HW12 is posted, due next Wed

~~16~~ 17: MULTIPLE REGRESSION

Rice 14.4

Prof Amanda Luby

1 The Hat Matrix

$$\begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{bmatrix}$$

$$\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$$
$$= Y - X\hat{\beta}$$
$$= Y - X(X^TX)^{-1}X^TY$$
$$= Y - HY$$

$H = X(X^TX)^{-1}X^T$

$\hat{Y} = HY$

puts a "hat" on Y

$$H^T = X[(X^TX)^{-1}]^T X^T$$
$$= X(X^TX)^{-1}X^T$$
$$= H$$

H = constant, only a function of X's

Note: The "hat matrix" has some nice properties: $H = H^T = H^2$ and $(I - H) = (I - H)^T = (I - H)^2$.

2 Estimation of σ^2

$Y = X\beta + \epsilon$ $\quad \epsilon_i \sim N(0, \sigma^2)$

$var(Y_i) = \sigma^2$

In Notes 15, two of the properties that we worked with were:

SLR: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

(1) $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$

(2) $S^2 = \frac{n}{n-2}\hat{\sigma}^2$

$\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$

$S^2 = \frac{1}{n-2}\sum(Y_i - \hat{Y}_i)^2$

$E(S^2) = \sigma^2$ (unbiased)

MLR: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}$

In matrix notation, we can write:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

where p is # of columns in X (includes β_0)

$$\sum \hat{\epsilon}_i^2 = \sum(Y_i - \hat{Y}_i)^2 = ||Y - HY||^2$$
$$= ||(I-H)Y||^2$$
$$= [(I-H)Y]^T(I-H)Y$$
$$= Y^T(I-H)^T(I-H)Y$$
$$= Y^T(I-H)(I-H)Y$$
$$= \underset{\text{random vector}}{Y^T}\underbrace{(I-H)}_{\text{constant matrix}}\underset{\text{random vector}}{Y}$$

Note: $||z||^2 = z^Tz$

Then, using some nice properties for finding means of matrices (see Rice 14.4), we can show that $E(||Y - \hat{Y}||^2) = (n-p)\sigma^2$. This leads to the unbiased estimate for σ^2 for the multiple regression case:

$$s^2 = \frac{||Y - \hat{y}||^2}{n-p} = \frac{1}{n-p} \Sigma (y_i - \hat{y}_i)^2$$

Errors vs Residuals:

Population model : $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$

Fitted Model : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$

Population / True error : $\epsilon_i \sim N(0, \sigma^2)$

Residuals : $\hat{\epsilon}_i = Y_i - \hat{y}_i$

β_j's are pop. parameters (constant but unknown)

$\epsilon \sim N(0, \sigma^2)$

Covariance matrix of the residuals:

$$\hat{\epsilon} = Y - \hat{y} = (I - H)Y$$

$$\begin{aligned} \Sigma_{\hat{\epsilon}} &= (I-H)\Sigma_Y (I-H)^T \\ &= (I-H)\Sigma_\epsilon (I-H)^T \\ &= (I-H)\sigma^2 I (I-H)^T \\ &= \sigma^2 (I-H)(I-H)^T \\ &= \sigma^2 (I-H) \end{aligned}$$

$\leftarrow$ correlation between $\hat{\epsilon}_i, \hat{\epsilon}_j$ depends on $H = X(X^TX)^{-1}X^T$

In population model:

$$Y = X\beta + \epsilon$$

$$\Sigma_\epsilon = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I$$

Cross-covariance matrix

Let X be a random vector of length n with covariance matrix Σ_X. If $Y = AX$ and $Z = BX$, where $A = p \times n$ and $B = m \times n$, then the cross-covariance matrix of Y and Z is given by:

$$\Sigma_{YZ} = \underset{(p\times n)}{A}\ \underset{(n\times n)}{\Sigma_X}\ \underset{(n\times m)}{B^T} \qquad \Sigma_{YZ} = p \times m$$

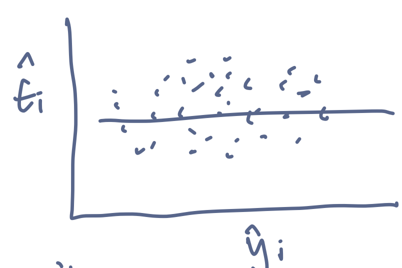
$cov(Y_1, Z_1)$ $\quad$ $cov(Y_1, Z_2)$

$$\Sigma_{YZ} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & & \vdots \\ \vdots & \vdots & & \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pm} \end{bmatrix}$$

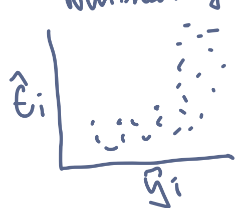
If the errors have covariance matrix $\sigma^2 I$, the residuals are uncorrelated with the predicted values

Proof: $\hat{\epsilon} = (I-H)Y \qquad \hat{y} = HY \qquad \Sigma_\epsilon = \sigma^2 I, \ \Sigma_Y = \sigma^2 I$

$$\begin{aligned} \Sigma_{\hat{\epsilon}\hat{y}} &= (I-H)\Sigma_Y H^T \\ &= (I-H)\sigma^2 I H^T \\ &= \sigma^2 [(I-H)H^T] \\ &= \sigma^2 [IH^T - HH^T] \\ &= \sigma^2 [H - H^2] \\ &= \sigma^2 [H - H] \\ &= \vec{0} \end{aligned}$$



Nonlinearity



Non-constant variance

Errors are correlated

3 CI's for β

Sampling distribution for $\hat{\beta}$

$$\hat\beta \sim MVN(\beta, \sigma^2 (X^T X)^{-1})$$

← fun fact about MVN: each component has a marginal normal distribution

Each $\hat\beta_j \sim N(\beta_j, \sigma^2 c_{jj})$ $\quad C = (X^T X)^{-1}$

In the simple LR case:

$$(X^T X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$\hat\beta_1 \sim N\left(\beta_1, \sigma^2 \cdot \frac{n}{n\sum x_i^2 - (\sum x_i)^2}\right)$$

$$N\left(\beta_1, \sigma^2 \cdot \frac{n}{n\sum (x_i - \bar{x})^2}\right)$$

→ Same as SLR derivation

$$S^2/\sigma^2 \sim \chi^2_{n-p}$$

$$U_j = \frac{\left(\frac{\hat\beta_j - \beta_j}{\sigma\sqrt{c_{ii}}}\right)}{\sqrt{S^2/\sigma^2(n-p)}} = \frac{\hat\beta_j - \beta_j}{S\sqrt{c_{jj}}} \sim t_{n-p}$$

Mon 12/11

- Homework 12 due on Wed
- Solutions by Fri morning
- Final Exam Sunday 9-12
- Project due Wed 20th at 11:59 pm
- OH this week
  - Mon 11:30-12:30
  - Wed 2:30-4
  - Fri 2-4 pm

4 CI's and PI's for predictions

Let $x^T = (1, x_1, ..., x_p)$ be a vector of predictors for a new observation Y.

Idea: we observe a new penguin w/ given body mass, bill length, etc. and want to draw inference about predicted flipper length.

$$\underset{\substack{\uparrow \\ \text{thing we want to make inference about}}}{\hat{Y}} = x^T \hat\beta = \underset{\substack{\uparrow \\ \text{new data}}}{x^T} \underbrace{(X^T X)^{-1} X^T Y}_{\text{old data}}$$

$\dim(\hat{Y}) = 1 \times 1$

$$\hat\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$E(\hat{Y}) = E(x^T \hat\beta) = x^T E(\hat\beta) = x^T \beta$$

$$\begin{aligned}
Var(\hat{Y}) &= V(x^T (X^T X)^{-1} X^T Y) \\
&= x^T (X^T X)^{-1} X^T V(Y) X \overbrace{(X^T X)^{-1}}^{((X^TX)^{-1})^T = ((X^TX)^T)^{-1} = (X^TX)^{-1}} x \\
&= x^T (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} x \\
&= \sigma^2 x^T (X^T X)^{-1} X^T X (X^T X)^{-1} x \\
&= \underbrace{\sigma^2 x^T (X^T X)^{-1} x}_{\text{variance we use for inference about } E(Y|X) \text{ inference for the line}}
\end{aligned}$$

For a prediction interval:

$$\begin{aligned}
V(Y - \hat{Y}) &= V(Y) + V(\hat{Y}) \\
&= \sigma^2 + \sigma^2 x^T (X^T X)^{-1} x \\
&= \underbrace{\sigma^2 [1 + x^T (X^T X)^{-1} x]}_{\text{variance for inference about new individual } Y_i\text{'s}}
\end{aligned}$$

Once again, we'll use $\frac{s^2}{\sigma^2} \sim \chi^2_{n-p}$

$$\frac{\dfrac{Y - \hat{Y}}{\sigma[1 + x^T(X^TX)^{-1}x]^{1/2}}}{\sqrt{\dfrac{s^2}{\sigma^2(n-p)}}} = t_{n-p}$$

$\rightarrow N(0,1)$

$\rightarrow \sqrt{\dfrac{\chi^2_\nu}{\nu}}$

5 Multiple R^2

Proportion of variability in Y variable explained by linear relationship w/ X variables

In the simple regression case, recall that

$R^2 = r^2 = \dfrac{\sum(y_i-\bar{y})^2 - \sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} = 1 - \dfrac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$

In simple linear regression, $R^2 = r^2$, where r is the sample correlation between X and Y. In the multiple regression case, we define $R = \text{Cor}(\hat{y}, y)$.

$\hat{y} = [\ ]$ $\quad y = [\ ]$

In multiple regression, whenever we add another predictor variable, R^2 *never gets worse*. The *Adjusted* R^2 is more often used in practice:

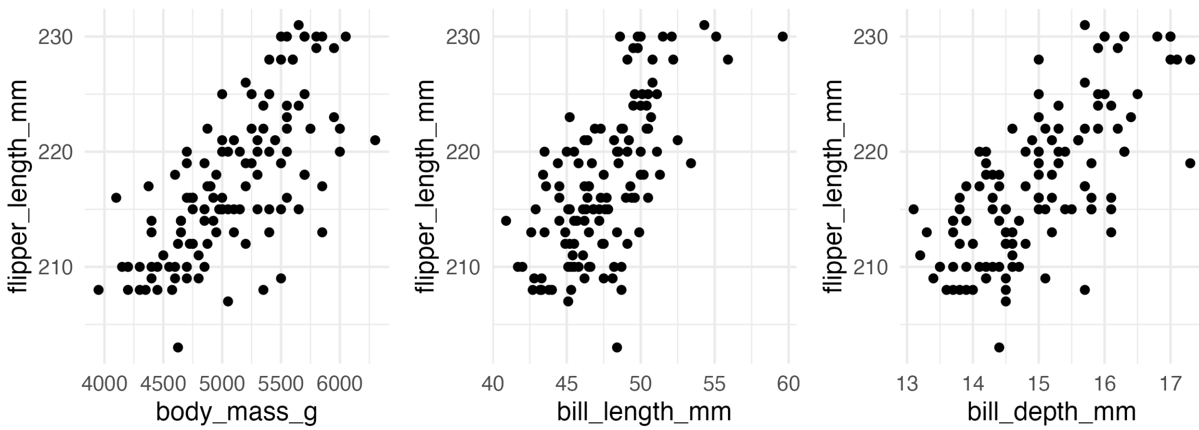
Adjusted $R^2 = 1 - \dfrac{\frac{1}{n-p}\sum(y_i-\hat{y}_i)^2}{\frac{1}{n-1}\sum(y_i-\bar{y})^2}$

worst case scenario: add another X
$\sum(y_i-\hat{y}_i)^2$ stays the same

as the number of predictors increase, what happens to the adjusted R^2?

P increases, adjusted R^2 decreases a small amount – $\frac{1}{n-p}$ is called a "penalty term"

6 Interpretation of β_i in Multiple Regression



$y = 171.3 + .0093x$ $\qquad y = 151 + 1.39x$ $\qquad y = 147.22 + 4.67x$

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

Coefficients:
(Intercept)  body_mass_g
  1.713e+02    9.039e-03


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

Coefficients:
   (Intercept)  bill_length_mm
       151.096           1.391
```

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

Coefficients:
  (Intercept)  bill_depth_mm
       147.22           4.67
```

Add all 3 predictors to the lm() model

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

Coefficients:
   (Intercept)     body_mass_g  bill_length_mm   bill_depth_mm
     139.99254         0.00382         0.52150         2.20463
```

summary(lm-mod)

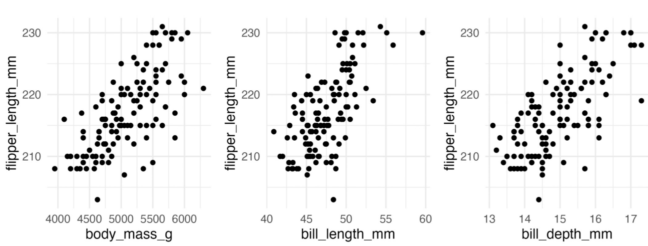
```
Call:
lm(formula = flipper_length_mm ~ body_mass_g + bill_length_mm +
    bill_depth_mm, data = gentoo)

Residuals:
    Min       1Q   Median       3Q      Max
-12.440   -2.492    0.023    2.829    8.322

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.400e+02  6.527e+00  21.448  < 2e-16 ***
body_mass_g    3.820e-03  1.153e-03   3.314 0.001217 **
bill_length_mm 5.215e-01  1.711e-01   3.047 0.002846 **
bill_depth_mm  2.205e+00  5.748e-01   3.836 0.000202 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.11 on 119 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6082,    Adjusted R-squared:  0.5983
F-statistic: 61.58 on 3 and 119 DF,  p-value: < 2.2e-16
```

On average, if body mass increases by 1, flipper length increases by .0038, if bill length & bill depth are held constant



$y = 171.3 + .0093x$ $\quad$ $y = 151 + 1.39x$ $\quad$ $y = 147.22 + 4.67x$

If X variables are correlated, $\hat{\beta}_j$ change SLR → MLR.

```
Call:
```