Homework 3

Released Tuesday, November 17, 2026

Due Tuesday, November 24 at 11:59 PM PT

Homework Instructions

- This homework is worth **40 points** and covers **Lectures 11–15 and Labs 6–8**: interactions, model selection, transformations, relative risks and odds ratios, and introductory logistic regression.
- There are **two Gradescope assignments with the same deadline**. Submit `answers.pdf` (37 points) to the **HW3 Written** assignment and `q4b.R` (3 points) to the separate **HW3 Q4(b) Code** assignment. **Use these exact filenames.** Submit the R file first so you can use its feedback when finishing the PDF.
- **Assign the correct PDF pages to each question and subpart in Gradescope.**
- In `answers.pdf`, answer Questions 1–3 and 4(a), 4(c), and 4(d). Write solutions in R Markdown and knit to PDF, or upload clearly scanned handwritten solutions. Include R code, relevant outputs, and plots for computational work. No code is needed for purely conceptual answers.
- For 4(b), submit only `q4b.R` to the code assignment. The autograder will run your function on a private confirmation dataset and grade the *numeric p-value returned by your function*. It accepts equivalent numeric notation, including decimals and scientific notation. Do not hard-code a p-value or a file name. The private data are not distributed. Use an unrounded (non-formatted) p-value.
- When 4(b) is accepted, the feedback will display **your confirmation p-value**. Use that value in **4(c)** to interpret the held-out test, and in **4(d)** to explain how it affects your provisional recommendation.
- Use only `Homework_3_DSC_152_Explore.tsv` for Questions 1–3 and 4(a). You may test your 4(b) function on Explore, but an Explore p-value is not a confirmation result.
- Give complete, concise answers. As a guide, use at most 100 words for 2-point parts, 130 for 3-point parts, and 150 for 4-point parts. Code, equations, tables, and plots do not count. Unless stated otherwise, use $\alpha = 0.05$ and 95% for requested confidence intervals. Report probability differences in percentage points. You may use base R or tidyverse code.

Introduction: Does AI help applicants get noticed?

In 2026, a job-application platform is considering two features. An optional AI coach helps applicants edit a truthful account of their experience. A skills-first layout puts projects and demonstrated abilities ahead of education history. Would either feature help, and is offering both better than offering them separately?

The platform conducts a hypothetical randomized experiment. Each applicant submits exactly one application and is randomly assigned to one of four feature combinations, with 600 applicants per combination. The platform records whether the employer subsequently invites the applicant to interview and how long a reviewer initially spends opening the application. The applicant's job family and original writing score are recorded before random assignment. The latter, recorded on a 0–100 scale, is called `prior_score`. All 2,400 records are synthetic; they do not represent actual applicants or employers.

Student data file: `Homework_3_DSC_152_Explore.tsv` contains 1,600 exploratory applications. The 800-row confirmation sample is not distributed. The supplied file has these columns; each row represents one application:

Column	Meaning
`application_id`	Unique application identifier.
`split`	`explore`; the confirmation split was assigned before outcomes were observed.
`coach`	Whether the AI coach was `on` or `off`.
`skills_first`	Whether the skills-first layout was `on` or `off`.
`prior_score`	Pre-treatment writing score, on a 0–100 scale.
`job_family`	Data, Design, or Operations.
`review_seconds`	Positive initial reviewer time in seconds.
`interview_invite`	1 if an interview invitation was received, and 0 otherwise.

The explore sample has equal numbers in each of the four feature combinations. Treat different applicants as independent. Use `off` as the reference for both features and Data as the reference job family, or explicitly transform equivalent contrasts if using other reference levels. The experiment can speak to the assigned features for this platform's participating applicants and employers; it does not directly measure job offers, hiring quality, or performance on the job.

Question 1 [10 points]

Invitations, risk, and odds. A product manager reports that the AI coach "triples the chance of an interview when skills-first is on" and urges immediate rollout. Use the explore data to evaluate that statement.

(a) [2 points] Summarize the number of invitations, number of applications, and invitation proportion for each of the four feature combinations. Identify the experimental unit.

(b) [3 points] Among applicants assigned the skills-first layout, compare AI coach on with off. Calculate the relative risk and the odds ratio for receiving an invitation, and interpret each in context. Explain why they are not the same number.

(c) [3 points] Calculate the AI coach's invitation-rate difference (on minus off) separately for each layout. Then compare those two differences. Does a larger observed difference establish a statistically significant interaction? Explain.

(d) [2 points] Suppose an auditor instead collects exactly 100 invited applicants and 100 non-invited applicants, sampling randomly within each invitation group. The auditor still knows which features each applicant used. Can invitation risks and relative risks be estimated directly from this new sample? What association measure from part (b) can still be estimated, and why?

Question 2 [10 points]

Does a shorter review mean a better process? The product team hopes applicants can communicate relevant experience more efficiently, but also worries that very short review times could mean reviewers are skipping important details.

(a) [2 points] Compare residual-versus-fitted plots from two regression models of `review_seconds`: one on its original scale and one on its log scale. In both, include the two assigned features, their interaction, `prior_score`, and `job_family`. Which scale appears more defensible for the usual linear-regression inference, and what does the evidence in your plots show?

(b) [4 points] Use `log(review_seconds)` as the response in R (the natural logarithm by default). Estimate the AI coach's association with review time when the skills-first layout is off and when it is on, keeping other predictors fixed. Report both comparisons as either exact percentage changes using $100(e^b - 1)$ or clearly labeled small-coefficient approximations using $100b$, where b is the appropriate natural-log contrast. Explain how the interaction changes the coach contrast on the log scale.
Do not treat coefficients as changes in seconds.

(c) [2 points] Is there evidence that the AI coach's association with log review time differs by layout? State the null hypothesis in terms of the fitted model, then report an appropriate statistic and conclusion.

(d) [2 points] A teammate says, "The log-time residual plot looks better, so our model is definitely correct and shorter review times must mean better application reviews." Give one statistical and one decision-making reason this conclusion is too strong.

Question 3 [10 points]

Modeling interview invitations. The outcome that matters most to the team is binary. Keep the same feature interaction, `prior_score`, and `job_family` predictors used above.

(a) [3 points] Fit a logistic regression for `interview_invite` using the two assigned features, their interaction, `prior_score`, and `job_family`. Report the interaction coefficient and its exponentiated value. Explain what the exponentiated value compares and why it is not a multiplier of interview probabilities.

(b) [3 points] For an applicant with `prior_score = 65` applying to a Data job, predict the probability of an interview invitation under all four feature combinations. Compare the AI coach's predicted benefit in percentage points with the layout off versus on. State what is held fixed.

(c) [2 points] For a Data applicant with the skills-first layout on, find the minimum prior_score giving a fitted interview probability of 50% with the AI coach on (allow fractional scores). Is 50% attainable with the coach off on the 0–100 score scale? Justify. Hint: $p = 0.5$ has log-odds zero.

(d) [2 points] The experiment used random assignment, yet the model adjusts for `prior_score` and `job_family`. Explain one useful reason to include these pre-treatment predictors and one important limit on what an interview-invitation model can establish.

Question 4 [10 points]

A result worth believing. Another analyst wants to explain initial review time using an automatic search for interactions. They start with an additive linear regression for `log(review_seconds)` that includes `coach`, `skills_first`, `prior_score`, and `job_family`. Then they separately add each of three candidate interactions:

- AI coach × skills-first layout.
- AI coach × job family.
- Skills-first layout × job family.

They plan to select the interaction with the smallest p-value and report that p-value as though the model had been fixed before looking at the data.

(a) [3 points] Using Explore, compare each candidate model with the *same additive baseline* using partial F-tests. Report three p-values, the number of added parameters for each test, and the selected interaction. In one sentence, explain why the smallest exploratory p-value should not be reported as though its hypothesis had been specified in advance.

(b) [3 points] In the separate **HW3 Q4(b) Code** assignment, submit `q4b.R` defining this function:

```
run_confirmation_test(confirm_path)
```

The function must read the TSV file named by its `confirm_path` argument, fit the additive model from (a) for `log(review_seconds)` and a model adding only your selected interaction. It must return the **unrounded (non-formatted) confirmation p-value** for the partial F-test as a single finite numeric value, not a printed or formatted string. Do not hard-code a result or file name. The autograder runs your function on private data; accepted feedback displays the p-value.

(c) [2 points] In `answers.pdf`, report the confirmation p-value shown in accepted feedback for 4(b). At $\alpha = 0.05$, what does it tell you about the selected log-review-time interaction, and what does it not establish? Why does a smaller p-value alone not prove a larger effect? State one benefit and one cost of using a held-out sample.

(d) [2 points] In `answers.pdf`, in at most three sentences, make a provisional recommendation about the features. Support it with one numerical invitation finding and one numerical review-time finding from Explore, state one limitation, and explain how the confirmation result *from 4(b)* affects your decision. The held-out test concerns review time, not invitations.