# Applied Regression and Time Series Analysis (2016 Fall): HW2 - Week 4

*Jeffrey Yau*

*September 17, 2016*

## Instructions:

The weekly assignment serves two purposes: (1) Review concepts, techniques, theories, statistical models covered during the week. (2) Extend the materials taught in the asynchronized lectures, assigned readings, and live sessions; some new concepts and/or techniques are introduced in the weekly assignment.

Below are specific instructions:

- **Due: 10/2/2016 (11:59pm PST)**
- You may complete this assignment on your own or in a group of no more than 3 students.
- When working in a group, you are strongly encouraged to complete the assignment on your own before discussing your group mates. Do not use the "division-of-labor" approach to complete the assignment.
- The homework is designed as a quantitative analysis. The instructions and questions are designed to guide you through the analysis of data using regression techniques. As such, you should think of it as a quantitative case study and the result of the study is a report with a set of well-written codes that can be used to reproduce the results in the report.
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. R-script or R markdown file
    2. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    - **SectionNumber_hw02_LastNameFirstInitial.fileExtension**
    - Examples:
      - Section1_hw02_YauJ.Rmd
      - Section1_hw02_YauJ.pdf
      - Section1_hw02_TiwariD_YauJ.Rmd
      - Section1_hw02_TiwariD_YauJ.pdf

## Overview:

The purpose of the homework assignment is to develop your skills in statistical inference in the context of classical linear regression. It is an importnat skill in building a regression model and testing hypothesis that may be based on a data science problem. Specifically, this homework will ask you to practice testing hypothesis testing of one regression parameter as well as testing linear restriction.

Remember that in the last homework, we used exploratory data analysis (EDA), generated insights learned from the EDA, estimated a linear regression model in *R*, interpretated model results, and conducted regression

diagnostics. One of the key techniques in regression model building is to leveraging insights learned from the EDA in regression model specification, which we will study in week 5. As such, for this assignment, we will just generate insights from the EDA and just think about (i.e. not need to try it yet) how you would use them in feature engineering (later).

You will continue to use the same data set from $hw01$

# Description of the Data:

The file **birthweight_w271.Rdata** contains data from the *1988 National Health Interview Survey*, which is modified by the instructor. This survey is conducted by the U.S. Census Bureau and has collected data on individual health metrics since 1957. Like all surveys, a full analysis would require advanced techniques such as those provided by the R survey package. For this homework, however, you are to treat the data as a true random sample. You will use this dataset to practice interpreting OLS coefficients.

### Testing a range of hypotheses

Assume that you have already examined the data structure, cleaned the data, and conducted the EDA. If you have not already done so (or did not do a good enough job in homework 1), you should conduct this analysis again before estimating the following model.

Estimate the following regression model:

$$bwgth = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fathereduc + \beta_6 male + \epsilon$$

# 1. Estimate this model.

```
model = lm(bwght ~ cigs + parity + faminc + motheduc + fatheduc + male,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = bwght ~ cigs + parity + faminc + motheduc + fatheduc +
##     male, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.597  -10.822    1.329   13.654  153.956
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 113.22153    4.30252  26.315  < 2e-16 ***
## cigs         -0.53358    0.12571  -4.245 2.36e-05 ***
## parity        2.33873    0.75094   3.114  0.00189 **
## faminc        0.09657    0.04171   2.316  0.02076 *
## motheduc     -0.33550    0.36416  -0.921  0.35708
## fatheduc      0.14830    0.32188   0.461  0.64508
## male          3.51988    1.31012   2.687  0.00732 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.53 on 1184 degrees of freedom
##   (197 observations deleted due to missingness)
## Multiple R-squared:  0.03518,    Adjusted R-squared:  0.03029
## F-statistic: 7.196 on 6 and 1184 DF,  p-value: 1.481e-07
```

## 2. Explain the coefficients of *cigs*, *faminc*, *motheduc*, and *male*.

- If you have two identical mothers except that first one smokes 1 more cigarette a day than the second, you'd expect the first one to have a birthweight by about .53 units less than the second mother
- If you have two identical mothers except that one earns $1,000 more in family income a year than the second, you'd expect the first one to have a birthweight by about .1 units more than the second mother
- If you have two identical mothers except that one has 1 more year of education than the second, you'd expect the first one to have a birthweight by about .34 units less than the second mother
- If you have two identical mothers except that one gives birth to male baby while the other gives birth to a female baby, you'd expect the male baby to have a birthweight 3.52 units more than the female baby.

## 3. Test the hypothesis that the average daily number of cigarettes the mother smoked during pregancy has no effect on birth weight. Interpret the results. Note that just conducting the analysis without any inte

The p value for cigs is highly significant, meaning that it is extremely unlikely that this data was observed given the null hypothesis of no effect. Since we assume that our model already past the GM assumptins during EDA, we have no reason to doubt the results of our model and reject the null hypothesis that average number of cigarettes smoked per day has no effect on birth weight.

## 4. Test the hypothesis that parents (i.e. both mother and father) education has no effect on birth weight.

To test whether parents education is significant, we need to check if the *motheduc* and *fatheduc* variables are jointly significant.

```
dataC = data[complete.cases(data),]
modelU = lm(bwght ~ cigs + parity + faminc + motheduc + fatheduc + male,data=dataC)
modelR = lm(bwght ~ cigs + parity + faminc + male,data=dataC)
anova(modelU,modelR)
```

```
## Analysis of Variance Table
##
## Model 1: bwght ~ cigs + parity + faminc + motheduc + fatheduc + male
## Model 2: bwght ~ cigs + parity + faminc + male
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1184 600987
## 2   1186 601420 -2   -432.41 0.4259 0.6533
```

The p value is not significant, so we accept the null hypothesis that parents education has no effect on birth weight. This makes sense, as

## 5. Test the hypothesis that an increase in family income by $10,000 has the same effect as an increase in father's education by 1 year.

First, I created a new independent variable called *famincfatheduc* that is the sum of *faminc*/10 and *fatheduc*. Then I perform the regression from problem 1, except that I replace *faminc* with *famincfatheduc*. If the coefficient on *fatheduc* is significantly different from 0, then we reject the hypothesis that the coefficients on *faminc*/10 and *fatheduc* are the same.

```
famincfatheduc = data$faminc/10 + data$fatheduc
model = lm(bwght ~ cigs + parity + famincfatheduc + fatheduc + motheduc + male,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = bwght ~ cigs + parity + famincfatheduc + fatheduc +
##     motheduc + male, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.597  -10.822    1.329   13.654  153.956
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    113.2215     4.3025  26.315  < 2e-16 ***
## cigs            -0.5336     0.1257  -4.245 2.36e-05 ***
## parity           2.3387     0.7509   3.114  0.00189 **
## famincfatheduc   0.9657     0.4171   2.316  0.02076 *
## fatheduc        -0.8174     0.5867  -1.393  0.16381
## motheduc        -0.3355     0.3642  -0.921  0.35708
## male             3.5199     1.3101   2.687  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.53 on 1184 degrees of freedom
##   (197 observations deleted due to missingness)
## Multiple R-squared:  0.03518,    Adjusted R-squared:  0.03029
## F-statistic: 7.196 on 6 and 1184 DF,  p-value: 1.481e-07
```

Since the coefficient on *fatheduc* is not significant, we accept the null hypothesis.

## 6. Test the overall significane of this regression.

```
model = lm(bwght ~ cigs + parity + faminc + motheduc + fatheduc + male,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = bwght ~ cigs + parity + faminc + motheduc + fatheduc +
##     male, data = data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.597  -10.822    1.329   13.654  153.956
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 113.22153    4.30252  26.315  < 2e-16 ***
## cigs         -0.53358    0.12571  -4.245 2.36e-05 ***
## parity        2.33873    0.75094   3.114  0.00189 **
## faminc        0.09657    0.04171   2.316  0.02076 *
## motheduc     -0.33550    0.36416  -0.921  0.35708
## fatheduc      0.14830    0.32188   0.461  0.64508
## male          3.51988    1.31012   2.687  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.53 on 1184 degrees of freedom
##   (197 observations deleted due to missingness)
## Multiple R-squared:  0.03518,    Adjusted R-squared:  0.03029
## F-statistic: 7.196 on 6 and 1184 DF,  p-value: 1.481e-07
```

*The p value shows that the model is overall significant. This means that it overall performs better than an intercept only model.*