# Simple Regression

STAT/BIOS 823

Homework 8

## Directions

Using `RMarkdown` in `RStudio`, complete the following questions. Launch RStudio and open a new RMarkdown file or use the class RMarkdown template provided and save it on your working directory as a `.Rmd` file. At the end of the activity, save your **pdf** generated from `RMarkdown+Knitr` and submit your homework on the Blackboard.

If you have questions, please post them on the lesson discussion board.

**All** questions are mandatory. Some **R-codes** and **output** from the code have been provided for you.

`R code` and output must be clearly shown.

Homework submitted after the due date will attract a penalty of 10 points per day after the due date.

## 1 Goodness of Fit

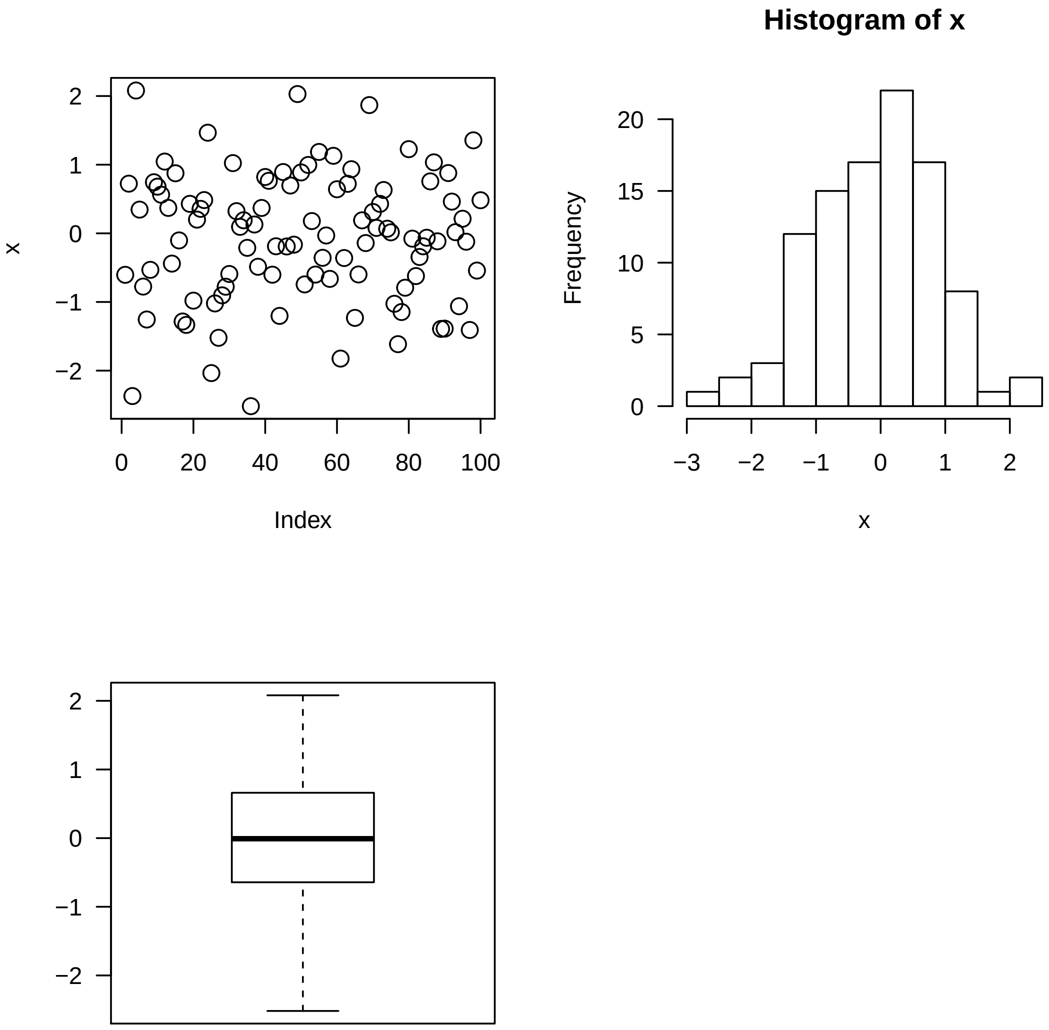
It is useful to have some measure of how well a model fits the data. One common choice for linear regression is $R^2$, the coeficient of determination, which is interpreted as the proportion of variance (of the outcome) explained by the model:

$$R^2 = 1 - \frac{\text{unexplained variation in y}}{\text{total variation in y}} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSE}{SSY}$$

Its range is $0 \leq R^2 \leq 1$. Values closer to 1 indicating better fits. For simple linear regression, $R^2 = r^2$ where $r$ is the linear correlation between $x$ and $y$. The linear correlation coeficient $r$ (Pearson's correlation coeficient) has a range $-1 \leq r \leq 1$, with values close to $-1$ indicating a strong negative linear association between $x$ and $y$, and values close to $+1$ indicating a strong positive linear association between $x$ and $y$. Let's investigate using simulation.

1. Generate a vector $x$ containing 100 random numbers from a standard normal distribution and visualize the data:

```r
x <- rnorm(100)
par(mfrow = c(2, 2))
plot(x, las = 1, cex = 1.5)
hist(x, las = 1, cex = 1.5)
boxplot(x, las = 1, cex = 1.5)
```



2. To induce a linear relationship between $x$ and a dependent variable $y$, generate the vector $y$ using a linear transformation of the vector $x$:

```r
y1 <- 0 + 1 * x  ## A perfect linear association
y2 <- 0 + 1 * x + rnorm(length(x))  # Add a little N(0,1) noise (error)
```

3. In this context, the `variance` of the Normal distribution used to generate the noise in the linear relationship is equivalent to the *error variance*, $\sigma^2$, in the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

   Using this connection, generate four more vectors $y_3$, $y_4$, $y_5$, and $y_6$ by varying the amount of error (unexplained) variation in the linear relationship with $x$. Choose values of the error variation that span the range $0 - 10$. For example, let $y_3 \sim N(x, 0.1)$:

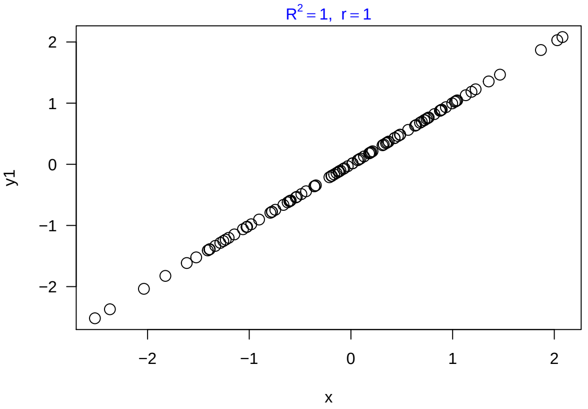
```
y3 <- 0 + 1 * x + rnorm(length(x), mean = 0, sd = sqrt(0.1))  #sigma^2 = 0.1
```

4. Generate scatterplots, $r$, and $R^2$ for each of the bivariate relationships: ($x$, $y_1$), ($x$, $y_2$), ($x$, $y_3$), ($x$, $y_4$), ($x$, $y_5$), ($x$, $y_6$). For example:

```
r.1 <- cor(x, y1)  # Calculates the correlation coefficient.
R2.1 <- r.1^2  # Relies on relationship in simple linear regression.
R2.1a <- summary(lm(y1 ~ x))$r.squared  # R^2 from linear regression model.
```

```
## Warning in summary.lm(lm(y1 ~ x)): essentially
## perfect fit: summary may be unreliable
```

```
plot(x, y1, las = 1, cex = 1.5)
mtext(bquote(paste(R^2 == .(R2.1a), ", ", ~r == .(r.1))),
    col = "blue")
```



5. Plot $r$ versus $\sigma^2$ and $R^2$ versus $\sigma^2$. Comment on the relationship.

# 2 Simple Linear Regression

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (`GPA`) at the end of the freshman year (`Y`) can be predicted from the `ACT` test score (`X`). Using the attached dataset `GPA`, answer the following questions.

```
##
```

```
## No. of observations = 120
##
##   Var. name obs. mean   median  s.d.   min.
## 1 GPA       120  3.07   3.08    0.64   0.5
## 2 ACT       120  24.73  25      4.47   14
##   max.
## 1 4
## 2 35
```

1. Write out the regression model that would explore the proposed relationship and state the model assumptions.
2. Using the methods discussed in the lesson, test each model assumption. If any assumptions appear to be violated, comment on the potential consequences of moving forward with the simple linear regression model.
3. Assume the simple linear regression model is appropriate and fit the model using `R` to obtain least squares estimates of $\beta_0$ and $\beta_1$. State the estimated regression function and interpret these estimates in the context of the problem.
4. Plot the estimated regression function over a scatterplot of the raw data. Does the estimated regression function appear to fit the data well?
5. What percentage of the variation in freshman GPA is explained by ACT test scores?
6. Obtain a 95% confidence interval for $\beta_1$ and interpret. Why might the director of admissions be interested in whether the confidence interval includes the value zero?
7. Test whether or not a linear association exists between student's `ACT score (X)` and `GPA` at the end of the freshman year `(Y)`. Use a level of significance of 0.01. State the hypotheses ($H_0$, $H_1$) and conclusion.
8. What is the $p$-value for the test? Interpret it (i.e., how exactly does it support the conclusion you reached?).