Lab 12: Multiple linear regression

American Community Survey

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.

The ACS received a surge of media attention in Spring 2012 when the House of Representatives voted to eliminate the survey. Daniel Webster, a first-term Republican congressman from Florida, sponsored the legislation citing the following reasons:

- "This is a program that intrudes on peoples lives, just like the Environmental Protection Agency or the bank regulators"
- "Were spending $70 per person to fill this out. Thats just not cost effective"
- "in the end this is not a scientific survey. Its a random survey."

In this lab 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.

Template for lab report

Write your report, or at least run the code and create the plots, as you go so that if you get errors you can ask your TA to help on the spot. Knit often to more easily determine the source of the error.

```
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Lab/lab12.Rmd", destfile = "lab12.Rmd")
```

The data

```
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Lab/acs.RData", destfile = "acs.RData")
load("acs.RData")
```

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`: gender, male or female
7. `citizens`: Whether respondent is a US citizen or not
8. `time_to_work`: Travel time to work
9. `married`: Whether respondent is married or not

10. `edu`: Education level, hs or lower, college, or grad
11. `disability`: Whether respondent is disabled or not
12. `birth_qrtr`: Quarter in which respondent is born, jan thru mar, apr thru jun, jul thru sep, or oct thru dec

Exercise 1 In this lab we will focus on predicting `income` for respondents who were employed in the last year. Create a new dataset called `acs_sub` that only contains such respondents. How many such respondents are there in the dataset? Note that this is a crucial step in the analysis since you'll be using this subsetted dataset in the rest of the lab.

Simple linear regression

Exercise 2 Income and hours worked: Fit a model, called `inc_hw`, predicting income based on number of hours worked per week. Include the model output in your response as well as writing out the linear model. Interpret the slope of hours worked, and determine if hours worked is a significant predictor of income.

Exercise 3 Income and gender: Fit a model, called `inc_g`, predicting income based on gender of the respondent. Include the model output in your response as well as writing out the linear model. According to this model how do salaries of males and females compare? Does there appear to be a significant difference in the average salaries of the two genders? Explain your reasoning.

Exercise 4 Hours worked and gender: Lastly, fit a model, called `hw_g`, predicting hours worked based on gender. Do these two variables appear to have a significant association with each other? What is the nature of the association between these variables?

Exercise 5 Explain why number of hours worked per week is a confounding variable in the relationship between income and gender.

Multiple linear regression

In order to determine if gender is a significant predictor of income, even after accounting for the number of hours worked per week, we can fit a multiple linear regression predicting income using both gender and hours worked.

```
inc_hw_g = lm(income ~ hrs_work + gender, data = acs_sub)
summary(inc_hw_g)
```

Exercise 6 Does gender appear to be a significant predictor of income even after accounting for the number of hours worked per week? Explain your reasoning and include the model output in your response.

Exercise 7 What is the predicted yearly income for a male who works 40 hours per week? For a female who works 40 hours per week? What is the difference?

We can easily get all the predicted (fitted) values for all individuals in our sample using the `predict` function in R:

```
inc_hat = predict(inc_hw_g)
```

To get interval estimates instead of just point estimates, we include the `interval` argument. You can generate confidence intervals and prediction intervals for all the data points with

```
predict(inc_hw_g, interval = "confidence")
predict(inc_hw_g, interval = "prediction")
```

The output of the predict function with the `interval` argument includes the first column, fit, which gives the predicted (or "fitted" values), and the next two columns give the lower and upper bounds of the interval estimate (condence or prediction, depending on which you specify). The default level of condence is 95%.

Exercise 8 Describe, in one sentence, the difference between a prediction and confidence interval.

We can also make predictions for new data. To generate a prediction interval for a male who works 40 hours per week, use the following code:

```
newdata = data.frame(hrs_work = 40, gender = "male")
predict(inc_hw_g, newdata, interval = "prediction")
predict(inc_hw_g, newdata, interval = "confidence")
```

Exercise 9 Interpret these intervals (include the interval outputs in your response). Which interval is wider? Why?

Model selection

We will start with a full model that predicts income based on hours worked per week, race, age, gender, citizenship status, travel time to work, laguage spoken at home, marriage status, education, disability status, and birth quarter.

Exercise 10 Which variable would you expect to have the highest p-value in this model? Why? *Hint:* Think about which variable would you expect to not have any association with income.

Let's run the model...

```
mlr_full = lm(income ~ hrs_work + race + age + gender + citizen + time_to_work +
              married + edu + disability + birth_qrtr, data = acs_sub)
summary(mlr_full)
```

Exercise 11 Confirm your suspicions from the previous exercise. Include the model output in your response.

Exercise 12 Interpret the slope coefficients associated with the two levels of the education variable.

Exercise 13 Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficients depend on the other variables that are included in the model.) If not, what does this say about whether or not the dropped variable was collinear with the rest of the other explanatory variables?

Exercise 14 Based on adjusted R^2 which model do you think is better? Justify your answer.

Exercise 15 Predict the income for a female who is 40 years old, completed college, is African American, works 60 hours per week, is a citizen, travels 30 minutes to work, is married, is not disabled, and born in July. Depending on your final model you may or may not use all this information for predicting her salary. Along with your estimate also provide a prediction interval.

Model diagnostics

Exercise 16 Check if conditions for multiple linear regression are met for your final model. You may want to refer to the previous lab if you're not sure how to make some of the plots.