Data Analysis #2

Dustin Leatherman

October 27, 2018

Introduction

The General Social Survey (GSS) is an American Instituion which "... conducts basic scientific research on the structure and development of American society with a data-collection program designed to both monitor societal change within the United States and to compare the United States to other nations." The dataset used is a subset of a survey conducted in 2012 which measures the hours spent on household chores between households with husbamds and wives alongside their respective education levels. The dataset in question contains responses where the wife has at least a high school degree.

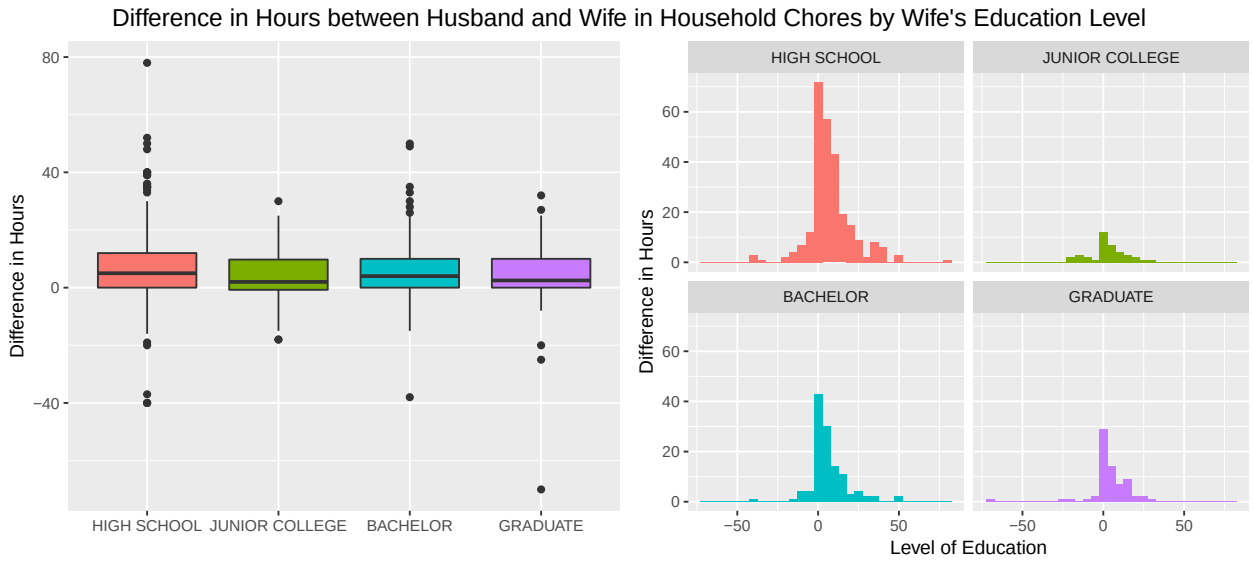
This paper is interested in discovering the answer to the following: * Is there evidence that the average difference in the number of hours of time spent on household work between the husband and wife depends on the wife's education level? * For each of the wife's education levels, by how many hours does the average difference in time spent between husband and wife on household work differ from that of the next lowest education category?

```
# reorder the wife's degree level so it shows up in order from the least amount of education to the mos
household$wife_degree = factor(
  household$wife_degree,
  levels = c("HIGH SCHOOL", "JUNIOR COLLEGE", "BACHELOR", "GRADUATE")
  )

hist <- ggplot(household, aes(x = diff_hours, fill = wife_degree)) +
  geom_histogram(position = "stack", binwidth=5, show.legend = FALSE) +
  xlab("Level of Education") + ylab("Difference in Hours") +
  guides(fill = guide_legend(title = "Level of Education"))  +
  facet_wrap(~wife_degree, ncol = 2)

bplot <- ggplot(household, aes(x = wife_degree, y = diff_hours, fill = wife_degree)) +
  geom_boxplot() +
  xlab("") +
  ylab("Difference in Hours") +
  guides(fill = FALSE)

grid.arrange(bplot, hist,
             widths = c(2, 2),
             top = textGrob("Difference in Hours between Husband and Wife in Household Chores by Wife's
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```




```
knitr::kable(household %>%
               group_by(wife_degree) %>%
               summarize_at(c("diff_hours"), funs(
                 mean,
                 sd,
                 min,
                 "25%"=quantile(diff_hours, probs = 0.25),
                 median,
                 "75%"=quantile(diff_hours, probs = 0.75),
                 max,
                 length)),
             col.names = c("", "Mean", "Std. Dev", "Min", "1st Quartile", "Median", "3rd Quartile", "Ma
             align = c('l'),
             caption = "",
             digits = 4)
```

	Mean	Std. Dev	Min	1st Quartile	Median	3rd Quartile	Max	Sample Size
HIGH SCHOOL	7.3939	13.7958	-40	0.00	5.0	12.00	78	264
JUNIOR COLLEGE	2.7895	11.3784	-18	-0.75	2.0	9.75	30	38
BACHELOR	5.9669	11.1385	-38	0.00	4.0	10.00	50	121
GRADUATE	3.9571	12.6782	-70	0.00	2.5	10.00	32	70

```
# Calculate the difference between the average of the current education level and the next lowest leve
knitr::kable(
  household %>%
  group_by(wife_degree) %>%
  summarise_at(c("diff_hours"), funs(mean)) %>%
  mutate(diff_hours_prev_wife_degree = diff_hours - lag(diff_hours, default = first(diff_hours))),
  align = c('l'),
  col.names = c("Education Level", "Avg Diff (Hrs)", "Difference in hours from previous education level
  digits = 4,
  caption = "Average difference in household hours between the current and next lowest education level"
)
```

Table 2: Average difference in household hours between the current and next lowest education level

Education Level	Avg Diff (Hrs)	Difference in hours from previous education level
HIGH SCHOOL	7.3939	0.0000
JUNIOR COLLEGE	2.7895	-4.6045
BACHELOR	5.9669	3.1775
GRADUATE	3.9571	-2.0098

Methodology

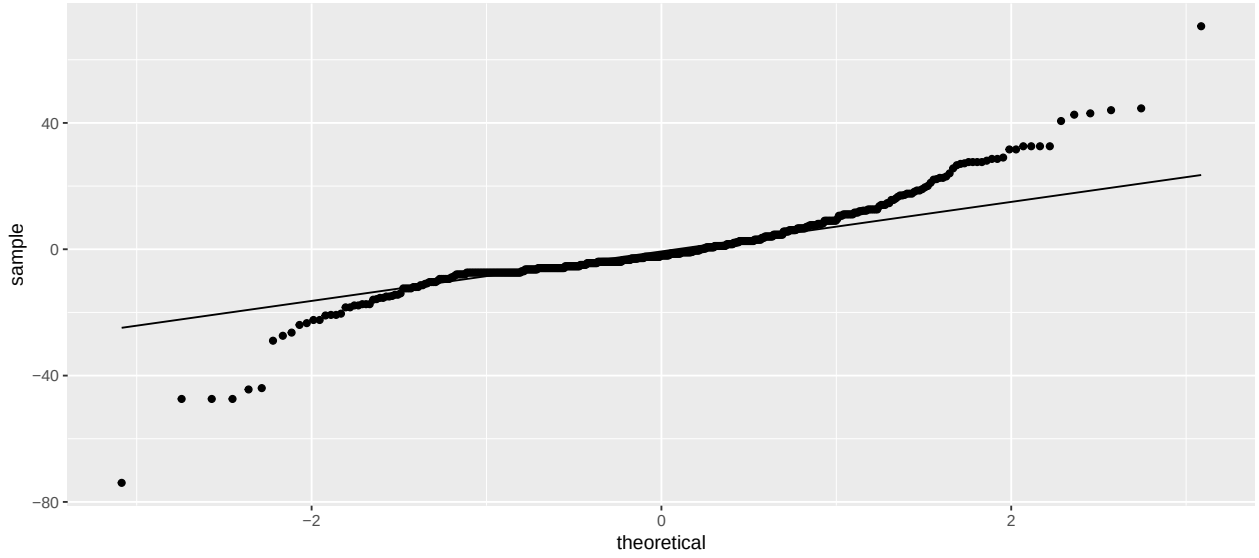
Is there evidence that the average difference in the number of hours of time spent on household work between the husband and wife depends on the wife's education level?

The summary graphs depict the difference in hours broken down by the wife's education level. In order to determine if the average difference varies between these groups, a One Way ANOVA F Test is the most appropriate test. ANOVA tests are used when comparing multiple groups together. As all tests, there are assumptions made by the ANOVA test that must be proven true before it can be used.

Normality

A normality plot of the residuals show that the extremes are non-normal. There is convincing evidence that the difference in hours worked is non-normal (Shapiro-Wilk. p-value $< 2.2e\text{-}16$). Despite the non-normality of the data, the sample size is large enough to continue forth with with a One Way ANOVA test.

```
model <- lm(diff_hours ~ wife_degree, data=household)
ggplot(model, aes(sample = .resid)) + stat_qq() + stat_qq_line()
```




```
shapiro.test(household$diff_hours)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  household$diff_hours
## W = 0.88119, p-value < 2.2e-16
```

Equal Standard Deviations

There is not enough evidence to suggest that the standard deviation of the difference of hours vary between groups of the wife's education level (Brown-Forsythe Levene's Test. p-value = 0.3412). While Levene's Test doesn't directly prove that the standard deviations are equal, the data is non-normal so Breusch-Pagan and Bartlett Test cannot be used. Levene's Test is more robust to departures from normality hence it is the best fit here.

```
levene.test(household$diff_hours, household$wife_degree)
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  household$diff_hours
## Test Statistic = 1.1134, p-value = 0.3432
```

Independence within and between groups

This sample is a simple random sample of this subset of americans which means that they are independent within and between groups. **For each of the wife's education levels, by how many hours does the average difference in time spent between husband and wife on household work differ from that of the next lowest education category?**

Since two groups are being compared with each other and the group comparisons are not linear combinations of each other, the appropriate test is a two sample t-test using pooled variance from all groups. The assumptions for a two sample t-test are the same as the assumptions above.

Normality

The consensus drawn earlier still stands. This data is non-normal but this can be waived given the large sample size.

Equal Standard Deviations

The consensus drawn earlier still stands. The standard deviations are equal.

Independence within and between groups

Same as above.

Summary

Is there evidence that the average difference in the number of hours of time spent on household work between the husband and wife depends on the wife's education level?

```
knitr::kable(lm(diff_hours ~ wife_degree, data=household) %>% anova, digits = 4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wife_degree	3	1177.034	392.3448	2.3738	0.0695
Residuals	489	80824.085	165.2844	NA	NA

There is weak evidence that the average difference in the number of hours spent on household work between the husband and wife varies based on the wife's education level (Extra Sum of Squares F Test on 489 and 3 DF. p-value = 0.06948).

For each of the wife's education levels, by how many hours does the average difference in time spent between husband and wife on household work differ from that of the next lowest education category?

```
# Get the standard deviation, sample size, and pooled standard deviation
base_stats <- household %>%
  group_by(wife_degree) %>%
  summarise_at(c("diff_hours"), funs(stdev=sd, n=n())) %>%
  mutate(pool.stdev = sqrt(sum((n - 1) * stdev^2) / sum(n - 1))) %>%
  select(wife_degree,pool.stdev)

# Calculate the lag on mean as well as the difference between the two
diff_hours_lag <- household %>%
  group_by(wife_degree) %>%
  summarise_at(c("diff_hours"), funs(mean)) %>%
  mutate(
    diff_hours_lag = lag(diff_hours, default = first(diff_hours)),
    diff_hours_prev_wife_degree = diff_hours - diff_hours_lag
  )

# Calculate the lag on n
n_lag <- household %>%
  group_by(wife_degree) %>%
  summarise_at(c("diff_hours"), funs(n=n())) %>%
  mutate(n_lag = lag(n, default = 0))

result <- inner_join(diff_hours_lag, base_stats, by = c("wife_degree")) %>%
    inner_join(n_lag) %>%
    mutate(
      stderr = pool.stdev * sqrt(1/n + 1/n_lag),
      df = sum(n) - length(wife_degree),
      lcl = diff_hours_prev_wife_degree - qt(0.975, df) * stderr,
      ucl = diff_hours_prev_wife_degree + qt(0.975, df) * stderr,
      t = diff_hours_prev_wife_degree / stderr,
      p_value = 2 * (1 -pt(abs(t), df))
      ) %>%
    # Remove High School row since we're not comparing it to a lower value
    slice(2:4)

# Join all of these together into the same dataframe and perform a two-tailed t-test with a 95% C.I
knitr::kable(
  result %>% select(wife_degree, diff_hours, diff_hours_lag,
                    diff_hours_prev_wife_degree, pool.stdev, n, n_lag, stderr, df),
  digits = 4,
  col.names = c("Degree", "Diff (Hrs)", "Diff Lag (Hrs)", "Estimate", "Pooled Stddev", "n", "n lag", "S
  align = c('l')
)
```

Degree	Diff (Hrs)	Diff Lag (Hrs)	Estimate	Pooled Stddev	n	n lag	Stderr	DF
JUNIOR COLLEGE	2.7895	7.3939	-4.6045	12.8563	38	264	2.2306	489
BACHELOR	5.9669	2.7895	3.1775	12.8563	121	38	2.3907	489
GRADUATE	3.9571	5.9669	-2.0098	12.8563	70	121	1.9306	489

```
knitr::kable(
  result %>% select(wife_degree, lcl, ucl, t, p_value),
  col.names = c("Degree", "LCL", "UCL", "T", "p-value"),
  align = c('l'),
  digits = 4
)
```

Degree	LCL	UCL	T	p-value
JUNIOR COLLEGE	-8.9872	-0.2217	-2.0642	0.0395
BACHELOR	-1.5199	7.8748	1.3291	0.1844
GRADUATE	-5.8031	1.7835	-1.0410	0.2984

Graduate vs Bachelor

There is not enough evidence to suggest that there is a difference in average household chore hours between couples where the wife has obtained a graduate degree and couples where the wife has obtained a bachelor degree (two-tailed t-test. p-value = 0.298). It is estimated that couples where the wife has a graduate degree does on average 2.01 hours less of household chores than couples where the wife has a bachelors degree. With 95% confidence, couples where the wife has obtained a graduate degree do between 5.80 hours less and 1.78 hours more in household chores than couples where the wife has obtained a bachelor degree.

Bachelor vs Junior College

There is not enough evidence to suggest that there is a difference in average household chore hours between couples where the wife has obtained a bachelor degree and couples where the wife has obtained a junior college degree (two-tailed t-test. p-value = 0.184). It is estimated that couples where the wife has a bachelor degree does on average 3.18 hours more in household chores than couples where the wife has a Junior college degree. With 95% confidence, couples where the wife has obtained a bachelors degree do between 1.52 hours less and 7.87 hours more in household chores than couples where the wife has obtained a junior college degree.

Junior College vs High School

There is moderate evidence suggesting that there is a difference in average household chore hours between couples where the wife has obtained a junior college degree and couples where the wife has obtained a high school degree (two-tailed t-test. p-value = 0.0395). It is estimated that couples where the wife has a junior college degree does on average 4.60 hours less in household chores than couples where the wife has a high school degree. With 95% confidence, couples where the wife has obtained a junior college degree do between 8.99 hours less and 0.222 hours less of household chores than couples where the wife has obtained a high school degree.

Final Thoughts

There appears to be some link between the average difference in hours spent on household chores between husbands and wives though it mostly appears between Junior College and High School level education. Given that this is a random sample from a population of american couples where the wife has at least a high school degree, inferences about this broader population can be made. Wives who have at least a junior college education tend to do less housework than those who only have a high school education. One possible explanation would be that those with only a high school education do not have enough credentials to receive a higher salary so they may be more likely to be stay at home moms and thus doing more household chores. It would be interesting to combine this with the household information (number of children) to see if there is a correlation between the number of children and the wife's education level.