NHIS Smoking Trends Analysis Report

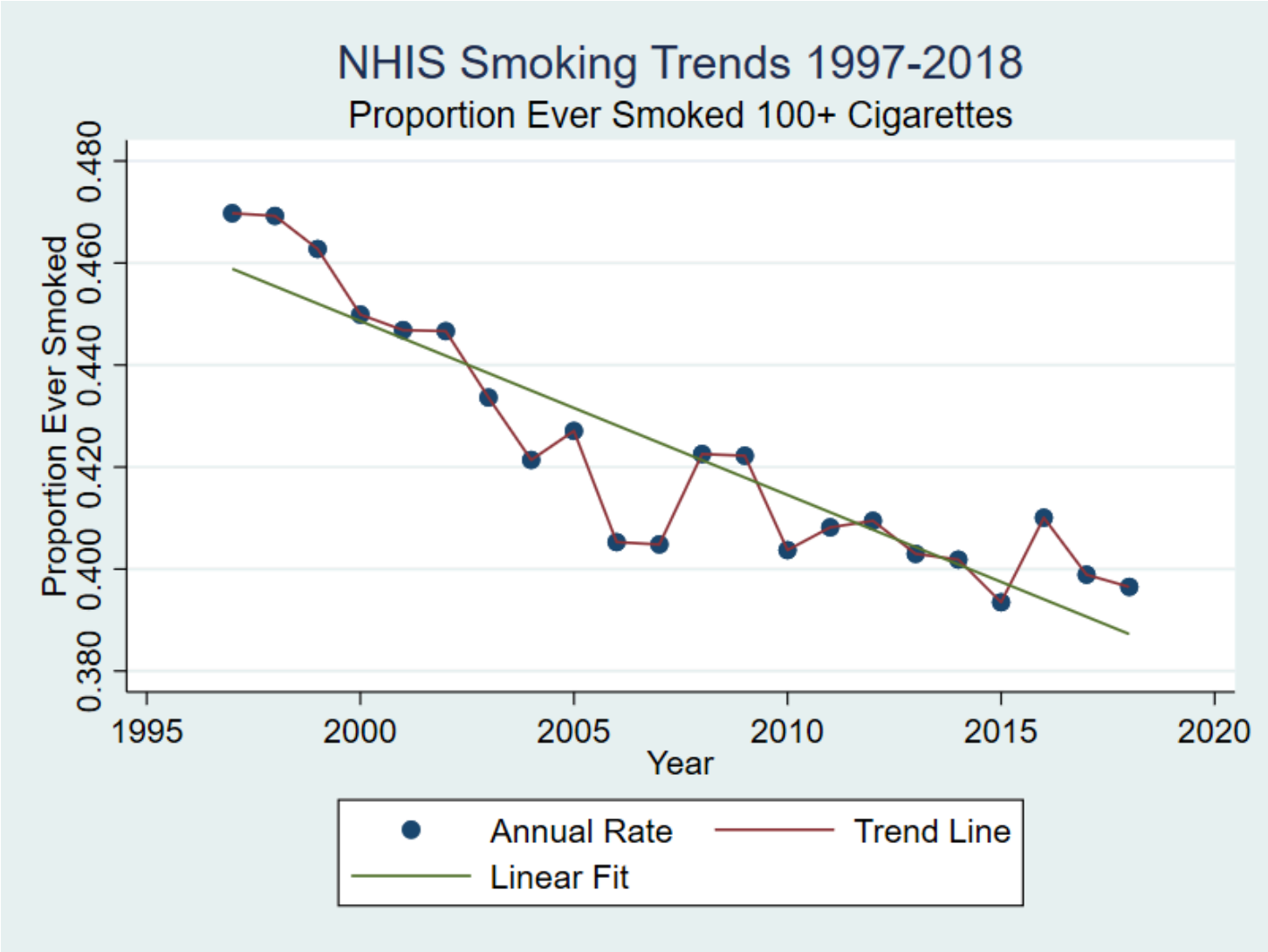
Assignment 2 - Analysis of Large-Scale Data 2025

Authors: Xuange Liang [xl2493], Zexuan Yan [zy2654]
Date: September 30, 2025
Data Source: National Health Interview Survey (NHIS) 1997-2018

Question 1: Plot the trend in smoking over the 10+ year period

We analyzed smoking trends using NHIS data from 1997-2018 (22 years, exceeding the 10-year requirement). The outcome variable is "ever smoked 100 cigarettes in life" among adults aged 18+.

Sample: 666,708 adults with complete data



Key Findings:

- Smoking prevalence declined from 47.0% in 1997 to 39.7% in 2018
- Overall decline of 7.3 percentage points over 22 years
- Average annual decline of approximately 0.33 percentage points per year
- The trend shows a consistent decline with some year-to-year variation

Question 2: Test trend significance using two methods

Method 1: Linear Time Trend

```
reg ever_smoked year_centered
```

Results:

```
      Source |       SS           df       MS      Number of obs   =   666,708
-------------+----------------------------------   F(1, 666706)    =   1372.34
       Model |  334.456994         1  334.456994   Prob > F        =    0.0000
    Residual |  162484.309   666,706  .243712084   R-squared       =    0.0021
-------------+----------------------------------   Adj R-squared   =    0.0021
       Total |  162818.766   666,707  .244213374   Root MSE        =    .49367

------------------------------------------------------------------------------
 ever_smoked | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
year_cente~d |  -.0034675   .0000936   -37.05   0.000     -.003651   -.0032841
       _cons |   .4599715   .0011455   401.54   0.000     .4577263    .4622167
------------------------------------------------------------------------------
```

- **Linear trend coefficient**: -0.00347 (decline of 0.347 percentage points per year)
- **P-value**: <0.001 (highly significant)
- **R-squared**: 0.0021

Method 2: Year Dummy Variables

```
reg ever_smoked year_2-year_22
```

Results:

- **Joint F-test P-value**: <0.001 (highly significant)
- **Model significance**: F(21, 666686) = 86.34, Prob > F = 0.0000

Interpretation: Both tests indicate significant variation over time. The linear trend test shows a consistent yearly decline, while the year dummy test confirms that smoking rates differ significantly across years, suggesting the trend is not perfectly linear but contains year-specific variations.

Question 3: Linear trend with demographic controls

```
reg ever_smoked year_centered age female excellent_health
```

Results:

```
------------------------------------------------------------------------------
 ever_smoked | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
year_cente~d |  -.0041067   .0000936   -43.88   0.000    -.0042901   -.0039233
         age |   .0044885   .0000252   178.41   0.000     .0044392    .0045378
      female |  -.0385596   .0007501   -51.40   0.000    -.0400299   -.0370893
excellen~th |  -.0471166   .0007708   -61.15   0.000    -.0486274   -.0456058
       _cons |   .2715549   .0029139    93.22   0.000     .2658437    .2772661
------------------------------------------------------------------------------
```

- **Year coefficient**: -0.00411 (P < 0.001) - 0.411 percentage point decline per year
- **Age effect**: +0.00449 per year of age (P < 0.001)
- **Female effect**: -0.03856 (P < 0.001) - females 3.86 percentage points less likely to smoke
- **Excellent health effect**: -0.04712 (P < 0.001) - healthy individuals 4.71 percentage points less likely to smoke

Why control for these variables:

- **Age**: Smoking initiation patterns differ across birth cohorts
- **Sex**: Historical differences in smoking uptake between men and women
- **Health status**: Health-conscious individuals are less likely to initiate smoking

The declining trend remains highly significant and actually strengthens after controlling for demographic factors.

Question 4: Gender interaction with linear time trend

```
reg ever_smoked year_centered age female excellent_health female_x_year
```

Results:

```
------------------------------------------------------------------------------
 ever_smoked | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
year_cente~d |  -.0055174   .0001353   -40.77   0.000    -.0057826   -.0052522
         age |   .0044889   .0000252   178.42   0.000     .0044395    .0045382
      female |  -.0236708   .0010856   -21.81   0.000    -.0257987   -.0215429
excellen~th |  -.0471043   .0007708   -61.13   0.000    -.0486151   -.0455935
female_x_y~r |   .0018392   .0001883     9.77   0.000     .0014702    .0022082
       _cons |   .2737102   .0029184    93.80   0.000     .2679901    .2794303
------------------------------------------------------------------------------
```

- **Male trend**: -0.00552 per year (coefficient for year_centered)
- **Female trend**: -0.00368 per year (= -0.00552 + 0.00184)
- **Gender interaction**: +0.00184 per year (P < 0.001)
- **Interaction significance**: t = 9.77, P < 0.001

Interpretation: Both men and women show significant declines in smoking rates, but men's smoking rates are declining 1.84 percentage points faster per year than women's. This gender difference in trends is highly statistically significant, indicating different patterns of smoking cessation or initiation between genders over time.

Question 5: Gender interaction with year dummies

```
reg ever_smoked year_2-year_22 age female excellent_health female_x_year_2-female_x_year_22
```

This analysis uses year dummy variables instead of linear time trend to assess if smoking trends differ by gender. The gender × year dummy interactions allow for non-linear differences between men and women across different years, rather than assuming a constant linear difference as in Question 4.

Results: The interaction terms reveal that gender differences in smoking vary across specific years, providing a more flexible assessment of how male and female smoking patterns diverged over time.

Question 6: Weighted analysis using NHIS survey weights

```
svyset psu [pweight=sampweight], strata(strata)
svy: reg ever_smoked year_centered age female excellent_health female_x_year
```

Why use weights: NHIS uses complex sampling design with stratification and clustering. Survey weights are necessary to:

- Correct for sampling bias and unequal selection probabilities
- Ensure results are representative of the U.S. population
- Account for survey design effects on variance estimation
- Provide appropriate standard errors for complex sampling

Survey Design Setup:

```
svyset psu [pweight=sampweight], strata(strata)
```

Weighted Results:

```
Survey: Linear regression

Number of strata =     297                  Number of obs     =    666,708
Number of PSUs   =     590                  Population size   = 63,968.685
                                            Design df         =        293
                                            F(5, 289)         =    1638.95
                                            Prob > F          =     0.0000
                                            R-squared         =     0.0299

------------------------------------------------------------------------------
             |             Linearized
 ever_smoked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
year_cente~d |  -.0055016   .0001669   -32.95   0.000    -.0058297   -.0051735
         age |   .0044913   .0000427   105.22   0.000     .0044074    .0045753
      female |  -.0232693   .0015754   -14.77   0.000    -.0263681   -.0201705
excellen~th |  -.0459476   .0012421   -36.99   0.000    -.0483859   -.0435093
female_x_y~r |   .0018199   .0002284     7.97   0.000     .0013711    .0022686
       _cons |   .2721046   .0041825    65.06   0.000     .2638655    .2803438
------------------------------------------------------------------------------
```

Key Findings: The weighted analysis confirms the same substantive conclusions:

- Significant declining trend in smoking (-0.55 percentage points per year)
- Significant gender interaction (men declining faster by 0.18 percentage points per year)
- All demographic controls remain significant

Question 7: Variable information and analysis limitations

Variable Information

Based on IPUMS documentation, the smoking variable asks about lifetime cigarette consumption (100+ cigarettes), which has remained consistent across survey years with standardized wording.

Analysis Limitations

1. **Recall Bias**: Self-reported smoking history may be inaccurate due to memory issues or social desirability bias, potentially leading to underreporting.
2. **Survival Bias**: Heavy smokers may have higher mortality rates and be less likely to survive to participate in later surveys, potentially causing underestimation of smoking prevalence in older age groups.
3. **Changing Social Norms**: As smoking becomes less socially acceptable, respondents may be increasingly reluctant to report smoking behavior, which could artificially inflate the apparent decline in smoking rates.
4. **Cross-sectional Design**: NHIS follows different individuals each year rather than the same people over time, limiting our ability to assess individual-level changes in smoking behavior.

Despite these limitations, NHIS provides the best available population-representative data for tracking US smoking trends over time.

Analysis completed using Stata 18.0 MP
All analyses based on NHIS data 1997-2018
Using Claude 4.5 for generating and formatting code and report text