Lab 02 - Simple Linear Regression and Confidence Intervals

PSC-103B

Marwin Carmo

2025-01-16

Read in Data

For today's class, we're going to use a fake dataset looking at the relation between temperature (in Fahrenheit), ice cream sales, and pool accidents.

```r
icecream <- read.csv("https://shorturl.at/qdZlt")
icecream
```

```
##    Temperature Sales      Pool
## 1        57.56   215 1.6101279
## 2        61.52   325 0.0000997
## 3        53.42   185 0.1348124
## 4        59.36   332 3.2882103
## 5        65.30   406 1.9550195
## 6        71.78   522 0.9825883
## 7        66.92   412 1.9631966
## 8        77.18   614 1.1486187
## 9        74.12   544 0.5999595
## 10       64.58   421 0.4459096
## 11       72.68   445 2.0538231
## 12       62.96   408 1.7382948
```

Simple Linear Regression

In our discussion of covariance and correlation, we've only talked about whether the two variables are associated with each other – asking if there's an association between temperature and sales, which is the same as asking if there's an association between sales and temperature. None of this looks at whether one variable has an effect on the other variable.

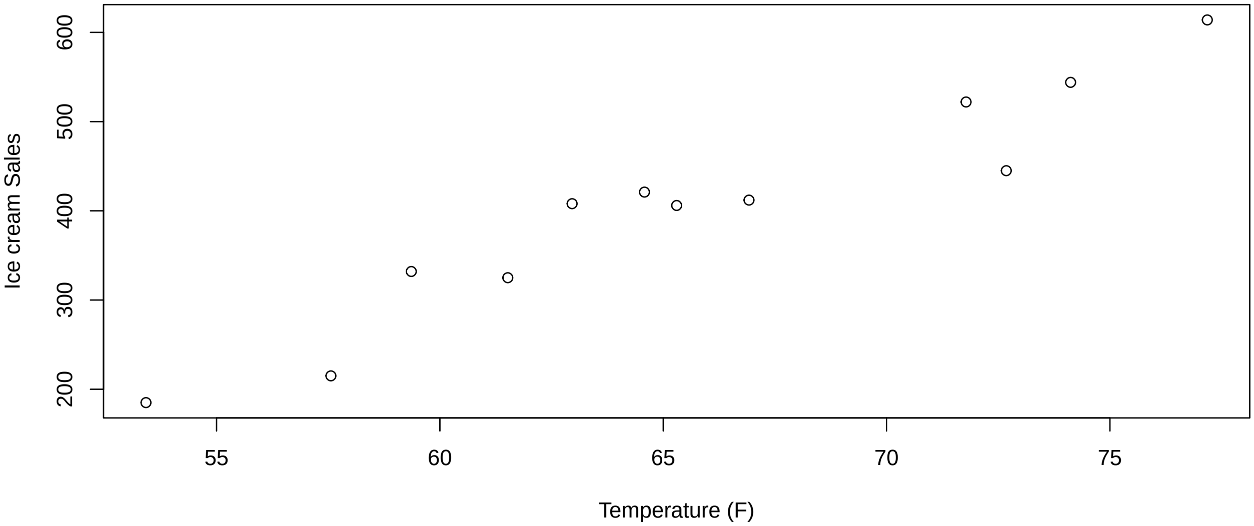
That's where linear regression comes in – now we're looking at the relation between a **dependent variable** and at least one **predictor variable**, so we've moved to the context of the predictor variable having an effect on the dependent variable.

Today's lab is going to focus on one predictor variable, which is called simple linear regression. Next week will be multiple regression, when we have at least 2 predictor variables.

For example, let us take our example of temperature and ice cream sales and see how temperature *affects* ice cream sales.

Let's plot our data again:

```r
plot(icecream$Temperature, icecream$Sales, xlab = "Temperature (F)", ylab = "Ice cream Sales")
```



What is the equation for the regression line?

$$\hat{Y}_i = b_0 + b_1 X_{1i}$$

This gives us the predicted value of our outcome for a given value of X. b_0 represents the intercept, which is the value of the outcome when the predictor is 0. b_1 represents the slope – the expected amount of change in our outcome when our predictor increases by 1 unit.

We can calculate estimates of these values using their formulas:

$$b_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$
$$b_0 = \bar{Y} - b_1 \bar{X}$$

```
b1_estimate <- cov(icecream$Temperature, icecream$Sales)/var(icecream$Temperature)
b0_estimate <- mean(icecream$Sales) - b1_estimate * mean(icecream$Temperature)

b1_estimate
```

```
## [1] 16.71548
```

```
b0_estimate
```

```
## [1] -694.3695
```

Let's write out our regression line:

$$Sales_i = -694.37 + 16.72 \times Temperature_i$$

Now we can also add this regression line to our plot:

```
plot(icecream$Temperature, icecream$Sales, xlab = "Temperature (F)",
     ylab = "Ice cream Sales")
abline(a = b0_estimate, b = b1_estimate, col = "red")
```

```
# abline() is a function to add a straight line to a plot
```

Of course, rather than calculating the values by hand, we can simply do it in R.

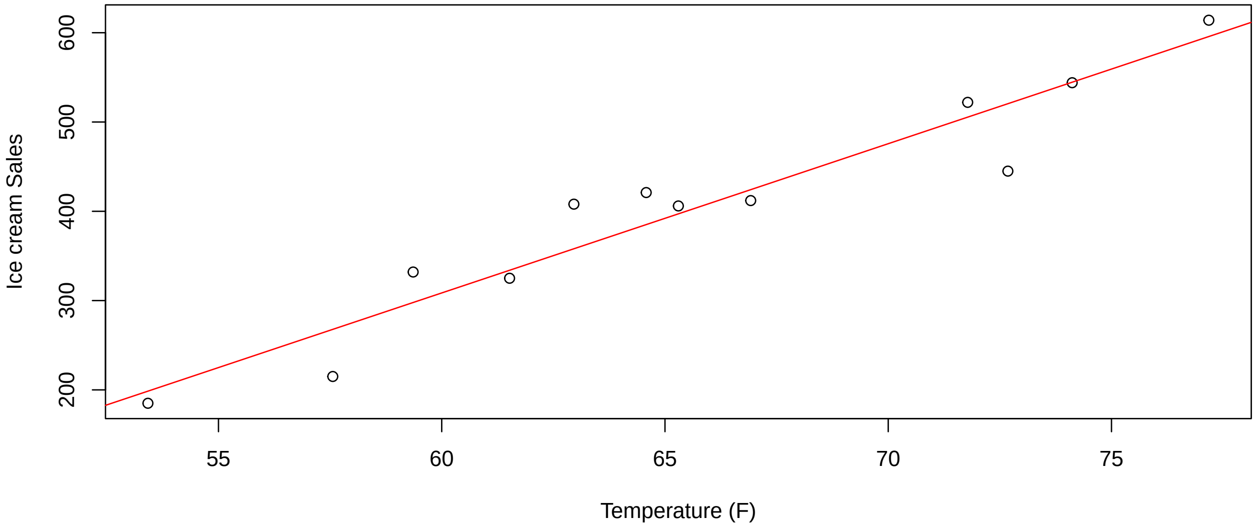


Figure 1: Simple linear regression of Ice cream sales predicted by temperature.

We use the `lm()` function (lm stands for “linear model”) And the function arguments are as follows: `lm(dependent_variable ~ independent_variable, data = data_name)`. This might look familiar - it’s pretty similar to how we conducted our t-tests in R!

```
simple_regression <- lm(Sales ~ Temperature, data = icecream)
```

Let’s take a look at the output using the `summary()` function:

```
summary(simple_regression)
```

```
##
## Call:
## lm(formula = Sales ~ Temperature, data = icecream)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.512 -12.566   4.133  22.236  49.963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -694.369    105.048   -6.61 6.00e-05 ***
## Temperature   16.715      1.592   10.50 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.13 on 10 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.9085
## F-statistic: 110.2 on 1 and 10 DF,  p-value: 1.016e-06
```

You’ll notice that we can get estimates for b_0 (`(Intercept)`) and for b_1 (`Temperature`). Do those match what we calculated before? How do we interpret these values?

Our intercept is the expected value of the dependent variable when the independent variable is 0. This means that when the temperature is 0 degrees Fahrenheit we expect to make -694 dollars in ice cream sales.

Note that this is a sort of ridiculous value - it never gets to 0 degrees Fahrenheit in Davis, so do we care about expected ice cream sales at that temperature? This is why centering your predictor is sometimes handy to improve interpretation of the intercept!

Our slope is the expected change in the dependent variable for a 1-unit change in the independent variable So as the temperature increases by 1 degree Fahrenheit, we expect ice cream sales to increase by $16.72.

Hypothesis Testing for the Slope and R-Squared

You'll notice that in addition to the estimates of the intercept and slope, we also have some other information In particular, we have a standard error, test statistic, and p-value.

These are for testing the null hypothesis that the parameter is equal to 0 versus the alternative that it's not equal to 0.

Although we get this test for the intercept as well that's typically not very interesting to us. Instead, we're more interested in the test for the slope Because that tells us whether the predictor is useful in predicting the outcome; otherwise, the slope is no different from 0 and there is no linear relation, no useful relation between X and Y that we can use to predict Y.

The hypotheses for this test are:

- H_0: $b_1 = 0$
- H_1: $b_1 \neq 0$

The t-statistic in this test is distributed as a t-distribution with N - 2 df. What is the df for our example?

Based on our `summary()` output, do we reject or fail to reject the null hypothesis? What does that tell us?

We could also create a confidence interval around our estimate of the slope using the information provided. Recall that the formula for a CI is:

$$\text{CI}_{95\%} = \hat{b_1} \pm t_{crit} \times \text{SE}$$

In a simple linear regression, we can get the standard error for the slope as:

$$\text{SE}_{b_1} = \frac{s_\varepsilon}{\sqrt{SS_X}}$$

where,

$$s_\varepsilon = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2},$$
$$SS_X = \sum_{i=1}^{N}(X_i - \bar{X})^2$$

As `R` code:

```
sd_error <- sqrt( (1/(nrow(icecream) - 2)) * sum((icecream$Sales - predict(simple_regression))^2) )
ss_x <- sum((icecream$Temperature - mean(icecream$Temperature))^2)
se_b1 <- sd_error/sqrt(ss_x)
```

Of course, we don't need to do all these calculations by hand. The standard error for the slope is given in the output when we call the `summary()` function on our model. Take a look at it again and try to find it!

To get the 95%CI we have to estimate ± $t_{crit} \times$ SE where t_{crit} is the critical value that cuts off the upper 2.5% of a t-distribution with N - 2 df:

```r
t_crit <- qt(.025, nrow(icecream) - 2, lower.tail = FALSE)

lower_limit <- b1_estimate - t_crit * 1.592
upper_limit <- b1_estimate + t_crit * 1.592

lower_limit
```

```
## [1] 13.16828
```

```r
upper_limit
```

```
## [1] 20.26268
```

We can also obtain the 95% CI using the `confint()` function:

```r
confint(object = simple_regression, level = 0.95)
```

```
##                 2.5 %     97.5 %
## (Intercept) -928.4318 -460.30714
## Temperature   13.1679   20.26305
```

R-Squared

One other piece of useful information that you get from the `summary()` output is R-squared, also called the coefficient of determination. R-squared is the proportion of total variability in our outcome that is explained by our predictor(s) This is calculated as sum of squares regression divided by total sum of squares: SS_{reg}/SS_{total} **or** $1 - (SS_{resid}/SS_{total})$

You could manually calculate SS_{resid} and SS_{total}, but I'm not going to bother with that You can see the R^2 value in the `summary()` output, and that value is 0.9168. This means that almost 92% of the variation in ice cream sales is explained by temperature.

You'll notice that there is also an adjusted R^2 This is more important in the context of multiple regression because as you add more predictors to the model, your R^2 will typically increase even if the added predictors are just explaining noise; therefore, adjusted R^2 corrects for the number of predictors in the model.

Why do we care about R^2? Because it may be the case that we have a significant predictor, but it doesn't actually explain a whole lot of the total variance So that's why in our write-up of the results, we'd want to report both the regression coefficients and R^2.

Here's what that write-up could look like:

We ran a simple linear regression to determine whether temperature predicts ice cream sales. The effect of Temperature is statistically significant and positive ($b = 16.72$, 95% CI [13.17, 20.26], $t(10) = 10.50$, $p < .001$). Temperature explained 91.68% of the total variability in ice cream sales.

Your Turn

Run a simple linear regression predicting the number of pool accidents from the temperature. Interpret the intercept and slope. Is the slope significant?