ISyE 3770 Assignment 7: Hypothesis Testing and Linear Regression

Due date: 11:59 PM, Friday, April 26, 2024.

Question 1 (Find type-I/type-II errors and p-value)**.** *A manufacturer is interested in the output voltage of a power supply used in a PC. Output voltage is assumed to be normally distributed, with standard deviation 0.25 volt, and the manufacturer wishes to test $H_0 : \mu = 5$ volts against $H_1 : \mu \neq 5$ volts, using $n = 8$ units.*

1) The acceptance region is $4.85 \leq \bar{x} \leq 5.15$. Find the value of α (Type I error).

2) Find the power of the test for detecting a true mean output voltage of 5.1 volts.

3) Find the p-value of the observed statistic is (i) $\bar{x} = 5.2$ and (ii) $\bar{x} = 4.7$, respectively.

Solution. 1) We can see $\bar{x} \sim N(5, 0.25^2/8)$ when H_0 is true. In this case, $\alpha = 1 - \mathbb{P}(4.85 \leq \bar{x} \leq 5.15) = 0.0897$

2) For the case $\mu = 5.1$, $\bar{x} \sim N(5.1, 0.25^2/8)$. In this case,

$$\text{Power} = 1 - \beta = 1 - \mathbb{P}(4.85 \leq \bar{x} \leq 5.15) = 0.288.$$

3) For the first case,

$$\text{p-value} = 2 \cdot \mathbb{P}(\bar{x} \geq 5.2) = 0.0236.$$

For the second,

$$\text{p-value} = 2 \cdot \mathbb{P}(\bar{x} \leq 4.7) = 0.000688.$$

□

Question 2 (Hypothesis testing for sample mean difference)**.** *A computer scientist is investigating the usefulness of two different design languages in improving programming tasks. Twelve expert programmers, familiar with both languages, are asked to code a standard function in both languages, and the time (in minutes) is recorded. It is assumed that the standard deviation of the first design language is* 13 *and that of the second is* 21*. Both design languages follow normal distribution. The data is presented as follows:*

Design Language 1 $17, 16, 21, 14, 18, 24, 16, 14, 21, 23, 13, 18$

Design Language 2 $18, 14, 19, 11, 23, 21, 10, 13, 19, 24, 15, 20$

1) *Find a 95% confidence interval on the difference in mean coding times. Is there any indication that one design language is preferable?*
2) *If hypothesis test is performed to answer the above question: whether one design language is preferable or not. Write down the null and alternative hypothesis. Report the* p*-value and draw a conclusion (either accept or reject* H_0*) under significance level* 0.05*.*

Solution. 1) Using the following code, we find the 95% confidence interval should be $[-13.31, 14.64]$. We cannot determine which design language is preferable.

```
data1 <- c(17,16,21,14,18,24,16,14,21,23,13,18)
data2 <- c(18,14,19,11,23,21,10,13,19,24,15,20)
lb <- mean(data1-data2) - qnorm(0.975) * sqrt(13^2/length(data1) + 21^2/length(data2))
ub <- mean(data1-data2) + qnorm(0.975) * sqrt(13^2/length(data1) + 21^2/length(data2))
```

2) $H_0: \ \mu_1 - \mu_2 = 0, \quad H_1: \quad \mu_1 - \mu_2 \neq 0$. Here the sample size $n = 12$, and the realization value of the testing statistic $\bar{X}_1 - \bar{X}_2$ (i.e., the sample mean difference between two groups) is 0.667. The p-value should be

$$2 * \mathbb{P}(\bar{X}_1 - \bar{X}_2 > 0.667 \mid H_0) = 2 * \mathbb{P}(\mathcal{N}(0, \frac{13^2}{n} + \frac{21^2}{n}) > 0.667) = 0.9254.$$

Since the p-value is greater than 0.05, we do not have enough evidence to reject H_0.

□

Question 3 (Programming Problem)**.** *Cloud seeding has been studied for many decades as a weather modification procedure (for an interesting study of this subject, see the article in* Technometrics*, "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification," Vol. 17, pp. 161 - 166). The rainfall in acre-feet from 23 clouds that were selected at random and seeded with silver nitrate is given below:*

$$49.44, 11.33, 12.33, 30.30, 30.50, 35.40, 36.40, 34.76, 33.21, 39.66, 37.53, 31.33$$

$$36.54, 38.54, 27.32, 26.43, 25.43, 26.65, 43.33, 42.33, 44.32, 41.32, 54.33$$

(Hint: you can load the data using the rain.txt file in the attachment).

It is assumed that the rain fall follows normal distribution with known standard deviation $\sigma = 10$*. Now, answer the following questions:*

1) *Suppose we want to test the claim that the mean rain fall is above 30. What is the null and alternative hypotheses that we should use to test this claim?*
2) *Perform a* Z*-test (see page 26 in slides Week11.pdf) to test the claim in question 3. Use a confidence level of 95%. What are your conclusions?*

   You may use:

   ```
   z.test(data, alternative = xxx, mu=xxx, sigma.x=xxx, conf.level=xxx)
   Where,
          data should be Rain,

          alternative is either alternative=c("greater"), or alternative=c("less")
          or alternative=c("two.sided") depending on your alternative hypothesis,

          mu should be 30,

          sigma.x should be 10,

          conf.level should be 0.95
   ```

3) *What was the* p*-value for the hypothesis test in part 2)?*

Solution. 1) $H_0: \mu = 30, H_1: \mu > 30$.

2) Based on the code below, we reject H_0 with confidence level 0.95.

```
> model = lm(y ~ x1 + x2 + x3 + x4)
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-14.098  -9.778   1.767   6.798  13.016

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -123.1312   157.2561  -0.783    0.459
x1             0.7573     0.2791   2.713    0.030 *
x2             7.5188     4.0101   1.875    0.103
x3             2.4831     1.8094   1.372    0.212
x4            -0.4811     0.5552  -0.867    0.415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.79 on 7 degrees of freedom
Multiple R-squared:  0.852,   Adjusted R-squared:  0.7675
F-statistic: 10.08 on 4 and 7 DF,  p-value: 0.00496
```

```
library(BSDA)
data <- read.table("rain.txt")
z.test(data, alternative=c("greater"), mu=30, sigma.x=10, conf.level = 0.95)
```

3) From the output in Part 2), the p-value is 0.01976.

□

Question 4 (Computation in simple linear regression)**.** *Regression analysis were used to analyze the data from a study investigating the relationship between roadway surface temperature (°F) (x) and pavement deflection (y). Summary quantities were* $n = 20$, $\sum y_i = 12.75$, $\sum y_i^2 = 8.86$, $\sum x_i = 1478$, $\sum x_i^2 = 143,215.8$, $\sum x_i y_i = 1083.67$.

1) Calculate the least squares estimate of the slope and intercept. Graph the regression line. Estimate σ^2.

2) *Use the equation of the fitted line to predict what pavement deflction would be observed when the surface temperature is 85F.*
3) *What is the mean pavement deflection when the surface temperature is 90F.*
4) *What change in mean pavement deflection would be expected for a 1F change in surface temperature.*

Solution.. 1)

$$\hat{\beta}_1 = 4.161 \cdot 10^{-3}, \qquad \hat{\beta}_0 = 0.330.$$

Please draw regression line using R command `abline(0.330, 4.161 * 10^-3)`.
$\widehat{\sigma}^2 = 7.692 \cdot 10^{-3}$.

2) When $x = 85$,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.6837.$$

3) When $x = 90$,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.7045.$$

4) The expected change would be $\hat{\beta}_1 \cdot 1 = 4.161 \cdot 10^{-3}$.

□

Question 5 (Simple Linear Regression)**.** *We have data on the quarterback ratings for the 2008 National Football League season (Source: The Sports Network). It is suspected that the rating y is related to the average number of yards gained per pass attempt (x).*

Getting the Data*: Data in this question is contained in file data113.csv. Once you have saved the data file in the working directory, read the data in R using the command*

data = read.csv("data113.csv",header=TRUE)

We can investigate the association of `Rating` *to* `Yds` *using linear regression. Define first*

```
y = data$Rating
x = data$Yds
```

The function to fit a linear regression model in R is `lm`*. We perform a linear regression with R as follows*

```
model = lm(y~x)
summary(model)
```

You can use the fitted results to help you answer above questions. Make sure to include the R output in your homework and state clearly how do you use the R output to solve the following questions.

1) *Calculate the least squares estimates of the slope and intercept. What is the estimate of* σ^2*? Graph the regression model.*
2) *Find an estimate of the mean rating if a quarterback averages* 7.5 *yards per attempt.*
3) *What change in the mean rating is associated with a decrease of one yard per attempt?*
4) *To increase the mean rating by* 10 *points, how much increase in the average yards per attempt must be generated?*
5) *Find* 95% *confidence intervals on the slope and intercept.*

   (Hint: you can get these CIs using R: `confint(model))`
6) *Find a* 95% *prediction interval on the rating when the average yards per attempt is* 8.0*.*

   (Hint: you can get these CIs using R: `predict(model, data.frame(x=8), interval = c("prediction"))`*).*

Solution. 1) $\hat{\beta}_0 = 14.195, \hat{\beta}_1 = 10.092, \hat{\sigma}^2 = 27.237$

2) $\hat{y} = \hat{\beta}_0 + 7.5 * \hat{\beta}_1 = 89.885$
3) The change is $\hat{\beta}_1 \cdot (-1) = -10.092$.
4) The increase in yards should be $\frac{10}{\hat{\beta}_1} = 0.991$.
5) Confidence interval for β_0: $[-4.30, 32.70]$
   Confidence interval for β_1: $[7.46, 12.72]$
6) $[83.79, 106.07]$.

□

Question 6 (Multiple Linear Regression)**.** *The electric power consumed each month by a chemical plant to thought to be related to the average ambient temperature* x_1*, the number of days in the month* x_2*, the average product purity* x_3*, and the tons of product produced* x_4*. The past year's historical data are available and are presented in the following table.*

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
270	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

Use R to find the multiple linear regression model. For example, you may create the variable x_1 in R using the following command

```
x1 = c(25,31,45,60,65,72,80,84,75,60,50,38)
```

You can fit a multiple linear regression model similar to simple linear regression using the following command

```
model = lm(y~x1+x2+x3+x4)
summary(model)
```

Based on the results *or R, answer the following questions:*

1) *Fit a multiple linear regression model to these data.*
2) *Estimate σ^2.*
3) *Compute the standard errors of the regression coefficients. Are all of the model parameters estimated with the same precision? Why or why not?*
4) *Predict the power consumption for a month in which $x_1 = 75$ F, $x_2 = 24$ days, $x_3 = 90\%$, and $x_4 = 98$ tons.*

```
> model = lm(y ~ x1 + x2 + x3 + x4)
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-14.098  -9.778   1.767   6.798  13.016

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -123.1312   157.2561  -0.783    0.459
x1             0.7573     0.2791   2.713    0.030 *
x2             7.5188     4.0101   1.875    0.103
x3             2.4831     1.8094   1.372    0.212
x4            -0.4811     0.5552  -0.867    0.415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.79 on 7 degrees of freedom
Multiple R-squared:  0.852,   Adjusted R-squared:  0.7675
F-statistic: 10.08 on 4 and 7 DF,  p-value: 0.00496
```

Solution. 1) As indicated in blue box, the fitted model is

$$\hat{y} = -123.13 + 0.7573x_1 + 7.5188x_2 + 2.4831x_3 - 0.4811x_4.$$

2) As indicated in red box, $\widehat{\sigma}^2 = 11.79^2 = 139.0041$.

3) As indicated in green box, the standard errors are

$$157.2561,\ 0.2791,\ 4.0101,\ 1.8094,\ 0.5552.$$

Since the standard errors give an absolute measurement of the error, from the above standard error values, we know that the parameters are estimated with different precision.

4) Plugging in x_1, x_2, x_3, x_4, we have

$$\hat{y} = -123.13 + 0.7573 \cdot 75 + 7.5188 \cdot 24 + 2.4831 \cdot 90 - 0.4811 \cdot 98 = 290.4487.$$

□