Exam 1

Problems

Exercise 1 (1 point).

In regression analysis, the model in the form $y = \beta_0 + \beta_1 x + \varepsilon$ is called

A. estimated simple linear regression equation
B. simple linear regression model
C. correlation equation

Solution: B

Exercise 2 (1 point).

A simple regression model has

A. only one independent variable
B. more than one dependent variable
C. more than one independent variable
D. at least two dependent variables

Solution: A

Exercise 3 (1 point).

In regression analysis, the variable that is being predicted is called the *response variable*. It is also called the

A. dependent variable
B. independent variable
C. intervening variable
D. usually x

Solution: A

Exercise 4 (1 point).

In simple linear regression (SLR), which of the following is **NOT** a required assumption about the error term ε?

A. The expected value of the error term is one.
B. The variance of the error term is constant.
C. The values of the error term are independent.
D. The error term is normally distributed.

Solution: A

Exercise 5 (1 point).

A regression analysis between **sales** (measured in \$1000) and **price** (measured in dollars) resulted in the equation $\hat{y} = 60 - 8x$. This equation implies that an

A. increase of \$1 in sales is associated with an increase of \$8 in price
B. increase of \$1 in sales is associated with a decrease of \$52000 in price
C. increase of \$1 in price is associated with an increase of \$52 in sales
D. increase of \$1 in price is associated with a decrease of \$8000 in sales

Solution: D

Exercise 6 (1 point).

The confidence interval estimate for an **average value of** y (in a linear regression model) will be _________ compared to the prediction interval estimate for a particular value.

A. the same
B. wider
C. narrower
D. sometimes narrower, and sometimes wider

Solution: C

Exercise 7 (1 point).

In simple linear regression analysis, which of the following is **NOT true**?

A. MSE equals SSE divided by the degrees of freedom for error.
B. The F test and the t test have the same p-value, thus yield the same conclusion.
C. The F test and the t test do not have the same p-value, thus do not yield the same conclusion.
D. The relationship between x and y is represented by a straight line.

Solution: C

Exercise 8 (10 points).

At temperatures approaching absolute zero (-273 C), helium exhibits traits that seem to defy many laws of Newtonian physics. An experiment has been conducted with helium in solid form at various temperatures near absolute zero. The solid helium is placed in a dilution refrigerator along with a solid impure substance, and the fraction (in weight) of the impurity passing through the solid helium is recorded. (This phenomenon of solids passing directly through solids is known as quantum tunneling.) The data with 10 observations are given in the table.

Please fill in the blanks in the R output below.

```
Analysis of Variance Table

       Df      SS      MS      F          P
Source  1 0.83089 0.83089 46.728 0.00013289
Error 1.___ 0.14225 2.____
Total   9 0.97315


Call:
lm(formula = PROPPASS ~ TEMP)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30732 -0.02940  0.03045  0.05943  0.17014
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.490347   2.073772  -6.505 0.000187 ***
TEMP         -0.052829   0.007728  3.____ 0.000133 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.____ on 8 degrees of freedom
Multiple R-squared:  0.8538,    Adjusted R-squared:  0.8356
F-statistic: 5.___ on 6.___ and 7.___ DF,  p-value: 8.___
```

Solution:

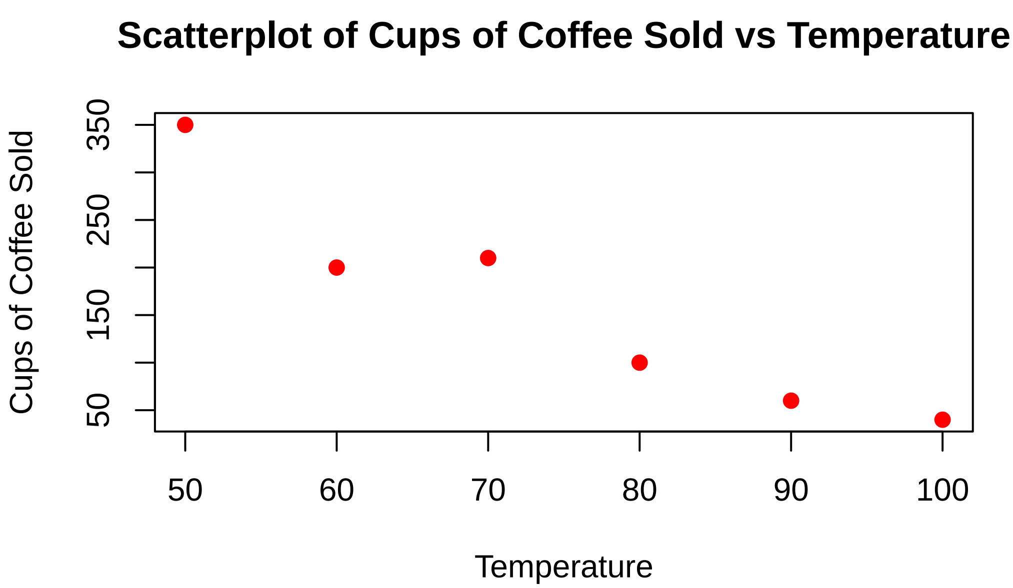
1. 8
2. $0.14225/8 \approx 0.0178$
3. $-0.052829/0.007728 \approx -6.836$
4. $\sqrt{0.0178} \approx 0.133$
5. 46.728
6. 1
7. 8
8. 0.000133

Exercise 9 (3 points)**.**

Jason believes that the sales of coffee (cups sold) at his coffee shop depend upon the temperature. He has taken a sample of 6 days. Some of the values were calculated based on the data collected.

Use the following information for any calculations.

$$\bar{x} = 75, \quad \bar{y} = 160$$
$$SS_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = -10400, \quad SS_{xx} = \sum(x_i - \bar{x})^2 = 1750$$
$$SS_{yy} = \sum(y_i - \bar{y})^2 = 68200, \quad SSE = \sum(y_i - \hat{y}_i)^2 = 6394$$



a. What does the scatter diagram (above) indicate about the relationship between cups of coffee sold and temperature?
b. What is the coefficient of determination r^2? (Display in percentage form, round to 1 decimal place.)
c. Interpret the coefficient of determination r^2.

Solution:

a. They have an approximately negative linear relationship.
b.

$$SSR = SST - SSE = 68200 - 6394$$
$$r^2 = \frac{SSR}{SST} = \frac{68200 - 6394}{68200} \approx 0.906 = 90.6\%$$

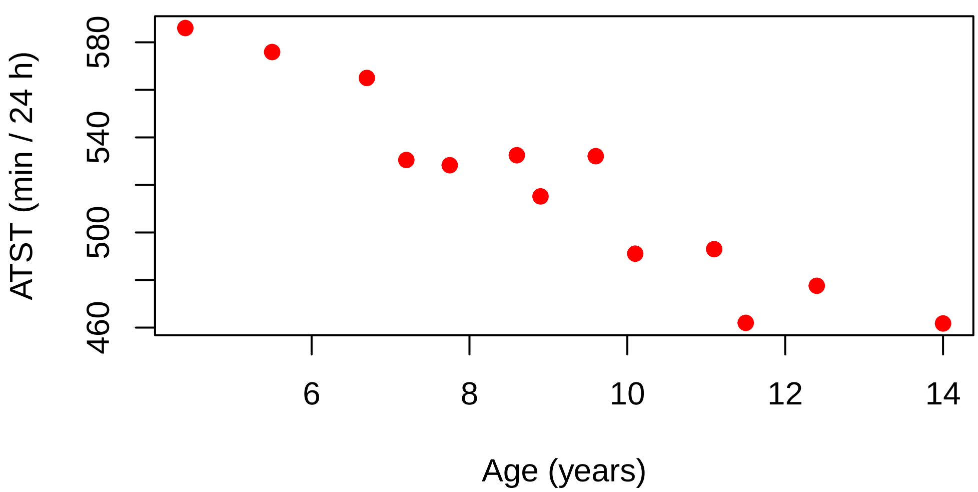
c. About 90.6% of the variance in coffee sales is explained by the model.

Exercise 10 (11 points)**.**

A group of 13 children and adolescents (considered healthy) participated in a psychological study designed to analyze the relationship between age and average total sleep time (ATST). To obtain a measure for ATST (in minutes), recordings were taken on each subject on three consecutive nights and then averaged. The results obtained are displayed in the following table.

Age (years)	ATST (min / 24 h)
4.40	586.00
14.00	461.75
10.10	491.10
6.70	565.00
11.50	462.00
9.60	532.10
12.40	477.60
8.90	515.20
11.10	493.00
7.75	528.30
5.50	575.90
8.60	532.50
7.20	530.50

The data is shown in the following plot. The r code output is also shown below.




```r
fit <- lm(ATST ~ AGE)
summary(fit)
```

```
Call:
lm(formula = ATST ~ AGE)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-23.011  -9.365   2.372   6.770  20.411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  646.483     12.918   50.05 2.49e-14 ***
AGE          -14.041      1.368  -10.26 5.70e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.15 on 11 degrees of freedom
Multiple R-squared:  0.9054,	Adjusted R-squared:  0.8968
F-statistic: 105.3 on 1 and 11 DF,  p-value: 5.7e-07
```

```r
anova_alt(fit)
```

```
Analysis of Variance Table

       Df      SS    MS      F       P
Source  1 18220.5 18221 105.33 5.7e-07
Error  11  1902.8   173
Total  12 20123.4
```

a. The psychologists were interested in the correlation between ATST (y) and AGE (x). Below is the scatterplot for these two variables. Do you observe a trend?

b. The regression of ATST on Age R output is as follows. Calculate the coefficient of correlation r relating ATST y to AGE x from the R output above (round to 3 decimal places).

c. Interpret the coefficient of correlation r you got in part (b). That is, is the linear relationship weak or strong? Is it a positive linear relationship or negative?

d. Conduct a test to determine whether ATST y is correlated with AGE x. Use $\alpha = .01$. What are your null and alternative hypotheses?

   A. $H_0 : \beta_1 \neq 0$, $H_a : \beta_1 = 0$
   B. $H_0 : \beta_0 = 0$, $H_a : \beta_0 > 0$
   C. $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$
   D. $H_0 : \beta_0 = 0$, $H_a : \beta_0 \neq 0$
   E. $H_0 : \beta_1 > 0$, $H_a : \beta_1 = 0$

e. Conduct a test to determine whether ATST y is correlated with AGE x. Use $\alpha = .01$. What is your conclusion?

   A. As p-value is greater than $\alpha = .01$, we fail to reject the null hypothesis.
   B. As p-value is smaller than $\alpha = .01$, we fail to reject the null hypothesis.
   C. As p-value is greater than $\alpha = .01$, we reject the null hypothesis.
   D. As p-value is smaller than $\alpha = .01$, we reject the null hypothesis.

f. Using the conclusion drawn in the previous step, determine whether there is sufficient evidence to support the claim of a linear correlation between the ATST and the AGE. Choose the correct answer.

   A. We conclude there is sufficient evidence to indicate the nonlinear correlation between the ATST and AGE at $\alpha = .01$.
   B. We conclude there is not enough evidence to indicate the linear correlation between the ATST and AGE at $\alpha = .01$.
   C. We conclude there is not enough evidence to indicate the nonlinear correlation between the ATST and AGE at $\alpha = .01$.
   D. We conclude there is sufficient evidence to indicate the linear correlation between the ATST and AGE at $\alpha = .01$.

g. Find $\hat{\sigma}^2$, the variance of the random error term in the model.

h. SSE = ? (round to 2 decimal places)

i. MSE = ? (round to 2 decimal places)

j. SE = ? (round to 2 decimal places)

k. Give a practical interpretation of the value of s.

Solution:

a. There is an approximately negative linear trend.
b. $-\sqrt{0.9054} \approx -0.952$
c. A strong negative linear relationship.
d. C
e. D
f. D
g. 173
h. 1902.80
i. 173
j. $\sqrt{173} \approx 13.15$
k. It is an estimate of the standard deviation of the random error term. Roughly, about 68% of the observed responses are expected to fall within s of the fitted value, and about 95% within $2s$ (assuming the error term is approximately normal).

Exercise 11 (4 points)**.**

Scientists conducted a series of trials to investigate the feeding habits of a kind of birds. They were deprived of food until their guts were empty, then were allowed to feed for 6 hours. For each of 42 feeding trials, the change in the weight of the birds after 2.5 hours was recorded as a percentage of initial weight. The other variable recorded was digestion efficiency (measured as a percentage).

The data for selected feeding trials are listed in the table below.

Trial	Weight Change (%)	Digestion Efficiency (%)
1	-6.0	0.0
2	-5.0	2.5
3	-4.5	5.0
4	0.0	0.0
5	2.0	0.0
38	9.0	59.0
39	12.0	52.5
40	8.5	75.0
41	10.5	72.5
42	14.0	69.0

```r
fit <- lm(weight_change ~ DIGEFF, data=dat)
summary(fit)
```

```
Call:
lm(formula = weight_change ~ DIGEFF, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5501 -0.3057  0.6570  1.1310  3.7291 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.72912    0.55135  -3.136  0.00321 ** 
DIGEFF       0.21039    0.01331  15.802  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.956 on 40 degrees of freedom
Multiple R-squared:  0.8619,    Adjusted R-squared:  0.8585
F-statistic: 249.7 on 1 and 40 DF,  p-value: < 2.2e-16
```

```r
anova_alt(fit)
```

```
Analysis of Variance Table

       Df      SS     MS     F          P
Source  1  955.34 955.34 249.7 8.5337e-19
Error  40  153.04   3.83
Total  41 1108.38
```

```r
coef <- coefficients(summary(fit))
coef_CI <- confint(fit, level=0.90)
cbind(coef, coef_CI)
```

```
              Estimate Std. Error  t value     Pr(>|t|)        5 %       95 %
(Intercept) -1.7291172 0.55135210 -3.13614 3.205569e-03 -2.6575120 -0.8007224
DIGEFF       0.2103902 0.01331432 15.80180 8.533747e-19  0.1879709  0.2328095
```

```r
new_digeff <- data.frame(DIGEFF = c(27.5, 35))
CI <- predict(fit, new_digeff, interval='confidence', level=0.90)
cbind(new_digeff, CI)
```

```
  DIGEFF      fit      lwr      upr
1   27.5 4.056613 3.523681 4.589545
2   35.0 5.634539 5.126261 6.142818
```

```r
PI <- predict(fit, new_digeff, interval='prediction', level=0.90)
cbind(new_digeff, PI)
```

```
  DIGEFF      fit      lwr      upr
1   27.5 4.056613 0.720138 7.393088
2   35.0 5.634539 2.301913 8.967165
```

Snow geese feeding trial R output.

a. Find a 90% confidence interval for the true slope of the line.
b. Based on the 90% confidence interval from the previous question, what can you conclude?

   A. The confidence interval contains 0. Then, β_1 is significantly different from 0.
   B. The confidence interval contains 0. Then, β_1 is not significantly different from 0.
   C. The confidence interval doesn't contain 0. Then, β_1 is significantly different from 0.
   D. The confidence interval doesn't contain 0. Then, β_1 is not significantly different from 0.

c. The researcher wants to estimate the average weight change of all the birds with digestion efficiency 35% (i.e., DIGEFF = 35). From the R output, find the values of the interval at the 90% confidence level that is desired by the researcher.
d. Give a practical interpretation of the interval in part (c).

Solution:

a. (0.093, 0.190) (0.1879709, 0.2328095)
b. C
c. (0.714, 2.847) (5.1262612, 6.1428177)
d. We are 90% confident that the true mean weight change for the birds with digestion efficiency 35% lies within this interval.