CS 418: Introduction to Data Science
Homework Assignment 02
Fall 2019

Instructions

This assignment is due Friday, October 18, at 11:59PM (Central Time).

Answers for this assignment must be entered on *Blackboard*. You have 1 attempt to submit the assignment. This attempt is not timed. The assignment will close automatically at the due date and no late submissions will be accepted.

This assignment is individual. Offering or receiving any kind of unauthorized or unacknowledged assistance is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.

Part I: Simple Linear Regression (60 pt.)

Problem 1 (10 pt.)

A credit card company charges $10 when a payment is late and $5 per day each day the payment remains unpaid. Find the linear equation that expresses the late payment fee in terms of the number of days the payment is late and answer the following questions:

a. **(2 pts.) What are the dependent and independent variables of the linear equation?**

b. **(3 pts.) What are the y-intercept and the slope of the linear equation?**

c. **(3 pts.) What is the late payment fee when the payment is two weeks late?**

d. **(2 pts.) Is the correlation between the late payment fee and the number of days the payment is late positive or negative?**

Problem 2 (10 pt.)

A cleaning company charges an equipment fee and an hourly fee. The total fee charged by the company (in $) is given by the linear equation $y = 100 + 50x$. Answer the following questions:

a. **(2 pts.) What are the dependent and independent variables of the linear equation?**

b. **(3 pts.) The y-intercept of the linear equation is 100 and the slope is 50. Explain, in two or three sentences, what this y-intercept and this slope represent.**

c. **(3 pts.) What is the total fee charged by the cleaning company for 4 hours of work?**

d. **(2 pts.) Is the correlation between the total fee of the cleaning company and the number of hours worked positive or negative?**

Problem 3 (40 pt.)

Given the following table with the number of endorsements and the annual salary for a random sample of professional athletes:

Number of endorsements	Annual salary (in millions of $)
0	2
3	8
2	7
1	3
5	13
5	12
4	9
3	9
0	3
4	10

Build a simple linear regression model to predict the annual salary of a professional athlete given their number of endorsements and answer the following questions:

a. (4 pts.) What is the sample mean and the sample standard deviation of the athletes' number of endorsements and annual salary?

b. (3 pts.) What is the correlation coefficient between the athletes' number of endorsements and their annual salary?

c. (2 pts.) What is your interpretation of this correlation coefficient?

d. (6 pts.) Fit the model using the least squares method. What is the y-intercept and the slope of the model?

e. (4 pts.) Explain, in two or three sentences, what this y-intercept and this slope represent.

f. (3 pts.) What is the predicted annual salary of an athlete with 3 endorsements?

g. (3 pts.) Suppose that an athlete with 3 endorsements has an annual salary of $7.5 million. What is the residual for this observation?

h. (3 pts.) If an athlete has an annual salary of $10 million, how many endorsements do you expect this athlete to have?

i. (6 pts.) What is the residual sum of squares, the explained sum of squares, and the total sum of squares of the model?

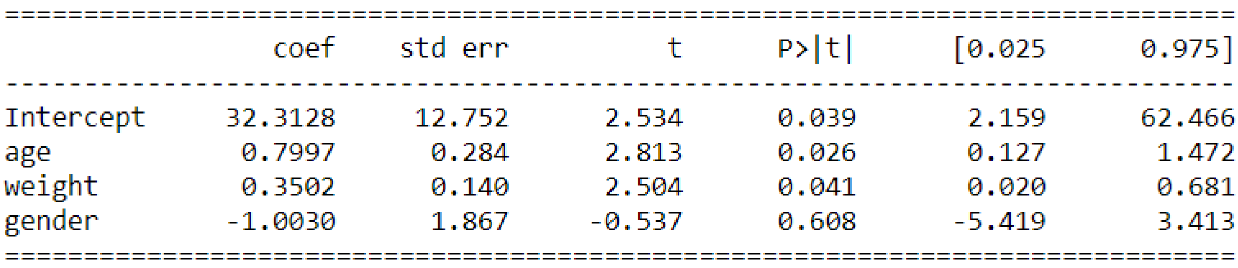
j. (2 pts.) What is the coefficient of determination of the model?

k. (4 pts.) Explain, in one or two sentences, what this coefficient of determination represents.

Part II: Multiple Linear Regression (30 pt.)

Problem 4 (30 pt.)

Suppose that you build a multiple linear regression model to predict a person's blood pressure (in mmHg) given their age (in years), weight (in pounds), and gender (male = 0, female = 1) using a dataset with 10 observations. The regression parameters obtained are shown below:



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     32.3128     12.752      2.534      0.039       2.159      62.466
age            0.7997      0.284      2.813      0.026       0.127       1.472
weight         0.3502      0.140      2.504      0.041       0.020       0.681
gender        -1.0030      1.867     -0.537      0.608      -5.419       3.413
==============================================================================
```

Use this information to answer the following questions:

a. (20 pts.) Which of the following statements is true?

- **For each additional mmHg in blood pressure, the age increases by 0.7997 years, assuming that the weight and the gender remain constant.**
- **For each additional year in age, the blood pressure increases by 0.7997 mmHg, assuming that the weight and the gender remain constant.**
- **If the weight increases by one pound, then the blood pressure increases by 0.3502 mmHg, assuming that the age and the gender remain constant.**
- **If the weight increases by 0.3502 pounds, then the blood pressure increases by one mmHg, assuming that the age and the gender remain constant.**
- **After accounting for age and weight, the blood pressure of females is 1.0030 mmHg more than that of males.**
- **After accounting for age and weight, the blood pressure of females is 1.0030 mmHg less than that of males.**
- **There is a statistically significant linear relationship between blood pressure and age at a significance level of 0.05.**
- **There is a statistically significant linear relationship between blood pressure and weight at a significance level of 0.05.**
- **There is a statistically significant linear relationship between blood pressure and gender at a significance level of 0.05.**

b. (3 pts.) What is the predicted blood pressure of a 50-year-old man weighing 195 pounds?

c. (3 pts.) Suppose that a 55-year-old woman weighing 180 pounds has a blood pressure of 150 mmHg. What is the residual for this observation?

d. (4 pts.) Suppose that the coefficient of determination of the model is 0.978. What is the adjusted R^2 of the model?

Part III: Regularization (10 pt.)

Problem 5 (5 pt.)

Explain, in two or three sentences, the goal of regularization techniques such as ridge regression, LASSO regression, and elastic net.

Problem 6 (5 pt.)

Suppose that you want to build a multiple linear regression model from a high-dimensional dataset using either ridge regression or LASSO regression. You want the model to fit the data and be interpretable. Which of these two regularization techniques would you choose and why?