Correlation in Multiple Regression

Applied Regression and Other Multivariable Methods
Sections 10-1 - 10-7

Review

- Recall correlation coefficient r in simple regression
- Describes **strength** of **linear** relationship
- Dimensionless statistic ranging between -1 and 1
- Has same sign as slope estimate

$$r = \tilde{\beta}_1 (S_X/S_Y)$$

- Statistic r^2 is coefficient of determination
- Describes %-age of SSY explained by variable X

$$r^2 = \text{(SSY-SSE)/SSY}$$

- Will now extend these ideas to multiple regression

Correlation Matrix

- In multiple regression, must also consider relationship between predictors
- The corr option in Proc REG gives corr matrix

```
/* From Topic 8 */
proc reg simple corr;
 model wgt = hgt;
 model wgt = age;
 model wgt = hgt age;
 model wgt = hgt agesq;
 model wgt = hgt age agesq;
run;
--------------------------------
                          Correlation
Variable        hgt         age      agesq        wgt
hgt          1.0000      0.6138     0.6154     0.8143
age          0.6138      1.0000     0.9944     0.7698
agesq        0.6154      0.9944     1.0000     0.7665
wgt          0.8143      0.7698     0.7665     1.0000
```

- Helpful when looking at pairwise relationships
- Need something else to assess strength of
  - Relationship between Y and a set of predictors - **multiple correlation coefficient**
  - Relationship between Y and X after adjusting for other predictors (Z's) - **partial correlation coefficient**
  - Relationship between Y and a set of predictors (X's) after adjusting for other predictors (Z's) - **multiple partial correlation coefficient**

Multiple Correlation Coefficient

- Multiple correlation coefficient denoted $R_{Y|X_1,X_2,...,X_k}$
- Similarly measures strength of "linear" association
- By "linear" association, however, we refer to the relationship between Y and $\hat{Y}$ where

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$$

- NOTE: $R_{Y|X_1,X_2,...,X_k}$ will always be ≥ 0. As Y increases so should $\hat{Y}$ if it is a good predictor.
- NOTE: Square of $R_{Y|X_1,X_2,...,X_k}$ is coefficient of determination when considering all the variables in the model.
- NOTE: Since least squares minimizes SSE, no other linear function of the X's will have a larger multiple correlation coefficient

Partial Correlation Coefficient

- Measures strength of relationship between Y and X after adjusting for other predictors (Z's)
  - The partial coefficient r_{YX} is zero order
  - The partial coefficient $r_{YX|Z_1}$ is 1st order
  - The partial coefficient $r_{YX|Z_1,Z_2}$ is 2nd order
  - The partial coefficient $r_{YX|Z_1,Z_2,\ldots,Z_p}$ is pth order
- Same as correlation of residuals from models Y vs Z's and X vs Z's

$$r_{YX|Z_1,Z_2,\ldots,Z_p} = r_{Y-\hat{Y},X-\hat{X}}$$

- $Y-\hat{Y}$ and $X-\hat{X}$ represent what is unexplained after letting the Z's explain as much as they can in both Y and X. Thus we're looking at the correlation between Y and X after adjusting for the Z's
- First order partial correlation coefficient is adjustment of the zero-order coefficient.

$$r_{YX|Z} = \frac{r_{YX} - r_{YZ}r_{XZ}}{\sqrt{(1-r_{YZ}^2)\,(1-r_{XZ}^2)}}$$

Partial Correlation Coefficient

- Similarly, pth order partial correlation coefficient is adjustment of $p-1$ order partial correlation coefficient

$$r_{YX|Z_1,Z_2,\ldots,Z_p} = \frac{r_{YX|Z_1,Z_2,\ldots,Z_{p-1}} - r_{YZ_p|Z_1,Z_2,\ldots,Z_{p-1}}r_{XZ_p|Z_1,Z_2,\ldots,Z_{p-1}}}{\sqrt{\left(1-r^2_{YZ_p|Z_1,Z_2,\ldots Z_{p-1}}\right)\left(1-r^2_{XZ_p|Z_1,Z_2,\ldots,Z_{p-1}}\right)}}$$

- Can obtain the squared coefficients from SAS
- The squared coeffs can be expressed

$$r^2_{\hat{Y}X|Z_1,Z_2,\ldots,Z_p} = \frac{(\text{Residual SS}_{\text{Small}} - \text{Residual SS}_{\text{Large}})}{\text{Residual SS}_{\text{Small}}}$$

where **Small** is Y vs Z's and **Large** is Y vs X, Z's

- Use partial partial F test to test $H_0 : \rho_{YX|Z} = 0$
- Can use multiple corr coef to calculate this

$$F = \frac{(r^2_{\text{Large}} - r^2_{\text{Small}})/(\text{Model df}_{\text{Large}} - \text{Model df}_{\text{Small}})}{(1-r^2_{\text{Large}})/\text{Residual df}_{\text{Large}}}$$

Example

```
/* Consider Exercise 10-1 */

data problem81;
 infile 'i:\.www\datasets502\EX0502.DAT' firstobs=2 dlm='09'x;
 input person sbp quet age smk;

/* These commands generate output in text */

proc reg corr;
 model sbp = age smk / pcorr1 pcorr2;
 model sbp = age quet/ pcorr1 pcorr2;
 model sbp = age smk quet / pcorr1 pcorr2;

/* Example that links first order regr coefficient
   with a regression of residuals */

proc reg noprint;
 model sbp = age;
 output out=resyz r=resyz;
proc reg noprint;
 model smk = age;
 output out=resxz r=resxz;
proc reg;
 model resyz = resxz;
run;
quit;
```


Correlation

Variable	age	smk	sbp	quet
age	1.0000	-0.1395	0.7752	0.8028
smk	-0.1395	1.0000	0.2473	-0.0714
sbp	0.7752	0.2473	1.0000	0.7420
quet	0.8028	-0.0714	0.7420	1.0000

The zero-order correlations are $r_{\text{SBP,AGE}} = 0.7752$, $r_{\text{SBP,SMK}} = 0.2473$, $r_{\text{SBP,QUET}} = 0.7420$. It appears that AGE has the strongest linear association.

Will now look at the partial correlation coefficients of (SBP,SMK) and (SBP,QUET) after adjusting for AGE.

Can use zero order correlations given above

$$\begin{aligned} r_{\text{SBP,QUET|AGE}} &= \frac{.7420 - .7752(.8028)}{\sqrt{(1-.7752^2)(1-.8028^2)}} \\ &= .3177 \\ r_{\text{SBP,SMK|AGE}} &= \frac{.2473 - .7752(-.1395)}{\sqrt{(1-.7752^2)(1-.1395^2)}} \\ &= .5682 \end{aligned}$$

Here is appears SMK is the next variable to add

```
SBP vs SMK        Analysis of Variance
                            Sum of          Mean
Source           DF        Squares        Square  F Value  Pr > F
Model             2     4689.68423   2344.84211    39.16  <.0001
Error            29     1736.28452     59.87188
Corrected Total  31     6425.96875

Root MSE              7.73769    R-Square     0.7298
Dependent Mean      144.53125    Adj R-Sq     0.7112
Coeff Var             5.35365

                        Parameter Estimates
                                                          Squared      Squared
               Parameter Standard                         Partial      Partial
Variable   DF  Estimate    Error t Value Pr > |t| Corr Type I  Corr Type II
Intercept   1  48.04960 11.12956   4.32   0.0002     .            .
age         1   1.70916  0.20176   8.47   <.0001   0.60094      0.71220
smk         1  10.29439  2.76811   3.72   0.0009   0.32291      0.32291
-----------------------------------------------------------------------
SBP vs QUET       Analysis of Variance
                            Sum of          Mean
Source           DF        Squares        Square  F Value  Pr > F
Model             2   4120.59225   2060.29612    25.92  <.0001
Error            29   2305.37650     79.49574
Corrected Total  31   6425.96875

Root MSE              8.91604    R-Square     0.6412
Dependent Mean      144.53125    Adj R-Sq     0.6165
Coeff Var             6.16893

                        Parameter Estimates
                                                          Squared      Squared
               Parameter Standard                         Partial      Partial
Variable   DF  Estimate    Error t Value Pr > |t| Corr Type I  Corr Type II
Intercept   1  55.32344 12.53475   4.4    0.0001       .            .
age         1   1.04516  0.38606   2.71   0.0113   0.60094      0.20175
quet        1   9.75073  5.40246   1.80   0.0815   0.10099      0.10099
```


- From SBP vs SMK

$$\begin{aligned} r^2_{\text{SBP,SMK|AGE}} &= .32291 \\ r^2_{\text{SBP,AGE|SMK}} &= .71220 \\ r^2_{\text{SBP,AGE}} &= .60094 \end{aligned}$$

- From SBP vs QUET

$$\begin{aligned} r^2_{\text{SBP,QUET|AGE}} &= .10099 \\ r^2_{\text{SBP,AGE|QUET}} &= .20175 \\ r^2_{\text{SBP,AGE}} &= .60094 \end{aligned}$$

Appears SMK would be the next variable to include. Can perform a partial F test to see if SMK beneficial.

$$\begin{aligned} F &= \frac{(.7298 - .60094)/(2-1)}{(1-.7298)/29} \\ &= 13.83 \end{aligned}$$

Compare with $F_{1,29}$. Notice that this is the same as the test $H_0 : \beta_{\text{SMK}} = 0$ since $3.72^2 = 13.83$. This means the P-value of this test is 0.0009.

From these partial correlations, one can also see that QUET and AGE are more correlated than AGE and SMK.


```
FULL MODEL        Analysis of Variance
                            Sum of          Mean
Source           DF        Squares        Square  F Value  Pr > F
Model             3     4889.82570   1629.94190    29.71  <.0001
Error            28     1536.14305     54.86225
Corrected Total  31     6425.96875

Root MSE              7.40691    R-Square     0.7609
Dependent Mean      144.53125    Adj R-Sq     0.7353
Coeff Var             5.12478

                        Parameter Estimates
                                                          Squared      Squared
               Parameter Standard                         Partial      Partial
Variable   DF  Estimate    Error t Value Pr > |t| Corr Type I  Corr Type II
Intercept   1  45.10319 10.76488   4.19   0.0003     .            .
age         1   1.21271  0.32382   3.75   0.0008   0.60094      0.33373
smk         1   9.94557  2.65606   3.74   0.0008   0.32291      0.33367
quet        1   8.59245  4.49868   1.91   0.0664   0.11527      0.11527
```

Here $r^2_{\text{SBP,QUET|AGE,SMK}} = .11527$. This is not very large and would suggest it shouldn't be included. We can construct the partial F test

$$\begin{aligned} F &= \frac{(.7609 - .7298)/(3-2)}{(1-.7609)/28} \\ &= 3.64 \end{aligned}$$

Notice this is the same as the parameter t-test $1.91^2 = 3.64$ which has a P-value of 0.0664.


```
The REG Procedure
Dependent Variable: resyz Residual
                            Analysis of Variance
                                   Sum of           Mean
Source              DF         Squares         Square    F Value    Pr > F
Model                1       828.05385       828.05385      14.31    0.0007
Error               30      1736.28452        57.87615
Corrected Total     31      2564.33837

Root MSE              7.60764    R-Square     0.3229
Dependent Mean    -2.4772E-15    Adj R-Sq     0.3003
Coeff Var         -3.07108E17

                        Parameter Estimates
                                  Parameter    Standard
Variable   Label        DF        Estimate       Error t Value  Pr > |t|
Intercept  Intercept     1   -2.6915E-15     1.34485   -0.00    1.0000
resxz      Residual      1      10.29439     2.72158    3.78    0.0007
```

Notice that R-square here is the same as the partial correlation coefficient for SBP vs SMK after adjusting for AGE.

Also, recall that the test for the slope is the same as the test whether $r = 0$. The reason the two tests do not completely agree is because SAS does not know the correct degrees of freedom to use. It uses 1 and 30 here instead of 1 and 29.