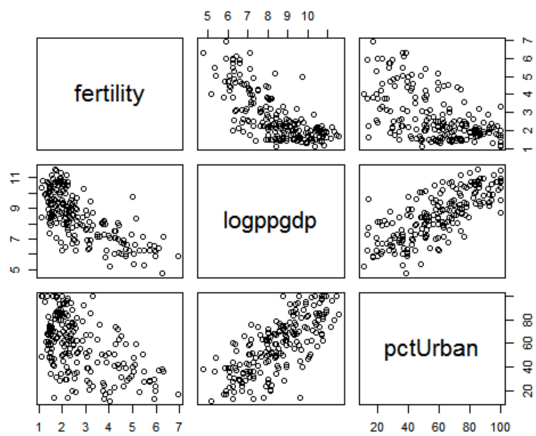
Sample solutions

Stat 8051 Homework 2

Problem 1: ALR Exercise 3.2



3.2.1 There seem to be sizeable linear dependency among all pairs of variables.

3.2.2

```
> m1 = lm(fertility~logppgdp, data=data32)
> m2 = lm(fertility~pctUrban, data=data32)
> summary(m1)

Call:
lm(formula = fertility ~ logppgdp, data = data32)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16313 -0.64507 -0.06586  0.62479  3.00517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.00967    0.36529   21.93   <2e-16 ***
logppgdp    -0.62009    0.04245  -14.61   <2e-16 ***
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9305 on 197 degrees of freedom
Multiple R-squared:   0.52,Adjusted R-squared:  0.5175
F-statistic: 213.4 on 1 and 197 DF,  p-value: < 2.2e-16

> summary(m2)

Call:
lm(formula = fertility ~ pctUrban, data = data32)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4932 -0.7795 -0.1475  0.6517  2.9029

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.559823   0.213681  21.339   <2e-16 ***
pctUrban    -0.031045   0.003421  -9.076   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.128 on 197 degrees of freedom
Multiple R-squared:  0.2948,Adjusted R-squared:  0.2913
F-statistic: 82.37 on 1 and 197 DF,  p-value: < 2.2e-16
```

3.2.3 `logppgdp` seems to be still useful after adjusting for `pctUrban`, but not the converse.

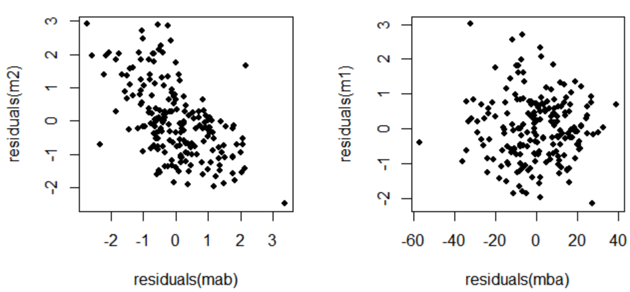


Figure 1: Added variable plots for `fertility` vs. (L) `logppgdp` and (R) `pctUrban`

A summary of the full model confirms this finding. The coefficient for `logppgdp` is significant, but not the one for `pctUrban`.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  7.9932699  0.3993367  20.016   <2e-16 ***
logppgdp    -0.6151425  0.0641565  -9.588   <2e-16 ***
pctUrban    -0.0004393  0.0042656  -0.103    0.918
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9328 on 196 degrees of freedom
Multiple R-squared:   0.52,Adjusted R-squared:  0.5151
F-statistic: 106.2 on 2 and 196 DF,  p-value: < 2.2e-16
```

3.2.4

```
> m.res2ba = lm(residuals(m2)~residuals(mab))
> m.res2ba$coef
   (Intercept) residuals(mab)
 -1.985664e-16  -6.151425e-01
```

The coefficient of slope term in the regression of appropriate residuals is same as that of `logppgdp` in the full model.

3.2.5 We only check the first few elements of the two residuals. They are same (use `View` command to view all residuals).

```
> two.resids = cbind(residuals(m.full), residuals(m.res2ba))
> head(two.resids)
         [,1]        [,2]
1  1.80647138  1.80647138
2 -1.39472572 -1.39472572
3 -0.65106464 -0.65106464
4  2.31728224  2.31728224
5 -0.08777158 -0.08777158
6 -0.16857634 -0.16857634
```

3.2.6

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.986e-16  6.596e-02   0.000        1
residuals(mab) -6.151e-01  6.399e-02  -9.613   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9305 on 197 degrees of freedom
Multiple R-squared:  0.3193,Adjusted R-squared:  0.3158
F-statistic:  92.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

In this regression with residuals, the t-statistic for the slope term is -9.613, while in the full model the coefficient for `logppgdp` had a t-statistic -9.613. This minor difference is because of slightly different degrees of freedom ($n-2$ and $n-3$, respectively).

Problem 2: ALR Exercise 4.2

4.2.1

```
> coef(M4)
(Intercept)          t1          t2           a           d
 144.369443    5.462057    2.034549          NA          NA
```

This is because a and d are linearly dependent on the first two predictors.

4.2.2

```
> coef(M1)
(Intercept)          t1          t2
 144.369443    5.462057    2.034549
> coef(M2)
(Intercept)           a           d
 144.369443    7.496605    1.713754
> coef(M3)
(Intercept)          t2           d
 144.369443    7.496605    5.462057
> coef(M4)
(Intercept)          t1          t2           a           d
 144.369443    5.462057    2.034549          NA          NA
```

All intercept terms are same, but coefficient estimates are different.

4.2.3 Because the other predictor variable is different.

Problem 3: ALR Exercise 4.10

We now that for $(X, Y) \sim$ Bivariate normal $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$,

$$Y|X = x \sim N\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1-\rho^2)\right)$$

Thus we get the following:

$$\beta_0 = \mu_y - \rho\mu_x\frac{\sigma_y}{\sigma_x} \quad (1)$$

$$\beta_1 = \rho\frac{\sigma_y}{\sigma_x} \quad (2)$$

$$\sigma^2 = \sigma_y^2(1-\rho^2) \quad (3)$$

We are given that $X \sim N(\mu_x, \sigma_x^2)$. From 1 and 2 we have $\mu_y = \beta_0 + \beta_1\mu_x$. Now squaring 2 and putting $\sigma_y^2 = \sigma^2/(1-\rho^2)$ from 3 gives

$$\beta_1^2 = \frac{\rho^2}{1-\rho^2}.\frac{\sigma^2}{\sigma_x^2} \quad \Rightarrow \quad \rho = \sqrt{\frac{\sigma_x^2\beta_1^2}{\sigma_x^2\beta_1^2 + \sigma^2}}$$

Using this in 3 we get

$$\sigma_y^2 = \frac{\sigma^2}{1-\rho^2} = \sigma^2 + \sigma_x^2\beta_1^2$$

Problem 4: ALR Exercise 4.12

4.12.1 The OLS line and major axis are slightly different.

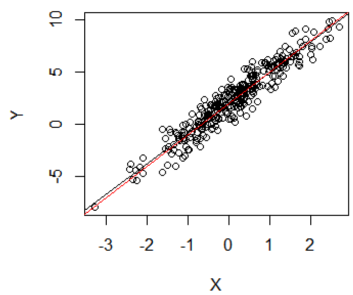


Figure 2: Scatterplot for $\sigma = 1$

4.12.2 The spread seems to be increasing with increasing σ.

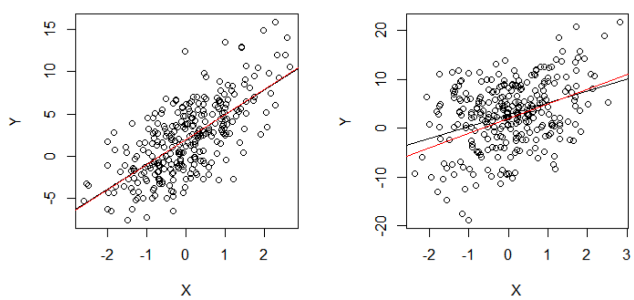


Figure 3: Scatterplot for (L)$\sigma = 1$ and (R) $\sigma = 6$

4.12.3 There are always some points far away from the OLS line. This happens because the Cauchy distribution has heavy tails.

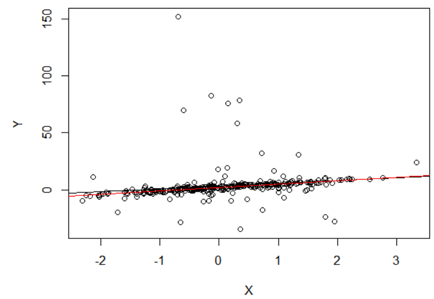


Figure 4: Scatterplot for (L)$\sigma = 1$ and standard Cauchy errors

Problem 5: ALR Exercise 5.8

5.8.1

```
> m1 = lm(Y ~ X1+X2+I(X1^2)+I(X2^2)+X1:X2, data=cakes)
> summary(m1)

Call:
lm(formula = Y ~ (X1 + X2)^2 + I(X1^2) + I(X2^2), data = cakes)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4912 -0.3080  0.0200  0.2658  0.5454

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
X1           2.592e+01  4.659e+00   5.563 0.000533 ***
X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
X1:X2       -4.163e-02  1.072e-02  -3.883 0.004654 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4288 on 8 degrees of freedom
Multiple R-squared:  0.9487,Adjusted R-squared:  0.9167
F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

5.8.2

```
> m2 = update(m1, ~.+block+X1*block+X2*block)
> summary(m2)
```

```
Call:
lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + block + X1:X2 +
    X1:block + X2:block, data = cakes)

Residuals:
       1        2        3        4        5        6        7
-0.01786 -0.01786 -0.01786 -0.01786  0.34714 -0.38286  0.10714
       8        9       10       11       12       13       14
 0.01786  0.01786  0.01786  0.01786 -0.31714  0.31286 -0.06714

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.202e+03  1.754e+02 -12.555 5.69e-05 ***
X1           2.575e+01  3.381e+00   7.616 0.000620 ***
X2           9.927e+00  8.466e-01  11.725 7.93e-05 ***
I(X1^2)     -1.569e-01  2.863e-02  -5.480 0.002758 **
I(X2^2)     -1.195e-02  1.145e-03 -10.437 0.000139 ***
block1      -5.677e+00  8.611e+00  -0.659 0.538883
X1:X2       -4.163e-02  7.779e-03  -5.351 0.003062 **
X1:block1    3.326e-01  1.100e-01   3.024 0.029298 *
X2:block1   -1.672e-02  2.200e-02  -0.760 0.481689
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3112 on 5 degrees of freedom
Multiple R-squared:  0.9831,Adjusted R-squared:  0.9561
F-statistic:  36.4 on 8 and 5 DF,  p-value: 0.0005155
```

Block effect is not significant itself, but has a significant interaction with the predictor `X1`.

Problem 6: ALR Exercise 5.12

5.12.1

```
> plot(HT18~HT9, data=BGSall)
> plot(HT18~HT9,
+      pch=ifelse(Sex==0,1,19),
+      data=BGSall)
> legend("topleft",c("Male","Female"), pch=c(1,19), title="Sex")
```

The two clusters appear to be somewhat different.

5.12.2

```
> m.add = lm(HT18 ~ HT9+Sex, BGSall)
> coefs = coef(m.add)
```

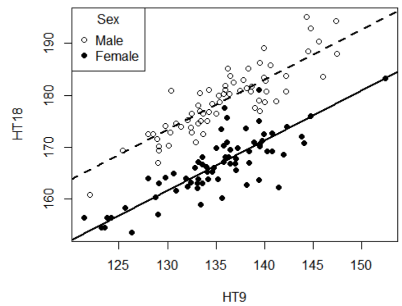


Figure 5: Scatterplot of heights at age 9 vs. age 18, for males and females

```
> abline(coefs[1],coefs[2], lty=2, lwd=2)
> abline(coefs[1]+coefs[3],coefs[2], lwd=2)
```

Calculating the additive model, and putting OLS lines for the two groups makes the distinction clearer.

A test for interaction can be formulated by obtaining the interactive model and comparing it to the additive one:

```
> m.int = update(m.add, ~.^2)
> anova(m.add, m.int  )
Analysis of Variance Table

Model 1: HT18 ~ HT9 + Sex
Model 2: HT18 ~ HT9 + Sex + HT9:Sex
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    133 1566.9
2    132 1532.5  1    34.409 2.9638 0.08749 .
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We fail to reject the null hypothesis of no interaction at 95% level. But the p-value is borderline so we cannot do that with too much emphasis. More so because there is evidence of the effect of sex in the scatterplot.

5.12.3 A 95% confidence interval of difference of intercepts in the additive model is simply that of the coefficient of sex in that model.

```
> confint(m.add, "Sex")
        2.5 %    97.5 %
Sex -12.86355 -10.52813
```