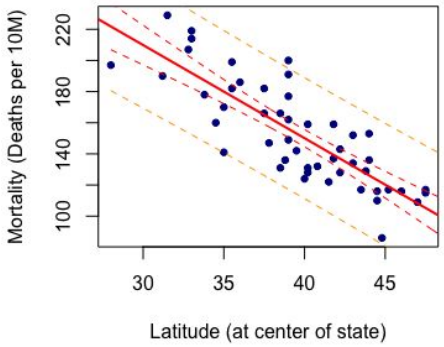
Sampling distributions of $\overrightarrow{\beta}_{\text{hat}}$

14.4 of Rice

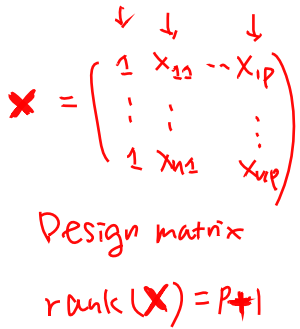
08/05/2021

Berkeley
UNIVERSITY OF CALIFORNIA

In the previous lecture,



- Simple linear regression (SLR):
  - $\left(\hat{\beta}_0, \hat{\beta}_1\right)^T$ is a bivariate Normal random variable;
  - $\left(\hat{\beta}_0, \hat{\beta}_1\right)^T$ is independent of residuals;
  - $\mathrm{RSS} = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \sim \chi^2_{n-2}.$
- Confidence intervals & bands
  - CIs of β_0 and β_1:
    $$\hat{\beta}_k \pm t_{n-2}(\alpha/2) \cdot \mathrm{se}\left(\hat{\beta}_k\right)$$
  - CIs of the regression line $\beta_0+\beta_1 x$:
    $$\hat{y} \pm t_{n-2}(\alpha/2) \cdot \mathrm{se}(\hat{y})$$
  - PIs of the response:
    $$\hat{y} \pm t_{n-2}(\alpha/2) \cdot \mathrm{se}(y - \hat{y})$$
- Multiple linear regression (MLR):
  - $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i\,,\ \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right).$
  - $\vec{\beta}_{\text{hat}} = (\mathbf{X^T X})^{-1} \mathbf{X^T} \vec{Y}$ ←
  - To avoid collinearity, the design matrix $\mathbf{X}$ must have rank $p+1$. ⇒ R^2 and VIF_i



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

$\text{Var}(A\vec{Y}) = A\Sigma A^T$

$\uparrow$

$\text{Var}(\vec{Y})$

Proposition D'. Under the MLR assumptions, $\vec{\beta}_{\text{hat}}$ is multivariate Normal with:

$$E\left(\vec{\beta}_{\text{hat}}\right) = \vec{\beta}, \text{var}\left(\vec{\beta}_{\text{hat}}\right) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

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

$\vec{\beta}_{hat} = (X^TX)^{-1}X^T\vec{Y} \sim N(\vec{\beta}, \sigma^2(X^TX)^{-1})$.

$\uparrow$

$E\,\vec{\beta}_{hat} = (X^TX)^{-1}X^TX\vec{\beta} = \vec{\beta}$.

$\text{var}(\vec{\beta}_{hat}) = (X^TX)^{-1}X^T\sigma^2 I \times (X^TX)^{-1}$

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

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

Sampling distribution of RSS

Projection matrix of span($\mathbf{X}$): $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

$\vec{\beta}_{hat} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$

Lemma D'. Under the MLR assumptions,

$\vec{\beta}_{\text{hat}} \perp\!\!\!\perp \vec{Y} - \mathbf{X}\vec{\beta}_{\text{hat}}$,

and $\text{RSS} = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \cdots - \hat{\beta}_p X_{ip}\right)^2 \sim \chi^2_{n-p-1}$.

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*.

$\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \cdots - \hat{\beta}_p X_{ip}$

$= Y_i - (1, X_{i1}, \cdots X_{ip})\,\vec{\beta}_{hat}$

$\begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} = \vec{Y} - \mathbf{X}\vec{\beta}_{hat}$.

$= \vec{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$

$= (I - H)\vec{Y}$.

$\text{cov}(\vec{\beta}_{hat}, \vec{Y} - \mathbf{X}\vec{\beta}_{hat})$

$= \text{cov}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}, (I-H)\vec{Y})$

$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \sigma^2 I (I-H)$

$= \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(I-H) = 0$.

$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \vec{Y}^T(I-H)(I-H)\vec{Y}$

$= \vec{Y}^T(I-H)\vec{Y} \sim \chi^2_{n-p-1}$.

$\text{rank}(I-H) = n - \text{rank}(X) = n - p - 1$.

CIs of $\vec{\beta}$

$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \vec{\beta}_{MLE} \sim N(\vec{\beta}, \sigma^2 (X^T X)^{-1})$

$\hat{\beta}_k \sim N(\beta_k, \sigma^2 (X^T X)^{-1}_{kk})$

Theorem D'. Under the MLR assumptions,

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

$\Rightarrow$ 95% exact confidence interval for β_k :

$$\hat{\beta}_k \pm t_{n-p-1}(\alpha/2) \cdot \text{se}\left(\hat{\beta}_k\right)$$

$$\text{RSS} = \vec{Y}^T (I - H) \vec{Y}$$

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}$$

$$\text{se}\left(\hat{\beta}_k\right) = \sqrt{\hat{\sigma}^2 \left[\left(\mathbf{X}^T \mathbf{X}\right)^{-1}\right]_{kk}}$$

CIs of $\vec{\beta}$

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

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

$\vec{Y} - \mathbf{X}\vec{\beta}_{\text{hat}}$

```
coefs <- solve(t(X)%*%X)%*%t(X)%*%y
Res <- y-X%*%coefs
RSS <- sum(Res^2)
s2_hat <- RSS/(n-ncol(X))

cov_diag_elements <- diag(solve(t(X)%*%X))
se_beta_vec <- sqrt(s2_hat*cov_diag_elements)
alpha = 0.05
t_alpha <- qt(0.05/2, n-ncol(X), lower.tail = FALSE)
cbind(coefs - t_alpha*se_beta_vec, coefs + t_alpha*se_beta_vec)
```

$\text{diag}(X^TX)^{-1}$

$$\text{RSS} = \vec{Y}^T(I-H)\vec{Y}$$
$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1}$$
$$\text{se}(\hat{\beta}_k) = \sqrt{\hat{\sigma}^2\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{kk}}$$

```
confint(fit, level = 0.95)
```

	2.5 %	97.5 %
(Intercept)	-16.6190567	239.3262733
Brain	0.9153051	3.2054285
Height	-5.2304287	-0.2334296
Weight	-0.3999266	0.4010465

fit ~ lm(Res ~ pred1 + pred2 + pred3, data = ..)

SE of the regression line

$E(\hat{y}) = E(\vec{x}^T \vec{\beta}_{hat}) = \vec{x}^T \vec{\beta}$

$var(\hat{y}) = \vec{x}^T \, var(\vec{\beta}_{hat}) \, \vec{x} = \sigma^2 \vec{x}^T (X^T X)^{-1} \vec{x}$

Theorem E. Under the MLR assumptions, we can use $\vec{\beta}_{\text{hat}}$ to predict at any $\vec{X} = \vec{x}$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \vec{x}^T \hat{\beta}_{\text{hat}}.$$

Then $E(\hat{y}) = \vec{x}^T \vec{\beta}$, $\text{var}(\hat{y}) = \sigma^2 \vec{x}^T (\mathbf{X}^T\mathbf{X})^{-1} \vec{x}$ and $\hat{y}$ is independent of RSS.

$\Rightarrow$ 95% exact confidence interval for $\beta_0 + \cdots + \beta_p x_p$:

$$\hat{y} \pm t_{n-p-1}(\alpha/2) \cdot \text{se}(\hat{y})$$

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1}$$

$$\text{se}(\hat{y}) = \sqrt{\hat{\sigma}^2 \vec{x}^T (\mathbf{X}^T\mathbf{X})^{-1} \vec{x}}$$

For prediction intervals of the response:

$$E(y - \hat{y}) = 0, \ \text{var}(y - \hat{y}) = \sigma^2 \left[1 + \vec{x}^T (\mathbf{X}^T\mathbf{X})^{-1} \vec{x}\right]$$ and y is independent of $\hat{y}$ and RSS.

CIs of population mean

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

Obtain a 95% CI for the mean IQ score when brain size = 96, height = 72, weight = 120.

```
> x_new <- c(1, 96, 72, 120)
> fitted <- x_new %*% coefs
> se_fitted <- sqrt(s2_hat*t(x_new)%*%solve(t(X)%*%X)%*%x_new)
> c(fitted - t_alpha*se_fitted, fitted + t_alpha*se_fitted)
```

```
> new <- data.frame(Intercept=1, Brain=96,
+                   Height=72, Weight=120)
> predict(fit, newdata = new, interval = 'confidence', level=0.95)
```

```
       fit      lwr      upr
112.5171 91.83446 133.1998
```

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1}$$

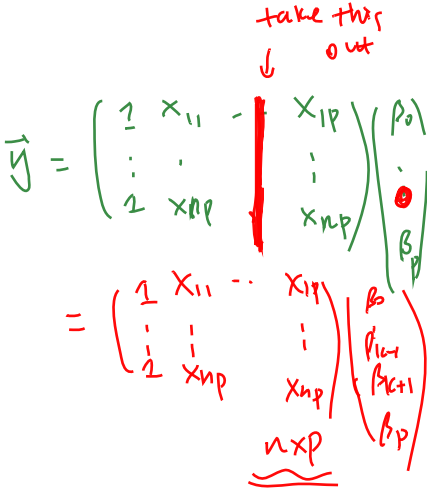
$$\text{se}(\hat{y}) = \sqrt{\hat{\sigma}^2 \vec{x}^T (\mathbf{X}^T\mathbf{X})^{-1} \vec{x}}$$

```
predict(fit, newdata = new, interval = 'prediction', level=0.95)
       fit      lwr      upr
112.5171 67.28453 157.7497
```

Testing $H_0 : \beta_k = 0$

Making sense of R output

08/05/2021



Hypothesis testing

Theorem G. Under the MLR assumptions, find the LRT for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}_{n \times (p+1)}$$

Solution: $\to \Theta_0 = \{ \beta_0 \in \mathbb{R}, \beta_2 \in \mathbb{R}, \cdots \beta_p \in \mathbb{R}, \sigma^2 > 0 \}$

$\to \Theta = \{ \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, \beta_2 \in \mathbb{R}^2, \cdots \beta_p \in \mathbb{R}, \sigma^2 > 0 \}$

$$\sup_{\Theta} L(\vec{\beta}, \hat{\sigma}^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \cdots - \hat{\beta}_p X_{ip})^2}{2\hat{\sigma}^2}}$$

RSS

$$\sup_{\Theta_0} L(\beta, \sigma^2) = \sup_{\Theta_0} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_2 X_{i2} - \cdots \beta_p X_{ip})^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right)^n e^{-\frac{\vec{Y}^T(I-H)\vec{Y}}{2\frac{1}{n}\vec{Y}^T(I-H)\vec{Y}}}$$

$$X_0 = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n2} & & X_{np} \end{pmatrix}_{n \times p} = \left(\frac{n}{2\pi \cdot RSS_0}\right)^{\frac{n}{2}} e^{-\frac{n}{2}} = \left(\frac{n}{2\pi \cdot RSS}\right)^{\frac{n}{2}} e^{\frac{n}{2}}$$

$$RSS_0 = \vec{Y}^T(I - H_0)\vec{Y}$$

$$H_0 = X_0(X_0^T X_0)^{-1} X_0^T$$

$$\lambda(Y_n) = \frac{\sup_{\Theta_0} L}{\sup_{\Theta} L} = \left(\frac{RSS}{RSS_0}\right)^{\frac{n}{2}} = \left(\frac{\vec{Y}^T(I-H)\vec{Y}}{\vec{Y}^T(I-H_0)\vec{Y}}\right)^{\frac{n}{2}}$$

SS_E

$$= \left(\frac{Y^T(I-H)Y}{Y^T(I-H)Y + Y^T(H-H_0)Y}\right)^{\frac{n}{2}}$$

$SS_E + SS_B$

$$R = \{\lambda(T_n) \le c\} = \left\{\left[\frac{1}{1+\frac{\vec{Y}^T(H-H_0)\vec{Y}}{\vec{Y}^T(I-H)\vec{Y}}}\right]^{\frac{n}{2}} \le c\right\} = \left\{\frac{\vec{Y}^T(H-H_0)\vec{Y}}{\vec{Y}^T(I-H)\vec{Y}} \ge c'\right\}$$

$\in SS_E$

① $\vec{Y}^T(I-H)\vec{Y} \sim \chi^2_{n-p-1}$.

② $\vec{Y}^T(H-H_0)\vec{Y} \sim \chi^2_{\dim H - \dim H_0}$ under H_0

$= \chi^2_1$

$\text{rank}(H-H_0) = \text{trace}(H-H_0) = \text{trace}\,H - \text{trace}\,H_0 = \text{rank}\,X - \text{rank}\,X_0$

$= 1$.

$\vec{Y}^T(H-H_0)\vec{Y} \sim \chi^2_1$.

Under H_0, $\dfrac{\vec{Y}^T(H-H_0)\vec{Y} / (\dim H - \dim H_0)}{\vec{Y}(I-H)\vec{Y} / (n-p-1)} \sim F_{\dim H - \dim H_0,\, n-p-1}$.

$= F_{1,n-p-1}$

$= t^2_{n-p-1}$

③ Independence:

It suffices to show $(I-H)\vec{Y} \perp\!\!\!\perp (H-H_0)\vec{Y}$

$\text{cov}((I-H)\vec{Y}, (H-H_0)\vec{Y}) = (I-H)(H-H_0)$

$= H - H_0 - H + HH_0$

$= H - H_0 - H + H_0$

$= 0$.

$HH_0 = H_0$.

$H X_0 (X_0^T X_0)^{-1} X_0^T = X_0$

↓

projecting column vectors of X_0 onto span(X). ⇒ projecting onto the span that contains all column vectors of X_0

Hypothesis testing

Theorem G'. Now find the LRT for $H_0: \beta_1 = 0$ and $\beta_2 = 0$ versus $H_1: H_0$ is not true.

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}_{n\times(p+1)}$$

$$X_0 = \begin{pmatrix} 1 & X_{13} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n3} & \cdots & X_{np} \end{pmatrix}_{n\times(p-1)}$$

$$\sup_{H} L(\mu, b^2) = \left(\frac{1}{2\pi \cdot RSS}\right)^{\frac{n}{2}} e^{-\frac{n}{2}}.$$

$$\sup_{H_0} L(\mu, b^2) = \left(\frac{1}{2\pi \cdot RSS_0}\right)^{\frac{n}{2}} e^{-\frac{n}{2}}$$

$$RSS_0 = \vec{Y}^T (I - H_0) \vec{Y}$$

$H_0 = X_0 (X_0^T X_0)^{-1} X_0^T$

$$R = \left\{ \frac{\vec{Y}^T (H - H_0) \vec{Y}}{\vec{Y}^T (I - H) \vec{Y}} \geq c' \right\}$$

Under H_0, $\dfrac{\vec{Y}^T (H - H_0) \vec{Y} / (\dim H - \dim H_0)}{\vec{Y}^T (I - H) \vec{Y} / (n - p - 1)} \sim F_{2,\, n-p-1}$.

Hypothesis testing

Theorem G'. Now find the LRT for $H_0:$ a certain set of β's are 0 versus $H_1: H_0$ is not true.

$$\mathbf{X} = \left(\begin{array}{c|c|c} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{array}\right)$$

$$F \text{ statistic} = \frac{\vec{Y}^T(\mathbf{H}-\mathbf{H}_0)\vec{Y}/m}{\vec{Y}^T(\mathbf{I}-\mathbf{H})\vec{Y}/(n-p-1)} \sim F_{m,\,n-p-1} \text{ under } H_0.$$

$m = \dim H - \dim H_0$

$= \#$ of zero β's in H_0

$X_0 =$ ( )

$H_0 = X_0 (X_0^T X_0)^{-1} X_0^T$

Output from R

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix}$$

Brain Height Weight

$H_0: \beta_{brain} = 0$

$F\text{-stat} \sim F_{1,\, n-ncol(X)}$

```
> summary(fit)

Call:
lm(formula = PIQ ~ Brain + Height + Weight, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-32.74 -12.09  -3.84  14.17  51.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
X <- cbind(1, dat$Brain, dat$Height, dat$Weight)
y <- dat$PIQ
H <- X%*%solve(t(X)%*%X)%*%t(X)

X0 <- X[, -2]
H0 <- X0%*%solve(t(X0)%*%X0)%*%t(X0)
I <- diag(n)
numerator <- t(y)%*%(H-H0)%*%y/1
denominator <- t(y)%*%(I-H)%*%y/(n-ncol(X))
F_stat <- numerator/denominator

F_stat
[1] 13.37159
pf(F_stat, 1, n-ncol(X), lower.tail = FALSE)
[1] 0.0008556322
sqrt(F_stat)
[1] 3.656719
```

$\vec{y}^T (H - H_0)\vec{y} / 1$

$\vec{y}^T (I - H)\vec{y} / (n - ncol(X))$

$\hat{\sigma} = \sqrt{\frac{RSS}{n-p-1}} \sim n-p-1$

$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n} (y_i - \bar{y}_n)^2}$

Output from R

$H_0: \beta_{brain} = \beta_{weight} = \beta_{weight} = 0$.

$F_{stat} \sim F_{3,\ n-ncol(X)}$

```
> summary(fit)

Call:
lm(formula = PIQ ~ Brain + Height + Weight, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-32.74 -12.09  -3.84  14.17  51.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
X0 <- X[, -c(2,3,4)]
H0 <- X0%*%solve(t(X0)%*%X0)%*%t(X0)
I <- diag(n)
numerator <- t(y)%*%(H-H0)%*%y/3
denominator <- t(y)%*%(I-H)%*%y/(n-ncol(X))
F_stat <- numerator/denominator

F_stat
[1] 4.740931
pf(F_stat, 3, n-ncol(X), lower.tail = FALSE)
[1] 0.00721527
```

< 0.05

Fail to reject H_0

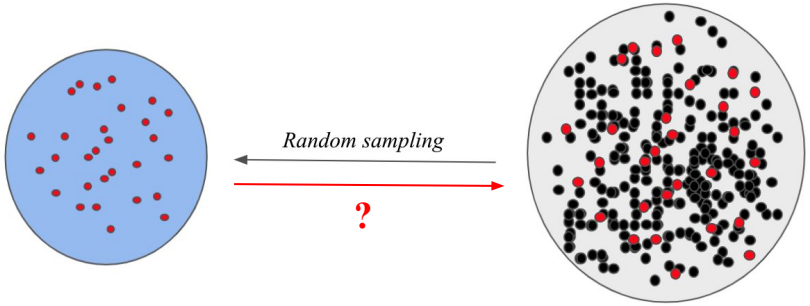
Bayesian statistics

8.6 of Rice

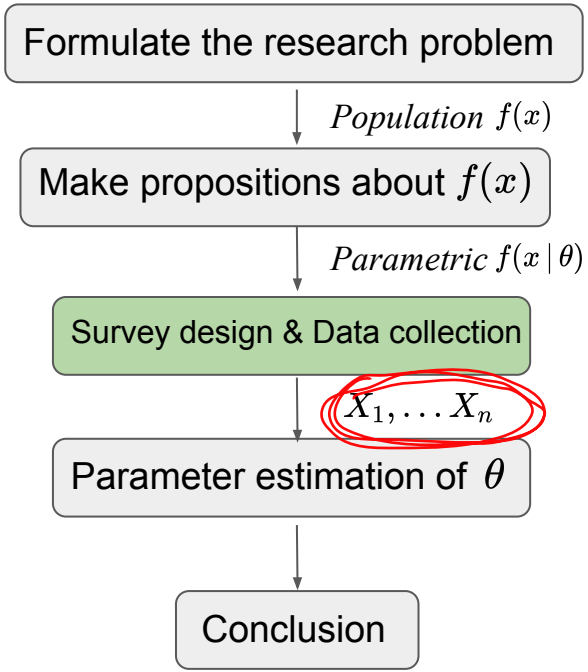
08/05/2021

Population vs. samples

How to learn from samples about $f(x)$?



- **Point estimate**: a particular value $\hat{\theta}$ that best approximates the parameter of interest.
  - Maximum likelihood estimators; Method of moment estimators;
  - Sufficiency; Estimation error quantification.
- **Interval estimate**: an interval $[\theta - a, \theta + b]$ that would contain the true parameter θ with a certain degree of *confidence*.
- **Hypothesis testing**: whether to reject a hypothesis.
  - $\theta > a$?
  - $\theta < a$?



- **Regression**: A special parametric model

$$Y \mid \beta, \sigma^2 = \underbrace{\mathbf{z}\beta}_{\text{Explanatory variables}} + N(0, \sigma^2)$$

Prior information

Example 4. What is a patient's probability of having liver disease if they are an alcoholic?

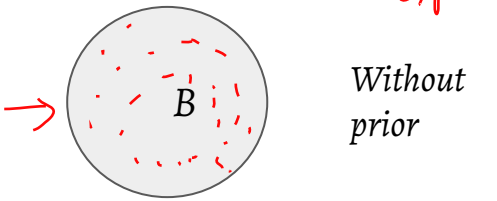
- Overall, 10% of patients have liver disease. P(A) = 0.10.
- Overall, 5% of patients are alcoholic. P(B)=0.05.
- Among those patients diagnosed with liver disease, 7% are alcoholics. P(B|A)=0.07.

from your experience / experiment

Data information *Prior information*

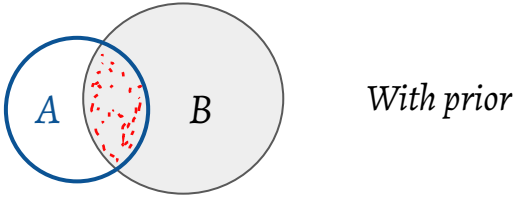
Bayes rule: $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$.

Posterior information *Marginalization*



Without prior

$P(A|B) = \dfrac{0.07 \times 0.1}{0.05} = 0.14 > P(A) = 0.1.$



With prior

Prior information

Example 5. We want to estimate the average height of UCB students, μ. Sample heights were collected from 100 students.

- **Model**:

$$x_j|\mu,\sigma^2 \sim N(\mu,\sigma^2), \quad ,j=1,\ldots,n.$$

- **Expert prior**: According to CDC.gov, average heights for US men and women (>20) are 69.0 inches and 63.5 inches.

Figure: 'Expert' prior $\pi(\mu)$

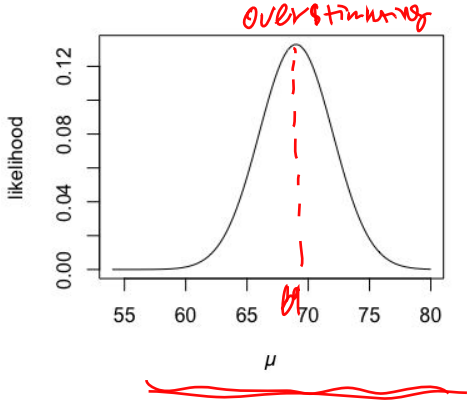
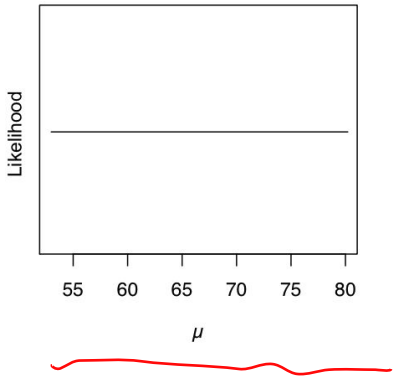


Figure: Naïve flat prior $\pi(\mu)$



$$\pi(\mu,\sigma^2\,|\,\mathbf{X}_n) = \frac{f(\mathbf{X}_n\,|\,\mu,\sigma^2)\pi(\mu)\pi(\sigma^2)}{f(\mathbf{X}_n)}.$$

- *Is the frequentist prior a good idea?*
- *Is the expert prior correct for the UCB student population?*

Posterior distribution

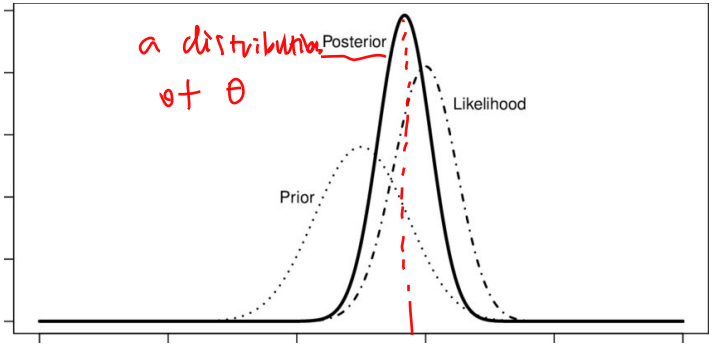
Definition. In the parametric model $f(x \mid \theta)$ that we assume, we regard θ as a random variable. Since θ is random, we assign a pdf/pmf which captures our belief about θ:

$$\pi(\theta) \rightarrow \text{ prior distribution on } \theta,$$

Then my modified belief about θ can be summarized in the posterior distribution

$$\pi(\theta \mid \mathbf{X}_n) = \frac{f(\mathbf{X}_n \mid \theta)\pi(\theta)}{\int_\Theta f(\mathbf{X}_n \mid \theta)\pi(\theta)d\theta} \propto f(\mathbf{X}_n \mid \theta)\pi(\theta).$$

$\Rightarrow$ $\hat{\theta}_B = E(\theta \mid \mathbf{X}_n)$ or $\text{median}(\theta \mid \mathbf{X}_n)$ or $\text{mode}(\theta \mid \mathbf{X}_n)$.



- *When n is small, prior has greater impact;*
- *When $n \rightarrow \infty$, data information dominates the prior information.*

Posterior distribution

$X \sim \text{Beta}(\alpha, \beta)$, $EX = \frac{\alpha}{\alpha+\beta}$, mode $X = \frac{\alpha-1}{\alpha+\beta-2}$.

Example 5. You flip a coin 20 times and get 13 heads. Do you have a fair coin?

Frequentist: $X_1, \dots X_{20}$ iid Bernoulli(p)

$f(\vec{x}_n \mid p) = p^{\sum_i x_i} (1-p)^{n - \sum_i x_i}$

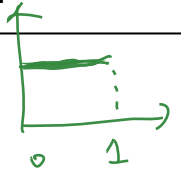
MLE: $\hat{p}_{MLE} = \bar{X}_n = \frac{13}{20} = 0.65$

MM

Estimators are based on $\vec{x}_n$

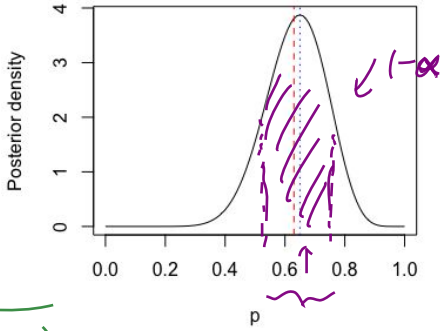
$\hat{p}_{MLE} \to$ sampling distribution

Bayesian: $p \sim U(0,1)$



$$\pi(p \mid \vec{x}_n) = \frac{f(\vec{x}_n \mid p)\,\pi(p)}{\int_0^1 f(\vec{x}_n \mid p)\,\pi(p)\,dp} = \frac{p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}}{\int_0^1 p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}\,dp} = \frac{p^{\sum x_i}(1-p)^{n-\sum_i x_i}}{\text{Beta}\left(\sum_i x_i + 1,\ n - \sum_i x_i + 1\right)}$$

$$\sim \text{Beta}\left(\sum_i x_i + 1,\ n - \sum_i x_i + 1\right)$$



$\hat{p}_{Bayes} = \text{mode}(p \mid \vec{x}_n) = \frac{\sum_i x_i}{n} = \bar{x}_n$ or $\hat{p}_{Bayes} = E(p \mid \vec{x}_n) = \frac{\sum_i x_i + 1}{n+2} = \frac{14}{22} = 0.63$

We can easily quantify the uncertainty of the Bayesian estimators.

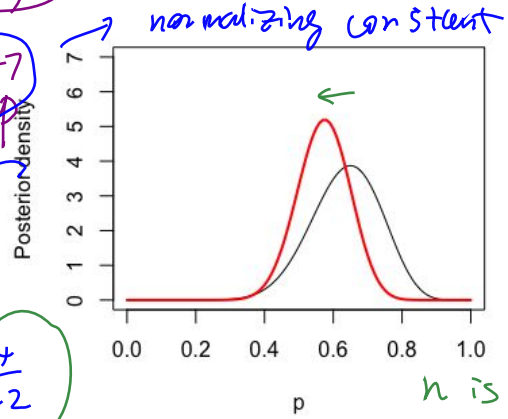
Posterior distribution with informative prior

Example 5. You flip the same coin 20 times again and get 10 heads. What is your modified belief on p?

Informative prior: $\pi(p) = \text{Beta}(14, 8)$.

$$\pi(p \mid \underline{x}_n) = \frac{f(\underline{x}_n \mid p)\,\pi(p)}{\int_0^1 f(\underline{x}_n \mid p)\,\pi(p)\,dp} = \frac{p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}\, p^{13}(1-p)^{7}}{\int_0^1 p^{\sum_i x_i+13}(1-p)^{n-\sum_i x_i+7}\,dp} = p^{\sum_i x_i+13}(1-p)^{n-\sum_i x_i+7} \times \text{normalizing constant}$$

$$\sim \text{Beta}\left(\sum_i x_i + 14,\ n - \sum_i x_i + 8\right)$$



$$\hat{p}_{\text{Bayes}} = \frac{\sum_i x_i + 14}{n + 22} = \frac{n}{n+22}\cdot\frac{\sum_i x_i}{n} + \frac{22}{n+22}\cdot\frac{14}{22}$$

$\bar{x}_n$: Frequentist estimator based on data information

prior mean: Estimator based on the prior.

$n \to \infty$, $\frac{n}{n+22} \to 1$, $\frac{22}{n+22} \to 0$, $\hat{p}_{\text{Bayes}} = \bar{x}_n$.

n is small $\Rightarrow \frac{22}{n+22}$ ↑

Our information on p is updated and more accurate.

Posterior distribution with informative prior

$P(\theta \in [L(\bar{X}_n), U(\bar{X}_n)]) \in$

$\bar{X}_n - z_{\alpha/2}\frac{S}{\sqrt{n}}$

Example 5. You flip the same coin 20 times again and get 10 heads. What is your modified belief on p?

Bayesian CI:

$P(\theta \in I \mid \mathbf{X}_n) = 1 - \alpha.$

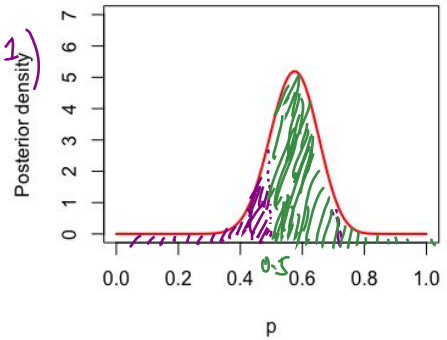
Credible interval

Bayesian HT:

Accept H_0 if $P(\theta \in \Theta_0 \mid \mathbf{X}_n) \geq P(\theta \in \Theta_1 \mid \mathbf{X}_n)$, else reject.

$H_0: p \leq 0.5$ vs. $H_1: p > 0.5$

$(0, 0.5]$ $(0.5, 1)$



$P(p \leq 0.5 \mid \bar{X}_n) =$

$\wedge$

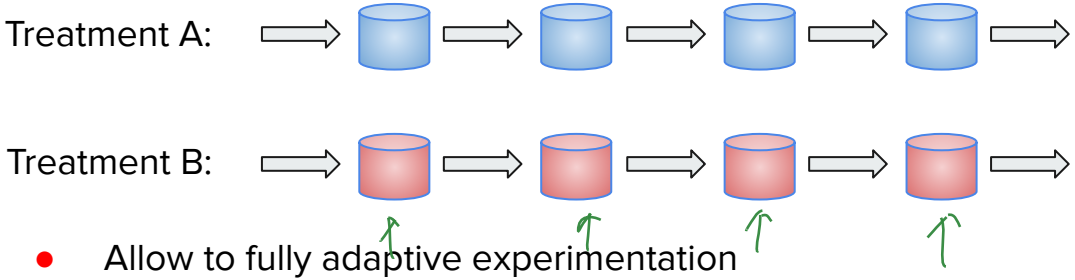
$P(p > 0.5 \mid \bar{X}_n) =$

Reject H_0 and conclude $p > 0.5$.

Application: Bayesian sequential update

Example 6. In large-scale A/B testing, it is costly and time-consuming to collect large number of samples. What if you want to save money or want to know the test results as the experimentation goes on.

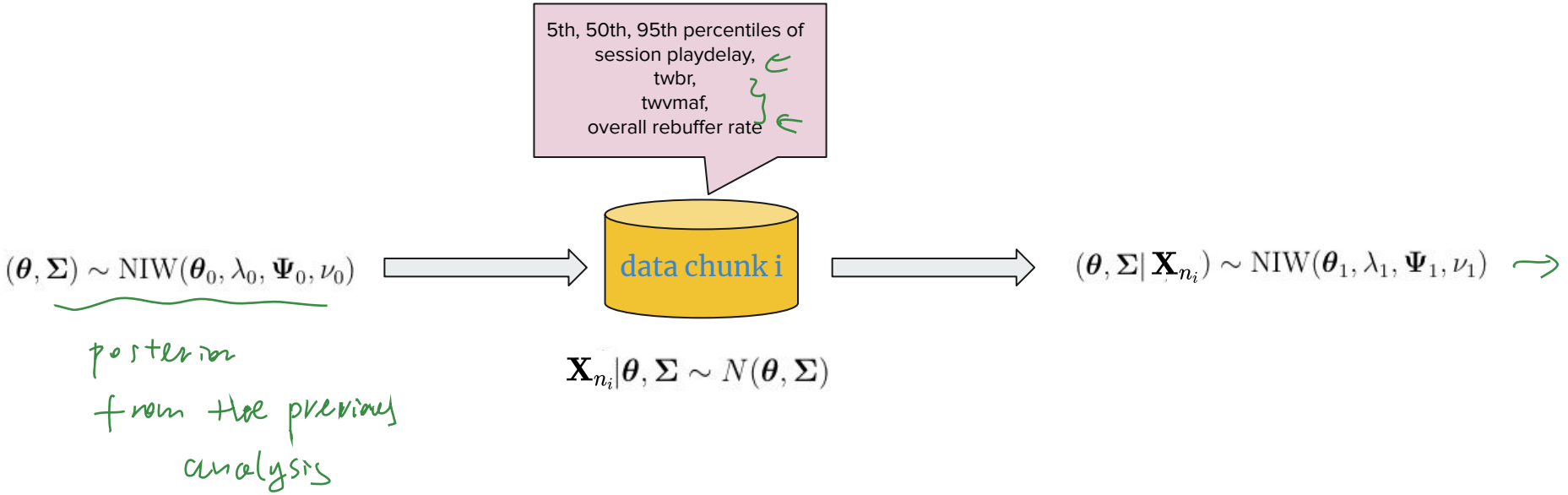
- Collect data chunks



- Allow to fully adaptive experimentation

Application: Bayesian sequential update

Example 6. In large-scale A/B testing, it is costly and time-consuming to collect large number of samples. What if you want to save money or want to know the test results as the experimentation goes on.



Application: Bayesian sequential update

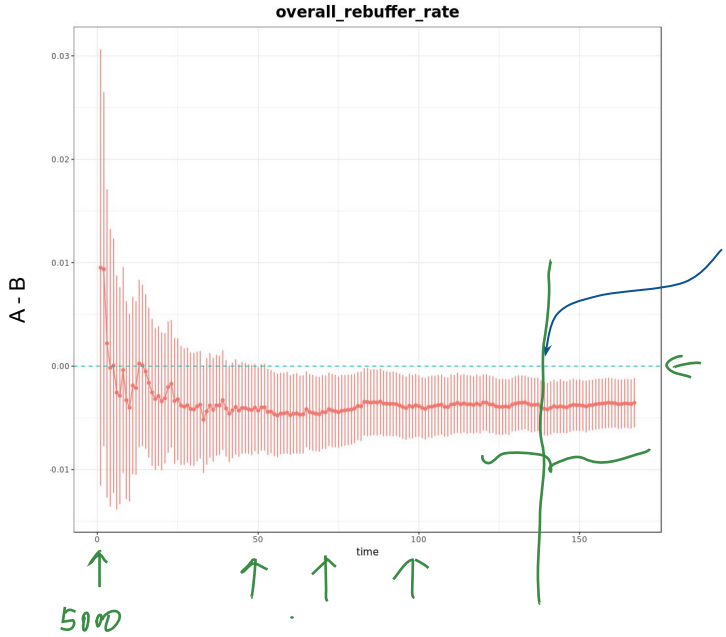
Example 6. In large-scale A/B testing, it is costly and time-consuming to collect large number of samples. What if you want to save money or want to know the test results as the experimentation goes on.

$\theta_A \mid \mathcal{X}_n$

$\theta_B \mid \mathcal{X}_n$

$\theta_A - \theta_B \mid \mathcal{X}_n$

< 0



Should be confident to stop the experiment at this point.

Next week ...

- Review session & practice final;
- Final exam.