Practice Quiz 3 - DSC 152, Spring 2026

Full Name: SOLUTIONS

PID:

Section: ◯ A00 ◯ B00 ◯ C00

Instructions:

- This quiz consists of 5 questions. 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.

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 3 is everything that is in Lectures 12 through 17, Labs 7-9 (but not the time series *regression with covariates* (other than time) part of Lab 9), and HW3.
- While an answer key will eventually be provided (around Monday 6/1), 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 a data analyst at a consortium of research universities. Your team is studying factors associated with graduate school admissions outcomes in **Statistics**, **Computer Science**, and **Data Science** programs. The dataset `grad` contains one row per applicant and includes the following columns:

- `gre` (double): GRE Quantitative score (130–170)
- `gpa` (double): undergraduate GPA (0.0–4.0)
- `field` (character): `"Statistics"`, `"CS"`, or `"DataSci"`
- `pubs` (double): number of undergraduate research publications
- `stipend` (double): annual stipend offered (in $1,000s), for admitted students
- `admitted` (double): 1 if admitted, 0 if not

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

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

gre	gpa	pubs	stipend	field	admitted
165	3.90	2	38.5	CS	1
158	3.60	0	31.0	Statistics	0
170	3.95	3	42.0	DataSci	1
152	3.20	1	28.0	CS	0
161	3.75	1	35.5	Statistics	1

Question 1

A colleague fits the following interaction model to understand how GRE score and field of study jointly predict the annual stipend (in $1,000s) offered to admitted students:

$$\begin{aligned}\widehat{\texttt{stipend}} = {} & \hat{\beta}_0 + \hat{\beta}_1\,\texttt{gre} \\ & + \hat{\beta}_2\,\texttt{CS} + \hat{\beta}_3\,\texttt{DataSci} \\ & + \hat{\beta}_4\,(\texttt{gre} \times \texttt{CS}) + \hat{\beta}_5\,(\texttt{gre} \times \texttt{DataSci})\end{aligned}$$

where `CS` equals 1 for Computer Science applicants (0 otherwise), `DataSci` equals 1 for Data Science applicants (0 otherwise), and `Statistics` is the reference category. The partial `summary()` output is:

	Estimate	Std. Error	t value	Pr(> $|t|$)
(Intercept)	−8.20	3.10	−2.65	0.009
`gre`	0.46	0.04	11.50	<0.001
`fieldCS`	12.80	5.40	2.37	0.019
`fieldDataSci`	6.30	5.10	1.24	0.217
`gre:fieldCS`	0.09	0.05	1.80	0.074
`gre:fieldDataSci`	−0.03	0.05	−0.60	0.550

a) Write out the estimated regression equation *specifically* for a CS applicant (simplify as much as possible by combining like terms).

> **Solution:** For CS, both `fieldCS` $= 1$ and `fieldDataSci` $= 0$, so:
> $\widehat{\texttt{stipend}} = (-8.20 + 12.80) + (0.46 + 0.09)\,\texttt{gre} = 4.60 + 0.55\,\texttt{gre}$

b) Give a careful interpretation of $\hat{\beta}_4 = 0.09$ in context.

> **Solution:** For CS applicants, each additional point on the GRE is associated with an average stipend that is \$90 *higher* compared with the corresponding increase for Statistics applicants, holding all other things equal. In other words, 0.09 is the *difference in slopes* between the CS and Statistics regression lines.

c) Your colleague wants to test whether the effect of GRE score on stipend is the *same* across all three fields. Write the H_0 for the appropriate partial $\mathcal{F}$-test in terms of model parameters, and state which coefficients would be constrained to zero under H_0.

> **Solution:** $H_0\colon \beta_4 = \beta_5 = 0$. Under H_0 both interaction terms are removed, so the slope on `gre` is identical for all three fields (parallel lines). The null model is `lm(stipend ~ gre + field, data=grad)`.

d) Based solely on the table above, can you immediately conclude whether H_0 from part (c) is rejected at $\alpha = 0.05$? Select the best answer and briefly justify.

○ Yes, because both interaction p-values exceed 0.05, so we fail to reject H_0.

● No, because the partial $\mathcal{F}$-test requires comparing the full and null models jointly; the individual p-values are not sufficient.

○ Yes, because at least one interaction p-value is below 0.05, so we reject H_0.

○ No, because interaction terms can never be tested with a partial $\mathcal{F}$-test.

Question 2

A researcher investigates whether GRE score (x) predicts stipend (y) using a **double-log** model. That is, she fits:

$$\log(\widehat{\texttt{stipend}}) = \hat{\beta}_0 + \hat{\beta}_1 \log(\texttt{gre})$$

and obtains $\hat{\beta}_1 = 2.84$.

a) Derive the interpretation of $\hat{\beta}_1$ in a double-log model. Your derivation should start from the fitted equation written at two values of `gre`—namely x and cx for some constant $c > 0$—and show algebraically what a c-fold increase in `gre` implies for the *predicted* stipend. Show all steps.

> **Solution:** At `gre` $= x$: $\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 \log(x)$.
> At `gre` $= cx$: $\log(\hat{y}^*) = \hat{\beta}_0 + \hat{\beta}_1 \log(cx) = \hat{\beta}_0 + \hat{\beta}_1 \log(c) + \hat{\beta}_1 \log(x)$.
> Subtracting: $\log(\hat{y}^*) - \log(\hat{y}) = \hat{\beta}_1 \log(c)$, i.e. $\log\left(\frac{\hat{y}^*}{\hat{y}}\right) = \hat{\beta}_1 \log(c)$, so $\frac{\hat{y}^*}{\hat{y}} = c^{\hat{\beta}_1}$.
> **Interpretation:** A c-fold (i.e. $100(c-1)\%$) increase in GRE score is associated with a predicted stipend that is $c^{\hat{\beta}_1}$ times as large (i.e. multiplied by $c^{2.84}$). For the common special case of a 1% increase in GRE ($c = 1.01$), stipend increases by approximately $\hat{\beta}_1 \times 1\% = 2.84\%$.

b) Using $\hat{\beta}_1 = 2.84$, give a specific numerical interpretation for a **1% increase** in GRE score. (You may use the approximation $\ln(1.01) \approx 0.01$.)

> **Solution:** A 1% increase in GRE score is associated with an approximately 2.84% increase in predicted stipend, holding all else equal.

c) A student argues: "Using `log(gre)` instead of `gre` changes the fitted curve, so the model is no longer *linear* regression." Is the student correct? Select the best answer.

- ○ Yes, because the relationship between the original `gre` and `stipend` is curved.
- ● No. The model is still linear regression because it is linear in log(`gre`); we simply transformed the predictor before fitting.
- ○ No, but only because we also log-transformed the outcome.
- ○ Yes, because `log(gre)` is a nonlinear function.

PID: ______________________________

Reminders: Write your **PID** on the top right of this page. Show all work where requested.

Question 3

A team studies how GRE score and GPA jointly predict stipend in an **interaction model with centered and scaled covariates**. They define:

$$\texttt{gre\_s} = \frac{\texttt{gre} - \overline{\texttt{gre}}}{2\, s_{\texttt{gre}}}, \qquad \texttt{gpa\_c} = \texttt{gpa} - \overline{\texttt{gpa}}$$

and fit:

$$\widehat{\texttt{stipend}} = \hat{\gamma}_0 + \hat{\gamma}_1\, \texttt{gre\_s} + \hat{\gamma}_2\, \texttt{gpa\_c} + \hat{\gamma}_3\, (\texttt{gre\_s} \times \texttt{gpa\_c})$$

a) Derive the interpretation of $\hat{\gamma}_0$ (the intercept) in this centered and scaled model. Your answer should reference specific values of `gre` and `gpa`.

> **Solution:** When $\texttt{gre} = \overline{\texttt{gre}}$ and $\texttt{gpa} = \overline{\texttt{gpa}}$, we have $\texttt{gre\_s} = 0$ and $\texttt{gpa\_c} = 0$, so $\widehat{\texttt{stipend}} = \hat{\gamma}_0$. Thus, $\hat{\gamma}_0$ is the estimated stipend for an applicant with average GRE score and average GPA.

b) Derive the interpretation of $\hat{\gamma}_1$ in this model. Your derivation should start from the fitted equation evaluated at two values of `gpa_c` and show what $\hat{\gamma}_1$ measures when $\texttt{gpa\_c} = 0$. Show all steps, and state your final interpretation clearly.

> **Solution:** Write the model as a function of `gre_s` for a fixed value of `gpa_c`:
>
> $$\widehat{\texttt{stipend}} = (\hat{\gamma}_0 + \hat{\gamma}_2\, \texttt{gpa\_c}) + (\hat{\gamma}_1 + \hat{\gamma}_3\, \texttt{gpa\_c})\, \texttt{gre\_s}.$$
>
> At $\texttt{gpa\_c} = 0$ (i.e. $\texttt{gpa} = \overline{\texttt{gpa}}$), the slope on `gre_s` is exactly $\hat{\gamma}_1$. Since `gre_s` was divided by $2s_{\texttt{gre}}$, a one-unit increase in `gre_s` corresponds to an increase of $2s_{\texttt{gre}}$ in the original `gre` score. Therefore, $\hat{\gamma}_1$ is the estimated difference in stipend associated with a $2s_{\texttt{gre}}$-point increase in GRE score, *among applicants whose GPA equals the sample mean GPA.*
>
> Note: the question was supposed to say "at two values of `gre_s`," not "at two values of `gpa_c`." Apologies for any confusion.

c) Now derive the interpretation of $\hat{\gamma}_3$ (the interaction coefficient). Your derivation should compare the slope on `gre_s` at two different values of `gpa_c`. Show all steps.

> **Solution:** From the re-written model above, the slope on `gre_s` for a given `gpa_c` is $\hat{\gamma}_1 + \hat{\gamma}_3\, \texttt{gpa\_c}$.
>
> At $\texttt{gpa\_c} = v$: slope $= \hat{\gamma}_1 + \hat{\gamma}_3 v$.
>
> At $\texttt{gpa\_c} = v + 1$: slope $= \hat{\gamma}_1 + \hat{\gamma}_3 (v + 1)$.
>
> Difference $= \hat{\gamma}_3$.

Interpretation: $\hat{\gamma}_3$ is the estimated change in the slope of `gre_s` on stipend for each 1-point increase in `gpa_c` (equivalently, for each 1-point increase in GPA above the mean). In other words, it quantifies how much the relationship between GRE score and stipend changes as GPA increases by one point.

d) One benefit of centering and scaling is that it allows more direct comparison of coefficient magnitudes. Suppose $|\hat{\gamma}_1| \gg |\hat{\gamma}_2|$. What would this suggest about the relative importance of GRE score vs. GPA in predicting stipend?

Solution: With GRE scaled and GPA now ranging from approximately -2 to 2 by just centering it, both covariates have been placed on roughly comparable scales (a one-unit change in each scaled variable represents a similar-sized shift in the original predictor), $|\hat{\gamma}_1| \gg |\hat{\gamma}_2|$ would suggest that GRE score has a substantially larger association with stipend than GPA does, at average levels of the other covariate.

Question 4

Your team now models the probability of admission using **logistic regression**. The R code and partial output are shown below.

```
model_logit <- glm(admitted ~ gre + gpa,
                   data = grad, family = "binomial")
summary(model_logit)
```

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	−14.82	2.07	−7.16	<0.001
`gre`	0.08	0.01	8.00	<0.001
`gpa`	0.77	0.19	4.05	<0.001

a) Write the fitted logistic regression model as a function of $\hat{p}$, where $\hat{p}$ is the estimated probability of admission.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -14.82 + 0.08\,\texttt{gre} + 0.77\,\texttt{gpa}$$

b) Interpret the coefficient for `gre` on the **odds ratio** scale, with an approximation.

Solution: For any small value x, $e^x \approx 1 + x$. Thus, $e^{0.08} \approx 1.08$. This indicates that each additional point on the GRE score is associated with a multiplicative increase of approximately 1.08 in the estimated *odds* of admission (roughly an 8% increase in the odds of admission), holding GPA constant.

c) A 95% confidence interval for $\beta_{\texttt{gre}}$ on the log-odds scale is (0.06, 0.10). Give the corresponding 95% confidence interval for the odds ratio associated with a 1-point increase in GRE score.

Solution: Exponentiate both endpoints: $(e^{0.06}, e^{0.10}) \approx (1.06, 1.10)$. We are 95% confident that a 1-point increase in GRE score multiplies the odds of admission by a factor between approximately 1.06 and 1.10.

d) Three conditions are required for valid logistic regression inference. List all three, and for each one briefly state how it could be checked or justified in this graduate admissions context.

Solution:

1. **Binary outcome**: `admitted` takes only the values 0 and 1. Satisfied by construction.
2. **Independence of observations**: Each row is a different applicant. This is approximately satisfied if applicants are not clustered (e.g. not many pairs

of applicants from the same program sharing advisors). Could be verified by checking the data collection design.

3. **Linearity on the log-odds scale**: The log-odds of admission should be approximately linear in `gre` and `gpa`. Checked via a deviance residual plot (fitted values vs. deviance residuals), looking for a loess smoother that tracks approximately the horizontal line at zero.

e) An applicant has a GRE score of 160 and a GPA of 3.5. Using the fitted model, write an expression for their *estimated probability* of admission. Your answer should be completely unsimplified; any expression equaling the correct quantity will receive full credit.

Solution:

Log-odds $= -14.82 + 0.08(160) + 0.77(3.5)$

Odds $= e^{-14.82+0.08(160)+0.77(3.5)}$.

$$\hat{p} = \frac{e^{-14.82+0.08(160)+0.77(3.5)}}{1 + e^{-14.82+0.08(160)+0.77(3.5)}}$$

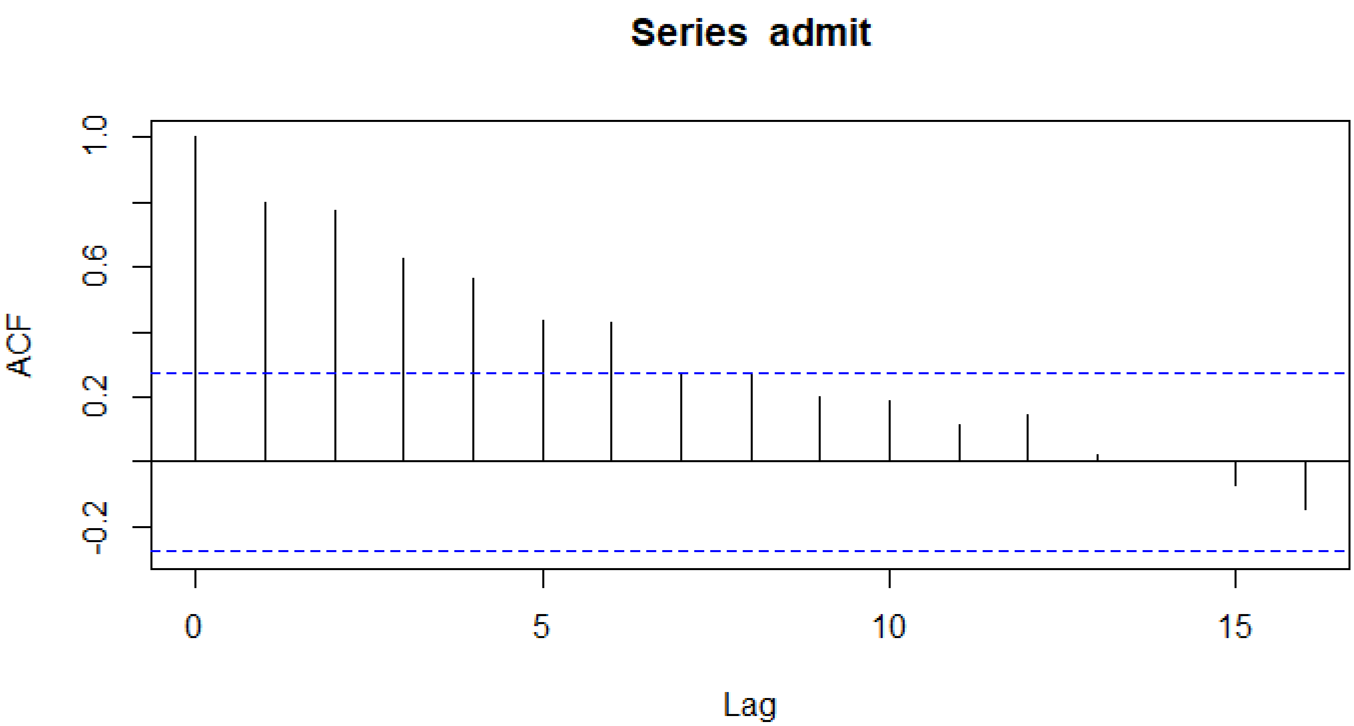
Question 5

Now, you are investigating the admissions rate of one CS program at one specific institution. Over the past 50 years, its admissions rate is stored in the variable `admit` below as a vector, and the first few values are shown:

```
> admit
[1] 0.23 0.24 0.31 0.24 0.12 0.20
```

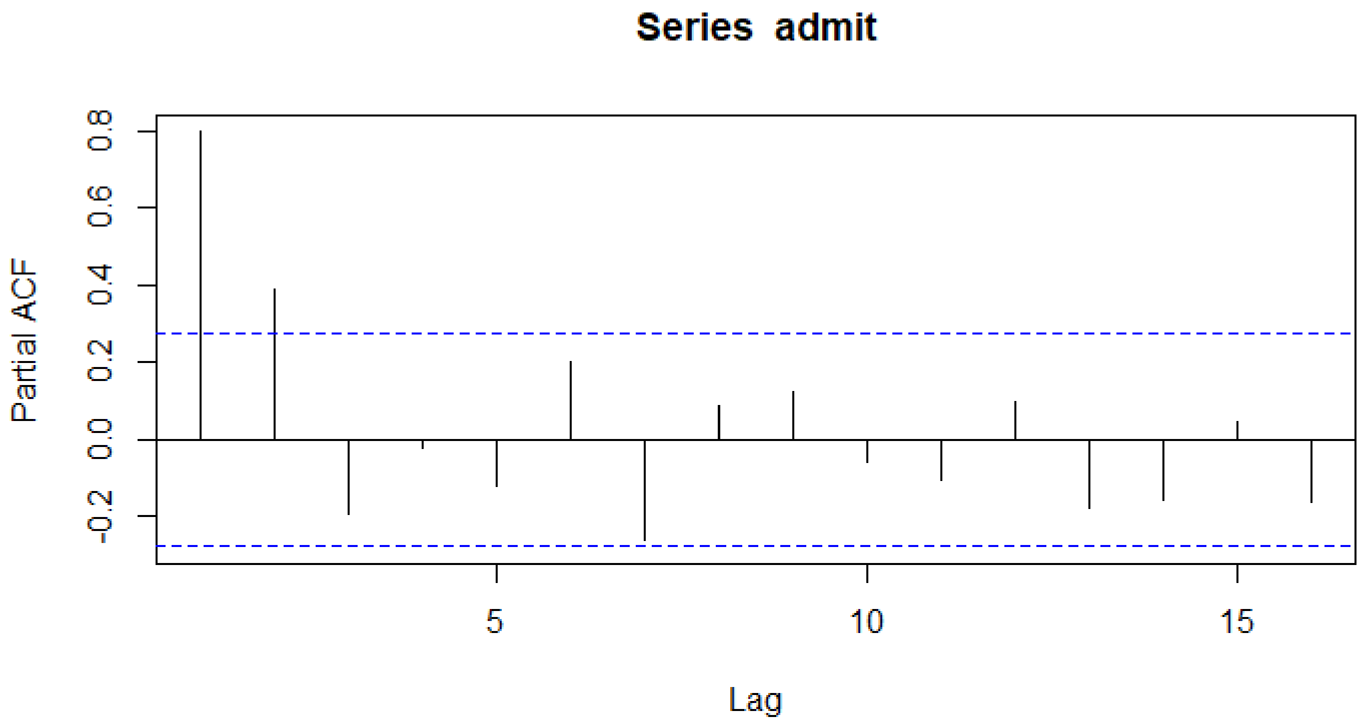
a) Write R code to produce the plot below.

Solution: `acf(admit)`



b) Write R code to produce the plot below.

> **Solution:** `pacf(admit)`



c) Explain why these two plots indicate that an AR(2) model may be appropriate for these data.

> **Solution:** The partial ACF plot shows spikes at the first two lags. Since the partial ACF factors out intermediate lags that are not directly causing any autocorrelation, the remaining two would indicate that these data might follow an AR(2) model.

d) The following output was obtained to determine whether there is a statistically significant decline in the admissions rate of this program, accounting for an AR(2) model:

```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1       0.4177405  0.1253017  3.3339 0.0008564 ***
ar2       0.4100506  0.1262192  3.2487 0.0011593 **
intercept 0.2421721  0.0397032  6.0996 1.064e-09 ***
xreg      0.0022312  0.0012686  1.7589 0.0785987 .
```

Write R code to produce this output.

Solution:

```
mod <- Arima(admit, order=c(2, 0, 0), xreg=1:50)
coeftest(mod)
```

e) From the output in the previous part, identify the p-value for the question stated in that part, and state your conclusions at $\alpha = 0.05$.

Solution: The p-value is 0.0785987. We would fail to reject H_0 and conclude that there is not statistically significant evidence of a decline in admissions rate while including an AR(2) term in the model.