STP 530: Applied Regression Analysis

Name : **Sai Swaroop Reddy Vennapusa**

Homework 3

Instructor : **Yi Zheng**

Due Date : 12th Sep 2023, 10:30AM

Question 2.1(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?

Answer:
The slope of the regression, representing the association between sales Y and population X, has an estimated value of .755048. The 95 percent confidence interval for the slope is (.452886, 1.05721). Given that the entire confidence interval is positive and does not contain zero, we can infer that there is a significant positive linear association between sales Y and population X at the 0.05 significance level. Therefore, the conclusion that there is a linear association between Y and X is warranted. The implied level of significance is 5% (since it's a 95% confidence interval).

Yes, the conclusion is warranted. The implied level of significance is 5%.

Question 2.1(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.

Answer:
The negative lower confidence limit for the intercept indicates that if there were a district with a population of zero, the 95% confidence interval for the sales in million dollars would range from -1.18518 to 16.0476. This doesn't make practical sense as sales can't be negative.

However, it's essential to understand that the intercept in a regression model represents the estimated outcome when the predictor is zero, which might not always have a practical interpretation, especially if there are no actual data points near zero. It's the slope of the regression that tells us about the relationship between Y and X. In many cases, the intercept might not be interpretable, especially if an X value of 0 is outside the range of observed data or doesn't make practical sense in the given context.

Conclusion:
The negative lower confidence limit for the intercept might be a result of the mathematical calculations of the regression model and does not necessarily imply a realistic scenario. In this context, having a district with zero population is also impractical. It's essential to interpret the intercept in light of the actual range of the data and the context of the problem.

Question 2.5(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?

Answer:

```
> summary(m)

Call:
lm(formula = Total.num.of.minutes.spent.by.servicemen ~ Num.of.copiers.serviced,
    data = mydata)

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
Num.of.copiers.serviced  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

> confint(m, level=0.9)
                             5 %      95 %
(Intercept)             -5.29378  4.133467
Num.of.copiers.serviced 14.22314 15.847352
```

1. Check the assumptions:
- We assume that the residuals are normally distributed, which is an assumption of linear regression.
- Independence of observations: The calls are independent of each other.
- Linearity: The relationship between the number of copiers serviced and the service time is linear.

2. Construct the hypotheses:
H_0: beta_1 = 0 (No linear association between X and Y)
H_a: beta_1 != 0 (There is a linear association between X and Y)

3. Calculate the test statistic:
The t-value for the coefficient `Num.of.copiers.serviced` is 31.123 (from the regression summary).

4. Find the p-value:
The p-value is < 2.2e-16 (from the regression summary), which is virtually 0.

5. Make the conclusion:
Given that the p-value is much smaller than the significance level of 0.10, we reject the null hypothesis. There is significant evidence to suggest a linear association between the number of copiers serviced and the service time.

Question 2.5(c). Are your results in parts (a) and (b) consistent? Explain.

Answer:
Given:

2.5(a)
For every additional copier serviced, the mean service time increases by approximately 15.0352 minutes. The 90% confidence interval for this increase ranges from about 14.22 minutes to 15.85 minutes.

2.5(b)
Based on the t-test, there's strong evidence of a linear association between the number of copiers serviced and the mean service time. The p-value is virtually 0, much less than the significance level of 0.10, leading us to reject the null hypothesis that there's no association.

Conclusion:
Yes, the results from parts (a) and (b) are consistent. In part (a), the estimated coefficient of `Num.of.copiers.serviced` is significantly different from zero, as indicated by its confidence interval which does not contain zero. This provides evidence of a linear association. In part (b), the t-test, which directly tests the significance of this relationship, also indicates a significant linear association. Both results point towards the same conclusion: There's a significant linear relationship between the number of copiers serviced and the mean service time.

Question 2.5(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?

Answer:

1. Check the assumptions:
- We assume that the residuals are normally distributed, which is an assumption of linear regression.
- Independence of observations: The calls are independent of each other.
- Linearity: The relationship between the number of copiers serviced and the service time is linear.

2. Construct the hypotheses:
H_0: $\beta_1 \not< 14$ (Mean service time increases by 14 minutes or less for each additional copier)
H_a: $\beta_1 > 14$ (Mean service time increases by more than 14 minutes for each additional copier)

R Code:

```r
# Question 2.5(d)

# Given values from regression summary
beta_estimate <- 15.0352
standard_error <- 0.4831
df <- 43  # degrees of freedom = n - 2

# Compute t-statistic
t_statistic <- (beta_estimate - 14) / standard_error

# Compute the one-sided p-value
p_value <- 1 - pt(t_statistic, df)

p_value
```

R Output:

```
> # Given values from regression summary
> beta_estimate <- 15.0352
> standard_error <- 0.4831
> df <- 43  # degrees of freedom = n - 2
> # Compute t-statistic
> t_statistic <- (beta_estimate - 14) / standard_error
> # Compute the one-sided p-value
> p_value <- 1 - pt(t_statistic, df)
> p_value
[1] 0.0189143
```

3. Calculate the test statistic:

`t_statistic <- (beta_estimate - 14) / standard_error`
beta_estimate is the estimated coefficient for the number of copiers serviced (15.0352), and SE is the standard error of the coefficient (0.4831).

4. Find the p-value:
From the R code, the computed one-sided p-value is 0.0189143.

5. Make the conclusion:
Given that the p-value (0.0189143) is less than the significance level of 0.05, we reject the null hypothesis H0. This means there is enough evidence at the 0.05 significance level to conclude that the mean increase in service time for each additional copier serviced is significantly more than 14 minutes.