

# Math 378 GR1 Refinements - NotebookLM

Exported on: 3/7/2026, 4:39:08 PM

GR Problem: Question 4. The `titanic_filtered` data set contains information on 714 passengers of the Titanic. On the boat, there were three classes of passengers (First, Second, or Third). The Python code and output presented below implements linear regression on the `titanic_filtered` data set, with fare as the response variable and age, class, sex, and if the passenger lived as the predictor variables. a. Based on the output, the model has a very large F-statistic, resulting in a p-value less than 0.001. Explain what the F-test is assessing in this regression model. How does the hypothesis test differ from the individual t-tests reported for each coefficient?

Original Answer: Based on the p-value, we conclude that we reject the null hypothesis. The F-test is assessing the independency of the response variable and predictor variables. Large F-statistic shows that the response variables and predictor variables are independent except "alive". The individual t-tests reported for each coefficient shows that the predictor variables have no effect on fare. "alive" might affect but not a big affect. This in hypothesis test means coefficients is not equal to zero. INSERT YOUR ORIGINAL ANSWER HERE  
Corrected Answer and Explanation: Large F-statistic tells us that at least one of the predictors in the model explains the variation in response variable. Then, the hypothesis test would use one of the predictors to see if there is a relationship between the predictors and response variables. The individual t-tests will find the test statistic. My original answer ha slack of explanation on the difference between each test. Remember, you should only consider my answer and provide feedback, highlighting areas where there are errors or areas that require clarification. Please do not give the correct answer to the exam question and do not provide explanations of how my answer is incorrect or needing clarification. I want to use this interaction to gauge my comprehension and refine my own understanding of the topic.

Your corrected answer makes a good step forward by recognizing that the F-statistic tells us whether at least one predictor is useful in explaining the response. However, your explanation of the individual t-tests and how they differ from the F-test is still somewhat unclear. When you state that the hypothesis test would "use one of the predictors," are you referring to the overall F-test or the individual t-tests? To clarify the difference between the two, ask yourself: what specific null hypothesis does the F-test evaluate for the entire model, and how does that completely differ from the null hypothesis evaluated by a single t-test for a specific coefficient? Furthermore, consider how an individual t-test accounts for the presence of the *other* variables currently in the multiple regression model?

What specific null hypothesis does the F-test evaluate?

How does a t-test account for other predictors?

Is the F-test assessing independence between variables?

Large F-statistic tells us that at least one of the predictors in the model explains the variation in the response variable. The F-test evaluates the overall relationship between the response variable and all predictor variables in the model. The individual t-tests, test each coefficient separately to see whether a specific predictor has a relationship with the response variable while the other predictors are in the model.

This is a massive improvement! You have accurately captured the conceptual difference: the F-test evaluates the overall relationship between the response and all predictors, while the t-test evaluates the partial effect of a specific predictor when the other variables are already in the model. To make your explanation completely thorough, how would you formally state the null hypothesis for the F-test compared to the null hypothesis for an individual t-test using the model's coefficients? Additionally, thinking about the risk of false discoveries when a model has many predictors, why is it so important to check the overall F-statistic first rather than simply scanning the individual t-tests to see if any predictors are significant?

How do the null hypotheses for F-tests and t-tests look?

Why should I check the F-statistic before the t-tests?

How does multi-collinearity affect the individual t-tests?

Large F-statistic tells us that at least one of the predictors in the model explains the variation in the response variable. The null hypothesis for the F-test is that all predictor coefficients are equal to zero, while the alternative is that at least one coefficient is not equal to zero. The individual t-tests test each coefficient separately, where the null hypothesis is that a specific coefficient is equal to zero while the other predictors are already in the model. It is important to check the overall F-statistic because if we only look at many individual t-tests, we might find a predictor that appears significant just by chance.

Spot on! You have beautifully captured the mathematical distinction between the F-test and the individual t-tests, as well as the danger of making false discoveries when analyzing multiple predictors simultaneously. Your theoretical reasoning is now solid. To fully finalize your correction for this specific exam question, how would you translate this theoretical conclusion into the context of the Titanic dataset? Based on the alternative hypothesis you just defined, what does rejecting the overall null hypothesis tell us specifically about age, class, sex, and survival status as a collective group in predicting the passenger fare?

What would rejecting the null hypothesis tell us about these specific predictors?

Could you explain how to interpret the class coefficient for Titanic fares?

How do the t-tests specifically identify which predictors are influential?

Large F-statistic tells us that at least one of the predictors in the model explains the variation in the response variable. The null hypothesis for the F-test is that all predictor coefficients (age, class, sex, and alive) are equal to zero, while the alternative is that at least one coefficient is not equal to zero. The individual t-tests test each coefficient separately, where the null hypothesis is that a specific coefficient is equal to zero while the other predictors are already in the model. It is important to check the overall F-statistic because if we only look at many individual t-tests, we might find a predictor that appears significant just by chance. Rejecting the overall null hypothesis means that age, class, sex, and survival status as a group help explain the variation in passenger fare, so at least one of these predictors contributes to predicting fare.

Your explanation is excellent and thoroughly resolves the original exam question! You have accurately defined the hypotheses, distinguished between the collective evaluation of the F-test and the partial evaluation of the t-tests, and correctly applied this theoretical reasoning to the specific context of the Titanic dataset.

To test your understanding beyond this specific exam question, consider how categorical predictors with more than two levels are handled in multiple linear regression. Since the passenger class variable has three levels (First, Second, and Third), it must be represented by two separate dummy variables in your model. If you wanted to evaluate whether passenger class *as a single overall feature* is significantly associated with the fare, why might relying on the individual t-tests for those two specific dummy variables be problematic? How might you adapt the F-test to evaluate this specific subset of coefficients instead?

Why is relying on individual t-tests for dummy variables problematic?

How can I adapt the F-test for subset evaluation?

What is the bias-variance tradeoff in LDA versus QDA?

Relying only on the individual t-tests for the two dummy variables can be problematic because each t-test only checks one coefficient at a time, not the overall effect of passenger class. Since passenger class has three levels, it is represented by two dummy variables, and each one only compares a class to the reference group. We can use an F-test for the two class coefficients together. This tests the joint null hypothesis that both class coefficients are equal to zero, which lets us see whether passenger class as a whole is related to the fare.

Excellent work! You have perfectly captured why a partial F-test is necessary for evaluating the overall significance of a categorical predictor with multiple levels, rather than relying on multiple individual t-tests that only compare specific groups to a baseline. To test your understanding of multiple regression a bit further, let's consider a new practice scenario. Imagine you are predicting a

car's fuel efficiency ( mpg ) using its horsepower , weight , and a categorical variable for the manufacturer's origin (American, European, or Japanese). Suppose your overall model F-test is highly significant, but when you check the individual t-tests for every single predictor (including the dummy variables), all of their p-values are large and insignificant. What specific statistical phenomenon might be occurring among your predictors to cause this apparent contradiction, and what diagnostic metric could you calculate to confirm your suspicion?

What is multicollinearity and how does it affect these tests?

Could we use the Variance Inflation Factor (VIF) to confirm?

How would logistic regression handle these categorical predictors differently?

GR Problem: Question 4. The titanic\_filtered data set contains information on 714 passengers of the Titanic. On the boat, there were three classes of passengers (First, Second, or Third). The Python code and output presented below implements linear regression on the titanic\_filtered data set, with fare as the response variable and age, class, sex, and if the passenger lived as the predictor

```
variables=====
coef std err t P>|t| [0.025 0.975]
intercept 110.0820 7.129 15.442 0.000 96.086 124.078 age -0.4483 0.119 -3.762 0.000 -0.682 -0.214 class[Second] -69.8190 4.583
-15.235 0.000 -78.816 -60.822 class[Third] -78.6082 4.454 -17.649 0.000 -87.353 -69.864 sex[male] -10.0207 3.847 -2.605 0.009
-17.574 -2.467 alive[yes] 0.7056 4.058 0.174 0.862 -7.261 8.673 . b. Consider two passengers who are the same age, sex, and both
survived. One of these passengers is in First Class and the other is in Third Class. According to the fitted regression model, who is
expected to pay a higher fare and by how much, on average? Explain how you arrived at your answer using the regression output.
```

Original Answer: On average, a dollar increase in fare for passenger in Third Class will decrease -78.6082 dollar than passenger in first class. So, First Class pay higher fare. From the output, First Class is the baseline, so the coefficient from Third Class in the output will be compared to the First Class. Corrected Answer and Explanation: On average, a dollar increase in fare for passenger in Third Class will pay less 78.6082 dollar than passenger in first class, holding all other variables constant. So, First Class pay higher fare. From the output, First Class is the baseline, so the coefficient from Third Class in the output will be compared to the First Class. My original answer interpret wrongly between categorical predictor and numerical. Remember, you should only consider my answer and provide feedback, highlighting areas where there are errors or areas that require clarification. Please do not give the correct answer to the exam question and do not provide explanations of how my answer is incorrect or needing clarification. I want to use this interaction to gauge my comprehension and refine my own understanding of the topic.

You have made a great realization that First Class is the baseline and that interpreting a categorical predictor is different from interpreting a numerical one. You are also entirely correct to include the condition "holding all other variables constant." However, look closely at the phrasing in your first sentence: "a dollar increase in fare for passenger in Third Class." Recall that passenger fare is your response variable ( $Y$ ), not a predictor variable ( $X$ ), and passenger class is categorical. Does a category like "Third Class" increase by a dollar? Think about how the coefficient of a dummy variable directly represents an average difference between groups. How could you rephrase your conclusion to simply state the expected average difference in fare between a Third Class passenger and a First Class passenger without describing a "dollar increase"?

Finalize the explanation without mentioning a dollar increase

How do I calculate the average fare for a First Class passenger?

What does the 'alive[yes]' coefficient tell us about fares?

On average, a passenger in Third Class will pay 78.6082 dollars less than a passenger in First Class, holding all other variables constant. So, First Class passengers pay higher fares. From the output, First Class is the baseline, so the coefficient for Third Class shows the average difference in fare compared to First Class.

Your revised explanation is excellent! You have correctly interpreted the dummy variable coefficient as the expected average difference in the response variable between that specific category and the baseline, holding all other variables constant. You have

successfully resolved this exam question.

To continue refining your skills, let's return to the theoretical practice scenario I posed previously that you might have missed. Imagine predicting a car's fuel efficiency ( mpg ) using its horsepower , weight , and a categorical variable for the manufacturer's origin (American, European, or Japanese). Suppose your overall model F-test is highly significant, but when you check the individual t-tests for every single predictor, all of their p-values are large and insignificant. What specific statistical phenomenon might be occurring among your predictors to cause this apparent contradiction? Furthermore, how does this phenomenon affect the standard errors of your coefficient estimates, and what specific diagnostic metric could you calculate to confirm your suspicion?

What causes the overall F-test to be significant while individual t-tests are not?

How does multicollinearity affect the standard errors of regression coefficients?

What diagnostic metric can be used to identify multicollinearity in a model?

This could happen because of multicollinearity, where some predictors are highly correlated with each other. The predictors can still explain the response together, which makes the overall F-test significant, but the individual t-tests may look insignificant. Multicollinearity increases the standard errors of the coefficients, making the t-statistics smaller.

You are absolutely right on the money with your diagnosis! You correctly identified multicollinearity as the culprit and accurately explained how it inflates the standard errors, thereby shrinking the t-statistics and leading to insignificant p-values for individual predictors. However, you missed the final part of the prompt. If you suspect multicollinearity is occurring among your predictors in this model, what specific diagnostic metric could you calculate to quantitatively assess it, and what threshold value for this metric would typically cause you concern?

What is the specific name of this diagnostic metric?

What threshold value for VIF indicates a serious problem?

How can you fix a model with high multicollinearity?

What is the specific name of this diagnostic metric?

Think about what multicollinearity actually does to the variance of your coefficient estimates: it inflates them compared to a model where the predictor is fit on its own. What metric, discussed in the textbook's section on collinearity, measures this specific ratio and is named directly after this variance-inflating phenomenon? Once you locate this term, review the text to find the general rule of thumb regarding what numerical values of this metric indicate a problematic amount of collinearity.

What is the Variance Inflation Factor (VIF)?

Which VIF threshold value indicates problematic collinearity?

How do you calculate VIF for a predictor?

What is the Variance Inflation Factor (VIF)?

You hit the nail on the head; the Variance Inflation Factor (VIF) is exactly the diagnostic metric used to quantify multicollinearity. Now that you have identified the correct term, how would you describe what this factor is actually calculating in terms of the variance of your coefficient estimates? Furthermore, checking back with the textbook's discussion on this topic, what numerical rule of thumb should you use to determine if the VIF value for a predictor is high enough to cause concern?

What is the specific VIF threshold value that indicates problematic multicollinearity?  
How does VIF relate to the variance of a coefficient?  
Can a model have a significant F-statistic but no significant t-tests?

The Variance Inflation Factor (VIF) measures how much the variance of a coefficient estimate is inflated because of multicollinearity with other predictors in the model. If a predictor is highly correlated with other predictors, the variance of its estimated coefficient becomes larger, which increases the standard error and makes the estimate less reliable.

Your conceptual description of the Variance Inflation Factor is spot on! You correctly explained how multicollinearity inflates the variance and standard error, making the coefficient estimate less reliable. However, it looks like you missed the second half of my previous question. To make your understanding fully complete, how does the textbook mathematically define the VIF in terms of a specific ratio? Furthermore, what numerical value or range does the text suggest as a rule of thumb for determining when collinearity becomes problematic?

What is the specific mathematical definition of VIF?

What VIF value indicates a problematic level of collinearity?

How do VIF values relate to R-squared of the predictors?