Lecture 9: Correlations

Reading Assignment:

- Muller and Fetterman, Chapter 6: “Correlations”

For two random variables X and Y, recall that the correlation ρ is defined as

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}},$$

where $-1 \leq \rho \leq 1$. We estimate the population correlation, ρ, using *Pearson's coefficient of correlation*, given by

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)}}.$$

Consider the simple linear model $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$.

- The *squared correlation coefficient*, R^2, measures the strength of the linear relationship between the variables $\mathbf{y}$ and $\mathbf{x}$. $R^2 \longrightarrow 1$ indicates a stronger linear relationship, and $R^2 \longrightarrow 0$ indicates a weaker linear relationship.

- The *sample correlation coefficient* R is a general measure of the *linear* relationship between two random variables.
  - If $R \approx 0$, there is little evidence of a linear association between two variables. (This does not mean there is no association between the two variables.)
  - As $R \longrightarrow 1$, the linear association is more positive (i.e.,subjects with high values of x will tend to have high values of y as well).
  - As $R \longrightarrow -1$, the linear association is more negative (i.e., subjects with high values of x will likely have low values of y).

We can interpret R as the expected change in the response (in units of standard deviation) associated with a change of one standard deviation in the predictor of interest.

To see this, standardize the response by subtracting its mean and dividing by its standard deviation, standardize the predictor x in the same way, and then regress the standardized response on the standardized predictor, omitting an intercept from the model. The resulting estimate of the regression coefficient is R. We verify this using the ozone data below.

```
proc corr;
var personal outdoor;
run;

proc standard data=ozone mean=0 std=1 out=ozone2;
var personal outdoor;
run;

proc glm data=ozone2;
model personal=outdoor/noint;
run;
****************************************************************************
```

```
               Pearson Correlation Coefficients, N = 64
                      Prob > |r| under H0: Rho=0

                               personal       outdoor
                personal       1.00000       0.39916
                                              0.0011

                 outdoor        0.39916       1.00000
                                 0.0011

                          The GLM Procedure

Dependent Variable: personal

                                     Sum of
Source                     DF       Squares    Mean Square   F Value
Model                       1    10.03745850    10.03745850    11.94
Error                      63    52.96254150     0.84067526
Uncorrected Total          64    63.00000000

                    Source                 Pr > F
                    Model                  0.0010
                    Error
                    Uncorrected Total
```

```
        R-Square     Coeff Var      Root MSE    personal Mean
       0.159325    -2.8725E17      0.916883        -0.000000

NOTE: No intercept term is used: R-square is not corrected for the
      mean.

Source                     DF      Type I SS    Mean Square   F Value
outdoor                     1    10.03745850    10.03745850     11.94

                    Source                 Pr > F
                   outdoor                0.0010

Source                     DF    Type III SS    Mean Square   F Value
outdoor                     1    10.03745850    10.03745850     11.94

                    Source                 Pr > F
                   outdoor                0.0010

                                    Standard
  Parameter         Estimate           Error    t Value    Pr > |t|
  outdoor       0.3991550301      0.11551646       3.46      0.0010
```

Example: Correlation Matrix for Ozone Data

The correlation matrix for the ozone data is provided below.

```
The CORR Procedure

         4  Variables:    personal outdoor  home     time_out

             Pearson Correlation Coefficients, N = 64

              personal       outdoor       home      time_out
 personal      1.00000       0.39916       0.53000       0.19452
 outdoor       0.39916       1.00000       0.55582       0.07818
 home          0.53000       0.55582       1.00000      -0.00714
 time_out      0.19452       0.07818      -0.00714       1.00000

             Pearson Correlation Coefficients, N = 64
```

How do we interpret these values?

These correlations may not give us exactly the information we want. Suppose that based on these correlations, we decide to perform a regression of personal exposure on home exposure. Before considering whether outdoor ozone or time spent outdoors should be added next, it would be helpful to know how associated each is with the response *after* we have controlled for home ozone.

The simple correlations do not tell us about

1. the relationship between $O_{PERSONAL}$ and $(O_{OUTDOOR}, O_{HOME},$ and $TIME_{OUT})$ as a group (multiple correlation coefficient),
2. the relationship between $O_{PERSONAL}$ and O_{HOME} after controlling for $TIME_{OUT}$ (partial correlation coefficient), or
3. the relationship between $O_{PERSONAL}$ and the combined effects of $O_{OUTDOOR}$ and O_{HOME} after controlling for $TIME_{OUT}$ (multiple partial correlation coefficient).

Interpreting ρ^2

Corrected ρ^2

Consider decomposing the variance of Y, namely σ_y^2. Because $\sigma_y^2 \geq 0$, it is sensible to ask what fraction (proportion) of the variance is explained by using a linear model to predict Y.

For a model that spans the intercept, one can define the *corrected correlation coefficient*

$$\rho_{\mathsf{c}}^2 = \frac{\sigma^2(\beta_0) - \sigma^2(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \ldots + \beta_{p-1} X_{p-1})}{\sigma^2(\beta_0)}$$

$$\widehat{\rho}_{\mathsf{c}}^2 = R_{\mathsf{c}}^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression}) + SSE}$$

$$= \frac{CSS(\text{Regression})}{CSS(\text{total})}$$

Under HILE Gauss, $\widehat{\rho}_{\mathsf{c}}^2$ is the MLE. Both $0 \leq \rho_c^2 \leq 1$ and $0 \leq R_{\mathsf{c}}^2 \leq 1$.

R_{c}^2 equals the squared univariate correlation between Y and $\widehat{Y}$ and is provided for the ozone data below.

Example: R_c^2 for ozone data

Verify that the value labeled "R-Square" on the SAS output is the corrected R_c^2. How do we interpret this value?

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
*****************************************************************************

                                      Sum of
Source                     DF        Squares    Mean Square   F Value
Model                       3     5034.90667     1678.30222      9.92
Error                      60    10148.19129      169.13652
Corrected Total            63    15183.09796

                   Source                      Pr > F
                   Model                       <.0001

        R-Square     Coeff Var      Root MSE    personal Mean
       0.331613      55.23389      13.00525        23.54578
```

“Uncorrected” ρ^2

The value of uncorrected correlations lies in evaluating models without an intercept and in evaluating the value of an intercept. We follow the default of considering only corrected correlations. Note that the discussion of partial correlations generalizes to uncorrected correlations with, as always, special attention paid to the presence or absence of the intercept.

Generalizing the Application of ρ^2

All GLH tests correspond to comparing two nested models, and all GLH tests correspond to comparing two correlations. We will show that the standard F test (with the SSE estimated from the larger model and not the full model in this case) can be written in terms of the R^2 from the models being compared.

$$
\begin{aligned}
F_{obs} &= \frac{\frac{[SSE(\text{smaller})-SSE(\text{larger})]}{[dfE(\text{smaller})-dfE(\text{larger})]}}{\frac{SSE(\text{larger})}{dfE(\text{larger})}} \\
&= \frac{\frac{[CSS(\text{larger})-CSS(\text{smaller})]}{[dfE(\text{smaller})-dfE(\text{larger})]}}{\frac{SSE(\text{larger})}{dfE(\text{larger})}} = \frac{\frac{[R^2(\text{larger})-R^2(\text{smaller})]}{a}}{\frac{[1-R^2(\text{larger})]}{dfE(\text{larger})}}.
\end{aligned}
$$

We assume the smaller model contains $\{X_1, \ldots X_{g_1}\}$ and the larger model differs by the addition of $\{X_{g_1+1}, \ldots X_{p-1}\}$.

Let $\Delta R^2 = R^2(\text{larger}) - R^2(\text{smaller})$, with

$$0 \leq \Delta R^2 \leq 1.$$

If all variables are scaled to have unit variance, ΔR^2 equals the covariance (with Y) of a group of predictors, adjusted for predictors already in the model. Examining that relationship leads to the study of partial correlations.

Correlation Formulae

Let $\boldsymbol{v}_j = \{v_{ij}\}$ indicate an $n \times 1$ vector of i.i.d.observations.

- Compute the mean as $\overline{v}_j = \mathbf{J}_n' \boldsymbol{v}_j / n$.
- Compute the sample variance as

$$s_j^2 = \sum_{i=1}^{n} \left(v_{ij} - \overline{v}_j\right)^2 / n = \boldsymbol{v}_j' \boldsymbol{v}_j / n - \overline{v}_j^2,$$

with $s_j^2 \left(\frac{n}{(n-1)}\right)$ an unbiased estimator of the population variance.

- Compute the sample covariance for two vectors j and j' as

$$c_{jj'} = \sum_{i=1}^{n} \left(v_{ij} - \overline{v}_j\right) \left(v_{ij'} - \overline{v}_{j'}\right) / n = \boldsymbol{v}_j' \boldsymbol{v}_{j'} / n - \overline{v}_j \overline{v}_{j'} \, ,$$

with $c_{jj'} \left(\frac{n}{(n-1)}\right)$ an unbiased estimator of the population covariance.

- Compute the sample correlation coefficient

$$r(v_{ij}, v_{ij'}) = \frac{c_{jj'}}{s_j s_{j'}}\,.$$

Partial Correlation

Partial correlations involve a predictor, X, a response, Y, and nuisance variables, $\mathbf{Z}$.

Partial correlations describe the strength of the linear relationship between two variables, Y and X, after controlling for the effects of other variables $\mathbf{Z}$. When multiple X and Z variables are involved, we call the partial correlation a *multiple partial correlation*. The *order* of a partial correlation depends on the number of variables for which we control: *first-order* partials control for $\mathbf{Z}_{(n\times 1)}$, *second-order* partials control for $\mathbf{Z}_{(n\times 2)}$, and p^{th}*-order* partials control for $\mathbf{Z}_{(n\times p)}$.

Example: Why do we want to consider a partial correlation? Consider computing the correlation between hair length and height in the general population.

Consider the following table of partial correlations for the ozone data.

Order	*Controlling Variables*	*Form of Correlation*	*Correlation Estimate*
0		$r_{PERS,HOME}$	0.53
0		$r_{PERS,OUT}$	0.40
0		$r_{PERS,TIME}$	0.19
1	HOME	$r_{PERS,OUT\|HOME}$	0.15
1	HOME	$r_{PERS,TIME\|HOME}$	0.23
1	OUT	$r_{PERS,HOME\|OUT}$	0.40
1	OUT	$r_{PERS,TIME\|OUT}$	0.18
1	TIME	$r_{PERS,HOME\|TIME}$	0.54
1	TIME	$r_{PERS,OUT\|TIME}$	0.39
2	HOME,OUT	$r_{PERS,TIME\|HOME,OUT}$	0.22
2	HOME,TIME	$r_{PERS,OUT\|HOME,TIME}$	0.13
2	OUT,TIME	$r_{PERS,HOME\|OUT,TIME}$	0.42

- Which variable has the greatest linear association with personal exposure?
- After that variable, what is the next most important variable?
- How do we tell whether the linear association is ”enough” to warrant inclusion in a statistical model? When do we stop adding variables?

At some point, we need to decide whether a particular partial correlation coefficient is significantly different from zero. Previously, we used F tests to determine whether adding a variable to a regression model was worthwhile, given that other variables were in the model. This type of test is also called a *partial F test*. This test is exactly equivalent to a test of significance for the corresponding partial correlation coefficient.

That is, a test of whether the population partial correlation coefficient $\rho_{PERS,TIME|HOME}$ is equal to 0 is exactly equivalent to a test of $H_0 : \beta_{TIME} = 0$ given that home exposure is already in the model.

Insight of Partial Correlation

Consider fitting two models, $\mathbf{y} = \mathbf{z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{x} = \mathbf{z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, to create $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ as well as predicted values $\widehat{\boldsymbol{y}} = \mathbf{z}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{x}} = \mathbf{z}\widehat{\boldsymbol{\gamma}}$.

Let $\widehat{\varepsilon}_{y_i} = y_i - \widehat{y}_i$ and $\widehat{\varepsilon}_{x_i} = x_i - \widehat{x}_i$.

Describe $r(\widehat{\varepsilon}_{y_i}, \widehat{\varepsilon}_{x_i})$, the correlation between these residuals, as the sample value of the (full) *partial correlation* between Y and X, controlling both for $\mathbf{Z}$. We use the notation $r_{YX|Z}$ for this partial correlation.

The partial correlation describes the simple linear correlation between Y and X after removing the linear effects of $\mathbf{Z}$.

Adjusting for Confounders

A very important application of partial F tests concerns controlling for confounders. Suppose that we have one main study variable of interest, $\mathbf{x}$, and q control variables (or confounders) $\mathbf{z}_1, \ldots, \mathbf{z}_q$. We can evaluate the effect of $\mathbf{x}$ on the outcome of interest, controlling for the potential confounders, $\mathbf{z}$, using the model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{z}_1 + \ldots + \beta_q \mathbf{z}_q + \beta_{q+1} \mathbf{x} + \boldsymbol{\varepsilon}.$$

We then test the hypothesis $H_0 : \beta_{q+1} = 0$, which is equivalent to testing whether the population partial correlation between $\mathbf{y}$ and $\mathbf{x}$, adjusting for the confounders $\mathbf{Z}$, is zero.

When we are interested in several variables in $\mathbf{X}$, we must determine which of these are important, and we may need to rank order them by their relative importance. We will discuss this strategy further when we talk about model selection.

Semi-Partial Correlation

We consider semi-partial correlations when we know that only one of X or Y is affected by the nuisance variables Z. For example, in a randomized clinical trial, the treatment variable X should be unaffected by nuisance factors Z (such as age or tumor type), although these nuisance factors may have an effect on the outcome.

Describe $r(\widehat{\varepsilon}_{y_i}, x_i) = r_{X(Y|Z)}$ as the semi-partial correlation between Y and X, controlling only Y for $\mathbf{Z}$.

Also define $r(y_i, \widehat{\varepsilon}_{x_i}) = r_{Y(X|Z)}$ as the semi-partial correlation between Y and X, controlling X for $\mathbf{Z}$.

Example: Semi-Partial Correlations in Ozone Data

Suppose we believe that while time spent outdoors may be related to personal ozone exposure, time spent outdoors and home ozone concentration are unlikely to be related (r=-0.007). In this case, when computing the correlation between personal ozone exposure and home ozone concentrations, we may wish to correct personal ozone exposure (but not home ozone concentrations) for the effect of time spent outdoors.

First, we will compute the partial correlation between personal ozone exposure and home ozone concentration, controlling both variables for the effect of time spent outdoors.

```
proc corr data=ozone nosimple /* noprob */;
var personal home;
partial time_out;
run;
*****************************************************************************
```

```
         Pearson Partial Correlation Coefficients, N = 64
               Prob > |r| under H0: Partial Rho=0

                                           personal          home

personal                                    1.00000       0.54175
Personal Ozone Exposure (ppb)                              <.0001

home                                        0.54175       1.00000
Home Indoor Ozone Concentration (ppb)        <.0001
```

So the partial correlation between personal ozone exposure and home ozone concentration, controlling both variables for the effect of time spent outdoors is 0.54.

Next, we control only personal ozone exposure (and not home ozone concentration) for the effect of time spent outdoors.

```
proc corr data=semipart nosimple noprob;
var home e_p;
run;
*****************************************************************************
                    Pearson Correlation Coefficients, N = 64

                                                 home           e_p
  home                                        1.00000       0.54173
  Home Indoor Ozone Concentration (ppb)
  e_p                                         0.54173       1.00000
```

The partial and semi-partial correlation coefficients are virtually identical. The simple correlation between home ozone concentration and personal ozone exposure is also similar at 0.53, suggesting that the linear influence of time spent outdoors on personal ozone exposure and/or on home ozone concentrations has little effect on the correlation between personal ozone exposure and home ozone exposure.

Choosing the Proper Correlation Coefficient

Considering the proper correlation coefficient to describe the relationship among variables Y and X along with nuisance (or confounding) variables Z depends on the nature of their relationship. The table below can be used to determine which correlation is appropriate.

Nuisance Relationship	*Preferred Correlation*
Neither X nor Y affected by Z	r_{xy}
Both X and Y affected by Z	$r_{yx\|z}$
Only X affected by Z	$r_{y(x\|z)}$
Only Y affected by Z	$r_{x(y\|z)}$

Computing Partial Correlations

Consider the following correlation matrix for three variables X, Y, and Z.

$$\boldsymbol{R} = \begin{array}{cc} & \begin{array}{ccc} X & Y & Z \end{array} \\ & \begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix} \begin{array}{c} X \\ Y \\ Z \end{array} \end{array}$$

Note that $r_{yx} = r_{xy}$.

A full partial correlation may be computed as:

$$r_{yx|z} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{(1 - r_{yz}^2)(1 - r_{xz}^2)}}.$$

In the same setting, semi-partials may be computed as:

$$r_{y(x|z)} = \frac{r_{yx} - r_{yz} r_{xz}}{\sqrt{(1)(1 - r_{xz}^2)}}$$

$$r_{x(y|z)} = \frac{r_{yx} - r_{yz} r_{xz}}{\sqrt{(1 - r_{yz}^2)(1)}} \, .$$

The (1) term represents the variance of a standardized variable.

Summary: Partial Correlation Coefficient

- The partial correlation $r_{YX|Z_1,\dots,Z_q}$ measures the strength of the linear relationship between X and Y while controlling for $\mathbf{Z}$.
- The square of the partial correlation $r_{YX|Z_1,\dots,Z_q}$ measures the proportion of the sum of squares for error in a model containing only $\mathbf{Z}$ that is accounted for by the addition of X to a regression model already containing $\mathbf{Z}$.
- The partial F statistic is used to test $H_0 : \rho_{YX|Z_1,\dots,Z_q} = 0$.
- The partial correlation $r_{YX|Z}$ can be defined as the correlation of the residuals of the straight-line regressions of Y on Z and of X on Z.
- Multiple partial correlations are tested using group-wise partial F tests.

Next: Assumption Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 7: “GLM Assumption Diagnostics”
- Weisberg, Chapter 9: “Regression Diagnostics”