Aren Ashlock
Department of Computer Science
Iowa State University

February 8, 2024

HW2 — Regression

1 Linear Regression

1. Consider the data set $x = \{1, 2, 3, 4, 5, 6, 10\}, y = \{0, 1, 3, 2, 20, -6, 80\}$.

   a) What are the sample means, $\bar{x}$ and $\bar{y}$?
   Answer: $\bar{x} = 4.43$, $\bar{y} = 14.29$

   b) What are the sample variances and covariance, s_x^2, s_y^2, and s_{xy}?
   Answer: $s_x^2 = 7.67$, $s_y^2 = 774.49$, $s_{xy} = 62.88$

   c) What is the solution (the values of a and b) for the simple linear regression using the given data set with the function of the form

   $$y = ax + b \tag{1-1}$$

   (please show your calculation process step-by-step)?
   Answer: $a = \frac{s_{xy}}{s_x^2} = \frac{62.88}{7.67} = 8.19$
   $b = \bar{y} - a\bar{x} = 14.29 - 8.19(4.43) = -22.00$

   d) Is your predictor for (1-1) better than, worse than, or the same with the trivial predictor $\hat{y} = \bar{y}$? Why?
   Answer: $RSS(a, b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2 = 1814.83$
   $R^2 \triangleq 1 - \frac{RSS/n}{s_y^2} \rightarrow R^2 \triangleq 1 - \frac{1814.83/7}{774.49} = 0.67$
   Since the goodness-of-fit is greater than 0, that means the predictor for (1-1) is **better than** the trivial predictor.

   e) What is the solution (the value of k) for the simple linear regression using the given data set with the function of the form

   $$y = kx \tag{1-2}$$

   (please show your calculation process step-by-step)?
   Answer: $RSS(k) = \sum_{i=1}^{n}(y_i - kx_i)^2 \rightarrow \frac{dRSS(k)}{dk} = 0 \rightarrow -2\sum_{i=1}^{n} x_i(y_i - kx_i) = 0 \rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - kx_i) = 0 \rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}k\sum_{i=1}^{n} x_i^2 = 0 \rightarrow k = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \rightarrow k = 4.62$

   f) Is your predictor for (1-2) better than, worse than, or the same with the previous predictor of the form (1-1)? Why?
   Answer: $RSS(k) = \sum_{i=1}^{n}(y_i - kx_i)^2 = 2767.86$
   $R^2 \triangleq 1 - \frac{RSS/n}{s_y^2} \rightarrow R^2 \triangleq 1 - \frac{2767.86/7}{774.49} = 0.49$
   Since the goodness-of-fit is less than 0.67 from (1-1), that means the predictor for (1-2) is **worse than** the predictor from (1-1).

   g) What is the solution (the values of $\alpha, \beta,$ and γ) for the regression using the given data set with the function of the form

   $$y = \alpha x + \beta x^2 + \gamma \tag{1-3}$$

   (please show your calculation process step-by-step)?
   I used quadratic regression because Bowen's email didn't come in before I didn't have time to redo this question...
   Equations: $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $S_{xx^2} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i^2 - \bar{x^2})$,

$S_{x^2x^2} = \sum_{i=1}^{n}(x_i^2 - \bar{x^2})^2$, $S_{x^2y} = \sum_{i=1}^{n}(x_i^2 - \bar{x^2})(y_i - \bar{y})$

Answer: $\alpha = \frac{S_{xy}S_{x^2x^2} - S_{x^2y}S_{xx^2}}{S_{xx}S_{x^2x^2} - (S_{xx^2})^2} = -9.32$, $\beta = \frac{S_{x^2y}S_{xx} - S_{xy}S_{xx^2}}{S_{xx}S_{x^2x^2} - (S_{xx^2})^2} = 1.58$, $\gamma = \bar{y} - \alpha\bar{x} - \beta\bar{x^2} = 12.42$

h) Is your predictor for (1-3) better than, worse than, or the same with the previous predictor of the form (1-1)? Why?

Answer: $RSS(\alpha, \beta, \gamma) = \sum_{i=1}^{n}(y_i - (\alpha x_i + \beta x_i^2 + \gamma))^2 = 642.68$
$R^2 \triangleq 1 - \frac{RSS/n}{s_y^2} \rightarrow R^2 \triangleq 1 - \frac{642.68/7}{774.49} = 0.88$
Since the goodness-of-fit is more than 0.67 from (1-1), that means the predictor for (1-3) is **better than** the predictor from (1-1).

i) The programming solution and its performance using `sklearn` for (1-1) can be obtained as:

```
import numpy as np
from sklearn import linear_model
X = np.array([1, 2, 3, 4, 5, 6, 10]).reshape(-1,1)
Y = np.array([0, 1, 3, 2, 20, -6,80])
"""fit y=ax+b"""
model = linear_model.LinearRegression(fit_intercept=True)
model.fit(X, Y)
print ("a = %s, b=%s, score=%s" &
        ("{:.3f}".format(model.coef_[0]),
         "{:.3f}".format(model.intercept_),
         "{:.3f}".format(model.score(X, Y))))
```

You can refer to the `sklearn` documentation (**link**) for more details and examples. Write a program to verify your solutions for (1-2) and (1-3). In your answer you should specify the key program (one or two lines of code) to obtain your solution, and whether the solution aligns with your previous answer for the respective functions ((1-2) and (1-3)).

Answer:

(1-2): The solution from the program aligns with my answer for the function. Below is the two lines used to get the solution...

```
model = linear_model.LinearRegression(fit_intercept=False)
model.fit(X, Y)
```

(1-3): The solution from the program aligns with my answer for the function. Below is the four lines used to get the solution... (sorry I have more than two, just wanted to make sure the program made sense)

```
poly = PolynomialFeatures(degree=2, include_bias=False)
poly_features = poly.fit_transform(X)
model = linear_model.LinearRegression(fit_intercept=True)
model.fit(poly_features, Y)
```

j) Is there a unique solution for the regression problem using the function of the form

$$y = \beta_0 + \sum_{i=1}^{10} \beta_i x^i \tag{1-4}$$

with the given data set? If yes, what is your solution? If not, why isn't the solution unique? Use the `sklearn` package and the `linear_model` method to verify your answer. Is your programmed solution aligning with your answer above? If not, why?

Answer: Since $n < d$, the solution is **NOT UNIQUE**, which is what I am getting when I use `sklearn`.

2 Logistic Regression

1. Should the following be tackled as regression or classification problems? Think carefully and explain your answers.

   a) Given a person's credit rating and zip code, predict how many children they will have.
   Answer: Regression. The number of children varies, not a choice between 2 values.

   b) Given a wildlife camera image, detect the number of legs on a creature.
   Answer: Classification. There are discrete values that can be detected meaning we can make classes for them and the predictor assigns each image to a class.

   c) Given a person's biometrics, predict whether they will live past the age of 80.
   Answer: Classification. Either the predictor says "Yes" or "No" meaning there are only 2 possibilities.

   d) Given a person's biometrics, predict the person's age.
   Answer: Regression. There are many possible ages, so this range isn't a binary set.

3 A "Bonus" Question, Again (1 pt)

From a scale of 1 to 5, how difficult is HW1? 1 is "I can do it in my sleep". 5 is "Bowen is ridiculous". 0 is "I refuse to answer this question". This is for my own reference to improve the quality of future assignments. Thanks!

Answer: I would say a 4. Getting over my anxiety of using Python made it easier to calculate answers. However, $'e'$ through $'j'$ in question 1 were kind of difficult since I wasn't sure how to approach them. Most of the issues surrounding those questions were due to my lack of understanding, but I *believe* I figured it out!

Submitted by Aren Ashlock on February 8, 2024.