Dr. Çetinkaya-Rundel
Duke University - Department of Statistical Science

Application exercise 7.2: Interpreting models with a transformed response

Team name: _____

Lab section: 8:30 10:05 11:45 1:25 3:05 4:40

Write your responses in the spaces provided below. WRITE LEGIBLY and SHOW ALL WORK! Only one submission per team is required. One team will be randomly selected and their responses will be discussed and graded. Concise and coherent are best!

Predicting income in the US

Each year since 2005, the US Census Bureau surveys about 3.5 million households with The American Community Survey (ACS). Data collected from the ACS have been crucial in government and policy decisions, helping to determine the allocation of more than $400 billion in federal and state funds each year. For example, funds for the Adult Education and Family Literacy Act are distributed to states taking into consideration data from the ACS on number of adults 16 and over without a high school diploma. This act is the primary source of federal funding for adults with low basic skills seeking further education or English language services, and Department of Education uses ACS data to ensure the efficient distribute funds.

In this application exercise we will analyze data from the ACS, and use the fact that it is "a random survey" to make inferences about the US population at large.

List of variables:

1. `income`: Yearly income (wages and salaries)
2. `employment`: Employment status, not in labor force, unemployed, or employed
3. `hrs_work`: Weekly hours worked
4. `race`: Race, White, Black, Asian, or other
5. `age`: Age
6. `gender`: Male or female
7. `citizens`: Whether respondent is a US citizen or not
8. `time_to_work`: Travel time to work
9. `lang`: Language spoken at home, English or other
10. `married`: Whether respondent is married or not
11. `edu`: Education level, hs or lower, college, or grad

12. `disability`: Whether respondent is disabled or not
13. `birth_qrtr`: Quarter in which respondent is born, Jan thru Mar, Apr thru Jun, Jul thru Sep, or Oct thru Dec

First, load the R Markdown template:

```
download("http://stat.duke.edu/courses/Spring16/sta101.001/rmd/app_Trans_MLR.Rmd",
  destfile = "app_Trans_MLR.Rmd")
```

1. Load data, and subset for those who were employed:

   ```
   load(url("http://stat.duke.edu/~mc301/data/acs.RData"))

   acs_emp <- acs %>%
     filter(employment == "employed", income > 0)
   ```

   Before you proceed, confirm that this leaves you with 787 observations.

2. Suppose we only want to consider the following explanatory variables: `hrs_work`, `race`, `age`, `gender`, `citizen`. Fit the full model using only the explanatory variables listed above, and report its adjusted R^2. Remember your response variable is `log(income)`, not `income`. Below is a neat trick for getting the just the adjusted R^2:

   ```
   m_full <- lm(log(income) ~ hrs_work + race + age + gender + citizen,
     data = acs_emp)

   summary(m_full)$adj.r.squared
   ```

3. Conduct model selection using the backwards adjusted R^2 method, and report the adjusted R^2 for the final model.
4. Check diagnostics for your final model using appropriate plots. *Hint:* Code from today's slides should be helpful.
5. Interpret the slope for one numerical and one categorical predictor. *Hint:* To exponentiate a value in `R`, use the `exp()` function, e.g. to calculate e^3, use

   ```
   exp(3)
   ```