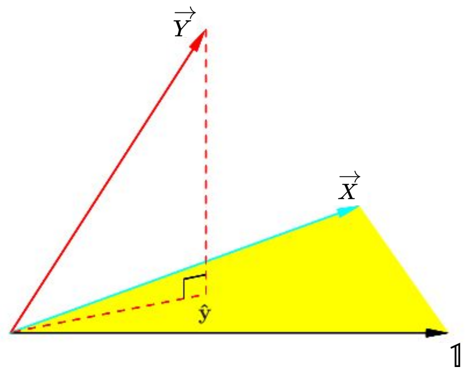
Sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

14.2 of Rice

08/04/2021

Berkeley
UNIVERSITY OF CALIFORNIA

In the previous lecture,



- Simple linear regression (SLR):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} N\left(0,\ \sigma^2\right).$$

  - Method of least squares;
  - Maximum likelihood estimation;
  - Projections onto the hyperplane $\mathrm{span}\left(\mathbb{1},\ \vec{X}\right)$.

$$\Rightarrow \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- Check model assumptions:
  - Linearity - Plotting Y_i vs X_i
  - Normality - QQ plot
  - Zero mean in error terms
  - Homoscedasticity
  - Independence

$\left.\begin{matrix} \text{Zero mean in error terms} \\ \text{Homoscedasticity} \\ \text{Independence} \end{matrix}\right\}$ Residual plot

Mean and variance

$$E(\bar{Y}_n) = \frac{E(Y_1)+\cdots+E(Y_n)}{n} = \frac{\sum_i^n \beta_0 + \beta_1 X_i}{n} = \beta_0 + \beta_1 \bar{X}_n$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0+\beta_1 X_1 \\ \vdots \\ \beta_0+\beta_1 X_n \end{pmatrix}, \sigma^2 I \right)$$

Proposition D. Under the SLR assumptions, $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ is bivariate Normal with:

$$E\left(\hat{\beta}_0\right) = \beta_0,\ \mathrm{var}\left(\hat{\beta}_0\right) = \frac{n^{-1}\sum_i X_i^2}{\sum_i \left(X_i - \bar{X}_n\right)^2}\sigma^2,$$

$$E\left(\hat{\beta}_1\right) = \beta_1,\ \mathrm{var}\left(\hat{\beta}_1\right) = \frac{1}{\sum_i \left(X_i - \bar{X}_n\right)^2}\sigma^2,$$

$$\mathrm{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{-\bar{X}_n}{\sum_i \left(X_i - \bar{X}_n\right)^2}\sigma^2.$$

Proof.

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum_i \left[\sum_i X_i^2 - X_i \sum_i X_i\right] Y_i}{n\sum X_i^2 - (\sum_i X_i)^2} = \sum_i b_i Y_i$$

$$b_i = \frac{\sum_i X_i^2 - X_i \sum_i X_i}{n\sum_i X_i^2 - (\sum_i X_i)^2}$$

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_i (X_i - \bar{X}_n)^2} = \frac{n\sum_i X_i Y_i - \sum X_i \sum Y_i}{n\sum_i X_i^2 - (\sum_i X_i)^2} = \frac{\sum_i \left[n X_i - \sum_i X_i\right] Y_i}{n\sum_i X_i^2 - (\sum_i X_i)^2} = \sum_i a_i Y_i$$

$$a_i = \frac{n X_i - \sum_i X_i}{n\sum_i X_i^2 - (\sum_i X_i)^2}$$

$$E(\hat{\beta}_1) = E\left(\sum_i a_i Y_i\right) = \sum_i a_i E(Y_i) = \sum_i \frac{n X_i - \sum_i X_i}{n\sum_i X_i^2 - (\sum_i X_i)^2} \times (\beta_0 + \beta_1 X_i)$$

$$= \beta_0 \underbrace{\sum_i \frac{n X_i - \sum_i X_i}{n\sum_i X_i^2 - (\sum_i X_i)^2}}_{=0} + \beta_1 \underbrace{\sum_i \frac{n X_i^2 - X_i \sum X_i}{n\sum_i X_i^2 - (\sum_i X_i)^2}}_{=1} = \beta_1$$

$$E(\hat{\beta}_0) = E(\bar{Y}_n + \hat{\beta}_1 \bar{X}_n)$$

$$= \beta_0 - \cancel{\beta_1 \bar{X}_n} + \cancel{\beta_1 \bar{X}_n}$$

$$= \beta_0.$$

Mean and variance

$$\hat\beta_1 = \frac{\sum_i (X_i-\bar X_n)(Y_i-\bar Y_n)}{\sum_i (X_i-\bar X_n)^2} = \frac{\sum (X_i-\bar X_n)Y_i - \bar Y_n \underbrace{\sum_i (X_i-\bar X_n)}_{=0}}{\sum_i (X_i-\bar X_n)^2} = \frac{\sum_i (X_i-\bar X_n)Y_i}{\sum_i (X_i-\bar X_n)^2}.$$

Proof *cont'd*. $\operatorname{var}(\hat\beta_1) = \dfrac{\sum_i (X_i-\bar X_n)^2 \operatorname{var}(Y_i)}{\left[\sum_i (X_i-\bar X_n)^2\right]^2} = \dfrac{\cancel{\sum_i (X_i-\bar X_n)^2}\, \sigma^2}{\left[\sum_i (X_i-\bar X_n)^2\right]^{\cancel{2}}} = \dfrac{1}{\sum_i (X_i-\bar X_n)^2}\sigma^2.$

$$\begin{aligned}
\operatorname{var}(\hat\beta_0) &= \operatorname{var}(\bar Y_n - \hat\beta_1 \bar X_n) \\
&= \operatorname{var}(\bar Y_n) + \bar X_n^2 \operatorname{var}(\hat\beta_1) - 2\bar X_n \operatorname{cov}(\bar Y_n, \hat\beta_1) \\
&= \frac{\sigma^2}{n} + \frac{\bar X_n}{\sum_i (X_i-\bar X_n)^2}\sigma^2 - 2\bar X_n \operatorname{cov}\left(\frac{\sum_i Y_i}{n}, \frac{\sum_i (X_i-\bar X_n)Y_i}{\sum_i (X_i-\bar X_n)^2}\right) \\
&= \frac{\sigma^2}{n} + \frac{\bar X_n^2}{\sum_i (X_i-\bar X_n)}\sigma^2 - 2\bar X_n \sum_{i=1}^n \frac{1}{n}\cdot\frac{X_i-\bar X_n}{\sum_i (X_i-\bar X_n)^2} \underbrace{\operatorname{var}(Y_i)} \\
&= \frac{\sigma^2}{n} + \frac{\bar X_n^2}{\sum_i (X_i-\bar X_n)^2}\sigma^2 - \frac{2\bar X_n \sigma^2}{n} \underbrace{\sum_{i=1}^n \frac{X_i-\bar X_n}{\sum_i (X_i-\bar X_n)^2}}_{=0} \\
&= \frac{\frac{1}{n}\sum_i X_i^2 - \bar X_n^2 + \bar X_n^2}{\sum_i (X_i-\bar X_n)^2}\sigma^2 = \frac{n^{-1}\sum_i X_i^2}{\sum_i (X_i-\bar X_n)^2}\sigma^2.
\end{aligned}$$

$$\mathrm{cov}(\hat\beta_0, \hat\beta_1) = \mathrm{cov}\left(\sum_i b_i Y_i, \sum_i a_i Y_i\right) = \sum_i a_i b_i \,\mathrm{var}(Y_i) = \sigma^2 \sum_i a_i b_i$$

$$= \frac{-\bar{X}_n}{\sum (X_i - \bar{X}_n)^2}\,\sigma^2 .$$

$Y \sim N(\mu, \Sigma)$

$AY \sim N(A\mu, A\Sigma A^T)$

$$\begin{pmatrix}\hat\beta_0 \\ \hat\beta_1\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}, \underbrace{\begin{pmatrix}\mathrm{var}(\hat\beta_0) & \mathrm{cov}(\hat\beta_0, \hat\beta_1) \\ \mathrm{cov}(\hat\beta_0, \hat\beta_1) & \mathrm{var}(\hat\beta_1)\end{pmatrix}}_{= \sigma^2 (X^T X)^{-1}}\right)$$

Recall that $\begin{pmatrix}\hat\beta_0 \\ \hat\beta_1\end{pmatrix} = (X^T X)^{-1} X^T \vec{Y}$ in which $X = \begin{pmatrix}1 & X_1 \\ \vdots & \vdots \\ 1 & X_n\end{pmatrix}$, $\vec{Y} = \begin{pmatrix}Y_1 \\ \vdots \\ Y_n\end{pmatrix}$

$$\mathrm{cov}\begin{pmatrix}\hat\beta_0 \\ \hat\beta_1\end{pmatrix} = (X^T X)^{-1} X^T (\sigma^2 I) \times (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} \cancel{X^T X (X^T X)^{-1}}$$

$$= \sigma^2 (X^T X)^{-1} .$$

Sampling distribution

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

Res $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$

$= Y_i - \underbrace{(1, X_i)}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$

$\Rightarrow \begin{pmatrix}\hat{\varepsilon}_1\\ \vdots \\ \hat{\varepsilon}_n\end{pmatrix} = \vec{Y} - X\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$

Lemma D. Under the SLR assumptions, $Y \sim N(\mu, \Sigma)$, $\mathrm{cov}(AY, BY) = A\Sigma B^T$.

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} \perp\!\!\!\perp \vec{Y} - X\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix},$$

and $\mathrm{RSS} = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \sim \sigma^2\chi^2_{n-2}$.

1. A matrix R is idempotent if $R^2 = R$.
2. If $Z \sim N(0, I)$ and R is symmetric and idempotent of rank r, then $Z^T R Z \sim \chi^2_r$.

Proof*. $\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = (X^TX)^{-1}X^T\vec{Y}$

$\vec{Y} - X\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \vec{Y} - X(X^TX)^{-1}X^T\vec{Y}$

$= \underbrace{(I - X(X^TX)^{-1}X^T)}_{H}\vec{Y}.$

$\mathrm{cov}\left[(X^TX)^{-1}X^T\vec{Y},\ (I-H)\vec{Y}\right] = \sigma^2(X^TX)^{-1}X^T(I-H)^T$

$= \sigma^2(X^TX)^{-1}X^T\left(I - X(X^TX)^{-1}X^T\right)$

$= \sigma^2\left[(X^TX)^{-1}X^T - (X^TX)^{-1}\cancel{X^TX}\cancel{(X^TX)^{-1}}X^T\right]$

$= 0.$

By definition,

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = (\hat{\varepsilon}_1, \cdots \hat{\varepsilon}_n) \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$$

$$= [(I-H)\vec{Y}]^T (I-H)\vec{Y} = \vec{Y}^T \underbrace{(I-H)^T(I-H)} \vec{Y} = \vec{Y}^T (I-H)\vec{Y}.$$

$$(I-H)^T(I-H) = (I-H)^2 = (I - X(X^TX)^{-1}X^T)(I - X(X^TX)^{-1}X^T)$$

$$= I - \underbrace{2X(X^TX)^{-1}X^T} + \underbrace{X(X^TX)^{-1}\cancel{X^TX(X^TX)^{-1}}X^T}$$

$$\Downarrow$$

$$H^2 = H$$

$$= I - H$$

$$\operatorname{rank}(I-H) \overset{\uparrow}{=} \operatorname{trace}(I-H) = n - \operatorname{trace}(H) = n - \operatorname{rank}(H) \rightarrow \underbrace{X(X^TX)X^T}$$

$$\operatorname{rank}(H) = \operatorname{rank}(X)$$

$$= n - \operatorname{rank}(X) = n-2.$$

$$RSS = \vec{Y}^T (I-H)\vec{Y} \sim \sigma^2 \chi^2_{n-2} \quad \text{and} \quad \underbrace{\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \perp\!\!\!\perp RSS.}$$

$$\downarrow$$

$$\vec{Y} \sim N(X\vec{\beta}, \sigma^2 I)$$

Sampling distribution

Theorem D. Under the SLR assumptions,

$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{se}\left(\hat{\beta}_k\right)} \sim t_{n-2},\ k = 0, 1.$$

95% exact confidence interval for β_k :

$$\hat{\beta}_k \pm t_{n-2}(\alpha/2) \cdot \mathrm{se}\left(\hat{\beta}_k\right)$$

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$$

$$\mathrm{se}\left(\hat{\beta}_0\right) = \sqrt{\frac{n^{-1}\sum_i X_i^2}{\sum_i \left(X_i - \bar{X}_n\right)^2}\hat{\sigma}^2}$$

$$\mathrm{se}\left(\hat{\beta}_1\right) = \sqrt{\frac{1}{\sum_i \left(X_i - \bar{X}_n\right)^2}\hat{\sigma}^2}$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{n^{-1}\sum_i x_i^2}{\sum_i (x_i - \bar{x}_n)^2}\sigma^2\right) \Rightarrow \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{n^{-1}\sum_i x_i^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\sigma^2}} \sim N(0,1)$$

$$\mathrm{RSS} \sim \sigma^2 \chi^2_{n-2}$$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2} \Rightarrow E(\hat{\sigma}^2) = \sigma^2.$$

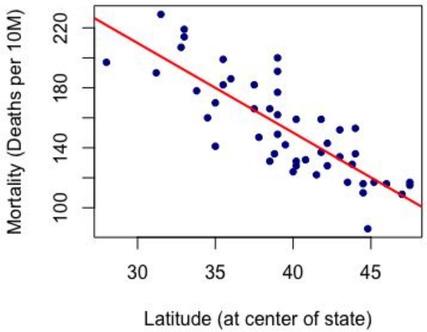
$$\frac{\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{n^{-1}\sum_i x_i^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\cancel{\sigma^2}}}}{\sqrt{\underbrace{\mathrm{RSS}/\cancel{\sigma^2}(n-2)}_{=\hat{\sigma}^2}}} \sim t_{n-2}$$

$$\Rightarrow \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{n^{-1}\sum_i x_i^2}{\sum_i (x_i - \bar{x}_n)^2}\hat{\sigma}^2}} \sim t_{n-2}.$$

Similarly for $\hat{\beta}_1$.

CIs of β_0 and β_1

Example 2 *cont'd*. During the 50s, data were collected to examine the relationship between the mortality rate due to skin cancer (number of deaths per 10 million people) and the latitude at the center of each of 48 states in the United States (Alaska and Hawaii were not yet states. And, Washington, D.C. was included in the data set even though it is not technically a state.)




```
> beta1 <- cov(Mort, Lat)/var(Lat)
> beta0 <- mean(Mort)-beta1*mean(Lat)
> c('beta0' = beta0, 'beta1' = beta1)
      beta0       beta1
389.189351  -5.977636
```

```
> n <- length(Mort)
> RSS <- sum(Res^2)
> s2_hat <- RSS/(n-2)
> se_beta0 <- sqrt(mean(Lat^2)*s2_hat/{(n-1)*var(Lat)})
> se_beta1 <- sqrt(s2_hat/{(n-1)*var(Lat)})
>
> alpha <- 0.05
> t_alpha <- qt(alpha/2, n-2, lower.tail = FALSE)
> c(beta0 - t_alpha*se_beta0, beta0 + t_alpha*se_beta0)
[1] 341.2852 437.0936
> c(beta1 - t_alpha*se_beta1, beta1 + t_alpha*se_beta1)
[1] -7.181404 -4.773867
```

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$$

$$\mathrm{se}\left(\hat{\beta}_0\right) = \sqrt{\frac{n^{-1}\sum_i X_i^2}{\sum_i \left(X_i - \bar{X}_n\right)^2}\hat{\sigma}^2}$$

$$\mathrm{se}\left(\hat{\beta}_1\right) = \sqrt{\frac{1}{\sum_i \left(X_i - \bar{X}_n\right)^2}\hat{\sigma}^2}$$

```
> confint(fit, 1, level = 0.95)
               2.5 %   97.5 %
(Intercept) 341.2852 437.0936
> confint(fit, 'Lat', level = 0.95)
        2.5 %    97.5 %
Lat -7.181404 -4.773867
```

SE of the regression line

$\beta_0 + \beta_1 x$

$\begin{pmatrix} \hat\beta_0 \\ \hat\beta_1 \end{pmatrix} \perp\!\!\!\perp \text{RSS}$

Theorem E. Under the SLR assumptions, we can use $(\hat\beta_0, \hat\beta_1)$ to predict at any $X = x$:

$$\hat y = \hat\beta_0 + \hat\beta_1 x,$$

Then $E(\hat y) = \beta_0 + \beta_1 x$, $\operatorname{var}(\hat y) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar X_n)^2}{\sum_{i=1}^n (X_i - \bar X_n)^2}\right]$ and $\hat y$ is independent of RSS.

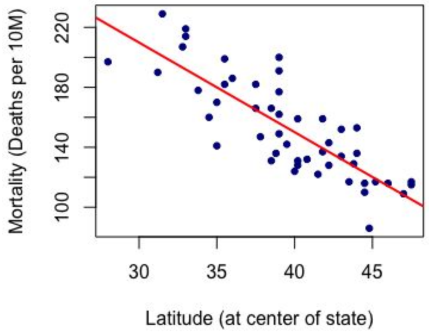
$\Rightarrow$ 95% exact confidence interval for $\beta_0 + \beta_1 x$:

$$\hat y \pm t_{n-2}(\alpha/2) \cdot \operatorname{se}(\hat y)$$

$$\text{RSS} = \sum_{i=1}^n \hat\epsilon_i^2 = \sum_{i=1}^n \left(Y_i - \hat\beta_0 - \hat\beta_1 X_i\right)^2$$

$$\hat\sigma^2 = \frac{\text{RSS}}{n-2}$$

$$\operatorname{se}(\hat y) = \sqrt{\left[\frac{1}{n} + \frac{(x - \bar X_n)}{\sum_i (X_i - \bar X_n)^2}\right]\hat\sigma^2}$$



$\operatorname{var}(\hat y) = \operatorname{var}(\hat\beta_0 + \hat\beta_1 x)$

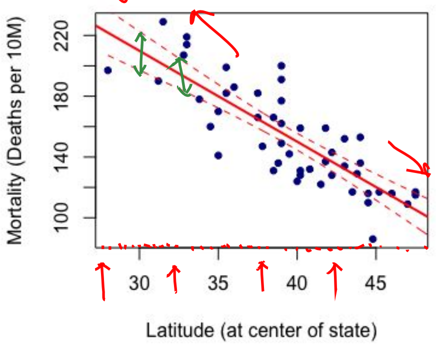
$= \operatorname{var}(\hat\beta_0) + \operatorname{var}(\hat\beta_1) x^2 + 2x \operatorname{cov}(\hat\beta_0, \hat\beta_1)$

$= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar X_n)^2}{\sum_{i=1}^n (X_i - \bar X_n)^2}\right]$

CIs of the population mean

Example 2 *cont'd*. During the 50s, data were collected to examine the relationship between the mortality rate due to skin cancer (number of deaths per 10 million people) and the latitude at the center of each of 48 states in the United States (Alaska and Hawaii were not yet states. And, Washington, D.C. was included in the data set even though it is not technically a state.)

confidence band




```
x_vec <- seq(28, 49, by=0.1)
fit_vec <- beta0 + beta1*x_vec
denom <- {(n-1)*var(Lat)}
se_vec <- sqrt({1/n + (x_vec-mean(Lat))^2/denom}*s2_hat)

lower_vec <- fit_vec - t_alpha*se_vec
upper_vec <- fit_vec + t_alpha*se_vec
lines(x_vec, lower_vec, lty=2, col='red')
lines(x_vec, upper_vec, lty=2, col='red')
```

```
new_dat <- data.frame(Lat=x_vec)
CIs <- predict(fit, newdata = new_dat,
               interval = 'confidence', level = 0.95)
lines(x_vec, CIs[,'lwr'], lty=2, col='red')
lines(x_vec, CIs[,'upr'], lty=2, col='red')
```

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$$

$$\mathrm{se}(\hat{y}) = \sqrt{\left[\frac{1}{n} + \frac{(x - \bar{X}_n)}{\sum_i \left(X_i - \bar{X}_n\right)^2}\right]\hat{\sigma}^2}$$

Prediction interval of the response

μ , σ^2 , ...

$\frac{y-\hat{y}}{\sqrt{\sigma^2+\mathrm{var}(\hat{y})}} \sim N(0,1)$

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

Theorem F. Under the SLR assumptions, the new observation at $X = x$ is a random variable:

$$y = \beta_0 + \beta_1 x + \epsilon,\ \epsilon \sim N(0, \sigma^2).$$

Then $E(y-\hat{y}) = 0$, $\mathrm{var}(y-\hat{y}) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{X}_n)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\right]$ and y is independent of $\hat{y}$ and RSS.

95% ~~exact confidence interval for $\beta_0 + \beta_1 x$~~: prediction interval for the response y:

$$\hat{y} \pm t_{n-2}(\alpha/2) \cdot \mathrm{se}(y-\hat{y})$$

$$\mathrm{RSS} = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$$

$$\mathrm{se}(y-\hat{y}) = \sqrt{\left[1 + \frac{1}{n} + \frac{(x-\bar{X}_n)}{\sum_i (X_i - \bar{X}_n)^2}\right]\hat{\sigma}^2}$$

More uncertainty

$y \sim N(\beta_0 + \beta_1 x,\ \sigma^2)$

$\hat{y} \sim N(\beta_0 + \beta_1 x,\ \mathrm{var}(\hat{y}))$

$y \perp\!\!\!\perp \hat{y}$

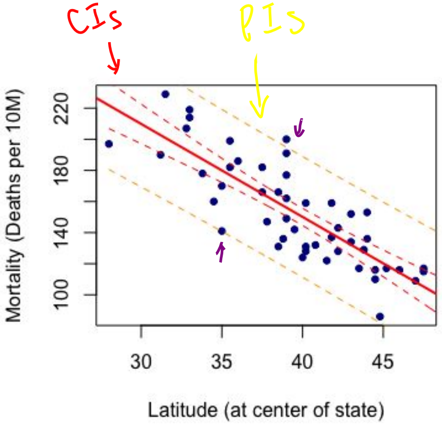
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$(Y_1, \ldots, Y_n \quad X_1, \ldots, X_n)$

indep $\Rightarrow$ $y - \hat{y} \sim N(0,\ \sigma^2 + \mathrm{var}(\hat{y}))$

Prediction interval of the response

Example 2 *cont'd*. During the 50s, data were collected to examine the relationship between the mortality rate due to skin cancer (number of deaths per 10 million people) and the latitude at the center of each of 48 states in the United States (Alaska and Hawaii were not yet states. And, Washington, D.C. was included in the data set even though it is not technically a state.)




```r
x_vec <- seq(28, 49, by=0.1)
fit_vec <- beta0 + beta1*x_vec
denom <- {(n-1)*var(Lat)}
se_pred_vec <- sqrt({1+1/n + (x_vec-mean(Lat))^2/denom}*s2_hat)

lower_pred_vec <- fit_vec - t_alpha*se_pred_vec
upper_pred_vec <- fit_vec + t_alpha*se_pred_vec
lines(x_vec, lower_pred_vec, lty=2, col='orange')
lines(x_vec, upper_pred_vec, lty=2, col='orange')
```

```r
new_dat <- data.frame(Lat=x_vec)
CIs_pred <- predict(fit, newdata = new_dat,
                    interval = 'prediction', level = 0.95)
lines(x_vec, CIs_pred[,'lwr'], lty=2, col='orange')
lines(x_vec, CIs_pred[,'upr'], lty=2, col='orange')
```

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$$

$$\mathrm{se}(y - \hat{y}) = \sqrt{\left[1 + \frac{1}{n} + \frac{(x - \bar{X}_n)}{\sum_i (X_i - \bar{X}_n)^2}\right] \hat{\sigma}^2}$$

Multiple linear regression

14.3 of Rice

08/04/2021

Multiple linear regression (MLR)

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \rightarrow \vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Response variable	Predictor variable 1	Predictor variable 2		Predictor variable p
Y_1	X_{11}	X_{12}	$\cdots$	X_{1p}
Y_2	X_{21}	X_{22}	$\cdots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Y_n	X_{n1}	X_{n2}	$\cdots$	X_{np}

Model assumption $(i = 1, \cdots, n)$:

$$Y_i = \underbrace{\beta_0}_{\text{common mean level}} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \underbrace{\epsilon_i}_{\overset{\text{iid}}{\sim} N(0,\sigma^2)} .$$

1. Linearity - Plotting $\vec{Y}$ vs $\vec{X}_j$
2. Normality - QQ plot
3. Zero mean in error terms
4. Homoscedasticity
5. Independence

(3–5) } Residual plot

Method of least squares

$f(\vec\beta)$ $\quad \frac{\partial f}{\partial \vec\beta} = \begin{pmatrix} \frac{\partial f}{\partial \beta_1} \\ \vdots \\ \frac{\partial f}{\partial \beta_n} \end{pmatrix}$

$f(\vec\beta) = \vec a^T \vec\beta, \quad \frac{\partial f}{\partial \vec\beta} = \vec a$

$f(\vec\beta) = \vec\beta^T A \vec\beta, \quad \frac{\partial f}{\partial \vec\beta} = (A^T + A)\vec\beta$

Proposition A'. Under the SLR assumptions, find estimators for $\vec\beta$ such that they minimize the sum of squared vertical deviations:

$$S\left(\vec\beta\right) = \sum_{i=1}^{n} (Y_i - \underbrace{\beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip}}_{})^2.$$

$X^T X$ is invertible as long as X is full rank, i.e. $\text{rank}(X) = p+1$.

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$$

$\leftarrow$ ith row vector $\vec X^i$

$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$

$= (1, X_{i1}, \cdots, X_{ip}) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \epsilon_i$

$\vec Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \mathbf{X}\vec\beta + \vec\epsilon$.

$S(\vec\beta) = \sum_{i=1}^{n} \left[Y_i - \vec X^i \vec\beta \right]^2$

$= (Y_1 - \vec X^1 \vec\beta, \cdots, Y_n - \vec X^n \vec\beta) \begin{pmatrix} Y_1 - \vec X^1 \vec\beta \\ \vdots \\ Y_n - \vec X^n \vec\beta \end{pmatrix}$

$\begin{pmatrix} Y_1 - \vec X^1 \vec\beta \\ \vdots \\ Y_n - \vec X^n \vec\beta \end{pmatrix} = \vec Y - \mathbf{X}\vec\beta$

$= (\vec Y - \mathbf{X}\vec\beta)^T (\vec Y - \mathbf{X}\vec\beta)$

$= \vec Y^T \vec Y - 2\vec Y^T \mathbf{X}\vec\beta + \vec\beta^T \mathbf{X}^T \mathbf{X}\vec\beta$

$\frac{\partial S}{\partial \vec\beta} = -2X^T\vec Y + (X^T X + X^T X)\vec\beta$

$= -2X^T\vec Y + 2X^T X\vec\beta = 0$

$\iff X^T X\vec\beta = X^T \vec Y$

$\iff \boxed{\vec\beta = (X^T X)^{-1} X^T \vec Y}$

Maximum likelihood estimation

Proposition B'. Under the SLR assumptions, calculate $\sup_{\Theta} L\left(\vec{\beta}, \sigma^2\right)$ and find MLEs of $\vec{\beta}$ and σ^2.

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$$

$$L(\vec{\beta}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_p X_{ip})^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \cdots - \beta_p X_{ip})^2}{2\sigma^2}}$$

$$\ell(\vec{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n}(Y_i - \beta_0 - \cdots - \beta_p X_{ip})^2}_{S(\vec{\beta})}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \cdots \hat{\beta}_p X_{ip})^2$$

Geometric approach

Proposition C'. We generalize to any $n \geq p+1$:

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}.$$

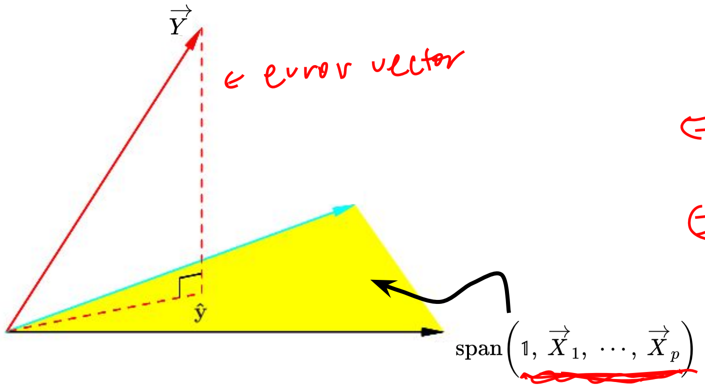
Find the best values of $\vec{\beta}$.

To ensure orthogonality,

$$X^T\underbrace{(\vec{Y} - X\vec{\beta})}_{} = 0$$

$$\iff X^T\vec{Y} - X^TX\vec{\beta} = 0$$

$$\iff \vec{\beta} = (X^TX)^{-1}X^T\vec{Y}.$$



dim = p+1

Calculate in R

Example 3. Interested in answering whether person's brain size and body size predictive of his or her intelligence, some researchers (Willerman, *et al*, 1991) collected the following data on a sample of n = 38 college students:

- Response (y): Performance IQ scores (PIQ) from the revised Wechsler Adult Intelligence Scale.Potential
- x_1: Brain size based on the count obtained from MRI scans (given as count/10,000).
- x_2: Height in inches.
- x_3: Weight in pounds

```
> dat <- read.table('~/Downloads/iqsize.txt', header = TRUE)
> head(dat)
  PIQ  Brain Height Weight
1 124  81.69   64.5    118
2 150 103.84   73.3    143
3 128  96.54   68.8    172
4 134  95.15   65.0    147
5 110  92.88   69.0    146
6 131  99.13   64.5    138
```

$$\vec{\beta}_{\text{hat}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\vec{Y}$$

```
> X <- cbind(1, dat$Brain, dat$Height, dat$Weight)
> y <- dat$PIQ
> solve(t(X)%*%X)%*%t(X)%*%y
              [,1]
[1,]  1.113536e+02
[2,]  2.060367e+00
[3,] -2.731929e+00
[4,]  5.599371e-04
```

```
> fit <- lm(PIQ ~ Brain + Height + Weight, data=dat)
> fit$coefficients
  (Intercept)         Brain        Height        Weight
 1.113536e+02  2.060367e+00 -2.731929e+00  5.599371e-04
```

Diagnostics

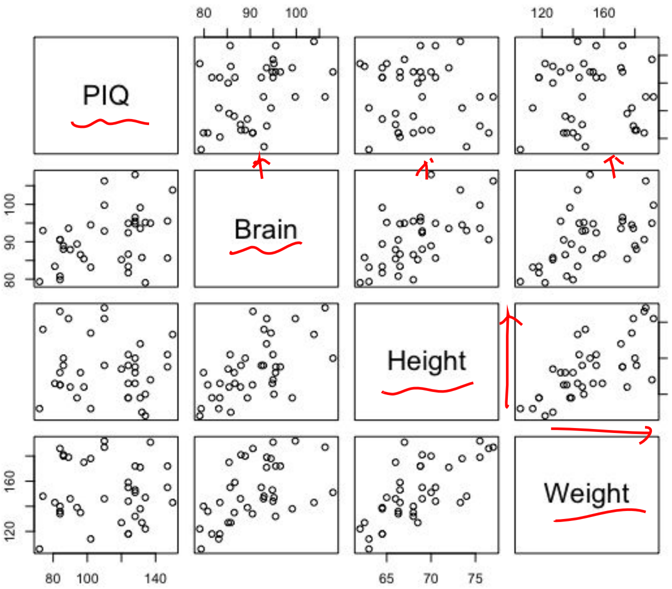
Model assumption $(i = 1, \cdots, n)$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \underbrace{\epsilon_i}_{\overset{\text{iid}}{\sim} N(0,\sigma^2)} . \quad \Rightarrow \quad \text{Residuals } \hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip}$$

1. Linearity - Plotting Y_i vs X_i
2. Normality - QQ plot
3. Zero mean in error terms
4. Homoscedasticity
5. Independence

(items 3–5) } Residual plot

```
> pairs(dat)
```



Height

$\approx a^* \text{weight}$

Diagnostics

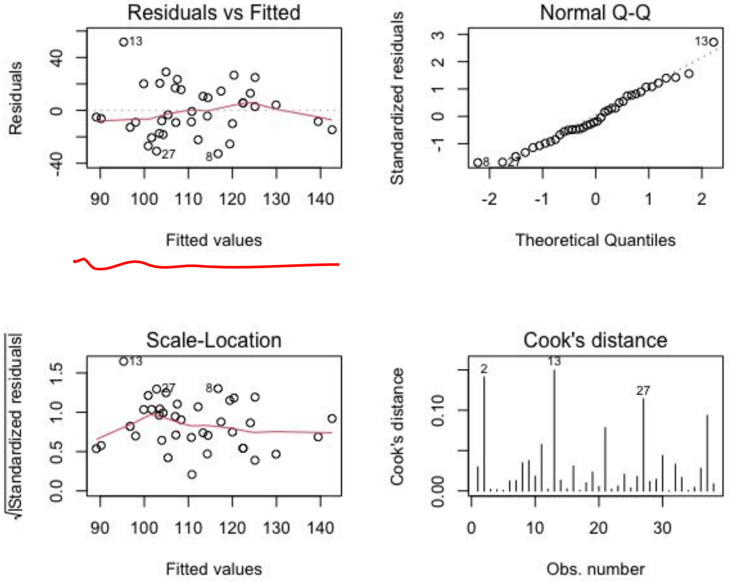
Model assumption $(i = 1, \cdots, n)$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \underbrace{\epsilon_i}_{\overset{\text{iid}}{\sim} N(0,\sigma^2)} . \quad \Rightarrow \quad \text{Residuals } \hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip}$$

1. Linearity - Plotting Y_i vs X_i
2. Normality - QQ plot
3. Zero mean in error terms
4. Homoscedasticity
5. Independence

} Residual plot

```r
> par(mfrow=c(2,2))
> plot(fit, which = 1)
> plot(fit, which = 2)
> plot(fit, which = 3)
> plot(fit, which = 4)
> par(mfrow=c(1,1))
```



Collinearity: redundant predictors

$\vec{X_1} = a\vec{X_2}$

What if $\vec{X}_1 \approx a\vec{X}_2$?

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$$

$\sum_i (Y_i - \bar{Y}_n)^2 = \sum_i (Y_i - \hat{\beta}_0 - \cdots \hat{\beta}_p X_{ip})^2$

$\downarrow SS_{tot}$ $\qquad \downarrow SS_R + SS_{pred}$

$SS_R = RSS$

$$R^2 = 1 - \frac{\text{RSS}}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} . = \frac{SS_{pred}}{SS_{tot}}$$

$\Downarrow$

$$R^2 = \text{Percent of variation in } \vec{Y} \text{ explained by } \vec{X}_1, \ldots, \vec{X}_p$$

```
> n <- nrow(dat)
> RSS <- sum(fit$residuals^2)
> R_sq <- 1-RSS/{(n-1)*var(y)}
> R_sq
[1] 0.2949392
```

```
> summary(fit)$r.squared
[1] 0.2949392
```

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$

$= \beta_0 + \beta_1 a X_{i2} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$

$= \beta_0 + (\beta_1 a + \beta_2) X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$

Collinearity: redundant predictors

$R_j^2 \uparrow$ $\vec{X}_j$ can be explain

$\vec{X}_j \sim \vec{X}_1, \vec{X}_2, \cdots \vec{X}_{j-1}, \vec{X}_{j+1}, \vec{X}_p$

What if $\vec{X}_1 \approx a\vec{X}_2$?

$$R^2 = 1 - \frac{\text{RSS}}{\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)^2}.$$

$$\Downarrow$$

$$R^2 = \text{Percent of variation in } \vec{Y} \text{ explained by } \vec{X}_1, \ldots, \vec{X}_p$$

$$R_j^2 = \text{Percent of variation in } \vec{X}_j \text{ explained by all other predictors}$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$$

fit ← lm(Res ~ Pred1 + Pred2 + ⋯ + Predp, data = ⋯)

```
> library(car)
Loading required package: carData
> vif(fit)
   Brain   Height   Weight
1.578524 2.276641 2.021541
```

- If R_j^2 is large (>0.9), then estimation of $\vec{\beta}$ will be difficult;
- The variance inflation factors (VIFs) are the most common diagnostics:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

Rule of thumb: VIF_j >10 indicates strong collinearity in $\vec{X}_j$.

Tomorrow ...

- Sampling distribution of $\vec{\beta}_{\text{hat}}$
- Bayesian statistics