Quiz 3 - DSC 152, Spring 2026

Full Name:	SOLUTIONS
PID:	

Quiz Time: ◯ 3pm ◯ 4pm

Instructions:

- This quiz consists of 5 questions worth a total of 70 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 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.

PID: _____

You are working with a nonprofit internship-prep program that offers optional one-on-one coaching to students applying for summer internships. The main dataset, `career_df`, contains one row per student and includes:

- `practice_hours` (double): number of hours spent on mock-interview practice
- `portfolio_score` (double): portfolio review score (0–100)
- `coaching` (character): whether the student used one-on-one coaching (`"yes"` or `"no"`)
- `stipend` (double): internship stipend (in $1,000s), for students who received offers
- `offer` (double): 1 if the student received an internship offer, 0 otherwise

Throughout this quiz, all regression models are fit in R.

A small sample of the `career_df` dataframe is shown below:

practice_hours	portfolio_score	coaching	stipend	offer
12	68	no	24.0	0
18	81	yes	29.5	1
10	74	no	23.0	0
25	88	yes	33.0	1
16	79	yes	27.5	1

PID: _____

Question 1

To understand how coaching may change the relationship between interview practice and stipend, a researcher fits the interaction model

```
model1 <- lm(stipend ~ practice_hours * coaching, data = career_df)
```

Assume `coaching = "no"` is the reference category. The estimated coefficients are:

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	18.00	2.10	8.57	< 0.001
`practice_hours`	0.20	0.05	4.00	< 0.001
`coachingyes`	-1.50	2.40	-0.63	0.62
`practice_hours:coachingyes`	0.10	0.04	2.50	0.014

a) (4 pts) Write the estimated regression equation specifically for students who *did* use coaching. Simplify as much as possible.

Solution: For coached students, `coachingyes` $= 1$, so

$\widehat{\texttt{stipend}} = 18.00 - 1.50 + (0.20 + 0.10)\texttt{practice\_hours} = 16.50 + 0.30\texttt{practice\_hours}.$

b) (3 pts) Which interpretation of the interaction coefficient 0.10 is most appropriate in context?

● For coached students, each additional practice hour is associated with an additional 0.10 thousand dollars in predicted stipend compared to uncoached students.

◯ For coached students, the estimated slope relating practice hours to stipend is 0.10 thousand dollars per hour.

◯ At a fixed number of practice hours, coached students are predicted to receive 0.10 thousand dollars more than uncoached students.

◯ The interaction coefficient means that coached and uncoached students have the same predicted stipend when `practice_hours = 0`.

◯ None of the above.

c) (6 pts) Suppose the researcher wants to test whether `coaching` contributes to the model *at all*, including its interaction with `practice_hours`.

- If the appropriate p-value for this test is available in the output above, state what it is and your conclusion at $\alpha = 0.05$.
- If it is not, write R code to carry out the required test.

Solution: The appropriate p-value is not available from the coefficient table above, since this is a joint test of the coaching main effect and the interaction term. Use a reduced model without any coaching terms:

```
reduced_model <- lm(stipend ~ practice_hours, data = career_df)
anova(reduced_model, model1)
```

d) (3 pts) Which statement best explains why the individual p-value for `coachingyes` is not enough, by itself, to answer the model-comparison question above?

◯ Because the individual t-test for `coachingyes` comes from the model that was run with an interaction term, this indicates the extent to which `coaching` impacts the `stipend`, while accounting for the interaction with `practice_hours` as required.

● The individual t-test for `coachingyes` only checks the coaching effect at `practice_hours = 0`; it does not jointly test the coaching main effect and interaction.

◯ The individual t-test for `coachingyes` tests the same null hypothesis as the model comparison, but reports it using a different test statistic.

◯ The individual t-test for `coachingyes` checks whether the slope for `practice_hours` is zero among students who did not use coaching.

◯ The individual t-test for `coachingyes` compares the residual standard error from the full model to the residual standard error from the reduced model.

◯ None of the above.

PID: _____

Question 2

Now, suppose that researchers are primarily interested in whether `practice_hours` is associated with `stipend`. They are considering `portfolio_score` as a secondary covariate.

a) (5 pts) A student proposes the following workflow:

> (1) Fit `lm(stipend ~ practice_hours + portfolio_score, data = career_df)`.
> (2) If the p-value for `portfolio_score` is above 0.05, drop it and then report the p-value for `practice_hours` from the smaller model.
> (3) If the p-value for `portfolio_score` is below 0.05, report the p-value for `practice_hours` from the model in (1) above.

Which of the following statements are true regarding this as a workflow for *statistical inference* about `practice_hours`? **Select all that apply.**

■ This is an incorrect workflow because the p-value for `practice_hours` from the smaller model is influenced by model-selection step based on the same data.

☐ This is an incorrect workflow because inference about `practice_hours` should always be based on the largest model available.

■ The decision of whether to include `practice_hours` in the model should not be based on this workflow, but rather on whether there is scientific justification for it as a confounding variable.

☐ This is a fine workflow because `portfolio_score` must be statistically significant before the coefficient on `practice_hours` can be interpreted.

☐ This is an incorrect workflow because dropping `portfolio_score` changes the definition of the outcome variable being modeled.

☐ None of the above.

b) (4 pts) After centering and scaling `practice_hours` and `portfolio_score`, a researcher refits the model in (1) above. Briefly state two benefits of centering/scaling for coefficient interpretation. Be specific about the units of a scaled predictor.

Solution: Centering makes the intercept correspond to a student with average predictor values, rather than a student with zero practice hours and zero portfolio score. Scaling also makes coefficients interpretable in standard-deviation units: a one-unit increase in a scaled predictor means a one-standard-deviation increase in the original predictor.

c) (4 pts) Which of the following are true about centering and scaling covariates? **Select all that apply.**

☐ Centering a covariate will not change the p-value for whether its coefficient is equal to 0, but scaling it will.

■ Neither centering a covariate nor scaling it will change the p-value for whether its coefficient is equal to 0.

☐ Centering a covariate or scaling it will both change the p-value for whether its coefficient is equal to 0.

■ Centering and scaling covariates will not change the R^2 value for the model.

☐ None of the above.

d) (4 pts) Now a researcher is comparing the two models:

```
model_raw  <- lm(stipend ~ practice_hours, data = career_df)
model_logy <- lm(log(stipend) ~ practice_hours, data = career_df)
```

Suppose the residual plot for `model_raw` shows clear curvature and increasing spread, while the residual plot for `model_logy` looks substantially more random and stable. Which of the following statements are most accurate? **Select all that apply.**

☐ This suggests that the researcher should use `model_logy` for statistical inference in the primary analysis on these data.

☐ `model_raw` should be the chosen model for primary statistical inference even if its residual plot raises concerns, because interpretation of regression coefficients on the raw scale is easier than on the log scale.

■ As a sensitivity analysis, these residual plots would suggest that future studies should consider using `model_logy` for statistical inference in the primary analysis.

■ Scientific rationale should also guide model choice; the residual plot should not be the only consideration.

☐ None of the above.

e) (3 pts) Which of the following statements is most accurate?

◯ The residual plots in the previous question do not actually matter; the best transformation is whichever one gives the largest R^2.

◯ Once a predictor is statistically significant, model selection (including whether or not to do a log transformation) no longer affects inference validity.

◯ Instead of a log transformation, centering and scaling can make coefficients easier to interpret, and they are a substitute for checking model conditions.

● None of the above.

PID: _____

Question 3

In a cross-sectional sample of students, the internship-prep program also records whether each student used coaching and whether they ultimately received an internship offer. The resulting table is:

	offer		
coaching	yes	no	total
yes	72	48	120
no	60	120	180
total	132	168	300

a) (3 pts) Compute the relative risk of receiving an offer for coached students compared to uncoached students. Your final answer may be left as a fraction or given as a decimal. Interpret your answer in one sentence.

Solution:

$$RR = \frac{72/120}{60/180} = \frac{3/5}{1/3} = \frac{9}{5} = 1.8.$$

Coached students are estimated to be 1.8 times as likely to receive an offer as uncoached students.

b) (3 pts) Compute the odds ratio for these data. Show your setup. Your final answer may be left as a fraction or given as a decimal. Interpret your answer in one sentence.

Solution:

$$OR = \frac{72/48}{60/120} = \frac{72 \cdot 120}{48 \cdot 60} = 3.$$

The odds of receiving an offer are estimated to be 3 times as large for coached students as for uncoached students.

c) (3 pts) With these data, is it expected that the OR and RR would be approximately equal to each other? Explain why or why not.

Solution: No. The OR is a good approximation of the RR when the outcome is rare. Here, the outcome (getting an offer) is not rare.

d) (3 pts) Suppose that, instead of a cross-sectional sample, a researcher had sampled 80 students with offers and 80 students without offers on purpose, then looked backward to see who used coaching. In that case, which measure from parts (a) and (b) would still be appropriate to interpret, and why? Explain your reasoning.

Solution: The odds ratio is still appropriate. In a case-control study, the numbers with and without offers are fixed by design, so the sample proportions cannot be used to estimate the risks of receiving an offer. The odds ratio can still be

interpreted as the association between coaching and offer status.

PID: _____

Question 4

Researchers fit a logistic regression model for the probability of receiving an internship offer. A cleaned-up excerpt of the coefficient output is:

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-5.20	1.30	-4.00	< 0.001
`coachingyes`	1.10	0.32	3.44	< 0.001
`portfolio_score`	0.07	0.015	4.67	< 0.001

a) (6 pts) Write one line of R code to run the model that would produce the output summarized above. Here, you are only required to write the model-fitting code, not the further code to extract/create this table.

Solution:
```
glm(offer ~ coaching + portfolio_score,
data = career_df,
family = binomial)
```

b) (4 pts) Provide an interpretation on the odds-ratio scale using the coefficient for `coachingyes`. Your answer may contain an unevaluated expression.

Solution: The odds ratio for coaching is $e^{1.10}$. Holding portfolio score fixed, coached students have $e^{1.10}$ times the odds of receiving an internship offer compared to uncoached students.

c) (4 pts) Consider two students with the same coaching status, but one has a portfolio score that is 10 points higher than the other. According to this model, by what multiplicative factor do their *odds* of receiving an offer differ? Show your work.

Solution: A 10-point increase in portfolio score changes the log-odds by $10(0.07) = 0.70$. Therefore, the odds are multiplied by $e^{0.70}$.

PID: _____

Question 5

In a separate analysis, the program tracks daily website traffic. Let `visits_t` denote the number of site visits on day t, and let `bookings_t` denote the number of coaching bookings on day t. For part (c), suppose the data are stored in a dataframe called `traffic_df`, with columns `visits`, `bookings`, and `day_of_week`. Assume `day_of_week` is a factor variable with Monday as the reference level.

A plot of log(`visits_t`) against day number shows a repeating weekly up-and-down pattern, but no obvious long-run upward trend.

a) (2 pts) Explain briefly why the repeating weekly pattern is a problem for ordinary linear-regression inference if time is ignored.

> **Solution:** If time is ignored, residuals may still contain a systematic weekly pattern. Errors from nearby days, or from the same day of the week, may not behave like independent random noise, so ordinary linear-regression standard errors, p-values, and confidence intervals may be misleading.

b) (3 pts) A researcher wants to adjust for the weekly up-and-down pattern without any AR(p) or MA(q) terms, but simply with day-of-week effects using Monday as the baseline day. Write R code to run a regression model for studying the relationship between `bookings` and `log(visits)` with adjustment for day of week.

> **Solution:** `lm(log(visits) ~ bookings + day_of_week, data = traffic_df)`

c) (2 pts) For the log website-traffic series, an AR(1) model has the form

$$y_t = c + \phi_1 y_{t-1} + \epsilon_t.$$

In words, what does it mean if ϕ_1 is close to 1?

> **Solution:** If ϕ_1 is close to 1, the series is highly persistent: today's value is strongly related to yesterday's value, and shocks or unusually high/low values decay slowly over time.