Leyli Garryyeva
Professor deCamp

Assignment 2

CSCI 688
September 20, 2019

Chapter 2 problems:
2.5, 2.8, 2.9, 2.14, 2.18, 2.24, 2.27, 2.28, 2.29, 2.47, 2.60

Problem 2.5

Refer to **Copier maintenance** Problem 1.20.

$X-$ number of copiers serviced
$Y-$ total number of minutes spent by service person $Y = 15.035X - 0.580$, where $b_0 = -0.580$ and $b_1 = 15.035$

a. Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90% confidence interval. Interpret your confidence interval.

Apply Confidence interval for $\boldsymbol{\beta_1}$***:***

$$b_1 \pm t_{1-\frac{\alpha}{2};n-2} s\{b_1\} \tag{1}$$

$b_1 = 15.035$
$s\{b_1\} = \sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}}$, where $MSE = 79.341$, and $\sum(X_i - \bar{X})^2 = 340.444$ from the Minitab
$s\{b_1\} = \sqrt{\frac{79.341}{340.444}} = \sqrt{0.233} = 0.48275$
$t_{1-\frac{\alpha}{2};n-2} = t_{0.95,43} = 1.6811$ (using the TI-84 command) $invT(0.95, 43)$

$$\begin{aligned} b_1 \pm t_{1-\frac{\alpha}{2};n-2} s\{b_1\} &= 15.035 \pm 1.6811 * 0.48275 = \\ &= \boxed{15.035 \pm 0.81154} \\ &= \boxed{(14.223, 15.847)} \end{aligned}$$

Then, with 90 percent confidence, we estimate that the mean service time increases somewhere between 14.233 and 15.847 minutes when the number of copiers serviced increases by one. Equivalently, with 90 percent confidence, we estimate that the mean service time will be within 0.8115 minutes from the point estimator $b_1 = 15.035$ *when the number of copiers serviced increases by one.*

b. Conduct a t test to determine whether or not there is a linear association between X and Y here; control the α risk at 0.10. State the alternative, decision rule, and conclusion. What is the $P-$value of your test?

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Decision rule:

$$\text{If } |t^*| \leq t_{1-\frac{\alpha}{2};n-2}, \text{ conclude } H_0$$
$$\text{If } |t^*| > t_{1-\frac{\alpha}{2};n-2}, \text{ conclude } H_a.$$

Then, for $t^* = \frac{b_1}{s\{b_1\}} = 15.035/0.48275 = 31.144$, and $t_{1-\frac{\alpha}{2};n-2} = 1.6811$ we conclude:
$\boxed{|t^*| > t_{1-\frac{\alpha}{2};n-2}, \text{ conclude } H_a, \text{ that } \beta_1 \neq 0 \text{ there exists a linear association between } X \text{ and } Y.}$

From the Minitab output:

$$P\left(t(43) > t^* = 31.144\right) \simeq 0.000$$

Then, since the two-sided $p-$*value* $= 0.000 * 2 \simeq 0+$ *is less than the specified level of control risk* $\alpha = 0.10$, *we conclude* H_a.

c. Are your results in parts (a) and (b) consistent? Explain.
Yes, results in parts (a) and (b) are consistent, because we are using the same $\alpha = 0.10$ *level for our calculations and the confidence interval for* β_1 *does not include* $\beta_1 = 0$.

d. The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at 0.05. State the alternatives, decision rule, and conclusion. What is the $P-$value of the test?

$$H_0 : \beta_1 \leq 14$$
$$H_a : \beta_1 > 14$$

Decision rule:

$$\text{If } |t^*| \leq t_{1-\alpha;n-2}, \text{ conclude } H_0$$
$$\text{If } |t^*| > t_{1-\alpha;n-2}, \text{ conclude } H_a.$$

Then, for $t^* = \frac{b_1 - \beta_1 0}{s\{b_1\}} = \frac{15.035-14}{0.48275} = 1.035/0.48275 = 2.144$, and $t_{1-\alpha;n-2} = t_{0.95,43} = 1.6811$ we conclude:
$\boxed{|t^*| > t_{1-\alpha;n-2}, \text{ conclude } H_a, \text{ that } \beta_1 > 14}$ and Tri-City does not satisfy the standard of the mean required time not not increasing by more than 14 minutes.

e. Does b_0 give any relevant information here about the "start-up" time on calls- i.e., about the time required before service work is begun on the copiers at a customer location?
No, because the Y *values are the total number of minutes spent by service person. This values don't tell us anything about the time when this services occurred and if they required any prep work.*

Problem 2.8

Refer to Figure 2.2 for the Toluca Company example. A consultant has advised that an increase of one unit in lot size should require an increase of 3.0 in the expected number of work hours for the given production item.

X_i- lot size
Y_i- work hours
$Y = 62.4 + 3.57X$
$b_{standard} = 3.0$

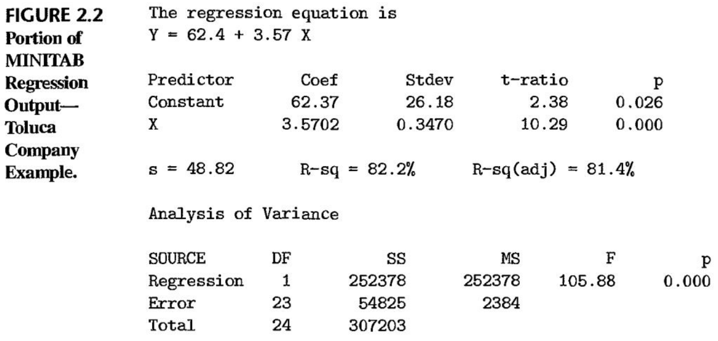


FIGURE 2.2 **Portion of MINITAB Regression Output— Toluca Company Example.**

```
The regression equation is
Y = 62.4 + 3.57 X

Predictor      Coef      Stdev     t-ratio        p
Constant      62.37      26.18        2.38    0.026
X            3.5702     0.3470       10.29    0.000

s = 48.82        R-sq = 82.2%      R-sq(adj) = 81.4%

Analysis of Variance

SOURCE       DF         SS         MS        F        p
Regression    1     252378     252378   105.88    0.000
Error        23      54825       2384
Total        24     307203
```

Figure 1

a. Conduct a test to decide whether or not the increase in the expected number of work hours in the Toluca Company equals this standard. Use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

$$H_0 : \beta_1 = 3$$
$$H_a : \beta_1 \neq 3$$

Decision rule:

$$\text{If } |t^*| \leq t_{1-\frac{\alpha}{2};n-2}, \text{ conclude } H_0$$
$$\text{If } |t^*| > t_{1-\frac{\alpha}{2};n-2}, \text{ conclude } H_a.$$

$t^* = \frac{b_1 - \beta_{standard}}{s\{b_1\}}$, where $s\{b_1\} = 0.3470$ and $b_1 = 3.57$. Then,

$$t^* = \frac{3.57 - 3}{0.3470} = 1.64265,$$

$$t_{1-\frac{\alpha}{2};n-2} = t_{0.975,23} = 2.06866.$$

$$p = 0.1139$$

Thus, we fail to reject H_0

b. Obtain the power of your test in part (a) if the consultant's standard actually is being exceeded by 0.5 hour. Assume $\sigma\{b_1\} = 0.35$.
$\delta = 0.5/0.35 = 1.43$, (number of standard deviations from H_0
Power $= P\{|t^*| > t_{1-\alpha/2,n-2}|\delta\} = P\{|t^*| > t_{0.975,23}|\delta\} = P\{|t^*| > 2.06866|\delta\} =$
$= 1 - P\{|t^*| \leq 2.06866|\delta\} = 1 - 0.7221 = 0.2779$
$\boxed{\text{Power} = 0.2779}$

c. Why is $F^* = 105.88$, given in the printout, not relevant for the test in part (a)?
$F^* = 105.88$ *given in the printout is not relevant for the test in part (a) because* F^* *is only useful in checking for linear relationship between* X *and* Y *as it helps with testing the hypothesis whether* β_1 *equals* 0 *or not.*

Problem 2.9

Refer to Figure 2.2. A student, noting that $s\{b_1\}$ is furnished in the printout, asks why $s\{\hat{Y}_h\}$ is not also given. Discuss.
$s\{b_1\} = 0.3470$
$s\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$, where $\hat{Y}$ is an unbiased estimator of $E\{Y_h\}$.
In order to get $s\{\hat{Y}_h\}$ we would need to know X_h which is the level of X for which we want to estimate the mean response $\hat{Y}_h$. This would be especially difficult to provide in the printout for continuous X variables as it would have to be long tables with all the corresponding values of $s\{\hat{Y}_h\}$ calculated for each possible value of X_h. Therefore, this might be why the value is not provided in the printout.

Problem 2.14

Refer to **Copier maintenance** Problem 1.20
$X-$ number of copiers serviced
$Y-$ total number of minutes spent by service person
$Y = 15.035X - 0.580$, where $b_0 = -0.580$ and $b_1 = 15.035$
$S_{xx} = 340.444$
$n = 45$
$\bar{X} = 5.111$
$MSE = 79.5$

a. Obtain a 90 percent confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your confidence interval.
$X_h = 6$, $\alpha = 0.1$

$$\begin{aligned}
&\hat{Y}_h \pm t_{1-\alpha/2;n-2} * s\{\hat{Y}_h\}\\
\hat{Y}_h &= b_0 + b_1X_h = -0.58 + 15.035 * 6 = 89.63\\
t_{1-\alpha/2;n-2} &= t_{0.95;43} = 1.6811\\
s\{\hat{Y}_h\} &= \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]} = \sqrt{79.5 * \left[\frac{1}{45} + \frac{(6-5.111)^2}{340.444}\right]} =\\
&= \sqrt{1.95122} = 1.3969\\
\hat{Y}_h \pm t_{1-\alpha/2;n-2} * s\{\hat{Y}_h\} &= 89.63 \pm 1.6811 * 1.3969 =\\
&= \boxed{89.63 \pm 2.348} =\\
&= \boxed{(87.282, 91.978)}
\end{aligned}$$

We conclude with confidence coefficient 0.90 that the mean service time on calls in which six copiers are serviced is somewhere between 87.282 and 91.978 minutes or within 2.348 minutes from the $E\{Y_h\}$.

b. Obtain a 90 percent prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be?

$$
\begin{aligned}
\hat{Y}_h &\pm t_{1-\alpha/2;n-2} * s\{pred\} \\
\hat{Y}_h &= 89.63 \\
t_{1-\alpha/2;n-2} &= t_{0.95;43} = 1.6811 \\
s\{pred\} &= \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]} = \sqrt{MSE + s^2\{\hat{Y}_h\}} = \\
&= \sqrt{79.5 + 1.95122} = \sqrt{81.45122} = 9.025 \\
\hat{Y}_h \pm t_{1-\alpha/2;n-2} * s\{\hat{Y}_h\} &= 89.63 \pm 1.6811 * 9.025 = \\
&= \boxed{89.63 \pm 15.172} = \\
&= \boxed{(74.458, 104.802)}
\end{aligned}
$$

The prediction interval is wider than the corresponding confidence interval in part (a), which is the expected outcome. Because, in part (a) we are estimating the mean of the distribution Y while in part (b) we are estimating an individual outcome drawn from the the distribution of Y. Therefore the latter will have higher uncertainty associated with it and the prediction interval will be bigger than the corresponding confidence interval. In addition, the $MSE = 79.5$ accounts for over 98 percent of the estimated prediction variable $s^2\{pred\}$ which means that there might be other "unaccounted" variables that affect Y.

c. Management wishes to estimate the expected service time *per copier* on calls in which six copiers are serviced. Obtain an appropriate 90 percent confidence interval by converting the interval obtained in part (a). Interpret the converted confidence interval.
$X_h = 6$
$E\{Y_h\}/X_h$ is the estimated expected service time per copier
$E\{Y_h\} \in (87.282, 91.978)$ (from part (a) solution).
Then,

$$
\begin{aligned}
E\{Y_h\}/X_h &\in \left(\frac{87.282}{6}, \frac{91.978}{6}\right) \\
E\{Y_h\}/X_h &\in \boxed{(14.5470, 15.3297)}
\end{aligned}
$$

or

$$
\begin{aligned}
E\{Y_h\}/X_h &\in \frac{89.63 \pm 2.348}{6} \\
E\{Y_h\}/X_h &\in \boxed{14.9383 \pm 0.3913}
\end{aligned}
$$

The expected service time per copier on calls in which six copiers are serviced is between 14.938 minutes and 15.3297 minutes.

d. Determine the boundary values of the 90 percent confidence band for the regression line when $X_h = 6$. Is your confidence band wider at this point than the confidence interval in

part (a)? Should it be?
$\hat{Y_h} = 89.63, s\{\hat{Y_h}\} = 1.3969, F(0.90, 2, 43) = 2.430407$ (Using TI-84 program to get F-inv)
$\hat{Y_h} \pm W * s\{\hat{Y_h}\}$

$$\begin{aligned} W^2 &= 2F(1-\alpha; 2, n-2) = 2F(0.90, 2, 43) = 2 * 2.4304 = 4.8608 \\ W &= 2.2047 \\ \hat{Y_h} \pm W * s\{\hat{Y_h}\} &= 89.63 \pm (2.2047 * 1.3969) = \\ &= \boxed{89.63 \pm 3.0797} \\ &= \boxed{(86.5503, 92.7098} \end{aligned}$$

Yes, as it is expected to be. In part (a) we are estimating the confidence interval for the mean response time while in part (d) we are calculating the confidence band for the entire regression line. Thus, it is wider than the confidence interval in part (a).

Exercise 2.18

For conducting statistical tests concerning the parameter β_1 why is the t test more versatile than the F test?
The t test is more versatile than the F test for conducting statistical tests concerning the parameter β_1 because the F test can be used only to test whether there is a linear association between X and Y. Thus, we test the hypothesis that $\beta_1 = 0$. On the other hand, t test allows us to test various hypothesis with regards to β_1. We can test whether β is $<, >$ to 0 or any other values specified by some standards (as given in previous problems).

Exercise 2.24

Refer to **Copier maintenance** Problem 1.20.

a. Set up the basic ANOVA table in the format of Table 2.2. Which elements of your table are additive? Also set up the ANOVA table in the format of Table 2.3. How do the two tables differ?

SS and DF are additive elements of the ANOVA table.

Table 1: ANOVA table for simple linear regression (format of Table 2.2)

Source of Variation	**SS**	**df**	**MS**	**E{MS}**
Regression	$SSR = 76960.4$	1	$MSR = 76960.4$	$\sigma^2 + 76960.423$
Error	$SSE = 3416.4$	$n-2=43$	$MSE = 79.45$	σ^2
Total	$SSTO = 80376.8$	$n-1=44$		

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2; SSE = \sum(Y_i - \hat{Y}_i)^2; SSTO = \sum(Y_i - -\bar{Y})^2; MSR = \frac{SSR}{1};$$
$$MSE = \frac{SSE}{n-2};$$
$$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2 = \sigma^2 + 76960.423$$

Table 2: Modified ANOVA table for simple linear regression (format of Table 2.3)

Source of Variation	**SS**	**df**	**MS**
Regression	$SSR = 76960.4$	1	$MSR = 76960.4$
Error	$SSE = 3416.4$	43	$MSE = 79.45$
Total	$SSTO = 80376.8$	44	
Correction for mean	$261, 749.488$	1	
Total, uncorrected	342,124	45	

$SS(\text{correction for mean}) = n\bar{Y}^2 = 45 * 76.267^2 = 261, 749.488$
$SSTOU = \sum Y_i^2 = 342, 124$

Table 1 contains expected MSR and MSE values which are absent in the Table 2. Moreover, Table 2 contains $SSTOU$ - the total uncorrected sum of squares- and SS(correction for mean) which are decomposed parts of the $SSTO$-the total sum of squares- from the Table 1.

b. Conduct an F test to determine whether or not there is a linear association between time spent and number of copiers serviced; use $\alpha = 0.10$. State the alternatives, decision rule, and conclusion.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Decision rule:

$$\text{If } |F^*| \leq F_{1-\alpha;1,n-2}, \text{ conclude } H_0$$
$$\text{If } |F^*| > F_{1-\alpha;1,n-2}, \text{ conclude } H_a.$$

$F_{1-\alpha;1,n-2} = F_{0.9,1,43} = 2.8259985 \simeq 2.826$
$F^* = MSR/MSE = 76960.4/79.45 = 968.6645689 \simeq 968.665$
$p \simeq 4 \cdot 10^{-31}$

Since $|F^*| > F_{1-\alpha;1,n-2}$, we conclude H_a that there is a linear association between time spent and number of copiers serviced at $\alpha = 0.10$.

c. By how much, relatively, is the total variation in number of minutes spent on a call reduced when the number of copiers serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

Definition: R^2 is the coefficient of determination which can be interpreted as the proportionate reduction of total variation associated with the use of the predictor variable X.
Based on this definition, our R^2 value is the measure that can help us determine the amount of reduction in variation in number of minutes spent on a call reduced when the number of copiers serviced is introduced into the analysis. From the Minitab printout, our $R^2 = 95.75\%$. Then, the total variation in the observations of Y - the number of minutes spent on call is reduced by 95.75% when the predictor variable X - the number of copiers serviced - is considered.

d. Calculate r and attach the appropriate sign.

$$r = \pm\sqrt{R^2} = +\sqrt{0.9575} = 0.978519 \simeq 0.9785$$

The attached sign is positive because the slope of the fitted regression is positive.

$$\boxed{r = 0.9785}$$

e. Which measure, r or R^2, has the more clear-cut operational interpretation?
R^2 has a more clear-cut operational interpretation explaining the percentage total variation accounted for by model.

Exercise 2.27

Refer to **Muscle mass** Problem 1.27.

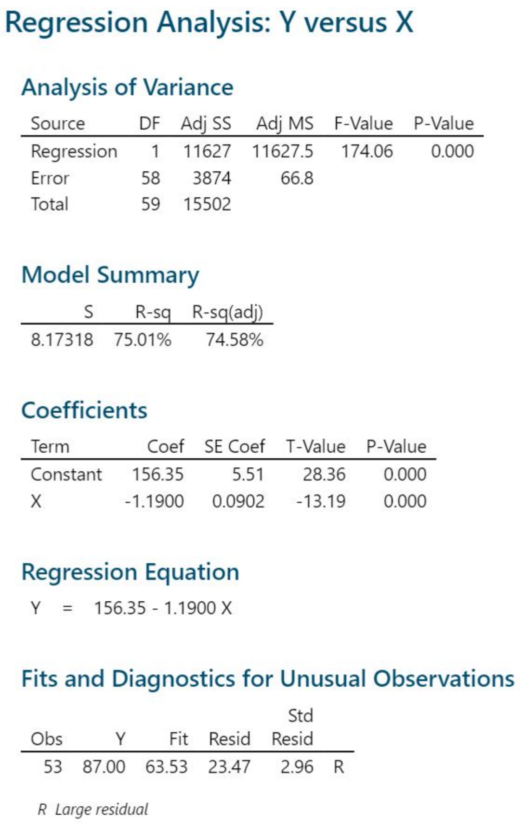
Sample: 15 women from each 10-year age group between age 40 and 79:
$n = 60$
$Age \in [40, 79]$
$X =$ age
$Y =$ muscle mass

a. Conduct a test to decide whether or not there is a negative linear association between amount of muscle mass and age. Control the risk of Type I error at 0.05. State the alternatives, decision rule, and conclusion. What is the $P-$value of the test?

$$H_0 : \beta_1 \geq 0$$
$$H_a : \beta_1 < 0$$

Decision rule: (note modified the decision rule because used the TI-84 to find the invT)

$$\text{If } t^* \geq t_{\alpha;n-2}, \text{ conclude } H_0$$
$$\text{If } t^* < t_{\alpha;n-2}, \text{ conclude } H_a.$$



```
Regression Analysis: Y versus X

Analysis of Variance

Source      DF  Adj SS   Adj MS  F-Value  P-Value
Regression   1   11627  11627.5   174.06    0.000
Error       58    3874     66.8
Total       59   15502

Model Summary

      S    R-sq  R-sq(adj)
8.17318  75.01%     74.58%

Coefficients

Term         Coef  SE Coef  T-Value  P-Value
Constant   156.35     5.51    28.36    0.000
X         -1.1900   0.0902   -13.19    0.000

Regression Equation

Y  =  156.35 - 1.1900 X

Fits and Diagnostics for Unusual Observations

                                Std
Obs      Y     Fit  Resid     Resid
 53  87.00   63.53  23.47      2.96  R

R  Large residual
```

Figure 2: Minitab output for the Problem 1.27

$s\{b_1\} = \sqrt{MSE/S_{xx}} = \sqrt{66.8008/8210.983} = 0.09019724$
$t^* = \frac{b_1}{s\{b_1\}} = \frac{-1.19}{0.09019724} = -13.1933$
$t_{\alpha;n-2} = t_{0.05,58} = -1.671552704 \simeq -1.67155$

Based on these calculations, $t^* < t_{\alpha;n-2}$ and we can conclude H_a that the β_1 is negative. Hence, there is a negative linear association between amount of muscle mass and age. Based on the Minitab output, the p−value is 0.000 ($2 \cdot 10^{-19}$).

b. The two-sided P-value for the test whether $\beta_0 = 0$ is $0+$. Can it now be concluded that b_0 provides relevant information on the amount of muscle mass at birth for a female child?
Our data only includes the ages between 40 and 79. Then, despite the fact that the two-sided P-value for the test whether $\beta_0 = 0$ is $0+$, we cannot conclude that b_0 provides relevant information on the amount of muscle mass at birth for a female child as $X = 0$ is not included in our model.

c. Estimate with a 95 percent confidence interval the difference in expected muscle mass for women whose ages differ by one year. Why is it not necessary to know the specific ages to make this estimate?

$$
\begin{aligned}
b_1 \pm t_{1-\alpha/2;n-2}s\{b_1\} & \\
&= -1.19 \pm (2.001717426 * 0.09019724) = \\
&= -1.19 \pm 0.180549 = \\
&\simeq \boxed{-1.9 \pm 0.18055} \\
&= \boxed{(-1.370549, -1.009451)}
\end{aligned}
$$

It is not necessary to know the specific ages to make this estimate because we are using S_{xx}, S_{xy} to calculate b_1, MSE, S_{xx} to get the $s\{b_1\}$, and α, n to get the $t-$value. None of these values require knowing specific age to make the estimate. Because even to get the S_{xx} and S_{xy} we only need the difference of the X_i's and $\bar{X}$ which would be still result in the same b_1 as long as the values are spaced similarly.

Exercise 2.28

Refer to **Muscle mass** Problem 1.27.

a. Obtain a 95 percent confidence interval for the mean muscle mass ($E\{Y_h\}$) for women of age 60. Interpret your confidence interval
$b_0 = 156.35; b_1 = -1.19; X_h = 60; MSE = 66.8008; S_{xx} = 8210.983; \bar{X} = 59.983$

$$
\begin{aligned}
\hat{Y_h} \pm t_{1-\alpha/2;n-2}s\{\hat{Y_h}\} & \\
\hat{Y_h} &= b_0 + b_1 Y_h = 156.35 - 1.19 * 60 = 84.95 \\
t_{0.975,58} &= 2.001717426 \\
s\{\hat{Y_h}\} &= \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}\right]} = \sqrt{66.8008 * \left[\frac{1}{58} + \frac{(60-59.983)^2}{8210.983}\right]} = \\
&= \sqrt{1.0733} = 1.073253835 \simeq 1.0733 \\
\hat{Y_h} \pm t_{1-\alpha/2;n-2}s\{\hat{Y_h}\} &= 84.95 \pm 2.0017 * 1.0733 = \\
&= \boxed{84.95 \pm 2.2148} \\
&= \boxed{(82.8017, 87.0983)}
\end{aligned} \tag{2}
$$

We conclude with confidence coefficient 0.95 that the mean muscle mass for women of age 60 is somewhere between 82.8017 and 87.0983 units or within 2.2148 units from the $E\{Y_h\}$.

b. Obtain a 95 percent prediction interval for the muscle mass of a woman whose age is 60. Is the prediction interval relatively precise?

$$\begin{aligned}
\hat{Y_h} &\pm t_{1-\alpha/2;n-2} * s\{pred\} \\
\hat{Y_h} &= 84.95 \\
t_{1-\alpha/2;n-2} &= t_{0.975;43} = 2.001717426 \\
s\{pred\} &= \sqrt{MSE\left[1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]} = \sqrt{MSE+s^2\{\hat{Y_h}\}} = \\
&= \sqrt{66.8008+1.15187} = \sqrt{67.95267} = 8.2433 \\
\hat{Y_h} \pm t_{1-\alpha/2;n-2} * s\{\hat{Y_h}\} &= 84.95 \pm 2.0017 * 8.2433 = \\
&= \boxed{84.95 \pm 16.5007} = \\
&= \boxed{(68.4493, 101.4507)}
\end{aligned}$$

The prediction interval is pretty wide with lower and upper bounds being within 20 percent (16.5007) of the point estimator $\hat{Y_h} = 84.95$. *Thus, it does not seem to be quite precise.*

c. Determine the boundary values of the 95 percent confidence band for regression line when $X_h = 60$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?
$\hat{Y_h} = 84.95, s\{\hat{Y_h}\} = 1.0733, F(0.95, 2, 58) = 0.051339$ (Using TI-84 program to get F-inv)
$\hat{Y_h} \pm W * s\{\hat{Y_h}\}$

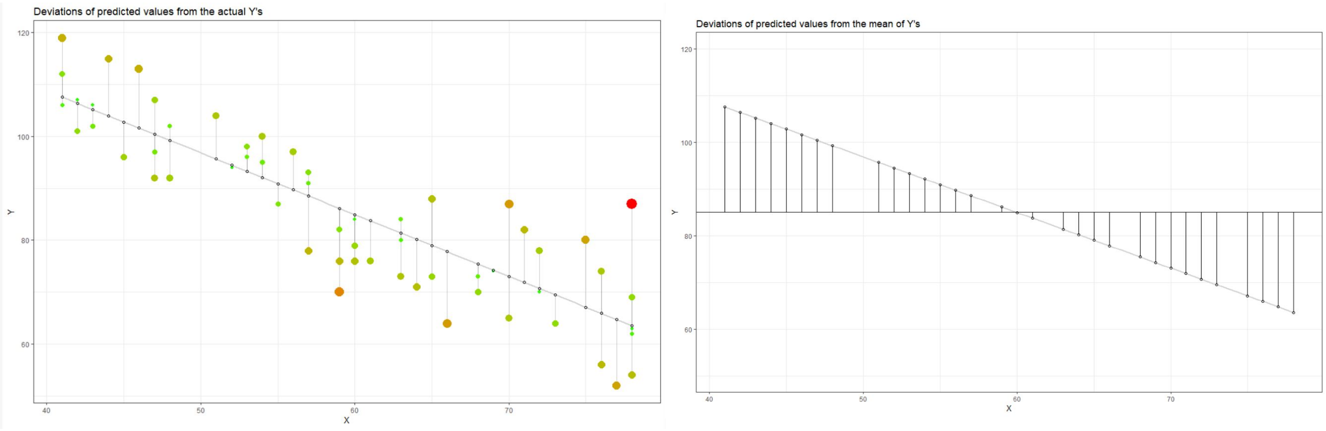
$$\begin{aligned}
W^2 &= 2F(1-\alpha; 2, n-2) = 2F(0.95, 2, 58) = 2 * 1.256 = 2.512 \\
W &= 2.512 \\
\hat{Y_h} \pm W * s\{\hat{Y_h}\} &= 84.95 \pm (2.512 * 1.0733) = \\
&= \boxed{84.95 \pm 2.651} \\
&= \boxed{(82.30, 87.60)}
\end{aligned}$$

No, as it is expected to be. In part (a) we are estimating the confidence interval for the mean response time while in part (d) we are calculating the confidence band for the entire regression line. From the textbook (p.62) we know that the boundary points of the confidence band are wider apart the further the X_h *is from the mean* $\bar{X}$ *os the* X *observations. In our case* $X_h = 60$ *and* $\bar{X} = 59.983$ *and, therefore, boundary points of the confidence bands are expected to be less wide.*

Exercise 2.29

Refer to **Muscle mass** Problem 1.27.

a. Plot the deviations $Y_i - \hat{Y_i}$ against X_i on one graph, Plot the deviations $\hat{Y_i} - \bar{Y}$ against X_i on another graph, using the same scales as in the first graph. From your two graphs, does



(a) The deviations $Y_i - \hat{Y}_i$ against X_i (SSE)

(b) The deviations $\hat{Y}_i - \bar{Y}$ against X_i (SSR)

Figure 3: Graphs of the deviation of the predicted Y's from the actual values of Y's and their mean

SSE or SSR appear to be the larger component of $SSTO$? What does this imply about the magnitude of R^2?

Based on the plots, SSR seems to be a larger component of the SSTO. Since, $R^2 = SSR/SSTO$*, this implies that the higher SSR will result in higher* R^2*.*

b. Set up the ANOVA table.

Table 3: ANOVA table for simple linear regression

Source of Variation	**SS**	**df**	**MS**
Regression	$SSR = 11627.486$	1	$MSR = 11627.486$
Error	$SSE = 3874.448$	58	$MSE = 66.8008$
Total	$SSTO = 15501.933$	59	
Correction for mean	433160.4065	1	
Total, uncorrected	448662	60	

SS(correction for mean) $= n\bar{Y}^2 = 60 * 84.9667^2 = 433160.4065$

c. Test whether or not $\beta_1 = 0$ using an F test with $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Decision rule:

$$\text{If } |F^*| \leq F_{1-\alpha;1,n-2}, \text{ conclude } H_0$$
$$\text{If } |F^*| > F_{1-\alpha;1,n-2}, \text{ conclude } H_a.$$

$F_{1-\alpha;1,n-2} = F_{0.95,1,58} = 4.006872653 \simeq 4.0069$
$F^* = MSR/MSE = 11627.486/66.8008 = 174.0620771 \simeq 174.0621$

Since $|F^*| > F_{1-\alpha;1,n-2}$*, we conclude* H_a *that there is a linear association between muscle mass and age* $\alpha = 0.05$*.*

d. What proportion of the total variation in muscle mass remains "unexplained" when age is introduced into the analysis? Is this proportion relatively small or large?
$SSE/SSTO = 3874.448/15501.933 = 0.2499331535 \simeq 0.2499 = 1 - SSR/SSTO$

Around 24.99% *of the total variation is in muscle mass remains unaccounted for when age is introduced into the analysis. Depending on the context, this might be considered a relatively small proportion if other models have higher error rates. However, without knowing anything else about the problem we might consider testing other models to decrease this proportion to less than 10 percent or even less.*

e. Obtain R^2 and r.
$R^2 = SSR/SSTO = 0.7500668465 \simeq 0.750$ or 75%
$r = \pm\sqrt{R^2} = -\sqrt{0.7500668465} = -0.8660639968 \simeq -0.866$
$\boxed{R^2 = 75\%}$
$\boxed{r = -0.866}$

Exercise 2.47

Refer to **Muscle mass** Problem 1.27. Assume that the normal bivariate model (2.74) is appropriate.

a. Compute the Pearson product-moment correlation coefficient r_{12}.

$$r_{12} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{S_{xx}S_{yy}}} = \frac{-9771.0333}{11282.11} = -0.86606398$$

$\boxed{r_{12} = -0.86606}$

b. Test whether muscle mass and age are statistically independent in the population; use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

$$H_0 : \rho_{12} = 0$$
$$H_a : \rho_{12} \neq 0$$

Decision rule:

$$\text{If } |t^*| \leq t_{1-\alpha/2;n-2}, \text{ conclude } H_0$$
$$\text{If } |t^*| > t_{1-\alpha/2;n-2}, \text{ conclude } H_a.$$

$t_{1-\alpha/2;n-2} = t_{0.975,58} = 2.001717426 \simeq 2.0017$
$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} = \frac{-0.86606*\sqrt{58}}{\sqrt{1-(0.86606)^2}} = -13.189$

Since $|t^*| > t_{1-\alpha/2;n-2}$ we conclude H_a that the muscle mass and age are not statistically independent.

c. The bivariate normal model (2.74) assumption is possibly inappropriate here. Compute the Spearman rank correlation coefficient, r_s.

Using Excel formulas we find the Spearman rank correlation coefficient $r_s = -0.8657217438$
$= RANK.AVG(B2, \$B\$2 : \$B\$61, 0)$ - example code for calculating the rank of a column;
$= CORREL(C2 : C61, D2 : D61)$ - example code for getting the correlation using the ranks.

$\boxed{r_s = -0.8657217438}$

d. Repeat part (b), this time basing the test of independence on the Spearman rank correlation computed in part (c) and test statistic (2.101). Use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

$$H_0 : X \text{ and } Y \text{are not associated}$$
$$H_a : X \text{ and } Y \text{are not associated}$$

Decision rule:

$$\text{If } |t^*| \geq t_{1-\alpha/2;n-2}, \text{ conclude } H_0$$
$$\text{If } |t^*| < t_{1-\alpha/2;n-2}, \text{ conclude } H_a.$$

$t_{1-\alpha/2;n-2} = t_{0.975,58} = 2.001717426 \simeq 2.0017$
$t^* = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{-0.8657217438*\sqrt{58}}{\sqrt{1-(-0.8657217438)^2}} = -13.1724$

Since $|t^*| > t_{1-\alpha/2;n-2}$ we conclude H_0 that the muscle mass and age are statistically independent.

e. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in parts (c) and (d)?
In part (a) we obtained Pearson product-moment correlation coefficient $r_{12} = -0.86606$ which is close to the Spearman rank correlation $r_s = -0.86572$ calculated in part (d). Both values indicate strong negative correlation between the muscle mass and age.
Interestingly, our test for statistical independence in part (b) indicate that the muscle mass and age are not statistically independent. However, when we drop the assumption that X and Y are drawn from the bivariate normal distribution, our test indicates that the muscle mass and age are not associated.

Exercise 2.60

Show that test statistics (2.17) and (2.87) are equivalent.
Formulas used and prep-work:

$$
\begin{aligned}
(2.17)\ t^* &= \frac{b_1}{s\{b_1\}} \\
(2.87)\ t^* &= \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} \\
b_1 &= \frac{S_{xy}}{S_{xx}} = \frac{r_{12}\sqrt{S_{yy}}}{S_{xx}} \\
r &= \frac{S_{xy}}{\sqrt{S_{xy}S_{xx}}} \ -> r^2 = \frac{S_{xy}^2}{S_{xy}S_{xx}} \\
SSR &= \frac{S_{xy}^2}{S_{xx}} = b_1^2 S_{xx} \\
SST &= S_{yy} \\
MSE &= \frac{SSE}{(n-2)} = \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) \div (n-2) = \frac{S_{yy}S_{xx} - S_{xy}^2}{(n-2)} \\
s\{b_1\} &= \sqrt{\frac{MSE}{S_{xx}}} = \sqrt{\frac{S_{yy}S_{xx} - S_{xy}^2}{(n-2)S_{xx}}}
\end{aligned}
$$

Proof:

$$
\begin{aligned}
t^* &= \frac{b_1}{s\{b_1\}} = r\frac{\sqrt{S_{yy}}}{S_{xx}} \div \sqrt{\frac{S_{yy}S_{xx} - S_{xy}^2}{(n-2)S_{xx}}} = r\frac{\sqrt{S_{yy}}}{S_{xx}} * \sqrt{\frac{(n-2)S_{xx}}{S_{yy}S_{xx} - S_{xy}^2}} = \\
&= r\sqrt{\frac{S_{yy}}{S_{xx}}} * \frac{\sqrt{n-2}}{\sqrt{S_{yy}S_{xx} - S_{xy}^2}} = r\sqrt{\frac{S_{yy}}{S_{xx}}} * \frac{\sqrt{n-2}}{\sqrt{S_{yy}S_{xx} - S_{xy}^2}} * \frac{\frac{1}{\sqrt{S_{yy}S_{xx}}}}{\frac{1}{\sqrt{S_{yy}S_{xx}}}} = \\
&= r\sqrt{n-2} * \frac{1}{\sqrt{(S_{yy}S_{xx} - S_{xy}^2)/\sqrt{S_{yy}S_{xx}}}} = \frac{r\sqrt{n-2}}{\sqrt{1 - \frac{S_{xy}^2}{S_{yy}S_{xx}}}} = \\
&= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}
\end{aligned}
$$

$\therefore$ (2.17) = (2.87)

A R Code for creating the graphs for the Exercise 2.29:

```
library(ggplot2)

muscle_data <- read.csv("~/Problem1.27.csv", header = TRUE)
muscle_data$y_i_y_bar <- muscle_data$Predicted_Y - mean(muscle_data$Y)

muscle.lm = lm(Y ~ X, data=muscle_data)
muscle.res = resid(muscle.lm)
Y_mean <- mean(muscle_data$Y)


# plot the deviations of predicted values from the actual Y's
ggplot(muscle_data, aes(x = X, y = Y)) +
    # regression line
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
    # draw line from point to line
  geom_segment(aes(xend = X, yend = Predicted_Y), alpha = .2) +
    # size of the points
  geom_point(aes(color = abs(Residuals), size = abs(Residuals))) +
    # colour of the points mapped to residual size - green smaller, red larger
  scale_color_continuous(low = "green", high = "red") +
    # Size legend removed
  guides(color = FALSE, size = FALSE) +
  geom_point(aes(y = Predicted_Y), shape = 1) +
  ggtitle("Deviations of predicted values from the actual Y's") +
  theme_bw()

# plot the deviations of predicted values from the mean of Y's
ggplot(muscle_data, aes(x = X, y = Y)) +
    # regression line
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
    # Size legend removed
  guides(color = FALSE, size = FALSE) +
  geom_point(aes(y = Predicted_Y), shape = 1) +
  geom_segment(aes(X, Predicted_Y, xend = X, yend = Y_mean)) +
  ylim(50, 120) +
  geom_hline(yintercept=Y_mean) +
  ggtitle("Deviations of predicted values from the mean of Y's") +
  theme_bw()
```