Statistics 2 2025/26 Resit Solutions

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`?

Answer:

```
df <- read.csv("employee-satisfaction.csv")
cov(df$y, df$num_projects)
```

```
[1] -0.04381449
```

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.

```r
m1 <- lm(y ~ promotion, data = df)
summary(m1)
```

```
Call:
lm(formula = y ~ promotion, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5460 -0.1719  0.0281  0.2081  0.3881

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.611895   0.002051 298.273  < 2e-16 ***
promotion   0.044124   0.014067   3.137  0.00171 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2486 on 14997 degrees of freedom
Multiple R-squared:  0.0006556, Adjusted R-squared:  0.000589
F-statistic: 9.839 on 1 and 14997 DF,  p-value: 0.001712
```

Question 2

What is the sample regression slope?

Answer:

```r
coef(m1)[2]
```

```
 promotion
0.04412371
```

Question 3

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

- Lower bound: _____
- Upper bound: _____

Answer:

```r
confint(m1, parm = "promotion", level = 0.99)
```

```
                0.5 %     99.5 %
promotion 0.007885075 0.08036235
```

Question 4

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

Answer:

The model is:

$$\mathbb{E}[Y_i|promotion_i] = \beta_0 + \beta_1 x_i$$

For employees that received a promotion, this is:

$$\mathbb{E}[Y_i|promotion_i = 1] = \beta_0 + \beta_1$$

Therefore we just need to add the estimate of the intercept and the slope together:

```
sum(coef(m1)[1:2])
```

```
[1] 0.6560188
```

We can check this by calculating the average manually:

```
mean(df$y[df$promotion == 1])
```

```
[1] 0.6560188
```

Which gives us the same answer!

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.

Answer:

According to our model estimates, a promotion leads to an increase in satisfaction of 0.0441 on average. The question is asking if this estimate is precise enough to say that a promotion leads to an average increase of more than 0.02 in the population. Therefore the claim is that $\beta_1 > 0.02$, which is our alternative hypothesis.

- Null hypothesis: $\beta_1 \leq 0.02$.
- Alternative hypothesis: $\beta_1 > 0.02$.
- Degrees of freedom: 14997 (can be read from model summary output). Can also be found by calculating $n - k - 1 = 14999 - 1 - 1 = 14997$.
- Value of the test statistic:

```
b_1 <- coef(summary(m1))[2, 1]
s_b_1 <- coef(summary(m1))[2, 2]
(t <- (b_1 - 0.02) / s_b_1)
```

```
[1] 1.714923
```

- p-value: This is an upper-tail test so we calculate this with:

```
1 - pt(t, 14997)
```

```
[1] 0.04318999
```

- Conclusion: The p-value is less than the significance level (0.05) so we reject H0. Therefore the correct option is: "Reject H0: There is 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.

```
m2 <- lm(y ~ num_projects + I(num_projects^2), data = df)
summary(m2)
```

```
Call:
lm(formula = y ~ num_projects + I(num_projects^2), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61118 -0.12118 -0.00118  0.15459  0.67728

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.3376667  0.0148693  -22.71   <2e-16 ***
num_projects       0.5566323  0.0077484   71.84   <2e-16 ***
I(num_projects^2) -0.0729804  0.0009504  -76.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2085 on 14996 degrees of freedom
Multiple R-squared:  0.2969,    Adjusted R-squared:  0.2968
F-statistic:  3166 on 2 and 14996 DF,  p-value: < 2.2e-16
```

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: _____

Answer: The question is asking for a confidence interval for the mean employee satisfaction given a value of the independent variables. Therefore we use the `interval = "confidence"` option in the `predict()` function.

```
df_p <- data.frame(num_projects = 3)
predict(m2, df_p, interval = "confidence", level = 0.95)
```

```
       fit       lwr       upr
```

```
1 0.6754064 0.6712987 0.6795142
```

Note: because we are using `I(num_projects^2)` in our model formula, we can create a dataframe with only the `num_projects` variable. If instead you estimated the model using a `num_projects_sq` variable that you created, you will need to also add a value for that variable (i.e. `data.frame(num_projects = 3, num_projects_sq = 9)`.

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.

Answer: There are several ways to do this question. One way is to predict the employee satisfaction for each of 2-7 projects. This could be done one-by-one in a similar way to the previous question. For example, we can try 2 projects:

```
df_p <- data.frame(num_projects = 2)
predict(m2, df_p)
```

```
        1
0.4836762
```

And 4 projects:

```
df_p <- data.frame(num_projects = 4)
predict(m2, df_p)
```

```
        1
0.7211757
```

We can continue this until we have the predicted satisfaction for all number of projects 2-7, and report the number of projects that gives the highest predicted satisfaction.

But a faster way would be to create a data frame with the sequence 2-7 and use that to predict the number of projects:

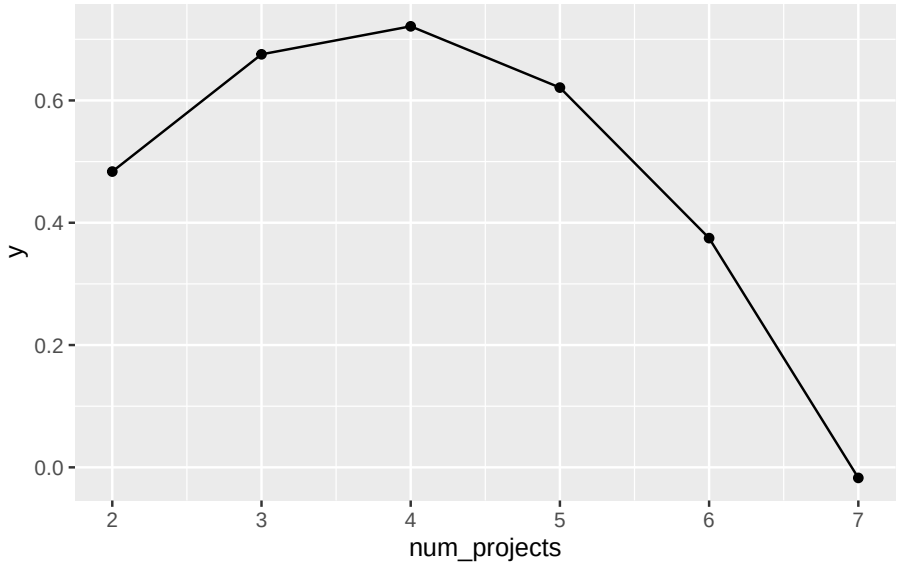
```
df_p <- data.frame(num_projects = 2:7)
df_p$y <- predict(m2, df_p)
df_p
```

```
  num_projects           y
1            2  0.48367622
2            3  0.67540641
3            4  0.72117575
4            5  0.62098423
5            6  0.37483187
6            7 -0.01728134
```

We can see that 4 projects gives the highest predicted satisfaction.

We can also plot these results to see it more clearly:

```
library(ggplot2)
ggplot(df_p, aes(num_projects, y)) + geom_point() + geom_line()
```



Another approach would use calculus. However, because the optimal number of projects needs to be a whole number, this approach is made a bit more complicated and is not what I recommend to do. Nevertheless, if you are interested, I provide the workings here.

The model is:

$$\mathbb{E}[Y_i|num\_projects_i] = \beta_0 + \beta_1 num\_projects_i + \beta_2 num\_projects_i^2$$

Taking derivatives:

$$\frac{\partial \mathbb{E}[Y_i|num\_projects_i]}{\partial num\_projects_i} = \beta_1 + 2\beta_2 num\_projects_i$$

The function $\mathbb{E}[Y_i|num\_projects_i]$ is at an extreme point when $\frac{\partial \mathbb{E}[Y_i|num\_projects_i]}{\partial num\_projects_i} = 0$. This happens at: $\beta_1 + 2\beta_2 num\_projects_i = 0$. Solving for $num\_projects_i$ we get $num\_projects_i = -\frac{\beta_1}{2\beta_2}$. Calculating this in our case gives:

```
- coef(m2)[2] / (2 * coef(m2)[3])
```

```
num_projects
    3.813573
```

Therefore 3.81 projects is an extreme point. The second-order derivative is $2\beta_2$. Because $b_2 < 0$, this means the second-order derivative is negative which means this extreme point is a maximum.

This means the ideal number of projects is 3.81. But because we can only work on a whole number of projects, we need to calculate which of 3 or 4 gives a higher satisfaction.

Our estimate of $\mathbb{E}[Y_i|num\_projects_i = 3]$ is:

```
coef(m2)[1] + coef(m2)[2] * 3 + coef(m2)[3] * 3^2
```

```
(Intercept)
  0.6754064
```

Our estimate of $\mathbb{E}[Y_i|num\_projects_i = 4]$ is:

```
coef(m2)[1] + coef(m2)[2] * 4 + coef(m2)[3] * 4^2
```

```
(Intercept)
  0.7211757
```

Because $0.7211757 > 0.6754064$, this shows that 4 is the optimal number of projects.

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.

Answer: The alternative hypothesis is heteroskedasticity. The p-value from the test is the p-value of the F-test from the auxiliary regression:

```
df$e2 <- m2$residuals^2
aux <- lm(e2 ~ num_projects + I(num_projects^2), data = df)
summary(aux)
```

```
Call:
lm(formula = e2 ~ num_projects + I(num_projects^2), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06150 -0.03527 -0.01841  0.00953  0.39284

Coefficients:
                    Estimate Std. Error t value           Pr(>|t|)
(Intercept)       -0.0064134  0.0043117  -1.487              0.1369
num_projects       0.0173018  0.0022468   7.701 0.0000000000000144 ***
I(num_projects^2) -0.0009965  0.0002756  -3.616              0.0003 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06046 on 14996 degrees of freedom
Multiple R-squared:  0.03557,   Adjusted R-squared:  0.03545
F-statistic: 276.6 on 2 and 14996 DF,  p-value: < 2.2e-16
```

The p-value is 0.000.

Because the p-value is less than the significance level of 0.05, the conclusion is "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.

```r
m3 <- lm(y ~ salary, data = df)
summary(m3)
```

```
Call:
lm(formula = y ~ salary, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54747 -0.17182  0.02925  0.19925  0.39925

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)   0.637470   0.007061  90.284    < 2e-16 ***
salarylow    -0.036717   0.007634  -4.809 0.00000153 ***
salarymedium -0.015653   0.007709  -2.031     0.0423 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2483 on 14996 degrees of freedom
Multiple R-squared:  0.002522,  Adjusted R-squared:  0.002389
F-statistic: 18.96 on 2 and 14996 DF,  p-value: 0.000000005967
```

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 $\bar{y}_{low} - \bar{y}_{medium}$).

Answer: High-salaried workers are the base category. Therefore we just need to report the intercept:

```
coef(m3)[1]
```

```
(Intercept)
  0.6374697
```

The difference between low- and high-salaried workers is measured by the coefficient on `"low"`. The reason for this is because:

$$\mathbb{E}[Y_i|salary = low] - \mathbb{E}[Y_i|salary = high] = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

```
coef(m3)[2]
```

```
  salarylow
-0.03671654
```

The difference between low- and medium-salaried workers is measured by the difference in coefficients between `"low"` and `"medium"`. The reason for this is because:

$$\mathbb{E}[Y_i|salary = low] - \mathbb{E}[Y_i|salary = medium] = (\beta_0 + \beta_1) - (\beta_0 + \beta_2) = \beta_1 - \beta_2$$

This is:

```
coef(m3)[2]-coef(m3)[3]
```

```
  salarylow
-0.02106349
```

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*.

Answer:

- H_0: All of $\beta_j = 0$ for $j = 1$ to 2.
- H_1: At least one of $\beta_j \neq 0$ for the same j.
- a is replacing k in the formula. This is 2 in this case.
- The value of the test statistic can be read from the `summary()` output (18.96) or obtained directly with:

```
summary(m3)$fstatistic
```

```
   value      numdf      dendf 
18.9609     2.0000 14996.0000 
```

- The critical value is:

```
qf(0.95, 2, 14996)
```

```
[1] 2.996331
```

where the numerator and denominator degrees of freedom (2 and 14996) can be read from the last line of the `summary()` output.

- Conclusion: *Reject* H_0. The model is *useful*.

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.

```
m4 <- lm(y ~ promotion + salary, data = df)
summary(m4)
```

```
Call:
lm(formula = y ~ promotion + salary, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54537 -0.17080  0.02957  0.19957  0.39957 
```

```
Coefficients:
              Estimate Std. Error t value   Pr(>|t|)
(Intercept)   0.635366   0.007107  89.398    < 2e-16 ***
promotion     0.036145   0.014122   2.560     0.0105 *
salarylow    -0.034939   0.007664  -4.559 0.00000519 ***
salarymedium -0.014564   0.007719  -1.887     0.0592 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2483 on 14995 degrees of freedom
Multiple R-squared:  0.002958,  Adjusted R-squared:  0.002759
F-statistic: 14.83 on 3 and 14995 DF,  p-value: 0.000000001227
```

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.

Answer:

The intercept gives an estimate of the mean value of the dependent variable when all independent variables are exactly equal to zero. When `promotion` is zero and `salary` is `"high"`, then the model is:

$$\mathbb{E}[Y_i|promotion_i = 0, salary_i = high] = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \beta_3 \times 0 = \beta_0$$

Therefore the answer is: "High-salary workers that did not receive a promotion in the last 5 years."

Question 12

Which variables are individually significant at the 1% level?

Answer:

All the variables with at least 2 `*`s in the summary output are significant at the 1% level. This means only `salarylow` is significant at the 1% level. The variable `promotion` has only 1 star and is only significant at the 5% level, not

the 1% level. The variable `salarymedium` is only significant at the 10% level (which is indicated by the `.`).

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.

Answer:

We can perform the partial F test by using the `anova()` function with the reduced model (`promotion` only) and the complete model (`promotion` and `salary`). The reduced model is actually model 1 from earlier, so there is no need to estimate it again.

```
anova(m1, m4)
```

```
Analysis of Variance Table

Model 1: y ~ promotion
Model 2: y ~ promotion + salary
  Res.Df    RSS Df Sum of Sq      F        Pr(>F)
1  14997 926.53
2  14995 924.39  2    2.1346 17.314 0.00000003087 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now have everything to answer the questions:

- H_0: All $\beta_j = 0$ for j from 2 to 3.
- H_1: At least one $\beta_j \neq 0$ for the same j.
- a is $k - g$, the number of variables in the complete model minus the number of variables in the reduced model (or the number of variables we are testing). This is $3 - 1 = 2$.
- Value of the test statistic: 17.314
- Critical value with $k-g = 2$ numerator and $n-k-1 = 14995$ denominator degrees of freedom:

```
qf(0.95, 2, 14995)
```

```
[1] 2.996331
```

- Conclusion: Reject H0. The variables are useful additions to the model (*reason:* the test statistic 17.314 is larger than the critical value of 2.996).