There is a discrepancy in R output from the functions `step`, `AIC`, and `BIC` over how to compute the AIC. The discrepancy is not very important, because it involves a difference of a constant factor that cancels when using AIC or BIC to compare two models. But it might be helpful to understand the differences so that you can compare output from these two functions.

AIC and BIC are based on the maximum likelihood estimates of the model parameters. In maximum likelihood, the idea is to estimate parameters so that, under the model, the probability of the observed data would be as large as possible. The likelihood is this probability, and will always be between 0 and 1. It is common to consider likelihoods on a log scale. Logarithms of numbers between 0 and 1 are negative, so log-likelihoods are negative numbers. It is also common to multiply log-likelihoods by -2, for reasons we will not explore.

In a regression setting, the estimates of the β_i based on least squares and the maximum likelihood estimates are identical. The difference comes from estimating the common variance σ^2 of the normal distribution for the errors around the true means. We have been using the best *unbiased* estimator of σ^2, $\hat{\sigma}^2 = \text{RSS}/(n-p)$ where there are p parameters for the means (p different β_i parameters) and RSS is the *residual sum of squares*. This estimate does not tend to be too large or too small on average. The maximum likelihood estimate, on the other hand, is RSS/n. This estimate has a slight negative bias, but also has a smaller variance.

Putting all of this together, we can write -2 times the log-likelihood to be

$$n + n\log(2\pi) + n\log(\text{RSS}/n)$$

in a regression setting. Now, AIC is defined to be -2 times the log-likelihood plus 2 times the number of parameters. If there are p different β_i parameters, there are a total of $p+1$ parameters if we also count σ^2. The correct formula for the AIC for a model with parameters $\beta_0, \ldots, \beta_{p-1}$ and σ^2 is

$$\text{AIC} = n + n\log 2\pi + n\log(\text{RSS}/n) + 2(p+1)$$

and the correct formula for BIC is

$$\text{BIC} = n + n\log 2\pi + n\log(\text{RSS}/n) + (\log n)(p+1)$$

This is what the functions `AIC` and `BIC` calculate in R. The AIC and BIC formulas in your textbook ignore the leading two terms $n + n\log 2\pi$ and use p instead of $p+1$. When comparing AIC or BIC between two models, however, it makes no difference which formula you use because the differences will be the same regardless which choice you make.

```
> case1201 = read.table("sleuth/case1201.csv", header = T, sep = ",")
> attach(case1201)
> keep <- STATE != "Alaska"
> x <- data.frame(SAT = SAT[keep], ltakers = log(TAKERS[keep]),
+     income = INCOME[keep], years = YEARS[keep], public = PUBLIC[keep],
+     expend = EXPEND[keep], rank = RANK[keep])
> detach(case1201)
> attach(x)
```

Example Computation in R `AIC` is part of the base package. You can find the `BIC` using the `AIC` function with the option $k = log(n)$, or, you can load the nonlinear mixed effects library and call the `BIC` function directly. Here is an example that demonsrates the above ideas.

```
> library(nlme)
```

```
Loading required package: nls
Loading required package: lattice
Loading required package: grid

> n <- nrow(x)
> fit0 <- lm(SAT ~ 1)
> summary(fit0)

Call:
lm(formula = SAT ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-158.45  -59.45   19.55   50.55  139.55

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   948.45      10.21   92.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.5 on 48 degrees of freedom

> rss0 <- sum(residuals(fit0)^2)
> n + n * log(2 * pi) + n * log(rss0/n) + 2 * 2

[1] 560.4736

> AIC(fit0)

[1] 560.4736

> n + n * log(2 * pi) + n * log(rss0/n) + log(n) * 2

[1] 564.2573

> AIC(fit0, k = log(n))

[1] 564.2573

> BIC(fit0)

[1] 564.2573

> fit1 <- lm(SAT ~ ltakers)
> summary(fit1)

Call:
lm(formula = SAT ~ ltakers)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-93.328 -21.380   4.154  22.614  50.794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1112.408     12.386   89.81   <2e-16 ***
ltakers      -59.175      4.167  -14.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.41 on 47 degrees of freedom
Multiple R-Squared: 0.811,        Adjusted R-squared: 0.807
F-statistic: 201.7 on 1 and 47 DF,  p-value: < 2.2e-16

> rss1 <- sum(residuals(fit1)^2)
> n + n * log(2 * pi) + n * log(rss1/n) + 2 * 3

[1] 480.832

> AIC(fit1)

[1] 480.832

> n + n * log(2 * pi) + n * log(rss0/n) + log(n) * 3

[1] 568.1491

> AIC(fit1, k = log(n))

[1] 486.5075

> BIC(fit1)

[1] 486.5075
```

The criteria AIC and BIC will not always lead to the same model. Compare the results from these forward selections.

Forward Selection using AIC

```
> step(lm(SAT ~ 1), SAT ~ ltakers + income + years + public + expend +
+     rank, direction = "forward")

Start:  AIC= 419.42
 SAT ~ 1

          Df Sum of Sq    RSS    AIC
+ ltakers  1    199007  46369    340
+ rank     1    190297  55079    348
```

```
+ income   1    102026 143350    395
+ years    1     26338 219038    416
<none>                245376    419
+ public   1      1232 244144    421
+ expend   1       386 244991    421

Step:  AIC= 339.78
 SAT ~ ltakers

         Df Sum of Sq   RSS   AIC
+ expend  1     20523 25846   313
+ years   1      6364 40006   335
<none>                46369   340
+ rank    1       871 45498   341
+ income  1       785 45584   341
+ public  1       449 45920   341

Step:  AIC= 313.14
 SAT ~ ltakers + expend

         Df Sum of Sq     RSS     AIC
+ years   1    1248.2 24597.6   312.7
+ rank    1    1053.6 24792.2   313.1
<none>                25845.8   313.1
+ income  1      53.3 25792.5   315.0
+ public  1       1.3 25844.5   315.1

Step:  AIC= 312.71
 SAT ~ ltakers + expend + years

         Df Sum of Sq     RSS     AIC
+ rank    1    2675.5 21922.1   309.1
<none>                24597.6   312.7
+ public  1     287.8 24309.8   314.1
+ income  1      19.2 24578.4   314.7

Step:  AIC= 309.07
 SAT ~ ltakers + expend + years + rank

         Df Sum of Sq     RSS     AIC
<none>                21922.1   309.1
+ income  1     505.4 21416.7   309.9
+ public  1     185.0 21737.1   310.7

Call:
lm(formula = SAT ~ ltakers + expend + years + rank)

Coefficients:
```

```
(Intercept)      ltakers       expend        years         rank
    399.115      -38.100        3.996       13.147        4.400
```

Forward Selection using BIC

```
> n <- nrow(x)
> step(lm(SAT ~ 1), SAT ~ ltakers + income + years + public + expend +
+     rank, direction = "forward", k = log(n))

Start:  AIC= 421.31
 SAT ~ 1

          Df Sum of Sq    RSS    AIC
+ ltakers  1    199007  46369    344
+ rank     1    190297  55079    352
+ income   1    102026 143350    399
+ years    1     26338 219038    420
<none>                 245376    421
+ public   1      1232 244144    425
+ expend   1       386 244991    425

Step:  AIC= 343.56
 SAT ~ ltakers

         Df Sum of Sq   RSS   AIC
+ expend  1     20523 25846   319
+ years   1      6364 40006   340
<none>                46369   344
+ rank    1       871 45498   347
+ income  1       785 45584   347
+ public  1       449 45920   347

Step:  AIC= 318.81
 SAT ~ ltakers + expend

         Df Sum of Sq     RSS     AIC
<none>                25845.8   318.8
+ years   1    1248.2 24597.6   320.3
+ rank    1    1053.6 24792.2   320.7
+ income  1      53.3 25792.5   322.6
+ public  1       1.3 25844.5   322.7

Call:
lm(formula = SAT ~ ltakers + expend)

Coefficients:
(Intercept)      ltakers       expend
   1028.582      -66.170        4.605
```

Comparison Notice that both approaches begin by first adding *ltakers* and then adding *expend*. But at this point, the AIC and BIC criteria lead to different decisions. The best new variable to add by other criterion is years. The change in log-likelihood is this.

```
> fit2 <- lm(SAT ~ ltakers + expend)
> rss2 <- sum(residuals(fit2)^2)
> fit3 <- lm(SAT ~ ltakers + expend + years)
> rss3 <- sum(residuals(fit3)^2)
> n * log(rss3/n) - n * log(rss2/n)

[1] -2.425431
```

This difference is larger than 2, but smaller than $\log(n) = 3.8918$, so BIC is not willing to pay the penalty, but AIC is.

Discussion Here is a summary of some key ideas.

1. The models selected by forward selection, backwards elimination, and stepwise regression might not be the same, even using the same model selection criterion.
2. In a forward selection or a backwards elimination procedure, BIC may result in fewer parameters in the model than AIC.
3. The forward selection, backward elimination, and stepwisse regression procedures are not guaranteed to find the best model according to the AIC or BIC criterion.
4. P-values in resultant models should be treated with more than usual caution, because they do not reflect the model selection process.
5. Generally, there may be several models that are highly similar in the quality of the fit.

A rich model with all of the variables can be made richer by considering interactions. You can try this example yourself and learn that the backward elimination procedure beginning with all main effects and two-way interactions produces a model with a very different set of variables than the analysis that assumes no interactions.

```
> step(lm(SAT ~ (ltakers + income + years + public + expend + rank)^2),
+     direction = "backward")
```