Data Science in Action
Linear & Logistic Regression in R

Jesse Gronsbell

Department of Statistical Sciences, University of Toronto

Take a peak at the data

```
> str(ice_cream_data)
'data.frame':        100 obs. of  2 variables:
 $ temperature    : int  69 91 61 76 85 72 67 83 79 79 ...
 $ ice_cream_sales: num  348 464 321 396 421 ...
> head(ice_cream_data)
  temperature ice_cream_sales
1          69           347.97
2          91           464.09
3          61           320.64
4          76           395.69
5          85           420.78
6          72           375.45
```

Take a peak at the data

```
> summary(ice_cream_data)
  temperature    ice_cream_sales
 Min.   :60.00   Min.   :289.2
 1st Qu.:68.50   1st Qu.:346.8
 Median :77.50   Median :393.8
 Mean   :77.15   Mean   :394.8
 3rd Qu.:85.50   3rd Qu.:440.3
 Max.   :95.00   Max.   :496.6
```

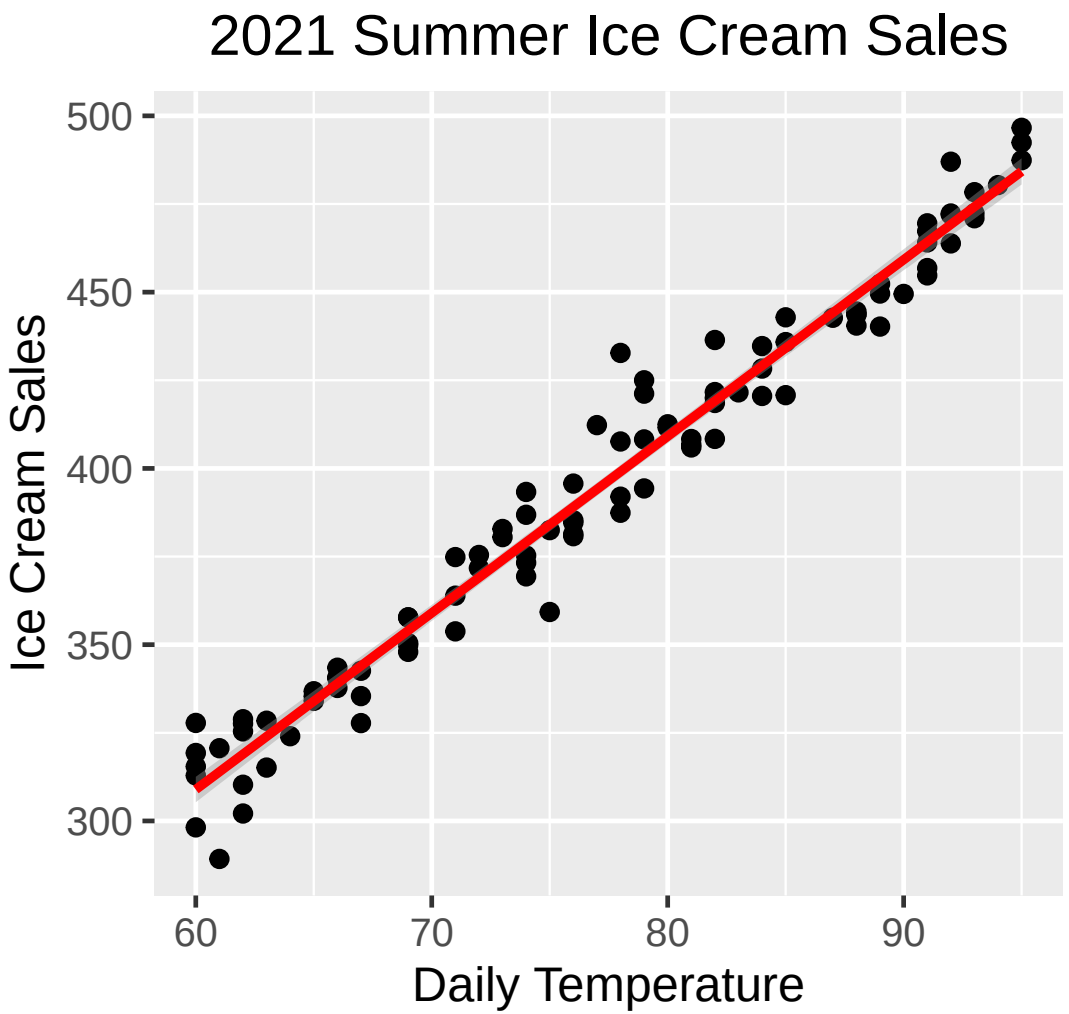
Fit the linear regression model

```
> my_fit <- lm(ice_cream_sales~temperature,
data = ice_cream_data)
> summary(my_fit)

Call:
lm(formula = ice_cream_sales ~ temperature,
data = ice_cream_data)
Residuals:
    Min      1Q  Median      3Q     Max
-24.787  -6.301  -0.115   5.374  33.666

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.38974    7.03602   1.192    0.236
temperature  5.00890    0.09036  55.433   <2e-16 ***
---
```

Plot the results

```
> # Load ggplot2 package
> library(ggplot2)
> ggplot(ice_cream_data, aes(y = ice_cream_sales,
+                            x = temperature)) +
+   geom_point() +
+   stat_smooth(method = "lm", col = "red") +
+   labs(y = 'Ice Cream Sales', x = 'Daily Temperature') +
+   ggtitle('2021 Summer Ice Cream Sales') +
+   theme(plot.title = element_text(hjust = 0.5))
```

Linear regression model results



Code least squares from scratch

```
> get_LS_estimates <- function(x, y){
+   ls_slope <-  (y-mean(y))%*%(x -
mean(x))/((x-mean(x))%*%(x - mean(x)))
+   ls_intercept <- mean(y) - ls_slope*mean(x)
+   return(list(ls_intercept = ls_intercept,
+               ls_slope = ls_slope))
+ }
```

Compare to your previous model fit

```
> my_LS_estimates <- get_LS_estimates(temperature,
ice_cream_sales)
> my_LS_estimates
$ls_intercept
         [,1]
[1,] 8.389741

$ls_slope
         [,1]
[1,] 5.008898

> my_fit$coefficients
(Intercept) temperature
   8.389741    5.008898
```

Fit the logistic regression model

```
> my_logistic_fit <- glm(made_400~temperature,
data = ice_cream_data, family = 'binomial')
> summary(my_logistic_fit)

Call:
glm(formula = made_400 ~ temperature, family = "binomial",
data = ice_cream_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90785  -0.00229   0.00000   0.00091   1.73556

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -112.8475    42.4827  -2.656  0.00790 **
temperature    1.4492     0.5448   2.660  0.00781 **
```

Plot the results

```
> ggplot(ice_cream_data, aes(y=made_400, x = temperature))
+     geom_point() +
+     stat_smooth(method="glm",
+                 method.args = list(family="binomial"),
+                 se=FALSE, fullrange=TRUE) +
+     labs(x="Daily Temperature",
+          y="Probability of Making $400") +
+     ggtitle('2021 Summer Ice Cream Sales') +
+     theme(plot.title = element_text(hjust = 0.5))
```

Logistic regression model results