Module 4 Honors

测验, 4 个问题

4/4 分 (100%)

✔ 恭喜！您通过了！

下一项

✔ 1 / 1 分

1。

Consider again the golf data from the regression quiz for Questions 1-4.

The data are found at http://www.stat.ufl.edu/~winner/data/pgalpga2008.dat and 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).

Now consider a multiple regression on the full data set, including both female and male golfers. Modify the third variable to be a 0 if the golfer is female and 1 if the golfer is male and fit the following regression:

$$E(y) = b_0 + b_1 x_1 + b_2 x_2$$

where x_1 is the average driving distance and x_2 is the indicator that the golfer is male.

- What is the posterior mean estimate of b_0 ? Round your answer to the nearest whole number.

147

正确回答

In this case b_0 has no meaningful physical interpretation because it represents the percentage accuracy of a female golfer ($x_2 = 0$) who drives the ball 0 yards on average ($x_1 = 0$).

✔ 1 / 1 分

2。

Golf data:

- The posterior mean estimates of the other two coefficients are $\hat{b}_1 = -0.323$, and $\hat{b}_2 = 8.94$. What is the interpretation of $\hat{b}_1$?

- Holding all else constant, each additional yard of distance is associated with a 0.323 increase in drive accuracy percentage.
- Holding all else constant, being male is associated with a 0.323 decrease in drive accuracy percentage.
- Holding all else constant, being male is associated with a 0.323 increase in drive

accuracy percentage.

Module 4 Honors

测验, 4 个问题

4/4 分 (100%)

- [x] Holding all else constant, each additional yard of distance is associated with a 0.323 decrease in drive accuracy percentage.

正确

Remember, we can't say distance causes loss of accuracy because these are observational data (they were not obtained through a randomized experiment).

✔ 1 / 1 分

3。

Golf data:

The standard error for b_1 (which we can think of as marginal posterior standard deviation in this case) is roughly $1/10$ times the magnitude of the posterior mean estimate $\hat{b}_1 = -0.323$. In other words, the posterior mean is more than 10 posterior standard deviations from 0. What does this suggest?

- [ ] The posterior probability that $b_1 < 0$ is about 0.5, suggesting no evidence for an association between driving distance and accuracy.
- [ ] The posterior probability that $b_1 < 0$ is very low, suggesting a negative relationship between driving distance and accuracy.
- [x] The posterior probability that $b_1 < 0$ is very high, suggesting a negative relationship between driving distance and accuracy.

正确

✔ 1 / 1 分

4。

4/4 分 (100%)

Golf data:

The estimated value of b_2 would typically be interpreted to mean that holding all else constant (for a fixed driving distance), golfers on the PGA tour are about 9% more accurate with their drives on average than golfers on the LPGA tour. However, if you explore the data, you will find that the PGA tour golfers' average drives are 40+ yards longer than LPGA tour golfers' average drives, and that the LPGA tour golfers are actually more accurate on average. Thus b_2 , while a vital component of the model, is actually a correction for the discrepancy in driving distances. Although model fitting can be easy (especially with software), interpreting the results requires a thoughtful approach.

It would also be prudent to check that the model fits the data well. One of the primary tools in regression analysis is the residual plot. Residuals are defined as the observed values y minus their predicted values $\hat{y}$. Patterns in the plot of $\hat{y}$ versus residuals, for example, can indicate an inadequacy in the model. These plots are easy to produce.

In R:

```
plot(fitted(mod), residuals(mod))
```

where "mod" is the model object fitted with the lm() command.

In Excel, residual plots are available as an output option in the regression dialogue box.

- Fit the regression and examine the residual plots. Which of the following statements most accurately describes the residual plots for this analysis?

- The residuals appear to exhibit a curved trend. There is at least one outlier (extreme observation) that we may want to investigate.
- The residuals appear to be random and lack any patterns or trends. There are no outliers (extreme observations).
- The residuals appear to be more spread apart for smaller predicted values $\hat{y}$. There are no outliers (extreme observations).
- The residuals appear to be random and lack any patterns or trends. However, there is at least one outlier (extreme observation) that we may want to investigate.

正确

Outliers can strongly influence model estimates. A thorough data analysis might include an investigation into whether this outlier value was due to some error (clerical or otherwise).