
Quiz 2 - DSC 152, Spring 2026

Full Name:

PID:

Quiz Time: 3pm 4pm

Instructions:

- This quiz consists of **4** questions, each worth 20 points (for a total of 80 points). You have **50 minutes** to complete it.
 - Please write **clearly** in the provided answer boxes; we will not grade work that appears elsewhere. Completely fill in bubbles and square boxes; if we cannot tell which option(s) you selected, you may lose points.
 - A bubble means that you should only **select one choice**.
 - A square box means you should **select all that apply**.
 - Show all work and R code where requested. Partial credit may be awarded.
 - You may use one double-sided handwritten sheet of notes. No calculators, no computers.
 - Assume we have already run all necessary `library()` calls in R.
-

By signing below, you are agreeing that you will behave honestly and fairly during and after this quiz.

Signature:

Version A

Please do not open your quiz until instructed to do so.

A public health organization is interested in understanding factors that influence recovery outcomes for patients undergoing treatment for tuberculosis (TB), the world's deadliest infectious disease. The organization collects observational data from a sample of patients with active TB enrolled in treatment programs across several clinics. Each row in the dataset represents a single patient's treatment record.

In addition to standard care, some patients are enrolled in a new adherence support program designed to improve medication compliance. The dataset is intended to evaluate how adherence and other patient characteristics relate to recovery time.

The resulting dataset, called `tb_data`, contains the following columns:

- `id` (double): unique identifier for each patient
- `adherence_rate` (double): percentage of prescribed doses taken during treatment (0–100)
- `time_to_recovery` (double): number of weeks until the patient is declared recovered
- `undernourishment` (character): insufficiency of patient's daily caloric intake for maintaining a normal, active, and healthy life ("yes", "no")
- `HIV_status` (character): HIV status of the patient ("positive", "negative")
- `intervention` (character): type of care received, whether standard treatment ("standard") or standard treatment plus the adherence program ("program")

Assume all categorical variables are already coded as factors in R with the following reference levels: `undernourishment = no`, `HIV_status = negative`, `intervention = standard`.

The first five rows of `tb_data` are shown below.

<code>id</code>	<code>adherence_rate</code>	<code>time_to_recovery</code>	<code>undernourishment</code>	<code>HIV_status</code>	<code>intervention</code>
1	80.3	28.6	yes	positive	program
2	68.3	24.7	no	negative	standard
3	74.7	19.5	yes	negative	program
4	84.7	17.0	no	negative	program
5	91.7	17.0	no	negative	program

Reminder: Write your **PID** on the top right of this page.

Question 1

Researchers wanted to see how adherence rates relate to recovery times. They fit the following simple linear regression in R:

```
model1 <- lm(time_to_recovery ~ adherence_rate, data = tb_data)
summary(model1)
```

The abridged output is as follows:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    38.1950     2.5255  15.123 < 2e-16 ***
adherence_rate -0.1892     0.0312  -6.063 6.65e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.42 on 198 degrees of freedom
Multiple R-squared:  0.1566,    Adjusted R-squared:  0.1523
F-statistic: 36.76 on 1 and 198 DF,  p-value: 6.646e-09
```

- a) (3 pts) Write a sentence giving the correct interpretation of the slope coefficient in context.

- b) (2 pts) What does the intercept represent in this model?
- The expected recovery time for a patient with 100% adherence
 - The expected recovery time for a patient with 0% adherence
 - The average recovery time in the dataset
 - The change in recovery time per 1% increase in adherence

c) (5 pts) The Residual Sum of Squares (RSS) is defined as:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

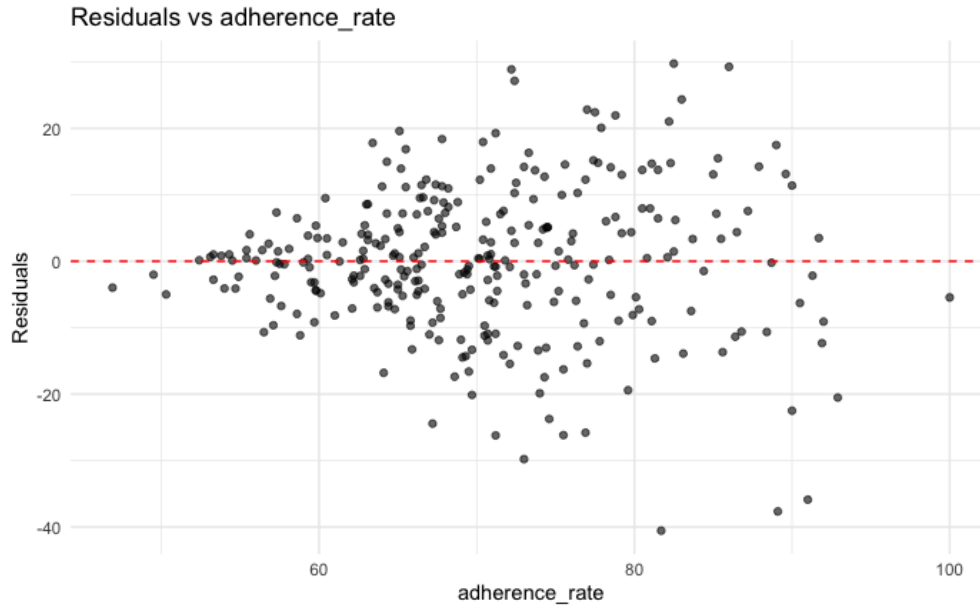
Which of the following are true? **Select all that apply**

- The RSS is used to find the least squares regression line, by finding the coefficient values that minimize it
- This quantity will always be larger when the relationship between x and y is non-linear compared to when it is linear
- This quantity only works for simple linear regression, and not multiple linear regression
- In simple linear regression, the definition of \hat{y}_i is the predicted value of y at the value of x_i
- If we add `HIV_status` to `model1` as a covariate and run a linear model, the RSS from this new model must be less than or equal to the RSS from the original `model1`

d) (6 pts) Write R code to produce a residual plot most suitable for checking the **linearity** condition as discussed in class. In doing so, you may extract required quantities from the `model1` object, or write R code to calculate them manually. Do not worry about adjusting axes labels, a title, or any other such elements for this question.

Reminder: Write your **PID** on the top right of this page.

- e) (4 pts) Researchers wanted to check the validity of using this linear model for statistical inference and create the following diagnostic plot. Are there any conditions that appear to be violated? **Select all that apply**



- Linearity
- Independence of observations
- Normality of ϵ_i
- Equal variance
- None of the above

Question 2

Researchers are studying tuberculosis treatment outcomes and are interested in whether adherence to prescribed medication is associated with faster recovery. They also suspect that **undernourishment** may influence recovery time, so they adjust for it in the analysis.

Researchers fit the following model:

```
mod_adj <- lm(time_to_recovery ~ adherence_rate + undernourishment,
              data = tb_data)
```

Assume `undernourishment = "no"` is the reference category. Their primary question of interest, then, is whether adherence to prescribed medication is associated with faster recovery, while adjusting for undernourishment.

- a) (4 pts) What must be true about **undernourishment** for it to be a confounder in the relationship between adherence rate and time to recovery?

- b) (3 pts) Suppose the estimated coefficient for `adherence_rate` in `mod_adj` is $\hat{\beta}_1 = -0.08$. Which of the following does this indicate, according to the linear model?
- Each 1% increase in adherence causes recovery time to decrease by 0.08 weeks for all patients.
 - Patients tend to recover 0.08 weeks slower per 1% increase in adherence, among patients with the same undernourishment status.
 - Patients tend to recover 0.8 weeks faster per 10% increase in adherence, among patients with the same undernourishment status.
 - Undernourished patients recover 0.08 weeks faster on average than well-nourished patients.
 - Adherence has no association with recovery time after adjusting for undernourishment.

Reminder: Write your **PID** on the top right of this page.

- c) (5 pts) Suppose the estimated coefficient for `undernourishmentyes` was $\hat{\beta}_2 = 1.5$. Then, the reference category is changed from "no" to "yes" and `mod_adj` is re-run. Which of the following would be true? **Select all that apply**
- $\hat{\beta}_2$ would be equal to -1.5
 - $\hat{\beta}_2$ would be equal to $\frac{1}{1.5}$
 - $\hat{\beta}_1$ could possibly change and be something different from -0.08
 - The p-value for $H_0 : \beta_2 = 0$ would be different in this new `mod_adj` vs. the original `mod_adj`
 - The RSS for this new `mod_adj` would be the same as the RSS for the original `mod_adj`
 - None of the above
- d) (4 pts) Which of the following are true about the overall Global F-test for the original `mod_adj`?
- It tests the $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$
 - It tests the $H_0 : \beta_1 = \beta_2 = 0$
 - It tests the $H_0 : \beta_1 = \beta_2$
 - It is an inappropriate test to do for the specific question of interest stated at the beginning of this question.
 - None of the above
- e) (4 pts) Suppose the p-value for `adherence_rate` in `mod_adj` is 0.003. State the null hypothesis being tested in terms of appropriate parameter(s). Also state the conclusion at $\alpha = 0.05$.

Question 3

Researchers want to investigate whether adherence rates differed between patients enrolled in the adherence support program and patients that were not. They fit the following model:

```
model2 <- lm(adherence_rate ~ intervention, data = tb_data)
summary(model2)
```

Assume `intervention = "standard"` is the reference category.

- a) (4 pts) State the null and alternative hypotheses for this question in terms of the appropriate parameter(s) from the linear model, and also in words.

- b) (5 pts) Write R code to perform a partial \mathcal{F} -test to test the hypotheses you wrote in part (a), by first fitting the null model corresponding to the null hypothesis and storing it as `null_model`, and then using `null_model` appropriately. Your output can be the entirety of the partial \mathcal{F} -test output; you do not need to isolate the p-value.

- c) (5 pts) Would an equivalent p-value to that of your partial \mathcal{F} -test in (b) appear anywhere in the `summary(model2)` output? If so, describe any and all locations of the equivalent p-value(s) in as much detail as possible. If not, explain why the partial \mathcal{F} -test was the only way to get the appropriate p-value here.

PID: _____

Reminder: Write your **PID** on the top right of this page.

- d) (6 pts) As the above approach in the earlier parts of this question utilized a linear model, it requires the **Equal variance** condition for validity. Write one line of R code to perform a test for the same hypotheses that you wrote in part (a) that does NOT require the **Equal variance** condition. If you cannot think of a way to do it in one line of R code, you may briefly state/describe a statistical test that would work, for partial credit.

Question 4

Researchers are now interested in whether tuberculosis recovery outcomes differ according to two patient characteristics: **HIV status** (“positive” or “negative”) and **undernourishment** (“yes” or “no”). They suspect that the relationship between HIV status and recovery time may differ between patients who are undernourished and those who are not.

The researchers fit the following model:

```
model3 <- lm(time_to_recovery ~ HIV_status + undernourishment +
             HIV_status*undernourishment,
             data = tb_data)
summary(model3)
```

Assume that R uses "negative" as the reference level for HIV_status and "no" as the reference level for undernourishment.

a) (4 pts) Which of the following best describes why an interaction term is included in this model? **Select all that apply**

- To account for baseline differences in recovery time between HIV status groups.
- To test whether the effect of HIV status on recovery time differs by undernourishment status.
- To control for confounding between HIV status and undernourishment.
- To test whether both predictors have additive effects on recovery time.
- None of the above

b) (3 pts) Complete the code below to calculate the mean recovery time for each combination of HIV status and undernourishment.

```
tb_data %>%
  ----- %>%
  summarize(mean_recovery = mean(time_to_recovery))
```

c) (3 pts) The overall Global \mathcal{F} -test p-value reported by `summary(model3)` corresponds to which null hypothesis?

- $H_0 : \beta_3 = 0$
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$
- $H_0 : \mu_{\text{yes}} = \mu_{\text{no}}$

Reminder: Write your **PID** on the top right of this page.

- d) (5 pts) Suppose researchers are primarily interested in determining whether HIV status has **any** relationship with recovery time, accounting for both the interaction term and undernourishment. If the appropriate p-value is in the `summary(model3)` output, identify as specifically as possible where you would find it. If not, write R code to obtain the appropriate p-value (in this case, any output that contains the appropriate p-value would be acceptable; that is, it does not need to output **ONLY** the p-value).

- e) (5 pts) When would it be appropriate to utilize the p-value for the `HIV_statuspositive` coefficient of `model3`? **Select all that apply**
- Any time we want to know something about the relationship between HIV status and time to recovery.
 - When we want to test whether HIV status is associated with time to recovery among those that are undernourished.
 - When we want to test whether HIV status is associated with time to recovery among those that are not undernourished.
 - When we want to test whether undernourishment status is associated with time to recovery, among those that are HIV positive.
 - When we want to test whether undernourishment status is associated with time to recovery, among those that are HIV negative.
 - There is no situation in which it would be correct to utilize this p-value.