Statistics 206

Homework 3

Due : October 16, 2019, In Class

1. Tell true or false of the following statements and provide a brief explanation to your answer.

   (a) If all observations Y_i fall on one straight line, then the coefficient of determination $R^2 = 1$.

ANS. True. If all Y_i are on one straight line then $SSE = 0$ and $R^2 = 1$.

   (b) A large R^2 always means that the fitted regression line is a good fit of the data, while a small R^2 always means that the predictor and the response are not related.

ANS. False. The predictor and response may be nonlinearly related and then R^2 would be misleading for such cases.

   (c) The scatter plot is the most effective graph to show a nonlinear relationship between two variables.

ANS. False. Residual plots are better.

   (d) The regression sum of squares SSR tends to be large if the estimated regression slope is large in magnitude or the dispersion of the predictor values is large.

ANS. True. $SSR = \hat{\beta}_1^2 \sum (X_i - \overline{X})^2$.

2. Under the simple linear regression model:

   (a) Show that the regression sum of squares

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Proof.

$$\begin{aligned} SSR &= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2 \\ &= \sum_{i=1}^{n} (\hat{\beta}_1 (X_i - \bar{X}))^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 \end{aligned}$$

□

(b) Derive $E(SSR)$.

$$E(SSR) = E(\hat{\beta}_1^2)\sum_{i=1}^{n}(X_i - \overline{X})^2$$
$$= (\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} + \beta_1^2)\sum_{i=1}^{n}(X_i - \overline{X})^2$$
$$= \sigma^2 + \beta_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2$$

3. Under the Normal error model: Show that SSE and SSR are independent.

Proof. From 2(a) of this homework, $SSR = \hat{\beta_1}^2(X_i - \bar{X})^2$ which is a function of $\hat{\beta}_1$ and $SSE = e'e$ is a function of e. Since e is independent of $\hat{\beta}_1$ (proved in homework 2), SSE and SSR are independent. □

4. Perform ANOVA on the "crime rate and education" data (from homework 2).

(a) Calculate sum of squares: $SSTO$, SSE and SSR. What are their respective degrees of freedom?

$$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}Y_i^2 - n(\overline{Y})^2 = 4796548849 - 84 * (7111.2)^2 = 548738952$$

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2 = \hat{\beta}_1^2[\sum_{i=1}^{n}X_i^2 - n(\overline{X})^2] \\ &= (-174.88)^2 \times (522098 - 84 \times 78.6^2) = 96316922 \end{aligned}$$

$$SSE = SSTO - SSR = 452422030$$

The degrees of freedom for $SSTO$, SSE and SSR are 83, 82 and 1 respectively.

(b) Calculate the mean squares.

$$MSR = \frac{SSR}{1} = 96316922$$

$$MSE = \frac{SSE}{n-2} = \frac{452422030}{82} = 5517342$$

(c) Summarize results from parts (a) and (b) into an ANOVA table.

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = 96316922$	d.f.$(SSR) = 1$	$MSR = 96316922$	$F^* = MSR/MSE$
Error	$SSE = 452422030$	d.f.$(SSE) = 82$	$MSE = 5517342$	=17.46
Total	$SSTO = 548738952$	d.f.$(SSTO) = 83$		

(d) Assume Normal error model, use the F test to test whether or not there is a linear association between crime rate and percentage of high school graduates at significance level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

$$H_0 : \beta_1 = 0 \; vs. \; H_a : \beta_1 \neq 0$$

Test statistic $F^* = MSR/MSE = 17.46$. Under null hypothesis, $F^* \sim F_{1,82}$. Since $F(0.99; 1, 82) = 6.95 < F^* = 17.46$, reject H_0 and conclude that there is a significant linear association between crime rate and percentage of high school graduates.

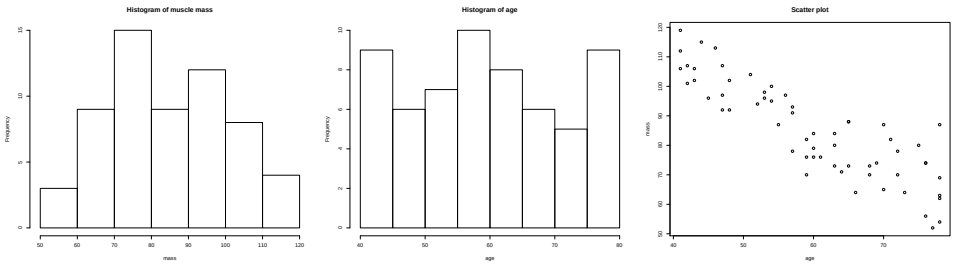
(e) Compare your calculation from part (d) with those from part (e) of Problem 6 of homework 2. What do you find?

$$F^* = 17.46 = (-4.178)^2 = (T^*)^2, \quad F(0.99; 1, 82) = 6.95 = 2.637^2 = (T(0.995, 82))^2$$

5. **A simple linear regression case study by R**. You must use R and the *lm* function and its associated functions to do this problem. Please also attach your R codes.

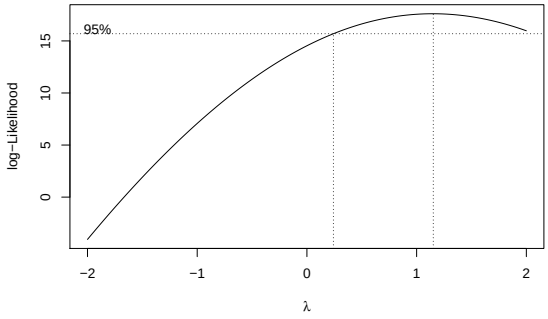
A person's muscle is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 *women from each of the four* 10*-year age groups, beginning with age* 40 *and ending with age* 79*. Two variables being measured are: age (X) and the amount of muscle mass (Y).* (The data is on smartsite under Resources/Homework/muscle.txt.)

(a) Read data into R. Draw histogram for muscle mass and age, respectively. Comment on their distributions. Draw the scatter plot of muscle mass versus age. Do you think their relation is linear? Does the data support the anticipation that the amount of muscle mass decreases with age?



The histogram of muscle mass is approximately bell-shaped; The histogram of age is pretty flat. Yes, the relation looks linear and it supports the anticipation that the amount of muscle mass decreases with age.

(b) Use the box-cox procedure to decide whether a transformation of the response variable is needed.



The suggested transformation by box-cox is a λ slightly bigger than 1 which means basically no transformation is needed.

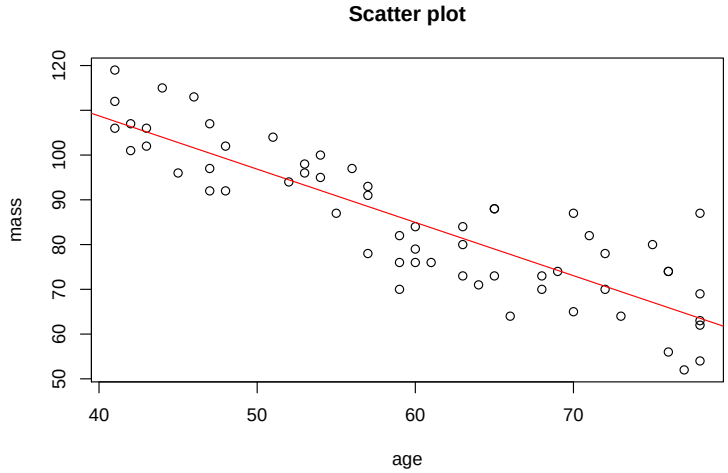
(c) Perform linear regression of the amount of muscle mass on age and obtain a summary. From the summary, obtain the estimated regression coefficients and their standard errors, the mean squared error (MSE) and its degrees of freedom.

$\hat{\beta}_0 = 156.3466$, $s\{\hat{\beta}_0\} = 5.5123$, $\hat{\beta}_1 = -1.19$, $s\{\hat{\beta}_1\} = 0.0902$, $MSE = 8.173^2 = 66.7979$, with 58 degrees of freedom.

(d) Write down the fitted regression line. Add the fitted regression line to the scatter plot. Does it appear to fit the data well?

Fitted regression line:

$$y = 156.3466 - 1.19x.$$

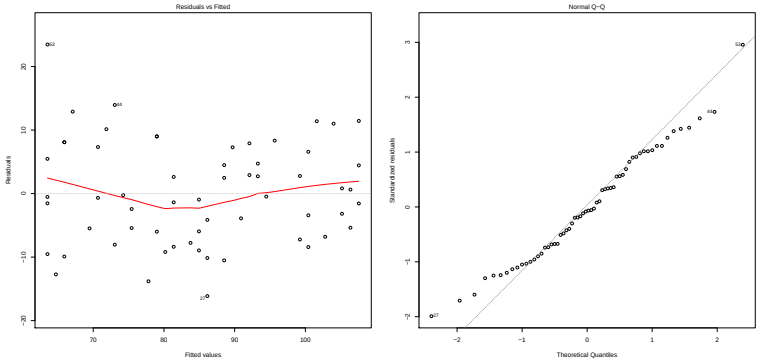


The regression line seems to fit the data well.

(e) Obtain the fitted values and residuals for the 6th and 16th cases in the data set.

$\widehat{Y}_6 = 107.5568$, $\widehat{Y}_{16} = 90.8968$, $e_6 = 11.4433$, $e_{16} = -3.8968$

(f) Draw the residuals vs. fitted values plot and the residuals Normal Q-Q plot. Write down the simple linear regression model with Normal errors and its assumptions. Comment on these assumptions based on the residual plots.



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Model assumptions: The random error terms ε_i are independently and identically distributed (i.i.d.) as $N(0, \sigma^2)$. From the residuals vs fitted values plot we can see that the expected value of the residuals is approximately zero and the variance is approximately constant. From the residuals Normal Q-Q plot we can see that the residuals are slightly light-tailed compared to Normal.

(g) Construct a 99% confidence interval for the regression intercept. Interpret your confidence interval.

$$\hat{\beta}_0 \pm t(0.995, 58) s\{\hat{\beta}_0\} = 156.3466 \pm 2.6633 \times 5.5123 = [141.6658, 171.0274]$$

(h) Conduct a test at level 0.01 to decide whether or not there is a negative linear association between the amount of muscle mass and age. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion. (Hint: which alternative you should use?)

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 < 0$$

$$T^* = \frac{\hat{\beta}_1}{s\{\hat{\beta}_1\}} = -13.19$$

Under null hypothesis, $T^* \sim t_{58}$. Since $T^* = -13.19 < -2.39 = t(0.01, 58)$, we reject the null hypothesis and conclude that there is significant negative linear association between the amount of muscle mass and age.

(i) Construct a 95% prediction interval for the muscle mass of a woman aged at 60. Interpret your prediction interval.

A 95% predication interval for the muscle mass of a woman aged at 60 is $[68.45, 101.44]$. We are 95% confident that the muscle mass of a woman aged at 60 is in between 68.45 and 101.44.

(j) Obtain the ANOVA table for this data. Test whether or not there is a linear association between the amount of muscle mass and age by an F test at level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = 11627.5$	d.f.$(SSR) = 1$	$MSR = 11627.5$	$F^* = MSR/MSE$
Error	$SSE = 3874.4$	d.f.$(SSE) = 58$	$MSE = 66.8$	$=174.06$
Total	$SSTO = 15501.9$	d.f.$(SSTO) = 59$		

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

Test statistic

$$F^* = 174.06$$

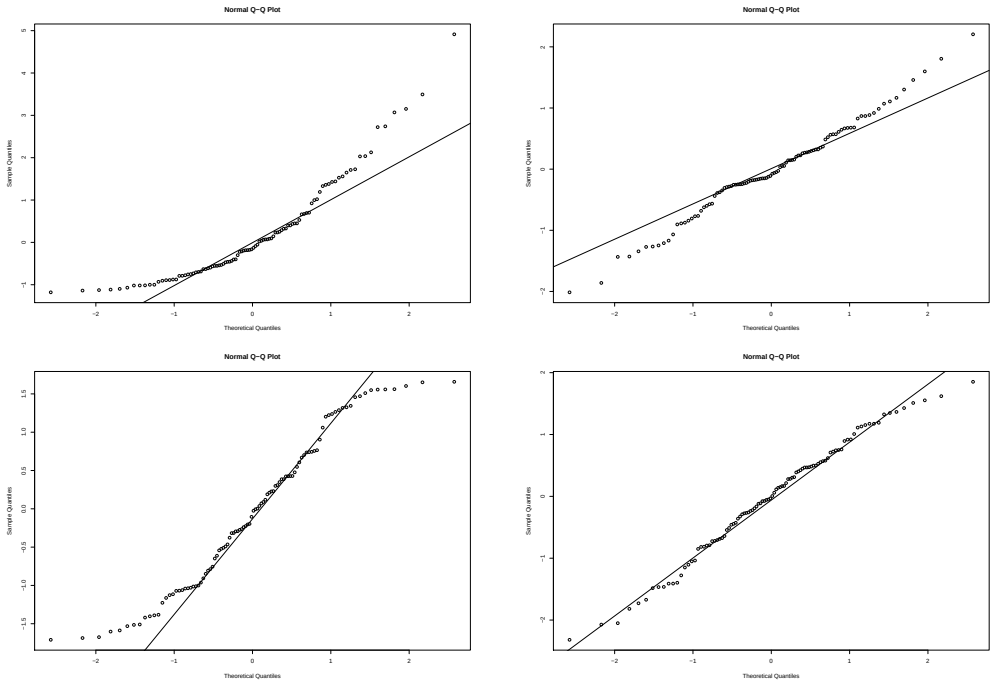
Under null hypothesis, $F^* \sim F_{1,58}$. Since $F^* = 174.06 > 7.09 = F(0.99; 1, 58)$, we reject H_0 and conclude that there is a significant linear association between the amount of muscle mass and age.

(k) What proportion of the total variation in muscle mass is "explained" by age? What is the correlation coefficient between muscle mass and age?

Since $R^2 = 0.7501$, about 75% of the total variation in muscle mass is "explained" by age. Since $\hat{\beta}_1 < 0$, the correlation coefficient r between the amount of muscle mass and age is $-\sqrt{R^2} = -0.8661$.

6. **Q-Q plots**. For each of the Q-Q plot in Figure 1, describe the distribution of the data (whether it is Normal or heavy tailed, etc.).

Figure 1: Q-Q plots



Looking at it in anticlockwise fashion,

* Top left: right skewed
* Bottom left: light tailed
* Bottom right: approximately normal
* Top right : heavy tailed

7. **Coefficient of determination** . Show that

$$R^2 = r^2, \quad r = \text{sign}\{\hat{\beta}_1\}\sqrt{R^2},$$

where R^2 is the coefficient of determination when regressing Y onto X and r is the sample correlation coefficient between X and Y.

$$R^2 = SSR/SSTO = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 / \sum (y_i - \bar{y})^2$$

$$= (\sum (x_i - \bar{x})(y_i - \bar{y}))^2 / \sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2 = r^2$$