Microsoft and ACS

Renata Gerecke

7/8/2021

Microsoft collects its own data on broadband internet access, reporting the percent of people in a county who are actually accessing "high-speed" internet (25 mbps upload/3 mbps download). These numbers could vary from the FCC estimates for several reasons, including:

1. The ISPs are reporting that they serve areas that they do not actually support;
2. The ISPs are reporting that they provide speeds that they do not actually achieve;
3. The ISPs serve the area at high speeds but at a price point that is too high for the residents to afford; or
4. The residents aren't interested in high speed internet access.

For this analysis I will be using open-source data from Microsoft that both reports the FCC estimate of availability and Microsoft's observed measure of accessibility. Then, I will use the 5-year ACS population estimates to assess the degree to which demographic characteristics are associated with high-speed internet access.

Setup

First, I read in two data files: Microsoft's data from November 2019 and the ACS 5-year estimate from 2015-2019.

```
df_ms_19 <- read_csv("../data/ms/broadband_data_2019November.csv",
                     na = "-")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ST = col_character(),
##   `COUNTY ID` = col_double(),
##   `COUNTY NAME` = col_character(),
##   `BROADBAND AVAILABILITY PER FCC` = col_double(),
##   `BROADBAND USAGE` = col_double()
## )
```

```
df_ms_20 <- read_csv("../data/ms/broadband_data_2020October.csv",
                     na = "-", skip = 18)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ST = col_character(),
##   `COUNTY ID` = col_double(),
##   `COUNTY NAME` = col_character(),
##   `BROADBAND AVAILABILITY PER FCC` = col_double(),
##   `BROADBAND USAGE` = col_double()
```

```
## )
```

```
df_acs <- read_rds("../data/acs/acs_5yr_ed.rds")

df_county <- st_read("../data/tl_2019_us_county/tl_2019_us_county.shp")
```

```
## Reading layer `tl_2019_us_county' from data source
##   `/Users/rgerecke/Desktop/rthings/broadband-access/data/tl_2019_us_county/tl_2019_us_county.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 3233 features and 17 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -179.2311 ymin: -14.60181 xmax: 179.8597 ymax: 71.43979
## Geodetic CRS:  NAD83
```

Munge

The main adjustment that needs to be made to the Micorosft data is the calculation of the difference between FCC's reported broadband availability and Microsoft's observed broadband usage. A difference of .2 would indicate that the FCC has overestimated availability by 20 percentage points; a difference of -.2 would indicate an underestimate of the same magnitude.

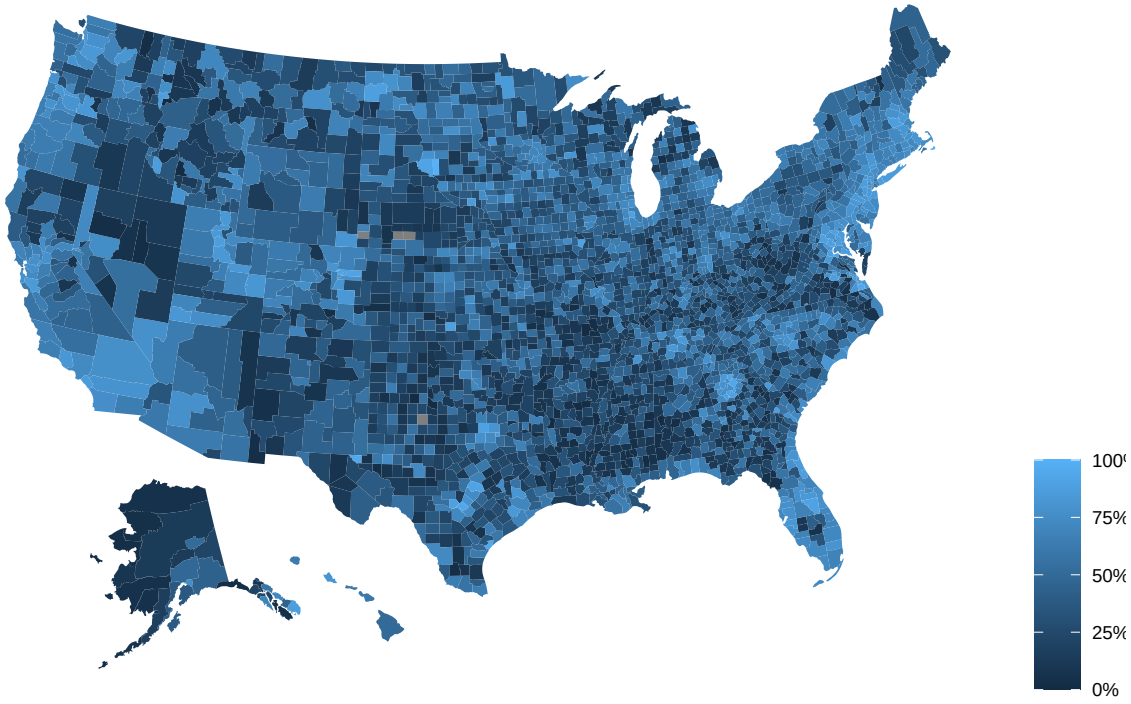
```
df_ms_mut <- bind_rows(df_ms_19, df_ms_20, .id = "year") %>%
    clean_names() %>%
    mutate(delta = broadband_availability_per_fcc - broadband_usage,
           fips = county_id,
           year = ifelse(year == 1, 2019, 2020)) %>%
    pivot_wider(
        names_from = year,
        values_from = c(starts_with("broadband"), delta),
        names_sep = "_"
    )
```

Map

```
plot_usmap(data = df_ms_mut, values = "broadband_usage_2020", color = "transparent") +
    scale_fill_gradient(label = scales::percent, limits = c(0,1)) +
    labs(title = "Percent of Residents Using Broadband Internet, October 2020",
         subtitle = "25 mbps download",
         fill = NULL,
         caption = "Source: Microsoft Broadband Data") +
    theme(legend.position = "right")
```

Percent of Residents Using Broadband Internet, October 2020

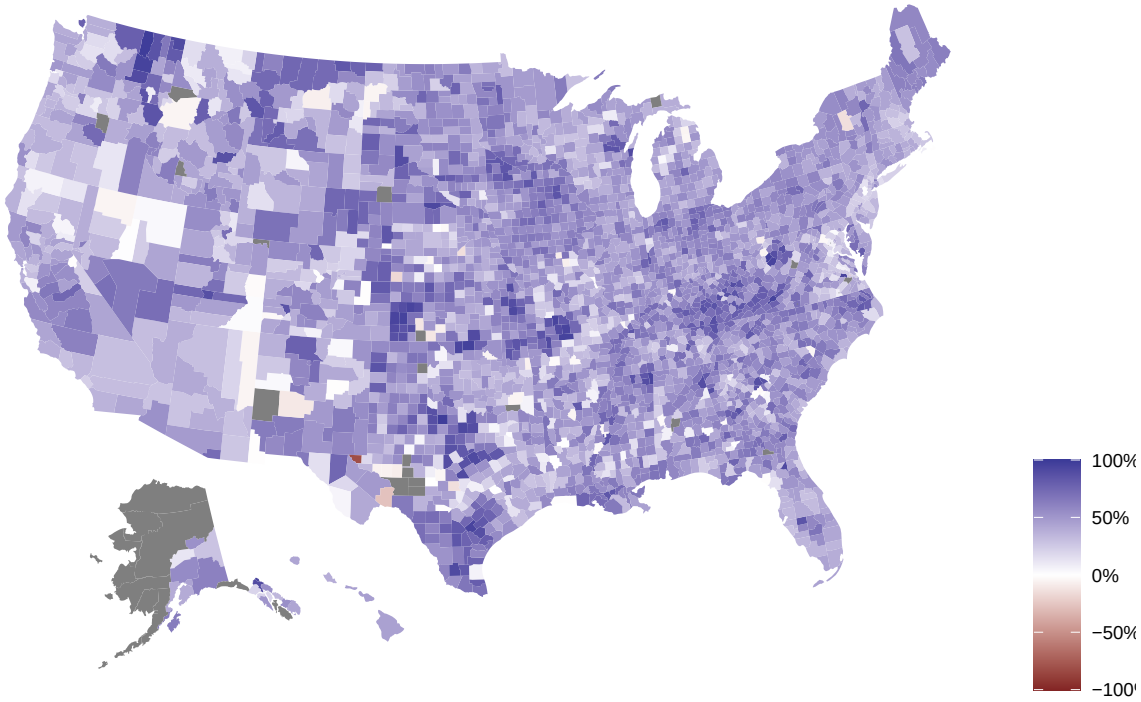
25 mbps download



Source: Microsoft Broadband Data

```
plot_usmap(data = df_ms_mut, values = "delta_2019", color = "transparent") +
    scale_fill_gradient2(label = scales::percent, limits = c(-1,1)) +
    labs(title = "Percentage Point Overestimation of Broadband Internet Access",
         subtitle = "November 2019",
         fill = NULL,
         caption = "Source: Microsoft Broadband Data") +
    theme(legend.position = "right")
```

Percentage Point Overestimation of Broadband Internet Access
November 2019

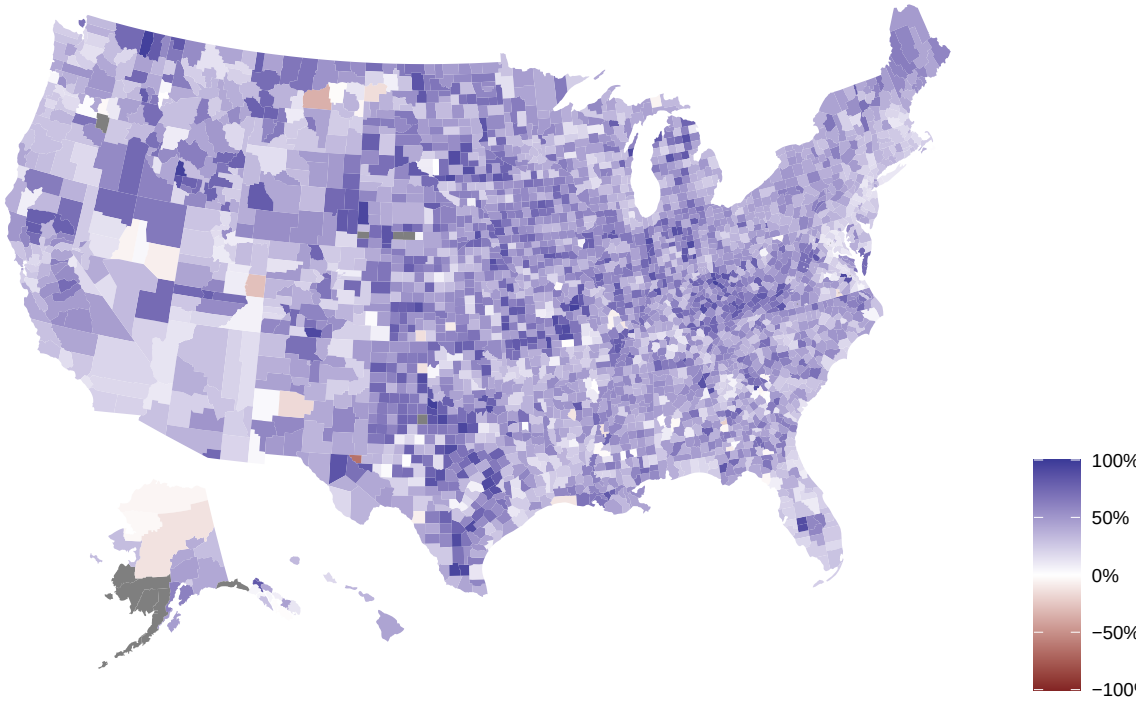


Source: Microsoft Broadband Data

```
# ggsave("../output/ms_fcc_delta.png")
```

```
plot_usmap(data = df_ms_mut, values = "delta_2020", color = "transparent") +
    scale_fill_gradient2(label = scales::percent, limits = c(-1,1)) +
    labs(title = "Percentage Point Overestimation of Broadband Internet Access",
         subtitle = "October 2020",
         fill = NULL,
         caption = "Source: Microsoft Broadband Data") +
    theme(legend.position = "right")
```

Percentage Point Overestimation of Broadband Internet Access
October 2020

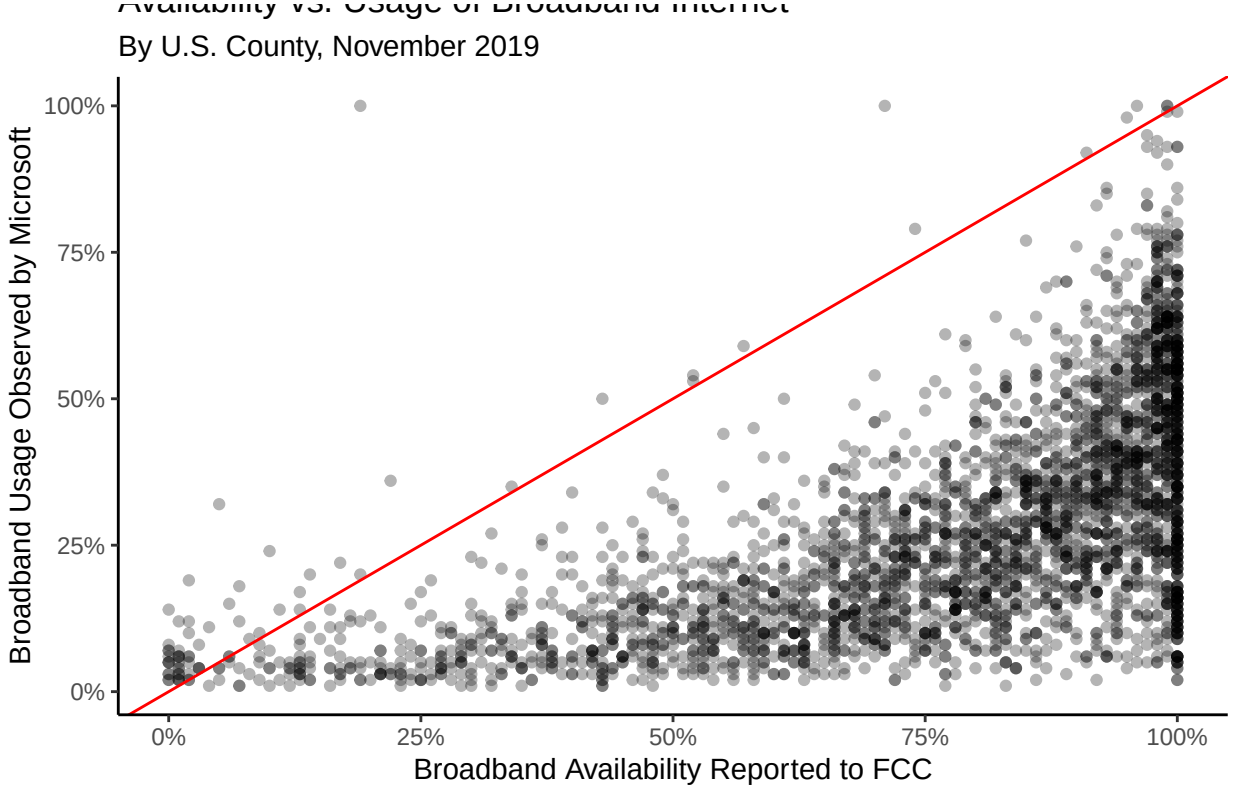


Source: Microsoft Broadband Data

Plot

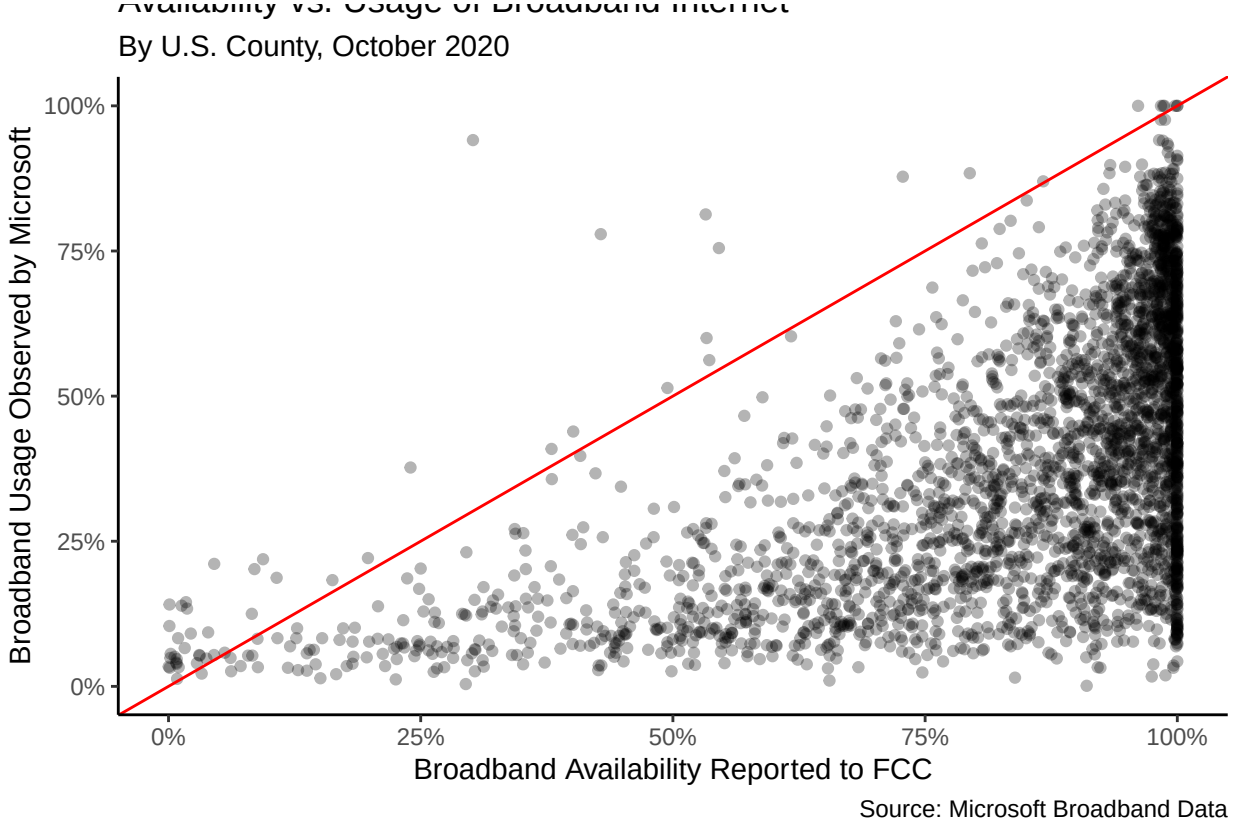
```
df_merge <- df_ms_mut %>%
    mutate(fips = as.character(county_id) %>%
               str_pad(5, side = "left", pad = "0")) %>%
    left_join(df_acs, by = "fips")

ggplot(df_merge) +
    aes(x = broadband_availability_per_fcc_2019, y = broadband_usage_2019) +
    geom_point(alpha = .3) +
    geom_abline(slope = 1, intercept = 0, color = "red") +
    scale_x_continuous(label = scales::percent) +
    scale_y_continuous(label = scales::percent) +
  labs(
    title = "Availability vs. Usage of Broadband Internet",
    subtitle = "By U.S. County, November 2019",
    caption = "Source: Microsoft Broadband Data",
    x = "Broadband Availability Reported to FCC",
    y = "Broadband Usage Observed by Microsoft"
  ) +
    theme_classic()
```

```
## Warning: Removed 41 rows containing missing values (geom_point).
```




```
ggplot(df_merge) +
    aes(x = broadband_availability_per_fcc_2020, y = broadband_usage_2020) +
  geom_point(alpha = .3) +
    geom_abline(slope = 1, intercept = 0, color = "red") +
    scale_x_continuous(label = scales::percent) +
    scale_y_continuous(label = scales::percent) +
  labs(
    title = "Availability vs. Usage of Broadband Internet",
    subtitle = "By U.S. County, October 2020",
    caption = "Source: Microsoft Broadband Data",
    x = "Broadband Availability Reported to FCC",
    y = "Broadband Usage Observed by Microsoft"
  ) +
    theme_classic()
```

```
## Warning: Removed 23 rows containing missing values (geom_point).
```



Single Regression

```
df_ms_long <- df_ms_mut %>%
  pivot_longer(
    starts_with("broadband")
  ) %>%
  mutate(year = as.character(parse_number(name)),
         source = case_when(
           str_detect(name, "fcc") ~ "fcc",
           TRUE ~ "ms"
         ),
         delta_2019 = NULL,
         delta_2020 = NULL,
         name = NULL,
         fips = as.character(county_id) %>%
               str_pad(5, side = "left", pad = "0")) %>%
    left_join(df_acs, by = "fips") %>%
  left_join(select(st_drop_geometry(df_county), fips = GEOID, area = ALAND), by = "fips") %>%
  mutate(pop_dense = tot_pop / (area / 2589988))
```

```
long_lm <- list(
  long_unadj = lm(
    value ~ source * year,
    data = df_ms_long
  ),
  long_adj1 = lm(
```

```
    value ~ source * year + race_blacknh + race_asiannh + race_othernh + race_hispanic,
    data = df_ms_long
  ),
  long_adj2 = lm(
    value ~ source * year + race_blacknh + race_asiannh + race_othernh + race_hispanic + male + age_18t
    data = df_ms_long
  )
)



stargazer(
  long_lm,
  covariate.labels = c("Source: Microsoft", "Year: 2020", "\\% Black Non-Hispanic", "\\% Asian Non-Hispa
  column.labels = c("Unadjusted", "Adjusted by Race", "Adjusted by Demos"),
  dep.var.labels = "\\% Residents in County with Broadband Access"
  # out = "../output/long_lm.htm"
)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 15, 2021 - 16:13:35

Original Regressions (Unused)

Match ACS data

TODO: Running all these models separately is interesting and says a lot but we can probably estimate the impact of being self-reported vs observed and year-on-year by combining all the data & running a single regression? To try later today.

```
all_unadjusted <- list(
  obs_19 = {
    lm(broadband_usage_2019 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic, data = df_merg
  },
    est_19 = {
    lm(broadband_availability_per_fcc_2019 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic
  } ,
  obs_20 = {
    lm(broadband_usage_2020 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic, data = df_merg
  },
  est_20 = {
    lm(broadband_availability_per_fcc_2020 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic
  }
)

stargazer(
  all_unadjusted,
  type = "html",
  covariate.labels = c("% Black Non-Hispanic", "% Asian Non-Hispanic", "% Other Non-Hispanic", "% Hispa
  dep.var.labels = c("Observed by MS, 2019", "Estimated by FCC, 2019", "Observed by MS, 2020", "Estimat
  out = "../output/unadjusted.htm"
)
```

Table 1:

	Dependent variable:		
	% Residents in County with Broadband Access		
	Unadjusted	Adjusted by Race	Adjusted by Demos
	(1)	(2)	(3)
Source: Microsoft	−0.488***	−0.487***	−0.487***
	(0.005)	(0.005)	(0.005)
Year: 2020	0.074***	0.074***	0.074***
	(0.005)	(0.005)	(0.005)
% Black Non-Hispanic		−0.188***	−0.314***
		(0.013)	(0.013)
% Asian Non-Hispanic		2.626***	1.466***
		(0.068)	(0.072)
% Other Non-Hispanic		−0.314***	−0.306***
		(0.024)	(0.023)
% Hispanic		0.001	0.039***
		(0.013)	(0.014)
% Male			−2.134***
			(0.080)
% Age 18-64			0.469***
			(0.078)
% Age 65+			−1.056***
			(0.061)
% Below Poverty Line			−1.637***
			(0.118)
Population Density (persons per sq. mi.)			0.00000***
			(0.00000)
Source: Microsoft * Year: 2020	0.037***	0.036***	0.036***
	(0.008)	(0.007)	(0.007)
Constant	0.767***	0.760***	1.855***
	(0.004)	(0.004)	(0.059)
Observations	12,509	12,508	12,508
R^2	0.551	0.608	0.658
Adjusted R^2	0.551	0.608	0.658
Residual Std. Error	0.216 (df = 12505)	0.202 (df = 12500)	0.188 (df = 12495)
F Statistic	5,122.951*** (df = 3; 12505)	2,768.811*** (df = 7; 12500)	2,006.337*** (df = 12; 12
Note:			*p<0.1; **p<0.05; ***p<

```
all_adjusted <- list(
  obs_19 = {
    lm(broadband_usage_2019 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic + male + age_18
  },
  est_19 = {
    lm(broadband_availability_per_fcc_2019 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic
  },
  obs_20 = {
    lm(broadband_usage_2020 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic + male + age_18
  },
  est_20 = {
    lm(broadband_availability_per_fcc_2020 ~ race_blacknh + race_asiannh + race_othernh + race_hispanic
  }
)
stargazer(
  all_adjusted,
  type = "html",
  covariate.labels = c("% Black Non-Hispanic", "% Asian Non-Hispanic", "% Other Non-Hispanic", "% Hispa
  dep.var.labels = c("Observed by MS, 2019", "Estimated by FCC, 2019", "Observed by MS, 2020", "Estimat
  out = "../output/adjusted.htm"
)
```