Predicting Course Grades Using Midterm Scores
Part III: Predicting Probability of Earning an A or A-
(with Solutions)

Let's continue to analyze real, historical, student performance data from the class. The goal here is to model the relationship between midterm and the probability of earning an A or A- in the class so that we can later predict given probability based on midterm scores. The dataset we will use is in the *grades.csv* file. Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is students.

variable	description
midterm	students' scores in the midterm (from 0 to 100 points)
final	students' scores in the final exam (from 0 to 100 points)
overall	students' scores in the class overall (from 0 to 100 points)
gradeA	identifies students who earned an A or an A minus in the class

Table 1: Variables in "grades.csv"

In this problem set, we practice fitting a line to make predictions when Y is binary, including computing correlations, creating scatter plots, adding the fitted line to the scatter plot, and computing R^2.

As always, we start by loading and looking at the data:

```
## load and look at the data
grades <- read.csv("grades.csv") # reads and stores data
head(grades) # shows first observations
##   midterm final overall gradeA
## 1   79.25 47.00    69.2      0
## 2   96.25 87.75    94.3      1
## 3   58.25 37.75    62.0      0
## 4   54.50 62.00    72.4      0
## 5   83.00 39.75    72.4      0
## 6   41.75 49.50    59.5      0
```

1. First, let's figure out what each observation represents, identify our X and Y variables, and explore whether they are moderately or strongly linearly associated with each other.

    a. In this dataset, what does each observation represent? (2.5 points)

    Answer: Each observation represents a student. (Note: As indicated above, the unit of observation in this dataset is students.)

b. What should be our X variable? In other words, which variable are we going to use as the predictor? Please provide the name of the variable and identify whether it is binary or non-binary. (2.5 points)

Answer: *midterm*, which is a non-binary variable

c. What should be our Y variable? In other words, which variable are we going to use as the outcome variable? Please provide the name of the variable and identify whether it is binary or non-binary. (2.5 points)

Answer: *gradeA*, which is a binary variable

d. Compute the correlation coefficient between X and Y. Is the relationship between X and Y moderately or strongly linear? A yes/no answer will suffice. (2.5 points)

R code:

```
cor(grades$gradeA, grades$midterm) # computes correlation
## [1] 0.6422328
```

(Recall: The function in R to compute a correlation coefficient is cor(). The only two required arguments are the code identifying the two variables. The order of the variables does not matter since cor(X,Y) = cor(Y,X).)

Answer: Yes, the relationship between the two variables is moderately linear since the correlation coefficient is closer to 1 than to 0.

2. Second, let's fit the linear model that we will use to make predictions.

a. Use the function lm() to fit a linear model to summarize the relationship between X and Y and store the output in an object called *fit*. Then, ask R to provide the contents of *fit* by running its name. (R code only.) (5 points)

R code:

```
# fit linear model and store it in an object called fit
fit <- lm(grades$gradeA ~grades$midterm)
```

```
fit # provides contents of object
##
## Call:
## lm(formula = grades$gradeA ~grades$midterm)
##
## Coefficients:
##     (Intercept)  grades$midterm
##         -1.34305         0.02122
```

(Recall: The function lm() fits a linear model. It requires a function of the type Y ~ X, where Y identifies the Y variable (*gradeA*, in this case) and X identifies the X variable

(*midterm*, in this case). To specify the dataframe where the variables are stored, we can use either the `$` operator (as in the code above) or the optional argument `data`. If we wanted to use the latter, the code to fit the linear model would be `lm(gradeA ~ midterm, data = grades)`.)

b. What is the fitted line? In other words, provide the formula $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$ where you specify each term (i.e., substitute Y for the name of the outcome variable, substitute $\widehat{\alpha}$ for the estimated value of the intercept coefficient, substitute $\widehat{\beta}$ for the estimated value of the slope coefficient, and substitute X for the name of the predictor.) (5 points)

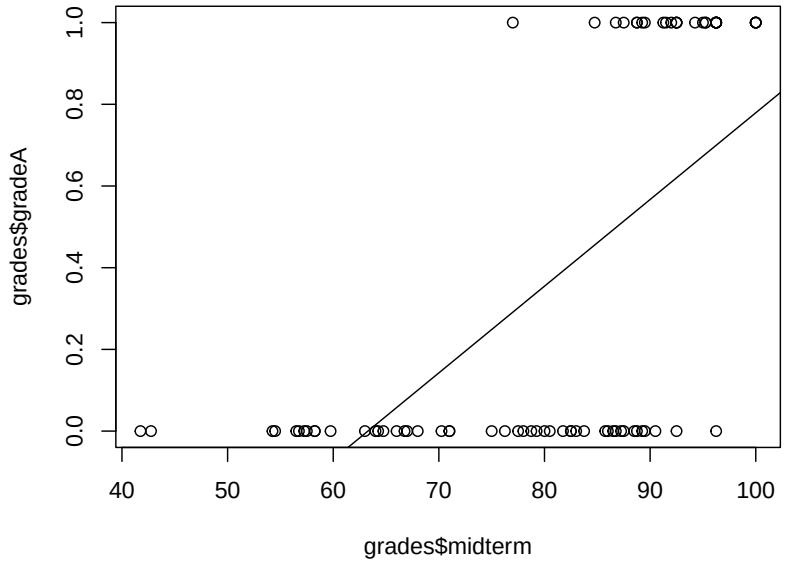
Answer: $\widehat{\text{gradeA}}$ = -1.3431 + 0.0212 *midterm*

(Note: The Y variable is *gradeA*, $\widehat{\alpha}$= -1.3431, $\widehat{\beta}$=0.0212, and the X variable is *midterm*.)

c. Create a visualization of the relationship between X and Y and add the fitted line to the graph using the function `abline()`. (R code only.). (5 points)

R code:

```
plot(grades$midterm, grades$gradeA) # creates scatter plot
abline( fit ) # adds fitted line
```



(Recall: a scatter plot is the graphical representation of the relationship between two variables. The function in R to create a scatter plot is `plot()`. It requires two arguments (separated by a comma) and in this particular order: (1) the code identifying the variable to be plotted along the x-axis, and (2) the the code identifying the variable to be plotted along the y-axis. We always plot the predictor along the X-axis and the outcome variable along the Y-axis. Alternatively, if we do not want the order of the arguments to matter, we could specify the names of the arguments, `x` and `y`, in the code. For example, `plot(x=grades$midterm, y=grades$gradeA)` and `plot(y=grades$gradeA,`

x=grades$midterm) would produce the same scatter plot as the one above. Recall: The function abline() adds lines to the most recently created graph. To add the fitted line, we specify as the main argument the name of the object where we stored the output of the lm() function, *fit* in this case.)

3. Now, let's use the fitted line to make some predictions.

   a. Computing $\widehat{Y}$ based on X: Suppose that you earn 80 points in the midterm. What would be your best guess of your predicted probability of earning an A or A- in the course based on your performance in the midterm? Please show your calculations and then answer the question with a full sentence (including units of measurement). (5 points)

   Calculations:

$$\begin{aligned}\widehat{\text{gradeA}} &= \widehat{\alpha} + \widehat{\beta}\,\text{midterm} \\ &= -1.3431 + 0.0212\,\text{midterm} \\ &= -1.3431 + 0.0212 \times 80 \quad \text{(if midterm=80)} \\ &= -1.3431 + 1.696 = 0.3529\end{aligned}$$

   Answer: If I earn 80 points in the midterm, I would predict that my probability/chances of earning an A or an A- in the course is of 35.29%, on average. (Note: $\widehat{Y}$ is in the same unit of measurement as $\overline{Y}$; in this case, Y is binary so $\overline{Y}$ and $\widehat{Y}$ are in percentages, after multiplying the outputs by 100. Note: We could have arrive at this same conclusion by looking at the scatter plot with the fitted line above. All we would need to do is (i) find 80 on the X-axis, (ii) go up to the line, and (iii) find the value on the Y-axis associated with that point on the line.)

   b. Computing $\widehat{Y}$ based on X: Now, suppose that you earn 90 points in the midterm. What would be your best guess of your predicted probability of earning an A or A- in the course based on your performance in the midterm? Please show your calculations and then answer the question with a full sentence (including units of measurement). (5 points)

   Calculations:

$$\begin{aligned}\widehat{\text{gradeA}} &= \widehat{\alpha} + \widehat{\beta}\,\text{midterm} \\ &= -1.3431 + 0.0212\,\text{midterm} \\ &= -1.3431 + 0.0212 \times 90 \quad \text{(if midterm=90)} \\ &= -1.3431 + 1.908 = 0.5649\end{aligned}$$

   Answer: If I earn 90 points in the midterm, I would predict that my probability/chances of earning an A or an A- in the course is of 56.49%, on average. (Note: $\widehat{Y}$ is in the same unit of measurement as $\overline{Y}$; in this case, Y is binary so $\overline{Y}$ and $\widehat{Y}$ are in percentages, after multiplying the outputs by 100. Note: We could have arrive at this same conclusion by looking at the scatter plot with the fitted line above. All we would need to do is (i) find 90 on the X-axis, (ii) go up to the line, and (iii) find the value on the Y-axis associated with that point on the line.)

c. Computing $\triangle\widehat{Y}$ based on $\triangle X$: What is the predicted change in the probability of earning an A or an A- in the class associated with an increase in midterm scores of 10 points? Please show your calculations and then answer the question with a full sentence (including units of measurement). (10 points)

Calculations:

$$\begin{aligned}\triangle\widehat{\text{gradeA}} &= \widehat{\beta}\,\triangle\text{midterm} \\ &= 0.0212 \times \triangle\text{midterm} \\ &= 0.0212 \times 10 \text{ (if } \triangle \text{ midterm=10)} \\ &= 0.212\end{aligned}$$

Answer: An increase of midterm scores of 10 points is associated with a predicted increase in the probability of earning an A or an A- in the class of 21.2 percentage points, on average. (Note: $\triangle\widehat{Y}$ is in the same unit of measurement as $\triangle\overline{Y}$; in this case, Y is binary and so $\triangle\overline{Y}$ and $\triangle\widehat{Y}$ are measured in percentage points, after multiplying the outputs by 100. Recall that percentage point is the unit of measurement of the arithmetic difference between two percentages. Note: We could have arrive at this same conclusion by looking at the scatter plot with the fitted line above. All we would need to do is (i) pick two values on the X-axis that differ by 10, for example: 90 and 80, (ii) find the values on the Y-axis associated with each of those two points, in this case: 0.5649 and 0.3529 and (iii) compute the difference between those two Y values: 56.49% - 35.29% = 21.2 percentage points.)

4. What is the R^2 of the fitted model? And, how would you interpret it? (Hint: the function cor() might be helpful here.) (5 points)

R code:

```
cor(grades$gradeA, grades$midterm)^2 # computes R^2
## [1] 0.4124629
```

Answer: The linear model using *midterm* as a predictor explains 41% of the variation of *gradeA*. (Note: R^2 measures the proportion of the variation of the outcome variable explained by the model. In the simple linear model: $R^2 = \text{cor(X,Y)}^2$. Since R^2 is relatively far from 1, it looks like it is not a very good predictive model. As we can see in the scatter plot above, the prediction errors—the vertical distance between the dots and the line—are relatively large.)