Exam 1 – Math 150

Ethan Ashby

due after 48 hours ~ March 19, 2021

PLEASE

Knit this file before doing anything! Make sure that the data load properly and that the packages work.

Guidelines

The exam is not be timed. You will pick up the exam sometime between Wednesday and Friday (from Gradescope), and it will be due 48 hours later. The general idea for what is allowed and what isn't is: all the class materials are allowed; all the non-class materials aren't allowed. Don't talk to anyone except the professor about the exam.

You may access the following:

- talking to **me**
- course textbook
- online course notes
- any notes you personally have written
- course videos
- applets
- HW / WU assignments & solutions
- the help files in R, e.g., `?geom_point()`
- previous Discord conversations
- posted sample exam and solutions
- graded assignments (on Gradescope)

You may not access the following:

- talking to **others**
- Google / online searching
- other textbooks / online materials
- other videos on the content

If you have any questions about R or any other types of questions whatsoever (even if they seem silly!), please, please, please email or DM me on Discord. I **love** the questions, and I'll try to answer as quickly as possible. If it doesn't seem like a fair question to ask, I'll just say that I can't answer it. No problem. But please ask!

Q1. Obstetrics (giving birth) and Periodontal (gum disease) Therapy[1]

###Note before reading

According to the `glm` documentation, by default, `glm` designates the first factor level as a failure and the second factor level as a success. So all following analyses will consider full term births (the second factor level for the `birth` variable), as successes.

The data dictionary is given here: https://pomona.box.com/s/n19jr6wevk8l5z61a6faeoao60uzba2h

> Maternal periodontal disease has been associated with an increased risk of preterm birth and low birth weight. We studied the effect of nonsurgical periodontal treatment on preterm birth.
>
> We randomly assigned women between 13 and 17 weeks of gestation to undergo scaling and root planing either before 21 weeks (413 patients in the treatment group) or after delivery (410 patients in the control group). Patients in the treatment group also underwent monthly tooth polishing and received instruction in oral hygiene. The gestational age at the end of pregnancy was the prespecified primary outcome. Secondary outcomes were birth weight and the proportion of infants who were small for gestational age.
>
> We designed the present trial to assess whether nonsurgical periodontal treatment in pregnant women reduces the risk of delivery before 37 weeks.

(a) Before considering the actual response variable (`birth`), assess whether or not the analysis should be done separately for the different `clinic`s. To focus the assessment, consider only the `age` variable. In particular do the following:

i. (+3 pts) Compute the average `age` for each of the 4 different `clinic`s. Example code:

```
opt %>%
  group_by(clinic) %>%
  summarize(mean_age=mean(age, na.rm = TRUE)) %>%
  kable()
```

clinic	mean_age
KY	24.73934
MN	27.19838
MS	25.05729
NY	26.76879

The mean age at the `KY` clinic was 24.7, `Mn` was 27.2, `MS` was 25.1, and `NY` was 26.8.

ii. (+4 pts) Run a single **linear** model to identify the significance of the `clinic` variable in predicting the average of the `age` variable (there will be more than one p-value). Include the `tidy()` model output.

```
fit<-lm(age~clinic, data=opt)
fit %>% tidy() %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	24.7393365	0.3765538	65.6993432	0.0000000
clinicMN	2.4590441	0.5127566	4.7957334	0.0000019
clinicMS	0.3179552	0.5455429	0.5828234	0.5601726
clinicNY	2.0294496	0.5610087	3.6175013	0.0003157

[1] reference provided in the solutions

The baseline level of `clinic` is the `KY` value. This linear regression output illustrates that the `MN` and `NY` clinic patients are significantly older (as evidenced by positive b_1 estimates and very small p-values) than those who attended the `KY` clinic. Those who attended the `MS` clinic were older than those at the `KY` clinic (as evidenced by the positive b_1 estimate), though the difference in ages was not significant (as evidenced by a p-value of 0.560).

iii. (+6 pts) Using the output (that is, the numbers) from the linear model, write out the prediction model (predicting age) *separately* for each `clinic` [there should be 4 prediction models]. Make sure to use appropriate notation indicating prediction as well as which variable is which, etc.

The models presented below designate b_0 as the intercept (mean age of someone at `KY` clinic), and $\alpha_1, \alpha_2, \alpha_3$ as the effects of each clinic (`MN`, `MS`, and `NY` respectively) on mean age relative to the mean age at the `KY` clinic.

Model 1 for predicting age at `KY` clinic:

$$\hat{y}_{\text{KY}} = b_0 = 24.74$$

Model 2 for predicting age at `MN` clinic:

$$\hat{y}_{\text{MN}} = b_0 + \alpha_1 = 24.74 + 2.46 = 27.2$$

Model 3 for predicting age at `MS` clinic:

$$\hat{y}_{\text{MS}} = b_0 + \alpha_2 = 24.74 + 0.32 = 25.06$$

Model 4 for predicting age at `NY` clinic:

$$\hat{y}_{\text{NY}} = b_0 + \alpha_3 = 24.74 + 2.03 = 26.77$$

iv. (+4 pts) Using the the estimates/coefficients from i. & iii. and the p-values from ii., do you think that the experimental results should be run/reported separately for the different `clinics`? Explain in 2-3 sentences.

I believe that results should be reported on a per-clinic basis. The relatively large, positive beta estimates and small p-values for the `MN` and `NY` clinics show that they have significantly older patients than the `KY` clinic. This significant difference in ages should be accounted for in the model.

```
lm(age~clinic, data=opt) %>% anova() %>% kable()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
clinic	3	962.5422	320.84741	10.72413	6e-07
Residuals	819	24503.0641	29.91827	NA	NA

Indeed, running an ANOVA test generates a very small p-value, suggesting that true mean ages among the different levels of `clinic` are significantly different. Thus, `clinic` should be included in our model of `age`.

(b) Did the treatment work? There are a **zillion** variables here, so don't go down a rabbit hole! A few suggestions are given to assess the question of whether or not the treatment worked, stick to those suggestions.

i. (+6 pts) Run a logistic regression model with only the treatment variable. Provide an interpretation of the model results[2] (use words like "periodontal treatment" and "full term birth", also communicate what the model actually is with a word like "average" or "probability" or "odds"). Show the estimated model along with your interpretation/conclusion.

[2] "interpretation of results" = a conclusion to a hypothesis test

```
glm(birth~group, data=opt, family="binomial") %>%
  tidy() %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.8961761	0.1473116	12.8718705	0.0000000
groupT	0.0723338	0.2109354	0.3429193	0.7316591

A logistic regression model using only the `group` variable (periodontal treatment status) to model the binary outcome of either `full term` or `early` birth failed to identify `group` as a significant variable in determining `birth`. In other words, we fail to reject the null hypothesis that the odds of a full term birth are the same between treatment and control groups.

To illustrate, the b_1 associated with `groupT` is 0.0723. So the odds ratio of a full term birth in groupT relative to groupC is: $\hat{OR} = \frac{\exp(b_0+b_1)}{\exp(b_0)} = \frac{\exp(1.90+0.0723)}{\exp(1.90)} = 1.075$. This odds ratios is very close to 1, and the large p-value from the `tidy` output demonstrates that it is not significantly different from 1. Thus, we are unable to claim from this logistic regression output that periodontal treatment significantly changes odds of full term birth.

ii. (+8 pts) Investigate whether or not `clinic` should be used in the model with the treatment variable. Comment on whether it should be used as a main effect (i.e., not interaction) or as a variable that interacts with the treatment or neither. Show work.

```
#calculate deviances for null model and full model
deviances<-c(glm(birth~group, data=opt, family="binomial") %>%
  glance() %>% select(deviance),
glm(birth~group+clinic, data=opt, family="binomial") %>%
  glance() %>% select(deviance))

#calculate drop in deviances
drop_in_dev=deviances[[1]]-deviances[[2]]
#df=3 b/c full model requires estimating 3 additional params
df=3
#calculate p-value from chi square distribution
pchisq(drop_in_dev, df=df, lower.tail=FALSE)
```

```
## [1] 0.08888425
```

The additive model yields a p-value associated with the `clinic` variable of approx. 0.089 (as determined by the drop-in-deviance test comparing the null model of `group` and the full model of `group+clinic`). This illustrates that `clinic` (as a main effect) is not significantly associated with the `birth` variable, and should not be included as a main effect variable in the final model.

```
#calculate deviances for null model and full model
deviances<-c(glm(birth~group, data=opt, family="binomial") %>%
  glance() %>% select(deviance),
glm(birth~group*clinic, data=opt, family="binomial") %>%
  glance() %>% select(deviance))

#calculate drop in deviances
drop_in_dev=deviances[[1]]-deviances[[2]]
#df=6 b/c full model requires estimating 6 additional params
df=6
#calculate p-value from chi square distribution
pchisq(drop_in_dev, df=df, lower.tail=FALSE)
```

```
## [1] 0.1319109
```

The interactive model yields a p-value associated with the `clinic` variable of 0.132 (as determined by the drop-in-deviance test). This illustrates that the interaction between `group` and `clinic` is not significant, and should not be included as a variable in the final model.

Thus, I conclude that when modeling `birth`, `clinic` **should not** be included in the model as a main effect nor an interaction term!

iii. (+ 6 pts) Choose one additional baseline variable (just pick one, don't spend hours trying to find a perfect variable!) from the list of "Baseline Risk Factors" or "Periodontal Summaries" (which start with the code `bl` for `baseline`). Does including that variable have any impact on your conclusions about the treatment effect? Explain / show work. [To the question of: "what should I do in part iii. about part ii.?" The answer is: whatever part ii. told you to do is what you should do here in part iii.]

I'm going to test if `diabetes` shows any association with preterm birth. To do this, I build a logistic regression model including the covariates `group` and `diabetes` (as main effects). The output of my model is shown below:

```
glm(birth~group+diabetes, data=opt, family="binomial") %>% tidy() %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.9436451	0.1493581	13.0133188	0.0000000
groupT	0.1204997	0.2136957	0.5638845	0.5728327
diabetesYes	-1.5127581	0.4377081	-3.4560887	0.0005481

According to the `tidy` output above, `diabetes` is significantly associated with `birth` (as evidenced by small p-value). The negative slope estimate means that having diabetes reduces the odds of achieving a full term birth. In other words, the odds ratio of having a full term birth in diabetic vs nondiabetics is $\exp(-1.51) = 0.22$. Put simply, if you are diabetic, your odds of having a full term birth are less than a quarter of the odds of full term birth if you are not diabetic.

However, inclusion of `diabetes` in the model does not change our view of the treatment effect. The p-value associated with `groupT` is still large, illustrating that treatment is not significantly associated with birth status.

iv. (+5 pts) Comment on what you've found in parts i., ii., iii. above. Address the question in (b): Did the treatment work?

Across all models surveyed, the `group` variable (periodontal treatment) did not have a significant effect on birth status (as measured by odds of full term birth). Thus, we cannot claim that periodontal treatment influences odds of preterm birth. `clinic` didn't warrant inclusion as a main effect nor interaction variable, illustrating that neither treatment effect nor full term birth odds varied across clinic sites. `diabetes` was found to be significantly associated with birth state, showing that diabetic mothers have significantly lower odds (less than 1/4) of having full term births than nondiabetic mothers.

Q2. Cross Validation and ROC

Below are 3 models, corresponding to 3 different subsets (each subset is a different 2/3 of the data) of the original dataset (as above). Note that the response variable is still whether the baby is full term (`birth`), but so as to distinguish from Q1, different explanatory variables are used. Do not change the code below at all.

Model AB

term	estimate	std.error	statistic	p.value
(Intercept)	4.4204067	0.7753462	5.701204	0.0000000
age	-0.0523978	0.0239041	-2.192001	0.0283794
bmi	-0.0367757	0.0163738	-2.246010	0.0247034

Model AC

term	estimate	std.error	statistic	p.value
(Intercept)	3.5482786	0.7470515	4.749711	0.0000020
age	-0.0249964	0.0233833	-1.068988	0.2850752
bmi	-0.0344355	0.0158295	-2.175403	0.0295999

Model BC

term	estimate	std.error	statistic	p.value
(Intercept)	3.4109312	0.8383593	4.068579	0.0000473
age	-0.0326599	0.0240580	-1.357549	0.1746069
bmi	-0.0243143	0.0199177	-1.220740	0.2221846

(a) (+7 pts) Given only the following 10 patients, predict their probability of full term birth. (You may want to do this by hand/calculator, or use the `predict()` function, see HW6.) Provide the 10 predictions. Do not modify the code below which provides the 10 individuals.

```
ROC_people %>% kable()
```

clinic	group	age	bmi	birth	diabetes	partition
KY	C	27	25	full term	No	a
MN	T	35	18	early	No	a
KY	C	33	27	full term	No	a
NY	C	30	38	full term	No	b
MS	C	22	42	full term	No	b
KY	T	24	30	full term	No	c
MN	C	31	32	full term	No	c
MN	T	37	19	full term	No	c
KY	C	20	32	full term	No	c
NY	T	23	24	early	No	c

```
ROC_pred<-cbind(ROC_people,
"pred"=c(
  predict(birth_bc, newdata=ROC_people %>% filter(partition=="a"), type="response"),
  predict(birth_ac, newdata=ROC_people %>% filter(partition=="b"), type="response"),
  predict(birth_ab, newdata=ROC_people %>% filter(partition=="c"), type="response")))
ROC_pred  %>% kable()
```

clinic	group	age	bmi	birth	diabetes	partition	pred
KY	C	27	25	full term	No	a	0.8722786
MN	T	35	18	early	No	a	0.8617830
KY	C	33	27	full term	No	a	0.8424631
NY	C	30	38	full term	No	b	0.8160540
MS	C	22	42	full term	No	b	0.8252122
KY	T	24	30	full term	No	c	0.8869129
MN	C	31	32	full term	No	c	0.8346921
MN	T	37	19	full term	No	c	0.8560608
KY	C	20	32	full term	No	c	0.8998563
NY	T	23	24	early	No	c	0.9115445

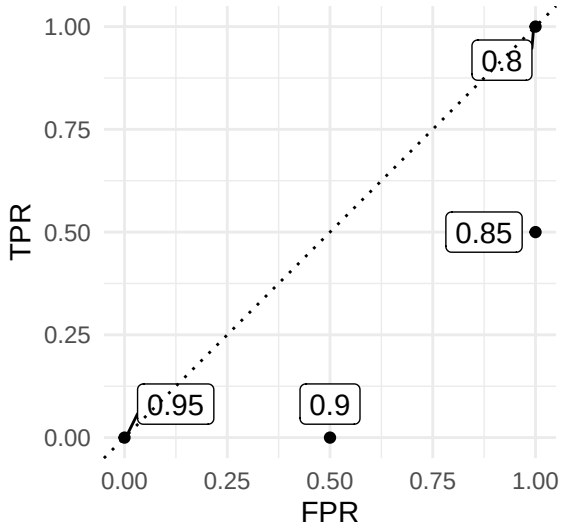
The `pred` column above shows the probabilities of full term birth for each individual, calculated using the model that DID NOT contain their particular partition.

(b) (+6 pts) Using the following four cutoff values: 0.8, 0.85, 0.9, and 0.95, plot (by hand is fine) **four** points on an ROC curve which correspond to the 10 predictions in part i. Show your work.

The following plot is the ROC curve for my predictions and the cutoff values listed above. The x-axis denotes the FPR, i.e. the number of samples predicted as full term births when they were early births over the total number of early births. The y-axis denotes the TPR, i.e. the number of correctly predicted full term births over all the full term births. The dotted black line denotes the expected curve observed by random guessing. The points are repel labelled according to their associated cutoff values.

```
ROC_data<-tibble(
  cutoff=c(0.8, 0.85, 0.9, 0.95),
  TPR=c(sum(ROC_pred$pred>0.8 & ROC_pred$birth=="full term")/sum(ROC_pred$birth=="full term"),
        sum(ROC_pred$pred>0.85 & ROC_pred$birth=="full term")/sum(ROC_pred$birth=="full term"),
        sum(ROC_pred$pred>0.9 & ROC_pred$birth=="full term")/sum(ROC_pred$birth=="full term"),
        sum(ROC_pred$pred>0.95 & ROC_pred$birth=="full term")/sum(ROC_pred$birth=="full term")),
  FPR=c(sum(ROC_pred$pred>0.8 & ROC_pred$birth=="early")/sum(ROC_pred$birth=="early"),
        sum(ROC_pred$pred>0.85 & ROC_pred$birth=="early")/sum(ROC_pred$birth=="early"),
        sum(ROC_pred$pred>0.9 & ROC_pred$birth=="early")/sum(ROC_pred$birth=="early"),
        sum(ROC_pred$pred>0.95 & ROC_pred$birth=="early")/sum(ROC_pred$birth=="early"))
)

ROC_data %>% ggplot(aes(x=FPR, y=TPR, label=cutoff))+
  geom_point()+
  xlim(0,1)+
  ylim(0,1)+
  coord_fixed()+
  geom_label_repel(point.padding=0.1)+
  theme_minimal() +
  geom_abline(aes(intercept=0, slope=1), lty=3)
```



Our ROC curve illustrates that our model is quite bad at predicting birth status (we are worse than random guessing).

(c) (+5 pts) Explain both **how** and **why** cross validation was used in this problem.

WHY: 3-fold cross validation was used in this experiment to gauge the performance of this model in the wild.

HOW: The data were partitioned into 3 groups: a, b, and c. Logistic regression models were fit to each pair of partitions (a and b, b and c, a and c), and predictions were generated for the left out observations for each model in our `ROC_people` table. For instance, the model built on data from the a and b subsets was used to generate predictions for data from the c subset. By building models and generating predictions for the held out data, we can get an idea of how the model will perform on a future dataset (as measured via AUC under an ROC curve for example).

Q3. (+9 pts) OR as RR????

Consider the situation where you have a case-control study evaluating an **extremely rare disease**. From the data, you estimate the OR (odds ratio), but you interpret it as if it were a RR (relative risk). Argue that your method is sound. Your answer should address:

- Why RR is not usually appropriate.
- Why OR is appropriate.
- Why **in this case** you can use the sample OR to estimate the true RR.

Suppose you have a 2x2 table containing the data from this case control study:

	Disease	Healthy
Exposed	a	b
Not Exposed	c	d

Relative Risk is not usually appropriate in case-control studies, because the study selects subjects on the basis of their disease status (response). Since the study selected subjects on the basis of disease status, we can only measure risk of exposure given outcome **but not** risk of outcome given exposure (which is what we need to calculate to determine relative risk).

Odds Ratio is appropriate for case control studies, because it is invariant to the choice of explanatory/response variable. In other words, we can obtain the desired quantity (odds ratio of disease between exposure statuses) from the odds ratio of exposure between disease states, because these values are equivalent.

The OR of the data in the 2x2 table above are calculated like so: $OR = \frac{a/b}{c/d}$. The RR of these data are calculated like so: $RR = \frac{a/(a+b)}{c/(c+d)}$. Note that in the case of rare diseases, where a, c are small relative to b, d, $a/b \approx a/(a+b)$ and $c/d \approx c/(c+d)$. In other words, when a disease is extremely rare, the risks and odds of infection are very similar in value. And if $a/b \approx a/(a+b)$ and $c/d \approx c/(c+d)$, then $OR = \frac{a/b}{c/d} \approx \frac{a/(a+b)}{c/(c+d)} = RR$. So $OR \approx RR$ in the case of rare diseases.

Q4. (+15 pts) CI for efficacy

Using the following table[3], find a 90% CI for the true efficacy[4] of the Pfizer vaccine. Include a one-sentence interpretation of the confidence interval using words like "Pfizer vaccine". [Hint, the efficacy is intimately related to one of the other statistics we've used in class / notes.]

Show work / all steps for creating the confidence interval.

		COVID		
		infected	not infected	total
treatment	placebo	162	21,668	21,830
	vaccine	8	21,822	21,830
		170	43,490	

[3] https://www.nytimes.com/2020/12/13/learning/what-does-95-effective-mean-teaching-the-math-of-vaccine-efficacy.html

[4] I called it the true efficacy instead of the effectiveness because effectiveness might also depend on how widely spread the disease is, mask protocols, etc.

Below shows the calculation of the point estimate for RR:

$$RR = \frac{\text{Risk}_{\text{vaccine}}}{\text{Risk}_{\text{placebo}}} = \frac{8/21830}{162/21830} = 0.049$$

Note that: $SE(\ln(RR)) \approx \sqrt{\frac{(1-\hat{p}_1)}{n_1\hat{p}_1} + \frac{(1-\hat{p}_2)}{n_2\hat{p}_2}} = \sqrt{\frac{(1-8/21830)}{21830 \cdot 8/21830} + \frac{(1-162/21830)}{21830 \cdot 162/21830}} = 0.362$.

A 90% CI for $\ln(RR)$ is computed by $\ln(0.049) \pm 1.645 \cdot 0.362$ which yields $[-3.604, -2.412]$. We backtransform by exponentiating these bounds to get a 90% CI for RR: $[0.027, 0.090]$.

We have our 90% CI for the RR. How do we get a 90% CI for efficacy? Using some algebra, we can show that efficacy is just 1-RR in this case:

$$\text{efficacy} = \frac{\text{Risk}_{\text{placebo}} - \text{Risk}_{\text{vaccine}}}{\text{Risk}_{\text{placebo}}} = 1 - \frac{\text{Risk}_{\text{vaccine}}}{\text{Risk}_{\text{placebo}}} = 1 - RR$$

By taking 1 minus the CI bounds for RR, we can find the 90% CI for efficacy: $[1 - 0.090, 1 - 0.027] = [0.910, 0.973]$. In short, we are 90% confident that the true efficacy of the Pfizer vaccine lies between 0.910 and 0.973.

Q5. (+4 pts each; +2 T/F, +2 explanation) TRUE or FALSE + one sentence explanation.

(a) Interaction happens when two variables are highly correlated.

FALSE. Interaction happens when a main predictor's effect on the response changes depending on the level of another variable. A classic example is modeling test scores by hours studied and academic year; academic year and hours studied may not be correlated (indeed, correlation can only exist between continous variables), but there may exist an interaction (ex. older students may show greater test score increases per hour studied because they are more efficient studiers).

(b) In logistic regression, if **in the population** the probability of success does not depend at all on the explanatory variable, then the calculated value of b_1 will be zero.

FALSE. While $\beta_1 = 0$ in the population, our b_1 is calculated from a *random sample* from the population. Due to the variability of a finite random sample, the b_1 value calculated from these data will not be 0.

(c) & (d) Consider the image (look at the pdf of the exam) which describes two logistic models:

Green: $logit(p(x)) = \beta_1 + \beta_2 x$
Black: $logit(p(x)) = \beta_3 + \beta_4 x$

(c) $\beta_2 = \beta_4$.

FALSE. The green curve is increasing wrt to x and the black curve is decreasing wrt x. Therefore, β_2 and β_4 must have opposite signs and therefore cannot be equal.

(d) β_1 and β_3 are both zero.

TRUE. We previously showed (in class and in the notes) that for $x = -\beta_1/\beta_2$ (or in the black line's case, $x = -\beta_3/\beta_4$), $p(x) = 0.5$. For both the green and black lines, $p(x) = 0.5$ at $x = 0$. Therefore, $x = 0 = -\beta_1/\beta_2 \Rightarrow \beta_1 = 0$. The same logic follows for β_3, so we conclude that both β_1 and β_3 must be 0.

```r
praise()
```

```
## [1] "You are sublime!"
```