Stat 145 – Notes on Inference for Regression

Thomas Scofield

April 26, 2021

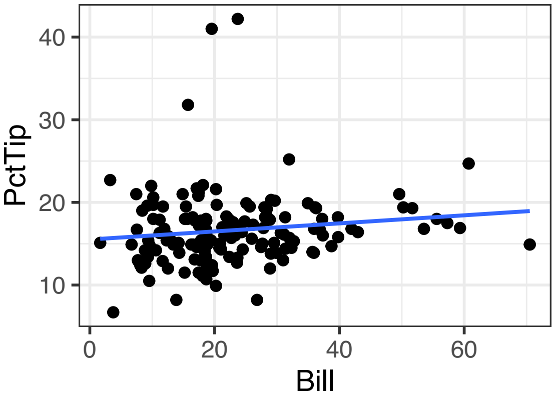
A summary of work to date: Inference for regression without randomization

In the context of **RestaurantTips** data, we have looked for an association between `Bill` (explanatory) and `PctTip` (response).

```
head(RestaurantTips)
```

```
##    Bill   Tip Credit Guests Day Server PctTip CreditCard
## 1 23.70 10.00      n      2 Fri      A   42.2         No
## 2 36.11  7.00      n      3 Fri      B   19.4         No
## 3 31.99  5.01      y      2 Fri      A   15.7        Yes
## 4 17.39  3.61      y      2 Fri      B   20.8        Yes
## 5 15.41  3.00      n      2 Fri      B   19.5         No
## 6 18.62  2.50      n      2 Fri      A   13.4         No
```

```
gf_point(PctTip ~ Bill, data=RestaurantTips) %>% gf_lm()
```



```
lm(PctTip ~ Bill, data=RestaurantTips)
```

```
##
## Call:
## lm(formula = PctTip ~ Bill, data = RestaurantTips)
##
## Coefficients:
## (Intercept)         Bill
##    15.50965      0.04881
```

Seeing that a linear relationship may be meaningful, we might charge into the **model utility test**

$$\mathbf{H}_0 : \beta_1 = 0 \qquad \text{vs.} \qquad \mathbf{H}_a : \beta_1 \neq 0.$$

The data produces the following ANOVA table.

Q: What is R^2 (called the coefficient of determination)

$$= \frac{SSModel}{SSTotal} = \frac{54.94}{54.94+2946.13}$$

$$= 0.018$$

```
anova(lm(PctTip ~ Bill, data=RestaurantTips))
```

```
## Analysis of Variance Table
##
## Response: PctTip
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Bill        1   54.94  54.936  2.8902 0.09112 .
## Residuals 155 2946.13  19.007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The corresponding F-statistic is 2.8902. In so far as we trust an F-distribution to give us the corresponding P-value, we have

```
1 - pf(2.8902, df1=1, df2=155)
```

```
## [1] 0.09112558
```

a number that R reported to us on the ANOVA table.

R gives us some diagnostic plots to help decide how concerned we should be about the use of theoretical tools (such as an F-distribution) in drawing inferences. These are all available through the command

```
plot( lm(PctTip ~ Bill, data=RestaurantTips) )
```

If you do it this way, you will have to hit to see the four plots one-by-one. If you prefer, then with the addition of the `par(...)` command, which instructs the plotter to use a 2-by-2 grid layout, you can see all four at once:

```
par(mfrow=c(2,2))
plot( lm(PctTip ~ Bill, data=RestaurantTips) )
```

A nice explanation of what to look for in these four plots is found at

https://data.library.virginia.edu/diagnostic-plots/

Model Utility test is a type of Inference on regression

Model Utility

$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

$H_0: \rho = 0$ $H_a: \rho \neq 0$

2nd go

Model Utility Test without randomization: Simple Linear Regression model

People have been conducting the Model Utility Test for a long while, since before computers were on every desktop, before it was feasible to generate randomization distributions, when only hand calculations were possible. Upon request, R will generate the kind of results those hand calculations produced. Naturally the `lm()` command is used, but so is `summary()`. In the case of the *PctTip-and-Bill* data, the request goes like this:

```
summary( lm(PctTip ~ Bill, data=RestaurantTips) )
```

anova(lm(- - -))

plot (lm(- - -))

```
##
## Call:
## lm(formula = PctTip ~ Bill, data = RestaurantTips)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9927 -2.3096 -0.6455  1.4679 25.5335
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.50965    0.73956   20.97   <2e-16 ***
## Bill         0.04881    0.02871    1.70   0.0911 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.36 on 155 degrees of freedom
## Multiple R-squared:  0.01831,    Adjusted R-squared:  0.01197
## F-statistic:  2.89 on 1 and 155 DF,  p-value: 0.09112
```

b_0

SE_{b_1}

b_1

Standardized test stats

$$t = \frac{b_1 - 0}{SE_{b_1}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0.04881}{0.02871} = 1.7$$

$[1 - pt(1.7, df=155)]*2$

$= 1 - pf(2.8902, 1, 155)$

done above

Notice the result contains information about both coefficients. The `(Intercept)` row says

> Our point estimate b_0 for β_0 is 15.50965, has $\mathrm{SE}_{b_0} = 0.73956$, and standardized t-score $t = 20.97$.

The other row, the one about *slope*, is always of greater interest to us. It says

> Our point estimate b_1 for β_1 is 0.04881, has $\mathrm{SE}_{b_1} = 0.02871$, and standardized t-score $t = 1.70$.

We see that the process our forbears devised in lieu of randomization led them to deal, again, with t-distributions. Specifically, since in the Model Utility Test we hypothesize that $\beta_1 = 0$, standardizing leads to the reported t-score:

$$t = \frac{0.04881 - 0}{0.02871} = 1.700.$$

To get the resulting P-value the way they (and this command) did, we find how often a t-score is as extreme or more so than this one—i.e., compute the tail area and double it. We get the tail area using a t-distribution, but with how many degrees of freedom? If there are n cases in the dataset, regression computes degrees of freedom in this way:

$$\text{df} = n - 1 - (\text{number of predictor variables}).$$

What is implied here is that it is possible to consider more than 1 explanatory/predictor variable. When we do, it is called **multiple regression**, a topic we will not cover in this course. Since we are considering only 1 predictor variable, the number of degrees of freedom is $\text{df} = n - 2$. So, in the case of *PctTip-as-predicted-by-Bill* data, which has $n = 157$ cases, R determined its P-value above by doing what the following command does:

```
2 * (1 - pt(1.7, df=155))
```

```
## [1] 0.09113667
```

This standard error SE_{b_1} can be used the way we have used standard errors in the past, not only in the calculation of a standardized t-statistic above, but also in the **construction of a confidence interval for the true slope** β_1:

$$\text{(point est.)} \pm t^* \text{ SE}, \qquad \text{or} \qquad b_1 \pm t^* \text{ SE}_{b_1}.$$

You choose the critical value t^* as in Chapter 6, but now using $n-2$ degrees of freedom. So, to get a 92% CI for β_1 for the *PctTip-as-predicted-by-bill* data, with sample size

```
nrow(RestaurantTips)
```

```
## [1] 157
```

we get our critical value

```
tstar <- qt(0.96, df=155); tstar
```

```
## [1] 1.76224
```

and our confidence interval

fancy way of obtaining the point estimate b_1.

$= SE_{b_1}$

```
lm(PctTip ~ Bill, data=RestaurantTips)$coefficients[2] + c(-1,1) * tstar * 0.02871
```

```
## [1] -0.001781462  0.099406336
```

As we have seen before, there is a connection between the P-value of a Model Utility Test and this confidence interval for β_1. Since the null value, 0, is inside a 92% CI, we would have expected correspondingly that the P-value of the Model Utility Test to be *larger* than 0.08, and above we found it to be 0.091, confirming that.

Question:

Our forbears also developed the formula for the standardized t-statistic of the sample correlation r to be

$$\boxed{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},}$$

which also has $n-2$ degrees of freedom when there is one predictor variable. What is the resulting P-value, and what hypotheses would we be testing?

Caution about using results based on a theoretical t-distribution: As in the previous chapters, beginning with Chapter 5, whenever we have turned to a theoretical distribution (a normal distribution, a t distribution, a chi-square distribution, or an F distribution) to compute a P-value, there have been conditions which validate the approach and, in the absence of such, leave us in some doubt about the conclusions. The same is true with the results above.

Conditions for the **Simple Linear Regression Model**. What has been assumed, and validates the approach, may be described as follows. Many different observed Y-values are possible for any fixed X-value, but

- they are normally distributed with mean $\mu_Y(X) = \beta_0 + \beta_1 X$, and
- the standard deviation is σ, a number that doesn't change with X.

The output from `summary( lm( ...) )` above includes an estimate for σ. It is the number reported as `Residual standard error`. For the *PctTip-as-predicted-by-Bill* data, the estimate for σ is 4.36

Some pictures from the textbook:

ANOVA for regression : Further details w/ examples of computations in R (OK to skim)

We can consider regression whenever we have bivariate quantitative data, a collection of n points which we can label

$$(x_1, y_1),\ \ (x_2, y_2),\ \ (x_3, y_3),\ \ \ldots\ \ (x_n, y_n).$$

Since Y is treated as a response variable, we call the y_1, y_2, ..., y_n **observed** values.

One can ignore the x-values and calculate things like the mean response-value

$$\overline{y} \;=\; \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \;=\; \frac{1}{n}\sum y_i,$$

or the **variance** (the square of the standard deviation)

$$s_y^2 \;=\; \frac{1}{n-1}\sum(y_i - \overline{y})^2.$$

And, since standard deviation and variance both measure how much variation there is between observed values, so does the quantity $(n-1)s_y^2$, which we again call SST, or

$$SSTotal \;=\; (n-1)s_y^2 \;=\; \sum(y_i - \overline{y})^2.$$

We can describe this relationship between SST and variance by saying that variance is SST divided by number of degrees of freedom $n - 1$.

The computed regression line with intercept b_0 and slope b_1 has been used by us back in Chapter 2 to compute **predicted**, also known as **fitted**, values. We use a "hat" to distinguish these values from the observed ones:

$$\widehat{y}_1 = b_0 + b_1 x_1,\ \ \widehat{y}_2 = b_0 + b_1 x_2,\ \ \ldots,\ \ \widehat{y}_n = b_0 + b_1 x_n.$$

It is perhaps not so surprising that these fitted values have the same mean as the observed ones—that is

$$\frac{1}{n}(\widehat{y}_1 + \widehat{y}_2 + \cdots + \widehat{y}_n) \qquad \text{also equals } \ \overline{y}.$$

So, if we did the same type of sum-of-squares calculation as above, but using fitted values instead of observed ones, that would give a measure of the variability among fitted values. We will call that SSM, or

$$SSModel \;=\; \sum(\widehat{y}_i - \overline{y})^2.$$

Our fitted values all lie on the line, and the difference between an observed value y_i and its predicted value $\widehat{y}_i$ at x_i,

$$\epsilon_i \;=\; y_i - \widehat{y}_i,$$

is what we call a **residual**. You may remember that, in finding the best-fit line, we chose, out of all possible lines, the one that had the smallest possible sum-of-squares-of-residuals, sometimes called $SSResid$, or SSE (since *error* and *residual* are synonyms):

$$SSE \;=\; \sum(y_i - \widehat{y}_i)^2.$$

As when we defined similar quantities in Chapter 8, the relationship between them is

$$SSTotal = SSModel + SSResid, \qquad \text{or} \qquad SST = SSM + SSE.$$

When SSE is small in comparison with SST, that is indicative of a strong linear relationship between the variables; the line does a good job of explaining the variation in observed values. A good measure of how well the variability in response values Y is explained by the linear model $b_0 + b_1 X$ is the ratio

$$R^2 \;=\; \frac{SSM}{SST} \;=\; \frac{SST - SSE}{SST},$$

known as the **coefficient of determination**. It might have been better to use a lower-case r, and call it r^2, since the coefficient of determination is equal to the square of the correlation.

Example: `InkjetPrinters`. In the text, the Locks propose using `PPM`, the number of pages a printer can turn out per minute, to explain `Price`. The first few rows of the raw data are as follows.

```
head(InkjetPrinters)
```

```
##                                   Model PPM PhotoTime Price CostBW CostColor
## 1 HP Photosmart Pro 8500A e-All-in-One 3.9        67   300    1.6       7.2
## 2                    Canon Pixma MX882 2.9        63   199    5.2      13.4
## 3                  Lexmark Impact S305 2.7        43    79    6.9       9.0
## 4               Lexmark Interpret S405 2.9        42   129    4.9      13.9
## 5                  Epson Workforce 520 2.4       170    70    4.9      14.4
## 6                  Brother MFC-J6910DW 4.1       143   348    1.7       7.9
```

A scatterplot makes a linear relationship appear reasonable. The black points represent the data, while the purple points, lying on the line, are the *fits*. By storing the result of the `lm()` command, R can be asked to provide the fitted values (in the same order as the original data)

```
lmRes <- lm(Price ~ PPM, data=InkjetPrinters)
lmRes$fitted
```

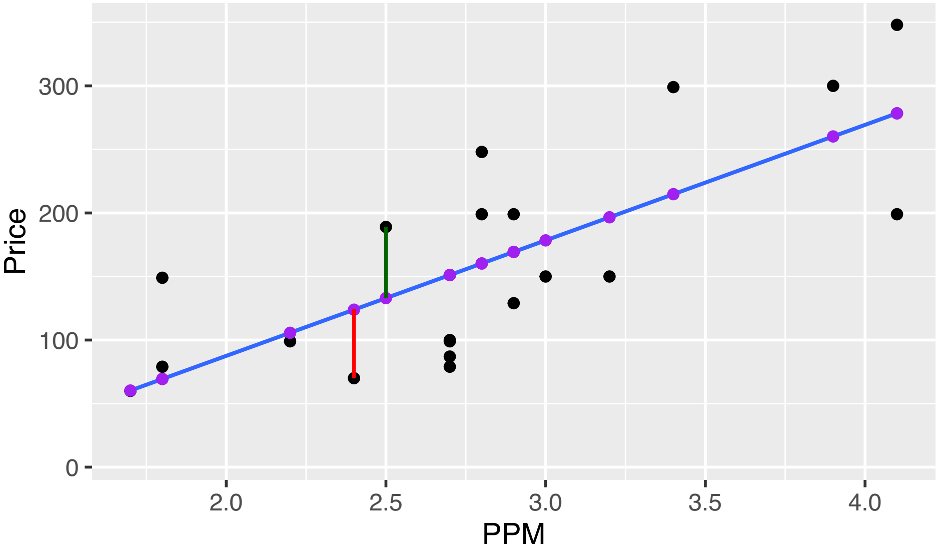
```
##         1         2         3         4         5         6         7         8 
## 260.20270 169.32464 151.14902 169.32464 123.88560 278.37832 214.76367 160.23683 
##         9        10        11        12        13        14        15        16 
## 178.41244 196.58806 151.14902 151.14902 105.70999 132.97341 151.14902  60.27095 
##        17        18        19        20 
## 160.23683  69.35876  69.35876 278.37832
```

as well as the residuals

```
lmRes$residuals
```

```
##           1           2           3           4           5           6 
##  39.7972965  29.6753642 -72.1490222 -40.3246358 -53.8856019  69.6216830 
##           7           8           9          10          11          12 
##  84.2363304  87.7631710 -28.4124425 -46.5880561 -64.1490222 -51.1490222 
##          13          14          15          16          17          18 
##  -6.7099884  56.0265913 -52.1490222  -0.2709545  38.7631710  79.6412387 
##          19          20 
##   9.6412387 -79.3783170
```

The vertical green line is from the 14th observed value, the observed price of \$189 for a Dell V715 w inkjet printer, down to its fitted price of \$132.97, a positive residual of $189 - 132.97 = 56.03$. The red vertical line is from the 5th observed value, the price of \$70 for an Epson Workforce 520 up to its fitted price of \$123.89, a negative residual of $70 - 123.89 = -53.89$.



Take a look at the output from the following commands applied to this data. First the summary from `lm()` (recall that `lmResult` stores output from `lm()`).

```
summary(lmRes)
```

```
##
## Call:
## lm(formula = Price ~ PPM, data = InkjetPrinters)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.38 -51.40  -3.49  43.85  87.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -94.22      56.40  -1.671 0.112086
## PPM            90.88      19.49   4.663 0.000193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.55 on 18 degrees of freedom
## Multiple R-squared:  0.5471, Adjusted R-squared:  0.522
## F-statistic: 21.75 on 1 and 18 DF,  p-value: 0.0001934
```

See there is a number reported at the bottom with the label `Multiple R-squared`. That is the coefficient of determination, R^2, we defined above. It says about 55% of the variability in sampled printer prices is "explained" by the variable `PPM` through the linear model

$$\widehat{\text{Price}} = -94.22 + 90.88(\text{PPM}).$$

Next look at the correlation:

```
cor(Price ~ PPM, data=InkjetPrinters)
```

```
## [1] 0.7396862
```

Squaring this correlation

$$(0.7396862)^2 \doteq 0.5471,$$

yields the same number as `Multiple R-squared` reported above.

Now look at an ANOVA table:

```
anova(lmRes)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## PPM        1  74540   74540  21.747 0.0001934 ***
## Residuals 18  61697    3428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The reported sum-of-squares values are $SSM = 74540$ and $SSE = 61697$, which means $SST = 74540+61697 = 136237$. We defined R^2 to be the ratio

$$\frac{SSM}{SST} = \frac{74540}{136237} = 0.5471,$$

again matching `Multiple R-squared`.

Question

Consider the command

```
sum( lmRes$residuals^2 )
```

```
## [1] 61696.79
```

Can you guess what number this will give and find its value using the ANOVA table? Run the command and check that you are correct.

Looking more carefully at this ANOVA table, we see that, out of the $n = 20$ cases (printers) in the **InkjetPrinters** dataset, $n - 2 = 18$ degrees of freedom have been "assigned" to the `Residuals`, and 1 degree of freedom to `PPM`, for a total of $18 + 1 = 19 = n - 1$. In simple linear regression (i.e., regression with just 1 predictor variable), the number of degrees of freedom on the residual row is always $n - 2$. The calculations of quantities such as MSM (we called it MSG in Chapter 8), MSE and F which appear in the ANOVA table are done exactly as in 1-way ANOVA:

$$MSModel = \frac{SSModel}{1}, \qquad MSE = \frac{SSE}{n-2}, \qquad \text{and} \qquad F = \frac{MSM}{MSE},$$

and the resulting P-value, obtained with the command like

```
1 - pf(fstatistic, df1=1, df2=n-2)
```

is exactly the same as P-value from the Model Utility Test, representing another way to compute it.

Prediction and confidence intervals

Done on another day

If the conditions for the simple linear model are met, and if we have rejected the null hypothesis in the Model Utility Test in favor of the alternative, that the explanatory variable has some usefulness as a predictor of values of the response variable, it is typical to see the model used that way. There are two sorts of *prediction*-type questions we might ask.