Exercises on Predictive Analytics

Carlo Gaetan [*]

January 7, 2019

Exercise 1. *Consider the joint density for the random variables X, Y*

$$f(x,y) = (x+y) I_{[0,1]\times[0,1]}(x,y),$$

1. *Find $p_{Y|X=x}(y|x)$.*

2. *Find $\mu(x) = \mathbb{E}\left[Y|X=x\right]$.*

3. *Find the best affine function $m(x) = b_0 + b_1 X$, i.e. find the values of b_0 and b_1 that minimize $MSE(b_0, b_1) = \mathbb{E}\left[(Y - b_0 - b_1 X)^2\right]$*

4. *Plot and compare them using R*

5. *Simulate 100 values from the random variable $\mathbb{E}\left[Y|X\right]$.*

Exercise 2. *Consider the pairs of data, (x_i, y_i) $i = 1, \dots, n$*

1. *Calculate the empirical correlation coefficient of (x_i, z_i) $i = 1, \dots, n$, where $z_i = a + b y_i$.*

Exercise 3. *Covariance, Correlation*

1. *Write a R function that calculates the empirical covariance and the empirical correlation coefficient according the formulas given in the slides.*

2. *Modify the previous function for delaing with missing data in one variable.*

[*] Some Problems (**RS**) are extracted from Fred Ramsey and Daniel Schafer (2013) *The Statistical Sleuth A Course in Methods of Data Analysis*, 3rd Edition, Cengage Learning

3. *Compare it with* `cov` *and* `cor`. *What is the difference ?*

4. *Download in R the file* `ex1.csv`, *calculate the empirical correlation between X and* Y.

5. *Now calculate the 5 correlations when* $Z = 1, 2, \ldots, 5$. *Could you give an explanation of this results*

6. *Find a real data example for three variables* X, Y, Z *for which* Z *is a confounder.*

Exercise 4. *Show that if* $\mathbb{E}\left[\epsilon|X = x\right] = 0$ *for all* x, *then* $\mathrm{Cov}\left[X, \epsilon\right] = 0$. *Would this still be true if* $\mathbb{E}\left[\epsilon|X = x\right] = a$ *for some other constant* a*?*

Exercise 5. *Find the variance of* $\hat{\beta}_0$. Hint: *Do you need to worry about covariance between* $\overline{Y}$ *and* $\hat{\beta}_1$*?*

Exercise 6. *Under the simple linear regression model, show that*

$$\mathbb{E}\left[(Y - (\beta_0 + \beta_1 X))^2\right] = \sigma^2$$

Exercise 7. *In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use* `set.seed(1)` *prior to starting part (1) to ensure consistent results.*

1. *Using the* `rnorm` *function, create a vector,* `x`, *containing 100 observations drawn from a* $N(0, 2)$ *distribution. This represents a predictor,* X.

2. *Using the* `rnorm` *function, create a vector,* `eps`, *containing 100 observations drawn from a* $N(0, 0.5)$ *distribution*

3. *Using x and eps, generate a vector y according to the model*

$$Y = -2 + 0.75X + \epsilon$$

4. *What is the value* $\mathrm{Cov}\left[X, Y\right]$

5. *Create a scatterplot displaying the relationship between x and y. Comment on what you observe.*

6. *Fit a least squares linear model to predict y using x. Comment on the model obtained.*

Exercise 8. *Consider the diamond data set from the* `UsingR` *package. You can install it with*

```
> install.packages("UsingR")
> data("diamond")
```

1. *Predict the diamond prices (in Singapore dollars) using a linear model in which the diamond weight in carat is a regressor. In particular calculate the prediction of the price at* 0.25 *carat*
2. *Consider the residuals of the model and verify numerically that are uncorrelated with the regressor*
3. *Calculate SS_{tot}, SS_{reg} and SS_{res} and R^2*
4. *What do you think about the goodness of fit ?*

Exercise 9. *Consider again the* `diamond` *data set*

1. *Calculate the standard error of the estimate of the slope using the formula in the slides (no devoted R commands)*
2. *Calculate a confidence interval of level 0.99 for the estimated slope using the formula in the slides (no devoted R commands)*
3. *Give the prediction interval of level 0.99 at $x = 337.5$*
4. *In the European market the price of the diamond ring is 3700 Singapore dollars per carat.*
   - (a) *Test if the data are consistent with a different quotation.*
   - (b) *Test if the data are consistent with a lower quotation.*

Exercise 10. *The aim of this exercise is to perform an empirical assesment of the fact that under the modelling assumption that*

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

where ϵ is independent across observations and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, independent of x, the distribution of the estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

is Gaussian.

1. *Set the number of observations* $n = 48$

2. *Set the 'true' values* $\beta_0 = -260$ *and* $\beta_1 = 3721$

3. *Simulate* $m = 1000$ *samples of size* $n = 48$ *with fixed values for* x *from the model (1)*

   ```
   > library(UsingR)
   > x<-diamond$carat
   ```

4. *For each sample estimate* β_1 *and compare the empirical distributon of the* $m = 1000$ *estimates with the theoretical one, by means of a q-q plot.*

Exercise 11 (This problem is taken from an exercise of Cosma Shalizi's)**.** *What if all null hypotheses were true? Draw a vector* Y *from a standard Gaussian distribution with 1000 observations. Draw a matrix* X *by setting* $p = 100$*, and giving each column* x_i *a standard Gaussian distribution.*

- *Regress* Y *on all 100* X*'s (plus an intercept).*

- *How many of the* β_i *s are significant at the 10% level ? At the 5% level? At the 1% level? What is the* R^2 *? The adjusted* R^2 *?*

- *Re-run the regression using just the variables which are significant at the 5% level. Plot a histogram of the change in coefficient for each variable from the old regression to the new regression. How many variables are now significant at the 1% level? What is the* R^2 *? The adjusted* R^2 *?*

Exercise 12. *Standard errors and correlations among the predictors. Assume that we have 2 predictors namely* $x_1 = (x_{11}, \ldots, x_{n1})^\top$ *and* $x_2 = (x_{12}, \ldots, x_{n2})^\top$*, so* $n^{-1}X^\top X$ *is a* 3×3 *matrix.*

- *Suppose that there is no sample covariance between the two predictors. Find* $(n^{-1}X^\top X)^{-1}$ *in terms of* $\overline{x}_i = n^{-1}\sum_{k=1} x_{ki}$ *and* $\overline{x}_i^2 = n^{-1}\sum_{k=1} x_{ki}^2$*,* $i = 1, 2$*. Simplify, where possible, to eliminate second moments in favor of variances.*

- *Give the general form of the inverse,* $(n^{-1}X^\top X)^{-1}$ *, without assuming there is no sample covariance between the two predictors. How, qualitatively, do the variances of the slope estimates depend on the variances and covariances of the predictors?*

Exercise 13. *For each of the following statements, indicate whether the statement is true or false, and explain your answer briefly.*

- *A predictor X cannot be statistically significant in a simple linear regression if the sample correlation of Y and X is 0.*
- *A predictor X_3 can be statistically significant in a multiple regression with X_1, X_2, X_3 even though the correlation of Y and X_3 is 0.*

Exercise 14. *1. In a simple linear regression show that*

   (a) $\hat{\beta}_1 = r(x, y)\dfrac{s_y}{s_x}$;

   (b) $R^2 = r(x, y)^2$.

2. In a simple linear regression, the F-statistic (F-ratio) tests the null hypothesis $H_0 : \beta_1 = 0$ and is equal to

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (\hat{y}_i - \hat{y}_i)^2/(n-2)}.$$

The t-statistic tests the null hypothesis $H_0 : \beta_1 = 0$ and is equal to

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}.$$

Because in a simple linear regression these two statistics are testing the same hypothesis, there should be a correspondence between the two. Demonstrate this correspondence by showing that the F-statistic is equal to the t-statistic squared. and use the definition of R^2.

Exercise 15 (Actually this taken from an assignment of Steve C. Wang, dataset `nenana.csv`)**.** *In 1917, railroad workers were building a bridge across the Tanana River in Alaska. As a diversion, they placed bets on when the ice in the river would start to break up, and a betting pool was started with a pot of \$800. Over the years, the contest has grown, and the Nenana Ice Classic (as it has come to be known) now sports a pot of over \$300,000 and is regulated by the state of Alaska as a legalized game of chance. Because of the large amount of money at stake, the exact moment of ice breakup has been recorded carefully each year, providing a consistent and high-quality source of data on local climatic change. The dataset gives the breakup date for each* *year from 1917 to 2007. The breakup date is given as the month/date/time, and also as a corresponding Julian date, which is a way of expressing the date in a numerical form that adjusts for leap years.*

1. *Is there any evidence for a warming trend? If so, over what time period has the trend occurred?*
2. *Is the trend linear or nonlinear?*
3. *Is there evidence for any short-term fluctuations or cycles?*

Exercise 16 (**RS**, dataset `ex0914.csv`). *Pace of Life and Heart Disease. Some believe that individuals with a constant sense of time urgency (often called type-A behavior) are more susceptible to heart disease than are more relaxed individuals. Although most studies of this issue have focused on individuals, some psychologists have investigated geographical areas. They considered the relationship of city-wide heart disease rates and general measures of the pace of life in the city. For each region of the United States (Northeast, Midwest, South, and West) they selected three large metropolitan areas, three medium-size cities, and three smaller cities. In each city they measured three indicators of the pace of life. The variable* ***walk*** *is the walking speed of pedestrians over a distance of 60 feet during business hours on a clear summer day along a main downtown street.* ***Bank*** *is the average time a sample of bank clerks takes to make change for two $20 bills or to give $20 bills for change. The variable* ***talk*** *was obtained by recording responses of postal clerks explaining the difference between regular, certified, and insured mail and by dividing the total number of syllables by the time of their response. The researchers also obtained the age-adjusted death rates from ischemic heart disease (a decreased flow of blood to the heart) for each city (****heart****).The variables have been standardized, so there are no units of measurement involved.*

1. *Obtain the least squares fit to the linear regression of* ***heart*** *on* ***bank****,* ***walk****, and* ***talk****.*
2. *Plot the residuals versus the fitted values. Is there evidence that the variance of the residuals increases with increasing fitted values or that there are any outliers?*
3. *Report a summary of the least squares fit. Write down the estimated equation with standard errors below each estimated coefficient.*

Exercise 17 (**RS**, dataset `ex0915.csv`). *The data on corn yields and rainfall,*

1. *We consider the corn yield and rainfall in six U.S. corn-producing states (Iowa, Nebraska, Illinois, Indiana, Missouri, and Ohio), recorded for each year from 1890 to 1927*
2. *Show that a a straight-line regression model for the corn yield where the rainfall is a predictor is not adequate.*
3. *Fit the multiple regression of corn yield on rain and rain*^{2} *.*
4. *Plot the residuals versus year. Is there any pattern evident in this plot? What does it mean? (Anything to do, possibly, with advances in technology?)*
5. *Fit the multiple regression of corn yield on rain, rain*^{2} *, and year. Write the estimated model and report standard errors, in parentheses, below estimated coefficients.*
   - (a) *How do the coefficients of rain and rain*^{2} *differ from those in the estimated model in (b)?*
   - (b) *How does the estimate of* σ *differ? (larger or smaller?)*
   - (c) *How do the standard errors of the coefficients differ? (larger or smaller?)*
   - (d) *Describe the effect of an increase of one inch of rainfall on the mean yield over the range of rainfalls and years.*
6. *Fit the multiple regression of corn yield on rain, rain*^{2} *, year, and year* × *rain. Is the coefficient of the interaction term significantly different from zero? Could this term be used to say something about technological improvements regarding irrigation?*

Exercise 18 (**RS**, dataset `ex0722.csv`). *As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces and the propodus heights of the claws on several crabs of three species.*

1. *Estimate the slope in the simple linear regression of log force on log height, separately for each crab species. Obtain the standard errors of the estimated slopes.*

2. *We want to compare the slopes for C. productus and L. bellus and then to compare the slopes for C. productus and H. nudus. Argue that the standard error for the difference in two slope estimates is the following:*

$$SE[\hat{\beta}_{1,i} - \hat{\beta}_{1,j}] = \sqrt{SE[\hat{\beta}_{1,i}]^2 + SE[\hat{\beta}_{1,j}]^2}$$

   where $\hat{\beta}_{1,j}$ represents the estimate of slope from the species j. Which hypothesis did you use ?

3. *Use a t-test*

$$T = \frac{\hat{\beta}_{1,i} - \hat{\beta}_{1,j}}{SE[\hat{\beta}_{1,i} - \hat{\beta}_{1,j}]}$$

   with the sum of the degrees of freedom associated with the two standard errors. What do you conclude?

4. *Draw a scatterplot of claw closing force versus propodus height (both on a log scale), with different plotting symbols to distinguish the three different crab species.*

5. *Fit the multiple regression of log force on log height and species (as a factor). Provide the estimated model including standard errors of estimated regression coefficients.*

6. *Plot the fitted model over the scatterplot in 4.*

7. *Use a model with interaction and explain how we can perform the previous tests, using the multiple regression model.*

Exercise 19. *Harrison and Rubinfeld in 1978 proposed an hedonic model for determining the willingness of house buyers to pay for clean air. An hedonic model is a model that decomposes the price of an item into separate components that determine its price. For example, an hedonic model for the price of a house may decompose its price into the house characteristics, the kind of neighborhood, and the location. The study of Harrison and Rubinfeld employed data from the Boston metropolitan area, containing 560 suburbs and 14 variables. The Boston dataset is available through the dataset `Boston` in the `MASS` package. The goals of this exercise are:*

1. *Identify the dummy variables are in the Boston dataset*

2. *Quantify the influence of the predictor variables without interaction in the housing prices.*

3. *Obtain the ”best possible” additive model for decomposing the housing prices and interpret it.*

4. *Consider the interactions in the ”best possible” model for decomposing the housing prices and interpret it.*

   Note: *A fast way of accounting interactions between predictors is to use the ˆ operator in* `lm`*:*

   ```
   > lm(y ~ (x1 + x2 + x3)^2)
   ```

   equals

   ```
   > lm(y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3)
   ```

   It is possible to regress on all the predictors and the first order interactions using

   ```
   > lm(y ~ .^2)
   ```

   Further flexibility in `lm` *is possible, e.g. removing a particular interaction with*

   ```
   > lm(y ~ .^2 - x1:x2)
   ```

   or forcing the intercept to be zero with

   ```
   > lm(y ~ 0 + .^2)
   ```

Exercise 20. *Harrison and Rubinfeld in their paper considered several nonlinear transformations of the predictors and the response are done to improve the linear fit. Also, different units are used for* `medv`*,* `black`*,* `lstat`*, and* `nox`*. The authors considered these variables:*

- *Response:* `log(1000 * medv)`
- *Linear predictors:* `age`*,* `black/1000` *(this variable corresponds to their* $(B-0.63)^2$ *),* `tax`*,* `ptratio`*,* `crim`*,* `zn`*,* `indus`*, and* `chas`*.*
- *Nonlinear predictors:* `rm^2`*,* `log(dis)`*,* `log(rad)`*,* `log(lstat/100)`*, and* `(10 * nox)^2`*.*

1. *Check if the model with such predictors corresponds to the one in the first column, Table VII, page 100 of Harrison and Rubinfeld. Call this model* `modelHarrison`

2. *Make a stepwise selection of the variables (using BIC) call this model* `modelHarrisonSel`*.*

3. *Which model has a larger R^2? And adjusted R^2? Which is simpler and has more significant coefficients?*

Exercise 21. *What is the formula for the residual deviance in case of*

1. *a logistic regression ?*

2. *a Poisson regression ?*

Exercise 22 (Exercise from Faraway)**.** *The ships dataset found in the* `MASS` *package gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation.*

1. *Develop a model for the rate of incidents, describing the effects of important predictors.*

Exercise 23. *Consider the wheezing data (available as data set* `ohio` *in the* `faraway` package*).*

1. *Fit a logistic regression model with age and smoke as factors.*

2. *Check the significance of the different effects using likelihood ratio test and omit nonsignificant effects.*

3. *Compare with a model with age as a covariate (i.e. a single slope parameter for age).*

4. *Compare the fit of this model with the previous model using a likelihood ratio test.*

5. *Take a look at deviance and Pearson residual plots.*

Exercise 24. *Challenger O-ring data (Dalal et al. 1989, Risk analysis of the space shuttle, Journal of the American Statistical Association, 84, 945-957.) The space shuttle Challenger suffered a catastrophic failure January 27, 1986. The solid rocket motor disintegrated after the several of o-rings failed to contain combustion gases. (There were two o-rings on each joint - at least one joint was breached, meaning that both the primary and secondary o-rings were breached. Each o-ring was 37.5 feet in diameter and .28 inches thick). Some testing had been done on the reliability of the o-rings, and it was believed that the resiliency depended on ambient air temperature and pressure. The lowest temperature at launch, before the January 27 event was 53 degrees F; at the time of the launch, the temperature was forecast to be 31 degrees F. Information on reliability was obtained by recovering 23 boosters used in 24 of the preceding flights. There are 6 primary o-rings that may fail, some of which did fail. None of the secondary o-rings failed in these 24 flights.*

The number of primary o-rings failures, number of rings, joint temperature, and field pressure are recorded in the Excel data set challenger.csv.

1. *Fit a logistic regression model to the Challenger data and investigate the relationship between o-ring failure and pressure and temperature.*

2. *In particular, use the fitted model to obtain graph showing the estimated probability of primary o-ring failure as a function of temperature over the range 30 to 85 degrees.*

3. *Estimate the temperatures at which the probability of o-ring failure is 0.2, 0.4, 0.6 and 0.8. Discuss your findings in a brief report.*

Final note

For more exercises, please follow this link.