BIS 505b – Problem Set 2 **Due 11:59 PM Eastern, 3/21/2021**

Instructions: Follow the homework instructions outlined in the syllabus. Round your answers to 2 decimal places. Perform all tests at the α = 0.05-level and follow the steps of hypothesis testing.

Assignment

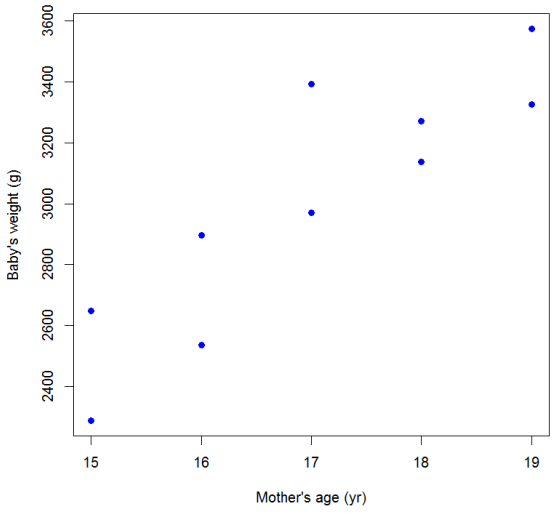
Question 1: Medical researchers have noted that adolescent females are much more likely to deliver low-birth-weight babies than are adult females. Because low-birth-weight babies have higher mortality rates, a number of studies have examined the relationship between birth weight and mother's age for babies born to young mothers. The following data on

x = maternal age (in years) and
y = birth weight of baby (in grams)

are consistent with data published by the National Center for Health Statistics.

	Observation										Summary Statistics
	1	2	3	4	5	6	7	8	9	10	
x	15	17	18	15	16	19	17	16	18	19	$\bar{x} = 17; s_x = 1.49$
y	2289	<<data values omitted from this table>>								3573	$\bar{y} = 3004.10; s_y = 413.55$

a. [5] A scatterplot of the data (right) shows a linear pattern and the spread of the y values appears to be similar across the range of x values. What two assumptions of linear regression does this information support?



1) Linearity
2) Homoscedasticity

b. [10] The sample correlation between x and y is $r = 0.884$. Find the equation of the estimated (i.e., fitted) simple linear regression line.

$$b = r(\frac{s_y}{s_x}) = 0.884 \times \frac{413.55}{1.49} = 245.35$$

$$a = \bar{y} - b\bar{x} = 3004.10 - 245.35 \times 17 = -1166.85$$

$$\hat{y} = a + bx = -1166.85 + 245.35x$$

The fitted simple linear regression line is $\hat{y} = -1166.85 + 245.35x$

c. [5] What is the expected birth weight of babies born to 18-year-old mothers?
$\hat{y} = a + bx^* = -1166.85 + 245.35 \times 18 = 3249.45$ (gram)

The expected birth weight of babies born to 18-year-old mothers is 3249.45 grams.

d. [5] What is the expected birth weight of babies born to 10-year-old mothers?

~~$\hat{y} = a + bx^* = -1166.85 + 245.35 \times 10 = 1286.65$ (gram)~~

We should not estimate this since it is outside the range of our data.

~~The expected birth weight of babies born to 10-year-old mothers is 1286.65 grams.~~

e. [10] Suppose you were given the following information from the ANOVA table for the simple linear regression model: SST = 1,539,182 and SSE = 337,212. Complete the ANOVA table and conduct the F test. Assume the ANOVA assumptions are met. What can you conclude as a result of your F test?

(1) State the null and alternative hypotheses

H_0: Model does not explain a significant amount of the variability in the response

vs. H_1: Model explains enough variability in the response to given evidence that the explanatory variable is linearly related to the response.

(2) Specify the significance level, α = 0.05

(2.5) ANOVA assumptions are met.

(3) Compute the test statistic

ANOVA table for the simple linear regression model

	SS	df	MS	F
Model	1,201,970	1	1,201,970	28.52
Error	337,212	8	42151.5	
Total	1,539,182	9		

$SSM = SST - SSE$ =1,539,182-337,212=1,201,970

$df_E = n - 2 = 10 - 2 = 8$

$$MSM = \frac{SSM}{df_M} = \frac{1{,}201{,}970}{1} = 1{,}201{,}970$$

$$MSE = \frac{SSE}{df_E} = \frac{337{,}212}{8} = 42151.5$$

$$F = \frac{MSM}{MSE} = \frac{1{,}201{,}970}{42151.5} = 28.52;\ F \sim F(1,8)$$

(4) Generate the decision rule

Given α = 0.05,

Reject H_0 if $F \geq F_{1-\alpha}(df_1, df_2) = F_{0.95}(1,8) = 5.32$ or if $p \leq 0.05$

(5) Draw a statistical conclusion, and state the conclusion in words in the context of the problem.

$F = 28.52 \geq 5.32 \rightarrow Reject\ H_0$

$or\ p = P(F \geq 28.52) = 0.0007 \leq 0.05 \rightarrow Reject\ H_0$

Conclusion: There is evidence to reject H_0 and conclude a significant amount of the total variability is explained by the model; there is a significant linear relationship between birth weight and mother's age ($p = 0.0007$).

f. [5] Calculate the R^2 of the model and interpret this value.
$R^2 = r^2 = 0.884^2 = 0.78$

or $R^2 = \frac{SSM}{SST} = \frac{1{,}201{,}970}{1{,}539{,}182} = 0.78$

Interpretation: The model explains 78% of the variability in the response variable (birth weight).

g. [5] Report the point estimate of $\sigma_{y|x}$.

$s_{y|x} = \sqrt{MSE} = \sqrt{42151.5} = 205.31$
The point estimate of $\sigma_{y|x}$ is 205.31.

h. [5] You are told that $s_b = 45.91$. Calculate the 95% confidence interval for the slope parameter.

$b \pm t_{n-2,1-\frac{\alpha}{2}} s_b = b \pm t_{8,0.975} s_b = b \pm \sqrt{F_{0.95}(1,8)} s_b = 245.35 \pm \sqrt{5.32}(45.91) = 245.35 \pm 105.89 = (139.46, 351.24)$

The 95% confidence interval for the slope parameter is $(139.46, 351.24)$.

i. [10] Is there a significant linear association between mother's age and baby's birth weight? [**Note**: There are several ways to answer this question; choose your preferred method. It's ok to call on results from a previous part of Question 1 to answer this question.]

There is a significant linear association between mother's age and baby's birth weight because the 95% CI for the slope $(139.46, 351.24)$ does not include 0.

Question 2 [5]**:** In words, what is $\sigma_{y|x}$?
$\sigma_{y|x}$ is the variability of the population of responses about the population regression line.

Question 3: In a study of the effects of split keyboard geometry on upper body postures, researchers examine if there is an association between surface angle of the keyboard and typing speed. Partial output from the simple linear regression model with x = surface angle (degrees) and y = typing speed (words per minute) is given below. The study had a very small sample size of $n = 5$.

- Least Squares Results:

	Estimate	Standard Error (SE)	t-value	p-value
Intercept	60.0286	0.2466		
Surface Angle	0.00357	0.03823		

- ANOVA Table:

	SS	df	MS	F	p-value
Model	0.0023				
Error	0.7857				
Total	0.7880				

a. [5] Based on the results above, report the equation of the fitted least squares regression line.

$\hat{y} = a + bx = 60.0286 + 0.00357x$

The fitted simple linear regression line is $\hat{y} = 60.0286 + 0.00357x$

b. [10] Assuming the basic assumptions of the simple linear regression model are reasonably met, carry out a hypothesis test to decide if there is a linear relationship between x and y.

(1) State the null and alternative hypotheses
$H_0: \beta = 0$ vs. $H_1: \beta \neq 0$

(2) Specify the significance level, α = 0.05

(3) Compute the test statistic
$t = \frac{b}{s_b} = \frac{0.00357}{0.03823} = 0.09 \sim t_3$

(4) Generate the decision rule
Given $\alpha = 0.05$ and a two-sided test is performed,

Reject H_0 if $|t| \geq t_{n-2,1-\frac{\alpha}{2}} = t_{3,0.975} = 0.76$

(5) Draw a statistical conclusion.
$|t| = 0.09 < t_{3,0.975}$ →Fail to reject H_0
$p = 2 \times P(T \geq 0.09) = 0.93 > 0.05$ →Fail to reject H_0

Conclusion: We fail to reject H_0 ,the data could not provide evidence that there is a significant linear association between surface angle of the keyboard and typing speed ($p = 0.93$).

c. [5] Is the value of R^2 consistent with the conclusion from part (b)? Explain.

$R^2 = \frac{SSM}{SST} = \frac{0.0023}{0.7880} = 0.0029$

The model could only explain 0.3% of the variability in the response variable (typing speed), which is consistent with the conclusion from part (b) that there is not a significant linear association between surface angle of the keyboard and typing speed.

Question 4: The following data resulted from a study looking at improving fermentation productivity with reverse osmosis where

x = fermentation time (days) for a blend of malt liquor and
y = glucose concentration (g/L)

x	1	2	3	4	5	6	7	8
y	74	54	52	51	52	53	58	71

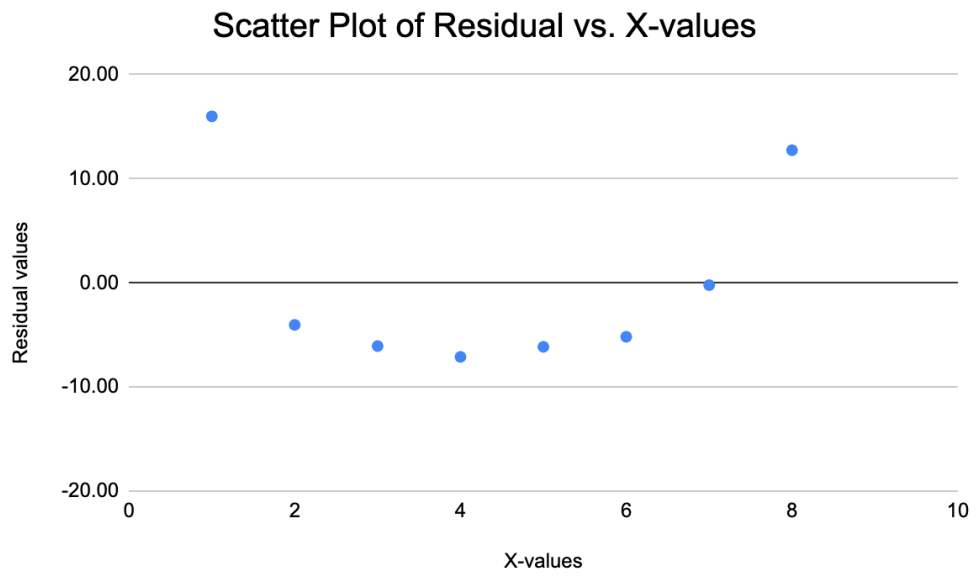
The fitted least squares regression line is: $\hat{y} = 57.964 + 0.0357x$.

a. [10] Using the estimated regression line above, complete the table below, computing the fitted values $\hat{y}_i$ and the residuals e_i. Construct a plot of the residuals vs. the x-values. [**Note**: You may construct this plot by hand or using software of your choice.]

x_i	y_i	$\hat{y}_i$	e_i
1	74	58.00	16.00

2	54	58.04	-4.04
3	52	58.07	-6.07
4	51	58.11	-7.11
5	52	58.14	-6.14
6	53	58.18	-5.18
7	58	58.21	-0.21
8	71	58.25	12.75

$\hat{y}_i = 57.964 + 0.0357x_i$
$e_i = y_i - \hat{y}_i$



b. [5] Based on the plot in part (a), do you think that the simple linear regression model is appropriate for describing the relationship between y and x? Explain.

I don't think that the simple linear regression model is appropriate for describing the relationship between y and x because of the violation of constant variance (the pattern is not random) and the violation of normality (there are two large outliers at left top corner and right top corner).