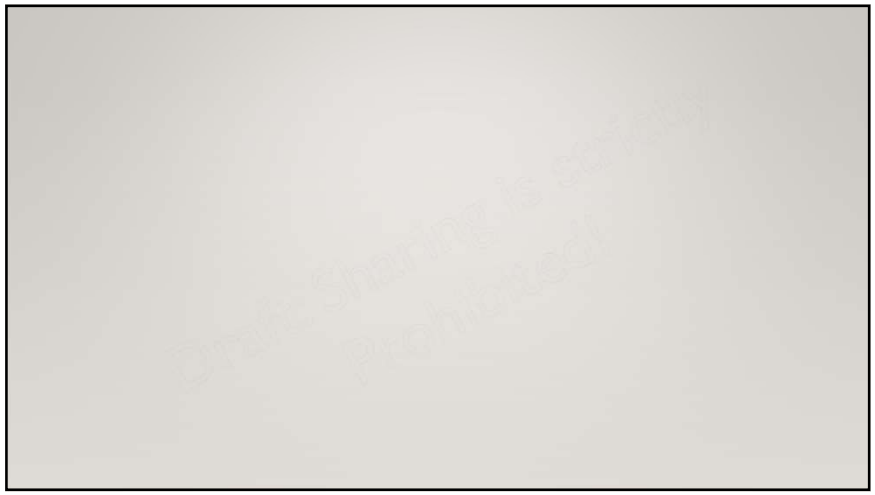
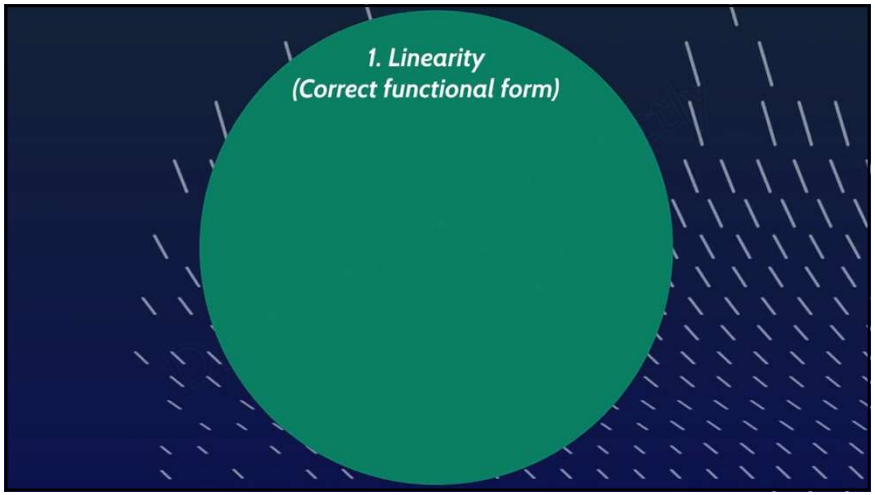
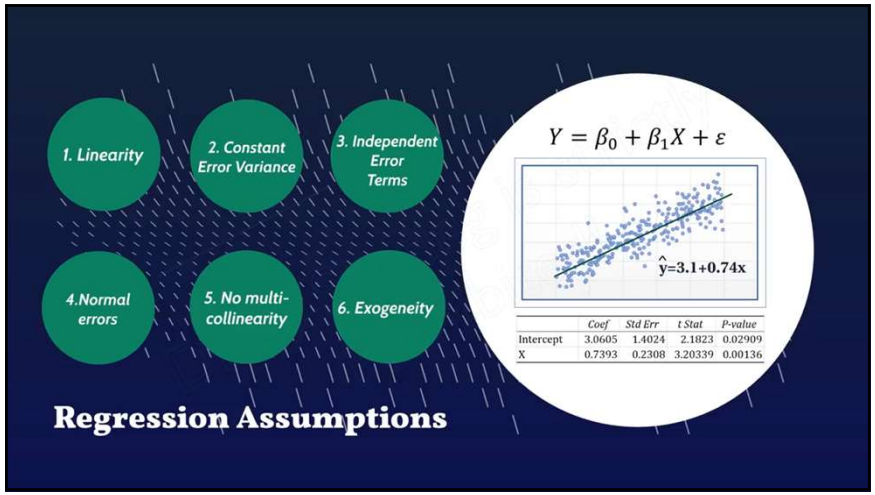
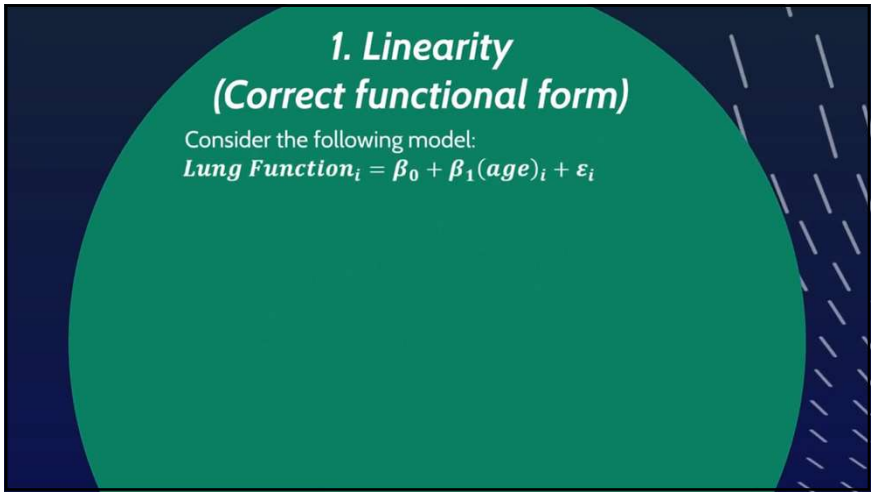



1. Linearity
(Correct functional form)

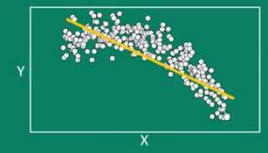

1. Linearity
2. Constant Error Variance
3. Independent Error Terms
4.Normal errors
5. No multi-collinearity
6. Exogeneity
$Y = \beta_0 + \beta_1 X + \varepsilon$
$\hat{y}=3.1+0.74x$
Coef Std Err t Stat P-value
Intercept 3.0605 1.4024 2.1823 0.02909
X 0.7393 0.2308 3.20339 0.00136
Regression Assumptions


1. Linearity
(Correct functional form)
Consider the following model:
$Lung\ Function_i = \beta_0 + \beta_1(age)_i + \varepsilon_i$

1. Linearity (Correct functional form)

Consider the following model:

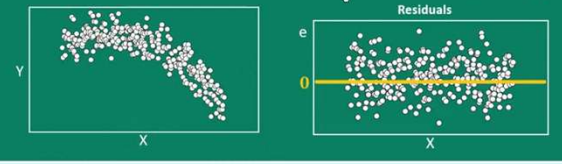
$$\textbf{Lung Function}_i = \beta_0 + \beta_1(age)_i + \varepsilon_i$$



1. Linearity (Correct functional form)

Consider the following model:

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



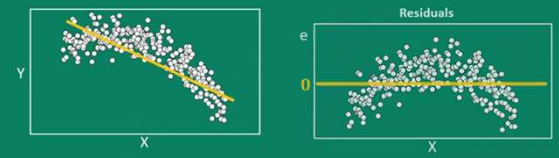
What's the issue?

- If functional form is incorrect, both the coefficients and standard errors in your output are unreliable

1. Linearity (Correct functional form)

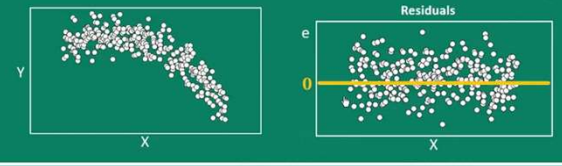
Consider the following model:

$$\textbf{Lung Function}_i = \beta_0 + \beta_1(age)_i + \varepsilon_i$$



Consider the following model:

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



What's the issue?

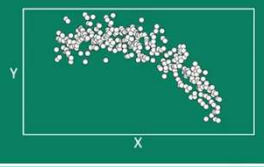
- If functional form is incorrect, both the coefficients and standard errors in your output are unreliable

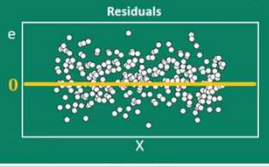
Detection:

- Residual plots
- Likelihood ratio (LR) test

Consider the following model:

$$Lung\ Function_i = \beta_0 + \beta_1(age)_i + \beta_2(age^2)_i + \varepsilon_i$$





What's the issue?

- If functional form is incorrect, both the coefficients and standard errors in your output are unreliable

Detection:

- Residual plots
- Likelihood ratio (LR) test

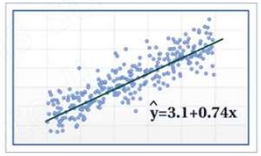
Remedies:

- Get the specification correct (trial and error)

Constant Variance
(no heteroskedasticity)

1. Linearity
2. Constant Error Variance
3. Independent Error Terms
4. Normal errors
5. No multi-collinearity
6. Exogeneity

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



	Coef	Std Err	t Stat	P-value
Intercept	3.0605	1.4024	2.1823	0.02909
X	0.7393	0.2308	3.20339	0.00136

Regression Assumptions

Constant Variance
(no heteroskedasticity)

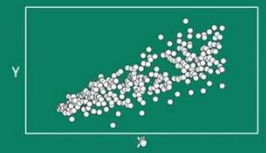
Consider the following model:

$$Expenditure_i = \beta_0 + \beta_1(Income)_i + \varepsilon_i$$

Constant Variance (no heteroskedasticity)

Consider the following model:

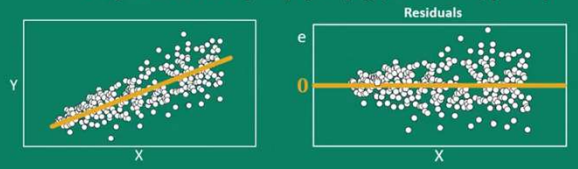
$$Expenditure_i = \beta_0 + \beta_1(Income)_i + \varepsilon_i$$



Constant Variance (no heteroskedasticity)

Consider the following model:

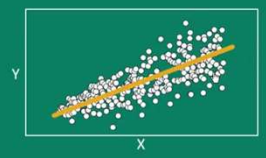
$$Expenditure_i = \beta_0 + \beta_1(Income)_i + \varepsilon_i$$



Constant Variance (no heteroskedasticity)

Consider the following model:

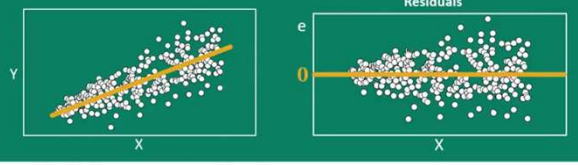
$$Expenditure_i = \beta_0 + \beta_1(Income)_i + \varepsilon_i$$



Constant Variance (no heteroskedasticity)

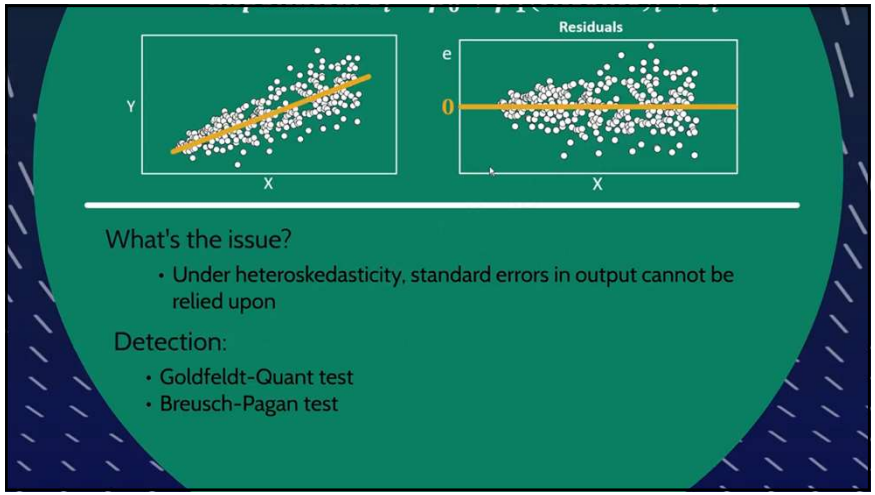
Consider the following model:

$$Expenditure_i = \beta_0 + \beta_1(Income)_i + \varepsilon_i$$



What's the issue?

- Under heteroskedasticity, standard errors in output cannot be relied upon

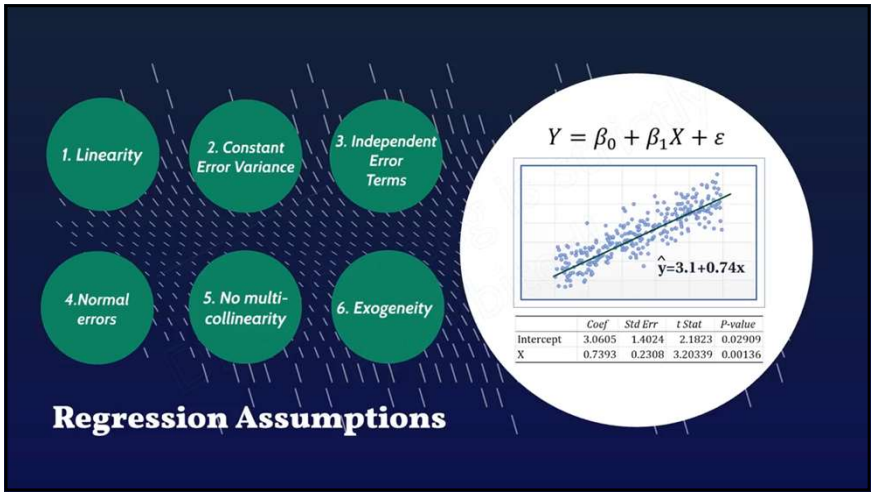


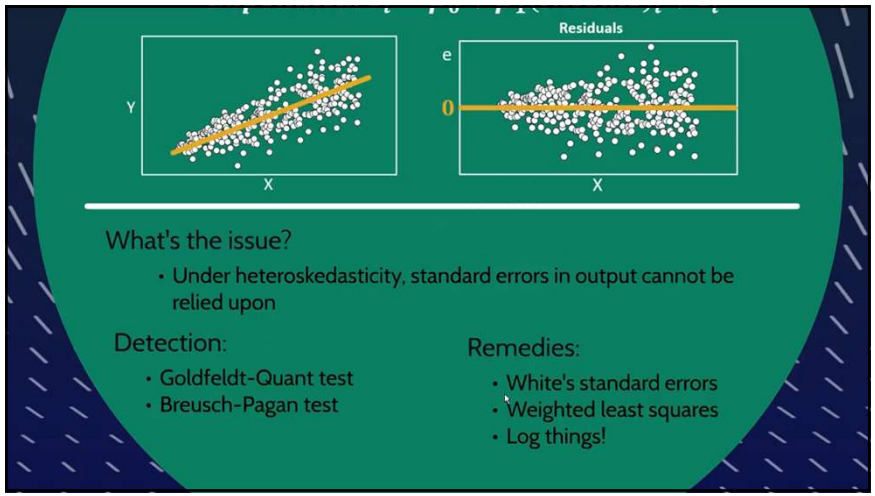
What's the issue?

- Under heteroskedasticity, standard errors in output cannot be relied upon

Detection:

- Goldfeldt-Quant test
- Breusch-Pagan test





What's the issue?

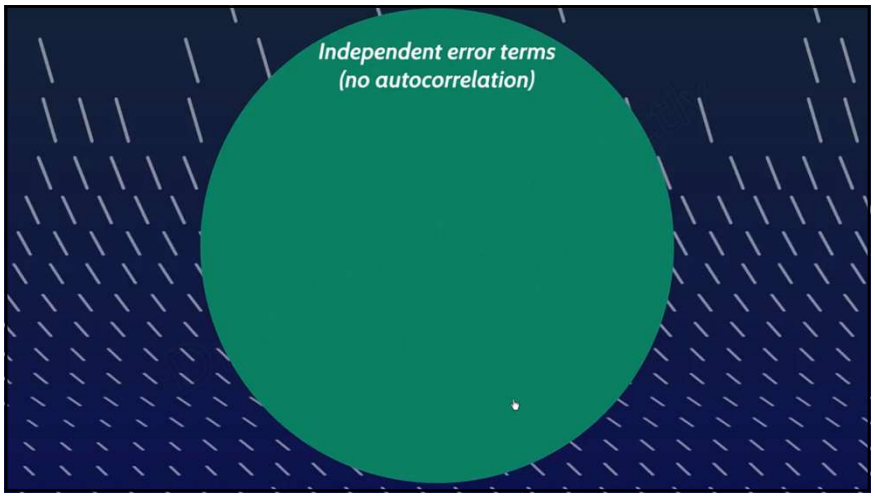
- Under heteroskedasticity, standard errors in output cannot be relied upon

Detection:

- Goldfeldt-Quant test
- Breusch-Pagan test

Remedies:

- White's standard errors
- Weighted least squares
- Log things!



Independent error terms (no autocorrelation)

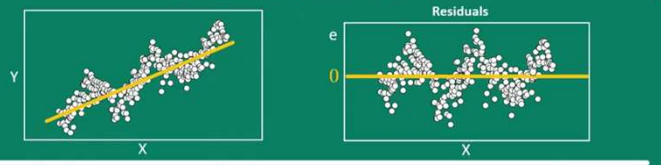
Consider the following model:

$$Stock\ Index_i = \beta_0 + \beta_1(Time)_i + \varepsilon_i$$

Independent error terms (no autocorrelation)

Consider the following model:

$$Stock\ Index_i = \beta_0 + \beta_1(Time)_i + \varepsilon_i$$



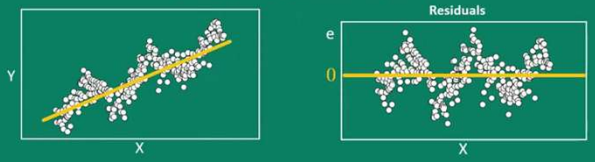
What's the issue?

- Under autocorrelation, standard errors in output cannot be relied upon

Independent error terms (no autocorrelation)

Consider the following model:

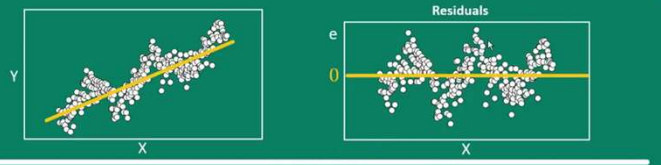
$$Stock\ Index_i = \beta_0 + \beta_1(Time)_i + \varepsilon_i$$



Independent error terms (no autocorrelation)

Consider the following model:

$$Stock\ Index_i = \beta_0 + \beta_1(Time)_i + \varepsilon_i$$



What's the issue?

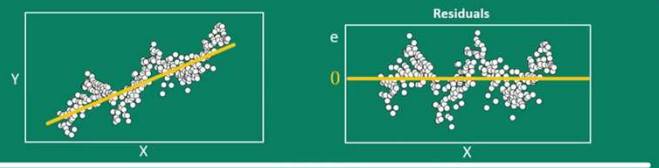
- Under autocorrelation, standard errors in output cannot be relied upon

Detection:

- Durbin-Watson test
- Breusch-Godfrey test

Consider the following model:

$$Stock\ Index_i = \beta_0 + \beta_1(Time)_i + \varepsilon_i$$



What's the issue?

- Under autocorrelation, standard errors in output cannot be relied upon

Detection:

- Durbin-Watson test
- Breusch-Godfrey test

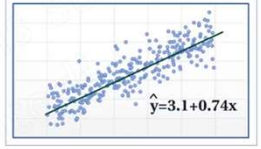
Remedies:

- Investigate omitted variables
- Generalised difference equation (Cochrane-Orchutt or AR(1) methods)

Normality of errors

1. Linearity
2. Constant Error Variance
3. Independent Error Terms
4. Normal errors
5. No multi-collinearity
6. Exogeneity

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



	Coef	Std Err	t Stat	P-value
Intercept	3.0605	1.4024	2.1823	0.02909
X	0.7393	0.2308	3.20339	0.00136

Regression Assumptions

Normality of errors

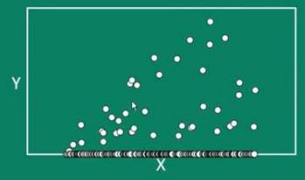
Consider the following model:

$$Medical\ Insurance\ Payout_i = \beta_0 + \beta_1(Customer\ Age)_i + \varepsilon$$

Normality of errors

Consider the following model:

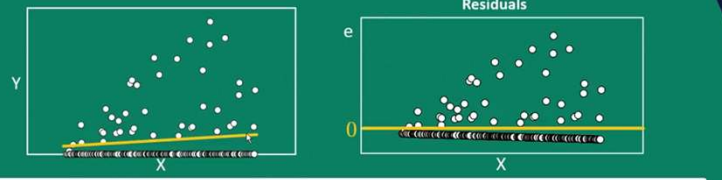
$$Medical\ Insurance\ Payout_i = \beta_0 + \beta_1(Customer\ Age)_i + \varepsilon$$



Normality of errors

Consider the following model:

$$Medical\ Insurance\ Payout_i = \beta_0 + \beta_1(Customer\ Age)_i + \varepsilon$$



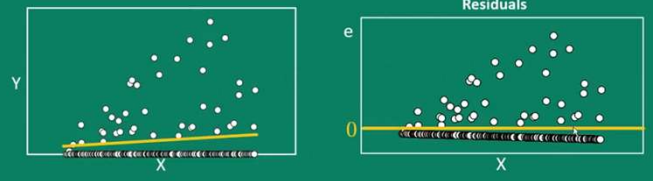
What's the issue?

- If normality is violated and n is small, standard errors in output are affected

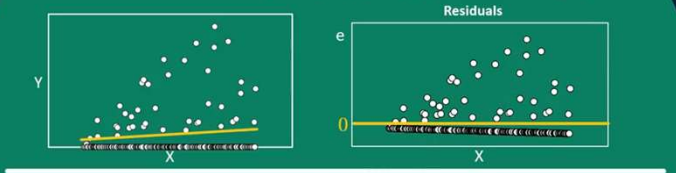
Normality of errors

Consider the following model:

$$Medical\ Insurance\ Payout_i = \beta_0 + \beta_1(Customer\ Age)_i + \varepsilon$$



$$Medical\ Insurance\ Payout_i = \beta_0 + \beta_1(Customer\ Age)_i + \varepsilon$$

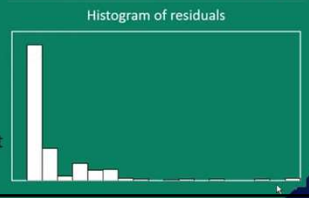


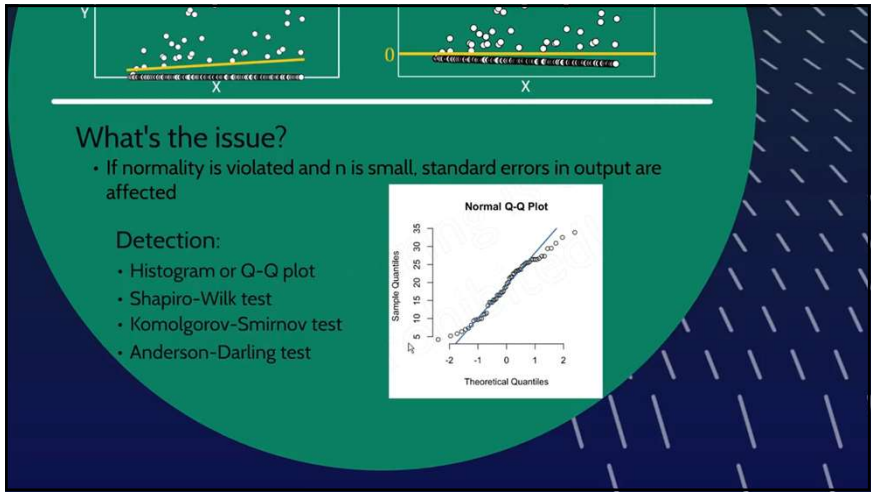
What's the issue?

- If normality is violated and n is small, standard errors in output are affected

Detection:

- Histogram or Q-Q plot
- Shapiro-Wilk test
- Komolgorov-Smirnov test
- Anderson-Darling test



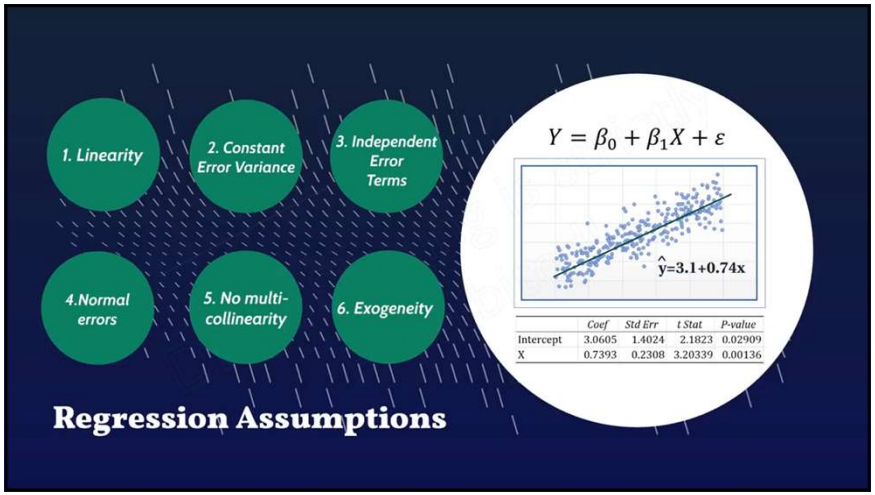


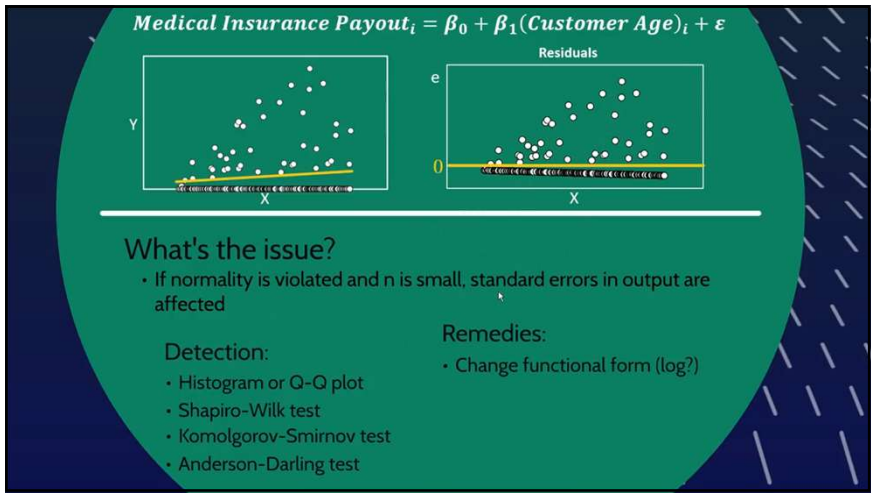
What's the issue?

- If normality is violated and n is small, standard errors in output are affected

Detection:

- Histogram or Q-Q plot
- Shapiro-Wilk test
- Komolgorov-Smirnov test
- Anderson-Darling test





What's the issue?

- If normality is violated and n is small, standard errors in output are affected

Detection:

- Histogram or Q-Q plot
- Shapiro-Wilk test
- Komolgorov-Smirnov test
- Anderson-Darling test

Remedies:

- Change functional form (log?)

No multicollinearity

Consider the following model:

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

$$i = suburb\ 1,2,3\ \ldots$$

No multicollinearity

Consider the following model:

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

$$i = suburb\ 1,2,3\ \ldots$$

Multi-collinearity occurs where the X variables are themselves related

What's the issue?

- Coefficients and standard errors of affected variables are unreliable.

No multicollinearity

Consider the following model:

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

$$i = suburb\ 1,2,3\ \ldots$$

Multi-collinearity occurs where the X variables are themselves related

No multicollinearity

Consider the following model:

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

$$i = suburb\ 1,2,3\ \ldots$$

Multi-collinearity occurs where the X variables are themselves related

What's the issue?

- Coefficients and standard errors of affected variables are unreliable.

Detection:

- Look at correlation (ρ) between X variables
- Look at Variance Inflation Factors (VIF)

Multi-collinearity occurs where the X variables are themselves related

What's the issue?

- Coefficients and standard errors of affected variables are unreliable.

Detection:

- Look at correlation (ρ) between X variables
- Look at Variance Inflation Factors (VIF)

Remedies:

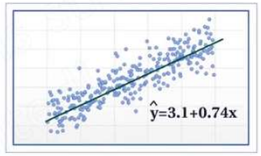
- Remove one of the variables

Exogeneity
(no omitted variable bias)

1. *Linearity*
2. *Constant Error Variance*
3. *Independent Error Terms*
4. *Normal errors*
5. *No multi-collinearity*
6. *Exogeneity*

Regression Assumptions

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



	Coef	Std Err	t Stat	P-value
Intercept	3.0605	1.4024	2.1823	0.02909
X	0.7393	0.2308	3.20339	0.00136

Exogeneity
(no omitted variable bias)

Consider the following model:

$$Salary_i = \beta_0 + \beta_1 (Years\ of\ education)_i + \varepsilon_i$$

Exogeneity
(no omitted variable bias)

Consider the following model:

$$Salary_i = \beta_0 + \beta_1(Years\ of\ education)_i + \varepsilon_i$$

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

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

What's the issue?

- Model can only be used for predictive purposes (can not infer causation)

Detection:

- Intuition
- Checking correlations

Consider the following model:

$$Salary_i = \beta_0 + \beta_1(Years\ of\ education)_i + \varepsilon_i$$

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

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

What's the issue?

- Model can only be used for predictive purposes (can not infer causation)

Detection:

- Intuition
- Checking correlations

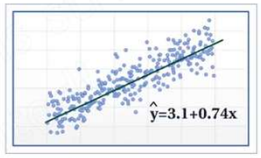
Remedy:

- Using instrumental variables

THANK YOU!

1. Linearity
2. Constant Error Variance
3. Independent Error Terms
4. Normal errors
5. No multi-collinearity
6. Exogeneity

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



	Coef	Std Err	t Stat	P-value
Intercept	3.0605	1.4024	2.1823	0.02909
X	0.7393	0.2308	3.20339	0.00136

Regression Assumptions