Math 4311 Lab 1: Regression

1. **Exploratory data analysis.**

   (a) The data for this lab consist of math and verbal SAT scores for a sample of students, stored in the file `sat-data.txt`. Save this data to your computer, and import it into Rstudio (upper-right corner of your screen). Make sure that "heading" is set to "yes".

   (b) You can always view a list of all variables in your workspace with the command `ls()`. Using this command, you will see a variable `sat.data` in your workspace.

   (c) How many students are represented in this data set? (Use `dim(sat.data)` to see the dimensions of your data set.)

   (d) Use `head(sat.data)` to see the first few rows of your data. This is a good way to see all of the variable names in the data set and their data types.

   (e) Throughout this lab, let's represent math and verbal SAT scores by X and Y, respectively. These variables can be summarized by their means, μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation coefficient ρ. Let's estimate these values as follows:

   - First, store the math and sat scores into their own variables with the commands `math=sat.data$math` and `verbal=sat.data$verbal`.
   - Compute the sample means, $\bar{x}$ and $\bar{y}$, using `mean(math)` and `mean(verbal)`.
   - Compute the sample standard deviations, s_x and s_y, using the `sd` command.
   - Compute the sample correlation coefficient, r, using `cor(math,verbal)`.

   (f) Generate histograms for `math` and `verbal` using the `hist` command. What approximate distributions do these variables have?

   (g) Finally, generate a scatterplot of math and verbal SAT scores using `plot(math,verbal)`.

2. **The regression model.**

   (a) Let's fit a regression model $Y = a + bX$ for predicting verbal SAT based on math SAT:

   `model=lm(verbal~math,data=sat.data)`.

   Here, `lm` stands for "linear model", which is another term for linear regression model.

   (b) View a summary of the model: `summary(model)`. Because we are relying on a sample of data rather than the entire population, this model estimates the `(Intercept)` term a and the coefficient of `math` b. We also have standard errors for these estimates, their t-statistics, and their observed significance levels/p-values, which will be discussed later in the course.

   (c) The true values of the parameters a and b satisfy these equations:

   - $b = \rho \frac{\sigma_Y}{\sigma_X}$
   - $a = \mu_Y - b\mu_X$

   Analogously, the estimates $\hat{a}$ and $\hat{b}$ are computed using these equations:

   - $\hat{b} = r \frac{s_y}{s_x}$
   - $\hat{a} = \bar{y} - \hat{b}\bar{x}$

   Verify that the estimates $\hat{a}$ and $\hat{b}$ given in the model summary satisfy these equations.

(d) Now, let's superimpose the regression line onto the scatterplot:

```
ahat=coef(model)[1]
bhat=coef(model)[2]

x.index=200:800
plot(math,verbal)
lines(x.index,ahat+bhat*x.index,col="red")
```

This plot allows us to assess the performance of the model visually. The more tightly the data fits the regression line, the more accurate the model will be. In this case, we have a correlation coefficient of $r = 0.46$, which is not great, but the correlation between math and verbal SAT is statistically significant (more on this later).

3. **Quantitative Model Performance.** We have now seen that the regression model doesn't fit the data perfectly. We can account for the noise in the model by introducing the *disturbance terms* ε_i

$$Y_i = a + bX_i + \varepsilon_i, \text{ for } i = 1, \ldots, n.$$

We are now accounting for the $n = 1,000$ rows of data in our sample. For each $i = 1, \ldots, n$, X_i and Y_i are the math and verbal SAT scores for that student, and ϵ_i is the corresponding disturbance term, which represents the regression model's error for that student.

Generally, $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be statistically independent variables with mean zero and variance $\sigma^2 > 0$, and some applications even require normality, $\varepsilon_i \sim N(0, \sigma^2)$. Of course, we can't directly observe a or b, so we also can't observe the ε_i's. Instead, we estimate them using *residual terms* $e_1, \ldots, e_n$:

$$Y_i = \hat{a} + \hat{b}X_i + e_i, \text{ for } i = 1, \ldots, n.$$

$$e_i = Y_i - (\hat{a} + \hat{b}X_i).$$

(a) **Store all 1,000 disturbance terms in a vector:** `e = verbal - (ahat + bhat*math)`.
For models with quantitative dependent variables Y, four commonly used performance metrics are the root mean square error (RMSE), normalized root mean square error (NRMSE), coefficient of determination R^2, and the mean absolute error (MAE).

$$\begin{aligned}
\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \\
\text{NRMSE} &= \frac{\text{RMSE}}{s_y} \\
R^2 &= 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n}{n-1}(\text{NRMSE})^2 \\
\text{MAE} &= \frac{1}{n}\sum_{i=1}^{n} |e_i|.
\end{aligned}$$

(b) **Compute these performance metrics.**
The first three equations above prove that RMSE, NRMSE, and R^2 are monotonic functions of each other, so they are essentially equivalent for assessing model performance. R^2 has the advantage that its maximum value is 1, indicating perfect model fit. As discussed in class, R^2 can produce misleading conclusions when applied to inappropriate models, which emphasizes the importance of exploratory data analysis.

(c) **Verify that $R^2 = r^2$ for the SAT regression model.**
This is true for all simple linear regression models, but it's not necessarily true for other machine learning algorithms, such as random forests or neural networks. In fact, R^2 can even be negative if model performance is sufficiently poor.