Inference and Prediction in Regression

Rob Root

4/22/2018

Elements of regression in R

Let's try using R to do some regression. Naturally we need some data to work with; let's use the data from the first exercise in the chapter.

```
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`)
```

This data offers the efficiency ratios for specimens of steel at different temperatures. (The efficiency ratio is the amount of phospate coating a specimen divided by the amount of steel lost creating the coating.) We can look at the individual variable; here are stemplots of the 24 observations of each.

```
stem(Ratio) #Stemplot for Efficiency Ratio
```

```
##
##   The decimal point is at the |
##
##   0 | 889
##   1 | 0011344
##   1 | 55668899
##   2 | 12
##   2 | 57
##   3 | 01
```

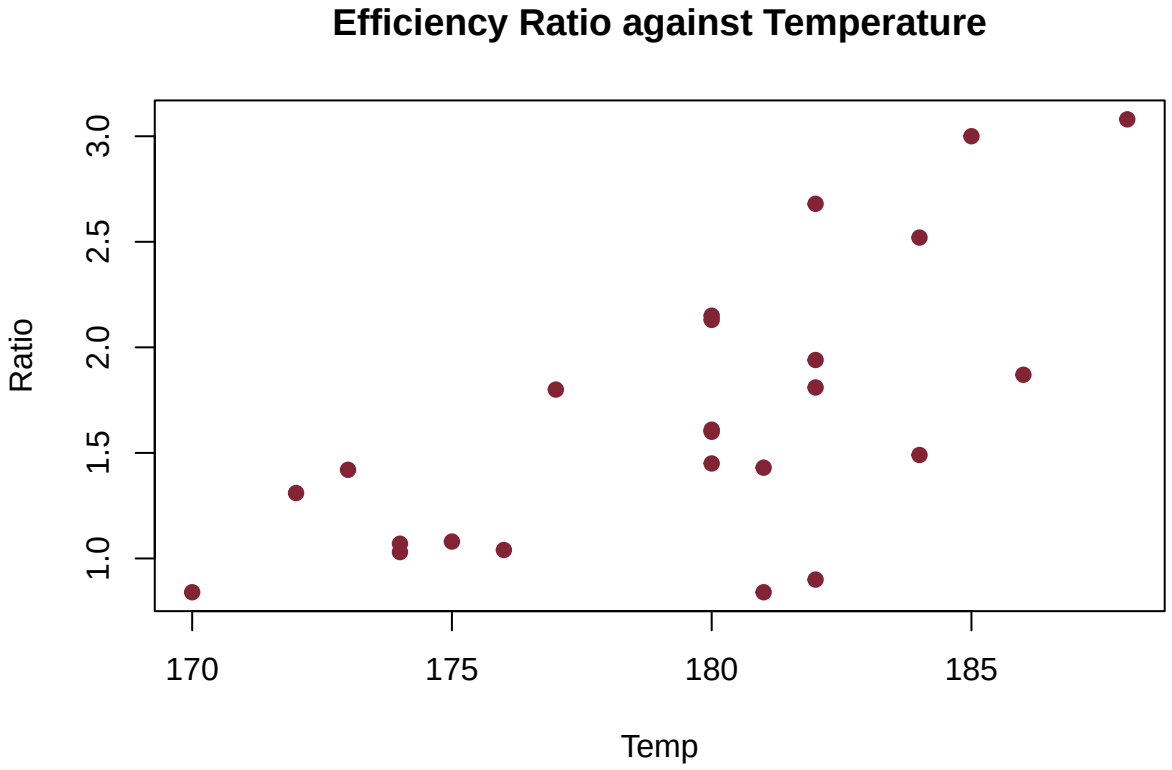
```
stem(Temp)  #Stemplot for Temperature
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   17 | 02344
##   17 | 567
##   18 | 0000011222244
##   18 | 568
```

Looking at these stemplots, we see that the efficiency ratios have a slight rightward skew, while the temperatures might be skewed leftward. It is worth noting that there are a large number of observations of the temperatures 180 and 182, while no other temperature of efficiency has more than 2 observations. (This will become significant later.) We can see how the two variables are related by plotting them in a scatterplot.

This shows that there is a moderately strong positive association between temperature and efficiency. That is, temperature is not a completely accurate predictor of efficiency, but it does clearly have an effect. Recall that "positive" means that large values of tmperature are associated with large values of efficiency, and small with small.

```
plot(Ratio~Temp,main="Efficiency Ratio against Temperature",pch=19,col=rgb(0.51,0.14,0.2))
```



Let's perform the regression and see what we can learn. First let's create the regression object and look at its summary.

```
reg<-lm(Ratio~Temp)
summary(reg)
```

```
##
## Call:
## lm(formula = Ratio ~ Temp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00601 -0.27580 -0.08906  0.37700  0.81128
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.24497    3.97705  -3.833 0.000905 ***
## Temp          0.09424    0.02215   4.255 0.000324 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4972 on 22 degrees of freedom
## Multiple R-squared:  0.4514, Adjusted R-squared:  0.4265
## F-statistic:  18.1 on 1 and 22 DF,  p-value: 0.0003239
```

The summary provides the call that generated the object. Next there are summary statistics for the residuals. Recall that if the estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ exactly matched their population parameters, the distributions of the residuals would be normal with mean 0 and variance σ^2. So we hope to see that the median residual is roughly 0, the first and third quartile are roughly negatives of one another, and the same for the maximum and minmum observations. Considering the relatively small size of this data set, those expectations hold for this data.

Next comes a table about $\hat{\beta}_0$ and $\hat{\beta}_1$. The first row of the table is labeled "(Intercept)," meaning y-intercept, or β_0 The column labeled "Estimate" has $\hat{\beta}_0$ for this data in it. The next column, "Std. Error," has $s_{\hat{\beta}_0}$. Next is the t statistic for the test of $H_0 : \beta_0 = 0$. The next column gives 2-sided p-values, and finally there is an indication of the significance of the assertion that $H_a : \beta_0 \neq 0$. In this case, the three asterisks (***) indicates that we can accept the alternative hypothesis as highly significant, with a p-value of less than 0.1%. (We will talk about this inference for β_0 at the end of class, on the blackboard.) The next row of the table is labeled for the x variable, in this case `Temp`. Here we have $\hat{\beta}_1$, $s_{\hat{\beta}_1}$, the model utility test t statistic, and its 2-sided p-value. Note that again this test suggests that the linear association with slope 0.09424 is highly significant, and is important in explaining the variation in efficiency. Note that the slope indicates that for every one degree increase in tmeperature, the average (or expected) efficiency ratio increases by 0.094.

After the key explaining the significance codes that can appear at the ends of each row come The residual standard error, which we have called $s = SSE/(n-2)$. The next line gives the coefficient of determination, or r^2. (If time permits, we will learn about the adjusted r^2 that is provided.) Finally, the F statistics, the square of the model utility t statistic is given, with its degrees of freedom and the p-value, which is the same as the one for β_1.

We can also perform the decompostion of variation for the regression using the command `aov` (for "analysis of variance").

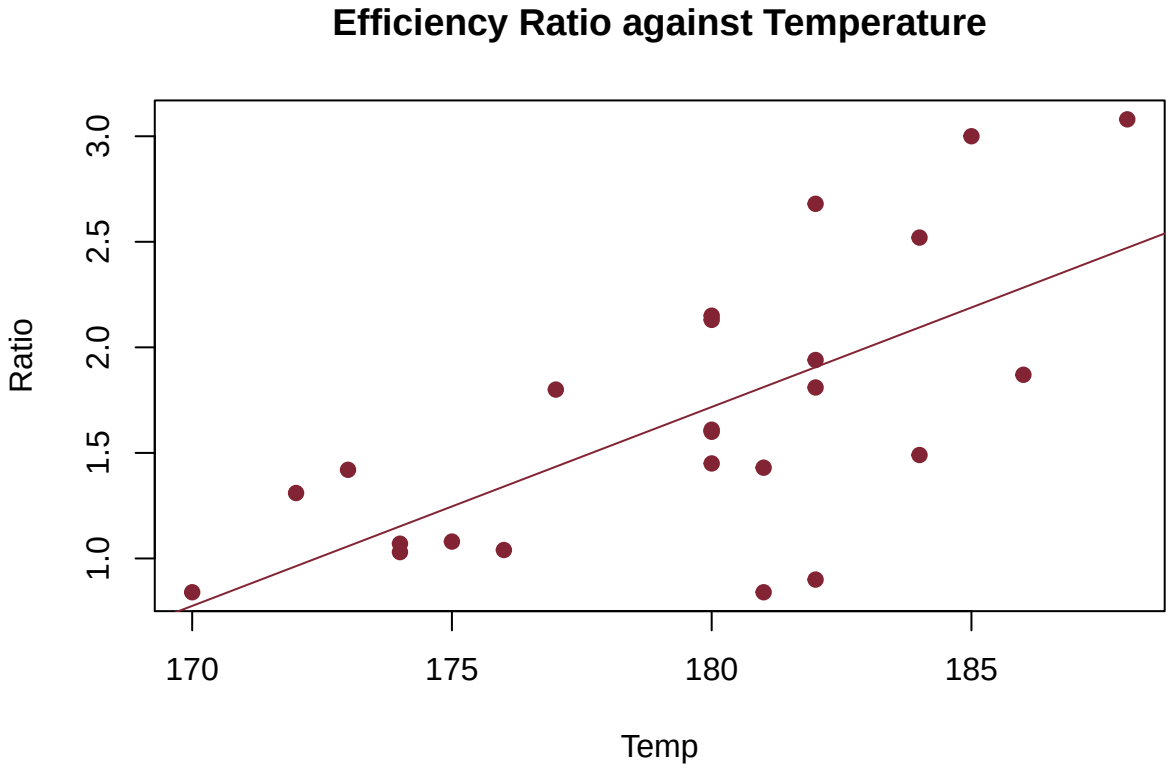
```
aov(reg)
```

```
## Call:
##    aov(formula = reg)
##
## Terms:
##                      Temp Residuals
## Sum of Squares  4.475744  5.439552
## Deg. of Freedom        1        22
##
## Residual standard error: 0.4972448
## Estimated effects may be unbalanced
```

Here, the row labeled "Sum of Squares" gives SSR followed by SSE, and the next row gives their respective degrees of freedom (as observations of χ^2 variables). The "Residual standard error" just $s = \sqrt{SSE/(n-2)}$, just as in the summary call. The final line, "Estimated effects may be unbalanced," is just R's way of saying that there are different numbers of observations at different temperatures. Recall that there are a few temperatures with many observations while most have only one or two. This places greater weight on the few temperatures (180 and 182) that have more observations, and so might "unbalance" our view of the underlying model.

Next, we might want to create a scatterplot that includes the regression line. Here are commands accomplishing that.

```
plot(Ratio~Temp,main="Efficiency Ratio against Temperature",pch=19,col=rgb(0.51,0.14,0.2))
abline(reg,col=rgb(0.51,0.14,0.2))
```



Note that the line does not pass through any point in the scatterplot, but recall that the sum of the squared vertical displacements (SSE) is minimized among all possible lines. Creating this scatterplot is a critically important part of the process of regression. You should always look at a regression model to see if it really is a good framework for the data at hand.

Inference for β_0

In the summary above we saw that there is a hypothesis test of $H_0 : \beta_0 = 0$ against a two-sided alternative, but we haven't investigated the distribution of $\hat{\beta}_0$, so let's do that now. We know that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$, where we are thinking of $\bar{Y}$ and $\bar{\beta}_1$ as random variables. The sample mean of the xs, $\bar{x}$ is not a random variable, because we consider the x observations are determined rather than random.

We have already seen that $\hat{\beta}_1 \sim N(\beta_1, \sigma/\sqrt{S_{xx}})$. We haven't looked at the distribution of $\bar{Y}$, but it is pretty easy to determine. Since $\bar{Y} = \frac{1}{n}\sum Y_i$, it is a linear combination of noraml random variables, so it is a normal random variable itself. Further we can compute its mean eithout much trouble.

$$\mu_{\bar{Y}} = E(\bar{Y}) = E\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n}\sum E(Y_i) = \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x}$$

and the sample mean of the Ys has expected value the y-coordinate of the point on the model line for $\bar{x}$. The variance is similarly easy. Recall that the Y_is are independent, so the variance of their sum is the sum of their variances. We get the same result as in the case when the Y_i are iid. Here are the computations:

$$V(\bar{Y}) = V\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n^2}\sum V(Y_i) = \frac{\sigma^2}{n}.$$

Thus the sample mean has distribution $\bar{Y} \sim N(\beta_0 + \beta_1\bar{x}, \sigma/\sqrt{n})$.

Now we can investigate the distribution of $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$. First note that, since $\bar{Y}$ and $\hat{\beta}_1$ are both normal, $\hat{\beta}_0$ is a linear combinations of normals and so normal itself. The expected value is easy to compute:

$$E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1\bar{x}) = E(\bar{Y}) - \bar{x}E(\hat{\beta}_1) = (\beta_0 + \beta_1\bar{x}) - \bar{x}\beta_1 = \beta_0.$$

So $\hat{\beta}_0$ is an unbiased estimator for β_0, just as we saw with β_1. Looking at the variance, it is tempting to simply say that $V(\hat{\beta}_0) = V(\bar{Y}) + \bar{x}^2 V(\hat{\beta}_1)$, but we don't know that $\bar{Y}$ and $\hat{\beta}_1$ are independent. (It is going to turn out that they are independent, but we need to show that.) There is a simple way to get around this problem: because both $\bar{Y}$ and $\hat{\beta}_1$ are linear combinations of the Y_i, so is $\hat{\beta}_0$. Recall from the April 20 lecture (on moodle)

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \sum \frac{x_i - \bar{x}}{S_{xx}} Y_i \text{ and } \bar{Y} = \sum \frac{1}{n} Y_i, \text{so, } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = \sum \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right) Y_i$$

From here we can compute the variance easily, because the Y_is are independent by assumption. We compute:

$$\begin{aligned}
V(\hat{\beta}_0) &= V\left(\sum \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right] Y_i \right) = \sum \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right)^2 V(Y_i) \\
&= \sum \left(\frac{1}{n^2} - \frac{2(x_i - \bar{x})\bar{x}}{nS_{xx}} + \frac{(x_i - \bar{x})^2 \bar{x}^2}{S_{xx}^2} \right) \sigma^2 \\
&= \left(\frac{1}{n} - \frac{2\bar{x}}{nS_{xx}} \sum (x_i - \bar{x}) + \frac{\bar{x}^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 \right) \sigma^2 \\
&= \left(\frac{1}{n} - \frac{2\bar{x}}{nS_{xx}} (0) + \frac{\bar{x}^2}{S_{xx}^2} S_{xx} \right) \sigma^2 = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}} \\
&= V(\bar{Y}) + \bar{x}^2 V(\hat{\beta}_1)
\end{aligned}$$

This provides us with two important pieces of information. First, we know the distribution $\hat{\beta}_0 \sim N(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}})$. Second we found that the variance of $\hat{\beta}_0$ is what we found when we assumed that $\bar{Y}$ and $\hat{\beta}_1$ were independent. This means that these variables are uncorrelated, which does not prove independence in general, but uncorrelated normal random variables are independent. So, we have demonstrated that $\bar{Y}$ and $\hat{\beta}_1$ are independent as well.

This is just a sidelight to our main investigation. We have determined the distribution of $\hat{\beta}_0$, and we can use that to perform inference. We have that

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim N(0,1)$$

and more importantly,

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2} \text{ where } S = \frac{SSE}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}.$$

This allows us to give a $1 - \alpha$ confidence interval for β_0:

$$\hat{\beta}_0 - t_{\alpha/2, n-2} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} < \beta_0 < \hat{\beta}_0 + t_{\alpha/2, n-2} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

It also allows use to test a hypothesis $H_0 : \beta_0 = \beta_{00}$ by using the t statistic

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}.$$

The p-value will depend on which alternative is being considered. In specific, the texttt{summary} command used above is testing $H_0 : \beta_0 = 0$ against a two-sided alternative.