Piecewise Regression

Andrew Zieffler
November 06, 2020

These notes draw from Berk (2016).

Preparation

The data in *tokyo-water-use.csv* were collected due to concerns about the provision of potable water to large metropolitan areas, which is a potential problem resulting from human-induced climate change. The data consist of 27 years worth of residential water use (measured in 1000s of cubic feet).

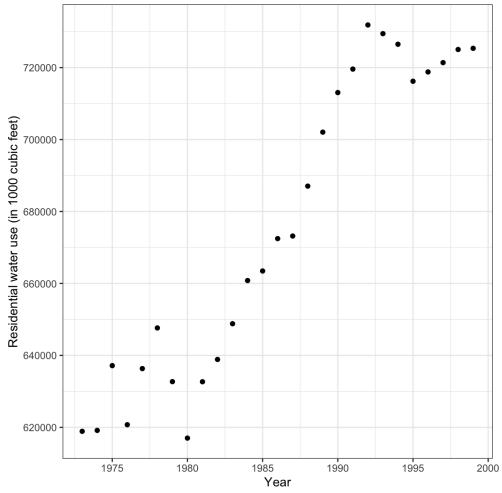
```
# Load libraries
library(broom)
library(patchwork)
library(tidyverse)


# Import data
tokyo = read_csv("../data/tokyo-water-use.csv")
head(tokyo)
```

```
## # A tibble: 6 x 2
##    year water_use
##   <dbl>     <dbl>
## 1  1973    618899
## 2  1974    619154
## 3  1975    637161
## 4  1976    620731
## 5  1977    636335
## 6  1978    647635
```

A plot of the data suggests that the water use is a non-linear function of time.

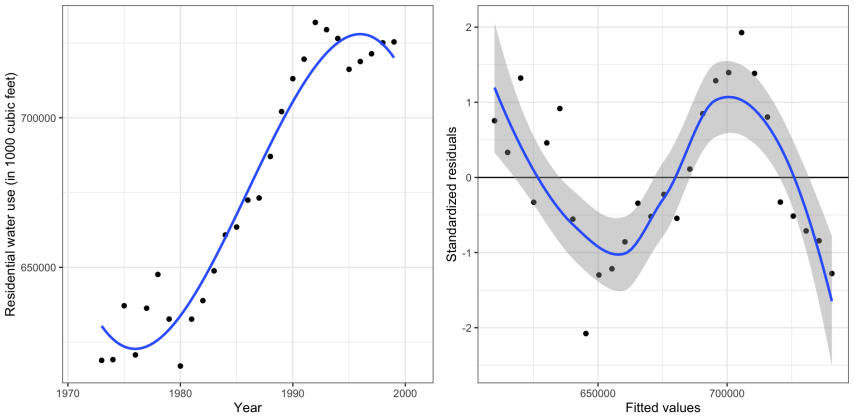
```
ggplot(data = tokyo, aes(x = year, y = water_use)) +
  geom_point() +
  scale_x_continuous(name = "Year", breaks = seq(from = 1970, to = 2005, by
        = 5)) +
  scale_y_continuous(
    name = "Residential water use (in 1000 cubic feet)",
    breaks = seq(from = 600000, to = 740000, by = 20000)
    ) +
  theme_bw()
```



The trend here suggests that residential water use was roughly the same from 1970 to 1980. After 1980, residential water use increased rapidly until about 1992, perhaps reflecting the economic and population growth during this time. Around 1992, the residential water use seems to plateau, perhaps corresponding to the introduction of more water-efficient technology.

The trend before and after these break points (at 1980 and 1992), however, seem to change quite abruptly, and it is unclear (substantively) why this is. One hypothesis is that the way in which the Tokyo Municipal Water Works collected or reported the data may have changed around those dates.

One could try fitting a third degree polynomial to these data, but the residuals from the fitted model are less than compelling. This is probably because the polynomials do not capture the abrupt changes in the relationship featured in these data.



LEFT: Water usage as a function of year. The cubic polynomial model is also shown. RIGHT: Standardized residuals versus the fitted values for the cubic polynomial model.

Piecewise Models

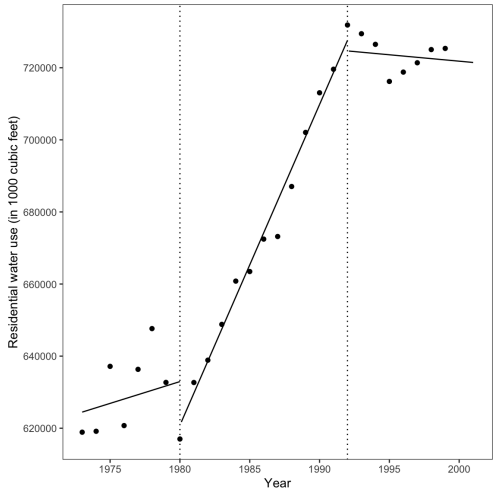
Piecewise models separate the range of the predictor(s) into K distinct regions and fit a separate model to each region. This is different from fitting the polynomial model which fits a model to a single region encompassed by the entire set of data.

For example, in the Tokyo data, we might split the data into three different regions:

- $\texttt{year} \leq 1980$
- $1980 < \texttt{year} \leq 1992$
- $\texttt{year} > 1992$

These region boundaries, 1980 and 1992, are referred to as **knots** or breakpoints. Note that in order to separate the data into K distinct regions we will need $K - 1$ knots.

While it is possible to fit models having different functional forms to each of the K regions, we typically impose the same functional form on each of the regions. For example, in our Tokyo data, it seems reasonable to fit a linear model to each of the three regions.

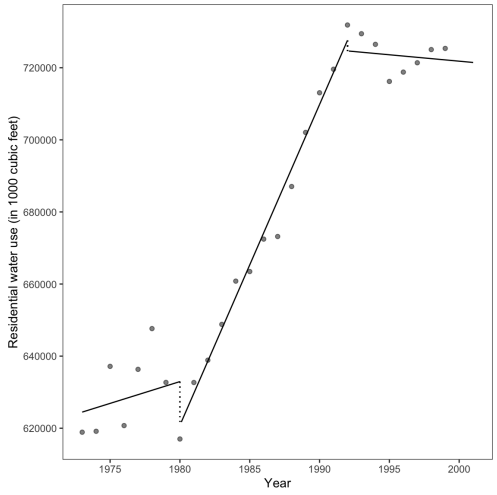


Residential water usage as a linear function of year fitted for three separate regions of data (1973–1980; 1981–1992; 1993–1999).

This plot suggests that the relationship between residential water use and time prior to 1980 (first region) is slightly positive; between 1980 and 1992 (second region) this relationship is positive and has a much higher magnitude; and after 1992 (third region) it is very slightly negative.

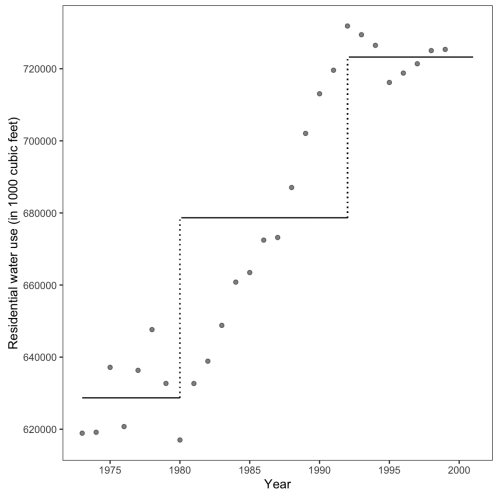
Step Functions: Discontinuous Piecewise Functions

This fitted piecewise function is *discontinuous*; the fitted lines do not intersect at the knots. Because of this, it is sometimes referred to as a step function (you have to "step" to a different y-value at each knot). This is often presented graphically by showing vertical dotted lines from the end of one fitted line to the next.



Residential water usage as a linear function of year fitted for three separate regions of data (1973–1980; 1981–1992; 1993–1999). The discontinuous steps are displayed as dotted lines in the plot.

In mathematics, the *step function* looks like a set of steps. This is shown in the figure below. In this traditional step function, we have fitted the intercept-only model to each of the K regions.



The conventional stepfunction fits the intercept-only model for three separate regions of data (1973–1980; 1981–1992; 1993–1999). The discontinuity is displayed as dotted lines in the plot.

Continuous Piecewise Models

While the linear step function seemed to model the data well, it is less mathematically compelling to have discontinuities in the fitted model. Because of this, many times we want to force the different fitted models to intersect. Having continuity also allows us to fit a single model (that incorporates the separate regions) to the dataset rather than having to split the dataset into K parts and fit a model to each part.

The steps to fitting a continuous piecewise function are:

- Identify where the knots (i.e., breakpoints) will be.
- Create a set of indicator variables that define the K regions. These are created as dummy variables (one for each knot) such that it takes a value of 1 if x is greater than the knot value and 0 otherwise.
- Fit a mean function that allows changes in both the intercept and slope (interaction) for each of the K regions.

Below we illustrate these steps for the Tokyo data.

Step 1: Identify the Knots

In our example we identified the breakpoints at 1980 and 1992. We did this by empirically examining a plot of the data and combining that with some substantive knowledge. In general we will define a set of $K-1$ knots, say $k_1, k_2, k_3, \ldots, k_{K-1}$ such that $k_1 < k_2 < k_3, < \ldots < k_{K-1}$. In our example,

$$
\begin{aligned}
k_1 &= 1980 \\
k_2 &= 1992
\end{aligned}
$$

Step 2: Create a Set of Indicator Variables

After identifying the knots, we want to create a set of indicator variables based on those knots, call them $I_1, I_2, I_3, \ldots, I_{K-1}$. Each of these indicators will be coded so that it takes a value of $x - k_i$ if x is greater than k_i and 0 otherwise. In our example,

$$
I_1 = \begin{cases} 0 & x \leq 1980 \\ x - 1980 & x > 1980 \end{cases}
$$

and

$$
I_2 = \begin{cases} 0 & x \leq 1992 \\ x - 1992 & x > 1992 \end{cases}
$$

Using R, we can create these indicator variables using the `if_else()` function.

```
tokyo = tokyo %>%
  mutate(
    I_1 = if_else(year <= 1980, 0, year - 1980),
    I_2 = if_else(year <= 1992, 0, year - 1992)
  )

tokyo
```

```
## # A tibble: 27 x 4
##     year water_use   I_1   I_2
##    <dbl>     <dbl> <dbl> <dbl>
##  1  1973    618899     0     0
##  2  1974    619154     0     0
##  3  1975    637161     0     0
##  4  1976    620731     0     0
##  5  1977    636335     0     0
##  6  1978    647635     0     0
##  7  1979    632707     0     0
##  8  1980    617011     0     0
##  9  1981    632682     1     0
## 10  1982    638877     2     0
## # … with 17 more rows
```

Step 3: Fit the Piecewise Model

Lastly, we want to fit a model for the mean function that allows changes in both the intercept and slope (interaction model) for each of the K regions. And, since we want this to be a continuous function, each adjoining set of lines need to intersect at the knots. To meet these conditions, we fit the following model:

$$Y_i = \beta_0 + \beta_1(x_i) + \beta_2(I_{1i}) + \beta_3(I_{2i}) + \ldots + \beta_{p+1}(I_{pi}) + \epsilon_i$$

Note that by substituting the indicator values for each of the three regions into this model, we end up with three equations that also intersect at the knots. For example, the fitted linear piecewise model for our Tokyo example with two knots at 1980 and 1992 is:

$$\hat{\text{Water Use}}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Year}_i) + \hat{\beta}_2(I_{1i}) + \hat{\beta}_3(I_{2i})$$

Year <= 1980

In this region, $I_1 = 0$, and $I_2 = 0$. The fitted equation here (dropping the i subscripts) would be,

$$\hat{\text{Water Use}}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Year}_i)$$

1980 < Year <= 1992

In this region, $I_1 = \text{Year}_\text{i} - 1980$, and $I_2 = 0$. The fitted equation here would be,

$$\begin{aligned}\hat{\text{Water Use}}_i &= \hat{\beta}_0 + \hat{\beta}_1(\text{Year}_i) + \hat{\beta}_2(\text{Year}_i - 1980) \\ &= \left[\hat{\beta}_0 - \hat{\beta}_2(1980)\right] + \left[\hat{\beta}_1 + \hat{\beta}_2\right](\text{Year}_i)\end{aligned}$$

Not only do these two regions have different intercepts and slopes, but the two lines intersect at the knot (Year = 1980). We can show this by seeing if the equations are equal when we substitute 1980 in for `year`.

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1(1980) &\stackrel{?}{=} \left[\hat{\beta}_0 - \hat{\beta}_2(1980)\right] + \left[\hat{\beta}_1 + \hat{\beta}_2\right](1980) \\ &\stackrel{?}{=} \hat{\beta}_0 - \hat{\beta}_2(1980) + \hat{\beta}_1(1980) + \hat{\beta}_2(1980) \\ &= \hat{\beta}_0 + \hat{\beta}_1(1980)\end{aligned}$$

Year > 1992

In this region, $I_1 = \text{Year}_\text{i} - 1980$, and $I_1 = \text{Year}_\text{i} - 1992$. The fitted equation here would be,

$$\begin{aligned}\hat{\text{Water Use}}_i &= \hat{\beta}_0 + \hat{\beta}_1(\text{Year}_i) + \hat{\beta}_2(\text{Year}_i - 1980) + \hat{\beta}_3(\text{Year} - 1992) \\ &= \left[\hat{\beta}_0 - \hat{\beta}_2(1980) - \hat{\beta}_3(1992)\right] + \left[\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3\right](\text{Year}_i)\end{aligned}$$

Again, this region has a different intercept and slope than than the other two regions, and, similarly, the second and third fitted lines intersect at the knot (Year = 1992).

Fitting the Model to the Tokyo Data

To fit the model

```
# Fit model
lm.pw = lm(water_use ~ 1 + year + I_1 + I_2, data = tokyo)

# Obtain coefficients
tidy(lm.pw)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept) 731101.   1828022.    0.400  0.693
## 2 year           -52.8      924.   -0.0571 0.955
## 3 I_1           8247.      1245.    6.62   0.000000932
## 4 I_2          -8607.      1245.   -6.91   0.000000478
```

The fitted model here is

$$\hat{\text{Water Use}}_i = 731,101 - 53(\text{Year}_i) + 8247(I_{1i}) - 8607(I_{2i})$$

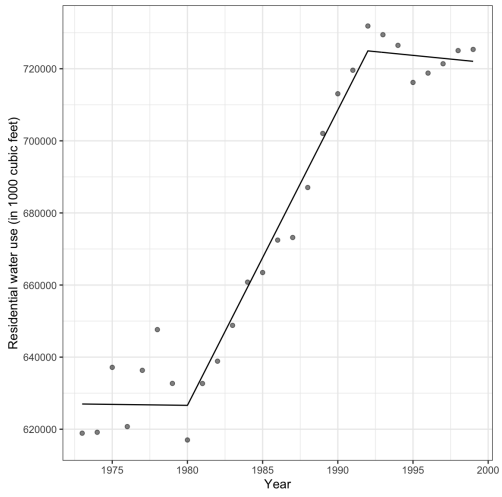
We do not interpret the coefficients, but rather use the fitted model to create a plot of the fitted model. Both of these help to interpret the relationships in the data. As Berk (2016) writes about interpretation of piecewise results,

> The point of the exercise is to superimpose the fitted values on a scatterplot so that the relationship between y and x can be more effectively visualized. The story is in the visualization not the regression coefficients (p. 60).

```
# Create data set for plot
plot_data = data.frame(
  year = seq(from = 1973, to = 1999, by = 0.1)
) %>%
  mutate(
    I_1 = if_else(year <= 1980, 0, year - 1980),
    I_2 = if_else(year <= 1992, 0, year - 1992),
  )

# Add in fitted values
plot_data = plot_data %>%
  mutate(
    yhat = predict(lm.pw, newdata = plot_data)
  )

# Plot
ggplot(data = plot_data, aes(x = year, y = yhat)) +
  geom_line() +
  geom_point(data = tokyo, aes(x = year, y = water_use), alpha = 0.5) +
  theme_bw() +
  scale_x_continuous(name = "Year", breaks = seq(from = 1970, to = 2005, by
        = 5)) +
  scale_y_continuous(
    name = "Residential water use (in 1000 cubic feet)",
    breaks = seq(from = 600000, to = 740000, by = 20000)
    )
```



Residential water usage as a continuous, piecewise linear function of year. Knots were chosen at 1980 and 1992.

The plot of the fitted equation indicates that there was rapid growth in residential water use between 1980 and 1992, and that prior to 1980, and after 1992, the growth was essentilly nil.

Graphical Inference

Since we are not interpreting any of the typical regression output, but instead focusing on the visualization of the fitted model, if we are interested in inference, we need to visualize the statistical uncertainty.

The `predict()` function takes an argument `se.fit=TRUE` that will compute standard errors for the fitted values. If this option is included, both the fitted values and standard errors are outputted as separate list elements named `fit` and `se.fit`.

```
plot_data2 = plot_data %>%
  mutate(
    yhat = predict(lm.pw, newdata = plot_data, se.fit = TRUE)$fit,
    se   = predict(lm.pw, newdata = plot_data, se.fit = TRUE)$se.fit
  )

head(plot_data2)
```

```
##     year I_1 I_2     yhat       se
## 1 1973.0   0   0 626995.8 4692.656
## 2 1973.1   0   0 626990.5 4612.411
## 3 1973.2   0   0 626985.3 4532.630
## 4 1973.3   0   0 626980.0 4453.340
## 5 1973.4   0   0 626974.7 4374.566
## 6 1973.5   0   0 626969.4 4296.336
```

We can then add columns to the plot data that give the lower- and upper-limits for a confidence envelope.

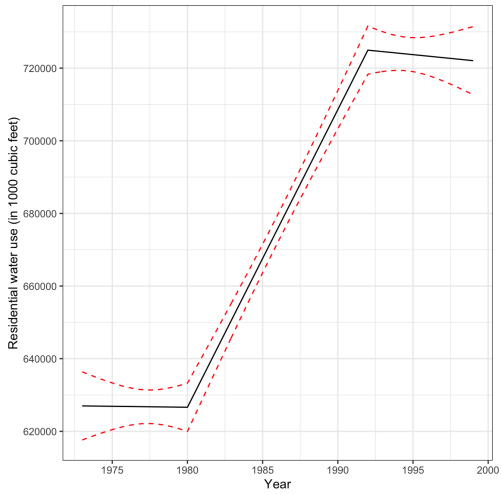
```
plot_data2 = plot_data2 %>%
  mutate(
    lower_limit = yhat - 2*se,
    upper_limit = yhat + 2*se
  )

head(plot_data2)
```

```
##     year I_1 I_2     yhat       se lower_limit upper_limit
## 1 1973.0   0   0 626995.8 4692.656    617610.5    636381.1
## 2 1973.1   0   0 626990.5 4612.411    617765.7    636215.4
## 3 1973.2   0   0 626985.3 4532.630    617920.0    636050.5
## 4 1973.3   0   0 626980.0 4453.340    618073.3    635886.7
## 5 1973.4   0   0 626974.7 4374.566    618225.6    635723.8
## 6 1973.5   0   0 626969.4 4296.336    618376.8    635562.1
```

We can then include lines showing the lower- and upper-bounds of the confidence envelope to our plot.

```
ggplot(data = plot_data2, aes(x = year, y = yhat)) +
  geom_line() +
  geom_line(aes(x = year, y = lower_limit), color = "red", linetype =
        "dashed") +
  geom_line(aes(x = year, y = upper_limit), color = "red", linetype =
        "dashed") +
  theme_bw() +
  scale_x_continuous(name = "Year", breaks = seq(from = 1970, to = 2005, by
        = 5)) +
  scale_y_continuous(
    name = "Residential water use (in 1000 cubic feet)",
    breaks = seq(from = 600000, to = 740000, by = 20000)
    )
```

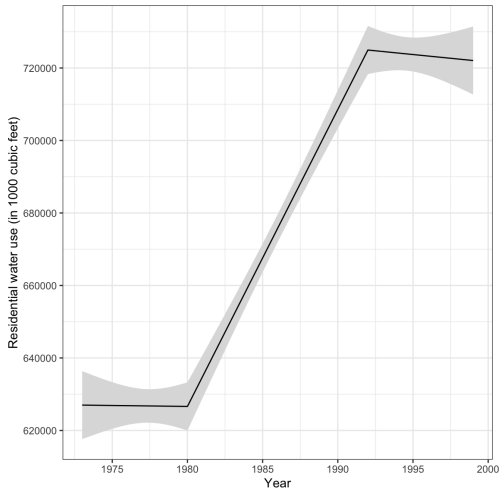


Residential water usage as a continuous, piecewise linear function of year. Knots were chosen at 1980 and 1992. The limits on the 95% confidence envelope are also displayed (red, dashed lines).

Visualizing the uncertainty helps us see that the change in residential water use prior to 1980 and after 1992 might be flat (no change); given that the a flat line is a reasonable model for those regions, as it would completely lie in the confidence bounds.

You can also use `geom_ribbon()` to sgade the confidence envelope, as seen in the syntax below.

```
ggplot(data = plot_data2, aes(x = year, y = yhat)) +
  geom_ribbon(aes(ymin = lower_limit, ymax = upper_limit), fill = "#dddddd")
        +
  geom_line() +
  theme_bw() +
  scale_x_continuous(name = "Year", breaks = seq(from = 1970, to = 2005, by
        = 5)) +
  scale_y_continuous(
    name = "Residential water use (in 1000 cubic feet)",
    breaks = seq(from = 600000, to = 740000, by = 20000)
    )
```



Residential water usage as a continuous, piecewise linear function of year. Knots were chosen at 1980 and 1992. The limits on the 95% confidence envelope are also displayed (grey, shaded area).

Re-examine the Residuals

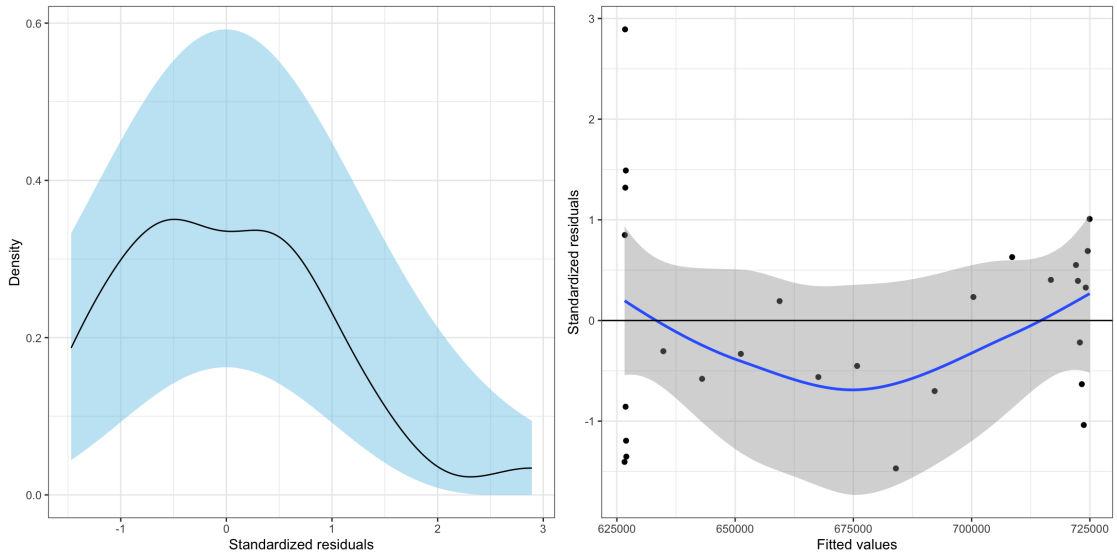
After fitting the piecewise model, it is important to re-examine the residuals. Here we use the `stat_density_confidence()` function from the **educate** package to show the expected uncertainty in the plotted density from a normal model. (To install the **educate** package see the *Installing Packages from GitHub* section in the Getting Started with R (https://zief0002.github.io/toolkit/getting-started-with-r.html#getting-started-with-r) chapter of Computational Toolkit for Educational Scientists (https://zief0002.github.io/toolkit/index.html).)

```
# Load educate package
library(educate)

# Get the residuals, fitted values, etc.
out_pw = augment(lm.pw)

# Check normality
ggplot(data = out_pw, aes(x = .std.resid)) +
  stat_density_confidence(model = "normal")  +
  stat_density(geom = "line") +
  theme_bw() +
  xlab("Standardized residuals") +
  ylab("Density")

# Check other assumptions
ggplot(data = out_pw, aes(x = .fitted, y = .std.resid)) +
  geom_point()  +
  geom_smooth() +
  geom_hline(yintercept = 0) +
  theme_bw() +
  xlab("Fitted values") +
  ylab("Standardized residuals")
```



Residual plots for the continuous, piecewise linear model fitted to the residential water usage data.

The normality assumption seems reasonably satisfied. The plot of the residuals versus the fitted values also suggest reasonable fit. At the very least, the piecewise model seems to fit better than the cubic polynomial model.

References

Berk, R. (2016). *Statistical learning from a regression perspective* (2nd ed.). Springer.