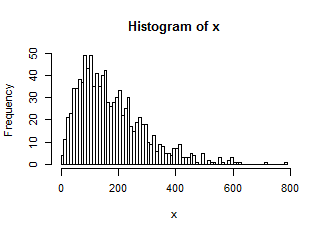
1. We are studying the predictors of income. The following picture shows the histogram of the income. What would you do in the first step of the analysis? Why?



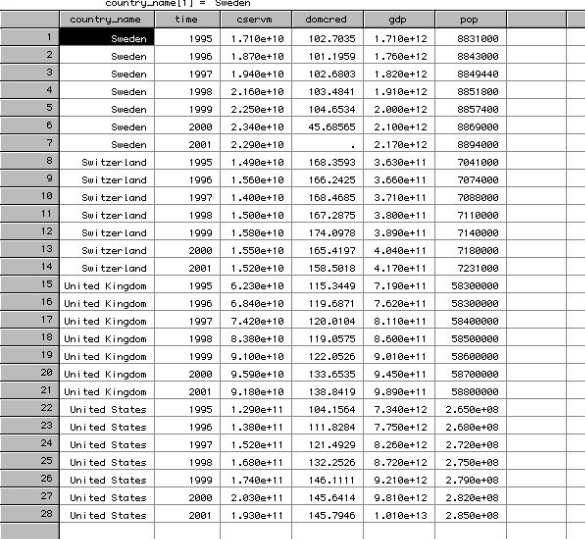
A. A log transformation, so that the income variable is comparable to other independent variables

B. Scale the income data, so that the income variable is comparable to other independent variables

C. A log transformation, so that the transformed income data approximates a normal distribution, and the changes in transformed data are smoother, the estimation results will be more accurate.

D. Scale the income data, so that the transformed income data approximates a normal distribution, and the changes in transformed data are smoother, the estimation results will be more accurate.

2. We have a panel data shown below. Which of the following is correct?



A. This is a balanced panel dataset, the cross section is “country-name”

B, This is an imbalanced panel dataset, the cross section is “country-name”

C. This is a balanced panel dataset, the cross section is “year”

D, This is an imbalanced panel dataset, the cross section is “year”

3. What is the benefit of using a panel dataset to analyze a problem?

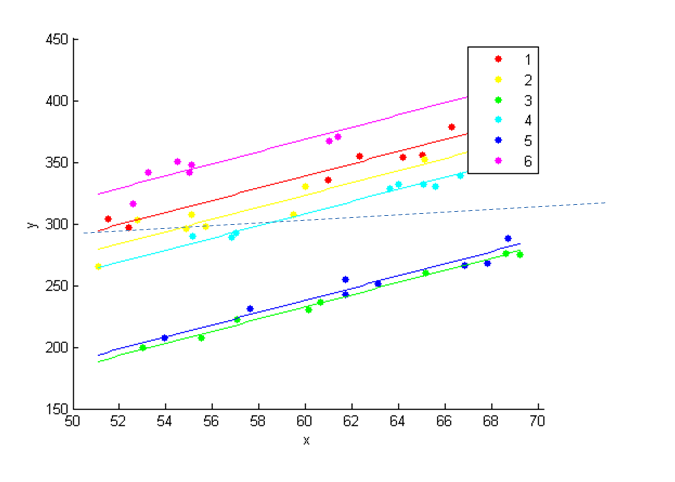
A. A panel dataset allows for the potential to conduct fixed effect model analysis, where we can control for the unobservable factors on a cross section level using some fixed effect dummies.

B. A panel dataset allows for the potential to conduct fixed effect model analysis, where we can control for the trend over time using time fixed effects

C. A panel dataset allows for logistic modeling, where we can control for the unobservable factors on a cross section level using some fixed effect dummies.

D. A panel dataset allows for logistic modeling, where we can control for the trend over time using time fixed effects

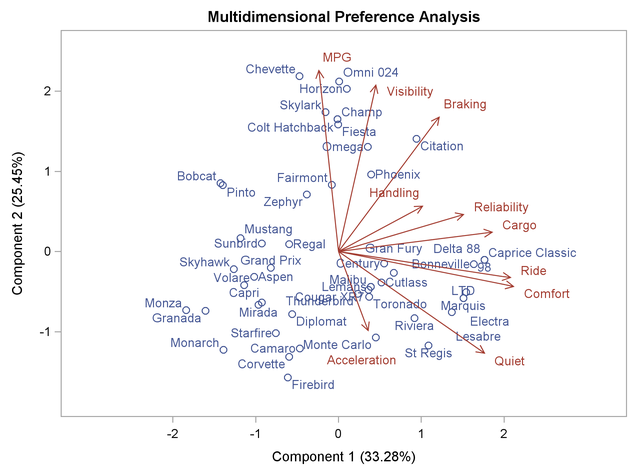
4. We are doing a fixed effect analysis; the dependent variable is y, and the independent variable is x. There are 6 cross sections in our panel dataset. The dotted line is the fitted line of the whole dataset, while the colored solid lines are fitted lines for each cross section. Which of the following is correct?



A. the variable x has almost no impact on y because the fitted line for the entire dataset is almost level.

B. the variable x is positively related to y because within each group, there is a upward sloping trend.

5. We are carefully designing a survey studying the most important aspects of a car that affect the final purchase decision. To start with, we let a group of car owners to rate their car on some different aspects (MPG, Visibility, Braking,……) using a scoring system from -3 to 3. We did a PCA and the following is the biplot. Which of the following is correct?



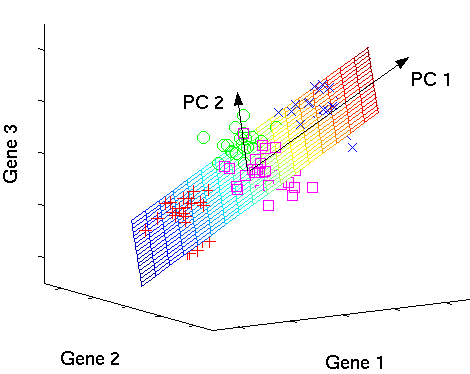
A. The first principal component is Acceleration, and the second principal component is MPG.

B. MPG is positively related to Acceleration because the angle between the two arrows is very wide.

C. The scoring on ride and comfort are very closely related, maybe we only need one of the two measures in the survey.

D. The principal component 2 is mainly a linear combination of Reliability, Cargo, Ride, and Comfort

6. We mapped the cells on a 3D space with 3 attributes (Gene1, Gene2, and Gene3), and then we did a PCA, which of the following is correct?

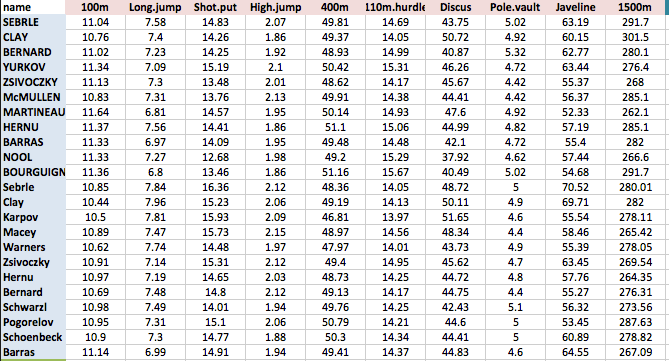


A. The direction along which the dataset has the highest variation is Gene1 direction.

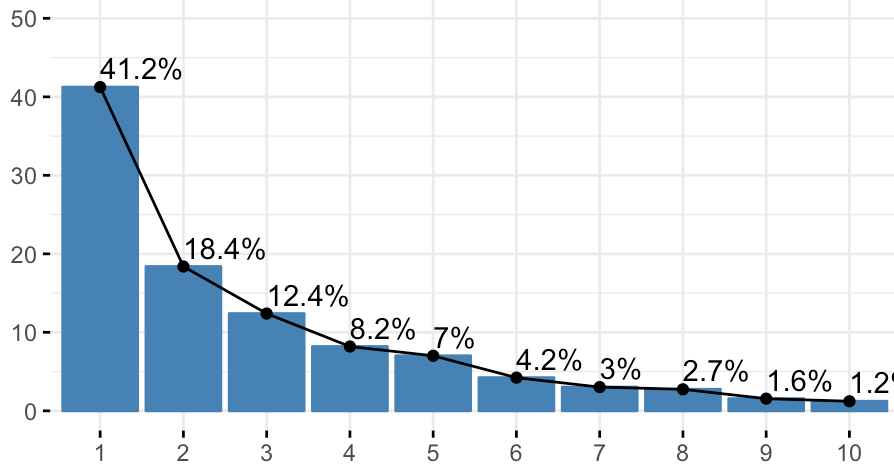
B. The direction along which the dataset has the highest variation is PC2 direction.

C. The cells’ readings on Gene1 and the cells’ readings on Gene2 have zero covariance

D. The cells’ readings on PC1 and the cells’ readings on PC2 have zero covariance

7. We got the data about the athletes’ performance on 100m, long jump……

After we completed the PCA, we got the following plot (Variance% ~ Principal Component):



Which of the following is incorrect?

A. the bar chart is called a scree plot

B. (the largest eigen value of the original covariance table/sum of eigen values) is 41.2%

C. (The largest value on the diagonal of the original covariance table/sum of the diagonal) is greater than 41.2%.

D. (The largest value on the diagonal of the original covariance table/sum of the diagonal) is smaller than 41.2%.

8. We are predicting customers’ purchasing decisions based on the following variables. The purchasing decision is a dummy variable.

1. the number of times the customer views the product descriptions (NV)
2. the number of times the customer reads positive product reviews (NP)
3. the number of times the customer reads negative product reviews (NN)

Which of the following model setup is appropriate?

A. NV + NP + NN +

B. NV + NP + NN +

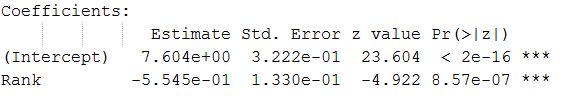
C. Prob( , where NV + NP + NN

D. log(purchase) = NV + NP + NN

9. We study estimate the following linear model:

We got the following estimation result:

Coefficients:



How to interpret the coefficient estimate on Rank?

A. one unit increase in rank is associated with 5.545 decrease in sales on average

B. one unit increase in rank is associated with 55.45 decrease in sales on average

C. one unit increase in rank is associated with 5.545% decrease in sales on average

D. one unit increase in rank is associated with 55.45% decrease in sales on average

10. The probability of raining is 40%, what is the odds of raining?

A. 0.667

B. 0.4

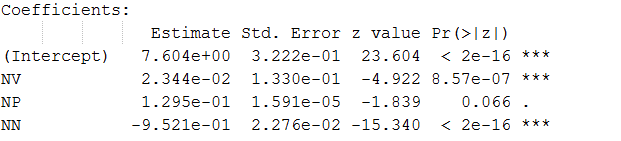
C. 0.6

D. 1.5

11. We are predicting customers’ purchasing decisions based on the following variables. The purchasing decision is a dummy variable.

1. the number of times the customer views the product descriptions (NV)
2. the number of times the customer reads positive product reviews (NP)
3. the number of times the customer reads negative product reviews (NN)

The logistic regression result is as follows:



How to interpret the coefficient estimate on NV?

A. one unit increase in NV leads to 2.344% increase in the purchasing odds

B. one unit increase in NV leads to 23.44% increase in the purchasing odds

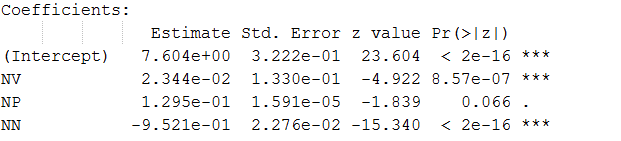
C. one unit increase in NV leads to 2.344% increase in the purchasing probability

D. one unit increase in NV leads to 23.44% increase in the purchasing probability

12. We are predicting customers’ purchasing decisions based on the following variables. The purchasing decision is a dummy variable.

1. the number of times the customer views the product descriptions (NV)
2. the number of times the customer reads positive product reviews (NP)
3. the number of times the customer reads negative product reviews (NN)

The logistic regression result is as follows:



How to interpret the coefficient estimate on NP?

A. one unit increase in NP leads to 12.95% increase in the purchasing odds

B. one unit increase in NP leads to 1.295% increase in the purchasing odds

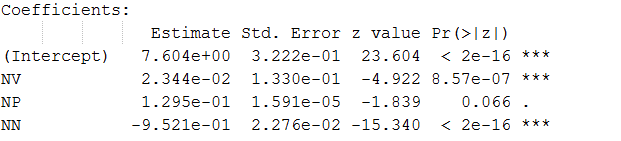
C. one unit increase in NP leads to 12.95% increase in the purchasing probability

D. the number of positive reviews read by the customers does not have statistically significant effect on the purchasing odds. P-value >5%

13. We are predicting customers’ purchasing decisions based on the following variables. The purchasing decision is a dummy variable.

1. the number of times the customer views the product descriptions (NV)
2. the number of times the customer reads positive product reviews (NP)
3. the number of times the customer reads negative product reviews (NN)

The logistic regression result is as follows:



There is a new customer. He has visited the product webpage 12 times, read 5 positive reviews and 8 negative reviews. What is the estimated purchasing probability?

A. 71.42%

B. 47.8%

C. 91.6%

D. 2.499

14. The centered coordinates of the data points in the original attribute space forms a matrix M, the covariance matrix of M is V, the PCA rotation matrix is P.

What is the coordinates of the data points in the principal component space?

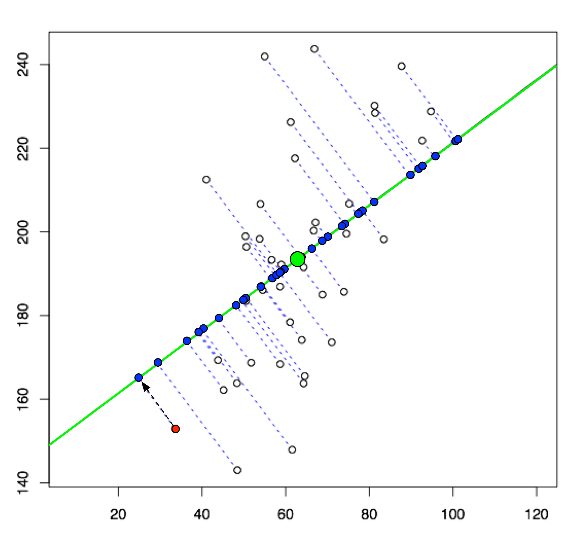
A. M·V·**P**

B. M·P

C. V·**P**

D. M·

15. We did a principal component analysis. The green line shows the direction of the principal component, and the green dot shows the central gravity point of all data points. What if the meaning of the blue dots?



A. The lamda value of the PCA

B. The coordinates of the data points in the original attribute space

C. The coordinates of the data points in the principal component space

D. The covariance matrix of the data points

16. We did a logistic regression, the confusion table is shown below:

|  |  |  |
| --- | --- | --- |
|  | Predicted Class = 1 | Predicted Class = 0 |
| Actual Class = 1 | 56 | 4 |
| Actual Class = 0 | 27 | 23 |

What is the sensitivity and specificity?

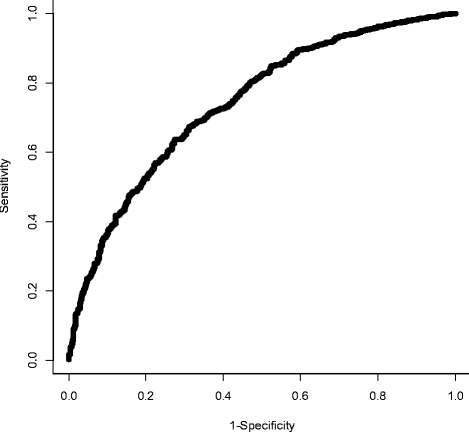
A. Sensitivity = 67.47%, Specificity = 14.81%

B. Specificity = 32.53%, Specificity = 85.19%

C. Sensitivity = 6.67%, Specificity = 54%

D. Specificity = 93.33%, Specificity = 46%

17. We did a logistic regression, and generated the SOC curve of the regression result:



How do we get this graph?

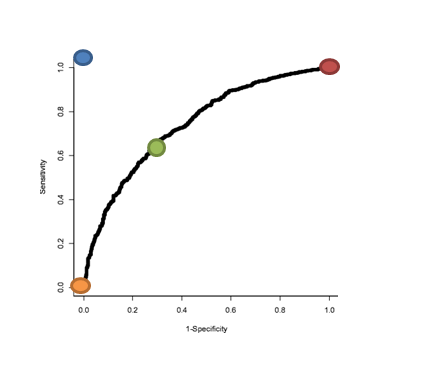
A. we try out different cutoff probability in the prediction, and then calculate the sensitivity and 1-specificity for each cutoff probability

B. we generate a scatter plot between the actually probability of Dependent Variable being 1, and the predicted probability of Dependent Variable being 1

C. we generate a scatter plot between the odds of Dependent Variable being 1 and the probability of Dependent variable being 1.

D. we generate a scatter plot between the odds of Dependent Variable being 1 and the probability of Independent variable being 1.

18. We did a logistic regression, and generated the SOC curve of the regression result:



Which point in the SOC curve represents the best case scenario for a logistic regression?

A. the Orange dot

B. the Blue dot

C. the green dot

D. the red dot