Inference for Regression Estimates

Rob Root

4/24/2018

Estimating y-values using a regression model

Review

Recall that the regression model assumes a collection of determined x values, x_i for $i = 1 \dots n$, and corresponding observations of random variables $Y_i = Y(x_i) \sim N(\beta_0 + \beta_1 x_i, \sigma)$. We have found unbiased estimators for all three parameters in this model, and they are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},\ \hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})x_i},\ \hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \sum(Y_i - \hat{Y}_i)^2 \Big/ (n-2)$$

Here, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the estimated mean for the variable Y_i. We have seen that $\hat{\beta}_1 \sim N(\beta_1, \sigma/\sqrt{S_{xx}})$, that $\hat{\beta}_0 \sim N(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}})$, and that $s \sim \sigma^2 \chi^2_{n-2}/(n-2)$.

Estimating the mean of $Y(x)$

We now want to consider how well we can estimate $E(Y(x)) = \beta_0 + \beta_1 x$. The book uses the notation $E(Y(x^*)) = \mu_{Y \cdot x^*}$ for this value. One might think that we can estimate this parameter *uniformly* in x, meaning that the variance in our estimate would not change as x changes. However, we will see that the variation in our estimates depends on the value of x. Our estimator has the obvoius form $\widehat{E(Y(x))} = \hat{\mu}_{Y \cdot x^*} = \hat{\beta}_0 + \hat{\beta}_1 x$ and from this we know that it is a normal random variable, since it is a linear combination of $\hat{\beta}_0$ and $\hat{\beta}_1$, both normal. Further, we know that it is an unbiased estimator, since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. That is, $E(\hat{\mu}_{Y \cdot x}) = \beta_0 + \beta_1 x$ To find its variance, we cannot simply say $V(\hat{\mu}_{Y \cdot x}) = V(\hat{\beta}_0) + V(\hat{\beta}_1)x^2$ because $\hat{\beta}_0$ and $\hat{\beta}_1$ are not independent. However, we can use $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ to create a sum of independent variables, recalling that we showed $\bar{Y}$ and $\hat{\beta}_1$ are independent in the notes from last lecture. (That's in the part we didn't get to in class.) We have

$$\begin{aligned}
\hat{\mu}_{Y \cdot x} &= \hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1 (x - \bar{x}) \\
V(\hat{\mu}_{Y \cdot x}) &= V(\bar{Y} + \hat{\beta}_1 (x - \bar{x})) = V(\bar{Y}) + V(\hat{\beta}_1)(x - \bar{x})^2 \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2 (x - \bar{x})^2}{S_{xx}}.
\end{aligned}$$

So, we have that $\hat{\beta}_0 + \hat{\beta}_1 x \sim N(\beta_0 + \beta_1 x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}})$. Notice that the standard deviation is smallest when $x = \bar{x}$, and that the variance grows with the square of the distance $|x - \bar{x}|$. So our estimate of the model line with the regression line is best when x is near $\bar{x}$, and gets worse the further away we are. It is

easy to see that we can standardize this normal variable, and by using s in place of σ create a t with $n-2$ degrees of freedom. The t random variable is

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 + \beta_1 x)}{S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} = \frac{\hat{\mu}_{Y\cdot x} - \mu_{Y\cdot x}}{S_{\mu_{Y\cdot x}}} \sim t_{n-2}.$$

So, for any x-value, we have a $1-\alpha$ confidence interval

$$\hat{\beta}_0 + \hat{\beta}_1 x - t_{\alpha/2,n-2} s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} < \mu_{Y\cdot x} < \hat{\beta}_0 + \hat{\beta}_1 x + t_{\alpha/2,n-2} s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}.$$

For the regression model to hold, all these variables $Y(x)$ need to be normal, so we can create a prediction interval, as we saw in §8.3 of the text. The radius of a prediction interval with $1-\alpha$ confidence is

$$t_{\alpha/2,n-2}\sqrt{s^2 + s^2_{\mu_{Y\cdot x}}} = t_{\alpha/2,n-2} s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}.$$

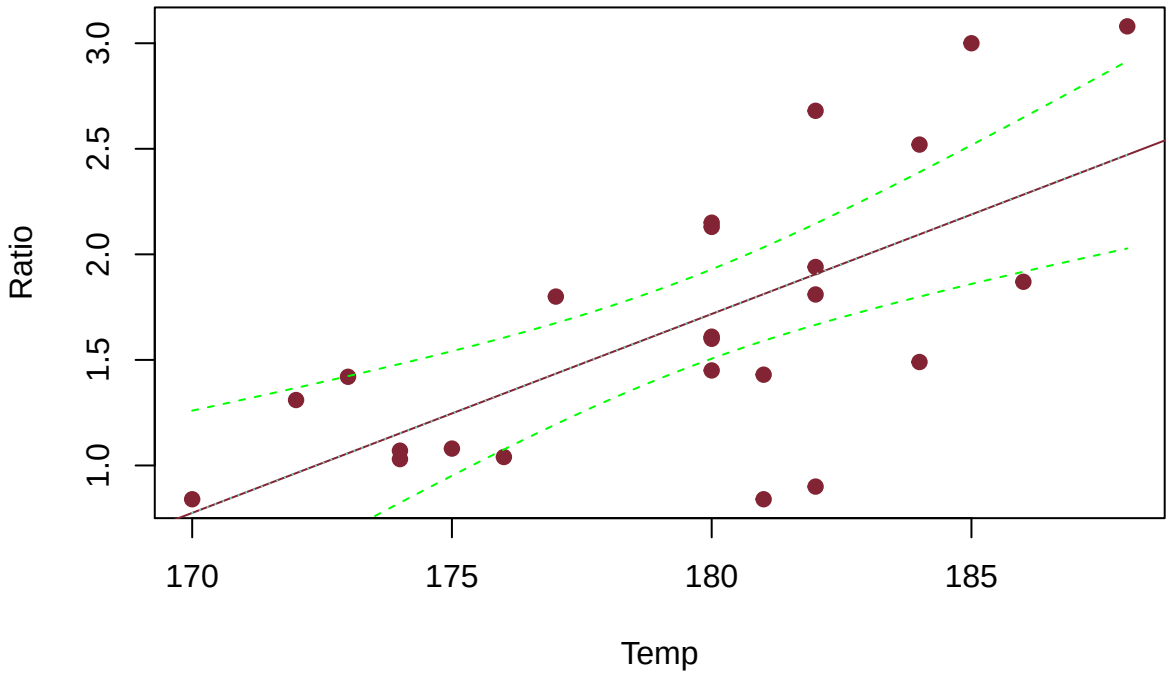
An Example

We can easily compute both the confidence and prediction intervals in R. Let's look again at the data from exercise 12.1, used in last class. We can easily replicate the work we did last time.

```
setwd("~/Documents/336 Spring 18/R/") #Change this to work on your computer
load("CH12/exe12001.RData")           #Note the data file is in a directory called CH12
attach('exe12-01')
reg<-lm(Ratio~Temp)
```

We have created the regression model, now we want to visualize the confidence interval. To do this we recreate the scatterplot and add the regression line, as in the last figure in the notes from last class. Then we create a list of one hundred x-values that fall in the range of the scatterplot. We then create the 95% confidence interval for each of those x-values using the `predict` command. Then we plot them.

```
plot(Ratio~Temp,main="Efficiency Ratio against Temperature",pch=19,
     col=rgb(0.51,0.14,0.2)) #Creates scatterplot
abline(reg,col=rgb(0.51,0.14,0.2))    #Adds regression line
xvals<-seq(min(Temp),max(Temp),length.out=100) #Creates list of 100 x values
yci<-predict(reg, newdata=data.frame(Temp=xvals), interval="confidence", level=0.95)
lines(xvals,yci[,"fit"],lty="dotted",col=rgb(0.49,0.86,0.8))
lines(xvals,yci[,"lwr"],lty="dashed",col="green")
lines(xvals,yci[,"upr"],lty="dashed",col="green")
```



Notice that the line of "fit" values falls right on the regression line, because that regression line was used to compute them. The upper ("upr") and lower ("lwr") limits of the confidence intervals fall on the bright green dashed lines. These form a hyperbola around the regression line, centered at $(\bar{x}, \bar{y})$. We can compute the interval for any particular x-value, for instance $x = 178$, using either the formula or the `predict` command:

```
beta0<-coefficients(reg)[1] #Beta0 from the lm
beta1<-coefficients(reg)[2] #Beta1 from the lm
yhat<-beta0 + beta1*178     #Fitted value for x = 178
s<-summary(reg)$sigma       #sample standard deviation from the lm
sxx<-23*var(Temp)           #S_xx
tcrit<-qt(1-0.025,22)       #critical t value
c(beta0, beta1, yhat, s, sxx, tcrit) #Check all the values
```

```
##  (Intercept)         Temp  (Intercept)
## -15.24496541   0.09423611   1.52906250   0.49724477 504.00000000
##
##   2.07387307
```

```
ciradius<-tcrit*s*(1/24 + (178 - mean(Temp))^2/sxx)^(1/2) #Radius of CI
myci<-c(yhat, yhat - ciradius, yhat + ciradius) #Compute CI
myci
```

```
## (Intercept) (Intercept) (Intercept)
##    1.529063    1.307575    1.750550
```

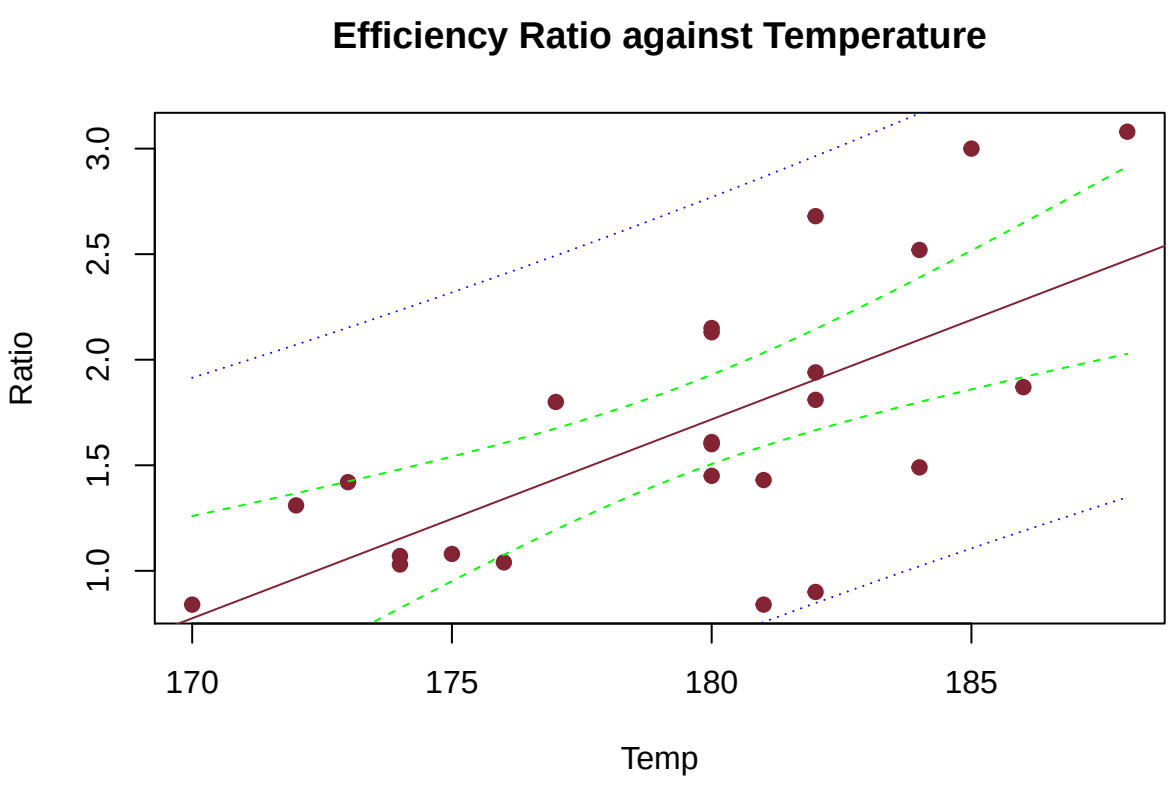
```
rci<-predict(reg, newdata=data.frame(Temp=c(178)), interval="confidence", level=0.95)
rci    #The easy way!
```

```
##        fit      lwr     upr
## 1 1.529063 1.307575 1.75055
```

Note that these intervals agree exactly, and that it is way easier to let R take care of the computations.

We can also get the prediction intervals for any x-value. Here is a scatterplot showing their limits with blue dotted lines.

```
plot(Ratio~Temp,main="Efficiency Ratio against Temperature",pch=19,
     col=rgb(0.51,0.14,0.2)) #Creates scatterplot
abline(reg,col=rgb(0.51,0.14,0.2))    #Adds regression line
lines(xvals,yci[,"lwr"],lty="dashed",col="green")
lines(xvals,yci[,"upr"],lty="dashed",col="green")
ypi<-predict(reg, newdata=data.frame(Temp=xvals), interval="prediction", level=0.95)
lines(xvals,ypi[,"lwr"],lty="dotted",col="blue")
lines(xvals,ypi[,"upr"],lty="dotted",col="blue")
```



Obviously the prediction intervals are much wider than the confidence intervals. Like the confidence intervals, they form a hyperbola around the regression line, but they look nearly linear in this view. We expect 95% of all observations to fall between them, so it isn't surprising that they include all 24 points in the data set. We can easily also create a single prediction interval:

```
piradius<-tcrit*s*(1 + 1/24 + (178 - mean(Temp))^2/sxx)^(1/2) #Radius of PI
mypi<-c(yhat, yhat - piradius, yhat + piradius) #Compute PI
mypi
```

```
## (Intercept) (Intercept) (Intercept)
##   1.5290625   0.4743225   2.5838025
```

```
rpi<-predict(reg, newdata=data.frame(Temp=c(178)), interval="prediction", level=0.95)
rpi    #The easy way!
```

```
##        fit       lwr      upr
## 1 1.529063 0.4743225 2.583803
```

Again, the agreement is perfect, though the upper limit is rounded a decimal place short in the interval resulting from the `predict` command.

Model checking for regression

It is natural to want to test whether the assumptions of the regression model are valid. Here are some reasonable questions to consider:

1. Are the means of the Y_i really linearly determined by the x values?
2. Are the Y_i really all normally distributed?
3. Are the standard deviations of the Y_i really all the same?
4. Is there any data that is having an undue effect on the regression model?

Whenever we do regression, we should be asking ourselves these questions. Our first recourse is always the scatterplot. Considering the first questions, if there is stark nonlinearity in the relationship between x and Y, we will be able to observe it in the scatterplot, for example. Considering the fourth question, we can spot strong outliers to a linear association or influential points by examining the scatterplot, as well. We should check the scatterplot for these problems before we even consider using a linear regression model.

Even after we have decided to try a linear regression model for our data, we should investigate how well the model fits the data. One of our basic tools is looking at the residuals; let's carefully consider their distribution next.

The distribution of residuals

We have determined the distribution of the fitted values, considered and random variables: $\hat{Y}(x) \sim N(\beta_0+\beta_1 x, \sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}})$, but we haven't considered the residuals. Obviously $Y_i-\hat{Y}_i$ is a linear combination of normals, and so is normal, and it is easy to see that $E(Y_i-\hat{Y}_i) = \beta_0-\beta_1 x_i-(\beta_0-\beta_1 x_i) = 0$, but the variance of a residual is harder. The fitting variable $\hat{Y}_i$ is not independent of Y_i, but we can still

determine the variance of their difference. We start by writing the residual as a linear combination of all the Y_js:

$$\begin{aligned} Y_i - \hat{Y}_i &= Y_i - (\bar{Y} + \hat{\beta}_1(x_i - \bar{x})) = Y_i - \frac{1}{n}\sum_j Y_j - \sum_j \frac{(x_j - \bar{x})}{S_{xx}} Y_j (x_i - \bar{x}) \\ &= Y_i - \sum_j \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} \right) Y_j \\ &= \sum_j \left[\delta_j^i - \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} \right) \right] Y_j \end{aligned}$$

Here δ_j^i is a *Kronecker delta*, meaning that it represents 1 when $j = i$ and 0 otherwise. It allows us to move our lone Y_i into our summation over j. Now we have written the residual as a linear combination of independent variables, so we can compute the variance,

$$\begin{aligned} V(Y_i - \hat{Y}_i) &= V\left\{ \sum_j \left[\delta_j^i - \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} \right) \right] Y_j \right\} \\ &= \sum_j \left[\delta_j^i - \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} \right) \right]^2 \sigma^2 \end{aligned}$$

Note that every j-term includes the quantity

$$\sigma^2 \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} \right)^2 = \sigma^2 \left(\frac{1}{n^2} + \frac{2(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} + \frac{(x_j - \bar{x})^2 (x_i - \bar{x})^2}{S_{xx}^2} \right).$$

We can sum this over j, remembering that the sum of deviations from a mean is always 0 and $S_{xx} = \sum_j (x_j - \bar{x})^2$. We get

$$\sum_j \sigma^2 \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} \right)^2 = \sigma^2 \left(\frac{1}{n} + 0 + \frac{(x_i - \bar{x})^2}{S_{xx}} \right).$$

This leaves only the terms that appear only in the term where $j = i$:

$$\left[1 - 2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right] \sigma^2$$

So, to find the variance, we sum these quantities to get

$$\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right] \sigma^2$$

as the variance of $Y_i - \hat{Y}_i$. Notice that this is a little *less* than the variance of Y_i on its own, which is just σ, demonstrating that the two variables are actually highly correlated. We can even see that since $S_{xx} = \sum_j (x_j - \bar{x})^2$, if $(x_i - \bar{x})^2$ makes up a large portion of that sum, it would be possible for the variance of the residual to be *much* less than the variance of the Y_i. This is part of what makes a point

influential in a regression; it is called a point's *leverage*. Points with outlying x-values have the leverage to draw the regression line close because their variance is constrained to be small.

We have established that residuals are distributed as $Y_i - \hat{Y}_i \sim N(0, \sigma\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}})$, so it is easy to see how to standardize them:

$$Z = \frac{Y_i - \hat{Y}_i}{\sigma\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \sim N(0, 1).$$

and more realistically

$$T = \frac{Y_i - \hat{Y}_i}{S\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

These standardized residuals are not as important in inference as they are in checking the aptness of the regression model, as we will see in the next section.

Diagnostic plots in R

The R language has elaborate built-in plots that allow us to assess validity of the assumptions of linear regression. There is no plot for each of the four questions posed to check a model.

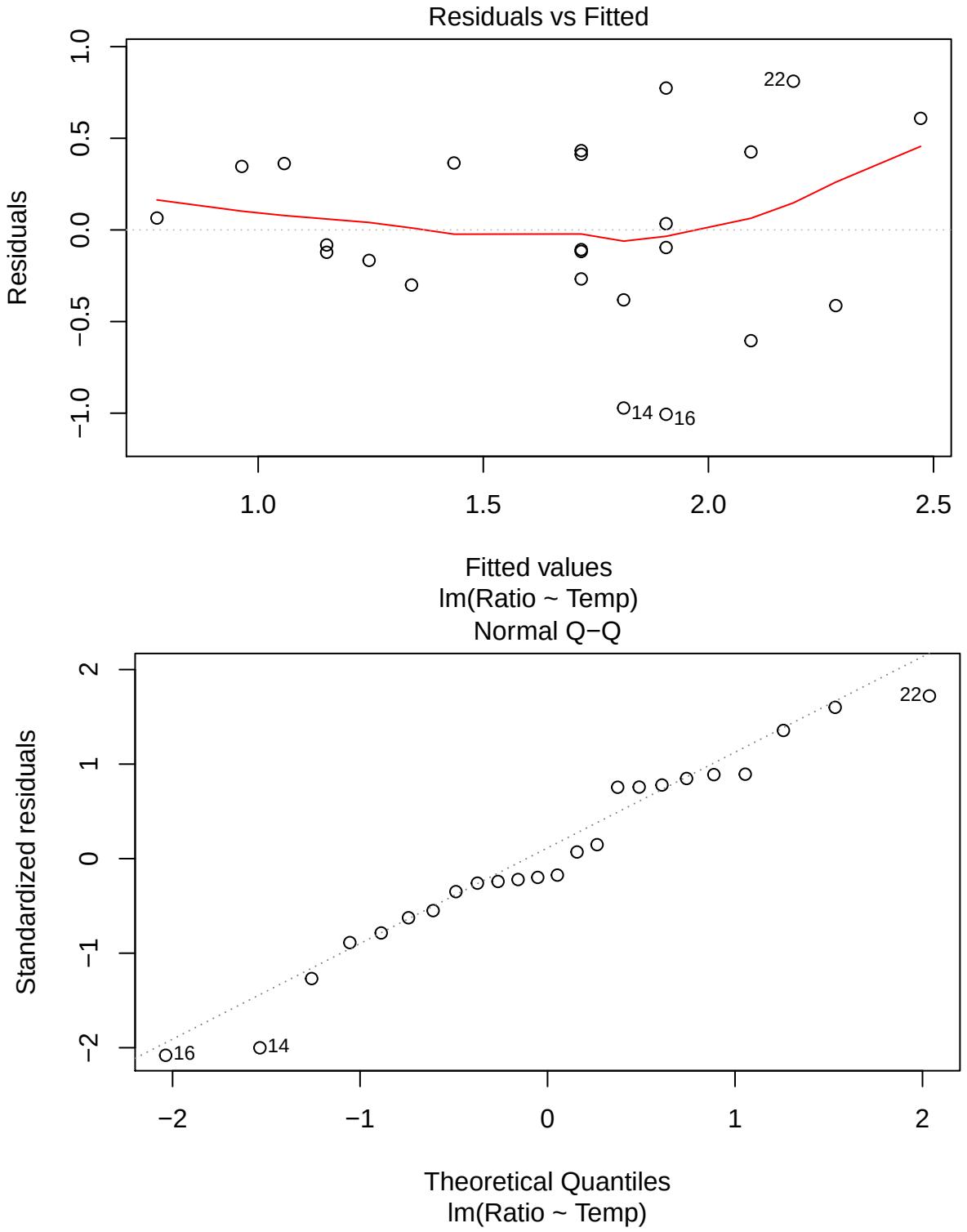
1. The first plot shows the (unstandardized) residuals against the fitted ($\hat{y}_i$) values. It allows us to spot any nonlinear pattern in the residuals that may have been hard to discern in the original scatterplot. The $\hat{y}_i$ values are used rather than the x_i because this works even if there is more than one explanatory (x) variable. For this example, the red fitting curve is very nearly horizontal, suggesting that the linear fit is an appropriate assumption.

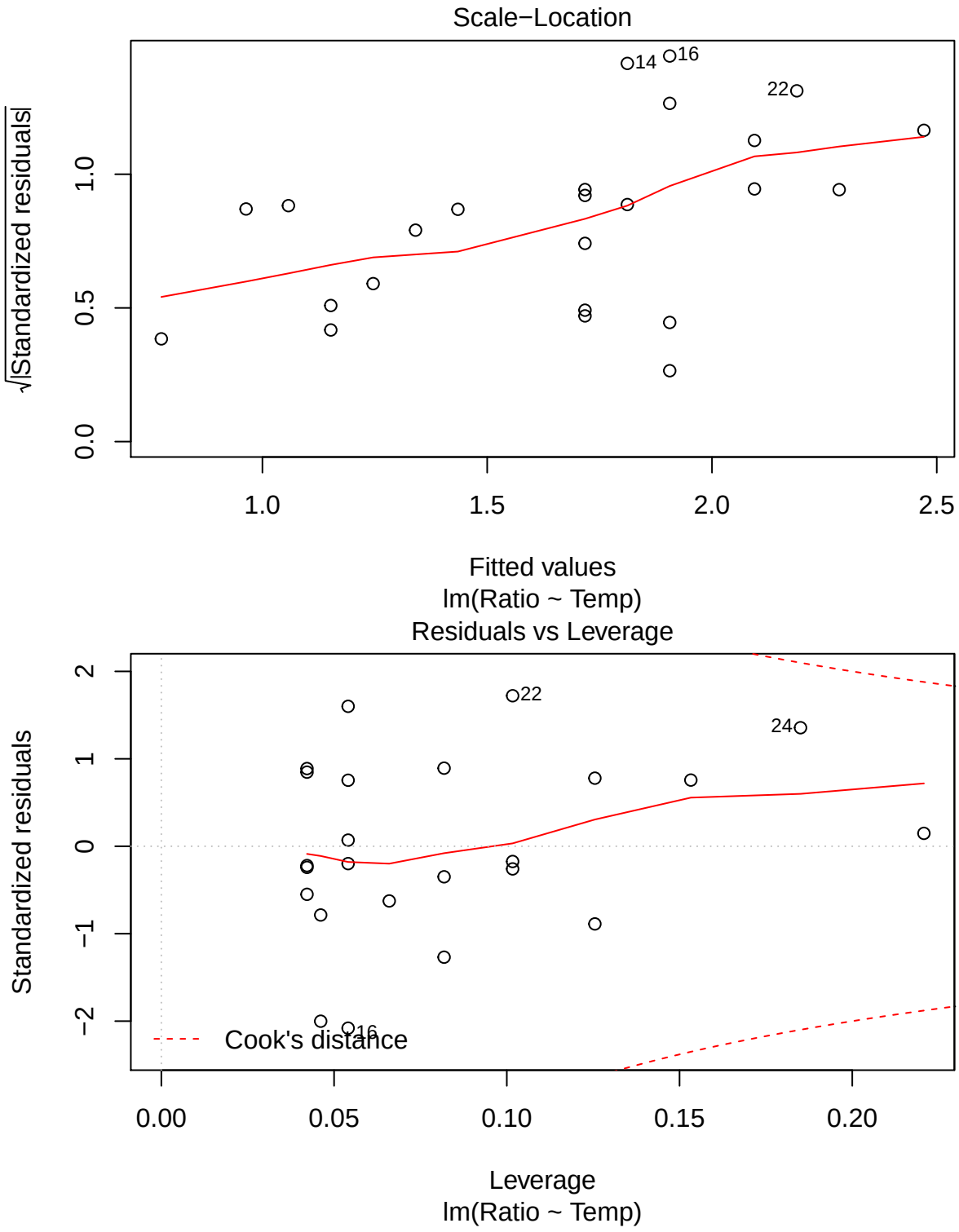
2. The second plot shows the standardized residuals on a normal quantile plot. Even though the standardized residuals are actually t distributed, they should fit a normal quantile plot fairly well, if the Y_is are really normally distributed. The residuals in this example all fall close to the line, suggesting that the assumption that the Y_is are normal is appropriate.

3. The third plot shows square root of the absolute values of the standardized residuals against the fitted values. The absolute value serves to make the residuals all nonnegative, while the square root serves to push them all toward magnitude 1. The resulting plot should not show any trend. If there is a clear pattern, that suggests that the variances of the Y_is are changing as x_i varies. Data with this difficulty is called *heteroskedastic*. The red fitting curve for this example has a marked increasing slope, suggesting that the data might be heteroskedastic.

4. The fourth plot shows standardized residuals against leverage. Leverage is just a measure of the distance from x_i to $\bar{x}$, the leverage associated with observations at x_i is $\frac{1}{n} + (x_i - \bar{x})^2/S_{xx}$. It varies symmetrically as the x_i value is more distant from $\bar{x}$ in either direction. Important features of this plot are the levels of Cook's distance that are shown by red contour lines. Cook's distance is an abstract measure of the degree to which removing any single data point alterings the model. Points with Cook's distance greater than 0.5 are considered at least marginally influential, while Cook's distance values of greater than 1 are considered influential. The this example there are no points that are even marginally influential.

```
plot(reg)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Ratio ~ Temp)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Ratio ~ Temp)
22
14
16



These diagnostic plots demonstrate that even what seems like a perfectly reasonable regression model has the potential for lurking problems. This model is usable, but seems to have heteroskedasticity. The easiest solution to this is to rerun the regression giving more weight to the data that seems to be less variable. In this case, since the curve in the third plot suggests that the magnitude of residuals roughly doubles between the left and right extremes, we could try a *weighted* regression, with the weight declining linearly from the

lowest x values to the highest. This should moderate or eliminate the heteroskedasticity. You are not expected to know this, or to know how to do it. If you want to know more, take the special topics course on regression.

There are competing standards for assessing the significance of Cook's distance, the quantity of interest in the last of the diagnostic plots. The plot is designed for the more lenient standard that the thresholds of $1/2$ and 1 make marginally influental and influential points. The more stringent standard is that Cook's distances larger than $4/n$ indicate points whose influence is significant. Since $n = 24$ in our example, this makes the threshold $4/24 = 1/6$ or roughly 0.167. This is not evident from the diagnostic plot, but of course we can use R to learn more.

```
cd<-cooks.distance(reg)
cd[cd>0.167]
```

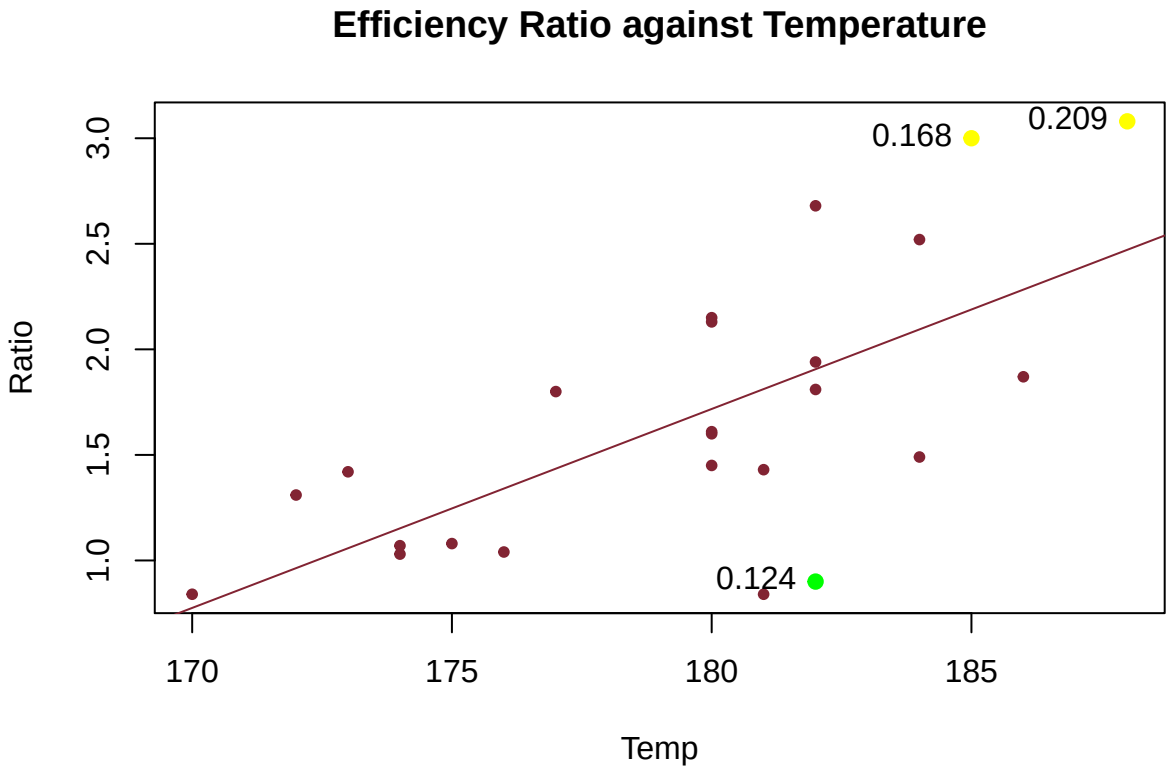
```
##        22        24
## 0.1677195 0.2086334
```

We can see that there are two data points with Cook's distances that are above this more stringent threshold. They are numbers 22 (barely) and 24 (safely). We can see that these points are labeled on the fourth diagnostic plot, along with point 16. These are the three data points with the largest Cook's distances, as we can see in the R output below. Let's look at these three data points on our original scatterplot.

```
sort(cd)[22:24]
```

```
##        16        22        24
## 0.1236651 0.1677195 0.2086334
```

```
plot('exe12-01', main="Efficiency Ratio against Temperature", pch=20, col=rgb(0.51,0.14,0.2))
abline(reg, col=rgb(0.51,0.14,0.2))
points(Temp[c(22,24)],Ratio[c(22,24)],pch=19,col="yellow")
points(Temp[16],Ratio[16],pch=19,col="green")
text('exe12-01'[c(16,22,24),],labels=round(cd[c(16,22,24)],3),pos=2)
```



We marked the two data points with Cook's distance over the threshold with yellow points, and the point with the next largest Cook's distance with a green point. We labeled the points with their Cook's distance values. Looking at the plot, we can see that the two most influential points are acting together to bring the regression to have a steeper positive slope. This combined influence is probably significant, and is worthy of mention in any thoughtful analysis of this regression model.

One response would be to go get more observations and bath temperatures above 185. There are only three, and two seem to be exerting unusual influence. With more data we will learn whether further observations match this pattern, or these are actually high outliers pulling the line toward them. It might be wise to do this before trying a new analysis using the weights suggested earlier to deal with the heteroskedasticity. The large positive residuals of observations 22 and 24 are probably a significant part of the pattern of increasing standardized residuals demonstrated in the third diagnostic plot.

Additional observations could offer confirmation of two possible situations. One outcome might be that the additional observations are just as spread out as datapoints 22, 23, and 24 are, demonstrating that the residuals for high temperatures really are larger, confirming heterskedasticity and requiring weighting or a more sophisticated remedy. Another might be that the additional observations are consistently high, close to those observed at data points 22 and 24, making data point 23 (between them) a low outlier. This would result in the regression line increasing its positive slope and displacing upward, resulting in the residuals at 22 and 24 shrinking, and so these points would become less influential. It would also reduce the heterskedasticity of the dataset, because residuals at lower x-values will also grow larger.

Finally, let's recall that the more lenient assessment of influence suggests that there is no need to concern ourselves with the potential of influential points in this data set. This more lenient approach would dictate that the influential points are of no concern.

More generally, It would not be unusual to find data like this modeled using linear regression in many client discipline journals with no mention of the deficiencies of the model. That does not mean that the weaknesses we have identified are inconsequential, just that many practioners apply regression without carefully considering the suitability of the model.