

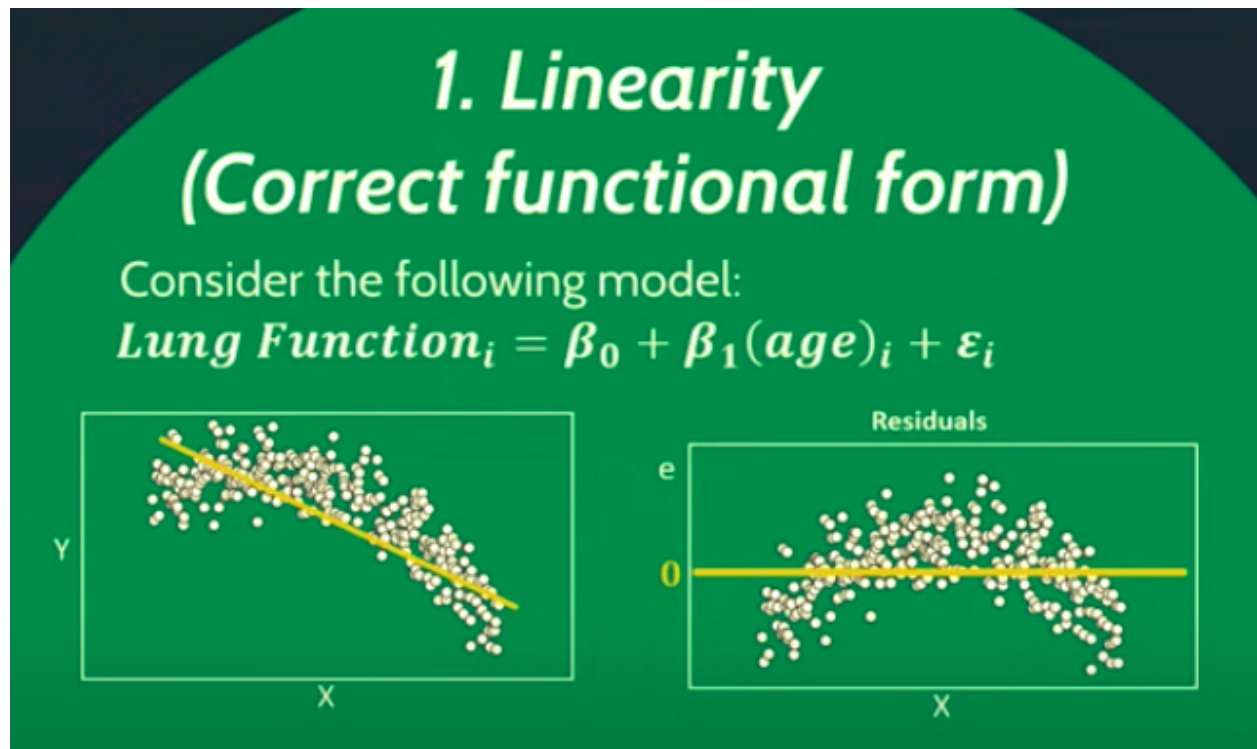
## Regression Assumptions Explained

### LINEAR REGRESSION NECESSITIES

- The most common function used in linear regression is “Least Squared Errors” function; which is the sum of squared errors ( $\sum(y_{\text{actuals}} - y_{\text{predicted}})^2$ ) over the training set, trying to maximize how far off the predictions are from the actuals.
- To calculate least squared errors, you must calculate the residuals.
- In order to find the least squared error you will have to find the optimal parameter values that minimize the sum of squared residuals. (best fit line)
- The optimization technique used in linear regression is Gradient Descent, which attempts to find a local or a global minimum of a cost function.
- Gradient descent finds the direction ‘gradient’ that the model line should take so that the errors will be reduced.
- The R squared is a measure of how close the data is to the fitted regression line. We want to maximize that value.
- The RMSE (root mean squared error) is the standard deviation on the residuals. Hence RMSE is a measure of how spread out your residuals are.
- The MAE (mean absolute error) is the average of all the absolute errors. The difference between the true value ( $y_{\text{train}}$ ) and the predicted value ( $y_{\text{pred}}$ ).
- Coefficients are the weights of the independent variables.
- A negative coefficient suggests that as the independent variables increase, the dependent variable tends to decrease.
- In regression with multiple independent variables, the coefficient tells you how much the dependent variable is expected to increase when that independent variable increases by one, holding all the other independent variables constant.

- Linear Regression is limited on hyperparameters; we must remove outliers, add new features.

## LINEARITY CORRECT FUNCTIONAL FORM



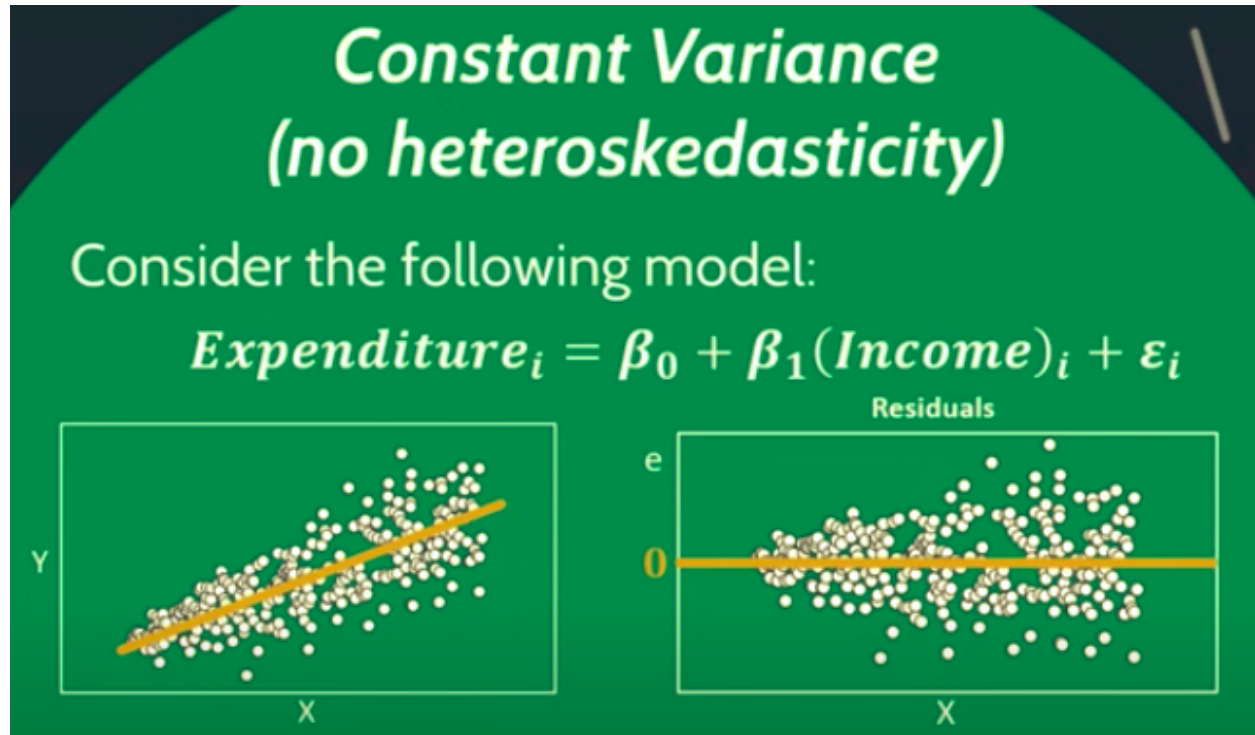
- Y is lung function, and X is age. If we were to try to run a linear regression aka find a gradient for the variable age as it relates to lung function we would essentially be drawing a straight line through the data.
- One way to assess how wrong we have our fitted line is to look at the residuals.

$$\text{Lung Function}_i = \beta_0 + \beta_1(\text{age})_i + \beta_2(\text{age}^2)_i + \varepsilon_i$$

- Our lung function is linear as it relates to age and age squared, but by including the age squared variable we have a quadratic function comparable to  $y = ax^2 + bx + c$

- If the functional form is incorrect, the gradient for our independent variable, or our best fit line. Then the coefficients and standard errors in the output are unreliable and useless.
- In order to detect bad functional forms we go to residual plots, and likelihood ratio tests.

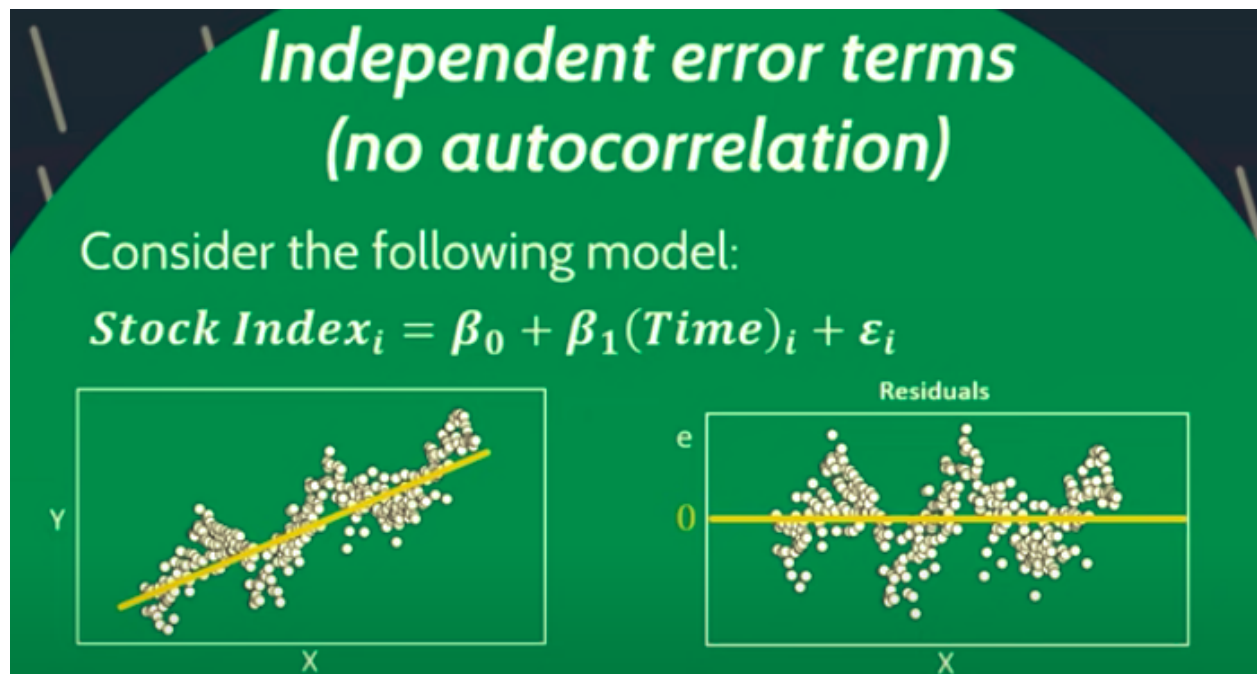
### CONSTANT ERROR VARIANCE (HOMOSKEDASTICITY)



- We consider expenditures in relation to income. As logic shows the variance for lower income is much smaller in regards to expenditures. The lower income group is limited to their paychecks, whereas the higher income group has a choice to spend or limit their expenditures.
- This poses a problem as the residuals grow larger in respect to expenditures. This problem is heteroskedasticity.
- Running a regression with heteroskedasticity present causes the standard errors to be unreliable.

- In order to detect this issue the goldfeldt quant test, and the breusch pagan test are options.
- The remedies are white's standard errors, weighted least squares, and logarithms transforming the variables.

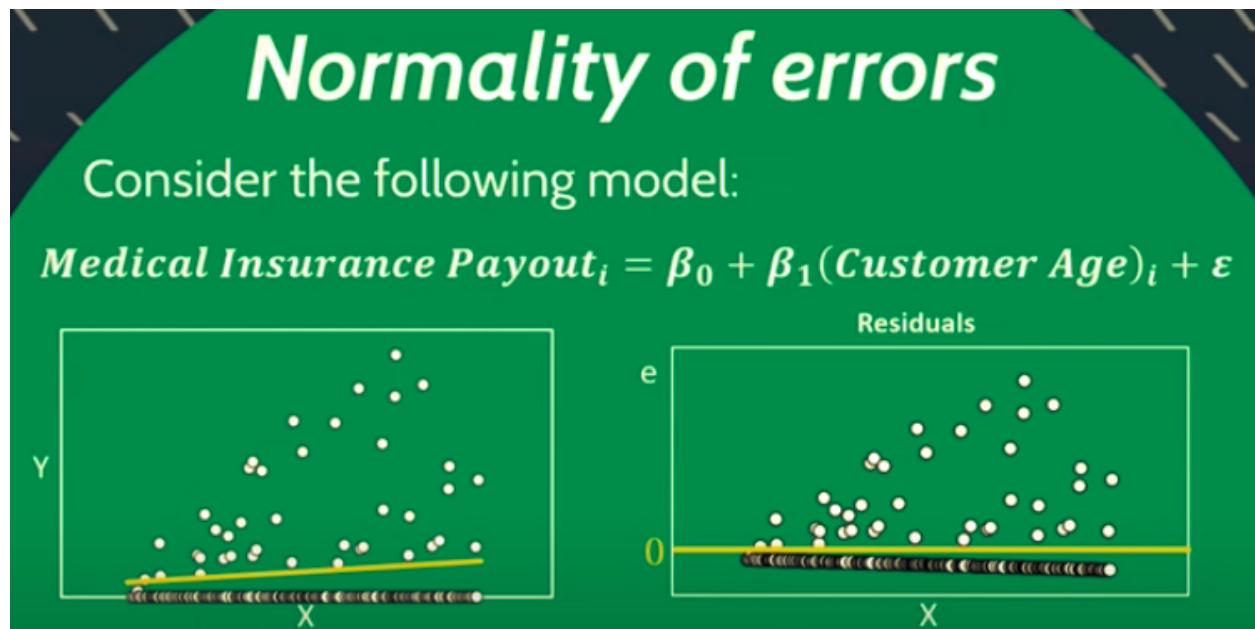
### INDEPENDENT ERROR TERMS (NO AUTOCORRELATION)



- Occurs when there is a natural order to X variable. This is present in time series analysis. The X variable is time, and we are mapping out a long term stock trend.
- Each of the residuals is affected by the one before it. If you know the point is positive, more than likely the next will be as well.
- Tests to detect are the durbin-watson test, and the breusch- godfrey test.
- Remedies are to investigate omitted variables, for example the natural business cycles of earnings and economic booms and busts. We can add an X variable in to account for these cycles.

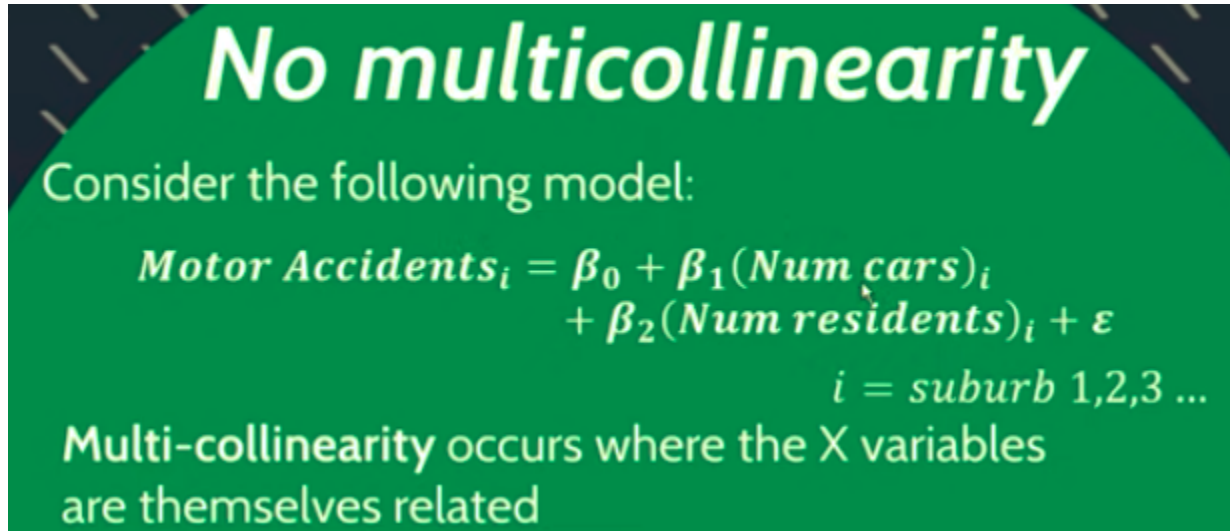
- Another option is a generalized difference equation. (cochrane-orchutt or AR(1) methods). Inputting a weighted value from your previous stock index, and inherently accounting for a difference and modeling it in the regression. Error term is no longer correlated.

## NORMALITY OF ERROR TERMS



- Most people who take out insurance will not take out a big insurance item, but some people may have to as displayed on the chart.
- We have heteroskedasticity going on, but also an issue with the normality of errors.
- This means that there should be a normal distribution , bell curve like where majority of residuals hug tight to the line, and outliers and few and far in between.
- If normality is violated, and the population is small, standard errors in output are affected. Because we have so many observations, we can offset the normality, but if there is a small # of observations than standard error in the output is affected.
- A way of assessing normality is through a QQ plot. Seeing if it has a bell curve.

## NO MULTICOLLINEARITY



**No multicollinearity**

Consider the following model:

$$\text{Motor Accidents}_i = \beta_0 + \beta_1(\text{Num cars})_i + \beta_2(\text{Num residents})_i + \varepsilon$$

$i = \text{suburb } 1, 2, 3 \dots$

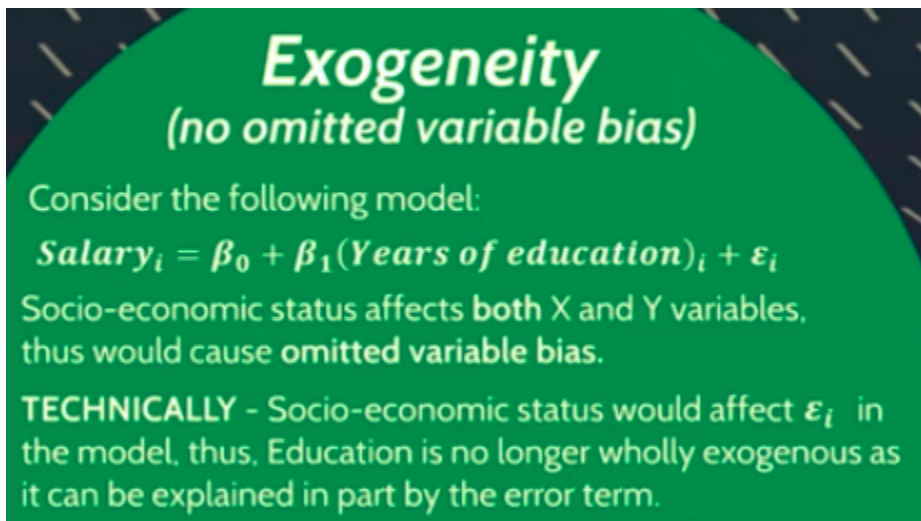
**Multi-collinearity** occurs where the X variables are themselves related

- This is a multiple linear regression with two X variables.
- The regression algorithm is attempting to isolate the individual effects that number of cars, and number of residents have on motor accidents.
- The way we would interpret the output values of  $\beta_1$ (num of cars) is the following: the expected effect on the number of motor accidents if the number of cars in the suburb increased by 1. WHEN WE SAY THIS WE ALWAYS SAY “HOLDING OTHER VARIABLES CONSTANT.”
- We can not do that here because we can not hold residents constant while we keep increasing the number of cars, logically it does not make sense that way.
- Our X variables will vary so closely together that our algorithm will have trouble identifying which made the difference in the number of motor accidents.
- The effects are unreliable standard errors, and unreliable coefficients. The model starts shutting down when multicollinearity is too high. Perfect multicollinearity includes two X

variables that are perfectly collinear, the two are exactly proportional then the model can not run at all.

- To detect this you can look at correlation between X variables, and look at variance inflation factors. VIF looks at each X variable and assesses the difference when we take it out versus put it into the regression. How is the variance of the regression affected? The higher the VIF number the more likely that variable information is already contained within the model hidden in other variables.
- The remedy is remove one of the variables.

## EXOGENEITY



**Exogeneity**  
(no omitted variable bias)

Consider the following model:

$$\text{Salary}_i = \beta_0 + \beta_1(\text{Years of education})_i + \varepsilon_i$$

Socio-economic status affects both X and Y variables, thus would cause omitted variable bias.

**TECHNICALLY** - Socio-economic status would affect  $\varepsilon_i$  in the model, thus, Education is no longer wholly exogenous as it can be explained in part by the error term.

- Mapping out someone's salary, in regards to the number of years of education.
- There's another variables that are actually fueling the relationship between salary, and years of education . For example, socio-economic status allows you to have more education, inherently meaning greater salary. The status affects both X and Y variables.
- The model can still be used for predictive purposes, but we can not infer causation between a high salary, and a high education.
- Detection - intuition, and checking correlations.

- Remedy - using instrumental variables.

## EXAMPLE OF COEFFICIENTS

```
[53] print('Coefficients:', model.coef_)
      print('Intercept:', model.intercept_)
      print('Mean squared error (MSE): %.2f'
            % mean_squared_error(Y_test, Y_pred))
      print('Coefficient of determination (R^2): %.2f'
            % r2_score(Y_test, Y_pred))
```

```
Coefficients: [ -3.38885045 -263.61109602  565.48999639  381.03358396 -697.5705492
                374.9251023   47.59537317  184.18561222  683.31463398  43.16955289]
```

- The coefficients give us a weighting for the impact of specified independent variables for our linear regression model.

```
[54] print(diabetes.feature_names)
```

```
['age', 'sex', 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6']
```

$Y = -3.38(\text{age}) - 263.61109602(\text{sex}) + 565.489(\text{bmi}) + \dots + 153.556$