Topic 16
Interval Estimation
Prediction Intervals & the Bootstrap

Outline

Prediction Intervals

Recall that a γ-level confidence set has the prescribed high probability of containing the true parameter value θ. Thus,

$$P_\theta\{\theta \in \hat{C}(\mathbf{X})\} \geq \gamma \quad \text{for all } \theta \in \Theta,$$

based on independent and identically observations $\mathbf{X} = (X_1, \ldots, X_n)$.

Complementary to this notion is the concept of a γ-level prediction set, $\hat{P}(\mathbf{x})$. This addresses the question of future observation X_* will fall, given that $\mathbf{X}$ has been observed.

$$P_\theta\{X_* \in \hat{P}(\mathbf{X})\} \geq \gamma \quad \text{for all } \theta \in \Theta,$$

Prediction Intervals

Example. Let $\mathbf{X} = (X_1, \ldots, X_n) \sim N(\mu, \sigma_0^2)$ with μ *unknown* and σ_0^2 known. The pivot $Q(\mathbf{X}, \mu) = \bar{X} - \mu$ was used to obtain γ-level confidence interval

$$\hat{C}(\mathbf{X}) = \left\{ \bar{X} - z_{(1-\gamma)/2} \frac{\sigma_0}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\gamma)/2} \frac{\sigma_0}{\sqrt{n}} \right\}$$

Now for a future observation $X_* \sim N(\mu, \sigma_0^2)$. We consider the distribution of $Q(\mathbf{X}, X_*) = \bar{X} - X_*$. Note that it is normally distributed, with mean 0, and variance

$$\mathrm{Var}(\bar{X} - X_*) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(X_*) = \frac{\sigma_0^2}{n} + \sigma_0^2.$$

giving us the standard normal $Z = \frac{\bar{X} - X_*}{\sigma_0 \sqrt{1 + 1/n}}$ and the γ-level prediction interval

$$\hat{P}(\mathbf{X}) = \left\{ \bar{X} - z_{(1-\gamma)/2} \sigma_0 \sqrt{1 + 1/n} < X_* < \bar{X} + z_{(1-\gamma)/2} \sigma_0 \sqrt{1 + 1/n} \right\}$$

Linear Regression

For ordinary linear regression, we have given least squares estimates for the slope β and the intercept α. For data $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, our model is

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where ϵ_i are independent $N(0, \sigma^2)$ random variables. Recall that the estimator for the slope

$$\hat{\beta}(x, y) = \frac{\text{cov}(x, y)}{\text{var}(x)}.$$

Exercise. Show that $\hat{\beta}$ is unbiased.

Linear Regression

First, because $E_{(\alpha,\beta)}\epsilon_i = 0$,

$$E_{(\alpha,\beta)} Y_i = E_{(\alpha,\beta)}[\alpha + \beta x_i + \epsilon_i] = \alpha + \beta x_i.$$

By the linearity property of expectation $E_{(\alpha,\beta)} \bar{Y} = \alpha + \beta \bar{x}$. Taken together,

$$E_{(\alpha,\beta)}[Y_i - \bar{Y}] = (\alpha + \beta x_i) - (\alpha + \beta \bar{x}) = \beta(x_i - \bar{x}).$$

To show that $\hat{\beta}$ is an unbiased estimator,

$$\begin{aligned} E_{(\alpha,\beta)}\hat{\beta} &= \frac{E_{(\alpha,\beta)}[\text{cov}(x, Y)]}{\text{var}(x)} = \frac{1}{(n-1)\text{var}(x)} E_{(\alpha,\beta)} \left[\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \right] \\ &= \frac{1}{(n-1)\text{var}(x)} \sum_{i=1}^{n} (x_i - \bar{x}) E_{(\alpha,\beta)}[Y_i - \bar{Y}] \\ &= \frac{\beta}{(n-1)\text{var}(x)} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = \beta. \end{aligned}$$

Linear Regression

Exercise. Show that the variance of $\hat{\beta}$ equals

$$\frac{\sigma^2}{(n-1)\text{var}(x)}.$$

Use the fact that $y_i - \beta x_i = \alpha + \epsilon_i$. Thus,

$$\begin{aligned}
\text{Var}_{(\alpha,\beta)}(\hat{\beta}) &= \text{Var}_{(\alpha,\beta)}\left(\frac{1}{(n-1)\text{var}(x)}\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\epsilon_i)\right) \\
&= \frac{1}{(n-1)^2\text{var}(x)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\text{Var}_{(\alpha,\beta)}(\alpha+\epsilon_i) \\
&= \frac{1}{(n-1)^2\text{var}(x)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2}{(n-1)\text{var}(x)}
\end{aligned}$$

Linear Regression

Generally, σ is unknown. However, the variance of the residuals,

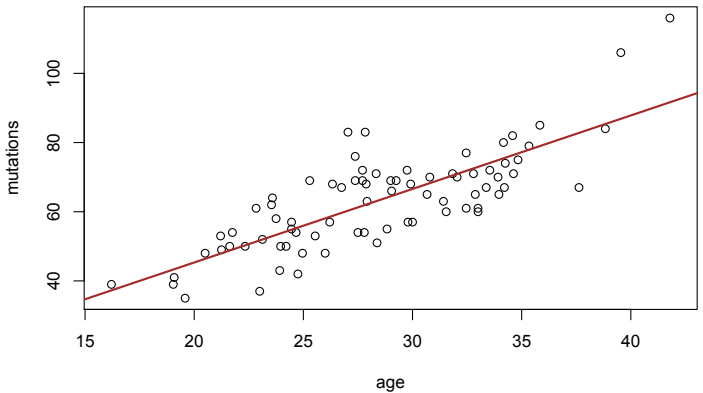
$$s_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\alpha} - \hat{\beta}x_i))^2$$

is an unbiased estimator of σ^2 and s_u/σ has a t distribution with $n-2$ degrees of freedom. This gives the t-interval

$$\hat{\beta} \pm t_{n-2,(1-\gamma)/2}\frac{s_u}{s_x\sqrt{n-1}}.$$

Linear Regression

Example. We nvestigate the relationship of age of parents to the *de novo* mutations in the offspring for the 78 Icelandic trios. We use the age of the parents to predict the number of mutations in the offspring. Thus, age is on the horizontal axis.



Constructing Confidence Intervals

For a general summary,

```
> summary(mutations.lm)
Call:
lm(formula = mutations ~ age)
Residuals:
     Min       1Q   Median       3Q      Max
-15.7849  -7.1364  -0.1244   5.1745  24.3591
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8145     5.5034   0.511    0.611
age           2.1255     0.1904  11.164   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 8.79 on 76 degrees of freedom
Multiple R-squared:  0.6212, Adjusted R-squared:  0.6162
F-statistic: 124.6 on 1 and 76 DF,  p-value: < 2.2e-16
```

Linear Regression

Similarly we find that

- $\text{Var}_{(\alpha,\beta)}(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\text{var}(x)} \right)$
- $\text{Cov}_{(\alpha,\beta)}(\hat{\alpha}, \hat{\beta}) = -\bar{x}\text{Var}_{(\alpha,\beta)}(\hat{\beta})$

For a new observation (x_*, Y_*). Set $\hat{Y}_* = \hat{\alpha} + \hat{\beta}x_*$ and $W = Y_* - \hat{Y}_*$. W is normally distributed. Its mean

$$E_{(\alpha,\beta)}[W] = E_{(\alpha,\beta)}[Y_* - \hat{Y}_*] = E_{(\alpha,\beta)}[Y_* - \hat{\alpha} - \hat{\beta}x_*] = E_{(\alpha,\beta)}[Y_*] - \alpha - \beta x_* = 0.$$

For the variance, note that Y_* and $\hat{Y}_*$ are independent. First, $\text{Var}_{(\alpha,\beta)}(Y_*) = \sigma^2$.

$$\begin{aligned} \text{Var}_{(\alpha,\beta)}(\hat{Y}_*) &= \text{Var}_{(\alpha,\beta)}(\hat{\alpha}) + 2x_*\text{Cov}_{(\alpha,\beta)}(\hat{\alpha}, \hat{\beta}) + x_*^2\text{Var}_{(\alpha,\beta)}(\hat{\beta}) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\text{var}(x)} + \frac{-2\bar{x}x_* + x_*^2}{(n-1)\text{var}(x)} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{(n-1)\text{var}(x)} \right) \end{aligned}$$

Linear Regression

Thus, $Y_*|x_* \sim N\left(\alpha + \beta x_*, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_*-\bar{x})^2}{(n-1)\text{var}(x)}\right)\right)$ We have the estimate s_u^2 for σ^2. This give us the two-sided prediction interval

$$\hat{P}(\mathbf{x}, x_*) = \left\{(\hat{\alpha} + \hat{\beta}x_*) \pm t_{n-2,(1-\gamma)/2} s_u \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{(n-1)\text{var}(x)}}\right\}.$$

For the Icelandic mutation data, here are the summaries for a 95% prediction interval.

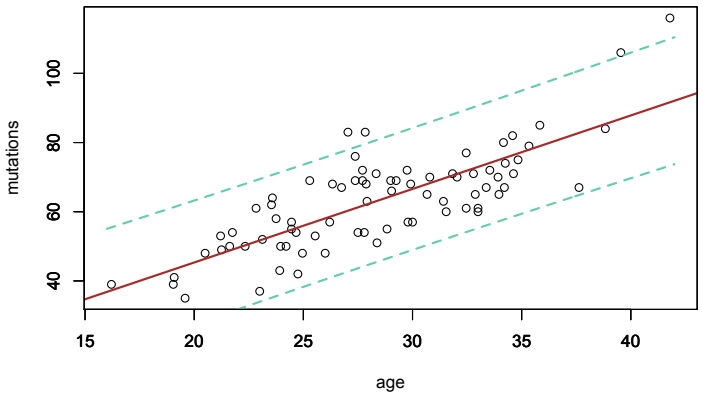
$$n = 78, \quad \hat{\alpha} = 2.8154 \quad \hat{\beta} = 2.1255 \quad \bar{x} = 28.43109$$

$$t_{76,0.025} = 1.9917 \quad s_u = 8.79, \quad \text{var}(x) = 27.6850$$

$$\hat{P}(\mathbf{x}, x_*) = \left\{(2.8154 + 2.1255x_*) \pm 1.9917 \times 8.79 \sqrt{\frac{79}{78} + \frac{(x_* - 28.4311)^2}{77 \times 27.6859}}\right\}$$

Linear Regression

The relationship of age of parents to the *de novo* mutations with the 95% prediction interval indicated by the teal colored dashed lines. The smallest prediction error is 17.62 mutations at the mean age 28.43 years.



The Bootstrap

The strategy of the bootstrap is to perform a calculation using the empirical cumulative distribution function $\hat{F}_n$ as an estimate of the calculation one would like to perform using F.

Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a simple random sample with state space $\mathcal{X}$. If the empirical distribution function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i)$$

is used, then the method is the nonparametric bootstrap.

If $\hat{\theta}_n$ is an estimate of the parameter $\theta \in \Theta$ and $\hat{F}_n(x) = F_{X_1}(x|\hat{\theta}_n)$ is used, then the method is the parametric bootstrap.

The Bootstrap

Let $\mathcal{F}$ be a set of cumulative distribution functions and let

$$R : \mathcal{X} \times \mathcal{F} \to \mathbb{R}$$

be some function of interest, e.g., the difference between the sample median of $\mathbf{X}$ and the median of F. Then the bootstrap replaces

$$R(\mathbf{X}, F) \qquad \text{by} \qquad R(\mathbf{X}^*, \hat{F}_n).$$

Note that

- $\mathbf{X}$ is a simple random sample of size n from F, and
- $\mathbf{X}^*$ is a resample of size n from $\hat{F}_n$.

The bootstrap was originally designed as a tool for estimating bias and standard error of a statistic.

The Bootstrap

Example Assume that the sample has real values from a cumulative distribution function F. Let

$$R(\mathbf{X}, F) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2 - \mu^2 \quad \mu = \int_{\mathbb{R}} x dF(x),$$

then

$$R(\mathbf{X}^*, \hat{F}_n) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^* \right)^2 - \bar{x}_n^2,$$

where $\bar{x}_n$ is the observed sample average. Use

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$$

as an estimate of the variance. Now

$$E[R(\mathbf{X}, F)] = \frac{1}{n} \sigma^2, \qquad E[R(\mathbf{X}^*, \hat{F}_n) | \mathbf{X} = \mathbf{x}] = \frac{1}{n} s_n^2.$$

The Bootstrap

Example. Snell's law tell us how light bends at an interface - the angle of incidence versus the angle of refraction - based on the ratio of the velocities of light in the two isotropic media.

If the angle of incidence of a laser beam in air is θ_1 radians and the angle of refraction in water is θ_2, then

$$\beta = \frac{\sin\theta_1}{\sin\theta_2}$$

where β is the velocity of light in water as a fraction of the speed of light in a vacuum.

Make repeated independent measurements in radians, $\theta_{1,1}, \theta_{1,2}, \cdots, \theta_{1,16}$ of the angle of incidence in air and $\theta_{2,1}, \theta_{2,2}, \cdots, \theta_{2,25}$ of the angle of refraction in water.

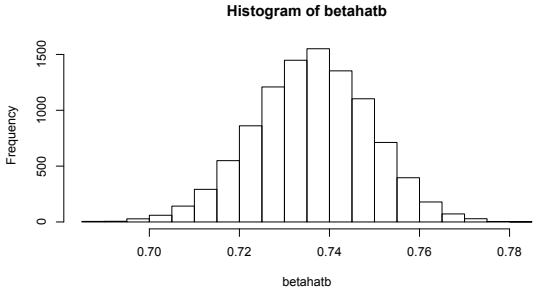
Constructing Confidence Intervals

Here are the data.

```
> theta2
 [1] 0.529 0.525 0.528 0.513 0.501 0.526 0.526 0.511 0.524 0.538 0.532
[12] 0.530 0.522 0.515 0.499 0.529
> theta1
 [1] 0.336 0.381 0.365 0.396 0.393 0.386 0.395 0.399 0.423 0.370 0.301
[12] 0.412 0.304 0.331 0.413 0.379 0.356 0.394 0.426 0.359 0.348 0.386
[23] 0.415 0.351 0.375
> (betahat<-sin(mean(theta1))/sin(mean(theta2)))
[1] 0.7363183
> for (i in 1:10000)theta1b<-sample(theta1,length(theta1),replace=TRUE);
  theta2b<-sample(theta2,length(theta2),replace=TRUE);
   betahatb[i]<-sin(mean(theta1b))/sin(mean(theta2b))
> summary(betahatb)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6880  0.7275  0.7363  0.7361  0.7450  0.7814
```

Snell's Law

```
> p<-c(0.5,1,2.5,50,97.5,99,99.5)/100
> quantile(betahatb,probs=p)
     0.5%         1%       2.5%        50%      97.5%        99%      99.5%
0.7019262 0.7050475 0.7102220 0.7362894 0.7607356 0.7652078 0.7681467
> sd(betahatb)
[1] 0.01288418
```



Snell's Law

Let's compare this to the delta method. Let $\beta = g(\theta_1, \theta_2)$

$$\begin{aligned}
\sigma^2_{\hat{\beta}} &\approx \left(\frac{\partial}{\partial\theta_1} g(\bar{\theta}_1, \bar{\theta}_2)\right)^2 \frac{s_1^2}{n_1} + \left(\frac{\partial}{\partial\theta_2} g(\bar{\theta}_1, \bar{\theta}_2)\right)^2 \frac{s_2^2}{n_2} \\
&= \left(\frac{\cos\bar{\theta}_1}{\sin\bar{\theta}_2}\right)^2 \frac{s_1^2}{n_1} + \left(\frac{\sin\bar{\theta}_1 \cos\bar{\theta}_2}{\sin^2\bar{\theta}_2}\right)^2 \frac{s_2^2}{n_2} \\
\sigma^2_{\hat{\beta}} &\approx 0.01296248
\end{aligned}$$

```
> quantile(betahatb,probs=p)
     0.5%         1%       2.5%        50%      97.5%        99%      99.5%
0.7019262 0.7050475 0.7102220 0.7362894 0.7607356 0.7652078 0.7681467

> qnorm(p,betahat,sqrt(sb2))
0.7029292 0.7061631 0.7109123 0.7363183 0.7617243 0.7664735 0.7697074
```