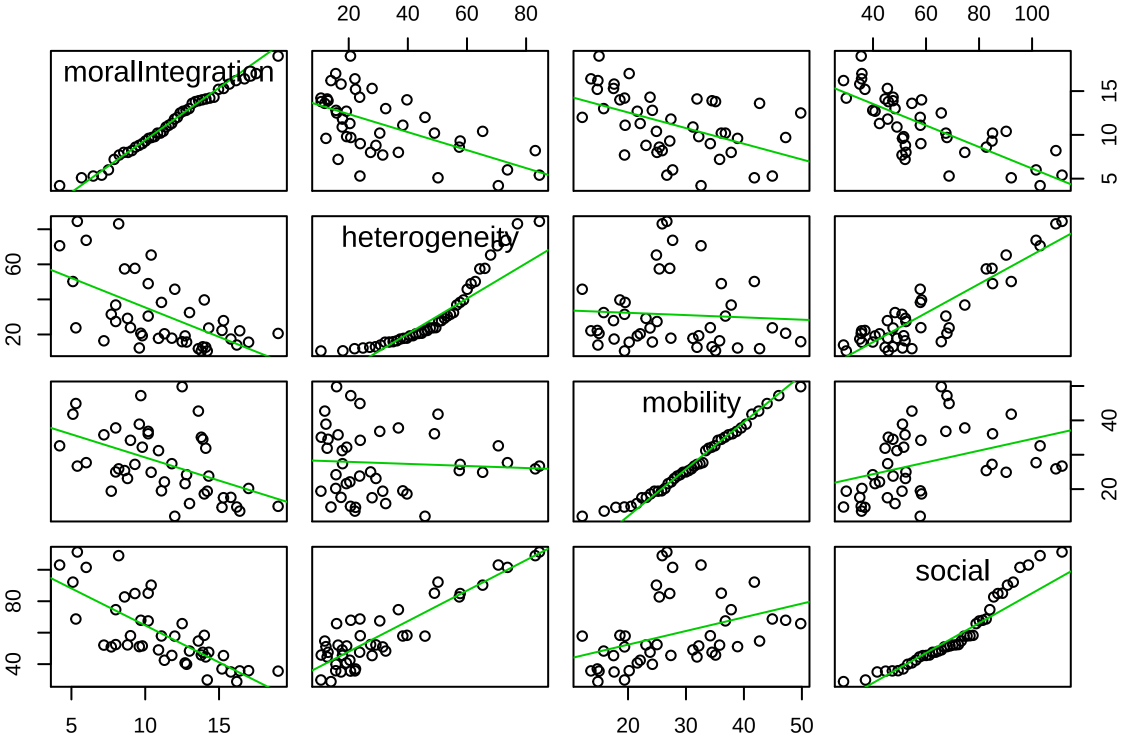
e. Run the provided code, and then perform a hypothesis test for *heterogeneity* and compare it with the one obtained in pard d. Assume we don't know how the predictor *social* was obtained, why do we see seemingly contradictory results?

```
set.seed(100)
n = dim(cities)[1]
cities$social = with(cities, heterogeneity+mobility+rnorm(n,0,.1))
mod1 = lm(moralIntegration ~ heterogeneity + mobility + social, data= cities)

summary(mod1)
```

```
##
## Call:
## lm(formula = moralIntegration ~ heterogeneity + mobility + social,
##     data = cities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2357 -1.1764 -0.2883  1.7623  4.3731
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     20.030      1.200  16.685   <2e-16 ***
## heterogeneity   -4.077      4.527  -0.900    0.373
## mobility        -4.165      4.531  -0.919    0.364
## social           3.968      4.527   0.876    0.386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.25 on 39 degrees of freedom
## Multiple R-squared:  0.6316, Adjusted R-squared:  0.6033
## F-statistic: 22.29 on 3 and 39 DF,  p-value: 1.427e-08
```

```
scatterplotMatrix(~moralIntegration + heterogeneity + mobility
                  + social, data = cities, smoother = FALSE, diagonal = "qqplot")
```



Asamsuming I am unaware of how *social* was obtained, I would guess that the reason why a hypothesis test for *heterogeneity* now fails to reject the null hypothesis that $\beta_1 = 0$ at a p-value of 0.373 when before we were able to reject the null is because the new variable introduced to the model is highly correlated with *heterogeneity* and the two variables are quite possibly explaining the same variation and so when we look at the case where *heterogeneity* is added to a model already containing *social*, it doesn't contribute much additional influence to the model and so shows up as an insignificant factor. This conclusion is further supported by the scatterplot of *social* on *heterogeneity*, which shows a clearly positive correlation between the two regressors.