HW3_2.5 bcd

2023-09-10

- 2.1: You do not need to state the 5 steps of hypothesis testing when answering 2.1. Simply answer Yes/No and give your reason. Check out this online resource for additional help.
- 2.5 (b) and (d): **State the 5 steps of hypothesis testing following the lecture slides (instead of answering those items mentioned in the question).**
- 2.5 (d): This is a one-sided test. Refer to Example 2 on textbook p.47. For additional hints on obtaining the correct p-value, see STP 530 prerequisite refresher.pdf (bottom of p.15 and top of p.16).
- For all questions you can either use R or hand calculation to arrive at the numbers you need for your answers. Either way you need to clearly show how you get there. Follow the general homework guideline for what's required.

2.1 A student working on a summer internship in the economic research department of a large corporation studied the relation between sales of a product (Y, in million dollars) and population (X, in million persons) in the firm's 50 marketing districts. The normal error regression model (2.1) was employed. The student first wished to test whether or not a linear association between Y and X existed. The student accessed a simple linear regression program and obtained the following information on the regression coefficients:

Parameter	Estimated Value	95 Percent Confidence Limits	
Intercept	7.43119	−1.18518	16.0476
Slope	.755048	.452886	1.05721

a. The student concluded from these results that there is a linear association between *Y* and X. Is the conclusion warranted? What is the implied level of significance?

β1= 0.755048 and 95% CI for β1 = [0.452886, 1.05721].
The regression analysis indicates that with 95% confidence, the sales (Y) increase about $453K – $1M per every 1M population increase in the 50 marketing districts.

In this β1 interval, it does not include β1 = 0.
The β1=0 implicates that there is no linear relationship between Y and X (when the slop is zero). Because the 95% CI for β1 is > zero, it can be concluded that the population in the 50 districts has a linear effect on sales of product; thus, yes, the linear association between Y and X is warranted, with 5% significance level.

b. Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.
β0 = 7.43119 and 95% CI for β0 = [-1.18518, 16.0476].
Indeed, in a district with zero population with $0 sales, the company might still have maintenance cost, thus leading to a negative β0. Alternatively, the negative β0 value could be due to extrapolation.

In some cases, an intercept, even if it doesn't make logical sense, can still be statistically estimated. This might result from extrapolation beyond the data's range or from not meeting the assumptions of the normal error regression model, $\varepsilon \sim iid\ N(0, \sigma^2)$.

- The model might be extrapolating beyond the range of observed population, causing the estimated β0 to be an extrapolation outside the range of the data.

- Or, it could be due to a violation of assumptions.
    1) Identically distributed: the variance of the residuals should be constant across all levels of the independent variable(s). All the error terms follow the same probability distribution. In this context, this means that the variability in sales should be roughly the same for districts with small populations and those with large populations. For example, if sales are more variable in larger districts, this would violate the assumption.
    2) Independent observation: error for one observation (sales in different marketing districts) is not influenced by the error term of any other observation, independent. If there are spatial or temporal correlations between districts, this assumption is violated. For example, if two districts are geographically close, an event affecting one might also affect the other, thus violating the independence assumption.
    3) Normality of errors: The residuals should be normally distributed with mean of 0 and constant variance σ^2. If they are not, it could affect the validity of some tests and confidence intervals. For example, if there are outliers in the data or if the sales distribution is skewed, this leads to non-normal errors. This can be checked using a variety of methods, including Q-Q plots.

2.5. Refer to Copier maintenance Problem 1.20.

1.20. Copier maintenance. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	43	44	45
X_i:	2	4	3	...	2	4	5
Y_i:	20	60	46	...	27	61	77

b. Conduct a t test to determine whether or not there is a linear association between X and Y here; control the a risk at .10. State the alternatives, decision rule, and conclusion. What is the P-value of your test?

Step 1: Check assumptions: We assume the error term follows $\varepsilon \sim$ iid $N(0, \sigma^2)$. Identical distribution. The variance of the residuals should be constant.

- Independent distribution: Independent observation. Data are independently collected. Here, each observation (service call) is independent.
- Normal error distribution: The residuals are normally distributed.

Step 2: Construct hypothesis

H0: β1 = 0

H1: β1 ≠ 0 (two-tailed test, there is a linear association between X and Y)

Step 3: Calculate t-statistic (t_{obs}).

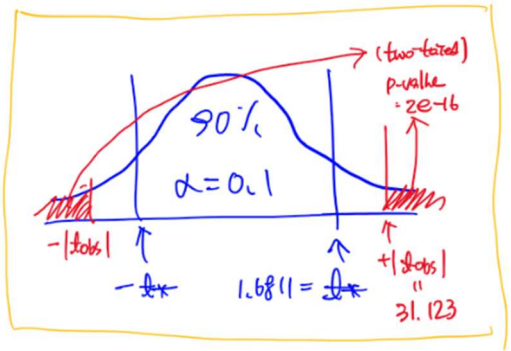
$t.obs = \frac{b1-\beta1|H0}{s\{b1\}} = \frac{15.0352-0}{0.4831} = 31.123$

α = 0.1

df = 45 – 2 = 43

t. critical (t*) = 1.681071 (by R: qt(1-0.05, 43))

t_{obs} >> t*



Step 4: Find p-value

p-value is shown in the summary output in R: Pr(>|t|) <2e-16

Alternatively, it can be found by R code: 1-pt(t.obs=31.123, 43).

p-value << α (significance level=0.1)

Step 5: Make conclusion

With p-value 2e-16 (< α = 0.1), there is enough evidence to reject H0 and conclude that the true population slop (β1) is not zero. This implies that there is a linear association between X (the No. of copiers served) and Y (total service time).

c. Are your results in parts (a) and (b) consistent? Explain.

Yes, consistent. Both (a) and (b) says that with 90% confidence, the true population slop is not zero (falls between 14.22 and 15.85), thus implying a linear relationship between X and Y.

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 .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Step 1: Check assumptions: $\varepsilon \sim iid\ N(0,\sigma^2)$

We assume the error term follows Identical, independent, normal error distribution.

Step 2: Construct hypothesis

'a risk of Type I error at 0.05' indicates the probability of falsely rejecting H0 is 0.05.

α = 0.05

H0: β1 ≤ 14 (mean service time increase by ≤14min per additional copier to be serviced)

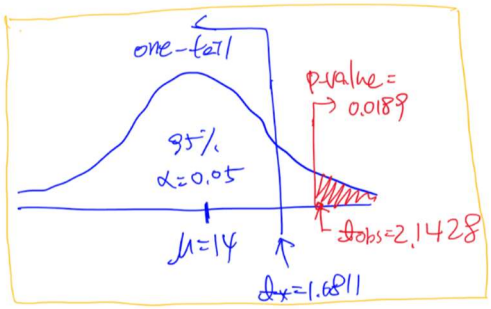
H1: β1 > 14 (one-tail test)

Step 3: Calculate t-statistic (t.obs).

t.obs = $\frac{b1-\beta 1|H0}{s\{b1\}} = \frac{15.0352-14}{0.4831} = 2.142828$

t. critical (t*) = 1.681071 (by R: qt(1-0.05, 43))

$t_{obs} > t^*$



Step 4: Find p-value.

p.value <- 1-pt(2.142828 , 43)

p-value = 0.01891429 < α , thus reject H0

Step 5: Make conclusion.

With p-value < 0.05, there is enough evidence to reject H0 and conclude the true population slop β1 > 14. Thus, with 95% confidence, the increase in mean required time for each additional copier is significantly greater than 14 minutes in Tri-City. Therefore, Tri-City does not seem to be satisfying the manufacturer's suggestion.

R code:

import data

```
setwd("C:/Users/jyang/OneDrive - Arizona State University/10
Classes_OneDrive/2023_STP530_Regression")
HW2.data <- read.table("CH01PR20.txt")
colnames(HW2.data) <- c("min", "ser.no")
```

regression model

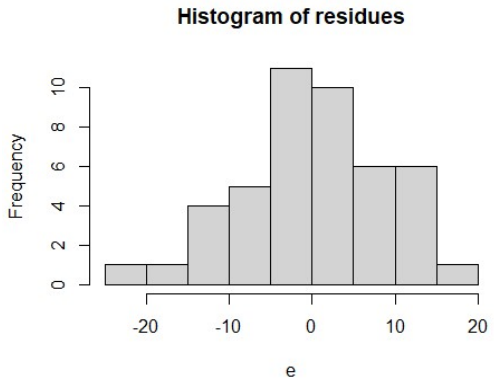
```
HW2.mod <- lm(min ~ ser.no, data = HW2.data)
summary(HW2.mod)
```

```
##
## Call:
## lm(formula = min ~ ser.no, data = HW2.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## ser.no       15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

normal error distribution (by basic histogram for ei)

ei = Yi - Yi^ y.hat = b0 + b1X

```
y.hat <- predict(HW2.mod)
e <- HW2.data$min - y.hat
breaks_vector <- seq(from=-25, to=25, by=5)
hist(e, main="Histogram of residues")
```



HW3. 2.5b

```
t.critical <- qt(1-0.1/2, 43)
t.critical

## [1] 1.681071

t.obs <- 31.123
p.value <- 1-pt(31.123, 43)
p.value

## [1] 0
```

#HW3. 2.5d

```
t.obs.14 = (15.0352-14)/0.4831
t.obs.14

## [1] 2.142828

p.value <- (1-pt(2.142828 , 43))
p.value

## [1] 0.01891429
```