S631 HW7

Erik Parker

October 13, 2017

ALR 4.1: Analyze the *BGSgirls* dataset with the new variables *ave = (WT2+WT9+WT18)/3*, *lin = WT18-WT2*, and *quad = WT2-2WT9+WT18* by regressing them against *BMI18* and comparing with the results in section 4.1.

```r
library(alr4)

girls <- BGSgirls

attach(girls)

girls$ave <- (WT2 + WT9 + WT18)/3
girls$lin <- WT18 - WT2
girls$quad <- WT2 - 2 * WT9 + WT18

detach(girls)

m1 <- lm(BMI18 ~ ave + lin + quad, data = girls)

summary(m1)
```

```
##
## Call:
## lm(formula = BMI18 ~ ave + lin + quad, data = girls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1037 -0.7432 -0.1240  0.8320  4.3485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.30978    1.65517   5.020 4.16e-06 ***
## ave         -0.06778    0.12751  -0.532    0.597
## lin          0.33704    0.07466   4.514 2.68e-05 ***
## quad        -0.02700    0.03976  -0.679    0.499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.333 on 66 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.767
## F-statistic: 76.73 on 3 and 66 DF,  p-value: < 2.2e-16
```

Compared to the results of the full model with all original regressors (model 1) outlined in section 4.1.3, we can see that this new model has fewer regressors which show up as significant in multiple regression terms (i.e. $\beta \neq 0$ at a significance level of $\alpha = 0.05$). In the model outlined in the book, both *WT2* and *WT18* are significant, while in our new model only the linear transformation *lin* (which again is the weight at age 18 - the weight at age 2) turns out to be significant. This result is similar when we compare the new model to the other two shown in the book, which each show two significant regressors (*DW9* and *DW18* in model 2, and *WT2* and *WT18* again in model 3, where *DW9* = *WT9* - *WT2* and *DW18* = *WT18* - *WT9*) compared to the one seen as significant in our model.

The lack of significance seen in *ave* and *quad* in our model makes sense, as they are both linear combinations of the same original predictors, just with slightly different transformations applied to them. While our significant regressor, *lin*, is a more unique linear transformation when compared to the other two which also features only the two predictors shown to be most significant in the models shown in the book (*WT2* and *WT18*).

ALR 4.2: Use the data file *Transact* to examine bank transactions and the time associated with them.

```r
bank <- Transact

bank$a <- (bank$t1 + bank$t2)/2
bank$d <- bank$t1 - bank$t2

m1 <- lm(time ~ t1 + t2, data = bank)
m2 <- lm(time ~ a + d, data = bank)
m3 <- lm(time ~ t2 + d, data = bank)
m4 <- lm(time ~ t1 + t2 + a + d, data = bank)
```

4.2.1: In the fit of M4, some of the coefficients estimates are labeled as "aliased", explain what this means and why it happens.

```r
summary(m4)
```

```
##
## Call:
## lm(formula = time ~ t1 + t2 + a + d, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## t1            5.46206    0.43327  12.607   <2e-16 ***
## t2            2.03455    0.09434  21.567   <2e-16 ***
## a                  NA         NA      NA       NA
## d                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

This means that the regressor under study is completely correlated with another regressor, or a linear combination of regressors, already contained in the model. Here specifically, the regressors *a* and *d* are listed as NA in model four because they are collinear as they are both just linear combinations of the other variables, *t2* and *t1*, already in the model.

4.2.2: What aspects of the fitted regressions are the same? What aspects are different?

```r
summary(m1)
```

```
##
## Call:
## lm(formula = time ~ t1 + t2, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## t1            5.46206    0.43327  12.607   <2e-16 ***
## t2            2.03455    0.09434  21.567   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```r
summary(m2)
```

```
##
## Call:
## lm(formula = time ~ a + d, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## a             7.4966     0.3654  20.514  < 2e-16 ***
## d             1.7138     0.2548   6.726 1.12e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```r
summary(m3)
```

```
##
## Call:
## lm(formula = time ~ t2 + d, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## t2            7.4966     0.3654  20.514   <2e-16 ***
## d             5.4621     0.4333  12.607   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```r
summary(m4)
```

```
##
## Call:
## lm(formula = time ~ t1 + t2 + a + d, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## t1            5.46206    0.43327  12.607   <2e-16 ***
## t2            2.03455    0.09434  21.567   <2e-16 ***
## a                  NA         NA      NA       NA
## d                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

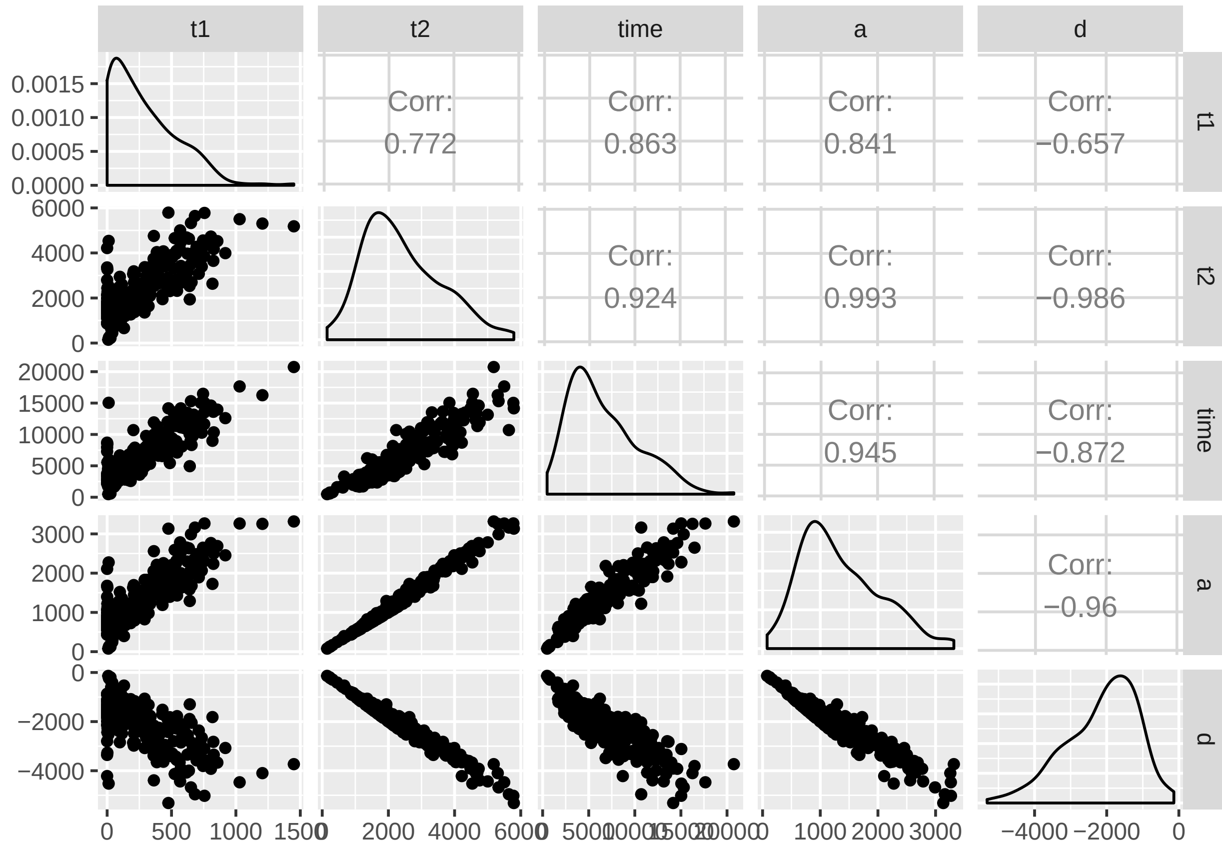
In every model the response variable is the same, as is the intercept estimate $\hat{\beta}_0$, though it is never significantly different from zero. Another similarity is that the same four regressors are used in different combinations in every model, furthermore the p-values and estimated slope coefficients for *t1* and *t2* are the same in models 1 and 4 as the other two regressors in model 4 are omitted so that model is effectively the same as model 1. A final similarity is that the $\hat{\beta}_1$ estimate and p-values in models 2 and 3 are the same, despite being estimates from two different regressors (a and *t2*). This seems to be because both models contain the regressor d which removes the effect of *t2*, so when it is added back in by either a or *t2* itself it has the same effect in both models (this relationship can also be seen in model 3 where d has the same estimated effect as *t1* in model 1, because in this case d is essentially just adding the effect of *t1* alone to the model because *t2* is already present).

The only real differences between these models is that the combinations of our regressors used in each model are different, and that the estimated slope coefficient for d in model 2 is unique. In this model, we noticed earlier that the slope coefficient for a is seen also in model 3 for *t2*, but here d represents something new - the added effect of *t1* without the influence of *t2*, after the average influence of *t1* and *t2* are already accounted for in the model.

4.2.3: Why is the estimate for *t2* different in M1 and M3.

```r
library(GGally)

ggpairs(bank)
```



As was briefly hinted at earlier, and as can be seen in the above plot, *t1* and *t2* are highly correlated. In model 1, the slope estimate for *t2* is calculated when *t1*, the other regressor it is correlated with, is already in the model. In model 3 though, the slope estiamate for *t2* is calculated when a regressor which is obtained by removing the effects of *t2* from *t1*, *d*, is already in the model. So, the estimated slope coefficient for *t2* is higher in model 3 as it represents the full explanatory infleunce of *t2* on the response when it is added to a model that doesn't contain any correlated regressors.

ALR 4.6: In the simple linear regression of log(fertility) on *pctUrban*, the fitted model is $log(\hat{fertility}) = 1.501 - 0.01 pctUrban$. Provide an interpretation of the estimated coefficient for *pctUrban*.

```r
100 * (exp(-0.01) - 1)
```

```
## [1] -0.9950166
```

In this model, we see that an interpretation of the estimated coefficient for *pctUrban* is: that for every additional unit increase in *pctUrban*, we see that the regressor, *fertility*, changes by $100(exp(-0.01) - 1)$ percent In other words, we can say that *fertiltiy* decreases by 0.99% for every percentage increase in Urbanization.

ALR 4.7: Verify that in the regression $log(fertility)$ $log(ppgdp) + lifeExpF$ a 25% increase in *ppgdp* is associated with a 1.4% decrease in expected fertility.

```
un <- UN11

mun <- lm(log(fertility) ~ log(ppgdp) + lifeExpF, data = un)

ppgdpco <- coef(summary(mun))[2, 1]

100 * (exp(log(1.25) * ppgdpco) - 1)
```

```
## [1] -1.449583
```

We see that yes, a 25% increase in *ppgdp* leads to a 1.4% decrease in expected *fertility*.