HW 11 due on wed
HW 12 due next wed
OH today 11:30-12:30
wed 2:30-4

16: CORRELATION AND MATRIX APPROACH

Larsen & Marx 11.4; Rice 14.3; 14.4

Prof Amanda Luby

1 Covariance and Correlation

When we started linear regression, we began with the simplest scenario from a statistical standpoint – the case where each (x_i, y_i) are just constants with no probabilistic structure. When we moved into inference for this setting, we treated x_i as constant and Y_i as a random variable. We'll now move into the next layer of complexity: assuming both X_i and Y_i are random variables.

Covariance

Let X and Y be two random variables. The *covariance* of X and Y is given by:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Let X and Y be two random variables with finite variances. Then,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_X$
"independent"

If X and Y are independent:
$f_{X,Y} = f_X f_Y$
$E(X \cdot Y) = E(X)E(Y)$
$\text{Cov}(X,Y) = 0$

all of these are if and only if statements → can show independence w/ any property

The covariance of two random variables gives us a sense of how/what direction they are "related", but it also depends on the scale of the mean/variance for each RV. The *correlation coefficient* gives us a similar measure that is comparable across all RV's:

Correlation coefficient

Let X and Y be two random variables. The correlation coefficient of X and Y is given by:

"rho"

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \text{Cov}(X^*, Y^*)$$

ρ = population parameter

where $X^* = \frac{X - \mu_X}{\sigma_X}$ $\quad Y^* = \frac{Y - \mu_Y}{\sigma_Y}$

Note: $|\rho(X, Y)| \leq 1$:

$$0 \leq \text{Var}(X^* \pm Y^*) = 1 + 1 \pm 2\text{Cov}(X^*, Y^*)$$
$$= 2 \pm 2 \cdot \rho(X,Y)$$
$$= 2[1 \pm \rho(X,Y)]$$
$$0 \leq 1 \pm \rho(X,Y)$$
$$\pm\rho(X,Y) \leq 1$$

Example: Suppose the correlation coefficient between X and Y is unknown, but we have observed n measurements $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. How could we use this data to estimate ρ?

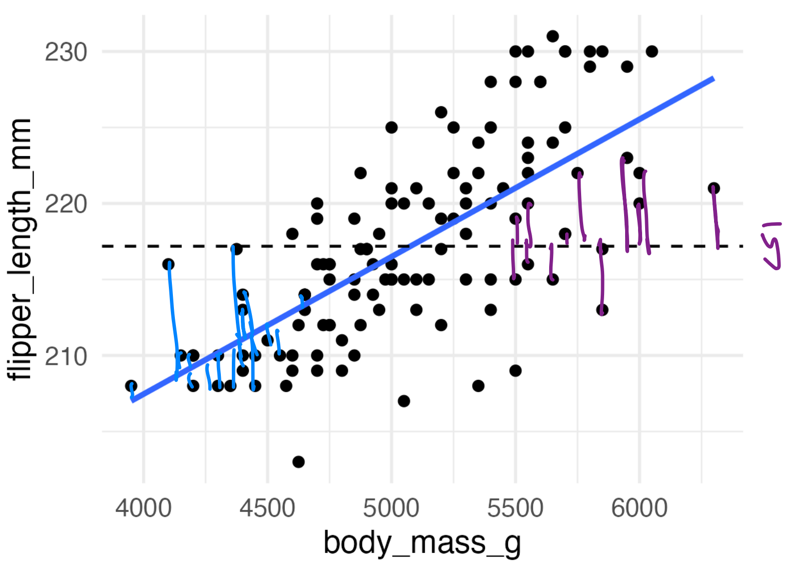
$$\rho(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$ } function of theoretical moments of X + Y. Idea: plug in the sample moments

$$r = \frac{\frac{1}{n}\sum X_iY_i - \frac{1}{n}\sum X_i \frac{1}{n}\sum Y_i}{\sqrt{\frac{1}{n}\sum(X_i-\bar{X})^2}\sqrt{\frac{1}{n}\sum(Y_i-\bar{Y})^2}} = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}}$$

* nice relationship between r + $\hat{\beta}_1$ (HW, yay!)

If we square the (estimated) correlation coefficient, we can simplify to: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$R^2 = r^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$



$\sum(y_i - \bar{y})^2$: total variability in Y_i's (SS_{total})

$\sum(y_i - \hat{y}_i)^2$: total variability "left over" after fitting regression model (SS_{resid})

Interpretation of R^2:

proportion of total variability in the Y_i's that is explained by the linear regression on X.

$$R^2 = \frac{\text{variability in } Y_i\text{'s that are explained by regression}}{\text{total variability in the } Y\text{'s}}$$

$r = .6 \rightarrow R^2 = .36 \rightarrow$ 36% of the variability in Y is explained by the regression on X (and therefore 64% is due to other factors) correlation

```r
cor(gentoo$body_mass_g, gentoo$flipper_length_mm, use = "complete.obs")
```

Deals w/ missing data

r =

```
[1] 0.7026665
```

y ~ x

```r
gentoo_lm = lm(flipper_length_mm ~ body_mass_g, data = gentoo)
summary(gentoo_lm)
```

```
Call:
lm(formula = flipper_length_mm ~ body_mass_g, data = gentoo)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0194  -2.7401   0.1781   2.9859   8.9806

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.713e+02  4.244e+00   40.36   <2e-16 ***
body_mass_g 9.039e-03  8.321e-04   10.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.633 on 121 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4937,    Adjusted R-squared:  0.4896
F-statistic:   118 on 1 and 121 DF,  p-value: < 2.2e-16
```

Not important until we have >1 x variable

2 Matrix Approach to Least Squares

2.1 Deriving the least squares solutions for 1 variable case

Define:

$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ $n \times 2$ "design matrix"

$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ $n \times 1$

$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ 2×1

$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$

$\hat{\mathbf{Y}} = \mathbf{X}\beta$

$n \times 2 \cdot 2 \times 1$

$n \times 1$

$$\begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix}$$

The least squares problem is to find $\hat{\beta}$ to minimize $L = \sum(y_i - (\beta_0 + \beta_1 x_i))^2$.

$= \|Y - X\beta\|^2$

$= \|Y - \hat{Y}\|^2$

Note: $\|\vec{u}\|^2 = \sum u_i^2$

$$Y - \hat{Y} = \begin{bmatrix} Y_1 - \hat{Y}_1 \\ Y_2 - \hat{Y}_2 \\ \vdots \\ Y_n - \hat{Y}_n \end{bmatrix}$$

In Notes14, we should that the least squares estimates satisfy:

$$\sum(y_i - (\beta_0 + \beta_1)x_i) = 0$$

$$\sum(y_i - (\beta_0 + \beta_1)x_i)x_i = 0$$

In matrix form, these equations are equivalent to:

$$X^T X\hat{\beta} = X^T Y$$

Which means that the least squares solution is (assuming $(X^T X)$ invertible)

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

2.2 Mean and Covariance of Vector-Valued RV's

Let $\mathbf{Y}$ be a random vector where $E(Y_i) = \mu_i$ and $Cov(Y_i, Y_j) = \sigma_{ij}$

Linear functions of random variables

Let $\mathbf{Z} = \mathbf{c} + \mathbf{A}\mathbf{Y}$. Then
$E(\mathbf{Z}) = \mathbf{c} + \mathbf{A}E(\mathbf{Y})$ and $\Sigma_Z = \mathbf{A}\Sigma_Y \mathbf{A}^T$

2.3 Mean and Covariance of Least Squares Estimates

Let $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where:

Mean and covariance of LS estimates (Matrix Form)

$E(\hat{\beta}) = \beta$
$\Sigma_{\hat{\beta}} = \sigma^2 (X^T X)^{-1}$