Assignment 9 Solution

1) **(1 point)** We are interested in predicting the price of a diamond (dollars), given its size (carats). The partial output from a simple regression of price vs. size is shown below. Fill in the values of the empty
(*Hint for the coefficient of size – look at the 95% CI*).

```
. regress price size
```

Source	SS	df	MS	Number of obs	=	49
				F(1, 47)	=	2134.72
Model	2131497.63	1	2131497.63	Prob > F	=	0.0000
Residual	46929.1471	47	998.492491	R-squared	=	0.9785
				Adj R-squared	=	0.9780
Total	2178426.78	48	45383.8912	Root MSE	=	31.599

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
size	3715.022	80.40654	46.20	0.000	3553.265	3876.779
_cons	-258.0504	16.93998	-15.23	0.000	-292.1292	-223.9715

2) **(2 points)** The following data shows one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days. We are interested in predicting heart rate (Y) based on swim time (X).

Mean(X) = 35.115
Mean(Y) = 141.4
Sample Standard Deviation(X) = 0.7879
Sample Standard Deviation(Y) = 9.5242
Correlation (X, Y) = -0.1236

a. Write out the population regression model and sample regression model.
Population Model: HRate = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}$STime + $\boldsymbol{\epsilon}$
Sample Model: (Estimated) HRate = $\hat{\alpha}$ + $\hat{\beta}$STime

b. Calculate the Slope of the Least Squares Regression Line
$\hat{\beta} = r_{XY}*(s_Y/s_X) = (-0.1236)*(9.5242/0.7879) = -1.494$

c. Calculate the Intercept of the Least Squares Regression Line
$\hat{\alpha} = \overline{Y} - \hat{\beta} * \overline{X} = 141.4 - (-1.494)*35.115 = 193.86$

d. Give a point estimate of the Heart rate (beats per minute) when swim time is 35 minutes.

Estimated Heart rate for STime=35 = 193.86+(-1.494)*35 = 141.57

e. Given that the Standard Error of the estimated slope is 4.2411, conduct the hypothesis test for the effect of swim time on heart rate. For the hypothesis test,

i. state the null and alternative hypothesis
H_0: $\boldsymbol{\beta} = 0$; There is no relationship between mean HRate and STime
H_a: $\boldsymbol{\beta} \neq 0$; There is a significant linear relationship between mean HRate and STime

ii. what is the underlying sampling distribution of the slope estimator?
The sampling distribution of $\widehat{\beta}$ is Normal ($\boldsymbol{\beta}$, 4.2411)

iii. what is the value of the test-statistic?
The test statistic is: $t_8 = -1.494/4.2411 = -0.3522$

iv. what is the p-value?
From t-table with 8 degrees of freedom, the p-value is between 0.50 and 0.1 (actual value 0.733 from R)

v. what is the conclusion of your test (in words) at a significance level of 0.05?
We do not reject the null hypothesis and conclude that *STime is not a significant (linear) predictor of HRate* or *There is no significant linear relationship between mean STime and HRate at a significance level of 0.05.*

3) The following table shows data on average *per capita wine consumption* (liters) and deaths from heart disease (in 1000's) in a random sample of 10 countries. The data is provided in a separate *Wine.csv file*.

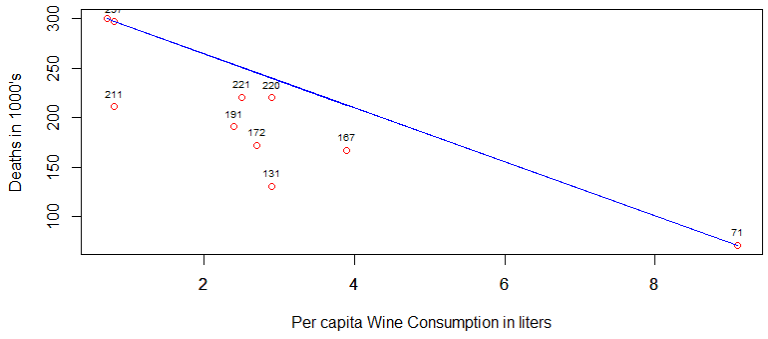
Using R:

a. **(2 points)** Perform a simple linear regression of deaths (Y) vs wine consumption (X).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  266.631     20.437  13.046 1.13e-06 ***
consumption  -23.878      5.544  -4.307  0.00259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.56 on 8 degrees of freedom
Multiple R-squared:  0.6987,    Adjusted R-squared:  0.6611
F-statistic: 18.55 on 1 and 8 DF,  p-value: 0.00259
```

b. Plot the original data along with the predicted regression line on the same graph.



c. Write out the sample regression equation based on the estimated slope and intercept.

Sample Model: (**Mean** Estimated) deaths = 266.63 - 23.878 consumption

d. Interpret the effect of wine consumption on deaths due to heart disease.

For each liter of per capita wine consumption, mean number of deaths from heart disease goes **down** by 23,878.

e. Conduct the hypothesis test of whether swine consumption is a good predictor of deaths from heart disease at $\alpha = 0.01$.

H_0: $\boldsymbol{\beta} = 0$; There is no relationship between mean number of deaths from heart disease and per capita wine consumption (in liters)

H_a: $\boldsymbol{\beta} \neq 0$; There is a significant linear relationship between mean number of deaths from heart disease and per capita wine consumption (in liters)

Since the p-vale for the test of the slope $0.003 < \alpha = 0.01$we conclude that: *"There is a significant linear relationship between mean number of deaths from heart disease and per capita wine consumption (in liters)"*

f. Perform a variability analysis using R. Show what the Total SS, Model SS, and Residual SS.

```
> # Variability Analysis - Using the anova() function
> #
> anova_mod <- anova(reg_model)
> print(round(anova_mod),12)
Analysis of Variance Table

Response: deaths
            Df Sum Sq Mean Sq F value     Pr(>F)
consumption  1  30528   30528      19 < 2.22e-16 ***
Residuals    8  13163    1645
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> print((anova_mod))
Analysis of Variance Table

Response: deaths
            Df Sum Sq Mean Sq F value  Pr(>F)
consumption  1  30528 30527.7  18.553 0.00259 **
Residuals    8  13163  1645.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #
```

Model SS = 30258, Residual SS = 13163 and Total SS = 43691.

g. Interpret the results of the F-test of Model Fit. What are the numerator and denominator degrees of freedom?

The F-test is a test of overall model fit and the F-statistic is calculated as: $(MS_{Model}/MS_{Residuals})$ = 30527.7/1645.4 = 18.533 with numerator df = 1 and Denominator df = 8.

In simple regression, the null and alternate hypothesis are the same as that for the slope.

H_0: $\boldsymbol{\beta} = 0$; There is <u>no relationship</u> between mean number of deaths from heart disease and per capita wine consumption (in liters)

H_a: $\boldsymbol{\beta} \neq 0$; There is a significant linear relationship between mean number of deaths from heart disease and per capita wine consumption (in liters)

The p-value for the F-test is that same as that for the slope = 0.00259 and our conclusion is that overall the model says that there "*is a significant linear relationship between mean number of deaths from heart disease and per capita wine consumption (in liters)"*

h. What percentage of the variability in heart disease deaths is **not explained** by wine consumption?

$(1-R^2)$ is the percentage of variability in heart disease deaths is **not explained** by wine consumption = 0.3013 or 30.13%

i. What is the expected death rate for wine consumption of 3 liters?

The expected death rate from heart disease (in 1000's) when the per capita wine consumption is 3 liters = 266.6308 + 3*(-23.878) = 195 i.e., 195000 deaths from heart disease.

```
> conump_3 <- data.frame(consumption=3)
> deaths_consump_3 <- predict(reg_model, conump_3)
> print(paste("Predicted Deaths in 1000s for Per capita Wine Consumption of 3 liters  = ", round(deaths_consump_3,2)))
[1] "Predicted Deaths in 1000s for Per capita Wine Consumption of 3 liters  =  195"
```