IEOR 165 – Lecture 5
Heteroscedasticity

1 Residuals Plot

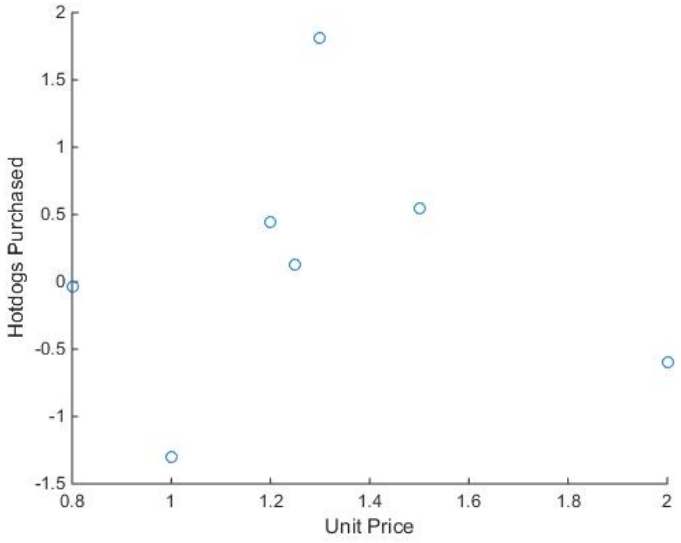
Another approach to evaluating the the quality of a linear model is to plot the residuals. That is we generate a scatter plot of the points $(x_i, y_i - \hat{y}_i)$. If the variation in the y-direction does not depend on x, and the average value in the y-direction is close to 0; then this provides evidence that the linear model accurately fits the measured data. Some examples are illuminating.

1.1 Example: Linear Model of Demand

The residuals are

$$y_1 - \hat{y}_1 = 91 - 91.0400 = -0.0400$$
$$y_2 - \hat{y}_2 = 1.8100$$
$$y_3 - \hat{y}_3 = -0.6000$$
$$y_4 - \hat{y}_4 = 0.1250$$
$$y_5 - \hat{y}_5 = 0.4400$$
$$y_6 - \hat{y}_6 = -1.3000$$
$$y_7 - \hat{y}_7 = 0.5500$$

And so plotting these gives

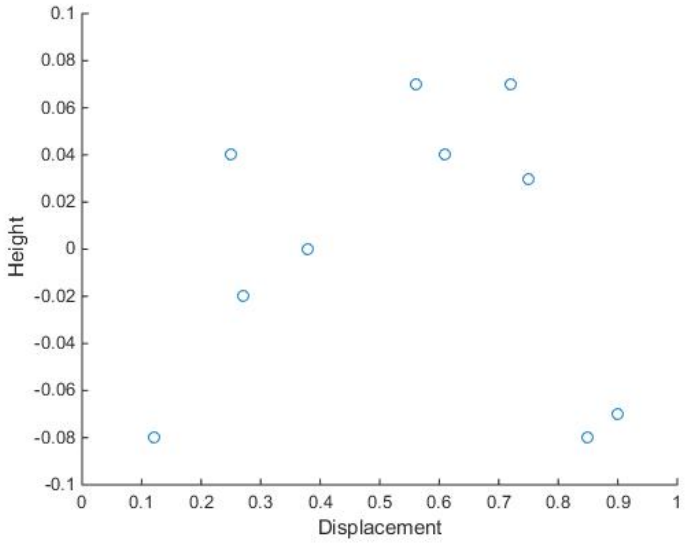


1.2 Example: Ball Trajectory

The residuals are

$$y_i - \hat{y}_i = \{0.07, 0.04, -0.08, 0.04, 0.07, -0.08, 0.00, -0.07, 0.03, -0.02\}.$$

And so plotting these gives



1.3 Example: Water Consumption

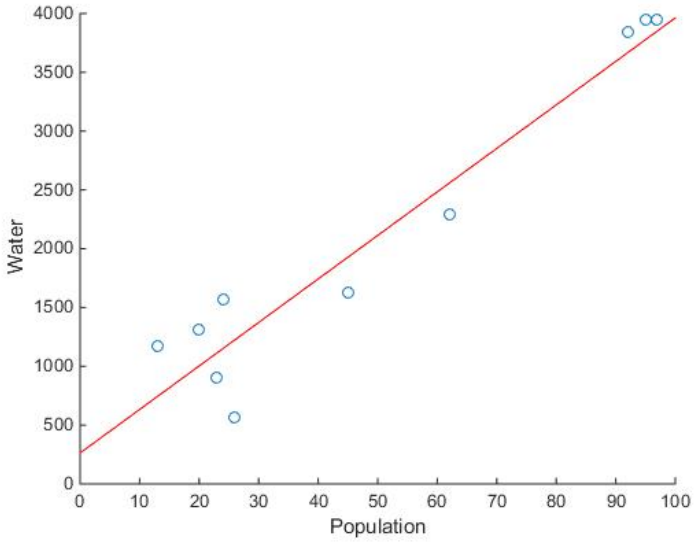
Imagine a scenario where we would like to construct a linear model to predict total water consumption c of a city in Southern California based on the population p of the city, and assume the data is given by

$$p = \{13, 92, 23, 62, 95, 26, 45, 97, 24, 20\}$$
$$c = \{1173, 3848, 902, 2290, 3946, 567, 1623, 3947, 1573, 1312\}.$$

The usual approach gives an estimated linear model of

$$\hat{c} = 37 \cdot p + 263.$$

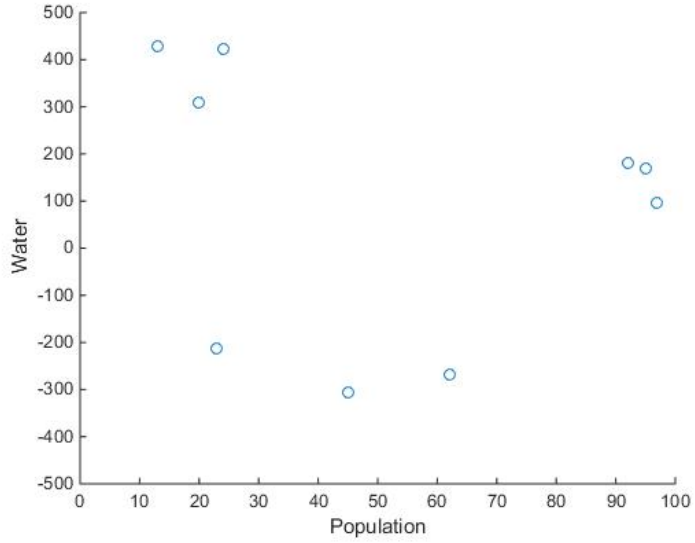
Plotting the estimated model on the scatter plot, we have



And the residuals are

$$c_i - \hat{c}_i = \{429, 181, -212, -267, 168, -658, -305, 9, 422, 309\}.$$

Plotting these gives



2 Heteroscedasticity

One striking feature of the residual plot (and the comparison of the estimated linear model to the scatter plot) in the *water consumption* example is that the measurement noise (i.e., noise in y) is larger for smaller values of x. When we are interested in estimation (as opposed to prediction) using linear models, this can be a problem because it is in conflict with the assumptions we have

made regarding the linear model. Recall that the statistical model is

$$y = x'\beta + \epsilon,$$

where ϵ is independent of x.

However, this model is not true when the measurement noise depends on the predictors. One situation of this behavior is known as *heteroscedasticity*, and it is a model where

$$y = x'\beta + \epsilon,$$

but $\mathbb{E}[\epsilon|x] = 0$. Now suppose $\hat{\beta} = (X'X)^{-1}X'Y$ are the OLS parameters, then its expectation is

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}((X'X)^{-1}X'Y) = \mathbb{E}((X'X)^{-1}X'(X\beta + e)) \\ &= \mathbb{E}\Big(\mathbb{E}[(X'X)^{-1}X'(X\beta + e)|X]\Big) = \mathbb{E}((X'X)^{-1}X'X\beta) = \beta.\end{aligned}$$

Thus, the OLS estimator is still estimating the correct value. However, it does not give an estimate with the minimum possible error in the value of $\hat{\beta}$. There are many approaches to improve the accuracy of estimates inn linear modesl with heteroscedasticity, and we will discuss two such approaches.

2.1 Weighted Least Squares

One approach assumes that we know the variance of ϵ conditioned on x, that is we know $\mathbb{E}[\epsilon^2|x]$. Strictly speaking, we only need to know a number that is proportional to this conditional variance. Suppose we define the function

$$w(x) = \frac{1}{\mathbb{E}[\epsilon^2|x]}.$$

In practice, we do not know this function value. It is common to approximate this function by inspecting the scatter plot and deciding reasonable values.

In the OLS context, we can define a matrix $W^{1/2}$ such that $W^{1/2}$ is a diagonal matrix whose i-th diagonal is equal to $\sqrt{w(x_i)}$. And then we solve the weight least squares problem

$$\hat{\beta} = \arg\min \|W^{1/2}(Y - X\beta)\|^2 = (X'WX)^{-1}X'WY.$$

In the special case of a linear model with a single predictor and a constant/intercept term (i.e., $y = mx + b$), this gives the following equations for estimating the model parameters:

$$\begin{aligned}\hat{m} &= \frac{\overline{wxy} - \overline{wx} \cdot \overline{wy}}{\overline{wx^2} - (\overline{wx})^2} \\ \hat{b} &= \frac{\overline{wy} - \hat{m} \cdot \overline{wx}}{\overline{w}}.\end{aligned}$$

2.2 Semiparametric Approach

Another approach is a two-step procedure. Let k be a constant we fix before starting the procedure. In the first step, we estimate $\mathbb{E}[y|x]$ by computing

$$\tilde{y}_i = \frac{1}{k} \sum_{j \text{ such that } x_j \text{ is one of the } k \text{ nearest points to } x_i} y_j.$$

In the second step, we estimate the model parameters by using the regular least squares equations with the data $(x_i, \tilde{y}_i)$ for $i = 1, \ldots, n$.

2.3 Example: Water Consumption

2.3.1 Weighted Least Squares Approach

Based on the scatter plot and the residuals plot, it roughly looks like the variance is double for points with $x_i < 70$ as compared to the variance for points with $x_i \geq 70$. As a result, we use the weighting function

$$w(x) = \begin{cases} 1/2, & \text{if } x < 70 \\ 1, & \text{if } x \geq 70. \end{cases}$$

Recall that the data is given by

$$p = \{13, 92, 23, 62, 95, 26, 45, 97, 24, 20\}$$
$$c = \{1173, 3848, 902, 2290, 3946, 567, 1623, 3947, 1573, 1312\}.$$

Then the weighted least squares estimates are determined by first computing

$$\overline{wx} = \tfrac{1}{10} \cdot (\tfrac{1}{2} \cdot 13 + 92 + \tfrac{1}{2} \cdot 23 + \tfrac{1}{2} \cdot 62 + \tfrac{1}{2} \cdot 95 + \tfrac{1}{2} \cdot 26 + \tfrac{1}{2} \cdot 45 + 97 + \tfrac{1}{2} \cdot 24 + \tfrac{1}{2} \cdot 20) = 39.05$$
$$\overline{wy} = \tfrac{1}{10} \cdot (\tfrac{1}{2} \cdot 1173 + 3848 + \tfrac{1}{2} \cdot 902 + \tfrac{1}{2} \cdot 2290 + \tfrac{1}{2} \cdot 3946 + \tfrac{1}{2} \cdot 567 + \tfrac{1}{2} \cdot 1623 + 3947 + \tfrac{1}{2} \cdot 1573 + \tfrac{1}{2} \cdot 1312) = 1646$$
$$\overline{wxy} = \tfrac{1}{10} \cdot (\tfrac{1}{2} \cdot 13 \cdot 1173 + 92 \cdot 3848 + \tfrac{1}{2} \cdot 23 \cdot 902 + \tfrac{1}{2} \cdot 62 \cdot 2290 + \tfrac{1}{2} \cdot 95 \cdot 3946 + \tfrac{1}{2} \cdot 26 \cdot 567 + \tfrac{1}{2} \cdot 45 \cdot 1623 + 97 \cdot 3947 + \tfrac{1}{2} \cdot 24 \cdot 1573 + \tfrac{1}{2} \cdot 20 \cdot 1312) = 127662$$
$$\overline{wx^2} = \tfrac{1}{10} \cdot (\tfrac{1}{2} \cdot 13^2 + 92^2 + \tfrac{1}{2} \cdot 23^2 + \tfrac{1}{2} \cdot 62^2 + \tfrac{1}{2} \cdot 95^2 + \tfrac{1}{2} \cdot 26^2 + \tfrac{1}{2} \cdot 45 + 97 + \tfrac{1}{2} \cdot 24 + \tfrac{1}{2} \cdot 20) = 3101$$
$$\overline{w} = \tfrac{1}{10} \cdot (\tfrac{1}{2} + 1 + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + 1 + \tfrac{1}{2} + \tfrac{1}{2}) = 0.65$$

Thus, our estimates are

$$\hat{m} = \frac{\overline{wxy} - \overline{wx} \cdot \overline{wy}}{\overline{wx^2} - (\overline{wx})^2} = 40$$
$$\hat{b} = \frac{\overline{wy} - \hat{m} \cdot \overline{wx}}{\overline{w}} = 116.$$

2.3.2 Semiparametric Approach

Suppose we choose $k = 3$, and recall that the data is given by

$$p = \{13, 92, 23, 62, 95, 26, 45, 97, 24, 20\}$$
$$c = \{1173, 3848, 902, 2290, 3946, 567, 1623, 3947, 1573, 1312\}.$$

Then, in the first step we compute

$$\tilde{y}_1 = \frac{1}{3} \cdot (1173 + 902 + 1312) = 1129 \qquad \tilde{y}_6 = \frac{1}{3} \cdot (567 + 902 + 1573) = 1014$$
$$\tilde{y}_2 = \frac{1}{3} \cdot (3848 + 3946 + 3947) = 3914 \qquad \tilde{y}_7 = \frac{1}{3} \cdot (1623 + 2290 + 567) = 1493$$
$$\tilde{y}_3 = \frac{1}{3} \cdot (902 + 567 + 1573) = 1014 \qquad \tilde{y}_8 = \frac{1}{3} \cdot (3947 + 3848 + 3946) = 3914$$
$$\tilde{y}_4 = \frac{1}{3} \cdot (2290 + 1623 + 3848) = 2587 \qquad \tilde{y}_9 = \frac{1}{3} \cdot (1573 + 902 + 567) = 1014$$
$$\tilde{y}_5 = \frac{1}{3} \cdot (3946 + 3848 + 3947) = 3914 \qquad \tilde{y}_{10} = \frac{1}{3} \cdot (1312 + 902 + 1573) = 1262$$

In the second step, we estimate the model parameters by using the regular least squares equations with the data $(x_i, \tilde{y}_i)$. This gives

$$\hat{m} = 38\hat{b} = 232.$$

2.4 Comparing the Estimates

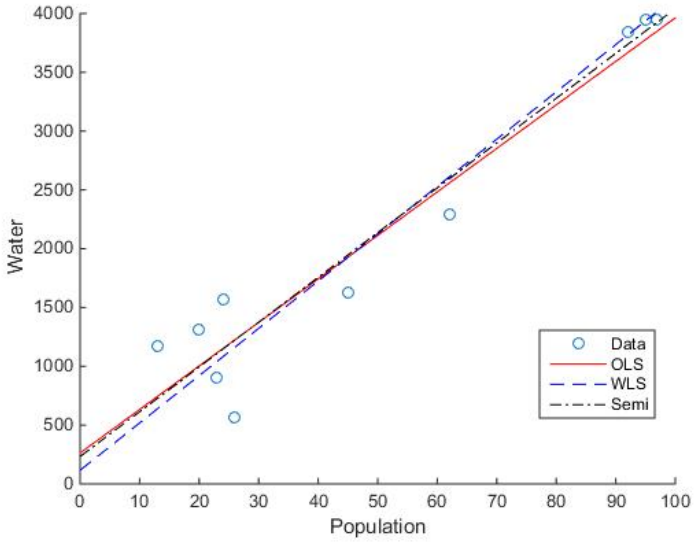
It is useful to compare the estimated parameters using the three methods, as well as plot the corresponding curves. The model that was used to generate the data (in a sense this represents the "true" model) is given by

$$c = 40 \cdot p + 100.$$

The models estimated by ordinary least squares, weighted least squares, and the semiparametric approach are

$$\hat{c} = 37 \cdot p + 263$$
$$\hat{c} = 40 \cdot p + 116$$
$$\hat{c} = 38 \cdot p + 232.$$

respectively. The weighted least squares estimates are closest to the true model, and the semiparametric estimates are better than those of ordinary least squares. However, the semiparametric estimates are not as good as those of weighted least squares. In general, the semiparametric approach works well when we have a lot of data. In our example, we only have $n = 10$ data points, and so the semiparametric approach would not be expected to work very well. Plotting all three estimated models on the scatter plot, we have



3 Estimating Nonlinear Models with Linear Regression

One reason for the popularity of linear regression is that it can often be used to estimate nonlinear models. There are a few different examples of this.

3.1 Nonparametric Regression

Suppose we would like to estimate the parameters of a polynomial model:

$$y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \cdots + a_k \cdot x^k.$$

It turns out that given measurements (x_i, y_i) for $i = 1, \ldots, n$, we can estimate the parameters $a_0, \ldots, a_k$ by solving an OLS using the following variables as predictors:

$$1, x, x^2, \ldots, x^k.$$

As another example, suppose we would like to estimate a model that is a partial sum of a Fourier series:

$$y = c + \sum_{m=1}^{k} \Big(a_m \cdot \sin(2\pi m x) + b_m \cdot \cos(2\pi m x) \Big).$$

We can estimate the parameters c, a_m, b_m by solving an OLS using the following variables as predictors:

$$1, \sin(2\pi m x), \cos(2\pi m x), \text{for all } m = 1, \ldots, k.$$

3.2 Nonlinear Transformation

In other cases, we can perform a nonlinear transformation on the data to convert the estimation problem into linear regression. For example, suppose the model we would like to estimate is

$$y = \exp(mx + b).$$

If we take the logarithm of y, then we have

$$\log y = mx + b.$$

Hence, given measurements (x_i, y_i) for $i = 1, \ldots, n$, we can estimate the parameters m, b by using our least squares formula but with the x_i as the predictor measurements and $\log y_i$ as the response measurements.
As another example, suppose the model we would like to estimate is

$$y = (m \log x + b)^2.$$

Then if we take the square root of y, then we have

$$\sqrt{y} = m \log x + b.$$

Thus, we can estimate the parameters m, b by using our least squares formula but with $\log x_i$ as the predictor measurements and $\sqrt{y_i}$ as the response measurements.