Regression Analysis, Fall 2001: Solutions to HW4

October 8, 2001

---

1. Problem 1

   (a) Quadratic forms

      i.

$$\begin{aligned} RSS &= \text{Y'Y - b'X'Y} \\ &= \text{Y'Y - (HY)'Y} \\ &= \text{Y'IY - Y'HY} \\ &= \text{Y'(I-H)Y} \end{aligned}$$

      ii.

$$\begin{aligned} SSReg &= b'X'Y - (1/n)Y'JY \\ &= Y'HY - (1/n)Y'JY \\ &= Y'(H - (1/n)J)Y \end{aligned}$$

      iii.

$$\begin{aligned} SYY &= Y'Y - (1/n)Y'JY \\ &= Y'IY - (1/n)Y'JY \\ &= Y'(I - (1/n)J)Y \end{aligned}$$

   (b) Trace of matrices

      i. Show: tr(AB) = tr(BA)

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

$$B = \begin{bmatrix} b_{11} & \dots & b_{1m} \\ b_{21} & \dots & b_{2m} \\ b_{n1} & \dots & b_{nm} \end{bmatrix}$$

$$AB = \begin{bmatrix} a_{11}b_{11} + \ldots + a_{1n}b_{n1} & \dots \\ \dots & a_{m1}b_{1m} + \ldots + a_{mn}b_{nm} \end{bmatrix}$$

$$BA = \begin{bmatrix} b_{11}a_{11} + \ldots + b_{1m}am1 & \dots \\ \dots & b_{n1}a_{1n} + \ldots + b_{nm}a_{mn} \end{bmatrix}$$

tr(AB)=$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ij}a_{ji}$ = tr(BA)

ii. $X'X$ is invertible.
$H = X(X'X)^{-1}X' \rightarrow$

$$\begin{aligned} tr(H) &= tr(X(X'X)^{-1}X') &= tr((X'X)^{-1}X'X) \\ &= tr(I(k+1)) &= k+1 \end{aligned}$$

---

2. Problem 2: Auction Clocks revisited.

(a) Model with interaction term:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Here is the results of the regression:

```
Call: lm(formula = Price ~ Age + Bidders + agebid)
Residuals:
  Min     1Q Median    3Q   Max
 -155 -70.43  2.069 47.88 202.3

Coefficients:
                Value Std. Error  t value  Pr(>|t|)
(Intercept)  320.4580  295.1413    1.0858    0.2868
        Age    0.8781    2.0322    0.4321    0.6690
    Bidders  -93.2648   29.8916   -3.1201    0.0042
     agebid    1.2978    0.2123    6.1123    0.0000

Residual standard error: 88.91 on 28 degrees of freedom
Multiple R-Squared: 0.9539
F-statistic: 193 on 3 and 28 degrees of freedom, the p-value is 0
```

Testing the hypothesis that the price-age slope increases as the number of bidders increases is equivalent to testing if the coefficient β_3 is 0. If you imagine fixing X_2 at different levels (increasing) then $\beta_3 X_2$ can be considered as part of the slope of X_1. Therefore, that slope can only increase if β_3 is positive.

- test 1: the Student's t test:
$$H_0 : \beta_3 > 0 H_1 : \beta_3 > 0$$
This is a one sided t-test, 95% CI: $\hat{\beta_3} \pm t_{0.05,28} = (1.2978 \pm 1.701 * 0.2123) = (0.9367, 1.6589)$. Clearly, this CI does not include the value 0, so I reject the hypothesis that β_3 equals 0.
- test 2: Use the F-tests as an alternative to Student's t for two tailed hypotheses about the regression coefficients. Here is the anova table:
```
> anova(lm(Price~Age+Bidders+Age*Bidders))
Analysis of Variance Table

Response: Price

Terms added sequentially (first to last)
            Df Sum of Sq Mean Sq  F Value        Pr(F)
Age          1   2555224 2555224 323.2092 0.000000e+00
Bidders      1   1727838 1727838 218.5535 0.000000e+00
Age:Bidders  1    295364  295364  37.3605 1.353474e-06
Residuals   28    221362    7906
```

The p-value associated to the F statistic is very small. Note that the F-statistic is the square of the t-statistic for the two sided hypothesis on β_3. The F test is appropriate for testing two sided hypothesis of a specific coefficient or for a set of coefficients. Therefore, the second method does not address directly the questionof interest, i.e. whether the coefficient corresponding to the interaction term is positive or not: it just tests that the coefficient is zero.

(b) Writing short report: Sample writing

Dear Sir:

I am writing to inform you about my updated analysis on the dataset you sent me a week ago. I have fitted two linear models to the data. In the first model, I found that the price of a given clock increases as either its age for the number of bidders at the auction increases, a result which agrees with my intuition: the more bidders there are, the higher the price could become. Likewise, the more antique a clock is, the more expensive it can be. If I keep the number of bidders unchanged, one year increase in the age of the clock causes an increse of \$12.74 of the auction price. If I treat the age of the clock as a constant, another person joining the auction would cause an increase of \$85.95 in the price of the clock. The age of the clock, together with the number of bidders, explain 89.23% of the variation in the price of that clock.

In the second model, I want to examine if there iss any iteraction between the two predictors: age of clocks and number of bidders. To my surprise, the coefficient of the number of bidders drops from \$85.95 in the previous model to \$-93.26. This number is not intuitive to me: can a bigger number of bidders taking part in the auction cause a decline in the selling price of the clock? The coefficient of the age of the clock remained positive but very small. Furthermore, I can no longer reject the hypothesis that this coefficient is zero. The level of significance of the interaction term is high and 95.39% of the variation in the prices of the clocks is explained by the regression model. At this stage of the analysis, I recommend the first model since it has a firmer statistical evidence. In the mean time, I am planning to design more complicated procedures for this problem. Should you have any questions, please do not hesitate to contact me.

Sincerely yours...

---

3. Problem 3: Canadians

(a) Numerical values for dummy variables.

- Using Dsex:

```
 Call: lm(formula = Meas.Wt ~ DSex, data = davis20)
Residuals:
   Min     1Q Median 3Q  Max
 -18.1 -9.275  -5.05  2 96.9

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  74.0000   7.8757     9.3959   0.0000
       DSex  -4.9000  11.1380    -0.4399   0.6652

Residual standard error: 24.91 on 18 degrees of freedom
Multiple R-Squared: 0.01064
F-statistic: 0.1935 on 1 and 18 degrees of freedom, the p-value is 0.6652
```

- Using CSex

```
Call: lm(formula = Meas.Wt ~ CSex, data = davis20)
Residuals:
   Min     1Q Median 3Q  Max
 -18.1 -9.275  -5.05  2 96.9
```

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  71.5500   5.5690    12.8479   0.0000
       CSex  -2.4500   5.5690    -0.4399   0.6652

Residual standard error: 24.91 on 18 degrees of freedom
Multiple R-Squared: 0.01064
F-statistic: 0.1935 on 1 and 18 degrees of freedom, the p-value is 0.6652
```

- Treating "Sex" as a factor

```
Call: lm(formula = Meas.Wt ~ Sex, data = davis20)
Residuals:
   Min     1Q Median 3Q  Max
 -18.1 -9.275  -5.05  2 96.9

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 71.5500  5.5690    12.8479  0.0000
        Sex  2.4500  5.5690     0.4399  0.6652

Residual standard error: 24.91 on 18 degrees of freedom
Multiple R-Squared: 0.01064
F-statistic: 0.1935 on 1 and 18 degrees of freedom, the p-value is 0.6652
```

The values of the coefficients are affected. However, the results of the t-tests are the same. Here is the anova table for F-test:

```
 > anova(lm(Meas.Wt~Sex,data=davis20))
Analysis of Variance Table
Response: Meas.Wt
Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq   F Value   Pr(F)
Sex        1    120.05 120.0500 0.1935441 0.66522
Residuals 18  11164.90 620.2722
```

This test says that Sex should not be included in the model. I do not believe in this test. I think that males are heavier than females (on average).

(b) Testing for outliers.

i. Regress reported weight on measured weight:

```
Call: lm(formula = Rept.Wt ~ Meas.Wt, data = davis20)
Residuals:
    Min     1Q Median    3Q   Max
 -19.52 -7.267 -1.417 7.685 32.12

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 60.6308  8.4471     7.1777  0.0000
    Meas.Wt  0.0897  0.1120     0.8007  0.4337

Residual standard error: 11.9 on 18 degrees of freedom
Multiple R-Squared: 0.03439
F-statistic: 0.6412 on 1 and 18 degrees of freedom, the p-value is 0.4337
```

I do not see $\beta_{measured\,weight}$ is significant. There is an outlier.

ii. Model with the dummy variable:

```
Call: lm(formula = Rept.Wt ~ Meas.Wt + D12)
Residuals:
    Min     1Q  Median    3Q   Max
 -7.073 -1.636 -0.3011 1.556 5.613

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)   -6.5526    4.7484    -1.3800    0.1855
    Meas.Wt    1.1142    0.0705    15.7977    0.0000
        D12 -122.4094    7.6872   -15.9238    0.0000

Residual standard error: 3.07 on 17 degrees of freedom
Multiple R-Squared: 0.9393
F-statistic: 131.6 on 2 and 17 degrees of freedom, the p-value is 4.521e-11
```

In this model, D12 is 0 except the 12th subject. This dummy variable shows that observation 12 is an outlier because the coefficients are significant when D12 is included in the model.

---

4. Problem 4: Basic Statistics of US states.

   (a) Simple models

Model	β_0	β_i	sd(β_0)	sd(β_i)	R^2	F
X1 vs. X2	17.71	0.055	28.87	0.008	0.446	39.39
X1 vs. X3	-20.42	0.60	94.66	0.26	0.097	5.26
X1 vs. X4	142.6	0.08	28.82	0.04	0.069	3.65

   X2 and X3 seem significant. The intercepts are not significant. Residual plots are good for diagnostics.

   (b) Multiple model

```
Call: lm(formula = x1 ~ x2 + x3 + x4)
Residuals:
    Min     1Q Median    3Q   Max
 -60.24 -15.74 -1.156 15.88 51.38

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -286.8388   64.9199    -4.4183    0.0001
         x2    0.0807    0.0093     8.6738    0.0000
         x3    0.8173    0.1598     5.1151    0.0000
         x4   -0.1058    0.0343    -3.0863    0.0034

Residual standard error: 26.69 on 47 degrees of freedom
Multiple R-Squared: 0.6896
F-statistic: 34.81 on 3 and 47 degrees of freedom, the p-value is 5.337e-12
```

   All variables are significant in explaining the variation in X1.

---

5. Problem 5

   (a)

```
Call: lm(formula = resid(lm(x1 ~ x3 + x4)) ~ resid(lm(x2 ~ x3 + x4)))
Residuals:
    Min     1Q Median    3Q   Max
 -60.24 -15.74 -1.156 15.88 51.38

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

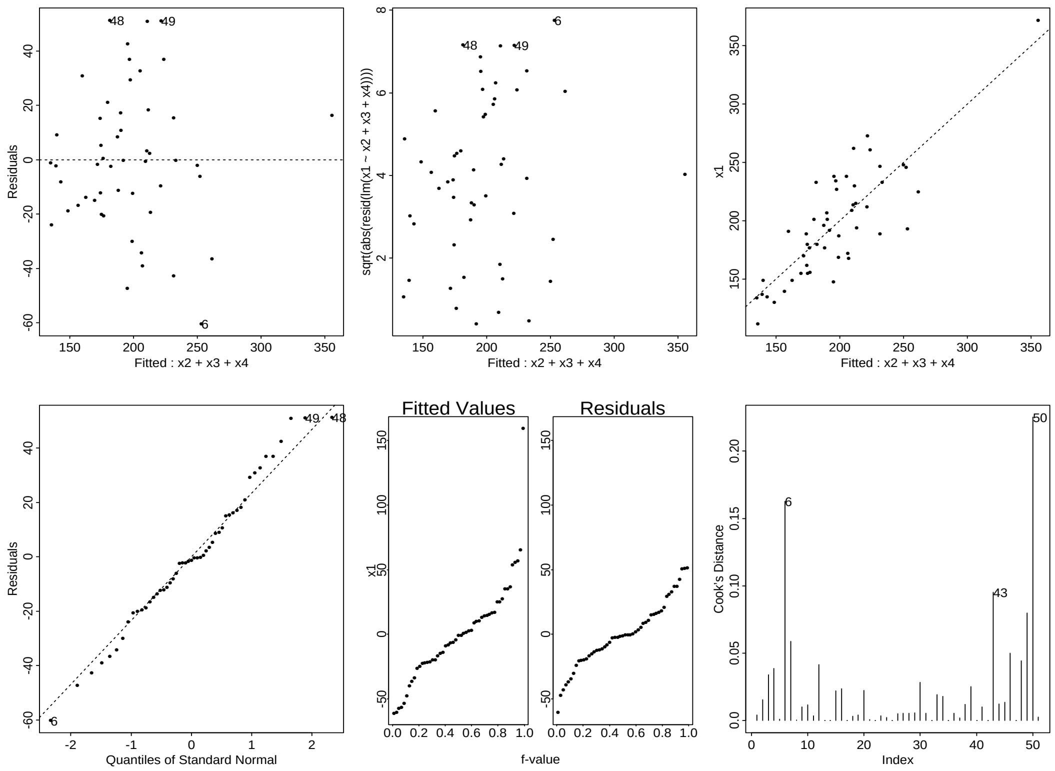

Figure 1: Diagnostics for Multiple Regression Model

```
              (Intercept) 0.0000 3.6608     0.0000  1.0000
resid(lm(x2 ~ x3 + x4)) 0.0807 0.0091     8.8564  0.0000

Residual standard error: 26.14 on 49 degrees of freedom
Multiple R-Squared: 0.6155
F-statistic: 78.44 on 1 and 49 degrees of freedom, the p-value is 9.661e-12
```

The coefficient for income is 0.0807 which matches with the result in problem 4.

(b) Partial regression plot: Income seems to be useful in predicting education expenditure in this problem.

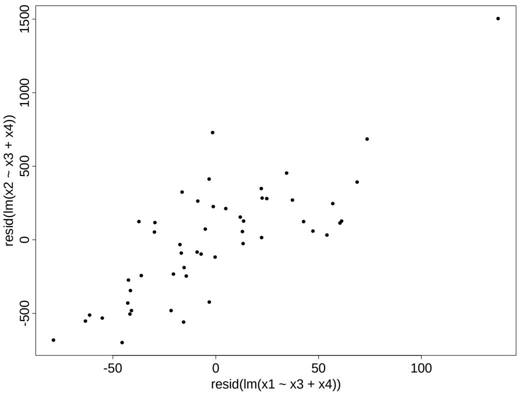

Figure 2: Partial Regression Plot for Income

(c) Intercept in (a) is 0. This is due to the fact that we are regressing random errors.

---

6. Problem 6

(a) using anova()

```
anova(lm(x1~x2+x3+x4),test="F")
Analysis of Variance Table

Response: x1

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
x2         1  48087.10 48087.10 67.48693 0.000000000
x3         1  19537.27 19537.27 27.41921 0.000003767
x4         1   6787.26  6787.26  9.52546 0.003392812
Residuals 47  33489.35   712.54
```

?I do not know how to get the F score for over all.

(b) Extra sum of square principle: Full model includes income, under18 and urban. Reduced model includes income only.
Testing H_0: $\beta_2 = \beta_3 = 0$

```
> (59813.88-33489.35)*(51-3-1)/(2*33489.35)
[1] 18.47233
> qf(.95,2,47)
[1] 3.195056
```

Thus, I reject the null.

(c)
```
> anova(lm(x1~x2+x3+x4),lm(x1~x2))
Analysis of Variance Table

Response: x1

         Terms Resid. Df      RSS   Test Df Sum of Sq  F Value        Pr(F)
1 x2 + x3 + x4        47 33489.35
2           x2        49 59813.88 -x3-x4 -2 -26324.53 18.47233 1.203543e-06
```

same result here...

---

7. Problem 7 The full model:

```
>summary(lm(moral~ethnic+geo+as.factor(region)))
Call: lm(formula = moral ~ ethnic + geo + as.factor(region))
Residuals:
    Min     1Q    Median    3Q   Max
 -3.969 -1.552 -0.000938 1.587 3.381

Coefficients:
                      Value Std. Error  t value Pr(>|t|)
       (Intercept)  15.2532   2.2351     6.8243   0.0000
            ethnic  -0.0665   0.0284    -2.3415   0.0247
               geo  -0.0673   0.0592    -1.1361   0.2632
as.factor(region)1  -1.2625   0.5378    -2.3476   0.0244
as.factor(region)2  -1.1090   0.5108    -2.1709   0.0364
as.factor(region)3  -0.3733   0.2819    -1.3240   0.1936

Residual standard error: 2.128 on 37 degrees of freedom
Multiple R-Squared: 0.6874
F-statistic: 16.27 on 5 and 37 degrees of freedom, the p-value is 1.768e-08
```

Testing the null hypothesis of no region effects:

```
anova(lm(moral~ethnic+geo+region),lm(moral~ethnic+geo))
Analysis of Variance Table

Response: moral

                  Terms Resid. Df      RSS    Test Df Sum of Sq  F Value
1 ethnic + geo + region        37 167.4922
2          ethnic + geo        40 201.2657 -region -3 -33.77355 2.486925
       Pr(F)
1
2 0.07562389
```

The p-value is 0.07562389. I do not reject the null at 5level.

---

8. Problem 8

- Moral vs. Mobility:

```
Call: lm(formula = moral ~ geo)
Residuals:
    Min     1Q Median    3Q   Max
```

```
 -6.104 -2.253 0.4246 2.355 5.543

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  16.1437   1.4518    11.1195   0.0000
        geo  -0.1791   0.0496    -3.6083   0.0008

Residual standard error: 3.149 on 41 degrees of freedom
Multiple R-Squared: 0.241
F-statistic: 13.02 on 1 and 41 degrees of freedom, the p-value is 0.00083
```

- Moral vs. Mobility and Region (no interaction)

```
Call: lm(formula = moral ~ geo + D1 + D2 + D3)
Residuals:
    Min    1Q Median   3Q   Max
 -4.269 -1.59 0.4398 1.69 3.642

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 7.4400 1.7998     4.1338  0.0002
        geo 0.0091 0.0523     0.1743  0.8625
         D1 7.7818 1.2933     6.0170  0.0000
         D2 4.6012 0.9137     5.0357  0.0000
         D3 2.7359 1.1277     2.4261  0.0201

Residual standard error: 2.25 on 38 degrees of freedom
Multiple R-Squared: 0.6411
F-statistic: 16.97 on 4 and 38 degrees of freedom, the p-value is 4.62e-08
```

- Moral vs. Mobility and Region (with interaction)

```
Call: lm(formula = moral ~ geo + D1 + D2 + D3 + geo:D1 + geo:D2 + geo:D3)
Residuals:
    Min     1Q Median    3Q   Max
 -4.267 -1.799 0.3841 1.436 3.785

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  8.4635  2.5527     3.3155  0.0021
        geo -0.0224  0.0763    -0.2939  0.7706
         D1  4.2297  5.5839     0.7575  0.4538
         D2  3.5409  3.5303     1.0030  0.3227
         D3 -1.6216  6.5037    -0.2493  0.8046
     geo:D1  0.1906  0.3178     0.5996  0.5526
     geo:D2  0.0329  0.1184     0.2783  0.7824
     geo:D3  0.1207  0.1754     0.6881  0.4959

Residual standard error: 2.319 on 35 degrees of freedom
Multiple R-Squared: 0.6487
F-statistic: 9.232 on 7 and 35 degrees of freedom, the p-value is 1.999e-06
```

END

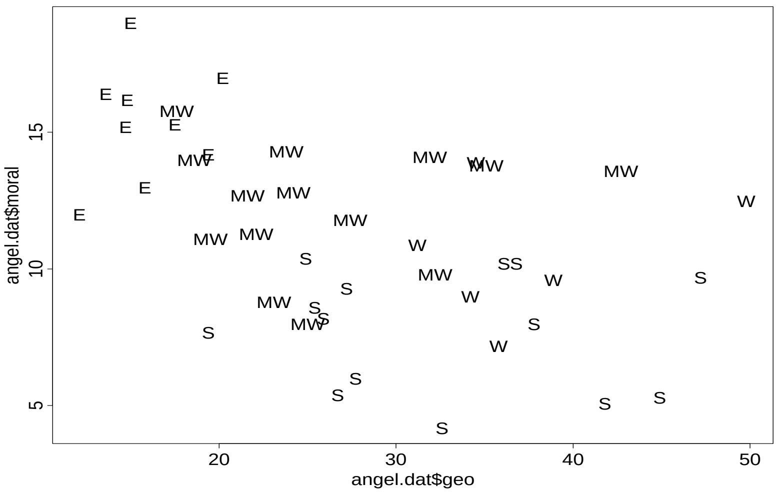

Figure 3: Moral Intergration vs. Mobility in Different Regions