Week 8 Quiz - Regression - Solutions

FIT5197 teaching team

Note you will need to use the unit Formula Sheet (https://lms.monash.edu/mod/resource/view.php?id=7439150) to answer the following questions.

Question 1

Last year, five randomly selected students took a math aptitude test before they began their statistics course. In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades.

Student	x_i	y_i
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

The Statistics Department has three questions:

(a) What linear regression equation best predicts statistics performance, based on math aptitude scores?

(b) If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

(c) How well does the regression equation fit the data?

(Adapted from the web page: Linear Regression Example (https://stattrek.com/regression/regression-example.aspx) - do not go to this webpage until you have attempted the solution.)

Answer 1

Note this solution uses different forms of the simple linear regression equations compared to those shown in the unit formula sheet in order to expose you to the idea that these equations can be expressed in different ways but still calculate the same things. As a separate exercise you can play around with the equations provided in the unit formula sheet to see how they can be written in the forms provided in this solution.

(a) The simple linear regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1 x$. To conduct a regression analysis, we need to find the parameter estimates $\hat{b}_0$ and $\hat{b}_1$. We will need to compute the difference between the student's score and the average score ($x_i - \bar{x}$, $y_i - \bar{y}$), the squares of the deviation scores ($(x_i - \bar{x})^2$, $(y_i - \bar{y})^2$) and the product of the deviation scores ($(x_i - \bar{x})(y_i - \bar{y})$).

Student	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	95	85	17	8	289	64	136
2	85	95	7	18	49	324	126
3	80	70	2	-7	4	49	-14
4	70	65	-8	-12	64	144	96
5	60	70	-18	-7	324	49	126
Sum	390	385			730	630	470
Mean	78	77					

Based on the table above, we can first solve the regression coefficient ($\hat{b}_1$):

$$\begin{aligned} \hat{b}_1 &= \sum[(x_i - \bar{x})(y_i - \bar{y})] / \sum[(x_i - \bar{x})^2] \\ \hat{b}_1 &= 470/730 \\ \hat{b}_1 &= 0.644 \end{aligned}$$

(b) Once we know the value of the regression coefficient estimate $\hat{b}_1$, we can solve for the regression y-intercept estimate $\hat{b}_0$:

$$\begin{aligned} \hat{b}_0 &= \bar{y} - \hat{b}_1 * \bar{x} \\ \hat{b}_0 &= 77 - 0.644 \times 78 \\ \hat{b}_0 &= 26.768 \end{aligned}$$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0644x$.

Once we have the regression equation, we can choose a value for the independent variable (x), perform the computation, and then have an estimated value ($\hat{y}$) for the dependent variable. In this example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade ($\hat{y}$) would be:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$
$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 \times 80$$
$$\hat{y} = 26.768 + 51.52 = 78.288$$

\textbf{Warning:} When you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called \textbf{extrapolation}, and it can produce unreasonable estimates. In this example, the aptitude test scores used to create the regression equation ranged from 60 to 95. Therefore, only use values inside that range to estimate statistics grades. Using values outside that range (less than 60 or greater than 95) is problematic.

(c) When we want to know how well the regression equation fits the data, One way to assess fit is to check the coefficient of determination, which can be computed from the following formula: [R^2 = {(1/N) * \sum[(x_i-\bar{x})(y_i-\bar{y})]/(\sigma_x*\sigma_y)}^2] where N is the number of observations used to fit the model, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

First, we compute the standard deviation of x:

$$\sigma_x = \sqrt{\sum (x_i - \bar{x})^2 / N}$$
$$\sigma_x = \sqrt{730/5} = \sqrt{146} = 12.083$$

Next, we find the standard deviation of y:

$$\sigma_y = \sqrt{\sum (y_i - \bar{y})^2 / N}$$
$$\sigma_y = \sqrt{630/5} = \sqrt{126} = 11.225$$

Note that this solution is using the biased estimate of the sample variance above, but our unit formula sheet uses the unbiased estimate of the sample variance. We prefer that you use the unbiased estimator, although the most important thing is that you tell us which estimator you are using when you are solving a problem so we can see why your answer might differ from what we might expect if the unbiased estimator was used.

Finally, we compute the coefficient of determination:

$$R^2 = \{(1/N) * \sum [(x_i - \bar{x})(y_i - \bar{y})]/(\sigma_x * \sigma_y)\}^2$$
$$R^2 = [(1/5) \times 470/(12.083 \times 11.225)]^2$$
$$R^2 = (94/135.632)^2 = (0.693)^2 = 0.48$$

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a reasonable fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

Question 2

Below is a table of the number of cricket chirps as a function of outdoor temperature:

Temperature (^{o}C)	# Cricket chirps
20	88.59
16	71.59
19.79	93.3
18.39	84.3
17.1	80.59
15.5	75.19
14.69	69.69
17.1	82
15.39	69.4
16.2	83.3
15	79.59
17.2	82.59
16	80.59
17	83.5
14.39	76.3

(a) Build a simple linear regression model to predict the number of cricket chirps as a function of outdoor temperature, and complete your solution by giving the linear prediction formula.

The key statistics for data in this table are:

$$\begin{aligned} n &= 15 \\ \overline{chirps} &= 80.04 \\ \overline{chirps^2} &= 6448.39 \\ \overline{temp} &= 16.65 \\ \overline{temp^2} &= 280.04 \\ \overline{chirps * temp} &= 1341.83 \end{aligned}$$

(b) What is the co-efficient of determination, R^2, of this model? Based on this R^2 value, is this a good model of the data? What is the reference model used in the calculation of R^2 in this case?

(c) Why is linear regression called linear regression?

Answer 2

Note this solution uses the simple linear regression equations given in the unit formula sheet in order to expose you to how you might answer a simple linear regression question if one is given in the exam.

(a) The solution for β_1 is

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\overline{XY} - \overline{X}\ \overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{1341.83 - 16.65 \cdot 80.04}{280.04 - 16.65^2} = 3.25$$

and the solution for β_0 is

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 80.04 - 3.25 \cdot 16.65 = 25.93$$

So prediction formula is $E[chirps_i | temp_i] = 25.93 + 3.25 temp_i$

(b) The $\mathbf{R^2}$ value is computed as

$$R^2 = 1 - \frac{RSS}{SS_{YY}} = \frac{SS_{XY}^2}{SS_{XX}SS_{YY}}$$

Using the equations in the unit formula sheet and the key statistics above we find $SS_{XY} = 9.16$, $SS_{XX} = 2.82$ and $SS_{YY} = 41.99$, so $R^2 = 0.71$. Thus the model is reasonably good since it is closer to 1 than it is to 0. The reference model in this case is $E[chirps_i] = \text{sample mean of } chirps = 80.04$

(c) It is called regression because it fits a model that maps input data to a continuous output variable. It is called linear because the models being fitted are linear with respect to the parameters. E.g. in the simple linear regression formula $y = \beta_0 + \beta_1 x$, if we consider y as a function of β_0 we see it is a line with slope 1 and y-intercept $\beta_1 x$, whereas if we consider y as a function of β_1 we see it is a line with slope x and y-intercept β_0.

R code hackers nail-biting challenge

a) Consider a random variable x from range of 0 to 2π. Using the 'runif' function for the uniform distribution, obtain $n = 10$ random observations of x. Now create a variable y from the observations of x using the sine function $y = \sin(x)$ and adding uniformly distributed noise selected on the interval $[-0.5, 0.5]$.

b) Plot a scatterplot for the x and y variables. Add a plot of the original sine function without noise. Create 4 polynomial linear regression models by fitting to the data using the 'lm' function where the degrees of the polynomials are 1, 2, 3 and 8. Discuss the differences. Which polynomials provide the best fits to the sine function? Which polynomial fit shows underfitting? Which polynomial fit shows overfitting?

c) Increase the number of observations of x and y from $n = 10$ to 100. Plot the 8-degree polynomial regression model again after fitting on these 100 points. Is it the same as previous plot? What are the differences and why?

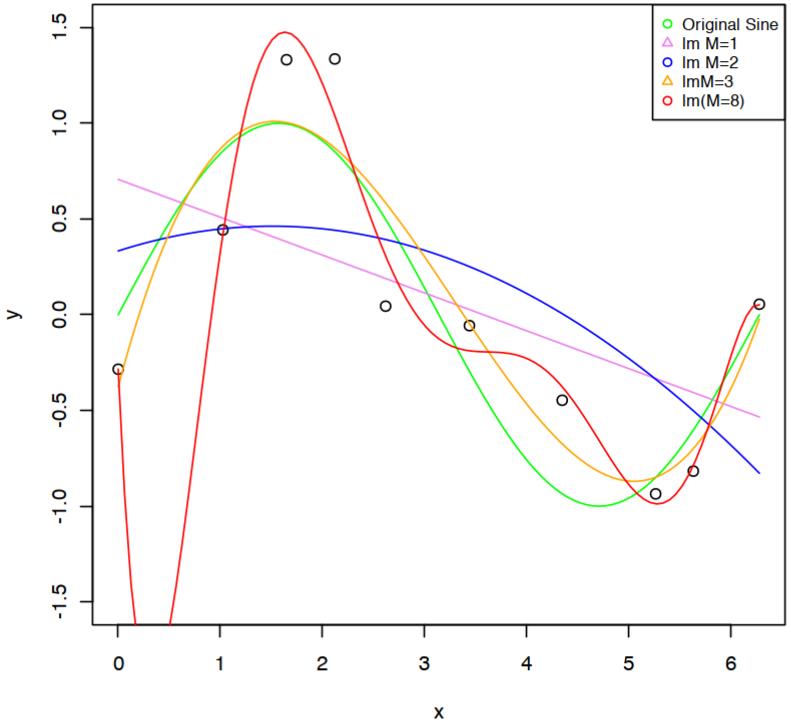
Answer

In [1]:

```
noisy_sine_data <- function(n){
    x <- runif(n-2, 0, 2*pi)
    x <- c(0, x, 2*pi)
    noises <- runif(n, 0, 0.5)
    y <- c()
    for(i in 1:n){
      xi <- x[i]
      noise <- noises[i]
      yi<- sin(xi)
      if(i%%2 == 0){
        yi <- yi + noise
      }
      else{
        yi <- yi - noise
      }
      y <- c(y, yi)
    }
    data <- data.frame(x, y)
    return(data)
}
```

In [5]:

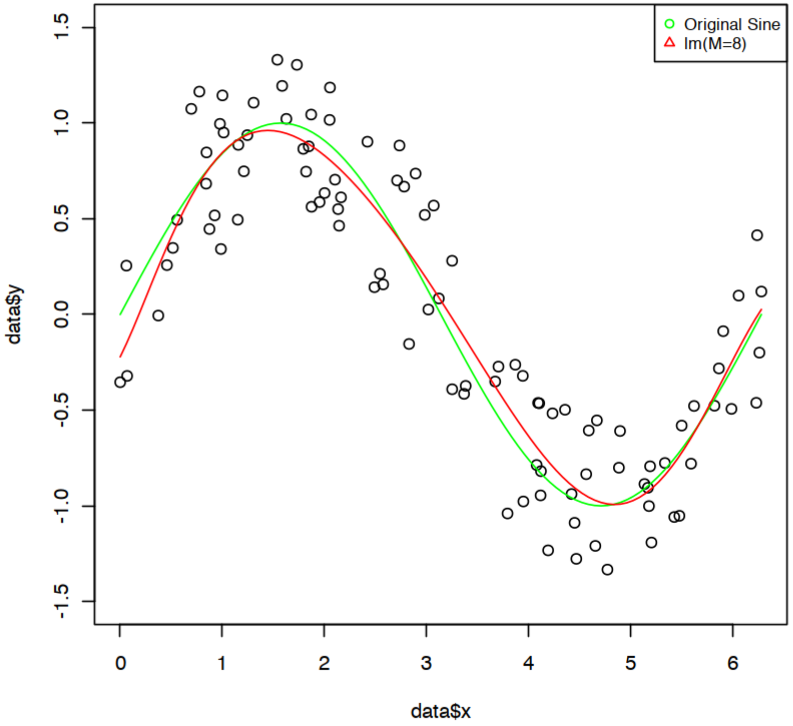
```
data <- noisy_sine_data(10)
x <- data$x
y <- data$y
plot(x, y, ylim=c(-1.5,1.5))
plot(sin, 0, 2*pi, add=T, type="l", col="green")
model8 <- lm(y ~ poly(x,degree=8,raw=TRUE), data = data)
model1 <- lm(y ~ poly(x,degree=1,raw=TRUE), data = data)
model2 <- lm(y ~ poly(x,degree=2,raw=TRUE), data = data)
model3 <- lm(y ~ poly(x,degree=3,raw=TRUE), data = data)
curve(predict(model1, data.frame(x=x)),add=T, col="violet")
curve(predict(model2, data.frame(x=x)),add=T, col="blue")
curve(predict(model3, data.frame(x=x)),add=T, col="orange")
curve(predict(model8, data.frame(x=x)),add=T, col="red")
legend("topright", c("Original Sine", "lm M=1", "lm M=2", "lmM=3", "lm(M=8)"),cex=.8,
       col=c("green", "violet", "blue", "orange", "red"),pch=c(1,2))
```



The model when M=3 i.e.(y=a+bx+cx^2+dx^3) fits the best. When M = 1 or 2, model is very simple. There are fewer parameters to learn. So, the model underfits the data giving high bias and low variance. When M = 8, there are relatively more parameters(9) to learn, so the model overfits giving low bias and high variance.

In [11]:

```
data <- noisy_sine_data(100)
x <- data$x
y <- data$y
plot(data$x, data$y, ylim=c(-1.5,1.5))
plot(sin, 0, 2*pi, add=T, type="l", col="green")
model8 <- lm(y ~ poly(x,degree=8,raw=TRUE), data = data)
curve(predict(model8, data.frame(x=x)),add=T, col="red")
legend("topright", c("Original Sine", "lm(M=8)"),cex=.8,
       col=c("green", "red"),pch=c(1,2))
```



The problem of overfitting can be resolved by increasing more training data. We can see here that when M = 8, the plot is very close to true function because we have increased training data from 10 to 100. This helps in solving the overfitting issue. If for some reason the ground truth sinusoid changed slowly over time, we would have to get new data and update the fit, otherwise we will encounter overfitting to the old data and the model will no longer be a good fit to new ground truth. This all depends on the how quickly the ground truth changes and how quickly you can get new data. In the real world most problems don't involve stationary problems, like a fixed sinusoid. Often the problems are nonstationary, i.e. the sinusoid keeps changing, but hopefully not too quickly so we can keep up.