Class Example 1. Linear Regression Example

The code below uses “BrisGCtemp.csv”, as appeared in a class example of assignment 3. This file contains temperature observations recorded in Brisbane and the GoldCoast.
We are looking for a model of the form:

$$\text{Gold Coast Temperature} = \beta_0 + \beta_1\text{Brisbane Temperature} + \epsilon.$$

We slice up the data over different months, yielding a different model for each month. That is, each month has it’s own β_0 and β_1.

As presented below, the code is for the month of May.

```
using PyPlot, DataFrames, GLM
table = readtable("BrisGCtemp.csv")
println(table)
#This is the desired month to filter by
desiredMonth = 5

#column 2 in the dataframe is the month, filter by it
xDat = table[table[2] .== desiredMonth, :][4] #4 is the column of Brisbane
yDat = table[table[2] .== desiredMonth, :][5] #5 is the column of Gold Coast
println("\nFor month: ", desiredMonth,
                ",observations: ", length(xDat),
                ", means: ", (round(mean(xDat)),round(mean(yDat))))

data = DataFrame(X = xDat, Y = yDat)
model = glm(@formula(Y ~ X), data, Normal(), IdentityLink())

PyPlot.scatter(xDat, yDat)
PyPlot.plot([minimum(xDat),maximum(xDat)],
                [1 minimum(xDat); 1 maximum(xDat)]*coef(model),"r");
model
```

The output presented on the next page shows:

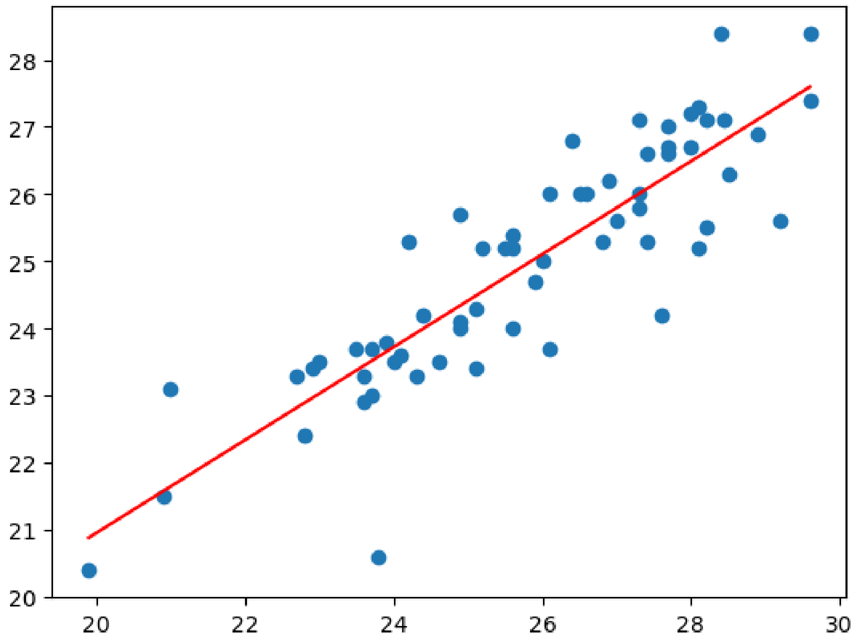
- A plot of the data and the regression line.
- A shortened version of the dataframe (table).
- Summary means (Brisbane,GoldCoast) and observation counts for the selected month (May in this case).
- The estimated model. In this case,

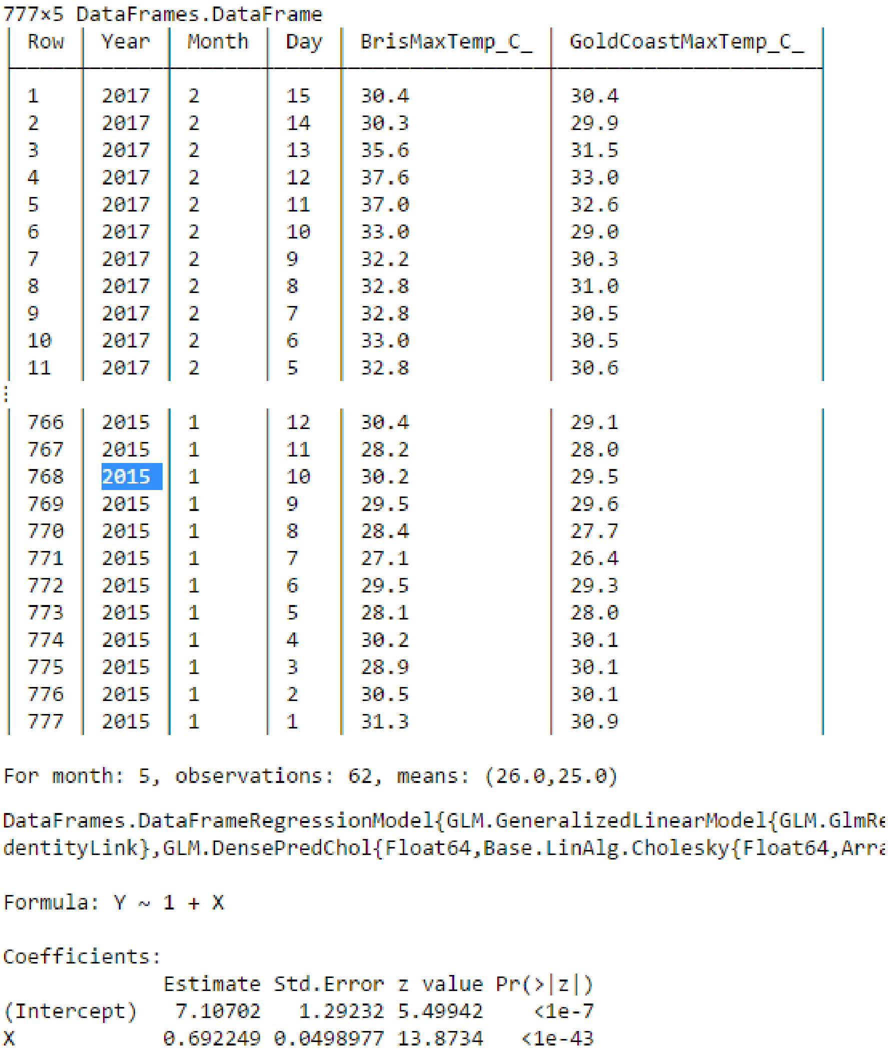
$$\text{Gold Coast Temperature} = 7.10 + 0.6922 \times \text{ Brisbane Temperature}.$$

(a) Run May and understand the result.

(b) Modify the code (desiredMonth) for a different month and analyse the results.

(c) Assume you accept this model and observe a temperate of 30 degrees in Brisbane in a given day. What is your estimate of the temperature in Gold Coast for that day? (i) In January? (ii) In October?





```
777×5 DataFrames.DataFrame
│ Row │ Year │ Month │ Day │ BrisMaxTemp_C_ │ GoldCoastMaxTemp_C_ │
├─────┼──────┼───────┼─────┼────────────────┼─────────────────────┤
│ 1   │ 2017 │ 2     │ 15  │ 30.4           │ 30.4                │
│ 2   │ 2017 │ 2     │ 14  │ 30.3           │ 29.9                │
│ 3   │ 2017 │ 2     │ 13  │ 35.6           │ 31.5                │
│ 4   │ 2017 │ 2     │ 12  │ 37.6           │ 33.0                │
│ 5   │ 2017 │ 2     │ 11  │ 37.0           │ 32.6                │
│ 6   │ 2017 │ 2     │ 10  │ 33.0           │ 29.0                │
│ 7   │ 2017 │ 2     │ 9   │ 32.2           │ 30.3                │
│ 8   │ 2017 │ 2     │ 8   │ 32.8           │ 31.0                │
│ 9   │ 2017 │ 2     │ 7   │ 32.8           │ 30.5                │
│ 10  │ 2017 │ 2     │ 6   │ 33.0           │ 30.5                │
│ 11  │ 2017 │ 2     │ 5   │ 32.8           │ 30.6                │
⋮
│ 766 │ 2015 │ 1     │ 12  │ 30.4           │ 29.1                │
│ 767 │ 2015 │ 1     │ 11  │ 28.2           │ 28.0                │
│ 768 │ 2015 │ 1     │ 10  │ 30.2           │ 29.5                │
│ 769 │ 2015 │ 1     │ 9   │ 29.5           │ 29.6                │
│ 770 │ 2015 │ 1     │ 8   │ 28.4           │ 27.7                │
│ 771 │ 2015 │ 1     │ 7   │ 27.1           │ 26.4                │
│ 772 │ 2015 │ 1     │ 6   │ 29.5           │ 29.3                │
│ 773 │ 2015 │ 1     │ 5   │ 28.1           │ 28.0                │
│ 774 │ 2015 │ 1     │ 4   │ 30.2           │ 30.1                │
│ 775 │ 2015 │ 1     │ 3   │ 28.9           │ 30.1                │
│ 776 │ 2015 │ 1     │ 2   │ 30.5           │ 30.1                │
│ 777 │ 2015 │ 1     │ 1   │ 31.3           │ 30.9                │

For month: 5, observations: 62, means: (26.0,25.0)

DataFrames.DataFrameRegressionModel{GLM.GeneralizedLinearModel{GLM.GlmRe
dentityLink},GLM.DensePredChol{Float64,Base.LinAlg.Cholesky{Float64,Arra

Formula: Y ~ 1 + X

Coefficients:
             Estimate Std.Error z value Pr(>|z|)
(Intercept)   7.10702   1.29232 5.49942    <1e-7
X            0.692249 0.0498977 13.8734   <1e-43
```

Class Example 2. Automation on All Months

The code below obtains a regression model for all 12 months.

```julia
using PyPlot, DataFrames, GLM
table = readtable("BrisGCtemp.csv")

function statsOfMonth(desiredMonth)
    xDat = table[table[2] .== desiredMonth, :][4]
    yDat = table[table[2] .== desiredMonth, :][5]
    data = DataFrame(X = xDat, Y = yDat)
    model = glm(@formula(Y ~ X), data, Normal(), IdentityLink())
    return (desiredMonth,length(xDat),
                (round(mean(xDat)),round(mean(yDat))),
                coef(model) )
end

[statsOfMonth(m) for m in 1:12]
```

(a) Run the code above and interpret the results.

(b) Replace “coef(model)” with “model” and re-run to obtain the model for all 12 months. Assume that for every month you are considering $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. What are your results for these 12 hypothesis tests?

Class Example 3. Logistic Regression.

Logistic regression is a mechanism to describe "0"/"1" outcomes (y) as a function of x. Here we are predicting the probability of an event happening ($y = 1$) as a function of x.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

The following code takes maximal weekly temperatures (x) and the event of having rain during that week (y). It then constructs a logistic regression model where C are the resulting coefficients.

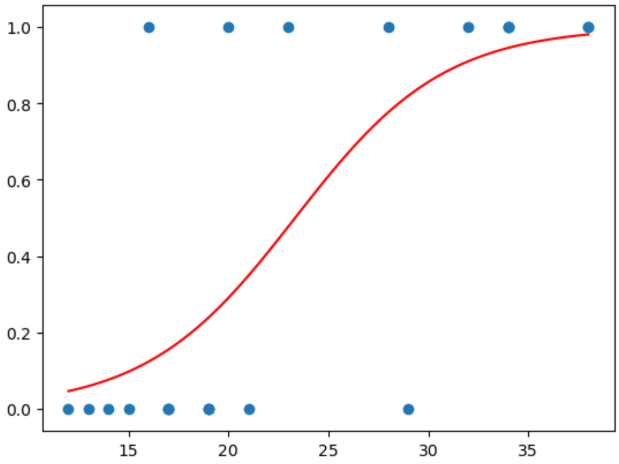
```julia
using PyPlot, GLM, DataFrames, Distributions

x = [23,32,17,34,29,38,16,14,19,34,38,19,20,21,19,17,13,12,28,34,15,17]
y = [1,1,0,1,0,1,1,0,0,1,1,0,1,0,0,0,0,0,1,1,0,0]
data = DataFrame(X = x, Y = y)

model = glm(@formula(Y ~ X), data, Binomial(), LogitLink())

C = coef(model)

xm = linspace(minimum(x),maximum(x),100)
ym = 1./(1+exp(-(C[1]+C[2].*xm)))
xlim = (minimum(x),maximum(x))
PyPlot.scatter(x,y)
PyPlot.plot(xm,ym,"r")
model
```



(a) Assume the temperate is 26 degrees, what is your estimate of the probability of having rain?

(b) Modify the code by adding a line "y = 1-y" after setting up y. Explain the output.

Question 1. Roadways

Regression methods were used to analyze the data from a study investigating the relationship between roadway surface temperature (x) and pavement deflection (y). Summary quantities were $n = 15, \sum y_i = 11.75, \sum y_i^2 = 7.86, \sum x_i = 1348, \sum x_i^2 = 123,324.6$ and $\sum x_i y_i = 983.67$.

(a) Calculate the least squares estimates of the slope and intercept. Graph the regression line. Estimate σ^2.

(b) Use the equation of the fitted line to predict what pavement deflection would be observed when the surface temperature is $75 \deg F$.

(c) What is the mean pavement deflection when the surface temperature is $95 \deg F$?

(d) What change in mean pavement deflection would be expected for a $1 \deg F$ change in surface temperature?

Question 2. House Selling Prices

An article in *Technometrics* by S.C. Narula and J.F Wellington [”Prediction, Linear Regression, and a Minimum Sum of Relative Errors” (1977, Vol. 19)] presents data on the selling price and annual taxes for 24 houses. The data is stored in (*11-6.csv*).

(a) Assuming that a simple linear regression model is appropriate, obtain the least squares fit relating selling price to taxes paid. What is the estimate of σ^2?

(b) Find the mean selling price given that the taxes paid are $x = 6.00$.

(c) Calculate the fitted value of y corresponding to $x = 5.8980$. Find the corresponding residual.

(d) Calculate the fitted $\hat{y}_i$ for each value of x_i used to fit the model. Then construct a graph of $\hat{y}_i$ versus the corresponding obversed value y_i and comment on what this plot would look like if the relationship between y and x was a deterministic (no random error) straight line. Does the plot actually obtained indicate that taxes paid is an effective regressor variable in predicting selling price?

Question 3. Rocket Motor

A rocket motor is manufactured by bonding together two types of propellants, an igniter and a sustainer. The shear strength of the bond y is thought to be a linear function of the age fo the propellant x when the motor is cast. The data is stored in(*11-13.csv*).

(a) Draw a scatter diagram of the data. Does the straight-line regression model seem to be plausible?

(b) Find the least squares estimates of the slope and intercept in the simple linear regression model. Find and estimate of σ^2.

(c) Estimate the mean shear strength of a motor made from propellant that is 20 weeks old.

(d) Obtain the fitted values $\hat{y}_i$ that correspond to each observed value y_i. Plot $\hat{y}_i$ versus y_i and comment on what this plot would look like if the linear relationship between shear strength and age were perfectly deterministic (no error). Does this plot indicate that age is a reasonable choice of regressor variable in this model?

Question 4. Regression without the Intercept Term

Suppose that we wish to fit a regression model for which the true regression line passes through the point (0,0). The appropriate model is $Y = \beta x + \epsilon$. Assume that we have n pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

(a) Find the least squares estimate of β. Show your working.

(b) Fit the model $Y = \beta x + \epsilon$ to the chloride concentration roadway area data stored in (*11-22.csv*). Plot the fitted model on a scatter diagram of the data and comment on the appropriateness of the model.

Question 5. Body Mass Index

The World Health Organization defines *obesity* in adults as having a body mass index (BMI) higher than 30. In a study of 250 men at Bingham Young University, 23 are by this definition obese. How good is waist (size in inches) as a predictor of obesity? A logistic regression model was fit to the data:

$$\log\left(\frac{p}{1-p}\right) = -41.828 + 0.9864 \text{ waist}$$

where p is the probability of being classified as obese.

(a) Does the probability of being classified as obese increase or decrease as a function of waist size? Explain.

(b) What is the estimated probability of being classified as obese for a man with a waist size of 36 inches?

(c) What is the estimated probability of being classified as obese for a man with a waist size of 42 inches?

(d) What is the estimated probability of being classified as obese for a man with a waist size of 48 inches?

(e) Make a plot of the estimated probability of being classified as obese as a function of waist size.

Question 6. Guest Lecture - Take Away

Choose one of the two guest lectures (either presented on the Monday or the Wednesday lecture). For this guest lecture, summarize in 1-3 paragraphs (a quarter to three quarters of a page) the main point of the lecture. Make sure that your summary is self contained, clear and well written.

Question 7. Guest Lecture - Statistics and Probability?

For the guest lecture that you chose, discuss some of the concepts of statistics and probability that came up. Relate those concepts to elementary concepts that you learned in the course where possible. Your write up should again be 1-3 paragraphs long, but make sure to use formulas or precise definitions where needed.

Question 8. And now for something completely different

Choose on of the items appearing in Unit 10 of the condensed lecture notes. This is an item covered in the course book, but not covered in the lectures. Summarize the item of your choice in 1-3 paragraphs, expanding on the short summary appearing in Unit 10 of the condensed notes. Give a real-life engineering example where possible.