PSC 204B – Winter 2024

The final is due on Thursday, March 21st by 5:00 PM PST, uploaded to Canvas.

Instructions:

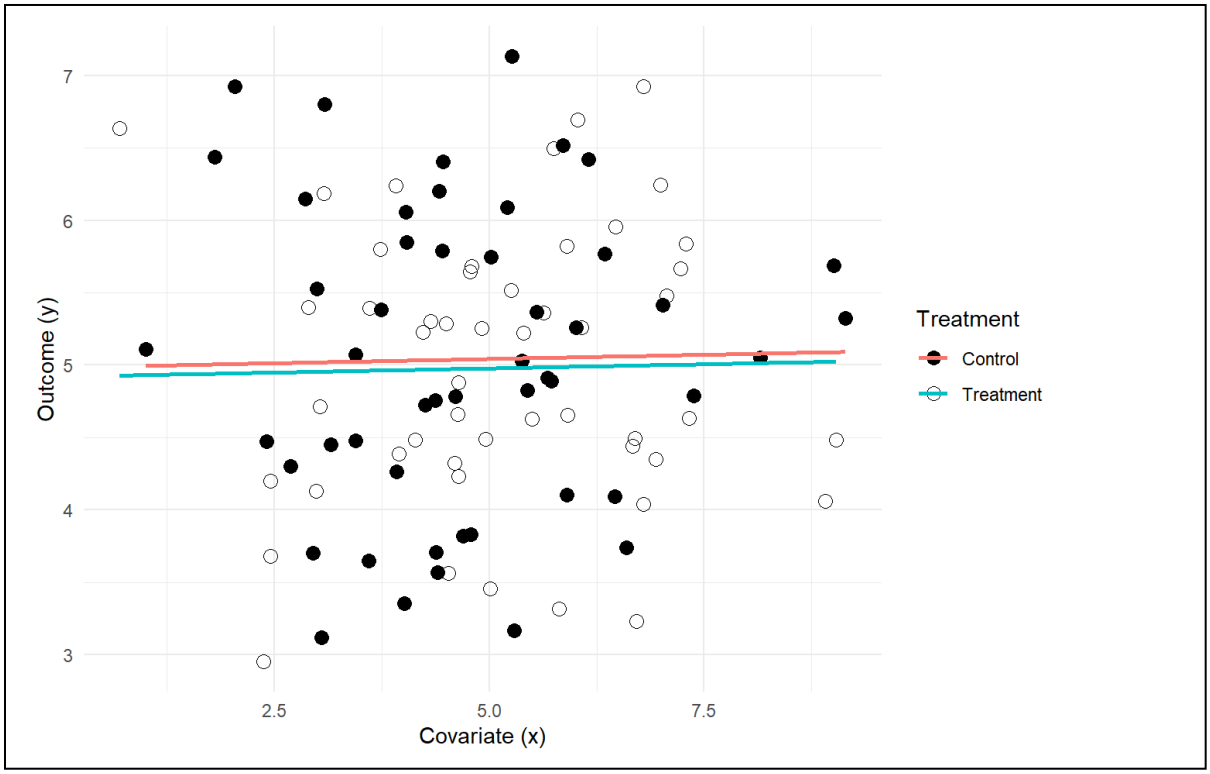
- This exam includes 8 questions and one bonus question summing up to 140 points. 100 points are needed for a perfect score. This exam includes a total of 40 bonus points, which may be used toward the exam portion of your final grade.
- **Please work independently.** You may consult with Philippe or Simran or Samuel concerning the final, but not with anyone else. You can also consult your book, class notes, lab material, recorded class and lab content, and search the web.
- Answers to all questions should be typed in the appropriate text box included with each question, and in red font (this is for ease of grading, and by default the text boxes are set to have red font. Note, you do not need to use red font for graphs). The only exception is for text boxes where you paste code; in these cases, the text boxes are formatted to have dark blue text (again, this is for ease of grading).
- Type your answers in complete sentences and answer questions thoroughly.
- Whenever possible, please answer using bullet points to list answers for each element of a question (for example, when interpreting regression coefficients, discussing similarities and differences, etc.).
- The text boxes are not necessarily scaled to the length of the expected answer, so in some cases you may need to expand the text box to fit your answers, and in other cases you may be able to answer a question without using all of the space in the text box.
- For questions requiring work in R, give your answers, code, and **relevant** output in the separate text boxes that are provided. Screen shots of output are preferred, but you can copy and paste too if it is easier.
- In general, round all numbers to 2 decimal places, unless you are specifically told to round to a different decimal place. However, if any of your numbers have several leading 0's (i.e., a very small decimal), you may either show a few additional decimal places for that one number or use a combination of rounding and scientific notation. For example, instead of 0.000027856, you could display 0.00003 or .29e-4 or < .001. This is preferred over rounding to 0.00. If you enter 0.00, it will be assumed to be a true 0, rather than a very small number, so please take note of this!
- Address every question – we can't give you points for skipped questions. Good luck!

Final Exam

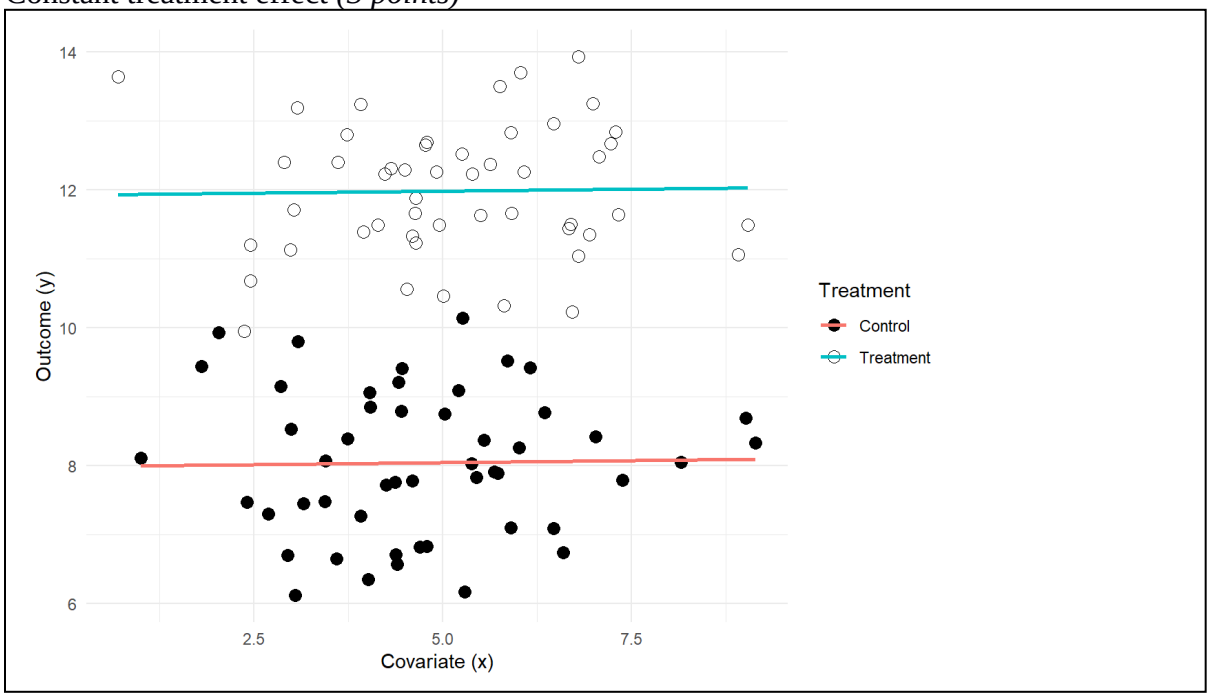
1. Treatment Effects (10 points)

Assume that a linear regression is appropriate for the regression of an outcome, *y*, on treatment indicator, *T* , and a single confounding covariate, *x*. Sketch (by hand, in paint, or in power point is fine) hypothetical data (plotting *y* versus *x*, with treated units indicated by open circles and control units black dots, respectively) and regression lines (for treatment and control groups) that represent each of the following situations:

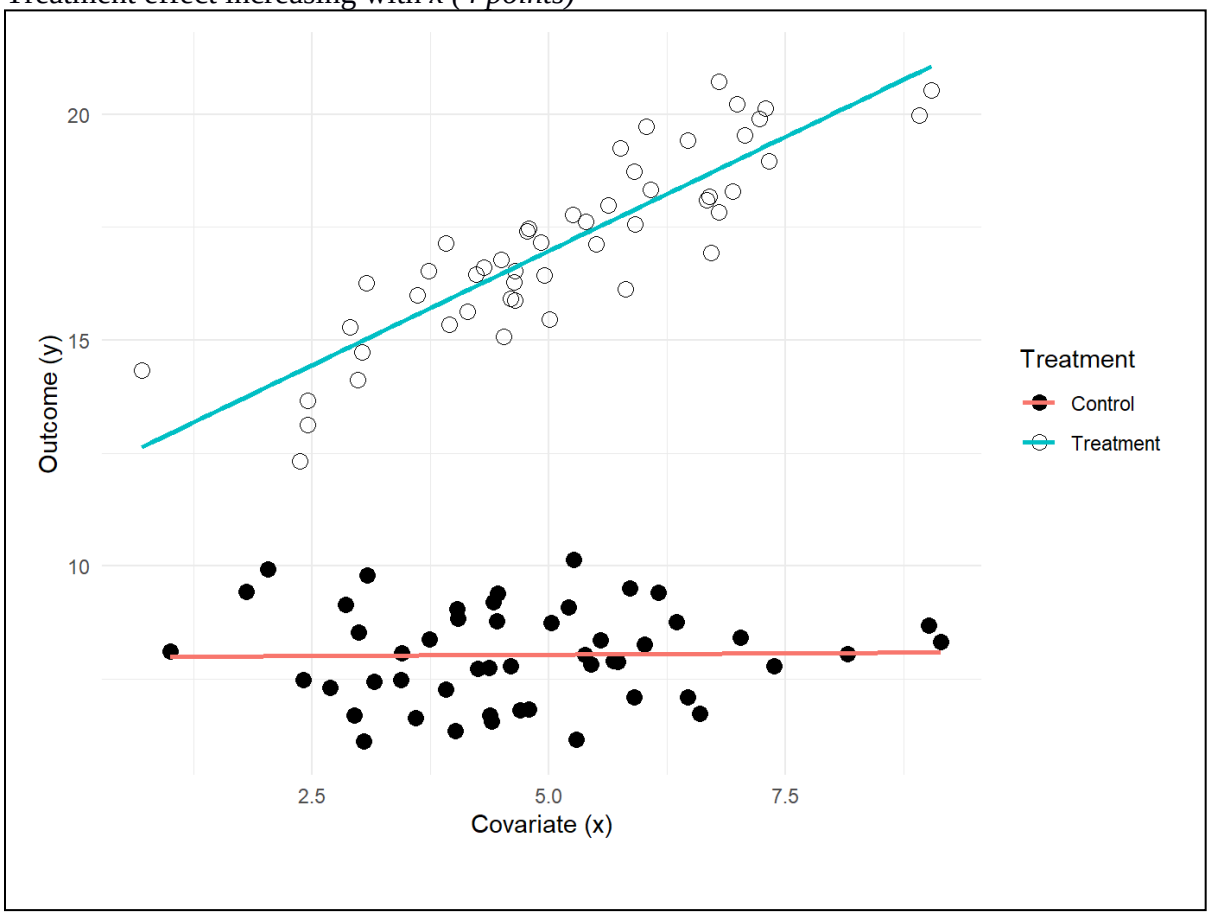
a. No treatment effect *(3 points)*



b. Constant treatment effect *(3 points)*



c. Treatment effect increasing with *x (4 points)*



2. Information Criteria (10 pts)
We talked about Akaike's information criteria; explain briefly the mechanism behind penalized information criteria.

Also, another criterion is the Bayesian Information Criterion (BIC) which is defined as BIC = D_{train} + ln(n)*k. From this definition, try to explain what the idea of BIC is and how it is different from and similar to AIC. In your answer, include an explanation under what conditions the penalty imposed by AIC is greater than the penalty imposed by BIC, and under what conditions the penalty imposed by AIC is less than the penalty imposed by BIC. Comment specifically on what the tipping point is.

Penalized information criteria help us select the best model among candidate models by balancing good fit to the data and parsimony. They serve to avoid overfitting by penalizing too complex models. Both AIC and BIC try to balance goodness-of-fit and complexity by adding a penalty term to the likelihood of the model, which increases with the number of parameters in the model. The key difference between the two is that BIC's penalty depends on sample size. This leads to a stronger preference for parsimonious models as the sample size grows. Therefore, the penalty imposed by AIC is greater than the penalty imposed by BIC when $2k > \ln(n) * k$, or $e^2 > n$, and less than the penalty imposed by BIC when $2k < \ln(n) * k$, or $e^2 < n$. The tipping point occurs when the number of observations is larger or smaller than $e^2 \cong 7$.

3. Logistic Regression (10 points)
Say you run a logistic regression with predictor variable X_1 and outcome variable Y. You fit the following model:

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X_1$$

a) What is the interpretation of B_0 in logits? What is the interpretation of B_1 in logits? *(4 points)*

B0 is the log odds of success when X1 is equal to zero. B1 represents the change in the log odds of success for a one-unit increase in X1.

b) How would you transform B_0 and B_1 to be expressed in odds? How would this change your interpretations of B_0 and B_1? *(3 points)*

To transform B0 and B1 in odds we need to exponentiate the coefficients.

Now, the interpretation of B0 is the odds of success when X1 is zero. It represents the baseline odds without the influence of X1. Then, B1 becomes the multiplier for the odds with each unit increase in X1. That is, it represents how much the odds of the event occurring are multiplied for every one-unit increase in X1.

Finally, you include the covariate within the analysis and fit the following model:

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X_1 + B_2X_2$$

c) What is the interpretation of each of the coefficients in this model? *(3 points)*

B0 is the log odds of success when both X1 and X2 are equal to zero.

B1 is the change in the log odds of success for a one-unit increase in X1, holding X2 constant.

B2 is the change in the log odds of success for a one-unit increase in X2, holding X1 constant.

4. Interactions (10 points)
It is common for researchers to incorporate interaction terms in regression models. For the following questions, please consider the following regression equation:

$$\hat{y} = B_0 + B_1 x + B_2 d + B_3 x d$$

a. Generally, what do the coefficients (B_0, B_1, B_2, & B_3) above represent? How would you interpret each one? Be very specific about how *x* and *d* are involved in each interpretation. *(3 points)*

B0: The expected value of y when both x and d are equal to zero.

B1: The change in y for a one-unit increase in x, when d equals zero.

B2: The change in y for a one-unit increase in d, when x equals zero.

B3: The change in the effect of x on y for a one-unit increase in d, or vice versa.

b. Imagine that *d* is a dummy coded variable indicating whether a participant is exposed to an experimental condition (1) or a control condition (0). Assume that *x* and *y* represent continuous scores. Under these conditions, what parameter or sum of coefficients (B_0, B_1, B_2, & B_3) correspond to the intercept and slope for both the control condition and for the experimental condition? *(2 points)*

Control Condition (d=0)

$$\hat{y} = B_0 + B_1 x + B_2(0) + B_3 x(0)$$
$$\hat{y} = B_0 + B_1 x$$

Intercept: B0 (the expected value of y when x is 0 in the control condition)
Slope: B1 (the change in y for a one-unit increase in x in the control condition)

Experimental Condition (d=1)

$$\hat{y} = B_0 + B_1 x + B_2(1) + B_3 x(1)$$
$$\hat{y} = (B_0 + B_2) + (B_1 + B_3)x$$

Intercept: B0 + B2 (the expected value of y when x is 0 in the experimental condition)
Slope: B1 + B3 (the change in y for a one-unit increase in x in the experimental condition)

c. What are some advantages and disadvantages to *z*-scoring variables before conducting regression (e.g., multiple regression, ANCOVA, including interaction terms) analysis? *(2 point)*

- Z-scoring standardizes variables to have a mean of 0 and a standard deviation of 1, placing them on a common scale. This makes it easier to assess the relative influence of independent variables. That is, they are useful for comparing effect size within the same study.
- It also improves interpretability of the coefficients. With z-scored variables, zero now represents the mean of that variable and one unit corresponds to one standard deviation.

- One disadvantage is that standardized parameters are dependent on the standard deviation of the respective parameters in the sample. Therefore, we might lose the ability to compare them across different samples.
- With bounded data, we often find nonlinear relationships near their limits. Standardizing the data can mask this true underlying relationship.

d. If the interaction term in the model above was not statistically significant but the main effects were, would you be able to interpret them? If yes, are there any caveats a researcher should be aware of? If all terms were significant, would you be able to interpret main effects? Explain your answers. *(3 points)*

If the interaction term is not significant, it suggests that the effect of one predictor on the outcome does not depend on the value of the other predictor. But we can still interpret the main effects, as they represent the average effect of each predictor on the outcome, regardless of the value of the other predictor.

5. Regression vs ANOVA (5 Points)

In general terms, explain the similarities and differences between linear regression and ANOVA. Please list each similarity and each difference as individual bullet points. You should comment on at least two similarities and at least two differences. *(5 points)*

Similarities:

- Both are used to explain the variance in a dependent variable by looking at the influence of independent variables. When all predictors of the linear regression are categorical, it is mathematically equivalent to an ANOVA.
- Both methods use the least squares method to estimate parameters.
- Both linear regression and ANOVA deal with continuous outcome variables.

Differences:

- To use categorical predictors in linear regression we need to create dummy variables. ANOVA, on the other hand, directly works with categorical predictors without requiring dummy variables.
- Linear regression's primary purpose is prediction. It aims to find the best-fitting line to predict the response variable based on the predictors. ANOVA focuses on group comparisons, assessing whether there are significant differences in means among different groups.
- The output of a linear regression provides coefficients for each predictor variable, while in ANOVA, the output provides an overall assessment of differences between groups.

6. Power Analysis (5 points)

Explain what is meant by the "Winner's Curse" in power analysis. What are some potential ways individual researchers could avoid this pitfall? What are some potential ways the field of psychology could avoid this pitfall?

It is a term that refers to the tendency to overestimate the effect size of a statistically significant result. That is, finding a significant result in a low-power study is probable to be an overestimation of the true effect size because it is highly unlikely to find the true effect size in such sample.
To avoid incurring in this mistake, individual researchers should always perform and report a priori power analysis to find the adequate sample size to detect the effect size of interest, refraining from conducting studies that don't have enough power. Also, reporting the effect size with its respective confidence interval for both significant and non-significant results.
As a field, Psychology should encourage replication studies, transparent reporting, and open data. Replication studies of previous significant findings is a way show whether these effects are possibly real or just an artifact. Data sharing allow other researchers to scrutinize and re-analyze findings, potentially uncovering inflated effects.

Applied Question

7. Logistic regression models (40 points)
In this question, you will use the *diabetes_pt1.csv* dataset (available on Canvas) to run a series of exploratory logistic regression models to try to find the best model to predict whether patients have diabetes (**Outcome**). You will then test the best model from these analyses in the *diabetes_holdout.csv* data file and explore whether a lasso regression obtains the same solution in the combined data file. The instructions below will explain what to do in further detail.

The dataset contains the following variables:

- **Pregnancies:** Number of times the patient has been pregnant
- **Glucose:** Plasma glucose concentration in an oral glucose tolerance test
- **Blood pressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction:** Diabetes pedigree function (family history of diabetes)
- **Age:** Age in years
- **Outcome:** Binary variable indicating whether patient has diabetes (Outcome = 1) or does not have diabetes (Outcome = 0)

More information about the dataset can be found at https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Follow the instructions in part a to fill in Table 6.1 *(4 points)*

a. Using the *diabetes_pt1.csv* data file, create four models to predict whether the patient has diabetes (Outcome). The first model should be a null model. The next model (Model 1) should include at least two predictors. The next model (Model 2) should include the same two predictors from Model 2 plus two additional predictors. Model 3 should be a full model and should include all eight predictors. Do not include interactions in any model.

In the table below (Table 7.1), report the model formula and BIC for each model you created (for consistency, use the BIC function in R). Example of model formula: y ~ Pregnancies + Glucose. The Null Model Formula and the Full Model Formula (for Model 3) are written for you. Round all numbers to two decimal places. Copy and paste your relevant code and output in the text boxes below.

Table 7.1 – Model Summaries from Part 1 of Data

	Model Formula	BIC
Null	y ~ 1	502.69
Model 1	y ~ Pregnancies + Glucose	398.77
Model 2	y ~ Pregnancies + Glucose + BloodPressure + SkinThickness	406.24

Model 3	y ~ Pregnancies + Glucose + BloodPressure + SkinThickness + + Insulin + BMI + DiabetesPedigreeFunction + Age	401.44

```
[1] 502.69 398.77 406.24 401.44
```

```
mod7a_null <- glm(Outcome ~ 1, data = diabetes, family =
"binomial")
mod7a_1 <- glm(Outcome ~ Pregnancies + Glucose, data =
diabetes, family = "binomial")
mod7a_2 <- glm(Outcome ~ Pregnancies + Glucose +
BloodPressure + SkinThickness, data = diabetes, family =
"binomial")
mod7a_3 <- glm(Outcome ~ Pregnancies + Glucose +
BloodPressure + SkinThickness + Insulin + BMI +
DiabetesPedigreeFunction + Age, data = diabetes, family =
"binomial")
models7a <- list(mod7a_null, mod7a_1, mod7a_2, mod7a_3)
round(sapply(models7a, BIC),2)
```

b. Based on the table created in part a, which model was the best at predicting whether patients had diabetes? Give evidence based on the fit statistics. *(5 points)*

Using BIC as criteria to select the best model, we would choose Model 1 among all the competing models since it had the lowest BIC value (398.77). That means this is the best model in terms of balancing good fit to the data and parsimony.

Fill out Table 7.2 based on the instructions given in parts c – e.

Table 7.2 - Model Estimates

	Data Part 1 (Best Model) (see Part c)		Hold Out Data (Best Model) (see Part d)	
Variable	Estimate (OR)	*p*	Estimate (OR)	*p*
Pregnancies	1.14	< .001	1.12	0.003
Glucose	1.04	< .001	1.04	< .001
BloodPressure	…	…	…	…
SkinThickness	…	…	…	…
Insulin	…	…	…	…
BMI	…	…	…	…
DiabetesPedigreeFunction	…	…	…	…
Age	…	…	…	…

c. In the table above (Table 7.2 – Model Estimates), report the Estimates (on the odds ratio scale) and the *p* values for each variable in the best-fitting model identified in part b (under the columns for "Part 1"). List each estimate in the appropriate row (you can rearrange the order of the variables on the table if you want). Note, if the full model (Model 3) was your best model, then each variable should have an estimate and p value filled in. If the full model was not the best model, leave the "…" in the table for that row to indicate the variable was not included in the model (do not delete the rows, you will need them later). Please round Estimates to 2 decimal places and round *p* values to 3 decimal places. In the case of *p* values below .001, please report these as < .001. Leave the column for "Hold Out" blank for now, you will fill these out in the next parts. You may include your code and any relevant output in the text boxes below.

```
            exp(coef) [confint] p
(Intercept) 0.00 [0.00, 0.01]   <0.001
Pregnancies 1.14 [1.07, 1.23]   <0.001
Glucose     1.04 [1.03, 1.05]   <0.001
```

```
tableone::ShowRegTable(mod7a_1, exp=TRUE))
```

d. Using the holdout dataset (*diabetes_holdout.csv*, available on Canvas), create a model to predict the likelihood of Outcome using the same parameters from the best fitting model that you identified in part c (i.e., the model that had the lowest BIC). Include all parameters that were in that original model (i.e., if the model with four predictors was the best model, use those same four predictors in this model with the holdout data). Report the Estimates (on the odds ratio scale) and the *p* values for each variable in the model in Table 6.2 above under the columns for "Hold Out". Follow the same guidelines for rounding outlined in part c. You may include your code and any relevant output in the text boxes below.

```
            exp(coef) [confint] p
(Intercept) 0.00 [0.00, 0.01]   <0.001
Pregnancies 1.12 [1.04, 1.20]    0.003
Glucose     1.04 [1.03, 1.05]   <0.001
```

```
mod7d <- glm(Outcome ~ Pregnancies + Glucose, data =
diabetes_test, family = "binomial")

tableone::ShowRegTable(mod7d, exp=TRUE)
```

Answer the following questions in parts f-h to interpret the results of the various analyses done above.

f. Compare how the parameter estimates from the exploratory analyses in Part 1 of the Data (determined in part c) changed in terms of size and significance when the model was repeated in the Hold Out Data (determined in part d). Did the results generally replicate? Do certain predictors stand out as important in both analyses? *(4 points)*

The results obtained in the exploratory analyses replicated in the hold out data. The predictors included in the model (Pregnancies and Glucose) were significant in both models, and their coefficients were generally the same. The only difference was a reduction of 0.02 in the exponentiated coefficient for Pregnancies in the hold out data. Their relative importance was unchanged in both analyses.

g. Comment on which solution (the one from Part 1 of the Data or the one from the Hold Out) you would rather use to make inferences with, and why. *(5 points)*

Since I used the data in Part 1 to derive the best model, it wouldn't be wise to make inferences from its coefficients.
The data used in Part 1 was used to derive the best model. It is not recommended to use their coefficients to make inferences because they are likely positively biased. That is, the coefficient values are likely to be larger due to overfitting. Therefore, it is best to use the estimates from the Hold Out model to make inferences because they provide a more accurate estimate of out-of-sample performance.

8. Multilevel Modeling (40 points)

Use cebu.Rds dataset to answer the following questions on multilevel modeling. To read this data into R, use the readRDS() function.

This data comes from the Cebu Longitudinal Health and Nutrition Survey. Babies were repeatedly assessed every month for twelve months, and the following measures were recorded:

id: Participant (Baby) ID
weightbb: Weight of the baby
HBBC: Height of the baby, centered at the sample mean
feedtype: Categorical variable indicating whether the baby was bottle-fed or breast-fed

a. Run a multilevel model to predict the baby's weight using their height and whether they were bottle-fed or breast-fed (no interaction between them). Include a random intercept and a random slope for the effect of height. Report the summary below. *(5 points)*

```
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 id       (Intercept) 0.2095698 0.45779
          HBBC        0.0008658 0.02942  0.39
 Residual             0.2449261 0.49490
Number of obs: 2491, groups:  id, 208

Fixed effects:
                    Estimate Std. Error t value
(Intercept)         8.034285   0.046024 174.565
HBBC                0.208788   0.002489  83.899
feedtypeBreast Fed -0.016616   0.063323  -0.262

Correlation of Fixed Effects:
            (Intr) HBBC
HBBC         0.231
fdtypBrstFd -0.689 -0.004
```

```
mod8a <- lmer(weightbb ~ HBBC + feedtype + (1 + HBBC|id),
data = cebu)
summary(mod8a)
```

b. Re-run the multilevel model in part (a), this time including an interaction between the baby's height and whether or not they were bottle-fed. (*5 points)*

```
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 id       (Intercept) 0.2073376 0.45534
          HBBC        0.0007709 0.02776  0.38
 Residual             0.2448635 0.49484
Number of obs: 2491, groups:  id, 208

Fixed effects:
                          Estimate Std. Error t value
(Intercept)               8.077191   0.046958 172.007
HBBC                      0.218820   0.003355  65.224
feedtypeBreast Fed       -0.102405   0.066376  -1.543
HBBC:feedtypeBreast Fed -0.020324   0.004788  -4.245
```

```
mod8b <- lmer(weightbb ~ HBBC * feedtype + (1 + HBBC|id),
data = cebu)
summary(mod8b)
```

c. Compare the fit of the two models using a likelihood ratio test. Based on this, report which model is the better fit, and explain why. (Hint: Use the anova() function in R). (*5 points*)

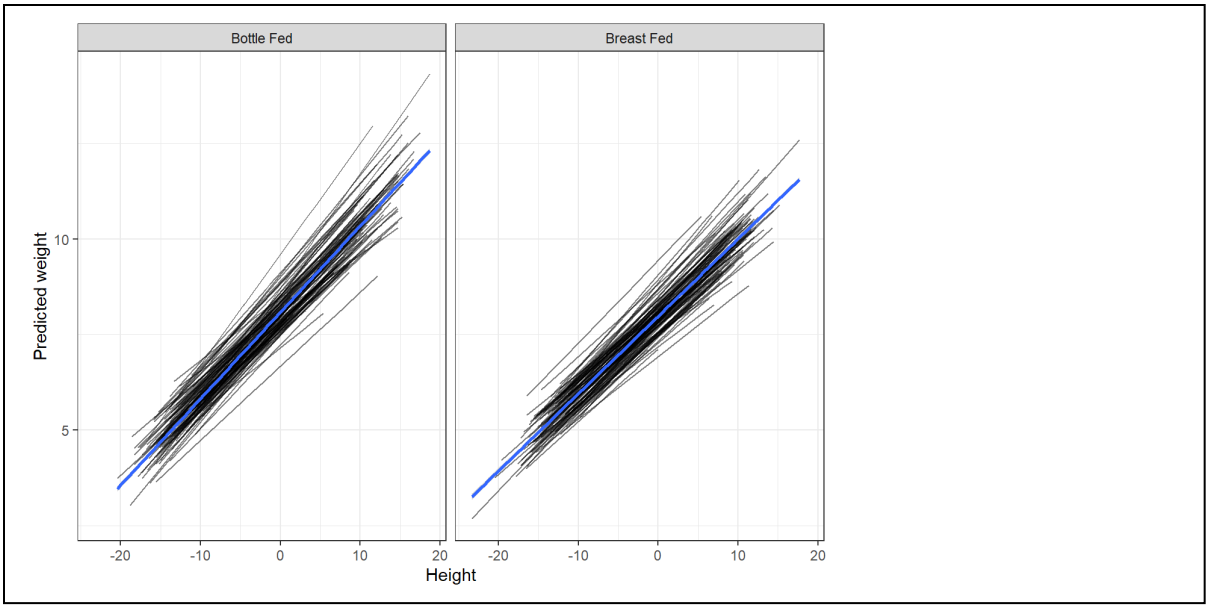
8c)
The model containing an interaction between baby's height and bottle-fed status fit better than the alternative model with only main effects. This is shown by its lower AIC and BIC ($AIC_{model1} = 4295.5, AIC_{model2} = 4280.1,\ BIC_{model1} = 4336.3,\ BIC_{model2} = 4326.7$) and statistically significant chi-square test of the difference ($\chi^2(1) = 17.45, p < 0.001$).

Code:
```
anova(mod8a, mod8b)
```

d. Interpret the fixed effects of the model you chose in part c). If your better-fitting model includes the interaction term, make sure to explain what the interaction means. *(20 points)*

8d)
The fixed effect of height tells us that for every unit increase in height the average change in baby weight is 0.22 units.
The fixed effect of feed type tells us that babies that were breast-fed weighted on average 0.10 units lower than babies that were bottle-fed.
The interaction term captures how the effect of height on weight differs between bottle-fed and breast-fed babies. The negative coefficient of -0.02 suggests that the rate of weight gain with increasing height is lower for breast-fed babies compared to bottle-fed babies.

e. Plot the average trajectory alongside individual baby's predicted trajectories for the model you chose in part d). If the model you chose includes the interaction term, facet your plot by feedtype as well. (*5 points*)



Code:

```
cebu$predicted <- predict(mod8b)

ggplot(cebu, aes(x = HBBC, y = predicted)) +
  geom_line(aes(group = id), alpha = 0.5) +
  geom_smooth(method = "lm") +
  theme_bw() +
  labs(x = "Height", y = "Predicted weight") +
  facet_wrap(~feedtype)
```

9. BONUS QUESTION: Car accidents (10 points)

This is an actual question I obtained two years ago from a former schoolmate. He works in an engineering business and they resurfaced a section on a highway between two cities in Switzerland. Here's the email I got (translated):

Dear Philippe,

….

I have been working in an engineering office for a few years now and mainly deal with risk analysis. Currently I'm working on a security report, which I then have to represent in court, and regarding a statistics question, I am a little uncertain. Since I do not want to embarrass myself in court, I am now looking for your advice.

I'm interested in three questions:

1. Is the accident frequency of sample 2 significantly smaller than that of sample 1?
2. Is the accident frequency on section S1 [a defined section of the road with a given length of some miles] of sample 2 significantly smaller than that of section S1 of sample 1?

3. Is the accident frequency in sections S2 and S5 and S6 and S8 of sample 2 significantly lower than the corresponding sections of sample 1?

The tests can be on a standard alpha level of .05.

What statistical tests would you use and what results do you obtain?

Very best,
Mänu

The file (highway.csv) contains all the information I got from Mänu.

- The month denotes the observation time.
  - Months 1-61 denote the observations before they renewed the surface. Months ranging from 91-105 denote the observations after they renewed the surface.
- S1 to S8 are given sections of the highway, each number denotes a crash (0=no accident, 1=one accident, 2=two accidents...)
- The variable "Sample" indicates the pre (coded as 1) and post (coded as 2) intervention – which can be derived from the month's as well.

Please answer the first point in Mänu's email, with your conclusion and your explanation on what you did and describe your results (hint: the outcome here is **count** of accidents--this is NOT a continuous variable--so you should take this into account when choosing a model).

Is the accident frequency of sample 2 significantly smaller than that of sample 1? Cite the appropriate statistics and p values to support your answer.

Because the number of accidents is count data, the most appropriate distribution for this variable is Poisson. Therefore, we can fit a Poisson regression with the number of accidents across the eight sections of the highway as outcome and the "sample" as predictor. The result of the regression analysis does not allow us to reject the null hypothesis that the accident frequency of sample 2 is lower from accident frequency of sample 1 ($b_{sample2} = -1.37, p = 0.060$).

9a) Output:

```
Call:
glm(formula = accidents ~ sample, family = "poisson", data =
highway_long)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7246     0.1768 -15.413   <2e-16 ***
sample2      -1.3698     0.7289  -1.879   0.0602 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 198.87  on 607  degrees of freedom
Residual deviance: 193.52  on 606  degrees of freedom
AIC: 264.14
```

9a) Code:

```
highway <- readr::read_csv("highway.csv")
highway_long <- highway |>
  tidyr::pivot_longer(cols = c(S1:S8),
                      names_to = "Section",
                      values_to = "accidents") |>
  dplyr::mutate(sample = factor(sample))
summary(glm(accidents ~ sample, family="poisson",
data=highway_long))
```