Regression

Quiz, 6 questions

6/6 points (100%)

✔ Congratulations! You passed!

Next Item

✔ 1 / 1 points

1.

For Questions 1-6, consider the following:

The data found at http://www.stat.ufl.edu/~winner/data/pgalpga2008.dat consist of season statistics for individual golfers on the United States LPGA and PGA tours. The first column reports each player's average driving distance in yards. The second column reports the percentage of the player's drives that finish in the fairway, measuring their accuracy. The third and final column has a 1 to denote a female golfer (on the LPGA tour), and a 2 to denote male golfer (on the PGA tour).

Load these data into R or Excel. In Excel, once you paste the data into a new worksheet, you may need to separate the data into columns using the "Text to Columns" feature under the "Data" menu.

If you wish to separate the LPGA and PGA data, one way in R is to use the subset function:

```
datF <- subset(dat, FM==1, select=1:2)
```

where "dat" is the name of the original data set (replace "dat" with whatever you named this data set), "FM" is the name of the third column (replace "FM" with whatever you named this column), and select=1:2 means to include columns 1 and 2 in the new data set "datF".

- Create two scatter plots with average drive distance on the x-axis and percent accuracy on the y-axis, one for female golfers and one for male golfers. What do you observe about the relationship between these two variables?

○ Drive distance and accuracy are positively correlated; greater distances are associated with greater accuracy.

● Drive distance and accuracy are negatively correlated; greater distances are associated with less accuracy.

Correct

Because these data are observational (not produced from a randomized experiment), we cannot conclude that greater distances *cause* less accuracy, only that greater distances are *associated* with less accuracy.

○ There is no association between driving distance and accuracy.

Regression

6/6 points (100%)

Quiz, 6 questions

1 / 1 points

2.
Golf:

Fit a linear regression model to the female golfer data only with drive distance as the explanatory variable x and accuracy as the response variable y. Use the standard reference (non-informative) prior.

Recall that in a linear regression, we are modeling $E(y) = b_0 + b_1 x$.

In this particular model, the intercept term is not interpretable, as we would not expect to see a 0-yard drive (but it is still necessary). Predictions should generally be made only within the range of the observed data.

- Report the posterior mean estimate of the slope parameter b relating drive distance to accuracy. Round your answer to two decimal places.

-0.26

Correct Response
This "significant" negative slope supports our previous claim that drive distance and accuracy are negatively correlated.

1 / 1 points

3.
Golf:

The posterior mean estimate of the slope from Question 2 is about five standard errors below 0. Hence, the posterior probability that this slope is negative is near 1.

- Suppose the estimate is b. How do we interpret this value?

- If x is the driving distance, we expect the percentage accuracy to be bx.
- For each additional yard of driving distance, we expect to see a decrease in percentage accuracy of $|b|$.

Correct
This is simply the interpretation of the slope in $E(y) = b_0 + b_1 x$. A unit increase in x would result in a b change in the expected value of y.

- For each additional yard of driving distance, we expect to see an increase in percentage accuracy of $|b|$.

If x is the driving distance, we expect the percentage accuracy to be $100bx$.

Regression

6/6 points (100%)

Quiz, 6 questions

1 / 1 points

4.
Golf:

- Use the posterior mean estimates of the model coefficients to obtain a posterior predictive mean estimate of driving accuracy for a new female golfer whose average driving distance is $x = 260$ yards. Round your answer to one decimal place.

64.20

Correct Response

This is $\hat{y}^* = \hat{b}_0 + \hat{b}_1 \cdot 260 = 130.9 - 0.256 \cdot 260 = 64.3.$

1 / 1 points

5.
Golf:

- Which of the following gives a 95% posterior predictive interval for the driving accuracy of a new female golfer whose average driving distance is $x = 260$ yards?

Hint: Modify the code provided with this lesson under "prediction interval."

(63.0, 65.4)

(55.4, 73.0)

(62.8, 65.6)

(53.7, 74.7)

Correct

This is $\hat{y}^* \pm t_{n-2}^{-1}(0.975) \cdot se_r \cdot \sqrt{1 + \frac{1}{n} + \frac{(260-\bar{x})^2}{(n-1)s_x^2}}$ where $\hat{y}^*$ is the prediction mean found in Question 4, $t_{n-2}^{-1}(0.975)$ is the 0.975 quantile of a standard t distribution with $n-2$ degrees of freedom, n is the number of data points, se_r is the residual standard error (estimate of σ), $\bar{x}$ is the sample mean of driving distance, and s_x^2 is the sample variance of driving distance.

Regression

Quiz, 6 questions

6/6 points (100%)

1 / 1 points

6. Golf:

- What is the correct interpretation of the interval found in Question 5?

If we select a new female golfer who averages 260 yards per drive, we are 95% confident that the posterior mean for her accuracy would be in the interval.

For all female golfers who average 260 yards per drive, our probability is .95 that the mean of their driving accuracy is in the interval.

If we select a new female golfer who averages 260 yards per drive, our probability that her driving accuracy will be in the interval is .95.

Correct

Predictive intervals provide probability statements about a new observation.

For all female golfers who average 260 yards per drive, we are 95% confident that all their driving accuracies will be in the interval.