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,

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

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:

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

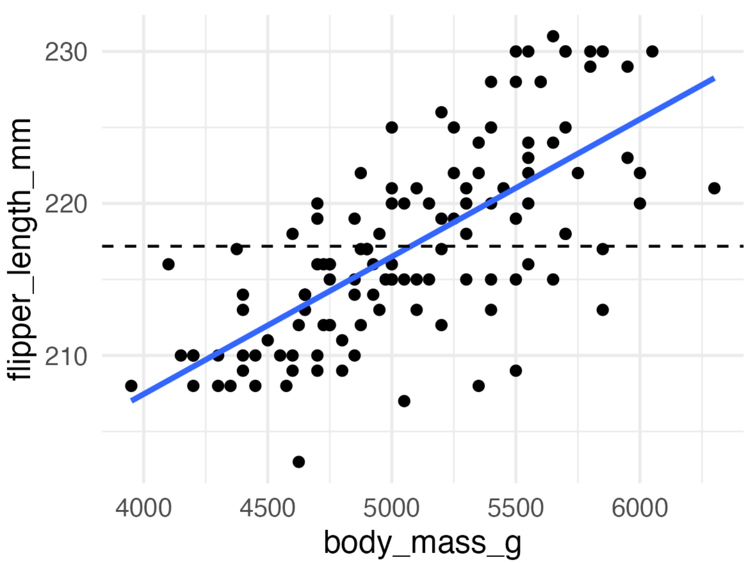
where

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

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?

If we square the (estimated) correlation coefficient, we can simplify to:

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



Interpretation of R^2:

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

```
[1] 0.7026665
```

```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
```

2 Matrix Approach to Least Squares

2.1 Deriving the least squares solutions for 1 variable case

Define:

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

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

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

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^TX) invertible)

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

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)

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