Review of Quiz 3

DSC 152

Jun. 4th, 2026

Question #1

Researchers want to understand how coaching may change the relationship between interview practicing and stipends. 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) Write the estimated regression equation for students who *did* use coaching.

The fitted interaction model for $y = \text{stipend}$ is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \text{practice\_hours} + \widehat{\beta}_2 \cdot \text{coaching} + \widehat{\beta}_3 \cdot (\text{practice\_hours} \cdot \text{coaching}).$$

Since "no" is the reference group, we can set coaching $= 1$ (to indicate "yes") and simplify:

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 \cdot (\text{practice\_hours}) + \widehat{\beta}_2 \cdot 1 + \widehat{\beta}_3 \cdot (\text{practice\_hours} \cdot 1) = \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3) \cdot (\text{practice\_hours}) = \\ &= (18 - 1.5) + (0.2 + 0.1) \cdot (\text{practice\_hours}) = 16.5 + 0.3 \cdot (\text{practice\_hours}).\end{aligned}$$

Question #1

(b) What is the correct interpretation of the interaction coefficient 0.10 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.
  TRUE: The interaction coefficient tells us how much steeper (or flatter) the relationship between interview practicing and stipends becomes under coaching.
- For coached students, the estimated slope relating practice hours to stipend is 0.10 thousand dollars per hour.
  FALSE: This is not true because, for coached students, the correct slope is $\beta_1 + \beta_3$; this answer choice confuses the interaction term with the full slope.
- At a fixed number of practice hours, coached students are predicted to receive 0.10 thousand dollars more than uncoached students.
  FALSE: This describes a main effect difference (shift in the intercept), where the difference in stipends at any fixed practice level depends on the number of practice hours (not just the 0.10 coefficient).
- The interaction coefficient means that coached and uncoached students have the same predicted stipend when `practice_hours = 0`.
  FALSE: The interaction has no effect at `practice_hours = 0`, since the difference between those who have coaching and those who do not is β_2.
- None of the above.
  FALSE: Since answer choice 1 is correct, this cannot hold.

Question #1

(c) If the researcher wants to see if coaching contributes *at all* (including its interaction), explain where the appropriate p-value is or how you would find this.

Because the requested test is a joint test of the coaching main effect and the interaction term (i.e., $H_0 : \beta_2 = \beta_3 = 0$, where β_2 corresponds to the coaching main covariate term and β_3 its interaction), the appropriate p-value is not available from the coefficient table.
To find it, we use a reduced model without any coaching terms:

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

Question #1

(d) Which best explains why p-value for `coachingyes` is not enough to answer 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.
  FALSE: Statement describes where estimate comes from but does not explain why the t-test fails to test both the main effect and interaction jointly.

- 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.
  TRUE: The coefficient for `coachingyes` tests for $H_0 : \beta_2 = 0$, which is the coaching difference when `practice_hours = 0`, and does not include the interaction term β_3. This means it cannot test the overall contribution of coaching to the model.

- The individual t-test for `coachingyes` tests the same null hypothesis as the model comparison, but reports it using a different test statistic.
  FALSE: The t-test only tests $H_0 : \beta_2 = 0$, while the model comparison tests $H_0 : \beta_2 = \beta_3 = 0$; clearly, this joint null is not the same as the null for β_2 only.

- The individual t-test for `coachingyes` checks whether the slope for `practice_hours` is zero among students who did not use coaching.
  FALSE: Statement confuses the coaching indicator with coefficient for `practice_hours`; test for `coachingyes` has nothing to do with slope of practice hours.

- The individual t-test for `coachingyes` compares the residual standard error from the full model to the residual standard error from the reduced model.
  FALSE: Individual t-tests do not compare model fit metrics; this is done in model comparisons (e.g. $\mathcal{F}$-tests).

Question #2

Suppose that researchers are primarily interested in whether `practice_hours` is associated with `stipend`, with `portfolio_score` as secondary covariate. Here is a proposed 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.

Question #2

(a) Which of the following statements are true regarding this as a *workflow for statistical inference* about `practice_hours`?

- [x] 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.
  TRUE: Using the data twice (first to decide whether to drop `portfolio_score`, then to test for `practice_hours`) will lead to post-selection inference bias, meaning the reported p-value is invalid.

- [ ] This is an incorrect workflow because inference about practice hours should always be based on the largest model available.
  FALSE: The issue is not always using the largest model, but rather avoiding model selection based on the data. Sometimes reduced or enlarged models can be appropriate, *if justified apriori*.

- [x] The decision of whether to include `portfolio_score` in the model should not be based on this workflow, but rather on whether there is scientific justification for it as a confounding variable.
  TRUE: Variable inclusion should be driven by causal, substantive, or domain-based reasoning (e.g. controlling for confounders), not by statistical significance in the dataset.

- [ ] This is a fine workflow because `portfolio_score` must be statistically significant before the coefficient on `practice_hours` can be interpreted.
  FALSE: Statistical significance of another covariate is not required to interpret the main covariate `practice_hours` correctly, and p-values themselves should not determine interpretability.

- [ ] This is an incorrect workflow because dropping `portfolio_score` changes the definition of the outcome variable being modeled.
  FALSE: Dropping a predictor changes the model specification, not the outcome variable, which remains the same (`stipend`) in both cases.

Question #2

(b) Suppose we refit the model in (1) after centering and rescaling. State two benefits for centering/scaling for interpretations.

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) Which of the following are true about centering and scaling covariates?

- ☐ Centering a covariate will not change the p-value for whether its coefficient is equal to 0, but scaling it will.
  FALSE: Scaling does not change the p-value for $H_0 : \beta_j = 0$; it just rescales the coefficient and its standard error proportionally.
- ■ Neither centering a covariate nor scaling it will change the p-value for whether its coefficient is equal to 0.
  TRUE: Centering/rescaling are linear transformations that do not change the underlying test statistic or the model's hypothesis test for a coefficient is zero.
- ☐ Centering a covariate or scaling it will both change the p-value for whether its coefficient is equal to 0.
  FALSE: Neither operation changes the test statistic for a coefficient, so the p-value is unchanged.
- ■ Centering and scaling covariates will not change the R^2 value for the model.
  TRUE: Centering/rescaling are linear transformations of the predictors. They do not change the space of the fitted values that the model can produce, and so they fitted values and the residuals remain unchanged with respect the original model, then the R^2 value will also remain unchanged.

Question #2

(d) If we fit two models (`model_raw` vs. `model_logy`), with the residual plot for the former showing clear curvature and increasing spread while the plot for the latter is more random and stable. Which of the following statements are most accurate?

- ☐ This suggests that the researcher should use `model_logy` for statistical inference in the primary analysis on these data.
  FALSE: Residual diagnostics alone suggest improved model fit under the transformation, but, by themselves, they do not justify selecting it as the primary analysis without considering scientific context (or even interpretability).
- ☐ `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.
  FALSE: Curvature and heteroskedasticity violate the assumptions for linear regression, which can mean standard errors and inference is invalid. Ease of interpretation is insufficient justification for choosing this model.
- ■ As a sensitivity analysis, these residual plots would suggest that future studies should consider using `model_logy` for statistical inference in the primary analysis.
  TRUE: Improved residual patterns are indications that the log-transformed model does a better job of satisfiying the modeling assumptions, making it a reasonable candidate for future primary analyses (or for checks for robustness).
- ■ Scientific rationale should also guide model choice; the residual plot should not be the only consideration.
  TRUE: Model selection should incorporate statistical diagnostics with domain -based and interpretability-based contexts, since the residual plots alone are insufficient for determining the correct scientific specification.
- ☐ None of the above.
  FALSE: Since answer choices 3 and 4 are correct, this cannot hold.

Question #2

(e) 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.
  FALSE: Model choice should be based on assumptions, not maximizing R^2 since it can increase even for mis-specified models.
- Once a predictor is statistically significant, model selection (including whether or not to do a log transformation) no longer affects inference validity.
  FALSE: Statistical significance does not guarantee a correctly specified model; transformations can still materially affect estimates, standard errors, and validity of inference.
- Instead of a log transformation, centering and scaling can make coefficients easier to interpret, and they are a substitute for checking model conditions.
  FALSE: Centering/rescaling only changes units and numerical stability; it does not fix issues with systematic curvature oor heteroskedasticity, meaning it cannot replace diagnostic checks or transformations.
- None of the above.
  TRUE: Since none of the other options apply, this one is true.

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) Compute the relative risk of receiving an offer for coached students compared to uncoached students. Interpret your answer.

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

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

Question #3

(b) Compute the odds ratio for these data. Interpret your answer.

The odds ratio is

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

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

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

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

Question #3

(d) 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..

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.

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) Write one line of R code to run the model that would produce the output summarized above.

```
# SOLUTION
glm(offer ~ coaching + portfolio_score, data=career_df,
    family='binomial')
```

Question #4

(b) Provide an interpretation on the odds-ratio scale using the coefficient for `coachingyes`.

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) 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?.

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

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.
A plot of $\log(\text{visits\_t})$ against day number shows a repeating weekly up-and-down pattern, but no obvious long-run upward trend.

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

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.

Question #5

(b) A researcher wants to adjust for the weekly up-and-down pattern without any $\mathrm{AR}(p)$ or $\mathrm{MA}(q)$ terms, but simply with day-of-week effects using Monday as the baseline day. Write R code to run a regression model for studying this relationship with adjustment for the day of the week.

```r
# SOLUTION
lm(log(visits) ~ bookings + day of week, data=traffic_df)
```

(a) For the log website-traffic series, an AR(1) model has the form:

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

What does it mean if ϕ_1 is close to 1?

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.