Correlation

Dustin Long, PhD

Department of Biostatistics
University of Alabama at Birmingham

October 24, 2019

Outline:

- Correlation

- If $Y, X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $Y|X = x \sim N(\mu_{Y|X=x}, \sigma^2_{Y|X=x})$
- where $\mu_{Y|X=x} = \mu_y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_x)$ and $\sigma^2_{Y|X=x} = \sigma^2_Y(1 - \rho^2)$ and ρ is the correlation between Y and X
- If we let $\beta_1 = \rho(\sigma_y/\sigma_x)$ and $\beta_0 = \mu_y - \beta_1\mu_x$ we have fit a straight line assuming $\mu_{Y|X=x} = \beta_0 + \beta_1 x$

- Thus, we can think of a correlation matrix for multiple regression
- $$\begin{array}{c} \\ Y \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{ccccc} Y & X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ & 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ & & 1 & \rho_{23} & \rho_{24} \\ & & & 1 & \rho_{34} \\ & & & & 1 \end{bmatrix} \end{array}$$

```
library(MASS)
library(plotly)
library(dplyr)
mu = c(0,0)
rho = 0.2
sigma = matrix(c(5,rho,rho,1),ncol=2,nrow=2,byrow=T)
dat = mvrnorm(n=10000,mu,sigma)
bob = kde2d(x=dat[,1],y=dat[,2])
image(bob,col=topo.colors(100))
contour(bob,add=T)
plot_ly(x = bob$x, y = bob$y, z = bob$z) %>% add_surface()
```

```
proc kde data=example;
  bivar wgt age / plots=all;
run;
```

- Simple correlations can be calculated in different ways
- Pearson correlation is related to simple linear regression as seen above
- $r_p = \frac{\sum(Y_i - \bar{Y})(\sum(X_i - \bar{X})}{\sqrt{\sum(Y_i - \bar{Y})^2 \sum(Y_i - \bar{Y})^2}}$
- Recall from above $\beta_1 = \rho(\sigma_y / \sigma_x)$

- Corrected Sums of Squares: CSS, SS from model with intercept
- Uncorrected SS: SS from model without intercept
- Corrected R^2 uses CSS, Uncorrected R^2 uses USS
- Both forms estimate ρ^2
- $R^2 = \frac{SSM}{SST}$

- R^2 is the amount of variability in the outcome that is explained by the model
- Recall that $SSM + SSE = SST$, thus $R^2 = 1 - \frac{SSE}{SST}$
- This is deceptive as R^2 always increases with more covariates, thus adding any variable will increase R^2

- R^2 does not measure the magnitude of the slope just the strength of association
- It does not measure the appropriateness of the straight line model
- Test for R^2, $H_0: \rho = 0$ is equivalent to $H_0: \beta_1 = 0$ for simple linear regression
- Test statistic is: $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$

- You can calculate partial correlations, or partial R^2 values
- $R^2_{X_1}$, $R^2_{X_2|X_1}$, etc. can be calculated
- SAS PCORR1 in PROC REG gives variables-added-in-order partial R^2 and PCORR2 gives variables-added-last R^2
- These partial correlations measure the strength of the linear relationship between Y and X_1 while controlling for the other variables.
- Can be tested with partial F tests which are located with either the TYPE I or TYPE III SS table based on which correlation you are interested in
- $R^2_{X_j|\boldsymbol{X}_{-j}} = \frac{SS_{X_j|\boldsymbol{X}_{-j}}}{SS_{X_j|\boldsymbol{X}_{-j}}+SSE}$

- There exist multiple partial correlations, $R^2_{X_2,X_3|X_1}$, which is the strength of association between Y and X_1 and X_2 controlling for X_3
- Frequently used when covariates are used together, such as polynomials
- Use the multiple partial F test, which needs to be calculated by hand
- To test $H_0 : \rho_{X_2,X_3|X_1} = 0$, use $F = \frac{(SSM_{full} - SS_{X_1})/2}{MSE_{full}}$ which is an F distribution with 2 and $n - p - 1$ d.f. under H_0

- Spearman correlation is based on the ranked data but is still a linear correlation
- Calculate rank of both Y (R_y) and X (R_x)
- $r_s = \frac{\sum(R_{iy} - \bar{R}_y)(\sum(R_{ix} - \bar{R}_x)}{\sqrt{\sum(R_{iy} - \bar{R}_y)^2 \sum(R_{iy} - \bar{R}_y)^2}}$
- Can be considered the pearson correlation of the ranks of Y and X