Question: Please explain how much your race affects your political views.

# Methodology

## 1.1 Data Cleaning

* **Tokenization**: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
* Words that have fewer than 3 characters are removed.
* All stopwords are removed using a regular stopwords sets[[1]](#footnote-0)
* Words are lemmatized — words in third person are changed to first person and verbs in past and future tenses are changed into present.
* Words are stemmed — words are reduced to their root form.
* After experimenting both BTM and LDA on the string dataset, both results showed that “race”, “black”, and “white” three words occupied every topic, which created extra difficulty in tagging, interpreting, and further regression analysis. To simplify the interpretation, I added these three words into stopwords.
* I tried several different combinations of stop words and generated different testing sets. The size of these testing sets are:
  + Original size: 8263
  + After removing regular stopwords: 3095
  + After removing regular stopwords and “Race”: 3082
  + After removing regular stopwords and “Race”, “White”, and “Black”: 3053
    - The 3053 dataset generated most of the significant results, which is consistent with above hypotheses.

## 1.2 Topic Model Methods

* LDA: A generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.
  + I did several rounds of LDA on the dataset. However, because of the data striction (most answers are short texts), the fitness of classification was not very good. *So we paid more attention to BTM at this stage.*
* BTM: A word co-occurrence based topic model that learns topics by modeling word-word cooccurrences patterns which are called biterms. This in contrast to traditional topic models like LDA which are word-document co-occurrence topic models. A biterm consists of two words co-occurring in the same short text window. This context window can for example be a twitter message, a short answer on a survey, a sentence of a text or a document identifier. The techniques are explained in detail in the paper *'A Biterm Topic Model For Short Text' by Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng (2013).*

## 1.3 Regression Analysis

* The regression analysis was mainly based on the BTM topic tagging
* There are two groups of topics that we cared the most in BTM tagging results:

| Group 1 | Topic 1 | Unfriendly Law |
| --- | --- | --- |
| Topic 2 | Social Inequality  Related to Race |
| Group 2 | Topic 4 | No Effect |
| Topic 5 | Race Should Not Matter |

* In the BTM result, every quote was labeled as different percentages of the six topics. So I created a series variable referring to “topic 1 or topic 2 being the highest related topic”, “topic 4 or topic 5 being the highest related topic” and so on for each quote. And I then regress these variables on the covariates we have in the original dataset.

# Result

## 2.1 LDA

* Word-topic relationship table
  + drawn from this question data by the LDA model. LDA identifies these six topics that could be understood as six distributions consisting of these corresponding words.[[2]](#footnote-1)
  + The rank of these topics are based on their contribution to the whole answer set. That is to say, the topic 1 is the most important one, and then topic 2, and so on.

|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Topic 1 | race | political | think | matter | black | prison | white | vote | racial | nothing |
| Topic 2 | dont | white | like | see | racial | know | prison | opinion | race | black |
| Topic 3 | race | white | much | politics | political | prison | one | believe | somewhat | see |
| Topic 4 | deal | great | race | racial | different | political | prison | get | everyone | way |
| Topic 5 | sure | race | racist | prison | one | know | black | question | think | politics |
| Topic 6 | color | right | know | feel | well | wrong | race | one | dont | man |

* Topic Interpretation and Contribution Table
  + The dominance percentage.
    - One answer could be made up of many topics. But there is usually a dominant topic. This index could basically tell us in what percentage of answers, the corresponding topic is the dominant topic.
  + The average contribution tells us, on average, how much does this topic contribute to the answers’ meaning.
  + The interpretation was conducted in a common-used “put-back and interpret” method.
    - Go back to the original string and translate the tag assigned. All the strings sharing the same tag are put together.

| Dominant Topic | Interpretation | Dominance Percentage | mean Contribution |
| --- | --- | --- | --- |
| Topic 1 | “Race Does Affect” | 0.22578693 | 0.59464665 |
| Topic 2 | Prison Experience | 0.12590799 | 0.79955337 |
| Topic 3 | Politics | 0.1722155 | 0.736013 |
| Topic 4 | “Race influence everything else” | 0.16404358 | 0.77916679 |
| Topic 5 | Racist Experience | 0.13559322 | 0.75656004 |
| Topic 6 | Interracial Interaction | 0.17645279 | 0.7849319 |

* Topic Dominance Among Different Racial Group
  + We could see that Topic 1 is the most important topic for almost every ethnicity group.
    - Topic 3 is the second important for black, white, Asian, Hwaiian, Latinx, Other and Not-sure people;
    - Again, the distribution of different topic in different ethnic groups are very similar, which may be the result of LDA not matching the data structure like mentioned above.

|  | Black | White | Native | Asian | Hawaiian | Latinx | Other | Not Sure |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Topic 1 | **0.22006** | **0.21447** | **0.27907** | **0.29167** | **0.20833** | **0.26984** | **0.19249** | **0.31579** |
| Topic 2 | 0.15042 | 0.12895 | 0.12558 | 0.125 | 0.125 | 0.10053 | 0.14554 | 0.18421 |
| Topic 3 | 0.17549 | 0.19079 | 0.14419 | 0.20833 | 0.25 | 0.21693 | 0.19249 | 0.10526 |
| Topic 4 | 0.15877 | 0.16711 | 0.15814 | 0.08333 | 0.125 | 0.13757 | 0.14085 | 0.15789 |
| Topic 5 | 0.12256 | 0.15 | 0.14419 | 0.125 | 0.125 | 0.13757 | 0.14085 | 0.13158 |
| Topic 6 | 0.1727 | 0.14868 | 0.14884 | 0.16667 | 0.16667 | 0.13757 | 0.18779 | 0.10526 |

## 2.2 BTM

I trained the BTM model on all three datasets.[[3]](#footnote-2) But the 3053 one always has the most significant results. The significant results generated from the other two datasets are mostly consistent with the 3053 one’s results. So here in this memo, I will only present the S3053’s result.

* Topic Word Table

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Topic 1 | never | political | see | know | even | need | attention | going | dont | since |
| Topic 2 | political | want | dont | never | get | care | know | country | really | world |
| Topic 3 | see | get | life | country | change | justice | never | want | one | vote |
| Topic 4 | see | get | vote | political | think | care | dont | help | change | want |
| Topic 5 | know | get | need | see | make | help | country | law | really | feel |
| Topic 6 | going | country | dont | political | vote | make | think | world | get | see |
| Topic 7 | political | never | life | see | news | good | get | pay | para | much |

* Topic Interpretation and Contribution Table

| **Topics** | **Dominance Percentage** | **Interpretation** |
| --- | --- | --- |
| Topic 1 | 0.173163 | Unfriendly Law |
| Topic 2 | 0.169515 | Social Inequality  Related to Race |
| Topic 3 | 0.169326 | Peer Impact |
| Topic 4 | 0.163664 | No Effect |
| Topic 5 | 0.162597 | Race Should Not Matter |
| Topic 6 | 0.161735 | Race Related News |

* Topic Dominance Among Different Racial Group
  + I used a chi-square test on the second important topics for different racial groups and it turns out to be significantly different.

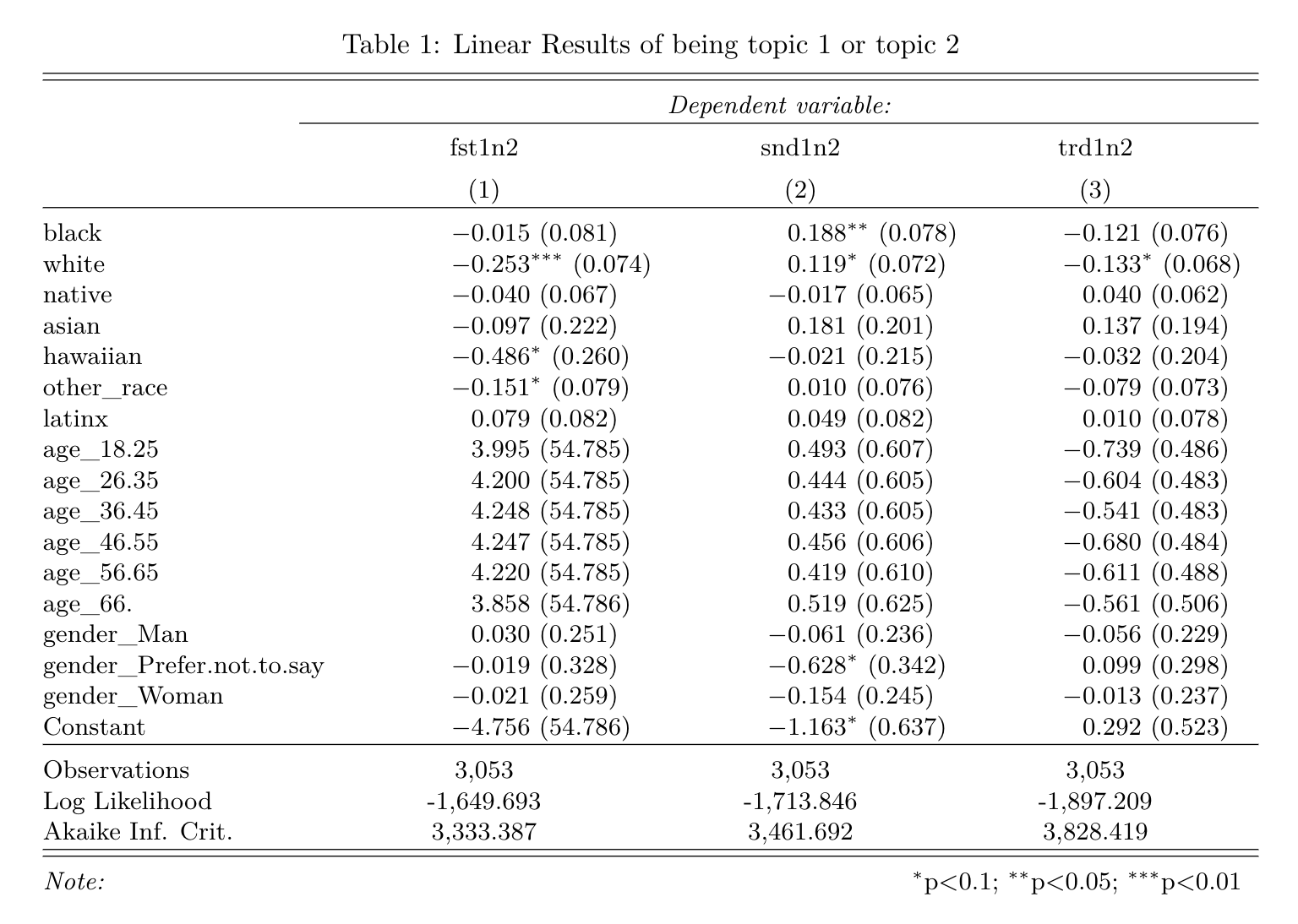
# Regression Analysis

Since there will be the issue of perfect multicollinearity, all regression models below excluded some variables to clear perfect multicollinearity. The variables excluded for perfect multicollinearity include "age\_Under 18" and "gender\_Gender non-conforming or non-binary or other".

## 3.1 “topic 1 or topic 2 being the most related topic” as dependent variables

### 3.1.1 Keep all variables

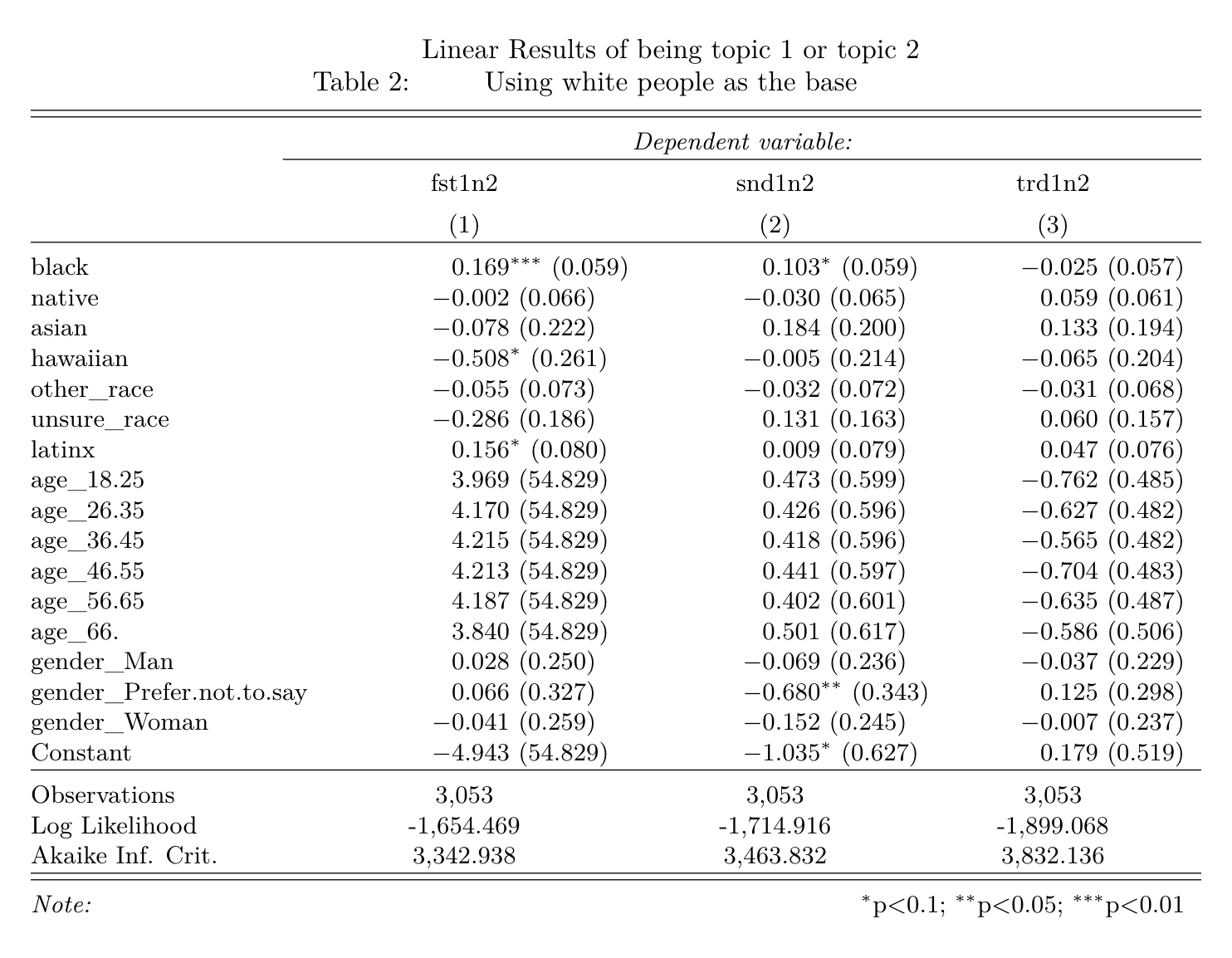
* These groups of models excluded “unsure\_race” for perfect multicollinearity.
* Ceteris paribus, **compared with “unsure race”**, it is 25.3% less possible for white people, 48.6% less possible for hawaiian and 15.1% less possible for other races to be identified topic 1 or topic 2 as the most dominant topic in their answers.
* Same as above, it is 18.8% more possible to find topic 1 or topic 2 as the second dominant topic in black people’s answers, and 11.9% more possible in white people’s answers.
* 13.3% less possible to find topic 1 or topic 2 as the third dominant topic in white people’s answers.



### 

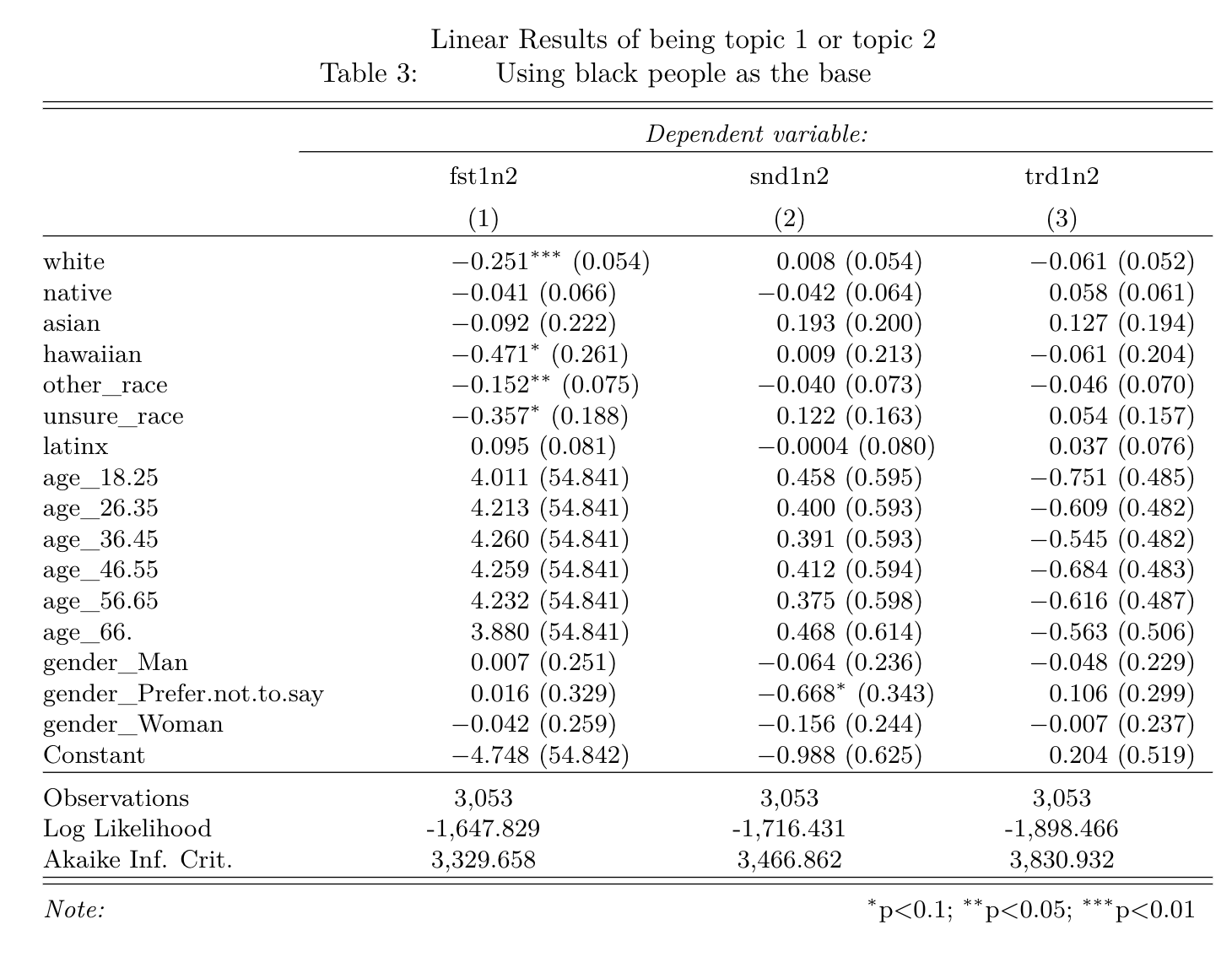
### 3.1.2 Using “white” as the base variables

* These groups of models excluded the “white” dummy variable.
* Ceteris paribus, **compared with white people**, it is 16.9% more possible for black people, 50.8% less possible for hawaiian and 15.6% more possible for latinx to be identified topic 1 or topic 2 as the most dominant topic in their answers.
* Same as above, it is 10.3% more possible to find topic 1 or topic 2 as the second dominant topic in black people’s answers.



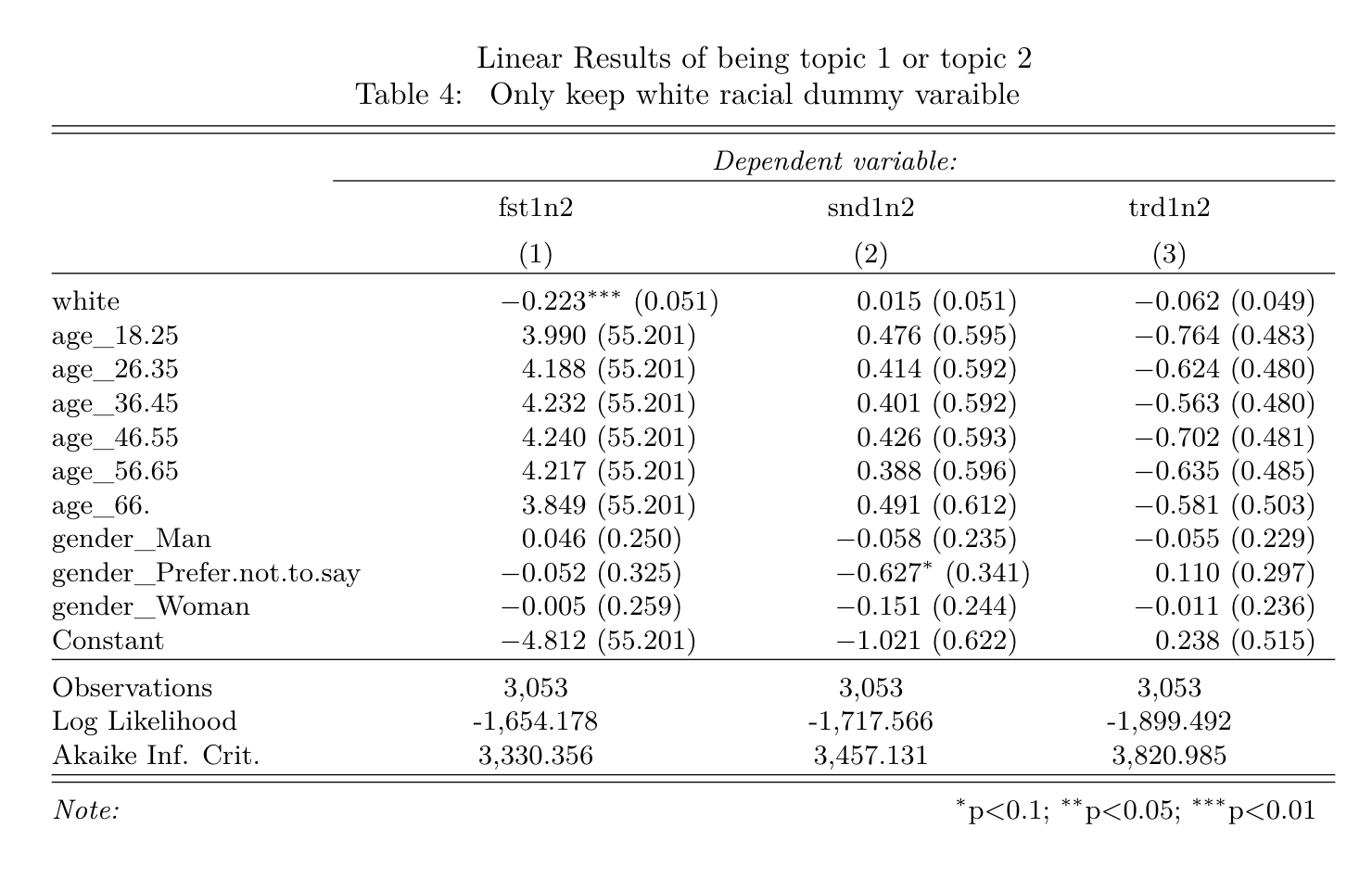
### 3.1.3 Using “black” as the base variables

* These groups of models excluded the “black” dummy variable.
* Ceteris paribus, **compared with black people**, it is 25.1% less possible for white people, 47.1% less possible for hawaiian, 15.2% less possible for other races, and 35.7% less possible for unsure race people to be identified topic 1 or topic 2 as the most dominant topic in their answers.



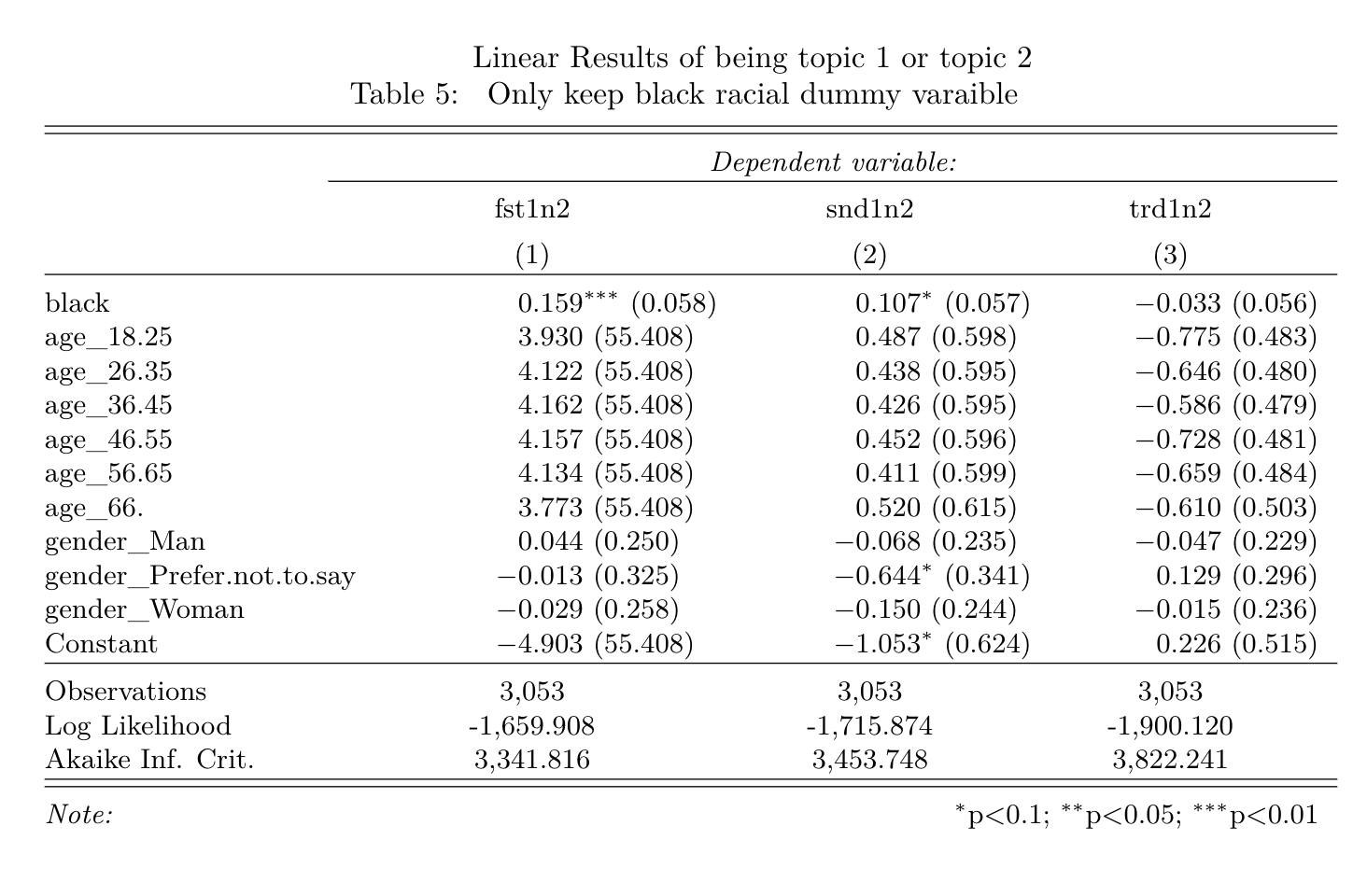
### 3.1.4 Only keeping a “white” racial dummy variable

* These groups of models excluded all racial dummy variables except “white”.
* Ceteris paribus, **compared with non-white people (minority groups)**, it is 22.3% less possible for black people to be identified topic 1 or topic 2 as the most dominant topic in their answers.



### 3.1.5 Only keeping a “black” racial dummy variable

* These groups of models excluded all racial dummy variables except “black”.
* Ceteris paribus, **compared with white people and other minority groups**, it is 15.9% more possible for black people to be identified topic 1 or topic 2 as the most dominant topic in their answers.
* Same as above, it is 10.7% more possible to find topic 1 or topic 2 as the second dominant topic in black people’s answers.

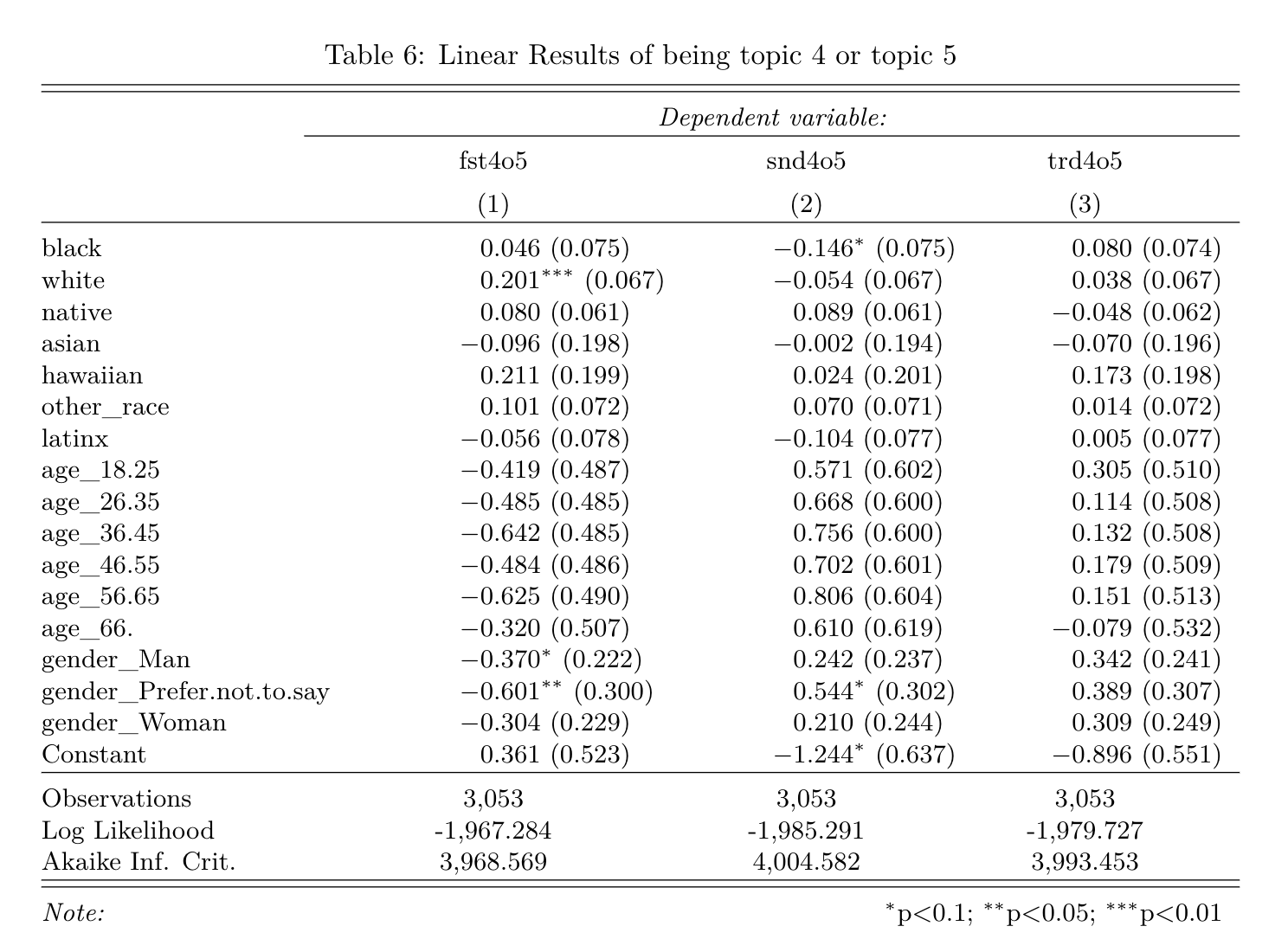


## 

## 3.2 “topic 4 or topic 5 being the most related topic” as dependent variables

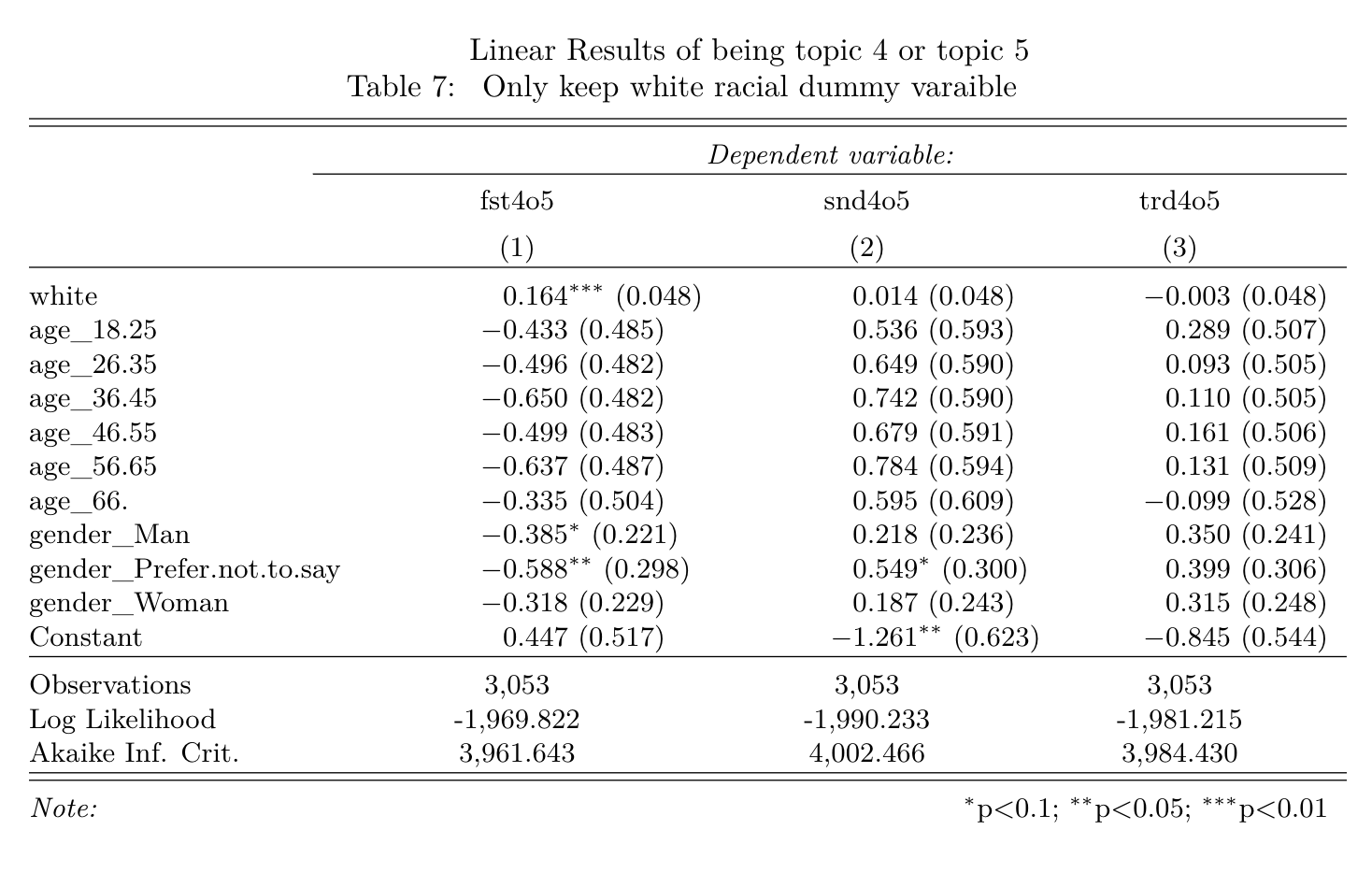
### 3.2.1 Keep all variables

* These groups of models excluded "unsure\_race" for perfect multicollinearity.
* Ceteris paribus, **compared with “unsure race”**, it is 20.1% more possible for white people to be identified topic 4 [Race does not matter] or topic 5 [Race should not matter] as the most dominant topic in their answers.
* Same as above, it is 14.6% less possible to find topic 4 or topic 5 as the second dominant topic in black people’s answers.



### 3.2.2 Only keeping a “white” racial dummy variable

* These groups of models excluded all racial dummy variables except “white”.
* Ceteris paribus, **compared with non-white people (minority groups)**, it is 16.4% more possible for black people to be identified topic 4 or topic 5 as the most dominant topic in their answers.

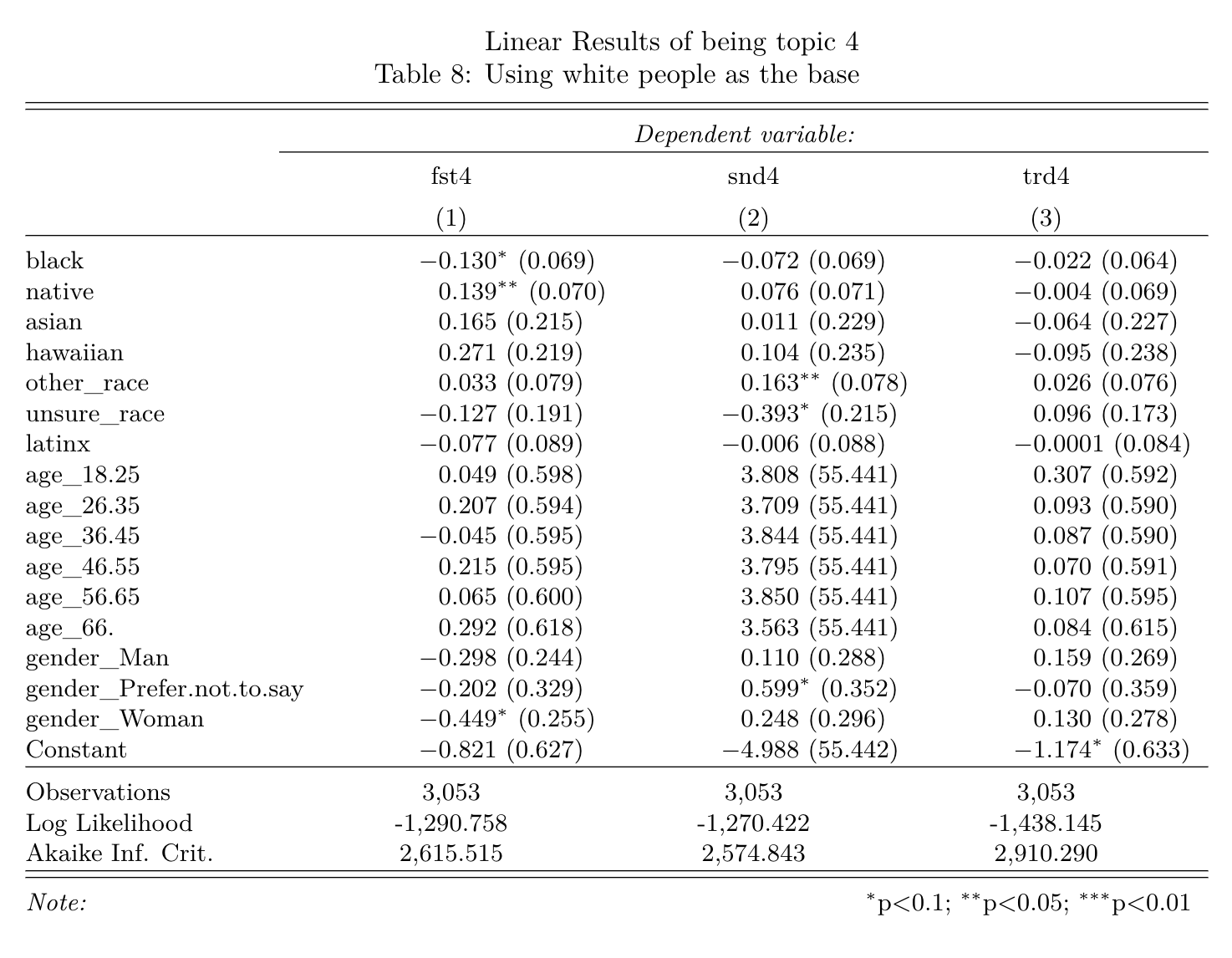


## 

## 3.3 “topic 4 being the most related topic” as dependent variables

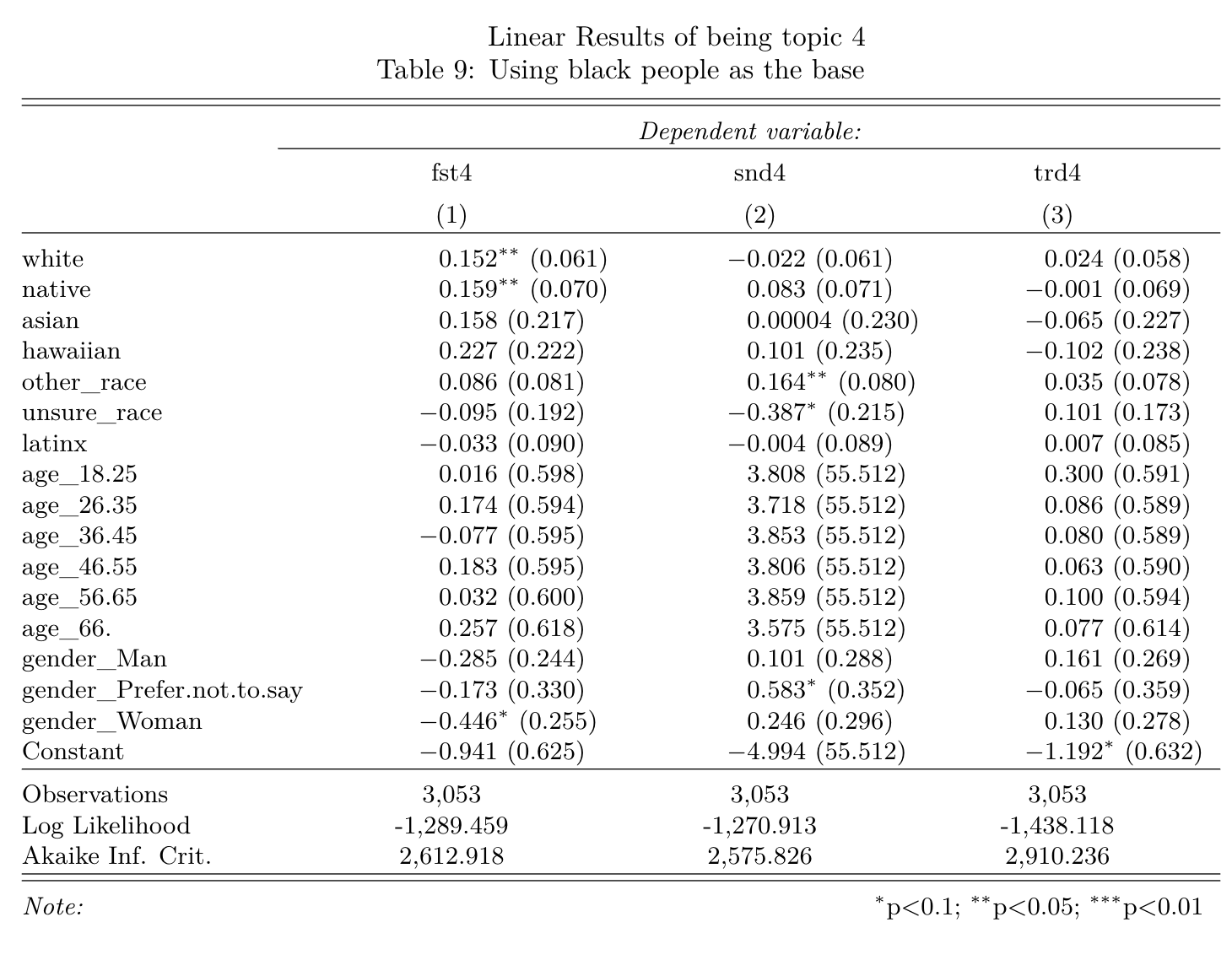
### 3.3.1 Using “white” as the base variables

* These groups of models excluded the “white” dummy variable.
* Ceteris paribus, **compared with white people**, it is 13.0% less possible for black people, 13.9% more possible for native American to be identified topic 4 [Race has no effect] as the most dominant topic in their answers.
* Same as above, it is 16.3% more possible to find topic 4 as the second dominant topic in other races’ answers and 39.3% less likely in unsure race’s answers.



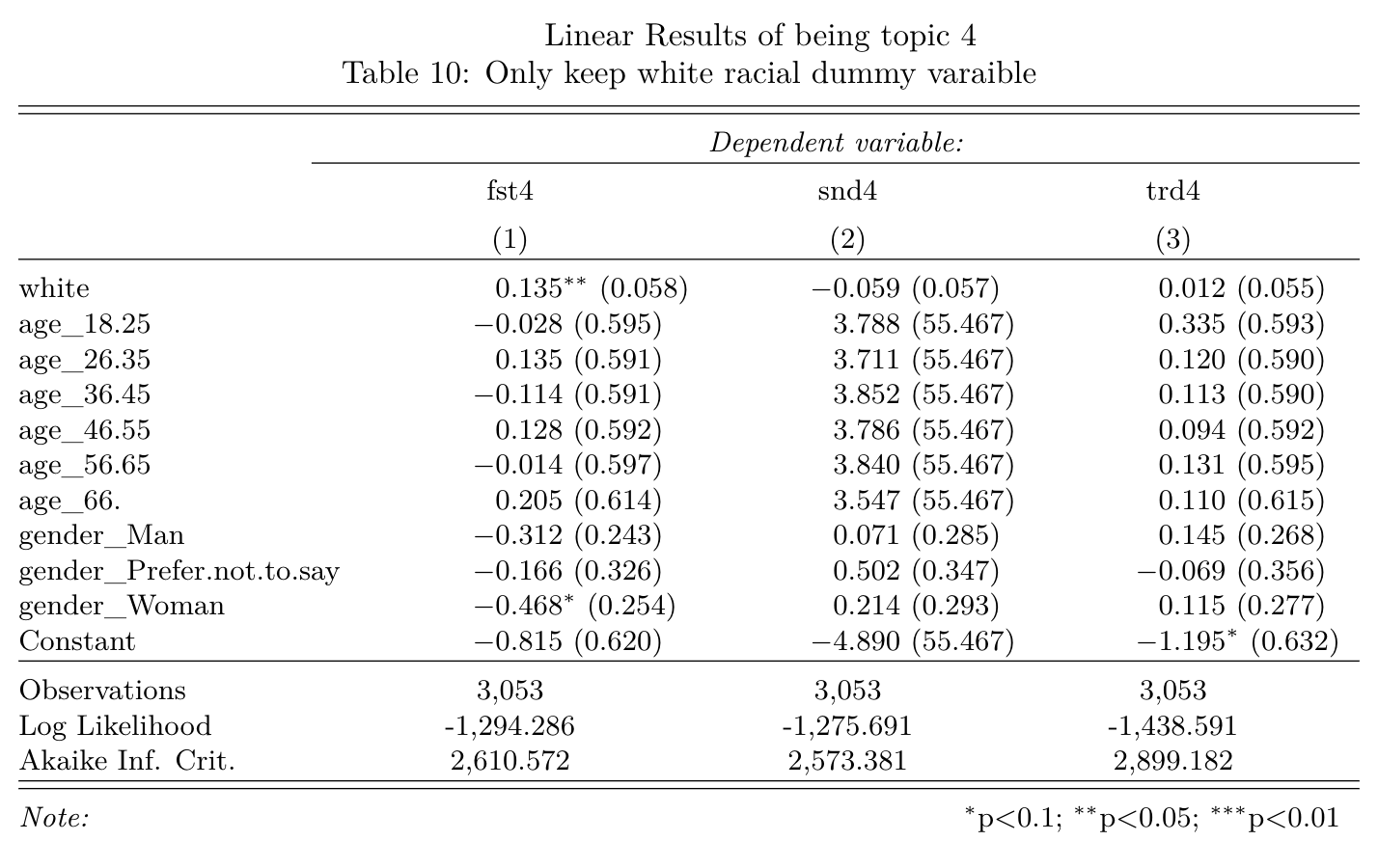
### 3.3.2 Using “black” as the base variables

* These groups of models excluded the “black” dummy variable.
* Ceteris paribus, **compared with black people**, it is 15.2% more possible for white people, 15.9% more possible for native American to be identified topic 4 as the most dominant topic in their answers.
* Same as above, it is 16.4% more possible to find topic 4 as the second dominant topic in other races’ answers and 38.7% less likely in unsure race’s answers.



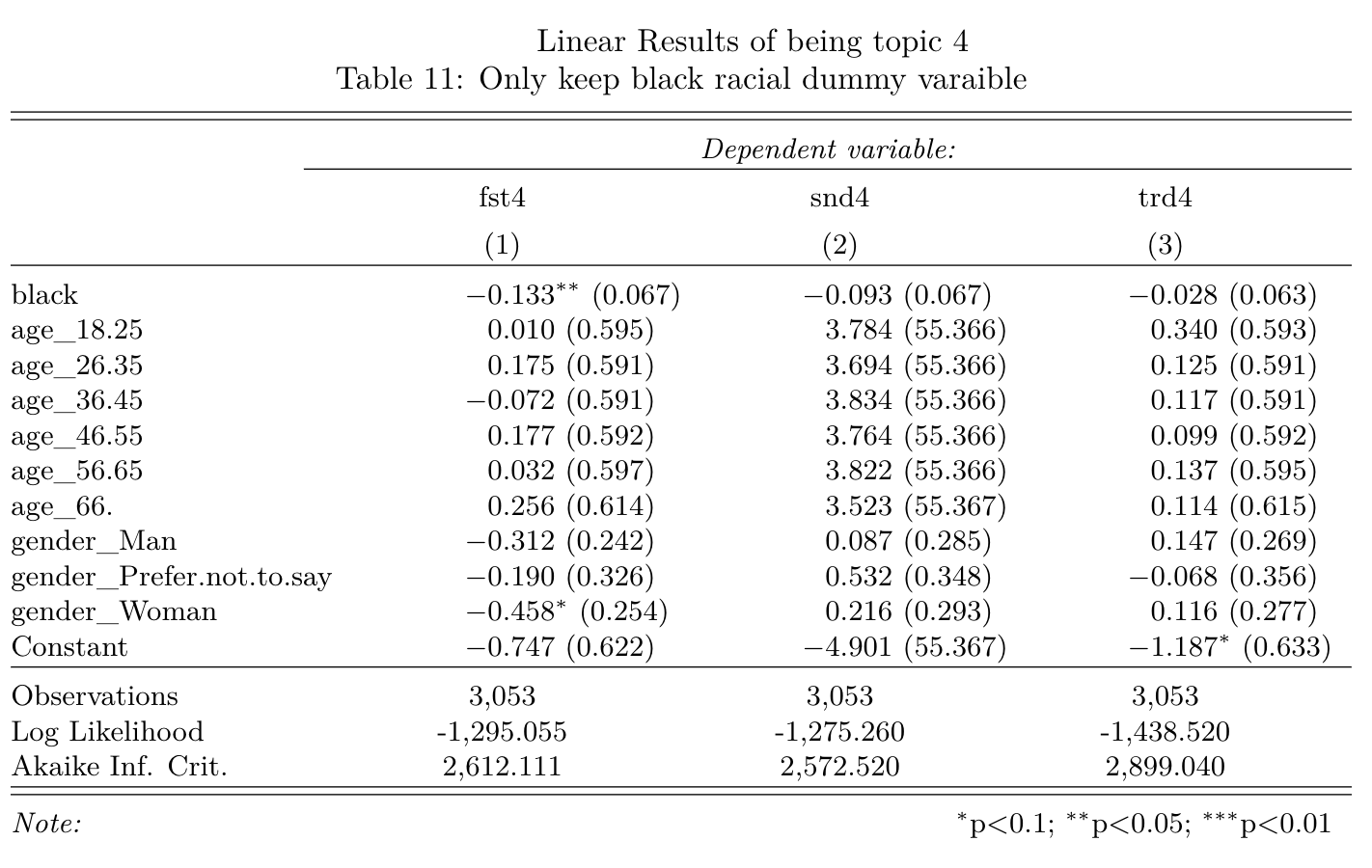
### 3.3.3 Only keeping a “white” racial dummy variable

* These groups of models excluded all racial dummy variables except “white”.
* Ceteris paribus, **compared with non-white people (minority groups)**, it is 13.5% more possible for white people to be identified topic 4 [Race has no effect] as the most dominant topic in their answers.



### 3.3.4 Only keeping a “black” racial dummy variable

* These groups of models excluded all racial dummy variables except “black”.
* Ceteris paribus, **compared with non-black people (white and other minority groups)**, it is 13.3% less possible for black people to be identified topic 4 [Race has no effect] as the most dominant topic in their answers.



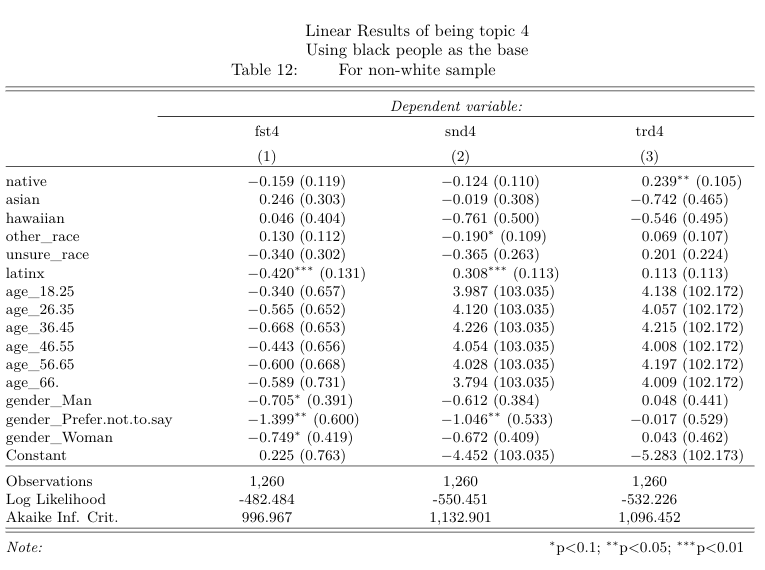
## 

## 3.4 “topic 4 being the most related topic” as dependent variables for the non-white sample set

Different from 3.3, in this part, I excluded all respondents whose answer to the race question is “white”.

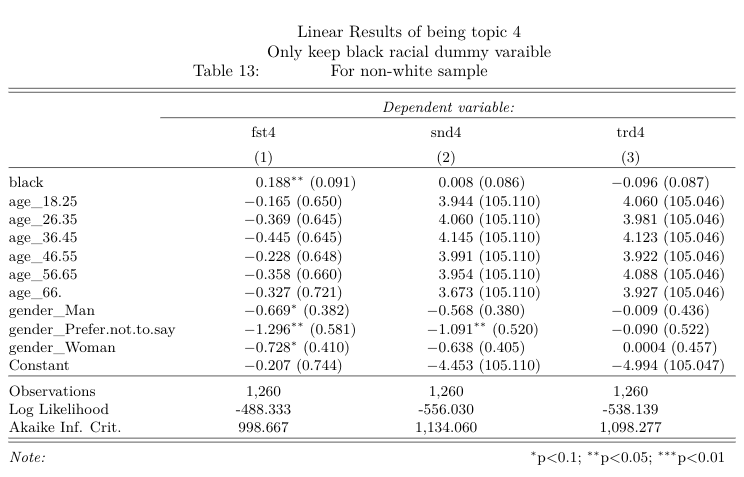
### 3.4.1 Using “black” as the base variables

* These groups of models excluded the “black” dummy variable.
* Ceteris paribus, **within non-white people, compared with black people**, it is 42.0% less possible for latinx to be identified topic 4 as the most dominant topic in their answers.
* Same as above, it is 19.0% less possible to find topic 4 as the second dominant topic in other races’ answers and 30.8% more likely in latinx’s answers.



### 3.4.2 Only keeping a “black” racial dummy variable

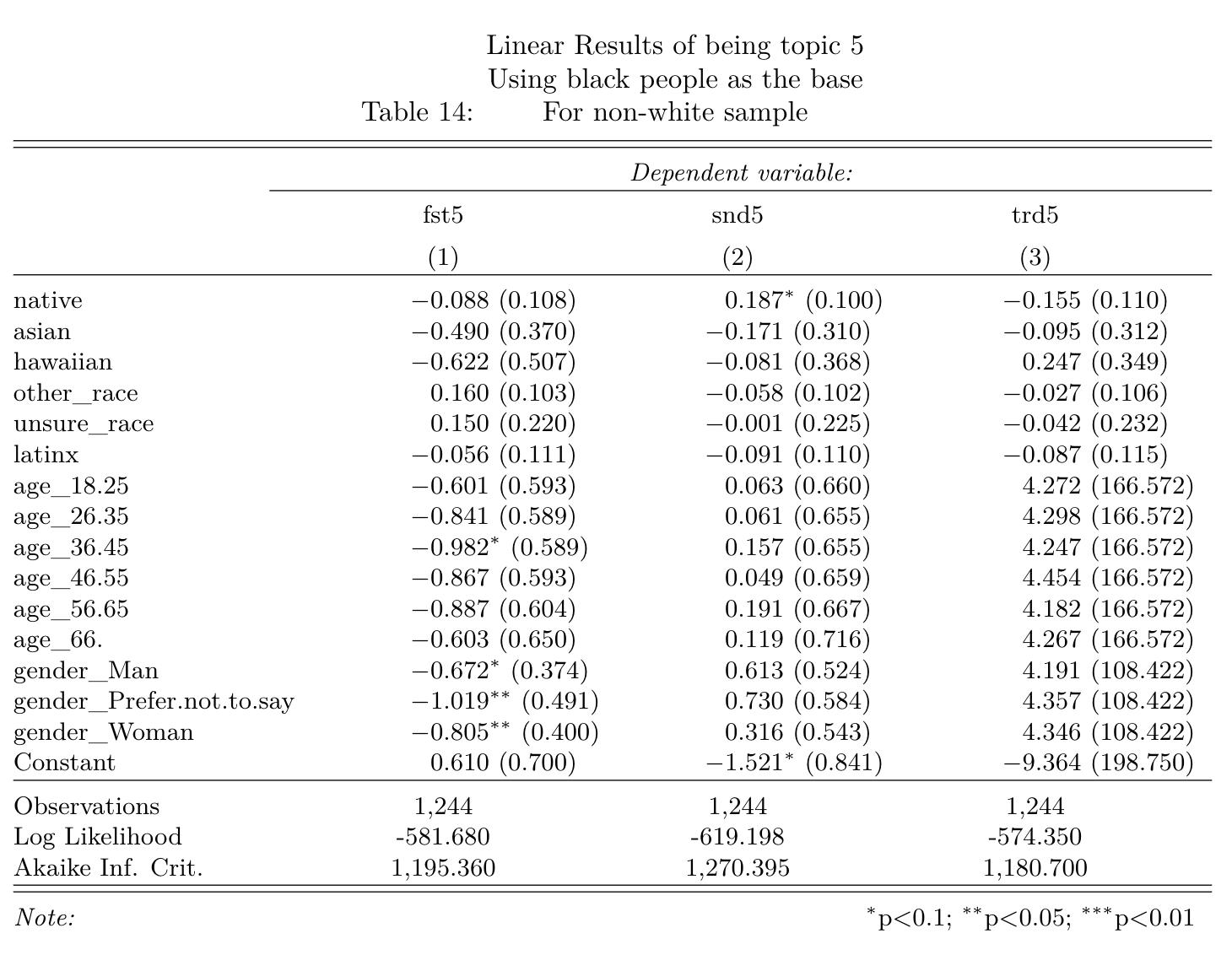
* These groups of models excluded all minority racial dummy variables except “black”.
* Ceteris paribus, **within non-white people, compared with other minority groups**, it is 18.8% more possible for black people to be identified topic 4 [Race has no effect] as the most dominant topic in their answers.



## 3.5 “topic 5 being the most related topic” as dependent variables for the non-white sample set

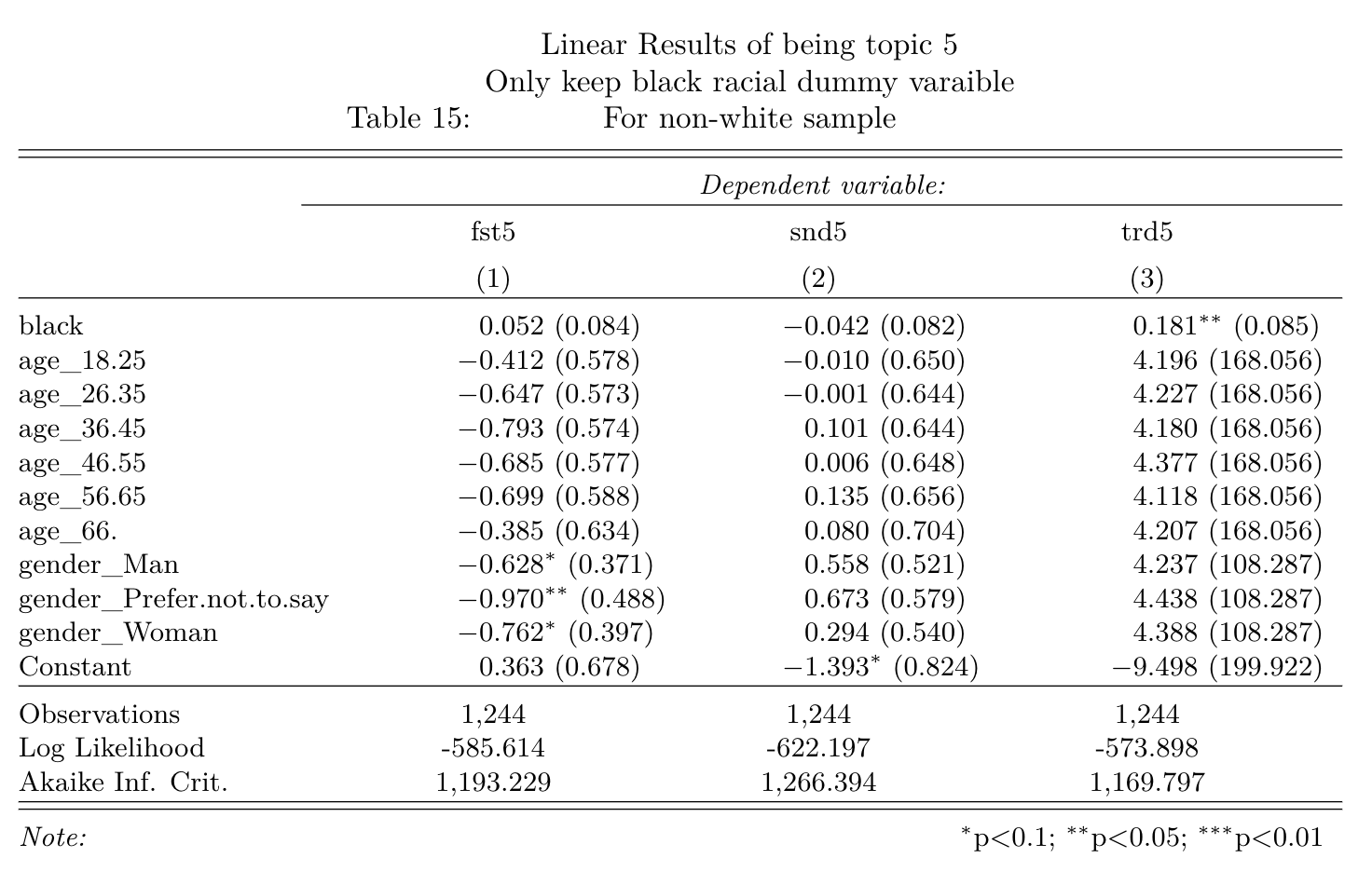
### 3.5.1 Using “black” as the base variables

* These groups of models excluded the “black” dummy variable.
* Ceteris paribus, **within non-white people, compared with black people**, it is 18.7% more possible for native American to be identified topic 5 [Race should have no effect] as the second dominant topic in their answers.



### 3.5.2 Only keeping a “black” racial dummy variable

* These groups of models excluded all minority racial dummy variables except “black”.
* Ceteris paribus, **within non-white people, compared with other minority groups**, it is 18.1% more possible for black people to be identified topic 5 [Race should have no effect] as the third dominant topic in their answers.



1. The stopword sets in the Python nltk module. [↑](#footnote-ref-0)
2. The Goodness of Fit was most significant when assumed there are six main topics in the answers set. [↑](#footnote-ref-1)
3. Removing regular stopwords [3095 samples]; removing regular stopwords and “Race” [3082 samples]; removing regular stopwords and “Race”, “White”, and “Black” [3053 samples] [↑](#footnote-ref-2)