Problem Set 3, Solutions

Stats 506, Fall 2018

Due: Monday November 19, 5pm

R scripts and source code for this document can be found at the Stats506_F18 repo (https://github.com/jbhender/Stats506_F18/blob/master/solutions/PS3).

Question 1 [60 points]

First, repeat question 3 parts a-c from problem set 1 using `data.table` for all computations and data manipulations.

Then, formulate and state a question answerable using the RECS data. Your question should be similar in scope to (one of) parts a-c above and should rely on one or more variables not previously used. Answer your question (using data.table) and provide supporting evidence in the form of nicely formatted graphs and/or tables.

Solution: The source code `ps3_q1.R` below as set up to allow copy and paste of previously used plotting and table commands.

```
source('./ps3_q1.R')
```

a. What percent of homes have stucco construction as the *major outside wall material* within each division? Which division has the highest proportion? Which the lowest?

```
caption = 'Proportion of homes with stucco construction within each census division in 2015. Es
timates are based on the residential energy consumption survey.'
p_stucco_tab =
  p_stucco[order(-p_stucco),
           .(`Census Division` = division,
             `% Stucco Homes (95% CI)` = sprintf('%4.1f%% (%4.1f, %4.1f)',
                                         100*p_stucco, 100*lwr, 100*upr)
            )]
p_stucco_tab %>%
  knitr::kable( align = 'r', caption = caption)
```

Proportion of homes with stucco construction within each census division in 2015. Estimates are based on the residential energy consumption survey.

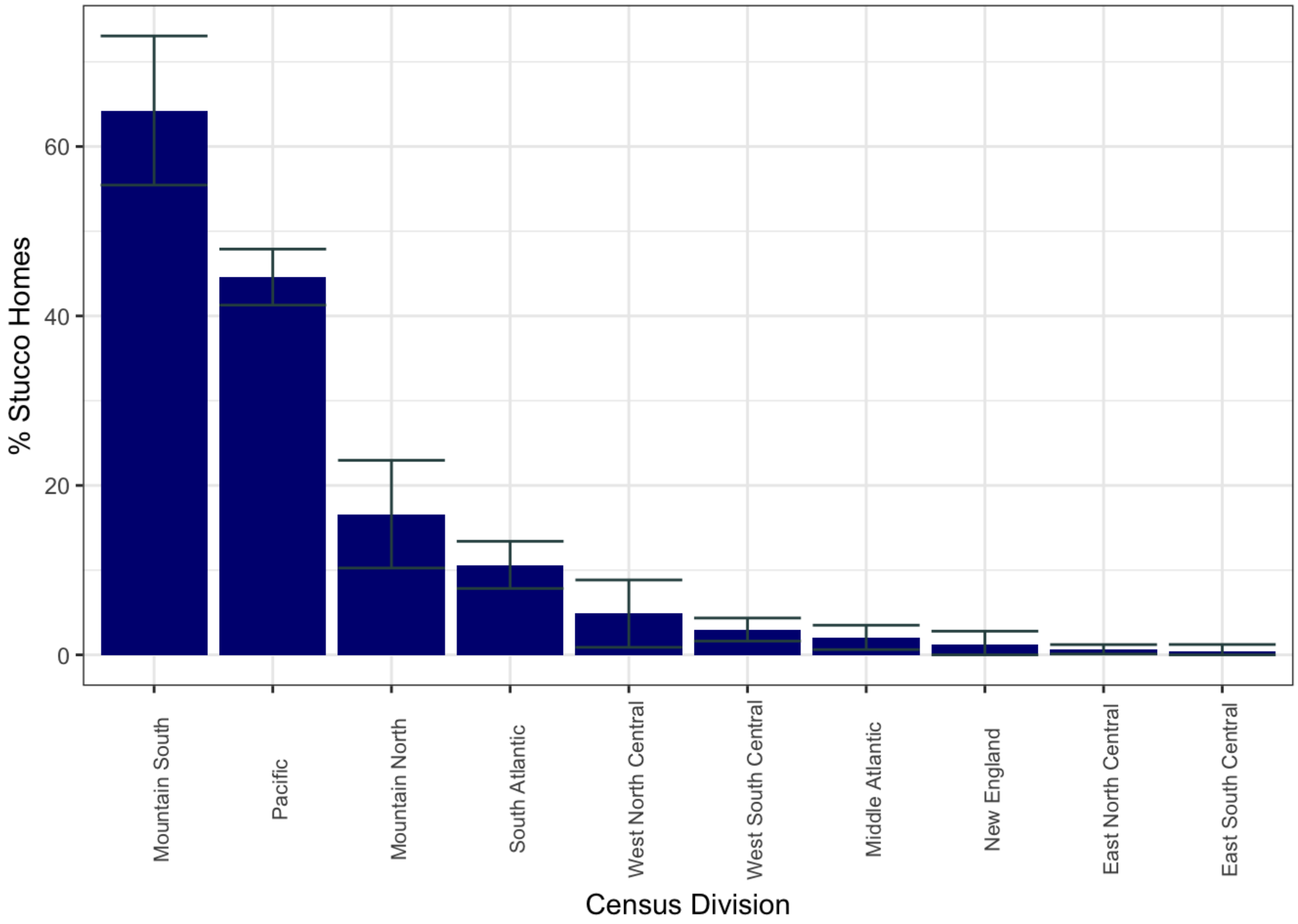
Census Division	% Stucco Homes (95% CI)
Mountain South	64.2% (55.4, 73.0)
Pacific	44.6% (41.3, 47.9)
Mountain North	16.6% (10.2, 23.0)
South Atlantic	10.6% (7.8, 13.4)
West North Central	4.9% (0.9, 8.8)
West South Central	3.0% (1.6, 4.3)
Middle Atlantic	2.1% (0.6, 3.5)

Census Division	% Stucco Homes (95% CI)
New England	1.2% (0.0, 2.8)
East North Central	0.7% (0.1, 1.2)
East South Central	0.4% (0.0, 1.2)

```
cap = ' Estimated percent of homes within each census division with major wall type of stucco.'

p_stucco = p_stucco[order(-p_stucco)]
p_stucco[ ,Division := factor(as.character(division),  as.character(division)) ]

p_stucco %>%
  ggplot( aes( x = Division,  y = 100*p_stucco) ) +
  geom_col( fill = 'navy' ) +
  geom_errorbar( aes( ymin = 100*lwr,  ymax = 100*upr),
                col = 'darkslategrey') +
  theme_bw() +
  ylab('% Stucco Homes') +
  xlab('Census Division') +
  theme( axis.text.x = element_text(size = 8,  angle = 90))
```



Estimated percent of homes within each census division with major wall type of stucco.

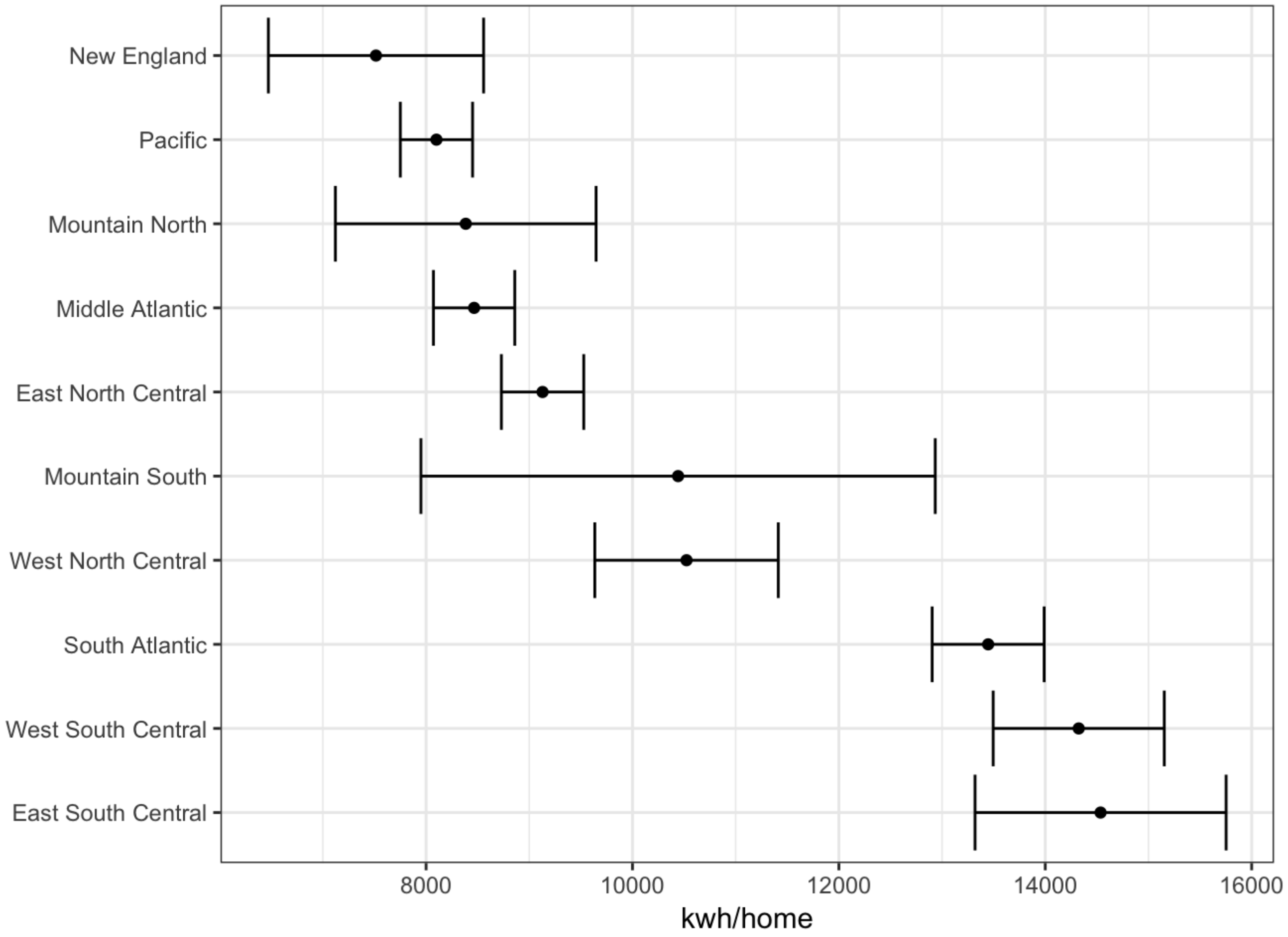
b. What is the average total electricity usage in kilowatt hours in each division? Answer the same question stratified by urban and rural status.

```
cap = 'Average annual electricity utilization by Census Division in kwh/home.'
# Multiplier for 95% CI
m = qnorm(.975)
# Pretty printing
pwc = function(x) format(round(x), big.mark = ',')
kwh[order(-est),
    .(`Census Division` = division,
      `Average Electricity Usage, kwh/home (95% CI)` =
       sprintf('%s, (%s - %s)', pwc(est), pwc(est - m*se), pwc(est + m*se) )
     ) ] %>%
 knitr::kable( align = 'r', caption = cap)
```

Average annual electricity utilization by Census Division in kwh/home.

Census Division	Average Electricity Usage, kwh/home (95% CI)
East South Central	14,536, (13,320 - 15,752)
West South Central	14,324, (13,495 - 15,153)
South Atlantic	13,447, (12,904 - 13,989)
West North Central	10,524, (9,635 - 11,413)
Mountain South	10,442, (7,950 - 12,934)
East North Central	9,129, (8,730 - 9,528)
Middle Atlantic	8,465, (8,071 - 8,860)
Mountain North	8,384, (7,121 - 9,648)
Pacific	8,100, (7,750 - 8,450)
New England	7,515, (6,472 - 8,557)

```
cap = 'Estimated average annual electricity usage in khw/home for
each of 10 census divisions.'
kwh = kwh[order(-est)]
kwh[ , div := factor(as.character(division), as.character(division))]
kwh %>%
  ggplot( aes(x = div, y = est) ) +
  geom_point() +
  geom_errorbar( aes(ymin = lwr, ymax = upr)) +
  coord_flip() +
  theme_bw() +
  ylab('kwh/home') +
  xlab('')
```



Estimated average annual electricity usage in khw/home for each of 10 census divisions.

```
cap = 'Average electricity utilization in kwh per home for urban and rural areas witihin each c
ensus division.'
# Order by simple average usage
kwh_div_urban[, div_avg := mean(est), division]
kwh_div_urban = kwh_div_urban[order(-div_avg)]
kwh_div_urban[ , div := factor(division, levels = unique(division) ) ]

# Table
kwh_div_urban[,
 .(`Census Division` = div,
   ci = sprintf('%s, (%6s - %6s)', pwc(est), pwc(est - m*se), pwc(est + m*se)),
   Rurality = ifelse(urban, 'Urban, kwh/home (95% CI)',
                    'Rural, kwh/home (95% CI)')
  )] %>%
  dcast(`Census Division` ~ Rurality, value.var = 'ci') %>%
  knitr::kable( align  = 'r', cap = cap)
```

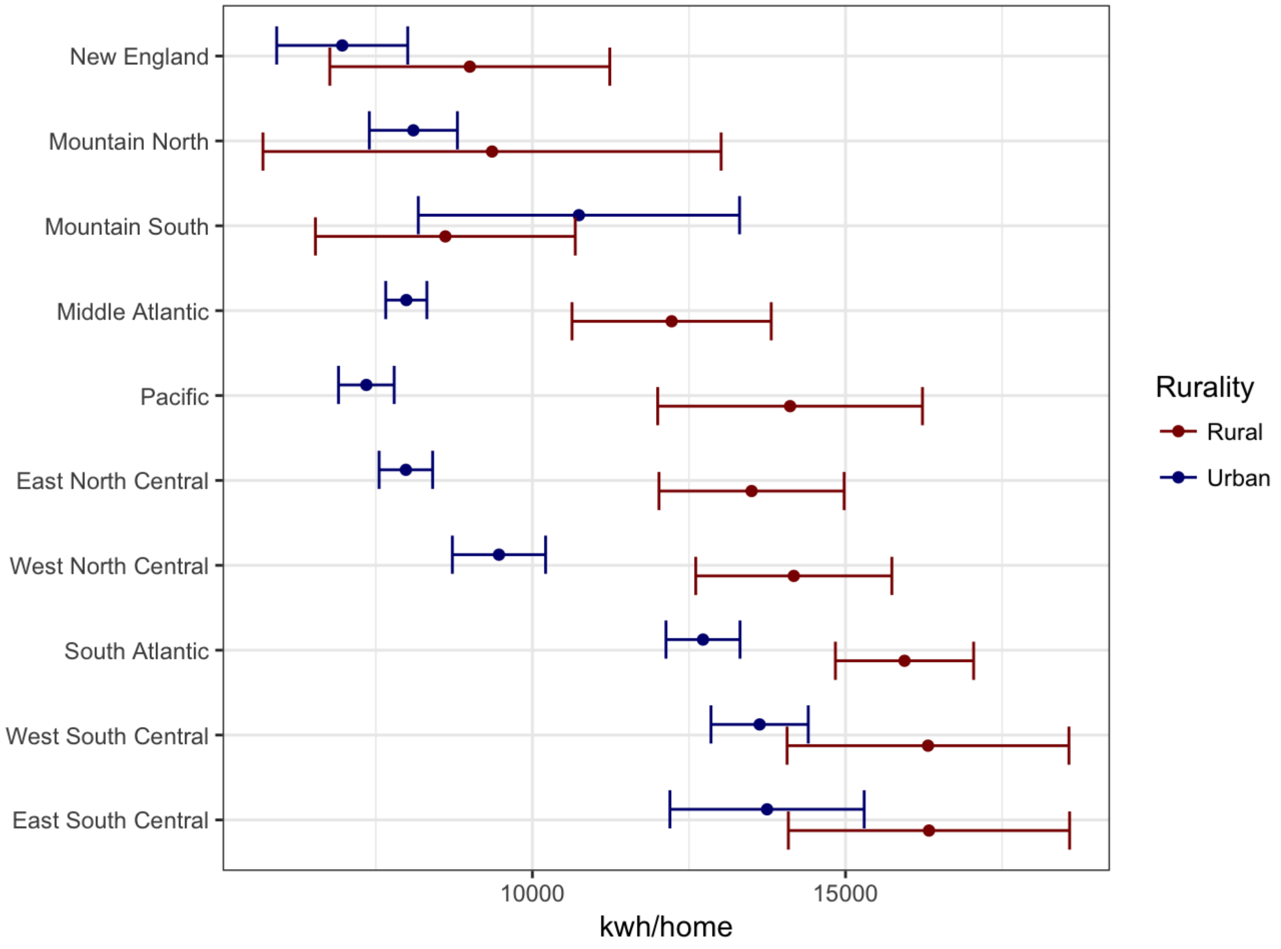
Average electricity utilization in kwh per home for urban and rural areas witihin each census division.

Census Division	Rural, kwh/home (95% CI)	Urban, kwh/home (95% CI)
East South Central	16,333, (14,088 - 18,578)	13,747, (12,197 - 15,298)
West South Central	16,317, (14,067 - 18,567)	13,629, (12,852 - 14,405)
South Atlantic	15,942, (14,839 - 17,045)	12,725, (12,134 - 13,316)
West North Central	14,174, (12,608 - 15,740)	9,467, (8,722 - 10,211)

Census Division	Rural, kwh/home (95% CI)	Urban, kwh/home (95% CI)
East North Central	13,500, (12,022 - 14,978)	7,980, (7,552 - 8,408)
Pacific	14,115, (12,001 - 16,229)	7,349, (6,905 - 7,793)
Middle Atlantic	12,223, (10,633 - 13,814)	7,987, (7,659 - 8,316)
Mountain South	8,610, (6,536 - 10,685)	10,743, (8,178 - 13,308)
Mountain North	9,356, (5,698 - 13,014)	8,099, (7,396 - 8,803)
New England	9,001, (6,766 - 11,236)	6,964, (5,918 - 8,010)

```
cap = 'Estimated average annual electricity usage in khw/home for
rural and urban areas in each of 10 census divisions.'
kwh_div_urban[,     Rurality := ifelse(urban, 'Urban', 'Rural')]

kwh_div_urban %>%
  ggplot( aes(x = div, y = est, color = Rurality) ) +
  geom_point( position = position_dodge(.5) ) +
  geom_errorbar( aes(ymin = lwr, ymax = upr),
                 position = position_dodge(.5)
  ) +
  scale_color_manual( values = c('navy', 'darkred')[2:1]) +
  coord_flip() +
  theme_bw() +
  ylab('kwh/home') +
  xlab('')
```



c. Which division has the largest disparity between urban and rural areas in terms of the proportion of homes with internet access?

```
cap = "Urban and rural disparity in internet access for the ten US Census Division in 2015. "
internet_disp %>%
  knitr::kable( align = 'r', caption = cap )
```

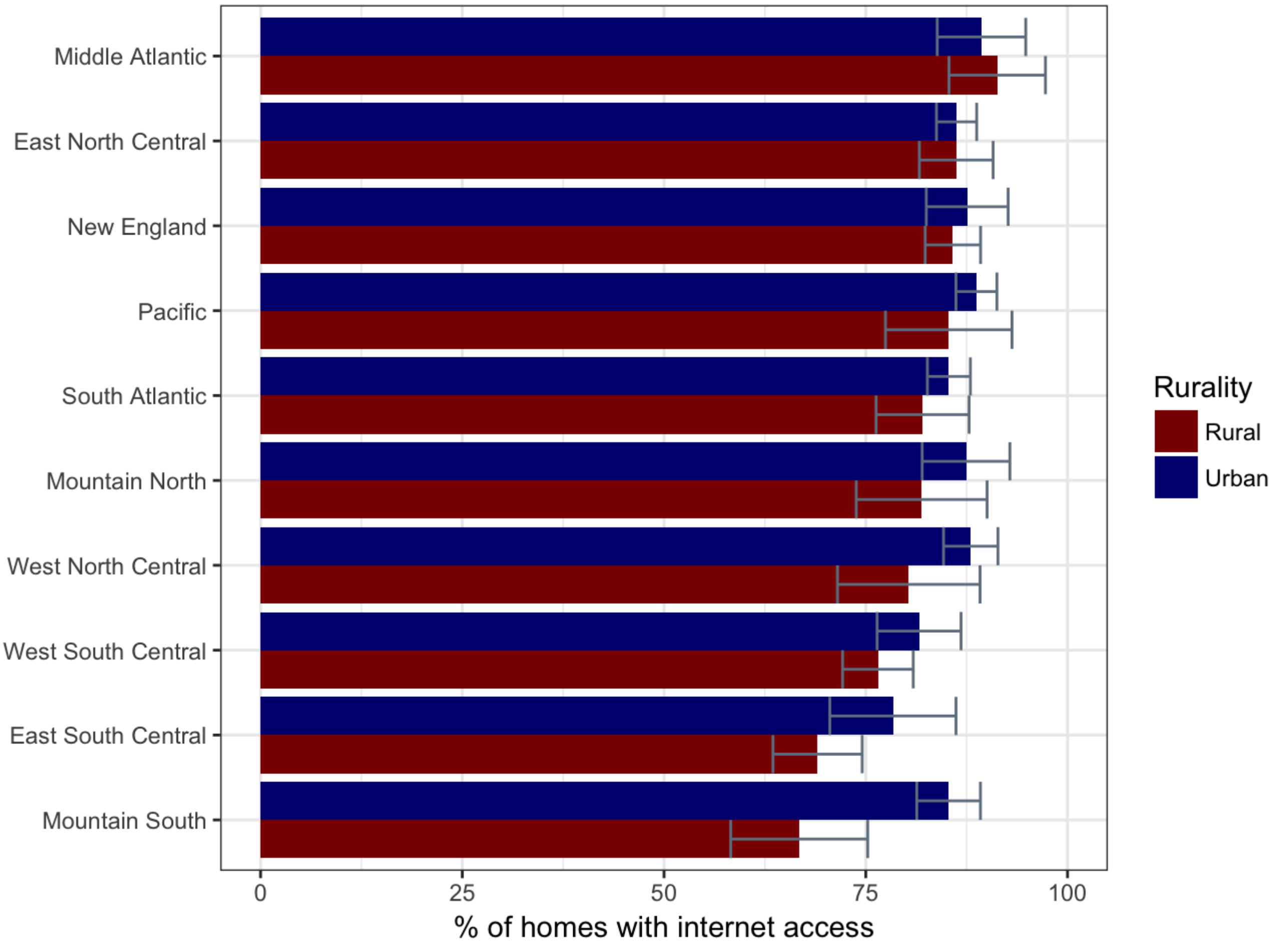
Urban and rural disparity in internet access for the ten US Census Division in 2015.

Division	Rural	Urban	Diff
Mountain South	66.7 (58.3, 75.2)	85.3 (81.3, 89.2)	18.5% (7.2, 29.8)
East South Central	69.0 (63.5, 74.6)	78.4 (70.5, 86.2)	9.3% (-1.4, 20.1)
West North Central	80.3 (71.5, 89.2)	88.0 (84.6, 91.4)	7.7% (-2.5, 17.8)
Mountain North	81.9 (73.8, 90.0)	87.4 (82.0, 92.9)	5.5% (-6.2, 17.2)
West South Central	76.5 (72.1, 80.9)	81.6 (76.4, 86.8)	5.1% (-2.3, 12.5)
Pacific	85.3 (77.4, 93.1)	88.7 (86.2, 91.2)	3.4% (-4.5, 11.4)
South Atlantic	82.0 (76.3, 87.8)	85.3 (82.6, 88.0)	3.3% (-3.5, 10.1)
New England	85.8 (82.4, 89.2)	87.6 (82.5, 92.6)	1.8% (-2.5, 6.0)
East North Central	86.2 (81.6, 90.8)	86.3 (83.8, 88.7)	0.0% (-5.3, 5.4)
Middle Atlantic	91.3 (85.3, 97.3)	89.3 (83.9, 94.8)	-1.9% (-9.1, 5.2)

In the Mountain South division there is an 18.5% disparity between Urban and Rural internet access. This is approximately twice as large as the next largest estimated disparity and the only estimate whose confidence interval does not include zero.

```
internet_ru[, `:=`(Rurality = ifelse(urban, 'Urban', 'Rural'),
                    division = factor( as.character(division),
            as.character(
              {internet_disp[order(Rural)]}$Division)))]

internet_ru %>%
  ggplot( aes(x = division, y = est, fill = Rurality) ) +
  geom_col( position = position_dodge() ) +
  geom_errorbar( aes(ymin = lwr, ymax = upr),
                position = position_dodge(),
                col = 'slategrey') +
  theme_bw() +
  xlab('') +
  ylab('% of homes with internet access') +
  ylim(c(0, 100)) +
  coord_flip() +
  scale_fill_manual(values = c('darkred', 'navy'))
```



Question 2 [25 points]

In this question you will design a Monte Carlo study in R to compare the performance of different methods that adjust for multiple comparisons (https://xkcd.com/882/). You can read more about each of these methods by referring to `help(p.adjust)` in R and the references listed there.

Throughout this question, let $n = 1000$, $p = 100$ and

$$\beta_i = \begin{cases} 1 & i \in \{1, \dots, 10\}, \\ 0 & \text{else.} \end{cases}$$

Let $X \in \mathbb{R}^{n \times p}$ with $X \sim N(0_p, \Sigma)$ and $Y \sim N(X\beta, \sigma^2 I_n)$ where I_n is an $n \times n$ identify matrix and Σ is a $p \times p$, symmetric, positive definite covariance matrix.

a. Write a function that accepts matrices `X` and `beta` and returns a `p` by `mc_rep` matrix of p-values corresponding to p-values for the hypothesis tests:

$$H_0 : \beta_i = 0, H_1 : \beta_i \neq 0.$$

In addition to `X` and `beta` your function should have arguments `sigma` (σ) and `mc_rep` controlling the error variance of Y and number of Monte Carlo replicates, respectively. Your function should solve the least-squares problems using the QR decomposition of $X'X$. This decomposition should only be computed once each time your function is called.

i. Refer to the course notes to find $\hat{\beta}$.

ii. Use Y and $\hat{Y} = X\hat{\beta}$ to estimate the error variance for each Monte Carlo trial m:

$$\hat{\sigma}^2_m = \frac{1}{n-p} \sum_{i=1}^{n} (Y_{im} - \hat{Y}_{im})^2$$

iii. Use the result from ii and the QR decomposition to find the variance of $\hat{\beta}_i$, $v_i = \hat{\sigma}^2(X'X)^{-1}_{ii}$. [Note: you will need to do some algebra to determine how to compute $(X'X)^{-1}$ using Q and R. Or you can use the function `chol2inv()` .]

iv. Form $Z_i = \hat{\beta}_i/\sqrt{v_i}$ and find $p = 2(1 - \Phi^{-1}(|Z_i|))$.

Test your function with a specific X and Y by comparing to the output from appropriate methods applied to the object returned by `lm(Y ~ 0 + X)` . It’s okay if there is some finite precision error less than ~1e-3 in magnitude. Hint: use `set.seed()` to generate the same Y inside and outside the scope of the function for the purpose of testing.

b. Choose Σ and σ as you like. Use the Cholesky factorization of Σ to generate a single X. Pass X, β, and σ to your function from the previous part.

c. Write a function `evaluate` that takes your results and a set of indices where $\beta \neq 0$, and returns Monte Carlo estimates for the following quantities:

- The family wise error rate
- The false discovery rate
- The sensitivity
- The specificity.

See this page (https://en.wikipedia.org/wiki/Sensitivity_and_specificity#Sensitivity_index) for additional details.

d. Apply your function from the previous part to the matrix of uncorrected P-values generated in part B. Use the function `p.adjust()` to correct these p-values for multiple comparisons using ‘Bonferroni’, ‘Holm’, ‘BH’ (Benjamini-Hochberg), and ‘BY’ (Benjamini-Yekuteli). Use your `evaluate()` function for each set of adjusted p-values.

e. Produce one or more nicely formatted graphs or tables reporting your results. Briefly discuss what you found.

Solution:

```
source(’./ps3_q2.R’)

cap = ’_Results from a Monte Carlo study comparing p-value correction methods in the context of
multiple regression._’

ps3_q2result[,.(Method = method, FWER = fwer, FDR = fdr,
                Sensitivity = sens, Specificity = spec)] %>%
  knitr::kable(caption = cap)
```

Results from a Monte Carlo study comparing p-value correction methods in the context of multiple regression.

Method	FWER	FDR	Sensitivity	Specificity
Holm	0.045 (0.041, 0.049)	0.012 (0.011, 0.014)	0.326 (0.323, 0.329)	0.999 (0.999, 1.000)
Bonferroni	0.043 (0.039, 0.047)	0.012 (0.011, 0.013)	0.322 (0.319, 0.325)	1.000 (0.999, 1.000)
BH	0.246 (0.238, 0.255)	0.045 (0.043, 0.047)	0.515 (0.511, 0.519)	0.997 (0.996, 0.997)
BY	0.035 (0.032, 0.039)	0.009 (0.008, 0.010)	0.285 (0.281, 0.289)	1.000 (1.000, 1.000)

Among the mehtods controlling the FWER, the Holm and Bonferroni methods are nearly identical and both control the FWER at slightly better than the nominal level. The BH and BY methods control the FDR; BH has FDR near the nominal .05 rate while BY is much more conservative. Because it gives up control of the FWER, the BH method has much higher sensitivity while BY has the lowest sesnsitivity as it guards against any form of dependence.

Question 3 (Optional) [30 points]

This is a bonus question related to problem 6 from the midterm. First, review the script written in Stata available here (https://github.com/jbhender/Stats506_F18/tree/master/solutions/PS3). In this question, you will work through various options for translating this analysis into R. You may submit all or some of these, but each part must be entirely correct to earn the points listed.

a. Write a translation using `data.table` for the computations. [5 pts]

b. Write a function to compute the univariate regression coefficient by group for arbitrary dependent, independent, and grouping variables. Use `data.table` for computations within your function. Test your function by showing it produces the same results as in part a. [10 pts]

c. Compute the regression coefficients using the dplyr verb `summarize_at()`. [5 pts]

d. Write a function similar to the one in part b to compute arbitrary univariate regression coefficients by group. Use `dplyr` for computations within your function. You should read the “Programming with dplyr” vignette at `vignette('programming', 'dplyr')` before attempting this. Warning: this may be difficult to debug! [10 points]

Solution: See `ps3_q3.R` for details.