BTRY 6020: Final Exam

INCLUDE YOUR NAME

5/5/2025

Instructions

You're almost done with the semester! Take a second to congratulate yourself on getting here. As a reminder, this final project is simply an (imperfect) way of measuring what you have learned throughout the semester. So take a deep breath and do your best, but also remember that it doesn't determine your value as a human being.

The exam is split into 4 sections: Module 1, 2 and 3 (6 questions), Modules 4 and 5 (3 questions), Module 6 (2 questions) and the final project. Most of the questions on this exam are short answers. You don't need to write out an overly long response (a sentence or so for each part of the question should be fine), but you should be specific in explaining your response. For example, if there is a question about whether the assumptions are reasonable. You shouldn't just say "from the plot we can see that the linearity assumption is (or is not) reasonable," but instead you should explain specifically why the plot leads you to believe the linearity assumption is (or is not) reasonable.

The exam is open notes so you **can** use any of the material or any of the notes you have taken throughout the class. You **cannot** discuss the exam (while it is in progress) with anyone else. You also **cannot** use any generative AI tools. Submissions will be sent by e-mail to **nbb45@cornell.edu** before **May 14th 11:59pm**.

Module 1, 2, and 3

In the questions for Modules 1, 2, and 3, we will look at data from SNCF, France's national railway. The data has been cleaned and made easily available by TidyTuesday. In particular, we have data on train delays from each month between 2015-2018 for each train route (i.e., from city A to city B). So each observation (i.e., row in the data) corresponds to a specific route in a specific year and month. In the dataset, we will be particularly interested in the following variables

For each row in the data, we have the following variables

- year : year of observation (2015, 2016, 2017 or 2018)
- month : month of observation (1, 2, . . . , 12)
- departure_station : station where the route begins (e.g., "PARIS NORD" or "MONTPELLIER")
- arrival_station : station where the route ends (e.g., "PARIS NORD" or "MONTPELLIER")
- journey_time_avg : average journey time in minutes for the route for that year and month
- avg_delay_all_departing : average delay in minutes for all departures for the route for that year and month (i.e., how many minutes the train was late to leave departure station)
- avg_delay_all_arriving : average delay in minutes for all arrivals for the route for that year and month (i.e., how many minutes the train was late to arrive at the arrival_station)

In the following questions, the model you fit or consider may change from question to question.

```
## Load in data and remove some outliers
train_data <- read.csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/
# removing some outliers
train_data <- train_data[-which(train_data$avg_delay_all_arriving < -30),]
train_data <- train_data[-which(train_data$avg_delay_all_departing > 30),]
# make month and year factors
train_data$month <- as.factor(train_data$month)
train_data$year <- as.factor(train_data$year)
```

Question 1 (2 pts)

Suppose we are interested in modeling the average delayed arrival; i.e., avg_delay_all_arriving is the outcome variable. Specifically, we would like to investigate the association between average delayed arrival and journey time (journey_time_avg) when controlling for the average departure delay (avg_delay_all_departing).

Fit the relevant linear model below and write 1 sentence interpreting the estimated coefficient for journey_time_avg.

Question 1 Answer

Question 2 (2 pts)

Some output for a **different model** is shown below. Using the output, predict the average arrival delay for a train route which has an average journey time of 200 minutes, has an average departure delay of 3 minutes, and took place in January (i.e., month == 1).

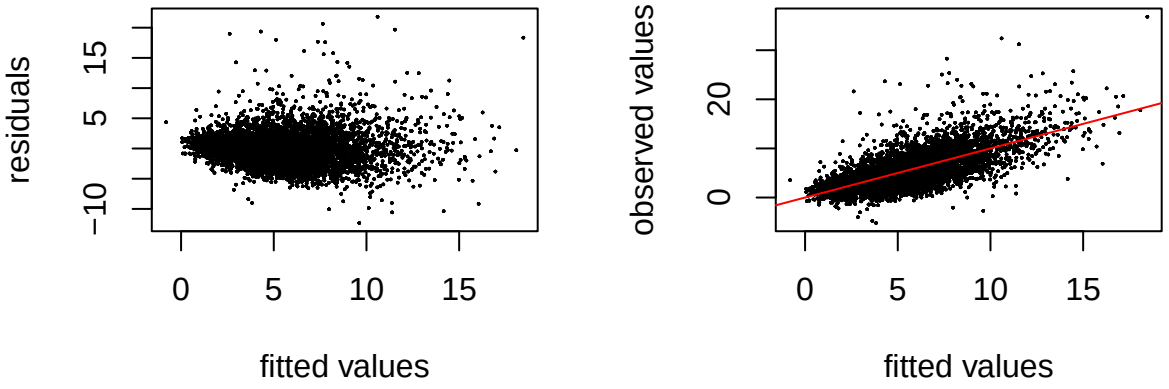
```
##                            Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)             -0.89153617 0.0625821018 -14.245865  6.529044e-46
## journey_time_avg         0.02215535 0.0001925167 115.082758  0.000000e+00
## avg_delay_all_departing  0.79854766 0.0067729269 117.902891  0.000000e+00
## month2                   0.45637989 0.0735194856   6.207605  5.444405e-10
## month3                  -0.47362582 0.0734957887  -6.444258  1.177825e-10
## month4                  -0.08049415 0.0735030454  -1.095113  2.734751e-01
## month5                   0.32511688 0.0735458330   4.420602  9.874291e-06
## month6                   1.86150670 0.0739347360  25.177701 1.475638e-138
```

```
## month7                 1.94714571 0.0742472465  26.225157 4.916823e-150
## month8                 0.75296105 0.0737335963  10.211913  1.908528e-24
## month9                 0.61577106 0.0735351836   8.373829  5.797172e-17
## month10                0.93242762 0.0734938111  12.687158  8.508994e-37
## month11                1.29220290 0.0736160461  17.553278  1.160711e-68
## month12                0.13960420 0.0810379232   1.722702  8.495186e-02
```

Question 2 Answer

Question 3 (6 pts)

Do the assumptions for linear regression seem reasonable for the model fit in Question 2? Explain why or why not? You should use the plots below to justify your answer.



Question 3 Answer

Question 4 (2 pts)

Suppose you think the association between arrival delay and journey time (i.e., the slope of journey time) may change from year to year. Fit a linear model below which would allow for that. For this problem, you **do not** need to consider adjusting for other variables in the model.

Question 4 Answer

Question 5 (3 pts)

Below, we fit a model which includes the covariates journey time, average departing delay and month. Suppose we want to test if the average arrival delay is associated with month after adjusting for journey time and average departure delay. For this problem, you don't need to consider interaction terms and you don't need to include other covariates. Describe how you would test this hypothesis. You don't need to actually perform any calculations or write any code, but specify which function in R you would use and be specific about what the inputs would be.

```
mod_year <- lm(avg_delay_all_arriving ~ journey_time_avg + avg_delay_all_departing + month,
               data = train_data)
summary(mod_year)
```

```
##
## Call:
## lm(formula = avg_delay_all_arriving ~ journey_time_avg + avg_delay_all_departing +
##     month, data = train_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3284  -1.6465  -0.1308   1.3595  21.8094
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.8915362  0.0625821 -14.246  < 2e-16 ***
## journey_time_avg          0.0221554  0.0001925 115.083  < 2e-16 ***
## avg_delay_all_departing   0.7985477  0.0067729 117.903  < 2e-16 ***
## month2                    0.4563799  0.0735195   6.208 5.44e-10 ***
## month3                   -0.4736258  0.0734958  -6.444 1.18e-10 ***
## month4                   -0.0804941  0.0735030  -1.095    0.273
## month5                    0.3251169  0.0735458   4.421 9.87e-06 ***
## month6                    1.8615067  0.0739347  25.178  < 2e-16 ***
## month7                    1.9471457  0.0742472  26.225  < 2e-16 ***
## month8                    0.7529611  0.0737336  10.212  < 2e-16 ***
## month9                    0.6157711  0.0735352   8.374  < 2e-16 ***
## month10                   0.9324276  0.0734938  12.687  < 2e-16 ***
## month11                   1.2922029  0.0736160  17.553  < 2e-16 ***
## month12                   0.1396042  0.0810379   1.723    0.085 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.748 on 32686 degrees of freedom
## Multiple R-squared:  0.4892, Adjusted R-squared:  0.489
## F-statistic:  2408 on 13 and 32686 DF,  p-value: < 2.2e-16
```

Question 5 answer

Question 6 (2 pt)

Suppose we fit the model below where we have used the log of journey_time_avg. Write 1 sentence interpreting the coefficient for journey time.

```
mod_log <- lm(avg_delay_all_arriving ~ log(journey_time_avg),
              data = train_data)
summary(mod_log)
```

```
##
## Call:
## lm(formula = avg_delay_all_arriving ~ log(journey_time_avg),
##     data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8915  -2.2847  -0.4961   1.5921  29.8838
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -11.06500    0.19401  -57.03   <2e-16 ***
## log(journey_time_avg)   3.29684    0.03868   85.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.477 on 32698 degrees of freedom
```

```
## Multiple R-squared:  0.1818, Adjusted R-squared:  0.1817
## F-statistic:  7264 on 1 and 32698 DF,  p-value: < 2.2e-16
```

Question 6 answer

Module 4 and 5

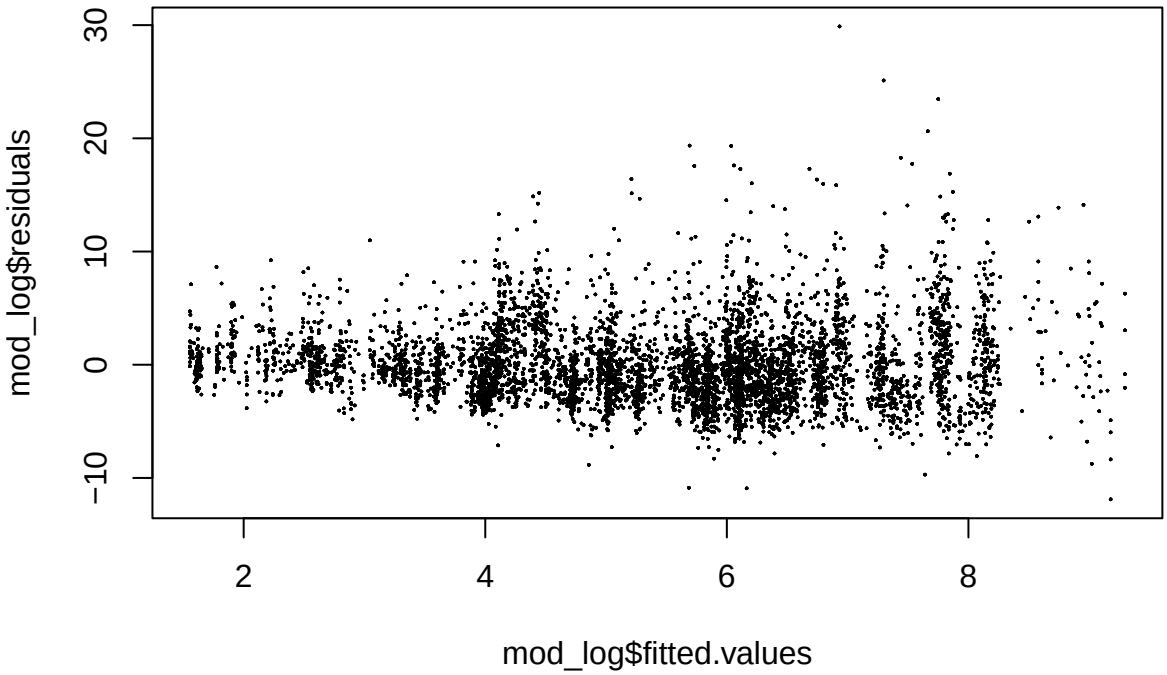
Question 7 (3 pts)

In the model you fit in Question 1, each observation in the dataset corresponds to a specific route observed in a specific month and year. Thus each route appears in the data multiple times. Explain why this might violate an assumption for linear regression. How could you fix this? If your suggestion involves additional covariates or a different modeling assumption, be specific about what you mean (i.e., say what covariates would you include, or what model you would fit). There is more than 1 reasonable answer for this question, but just pick one.

Question 7 answer

Question 8 (3 pts)

Using the model from Question 5, we plot the fitted values vs the residuals below. Explain why you might want to use robust standard errors. What might be the advantages and disadvantages of using the robust standard errors as opposed to the model based errors (the ones that come out of `summary`)?



Question 8 answer

Question 9 (3 pts)

Suppose you are taking a train tomorrow from Lille to Paris Nord and want to predict the delay in arrival. You want to be very sure about the prediction, so you gather data for 1000 different variables you think might be relevant (temperature, whether it is raining, GDP of France per month/year, the win/loss record of the soccer team in Lille, etc). You then regress average arrival delay onto all of those variables, and use it to

predict the arrival delay for tomorrow's train. Explain why this might not give a good prediction. What might you do instead? 2-3 sentences for this answer is fine.

Question 9 answer

Module 6

For the following questions, suppose we are analyzing data for Big Red Airlines, Cornell's latest idea for getting people to and from Ithaca. The dependent variable is whether or not a flight took off on time. In the `OnTime` variable: 1 indicates that the flight took off on time, 0 indicates that it was delayed. The covariates we have recorded include Temperature (in degrees), TimeOfDay (Evening, Midday, Morning), and Rain (FALSE, TRUE).

```
airlineData <- read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lab11/airline.csv
names(airlineData)
```

```
## [1] "OnTime"      "Temperature" "TimeOfDay"   "Rain"
```

Question 10 (2 pts)

What is the appropriate type of regression for modeling the binary data? What is being predicted by the linear model we are fitting? i.e., if the model we set up is

$$? = b_0 + b_1 X_{1,i} + b_2 X_{2,i} \dots$$

what is on the left side of the equation (you can write it out in words instead of typing out the math)?.

Question 10 answer

Question 11 (2 pts)

We fit the model below. How would you interpret the coefficient associated with `Temperature`?

```
mod <- glm(OnTime ~ Temperature + TimeOfDay + Rain,
           data = airlineData, family = "binomial")
summary(mod)
```

```
##
## Call:
## glm(formula = OnTime ~ Temperature + TimeOfDay + Rain, family = "binomial",
##     data = airlineData)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.46094    0.37880   3.857 0.000115 ***
## Temperature      -0.05248    0.00652  -8.050 8.29e-16 ***
## TimeOfDayMidday   0.18066    0.22978   0.786 0.431717
## TimeOfDayMorning -0.59611    0.25310  -2.355 0.018511 *
## RainTRUE         -0.52725    0.19977  -2.639 0.008308 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 765.46  on 892  degrees of freedom
```

```
## Residual deviance: 671.25  on 888  degrees of freedom
## AIC: 681.25
##
## Number of Fisher Scoring iterations: 5
```

Question 11 answer

Final Project (30 pts)

Introduction

This final project is designed to demonstrate your mastery of linear regression techniques on real-world data. You will apply the theoretical concepts we've covered in class to a dataset of your choice, perform a comprehensive analysis, and present your findings in a professional format suitable for showcasing to potential employers.

Objectives

By completing this project, you will:

- Apply linear regression techniques to solve real-world problems
- Demonstrate your ability to verify and address regression assumptions
- Perform meaningful feature selection and hypothesis testing
- Communicate the practical significance of your statistical findings
- Create a professional portfolio piece for future employment opportunities

Project Requirements

Dataset Selection

1. Choose a dataset from Kaggle
2. Your dataset must have a continuous target variable suitable for linear regression
3. The dataset should contain multiple potential predictor variables
4. Choose a dataset that interests you and has meaningful real-world applications

Analysis Requirements

Your analysis must include the following components:

Exploratory Data Analysis

- Summary statistics of variables
- Visualization of distributions and relationships
- Identification of missing values and outliers
- Data cleaning and preprocessing steps

Regression Assumptions Verification

- Linearity assessment
- Normality of residuals
- Homoscedasticity (constant variance of residuals)
- Independence of observations
- Multicollinearity assessment

Assumption Violation Handling

- Apply appropriate transformations when assumptions are violated
- Document your approach to each violation
- Compare models before and after corrections

Variable Selection & Hypothesis Testing

- Implement at least two different variable selection techniques
- Perform hypothesis tests on coefficients

- Assess model performance with metrics (R^2, adjusted R^2, RMSE, etc.)
- Validate your model using appropriate cross-validation techniques

Feature Impact Analysis

- Quantify and interpret the impact of each feature on the target
- Provide confidence intervals for significant coefficients
- Explain the practical significance of your findings in the context of the dataset

Deliverables GitHub Repository containing:

- All code (well-documented Rmd files)
- README.md with clear instructions on how to run your analysis
- Data folder (or instructions for accessing the data)
- Requirements.txt or environment.yml file

Final Report (PDF) containing:

- Introduction: dataset description and problem statement
- Methodology: techniques used and justification
- Results: findings from your analysis
- Discussion: interpretation of results and limitations
- Conclusion: summary and potential future work
- References: cite all sources used

Evaluation Criteria

Your project will be evaluated based on:

- Correctness of statistical analysis and procedures
- Proper handling of regression assumptions
- Quality of variable selection and hypothesis testing
- Clarity of interpretation and insights
- Organization and documentation of code
- Professional presentation of findings

Timeline and Submission

- Release Date: May 5th, 2025
- Due Date: Wednesday, May 14th, 2025 (11:59 PM EST)
- Submission: Email your GitHub repository link and PDF report to nbb45@cornell.edu with the subject line "Final Project - [Your Name]"

Resources

- Course materials and lecture notes
- Kaggle Datasets
- GitHub tutorial and GitHub documentation for repository setup.

Academic Integrity

This is an individual project. While you may discuss general concepts with classmates, all submitted work must be your own. Proper citation is required for any external resources used.

Good luck with your project! This is an opportunity to demonstrate your skills and create a valuable addition to your professional portfolio.

Finished

You're done, congratulations!