

---

## Practice Quiz 2 - DSC 152, Spring 2026

---

Full Name:

SOLUTIONS

PID:

Quiz Time:  3PM  4PM

### Instructions:

- This quiz consists of 5 questions. You have a total of 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**.
- You may use one handwritten sheet of notes. No calculators and no computers.
- Assume we have already run all necessary `library()` calls in R.

---

### Additional Practice Quiz notes:

- This practice quiz is meant to reflect the style, difficulty, and possible content that may appear on your actual quiz.
- However, please note that it is NOT meant to reflect comprehensive coverage of all concepts that may appear on your actual quiz (as there is no way to put all of that on one quiz). The content for Quiz 2 is everything that is in Lectures 6 through 11, Labs 4-6 (only the first two parts of Lab 6), and HW2.
- While an answer key will eventually be provided, it is recommended that you do not simply read through the key. It will be much better preparation if you: (1) Actually do the practice quiz as if it were the real thing; (2) Check your answers with the key, but make sure that you actually understand WHY each answer is true.

You are analyzing a dataset of entry-level data science job listings. The `ds_jobs` data frame (first several rows shown below) contains one row per job listing, with the following variables: `"title"` (job title, string), `"emp_n"` (number of employees at company, int), `"sal"` (starting salary in USD, int), `"city"` (city and state, string), `"avg_age"` (average age of employees in years, numeric), `"loc_type"` (urban/suburban/rural, string), `"modality"` (in-person/remote/hybrid, string), and `"rent_1br"` (average monthly rent for a 1-bedroom apartment in that city, int).

title	emp_n	sal	city	avg_age	loc_type	modality	rent_1br
Data Analyst	320	72000	Austin, TX	29	urban	hybrid	1450
ML Engineer	4800	118000	San Francisco, CA	31	urban	remote	3100
Data Scientist	150	95000	Raleigh, NC	28	suburban	in-person	1280
Analytics Eng	9200	105000	Seattle, WA	33	urban	hybrid	2150
Data Analyst	75	61000	Columbus, OH	27	suburban	in-person	980
Data Scientist	2100	99000	Chicago, IL	30	urban	remote	1850

## Question 1

You fit the following simple linear regression in R:

```
model1 <- lm(sal ~ rent_1br, data = ds_jobs)
summary(model1)
```

The output is as follows:

```
Call:
lm(formula = sal ~ rent_1br, data = ds_jobs)

Residuals:
    Min       1Q   Median       3Q      Max
-31842  -9204   -814    8531   44283

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 42000.00    5831.24   7.203 3.1e-11
rent_1br      28.50       5.12    5.566 1.0e-04

Residual standard error: 14210 on 118 degrees of freedom
Multiple R-squared:  0.2079, Adjusted R-squared:  0.2011
F-statistic: 30.98 on 1 and 118 DF,  p-value: 1e-04
```

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

**Solution:** Every \$1 increase in average monthly rent for a 1-bedroom apartment is associated with an average difference in starting salary of \$28.50.

- b) A classmate concludes: “There is statistically significant evidence that living in a more expensive city causes higher starting salaries.” Explain why this statement is either correct or incorrect.

**Solution:** This statement is **incorrect**. This is an observational dataset—we did not randomly assign jobs to cities. A significant p-value indicates a statistically significant association, not causation. Confounders such as `loc_type`, industry type or other variables that we don’t even have data on may be driving both rent prices and salaries.

- c) Write R code to (i) extract the residuals from `model1` and store them as a new column `resid` in `ds_jobs`, and (ii) produce a residual plot suitable for checking the **Equal variance** condition.

**Solution:**

```
ds_jobs$resid <- model1$residuals  
ggplot(data = ds_jobs, aes(x = rent_1br, y = resid)) + geom_point()
```

## Question 2

You suspect the positive association between `emp_n` and `sal` may be confounded by `loc_type`.

- a) What must be true about `loc_type` for it to be a confounder in the relationship between `emp_n` and `sal`?

**Solution:** `loc_type` must be associated with **both** the primary predictor (`emp_n`) and the outcome (`sal`). That is, location type must be related to company size *and* to starting salary.

- b) Write R code to fit (1) a simple linear regression of `sal` on `emp_n` only, stored as `mod_simple`, and (2) a multiple linear regression adjusting for `loc_type`, stored as `mod_adj`. Then display a clean coefficient table for `mod_adj` using `kable`.

**Solution:**

```
mod_simple <- lm(sal ~ emp_n, data = ds_jobs)
mod_adj <- lm(sal ~ emp_n + loc_type, data = ds_jobs)
kable(summary(mod_adj)$coefficients)
```

- c) The output of `mod_adj` includes a coefficient for `loc_typesuburban`. What is the reference category, and how does R choose it? Write a sentence interpreting this coefficient, assuming its estimated value is  $-8,200$ .

**Solution:** The reference category is **rural**, chosen because it comes first alphabetically among the levels of `loc_type`. The coefficient  $-8,200$  means that suburban jobs are associated with a \$8,200 lower average starting salary than rural jobs, among companies with the same number of employees.

### Question 3

You want to test whether `modality` (in-person / hybrid / remote) is associated with `sal`, adjusting for `rent_1br`.

- a) Write R code to find the p-value. In this case, it is ok if your output contains more than just the p-value, as long as the appropriate p-value is in it somewhere.

**Solution:**

```
null_mod <- lm(sal ~ rent_1br, data = ds_jobs)
full_mod <- lm(sal ~ modality + rent_1br, data = ds_jobs)
anova(null_mod, full_mod)
```

- b) Explain in words why the test statistic in this situation will be *large* when  $H_0$  is false and *small* when  $H_0$  is true.

**Solution:** The partial  $\mathcal{F}$ -statistic is driven by the reduction in residual sum of squares when `modality` is added to the model. When  $H_0$  is false, `modality` genuinely explains variation in `sal`, so  $RSS_{\text{full}} \ll RSS_{\text{null}}$  and the numerator is large. When  $H_0$  is true, `modality` carries no signal, so  $RSS_{\text{full}} \approx RSS_{\text{null}}$  and the statistic is small.

- c) Running the code above gives  $p = 0.041$ . Using  $\alpha = 0.05$ , which conclusion is correct? Select one.
- Reject  $H_0$ ; there is statistically significant evidence that at least one modality group differs in mean salary, adjusting for `rent_1br`.
  - Fail to reject  $H_0$ ; the p-value is too close to 0.05 to draw any conclusion.
  - Reject  $H_0$ ; we conclude that modality causes salary differences.
  - The test is invalid; a  $t$ -test on each modality dummy coefficient should be used instead.

### Question 4

A researcher restricts `ds_jobs` to in-person and remote listings only (stored as `ds_jobs_2`) and wants to model the interaction between `modality` and whether the job is located in a high-rent city.

- a) Write one line of R code to create a new column in `ds_jobs_2` called `hi_rent` that is TRUE if `rent_1br` is greater than \$2,000 and FALSE otherwise.

**Solution:** There are many ways to do this and anything that works would be accepted, but here is one:

```
ds_jobs_2$hi_rent <- ds_jobs_2$rent_1br > 2000
```

- b) Using `hi_rent` instead of `rent_1br`, the researcher fits the following interaction model and obtains the output below:

```
mod_int <- lm(sal ~ hi_rent * modality, data = ds_jobs_2)
summary(mod_int)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	74000	3800	19.47	<0.001
hi_rentTRUE	21000	4600	4.57	<0.001
modalityremote	6500	3100	2.10	0.038
hi_rentTRUE:modalityremote	9800	4200	2.33	0.021

Write out the fitted regression equation for **remote** jobs only, as a function of `hi_rent`. Simplify as much as possible.

**Solution:**  $\widehat{\text{sal}} = 80500 + 30800 \cdot \text{hi\_rentTRUE}$

- c) In one sentence, interpret the **intercept** (74,000) in context.

**Solution:** The predicted starting salary for an **in-person** job in a **low-rent** city (`hi_rent = FALSE`) is \$74,000.

- d) In one sentence, interpret the coefficient 21,000 for `hi_rentTRUE` in context.

**Solution:** Among **in-person** jobs, those located in high-rent cities are associated with a \$21,000 higher predicted starting salary than those in low-rent cities, on average.

- e) In one sentence, interpret the coefficient 6,500 for `modalityremote` in context.

**Solution:** Among jobs in **low-rent** cities, remote positions are associated with a \$6,500 higher predicted starting salary than in-person positions, on average.

- f) Interpret the coefficient 9,800 for `hi_rentTRUE:modalityremote` in one or two sentences.

**Solution:** The salary advantage of being in a high-rent city is \$9,800 larger for remote jobs than for in-person jobs. Equivalently, the difference in predicted salary between high- and low-rent cities is \$21,000 for in-person jobs but \$30,800 for remote jobs.

- g) Suppose that your primary interest is actually in the impact of modality. A colleague suggests that you should just look at the p-value of 0.038. Explain to your colleague why you do not want to do this.

**Solution:** The p-value of 0.038 is for the impact of modality ONLY in low rent locations, and we probably care about the impact of modality in ALL locations.

- h) Write the R code for the correct statistical test if modality is our primary interest, and state explicitly what  $H_0$  is in terms of the model coefficients.

**Solution:**

```
null_mod2 <- lm(sal ~ hi_rent, data = ds_jobs_2)
full_mod2 <- lm(sal ~ hi_rent * modality, data = ds_jobs_2)
anova(null_mod2, full_mod2)
```

$H_0: \beta_{\text{remote}} = \beta_{\text{hi\_rent:remote}} = 0$  (simultaneously). The null model retains `hi_rent` but removes both the main effect of `modality` and its interaction.

## Question 5

A recruiter wants to evaluate two interview formats: Format A (standard technical screen) and Format B (take-home project). She randomly assigns 80 applicants to one of the two formats and records each applicant's 6-month performance rating after being hired.

- a) A colleague suggests letting applicants choose their preferred format instead of randomizing. Explain specifically why this would be a problem for the analysis.

**Solution:** If applicants self-select their format, variables such as self-confidence or prior experience could be associated with both the format choice and the performance rating. This makes it impossible to determine whether any observed difference in ratings is due to the interview format itself or to pre-existing differences between the groups.

- b) The collected data are stored in a data frame called `interview_df`, which has a column `score` (the 6-month performance rating, numeric) and a column `format` (a character variable with values "A" and "B"). Write the null and alternative hypotheses for this A/B test in terms of  $\mu_A$  and  $\mu_B$ , the true mean performance scores for each format. Then write **one line of R code** that runs the appropriate parametric two-sample test.

**Solution:** First, the question was intended to state that the recruiter is looking for any difference between the impact of the two interview formats. As such, this is a two-sided test:

$$H_0: \mu_A = \mu_B \quad H_A: \mu_A \neq \mu_B$$

```
t.test(score ~ format, data = interview_df)
```

- c) Which of the following correctly describe the permutation test as an alternative to the two-sample  $t$ -test? **Select all that apply.**

It does not require the data to follow any particular distribution.

It generates a null distribution by repeatedly shuffling the group labels and recomputing the test statistic.

It always has higher statistical power than the two-sample  $t$ -test.

It requires equal sample sizes in both groups.

- d) Write R code that uses the `calc_stat` function defined below, to complete the permutation test started below on these data for the same hypotheses stated in part (b), and calculate the appropriate p-value.

```
calc_stat <- function(df){  
  group_means <- (df %>% group_by(group) %>%  
    summarize(mean_diff=mean(score_diff)))$mean_diff
```

```
    return(abs(group_means[1] - group_means[2]))
  }

  reps <- 10000
  stat <- NA

  for(i in 1:reps){
    ...
```

**Solution:** Anything that works would be counted for full credit. Here is how we did it in class:

```
perm_df <- interview_df

perm_df$score_diff <- sample(perm_df$score)

stat[i] <- calc_stat(perm_df)
}
interview_df$score_diff <- interview_df$score
sum(stat >= calc_stat(interview_df) ) / reps
```

Also note that there was an unintentional inconsistency in variable naming: the question stated that the performance rating is scored in the column called `score`, but the `calc_stat` function referred to a column called `score_diff`. As such, the solution code made this switch so that the `calc_stat` function would run (in the 2nd line and 2nd-to-last line above), but it was not intended that you would have to do that. If we made a similar mistake on the writing of a question on the actual quiz, we would take any answers that just assumed that the column was called `score`, `score_diff`, or flipped between any combination of the two.

### Reminders:

- **Write your PID** on the front page and on the top right corner of each subsequent page.
- Fill in bubbles and square boxes **completely and darkly**; partially filled marks will not be graded.
- Show all written work and code clearly inside the response boxes; work outside boxes will not be graded.
- Stay in your seat until the quiz is over. No leaving early.