STAT-S631

Assignment 7

John Koo

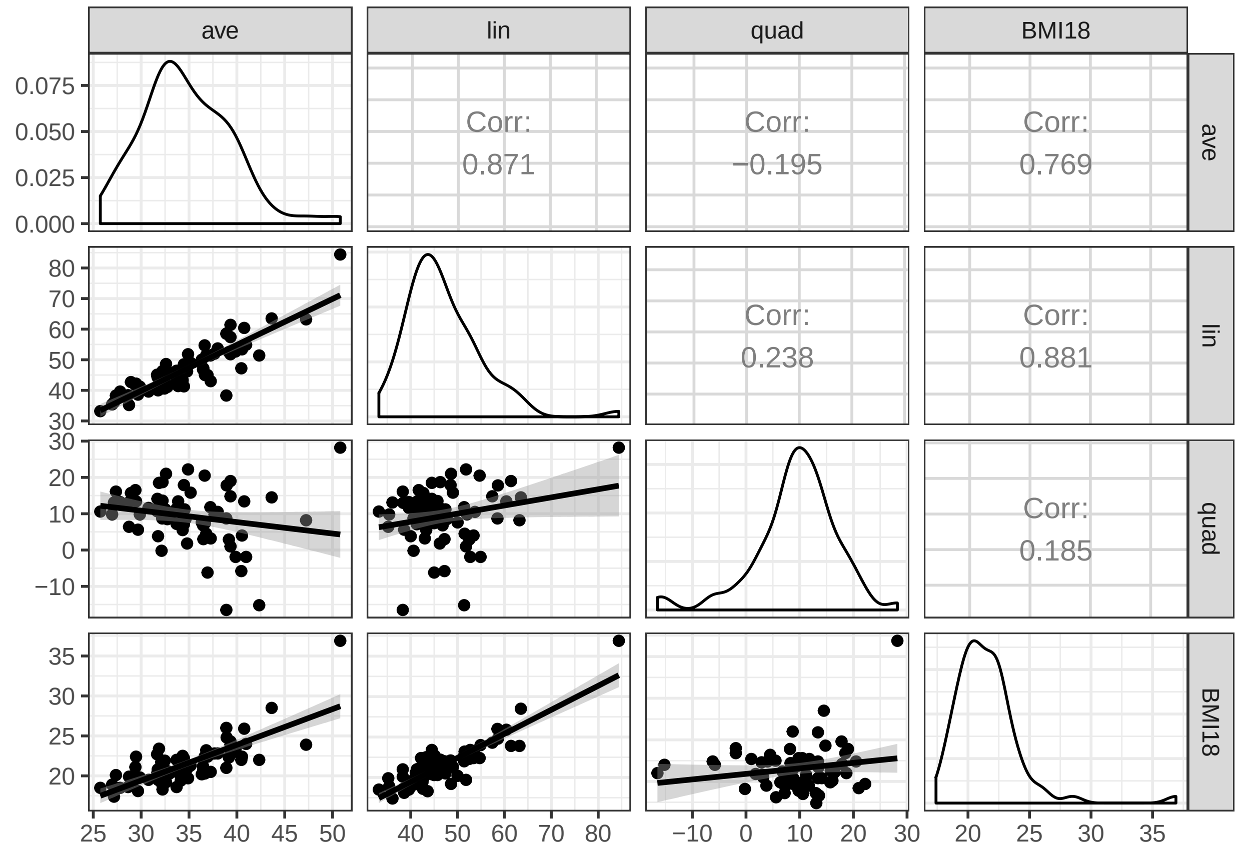
```r
dp <- loadNamespace('dplyr')
import::from(magrittr, `%>%`, `%<>%`)
library(ggplot2)
theme_set(theme_bw())
import::from(GGally, ggpairs)
```

Problem 1

[From ALR 4.1]

```r
# set up data
bgsgirls.df <- alr4::BGSgirls %>%
  dp$mutate(ave = (WT2 + WT9 + WT18) / 3,
            lin = WT18 - WT2,
            quad = WT2 - 2 * WT9 + WT18)

# pairwise scatterplot
bgsgirls.df %>%
  dp$transmute(ave, lin, quad, BMI18) %>%
  ggpairs(lower = list(continuous = 'smooth'))
```




```r
# fit model
model.1 <- lm(BMI18 ~ ave + lin + quad, data = bgsgirls.df)
summary(model.1)
```

```

Call:
lm(formula = BMI18 ~ ave + lin + quad, data = bgsgirls.df)

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
```

```r
# also model from 4.1
model.4.1 <- lm(BMI18 ~ WT2 + WT9 + WT18, data = bgsgirls.df)
summary(model.4.1)
```

```
Call:
lm(formula = BMI18 ~ WT2 + WT9 + WT18, data = bgsgirls.df)

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 ***
WT2         -0.38663    0.15145  -2.553    0.013 *
WT9          0.03141    0.04937   0.636    0.527
WT18         0.28745    0.02603  11.044  < 2e-16 ***
---
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
```

From the scatterplot matrix, we know that while both `ave` and `lin` correlate fairly strongly and positively with `BMI18`, they also correlate with each other. On the other hand, `quad` doesn't seem to correlate very strongly with any of the variables.

From the model summary, we can see that only the intercept term and the coefficient estimate for `lin` are significant at reasonable significance levels (e.g., $\alpha = 0.01$ or 0.05). This is consistent with what we observed in the scatterplots: `quad` has no correlation with `BMI18`, and although both `ave` and `lin` correlate with `BMI18`, they correlate with each other, so we expect some unexpected behavior. In this case, the model fit `BMI18` on `lin` and not on `ave`.

Comparing the results here to the example in Section 4.1, we can see that $\hat{\beta}_0$ is the same here as it is in the example. This is expected, since the regressors are all from the same set of three predictors, and although they end up being at different scales, this can just be adjusted by the estimates for the βs. Summary statistics such as the residual standard error, R^2, and F-statistic are identical, which is, again, expected, since the regressors are derived from the same set of predictors.

Problem 2

[From ALR 4.2]

```r
transact.df <- car::Transact %>%
  dp$mutate(a = (t1 + t2) / 2,
            d = t1 - t2)

m1 <- lm(time ~ t1 + t2, data = transact.df)
summary(m1)
```

```
Call:
lm(formula = time ~ t1 + t2, data = transact.df)

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
m2 <- lm(time ~ a + d, data = transact.df)
summary(m2)
```

```
Call:
lm(formula = time ~ a + d, data = transact.df)

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
m3 <- lm(time ~ t2 + d, data = transact.df)
summary(m3)
```

```
Call:
lm(formula = time ~ t2 + d, data = transact.df)

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
m4 <- lm(time ~ t1 + t2 + a + d, data = transact.df)
summary(m4)
```

```

Call:
lm(formula = time ~ t1 + t2 + a + d, data = transact.df)

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
```

Part 1

Coefficients are omitted when the model matrix isn't full rank. Since `a` and `d` are linear combinations of `t1` and `t2`, the model matrix of `m4`, which has five columns, isn't of rank 5.

Part 2

All of the models have the same $\hat{\beta}_0$ (including standard error), R^2, residuals (and therefore, RMSE), F-statistic, and degrees of freedom (since two of the four regressors are omitted in `m4`). However, they do not have the same $\hat{\beta}_1$ or $\hat{\beta}_2$, even when the corresponding regressors are the same.

Part 3

`d` is a linear combination of `t1` and `t2`. Therefore, `m2` and `m3` are equivalent models.

Starting with model 2:

$$
\begin{aligned}
E[Y|a,d] &= \beta_0 + \beta_{21}a + \beta_{22}d \\
&= \beta_0 + \beta_{21}\frac{t_1+t_2}{2} + \beta_{22}(t_1 - t_2) \\
&= \beta_0 + (\frac{\beta_{21}}{2} + \beta_{22})t_1 + (\frac{\beta_{21}}{2} - \beta_{22})t_2
\end{aligned}
$$

Therefore, for model 3, $\beta_{31} = \frac{\beta_{21}}{2} + \beta_{22}$ and $\beta_{32} = \frac{\beta_{21}}{2} - \beta_{22}$.

Problem 3

[From ALR 4.6 and 4.7]

$$\log(\hat{\text{fertility}}) = 1.501 - 0.01\text{pctUrban}$$

So . . .

$$\begin{aligned}\hat{\text{fertility}} &= \exp(1.501 - .01\text{pctUrban}) \\ &= 4.486e^{-0.01\text{pctUrban}}\end{aligned}$$

Then for one unit increase in `pctUrban`, `fertility` is multiplied by $4.487e^{-0.01}$ (on average).

```r
un11.df <- alr4::UN11
model.4.7 <- lm(log(fertility) ~ log(ppgdp) + lifeExpF, data = un11.df)
summary(model.4.7)
```

```
Call:
lm(formula = log(fertility) ~ log(ppgdp) + lifeExpF, data = un11.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61778 -0.16891  0.03731  0.17591  0.61072

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50736    0.12707  27.601  < 2e-16 ***
log(ppgdp)  -0.06544    0.01781  -3.675 0.000307 ***
lifeExpF    -0.02824    0.00274 -10.306  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.248 on 196 degrees of freedom
Multiple R-squared:  0.6926,    Adjusted R-squared:  0.6894
F-statistic: 220.8 on 2 and 196 DF,  p-value: < 2.2e-16
```

The model here is:

$$E[\log Y | x_1, x_2] = \beta_0 + \beta_1 \log x_1 + \beta_2 x_2$$

This is equivalent to (after some algebra):

$$E[Y | x_1, x_2] = \gamma_0 x_1^{\beta_1} e^{\beta_2 x_2}$$

Where $\gamma_0 = e^{\beta_0}$.

Then if `ppgdp`, which is x_1 in this case (and replacing our parameters with estimates) we can verify that a 25% increase in x_1 would yield:

$$\frac{\hat{\gamma}_0(1.25x_1)^{\hat{\beta}_1}e^{\hat{\beta}_2x_2})}{\hat{\gamma}_0x_1^{\hat{\beta}_1}e^{\hat{\beta}_2x_2}}$$

$$= \frac{(1.25x_1)^{\hat{\beta}_1}}{x_1^{\hat{\beta}_1}}$$

$$= 1.25^{\hat{\beta}_1} \approx 0.9855$$

And $1 - 0.9855 = 0.014497$