YData: Introduction to Data Science

Lecture 32: residuals

Overview

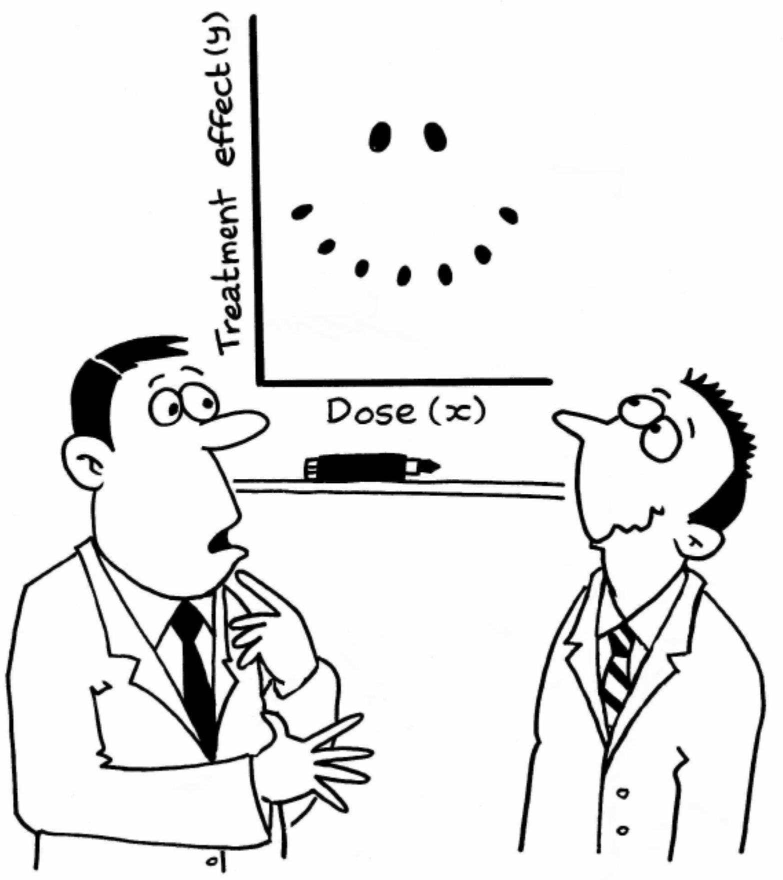
Linear regression continued…

Review and continuation of examining the RMSE

Minimizing the RMSE

If there is time

- Residuals
- Polynomial regression



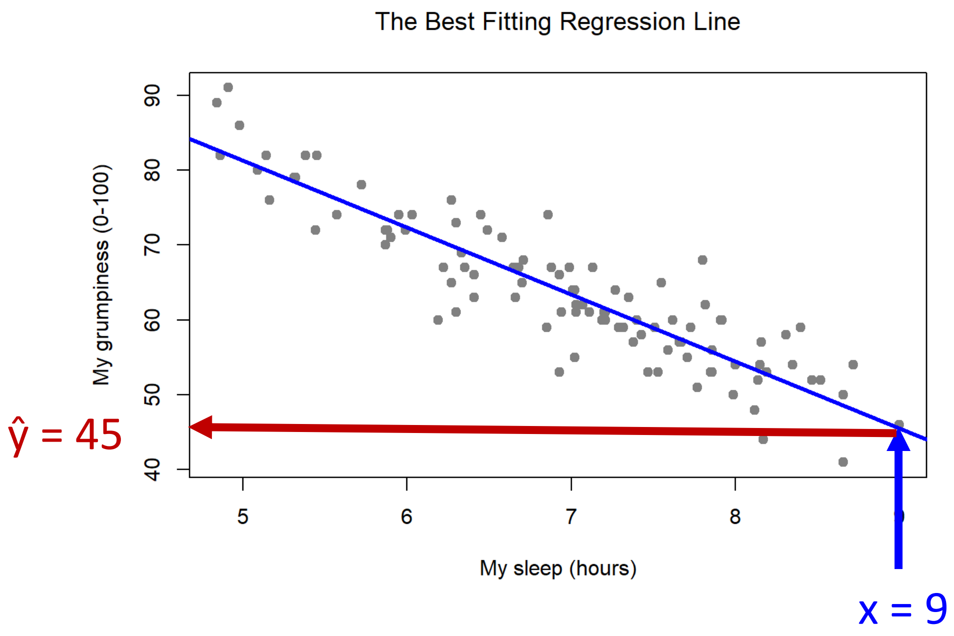
"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Review of Linear regression

Regression

Regression is method of using one variable ***x*** *<u>to predict</u>* the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$



In **linear regression** we fit a <u>line</u> to the data, called the **regression line**

Lines can be expressed by a slope and intercept:

$$\hat{y} = \text{slope} \cdot x + \text{intercept}$$

Regression line

Our equation for the regression line in standardized units is:

Regression Line

$$y_{(\mathrm{su})} = \underbrace{r}_{\text{Correlation}} \times x_{(\mathrm{su})}$$

Expanding the definition of standardized units we have:

$$(\hat{y} - \bar{y}) / SD_y = r \cdot (x - \bar{x}) / SD_x$$

Solving in our original units: $\hat{y} = \text{slope} \cdot x + \text{intercept}$

$$\text{Slope} = r \cdot SD_y / SD_x$$

$$\text{Intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

Residuals

Residuals

Technical definitions:

error = actual value - population line value

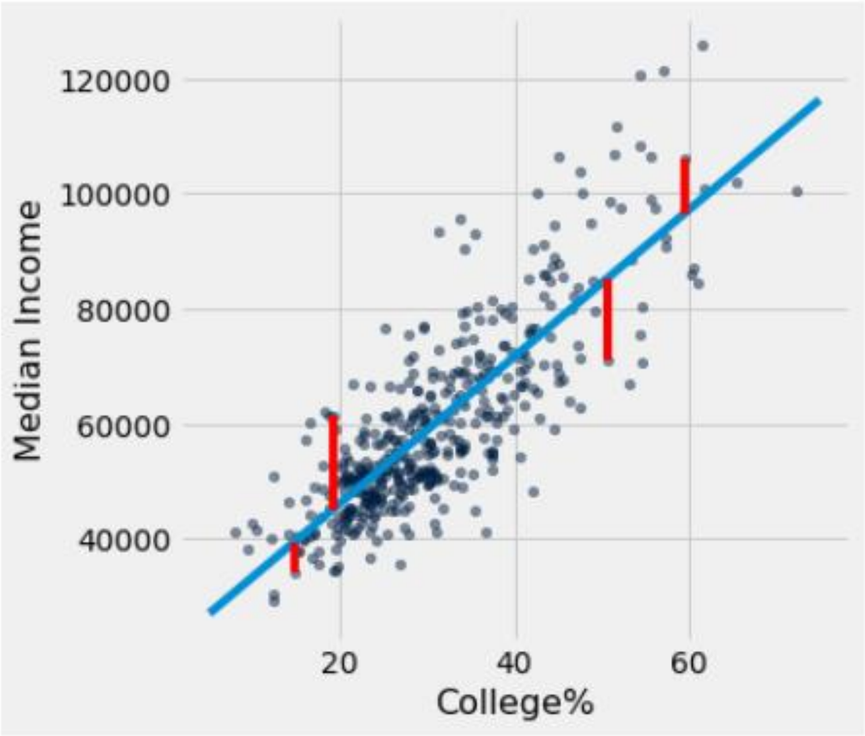
$\varepsilon_i = y_i - \mu(x_i)$

residual = actual value - sample line estimate

$e_i = y_i - \hat{y}_i$

We will not make much of a distinction between residuals and errors in this class

- Both capture the distance from an observed value y and the predictions from a regression line.



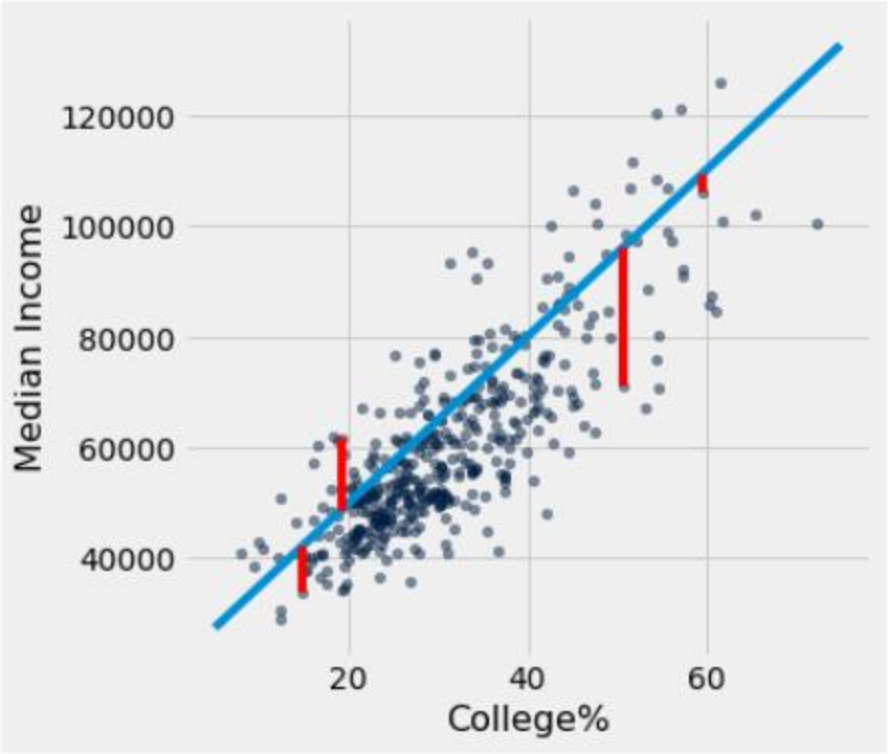
Let's explore this in Jupyter!

Least squares

Least squares estimation

residual = actual value - estimate

$e_i = y_i - \hat{y}_i$



Typically, some errors are positive and some negative

To measure the rough size of the errors we calculate the root mean square error (RMSE):

- Square the errors to eliminate cancellation
- Take the mean of the squared errors
- Take the square root to fix the units

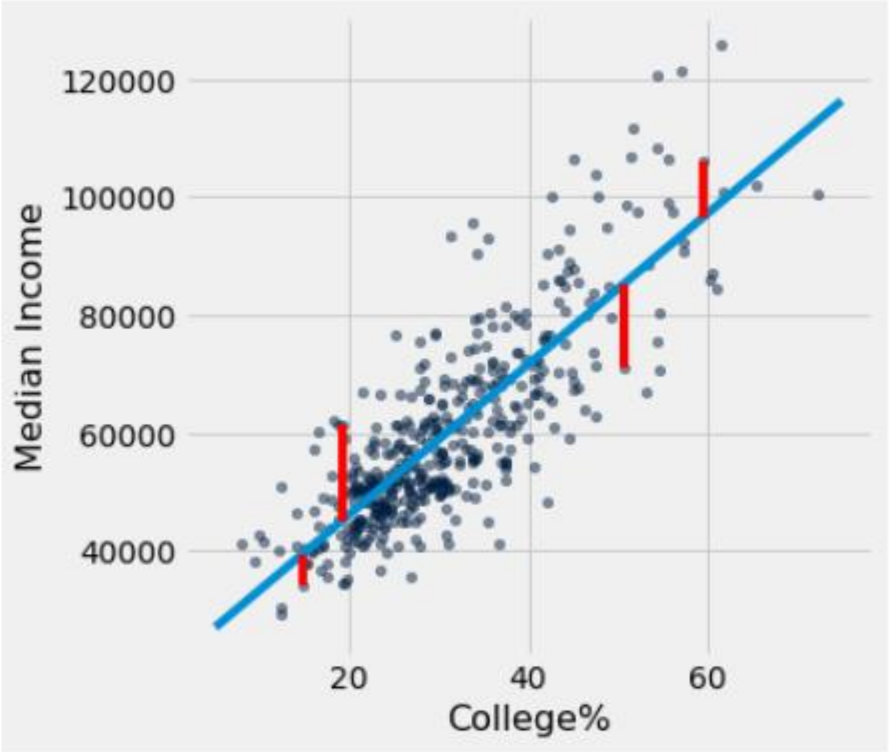
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Let's explore this in Jupyter!

Least Squares Line

Minimizes the root mean squared error (RMSE) among all lines

- Equivalently, minimizes the mean squared error (MSE) among all lines



Names:

- “Best fit" line
- Least squares line
- Regression line

Numerical optimization

Numerical minimization is approximate but effective

Much of machine learning is based on numerical minimization

If the function mse(a, b) returns the MSE of estimation using the line “estimate = ax + b"

- then minimize(mse) returns array [a_0, b_0]
- a_0 is the slope and b_0 the intercept of the line that minimizes the MSE among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

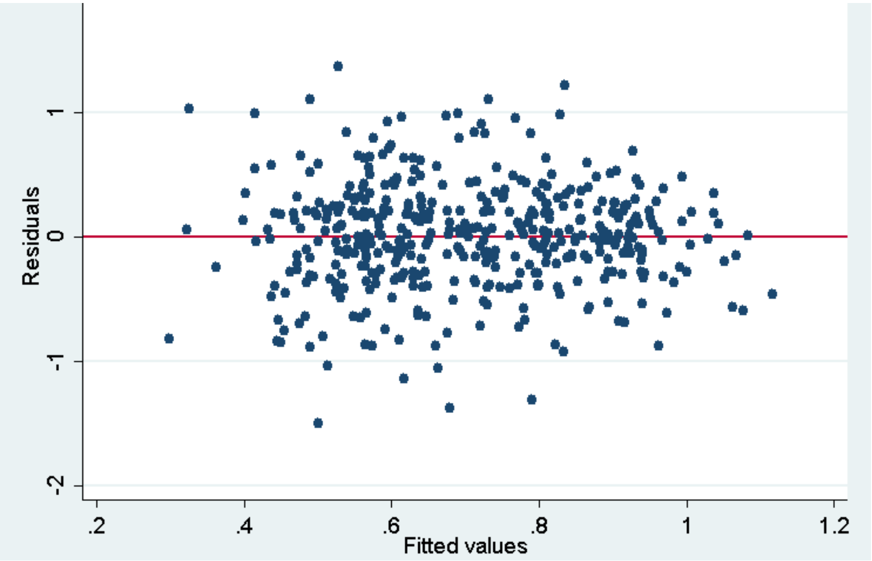
Let's explore this in Jupyter!

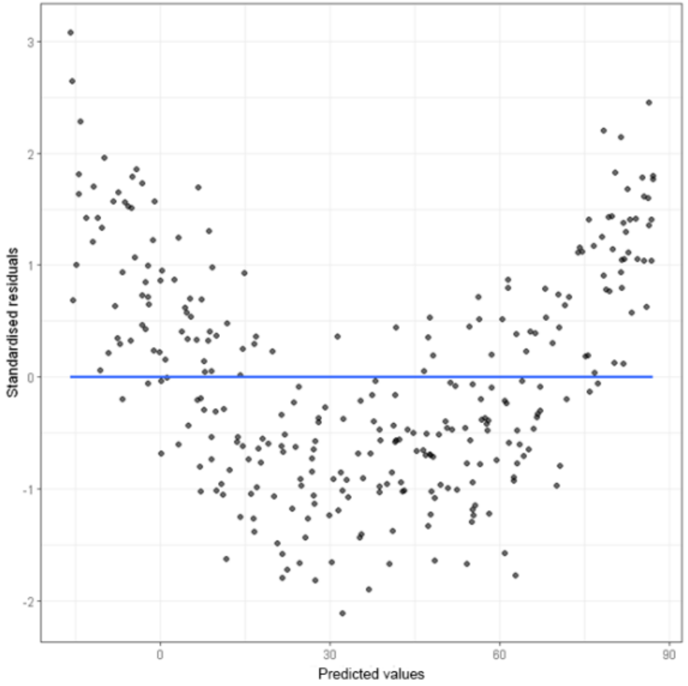
Regression diagnostics

Residual plot

A scatter diagram of residuals

- Should look like an unassociated blob for linear relations
- But will show patterns for non-linear relations
- Used to check whether linear regression is appropriate





Let's explore this in Jupyter!

Polynomial regression

Polynomial regression

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of predictors

$$
\begin{aligned}
\hat{y} = b_0 &+ b_1 \cdot x \\
&+ b_2 \cdot (x)^2 + \\
&+ b_3 \cdot (x)^3 + \varepsilon
\end{aligned}
$$

Still a linear equation but non-linear in original predictors

Polynomial regression

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of predictors

$$\text{child} = b_0 + b_1 \cdot \text{MidParent} + b_2 \cdot (\text{MidParent})^2 + b_3 \cdot (\text{MidParent})^3 + \varepsilon$$

Still a linear equation but non-linear in original predictors

Let's explore this in Jupyter!