STAT-S631

Assignment 6

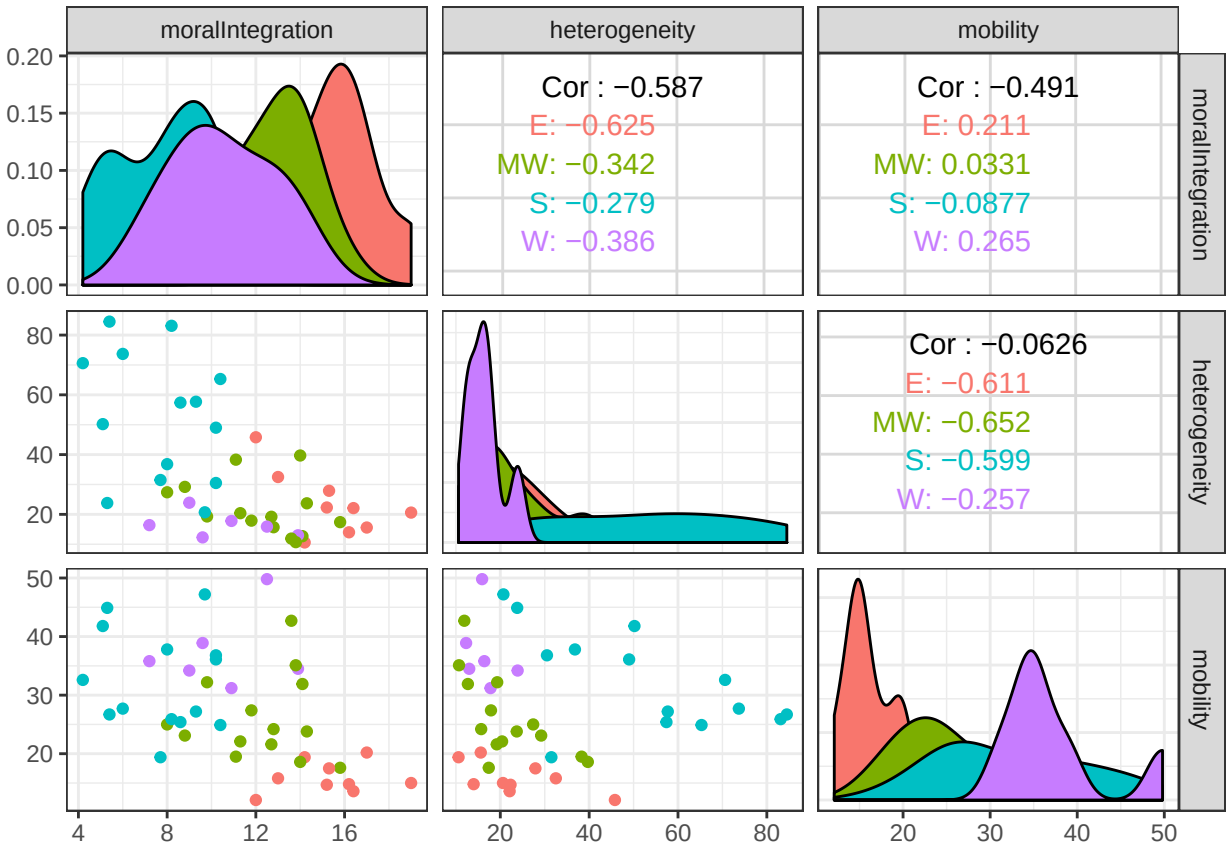
John Koo

```
dp <- loadNamespace('dplyr')
import::from(magrittr, `%>%`, `%<>%`)
library(ggplot2)
theme_set(theme_bw())
import::from(GGally, ggpairs)
```

```
angell.df <- read.table('~/dev/stats-hw/stat-s631/Angell.txt',
                        stringsAsFactors = FALSE) %>%
  dp$mutate(city = rownames(.))
```

Part a

```
ggpairs(angell.df,
        columns = c('moralIntegration', 'heterogeneity', 'mobility'),
        aes(colour = region))
```



There appears to be a linear trend between `moralIntegration` and each of the predictors, `heterogeneity` and `mobility`. At the very least it can be said that there is no reason to not believe that the trend is

linear—there are no obvious changes in the slope, and there is no obvious change in the variance of the response variable for various values of each predictor. There is also no obvious trend between the two predictors. OLS appears to be a plausible method for a model in this case.

Part b

```
model.b <- lm(moralIntegration ~ heterogeneity, data = angell.df)
summary(model.b)
```

```
Call:
lm(formula = moralIntegration ~ heterogeneity, data = angell.df)

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
```

Here $\hat{\beta}_0 \approx 14.426$. If `heterogeneity` $= 0$, then the response `moralIntegration` would be estimated to equal this value. However, in this case, we cannot exactly say this since `heterogeneity` never crosses 0 in the data.

$\hat{\beta}_1 \approx -0.103$. The model implies that for each unit increase in `heterogeneity`, on average, `moralIntegration` decreases by ~ 0.103. We do not need to account for other predictors in this case.

$R^2 \approx 0.345$. This means that around 34% of the variance in `moralIntegration` is explained by OLS model using `heterogeneity`.

Part c

Before we build the full model, from the scatterplot of `heterogeneity` vs. `mobility`, we can see that there is little to no correlation between the two. So we would expect that adding `mobility` to the model won't take any of `heterogeneity`'s influence in the model. In other words, we do not expect $\hat{\beta}_1$ to change much (here we will say β_1 corresponds to `heterogeneity` while β_2 corresponds to `mobility`).

```
model.c <- lm(moralIntegration ~ heterogeneity + mobility, data = angell.df)
summary(model.c)
```

```
Call:
lm(formula = moralIntegration ~ heterogeneity + mobility, data = angell.df)

Residuals:
   Min     1Q Median     3Q    Max
```

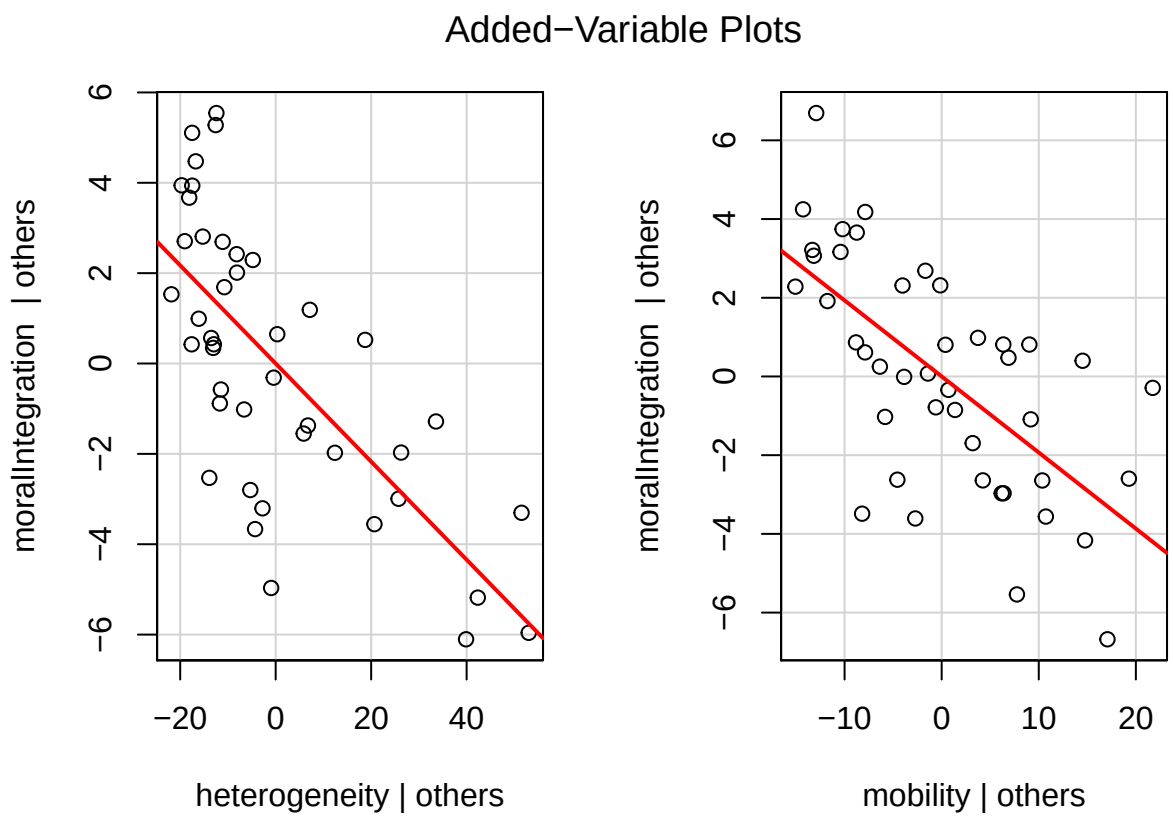
```
-5.071 -1.194 -0.206  1.738  4.195

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
```

As expected, $\hat{\beta}_1$ did not change much. Given fixed `mobility`, one unit increase in `heterogeneity` would result in, on average, a ~ 0.109 decrease in `moralIntegration`. Likewise, given fixed `heterogeneity`, one unit increase in `mobility` would result in, on average, a ~ 0.193 decrease in `moralIntegration`, given our $\hat{\beta}_2$.

Here $\hat{\beta}_0 \approx 19.941$. Like in part (b), this can be interpreted as "If our inputs are 0s, then we expect the response, `moralIntegration` to be equal to 19.941." Of course, like in part (b), the input values in the data never cross 0, so we cannot exactly claim this. One thing to note is that $\hat{\beta}_0$ increased by around 5, which is approximately equal to the sample mean of `mobility` times its coefficient, $\hat{\beta}_2$. This is consistent with our previous observation where the two predictors were uncorrelated.

$R^2 \approx 0.624$. In other words, the full model using both predictors explains around 62% of the variability in `moralIntegration`.

```
car::avPlots(model.c)
```



The added-variable plots show the contribution of each predictor given the other predictor. Both plots are linear, implying that they both contribute to the model, which is consistent with our previous observations about the two predictors being uncorrelated. If instead they were highly correlated, one of the plots would look like a null plot.

Part d

The output of `summary(model.c)` includes the estimate for β_1 as well as the corresponding standard error. These can be used to compute the t-value with the null hypothesis that $\beta_1 = 0$. Here the magnitude of the t-value is very large, which corresponds to a small p-value. Thus we can reject the null hypothesis that $\beta_1 = 0$.

If we want to find a 97% confidence interval for $\hat{\beta}_1$:

```
# confidence level
p <- .97

# compute t
deg.freedom <- model.c$df.residual
t.val <- qt((p + 1) / 2, deg.freedom)

# find estimate and standard error
estimate <- coef(model.c)['heterogeneity']
std.err <- summary(model.c)$coefficients['heterogeneity', 'Std. Error']

# confidence interval
c(-t.val, t.val) * std.err + estimate
```

```
[1] -0.14679334 -0.07032628
```

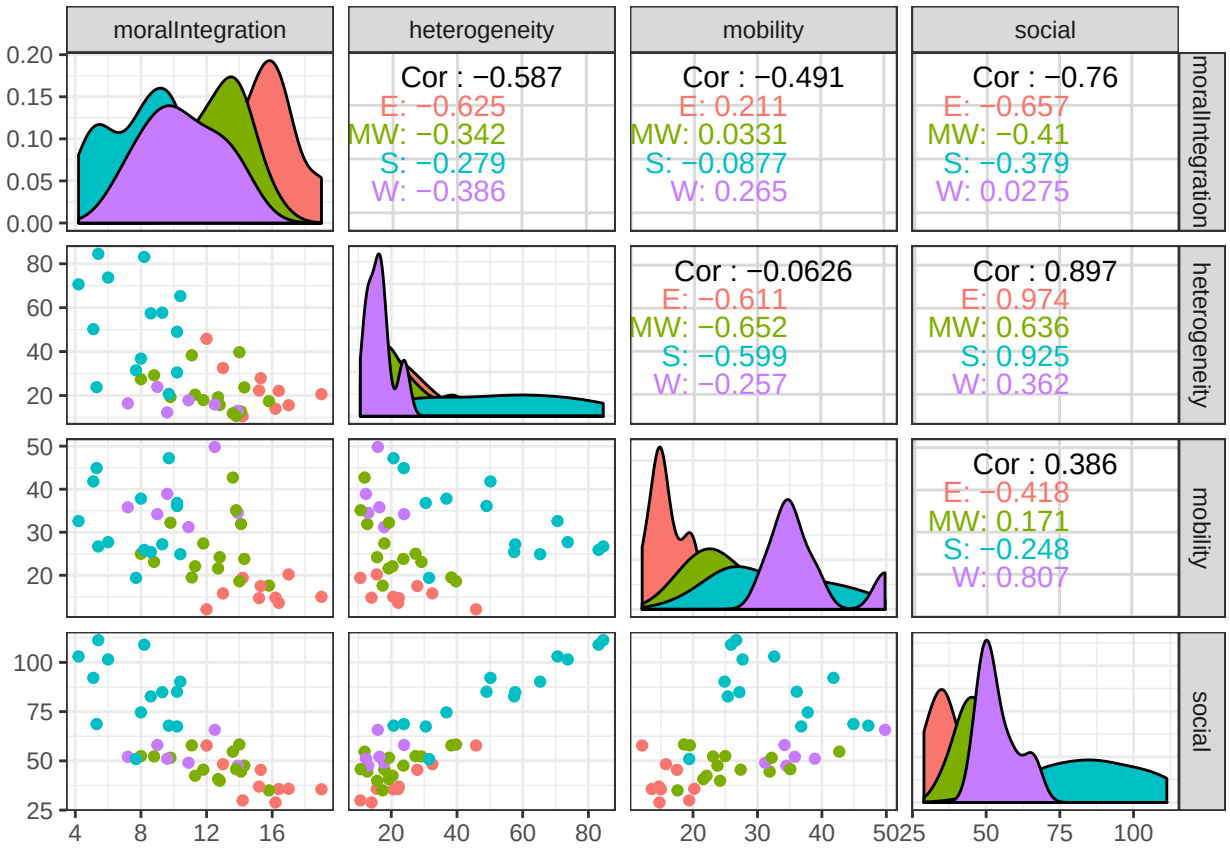
Part e

```
set.seed(100)

# new column that's just a linear combination of the covariates
angell.df %<>%
  dp$mutate(social = heterogeneity + mobility + rnorm(nrow(.), 0, .1))

# model using this new covariate
model.e <- lm(moralIntegration ~ heterogeneity + mobility + social,
              data = angell.df)

summary(model.e)
```

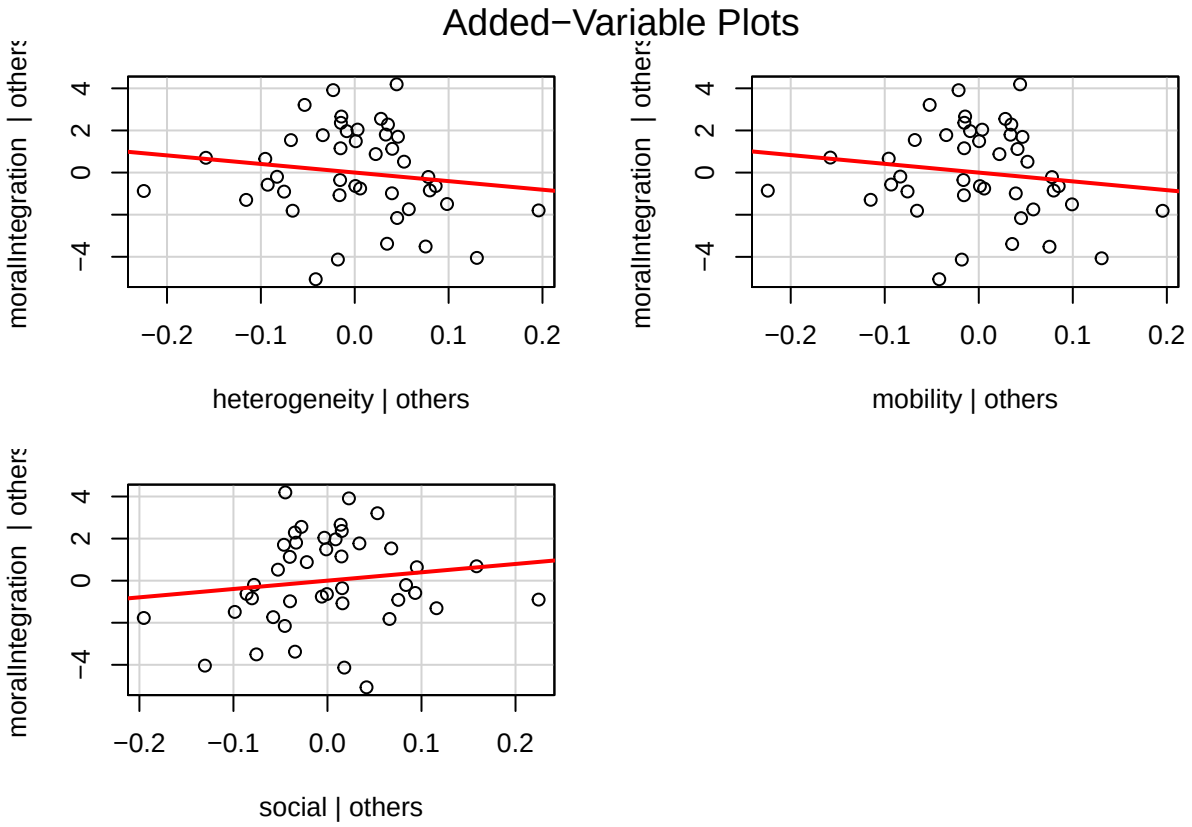
```

Call:
lm(formula = moralIntegration ~ heterogeneity + mobility + social,
    data = angell.df)

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
```

```r
ggpairs(angell.df,
        columns = c('moralIntegration', 'heterogeneity', 'mobility', 'social'),
        aes(colour = region))
```




```r
car::avPlots(model.e)
```



Assuming a reasonable value of α (say, $\alpha = 0.05$), we would fail to reject the null hypothesis that $\beta_1 = 0$. This is true of all of our estimates for the coefficients (except for β_0). On the other hand, our R^2 value is very high, and the F-statistic related to the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$ corresponds to a very low p-value, implying that this model is much better than an intercept-only model. It's also worth noting that the t-test can only make conclusions about the βs individually, not all at once (for that we would need the F-test).

This occurs because one of our predictors is highly correlated with two of the others. We can see the effect of this by comparing the pairwise scatterplots vs. the added-variable plots.