Statistics 2 2025/26 Resit

Introduction

The human resources (HR) department in your large firm has collected data on employee satisfaction from a random sample of $n = 14999$ employees. They are interested in the factors driving this. They are asking for your help to perform the statistical analysis.

Here is a link to download the dataset: `employee-satisfaction.csv`.

The definitions of all the variables are as follows:

- `y`: Employee satisfaction score (between 0 and 1) where a higher number means more satisfied.
- `num_projects`: The number of projects the employee has worked on in the last year.
- `salary`: A categorical variable indicating the employee's salary (`"high"`, `"medium"` or `"low"`).
- `promotion`: A dummy variable that equals 1 if the employee received a promotion in the last 5 years and 0 otherwise.

Question 1

What is the sample covariance between `y` and `num_projects`?

Model 1

The HR department want to know how promotions affect employee satisfaction. Estimate a linear regression model with `y` as the dependent variable and `promotion` as the independent variable.

If you estimated the model correctly, you should have an estimated intercept of 0.611895.

Question 2

What is the sample regression slope?

Question 3

Provide a 99% confidence interval for the sample regression slope.

- Lower bound: _____
- Upper bound: _____

Question 4

According to the estimated model, what is the average employee satisfaction of employees that received a promotion in the last 5 years.

Question 5

Test the following claim at the 5% level:

"Employees that received a promotion in the last 5 years on average have a higher satisfaction level of *more than* 0.02 compared to employees that did not receive a promotion in the last 5 years."

Perform this test by answering the questions below.

- What is the null hypothesis? $\beta_1 < / \leq / > / \geq / = / \neq$ _____ (choose one comparison operator and fill in a value in the blank).
- What is the alternative hypothesis? $\beta_1 < / \leq / > / \geq / = / \neq$ _____ (choose one comparison operator and fill in a value in the blank).
- Under the null hypothesis, the test statistic $T = (B_1 - b)/S_{B_1}$, where b is the hinge, follows a t distribution with how many degrees of freedom? _____
- What is the value of the test statistic? _____
- What is the associated p-value? _____
- What is your conclusion? Choose an option below:
  - Reject H0: There is sufficient evidence for the claim.
  - Reject H0: There is not sufficient evidence for the claim.
  - Don't reject H0: There is sufficient evidence for the claim.
  - Don't reject H0: There is not sufficient evidence for the claim.

Model 2

The HR department is interested in the relationship between employee satisfaction and work pressure. They have the following theory:

- Too much work increases stress and is bad for employee satisfaction.
- Too little work makes work boring and is also bad for employee satisfaction.

- There is an optimal amount of work in between these extremes that is best for employee satisfaction.

To test this theory you estimate a regression model with `y` as the dependent variable and the following two independent variables:

- `num_projects`
- The square of `num_projects`.

Note: it is possible to add the number of projects squared to your model by adding `I(num_projects^2)` to your model formula in the `lm()` function.

If you estimated the model correctly, you should have an estimated intercept of -0.3376667.

Question 6

Use the estimated model to predict the employee satisfaction of an employee with 3 projects.

Also use the model to provide the lower and upper bound of a 95% confidence interval for the mean employee satisfaction for employees with 3 projects.

- Predicted employee satisfaction: _____
- Confidence interval lower bound: _____
- Confidence interval upper bound: _____

Question 7

Employees in the data work on between 2 and 7 projects. According to the estimated model, what number of projects is on average best for employee satisfaction? The answer must be a whole number.

Question 8

Perform a formal test for heteroskedasticity by regressing the square of the residuals from your model on `num_projects` and the square of `num_projects` (the same explanatory variables as in the model).

Use a 5% significance level.

In this test, the alternative hypothesis is that the model exhibits which of the following:

- Homoskedasticity
- Heteroskedasticity
- Serial correlation
- Serial uncorrelation

What is the p-value from this test? _____

What is the conclusion? Choose one of the answers below:

- The model exhibits homoskedasticity. Therefore the coefficients are not reliable.
- The model exhibits homoskedasticity. Therefore the standard errors are not reliable.
- The model exhibits heteroskedasticity. Therefore the coefficients are not reliable.
- The model exhibits heteroskedasticity. Therefore the standard errors are not reliable.

Model 3

The HR department is also interested in how salary affects satisfaction. Regress employee satisfaction on the following 2 independent variables:

- A dummy variable for salary being `"low"`.
- A dummy variable for salary being `"medium"`.

Note that by using the `salary` variable in your model formula *as is* the `"high"` salary value will automatically be chosen as the base category. Therefore you can use the `salary` variable in your formula and do not need to make any adjustments to the levels.

If you estimated the model correctly, your estimated regression intercept should equal 0.637470.

Question 9

Define the following:

- $\bar{y}_{low}$: The average employee satisfaction of low-salaried workers.
- $\bar{y}_{medium}$: The average employee satisfaction of medium-salaried workers.
- $\bar{y}_{high}$: The average employee satisfaction of high-salaried workers.

The question below tests your ability to interpret the coefficients from the estimated model. Fill in the blanks below.

According to the estimated model, the average employee satisfaction of high-salaried workers is _____ (value of $\bar{y}_{high}$).

According to the estimated model, the difference in average employee satisfaction between low- and high-salaried workers is _____ (value of $\bar{y}_{low} - \bar{y}_{high}$).

According to the estimated model, the difference in average employee satisfaction between low- and medium-salaried workers is _____ (value of $\bar{y}_{low} - \bar{y}_{medium}$).

Question 10

Perform an appropriate hypothesis test to test the usefulness of the model. Use a 5% significance level.

- The null hypothesis is that *at least one/all/none* (choose one) of $\beta_j <$ / $\leq$ / $>$ / $\geq$ / $=$ / $\neq$ _____ for $j =$ _____ to _____ (choose one comparison operator and fill in values in the blank spaces).
- The alternative hypothesis is that *at least one/all/none* (choose one) of $\beta_j <$ / $\leq$ / $>$ / $\geq$ / $=$ / $\neq$ _____ for the same j (choose one comparison operator and fill in a value in in the blank space).
- The formula for the test statistic is of the form:

$$\frac{\frac{SST-SSE}{a}}{\frac{SSE}{n-k-1}}$$

  What is the value of a in the estimated model? _____
- What is the value of the test statistic? _____
- What is the critical value? _____
- What is your conclusion? (choose one option below):
  - *Reject* H_0. The model is *useful*.
  - *Reject* H_0. The model is *useless*.
  - *Don't reject* H_0. The model is *useful*.
  - *Don't reject* H_0. The model is *useless*.

Model 4

Estimate a multiple linear regression model explaining employee satisfaction by:

- The `promotion` dummy.
- The `"low"` salary dummy.
- The `"medium"` salary dummy.

As in Model 3, by using the `salary` variable as is the `"high"` salary will automatically be chosen as the base category.

If you estimated the model correctly, your estimated regression intercept should equal 0.635366.

Question 11

The estimated intercept 0.635366 is the average employee satisfaction among which subgroup?

- High-salary workers that received a promotion in the last 5 years.
- High-salary workers that did not receive a promotion in the last 5 years.
- Medium-salary workers that received a promotion in the last 5 years.
- Medium-salary workers that did not receive a promotion in the last 5 years.
- Low-salary workers that received a promotion in the last 5 years.
- Low-salary workers that did not receive a promotion in the last 5 years.

Question 12

Which variables are individually significant at the 1% level?

Question 13

Test the joint usefulness of the salary variables, which are variables 2-3 in your the model. Use a 5% significance level.

Choose one of the options in *italics* and fill in the blanks.

- The null hypothesis is that *all/at least one/none* of β_j _____ _____ for j from _____ to _____.
- The alternative hypothesis is that *all/at least one/none* of β_j _____ _____ for the same j.

The test statistic is of the form:

$$\frac{\frac{SSE_r - SSE_c}{a}}{\frac{SSE_c}{n-k-1}}$$

What is the value of a in the test? _____

What is the value of the test statistic? _____

What is the critical value? _____

Which of the 4 options below is the correct conclusion from the test?

- Reject H0. The variables are useful additions to the model.
- Reject H0. The variables are not useful additions to the model.
- Don't reject H0. The variables are useful additions to the model.
- Don't reject H0. The variables are not useful additions to the model.