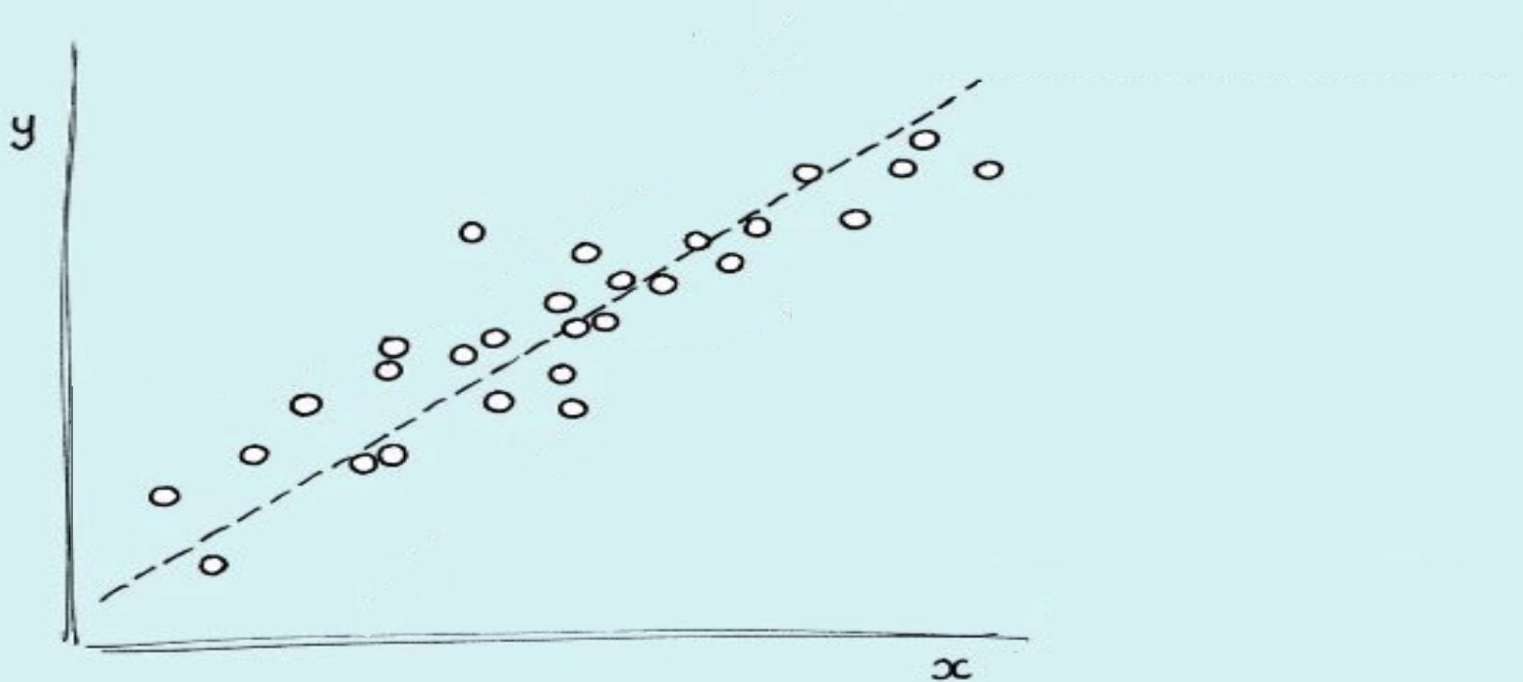


Linear Regression

Guided Practice



Poll

☐ A. Project Score

☐ B. Hours of Sleep

How to Edit

Click [Edit This Slide](#) in the plugin to make changes.

Don't have the Nearpod add-on? Open the "Add-ons" menu in Google Slides to install.





Matching Pairs

^ Instructions

Alternative
Hypothesis

Null Hypothesis

There is "some"
relation between
the amount spent
on TV

There is no
relationship
between the
amount spent on

How to Edit

Click [Edit This Slide](#) in the plugin to make changes.

Don't have the Nearpod add-on? Open the "Add-ons" menu in Google Slides to install.



Fill in the Blanks

Describe the each of the variables of the equation:

\hat{y} predicted value of our [] variable

b_0 [] when $x=0$

b_1 [] of the regression line

x an [] variable

How to Edit

Click [Edit This Slide](#) in the plugin to make changes.

Don't have the Nearpod add-on? Open the "Add-ons" menu in Google Slides to install.



Linear Regression Goals

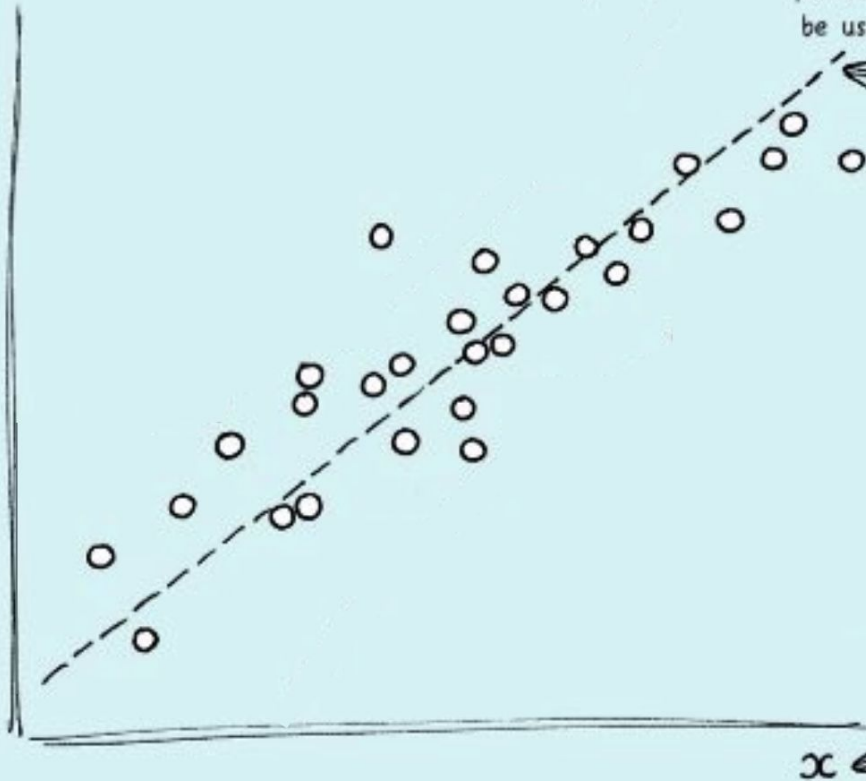
The thing we want
to explain

DEPENDENT
VARIABLE

y

If you only had data on x , this line
provides your best estimate of y . If the
fit is strong and no major outliers, x could
be used as a surrogate or forecast of y .

LINE OF BEST FIT



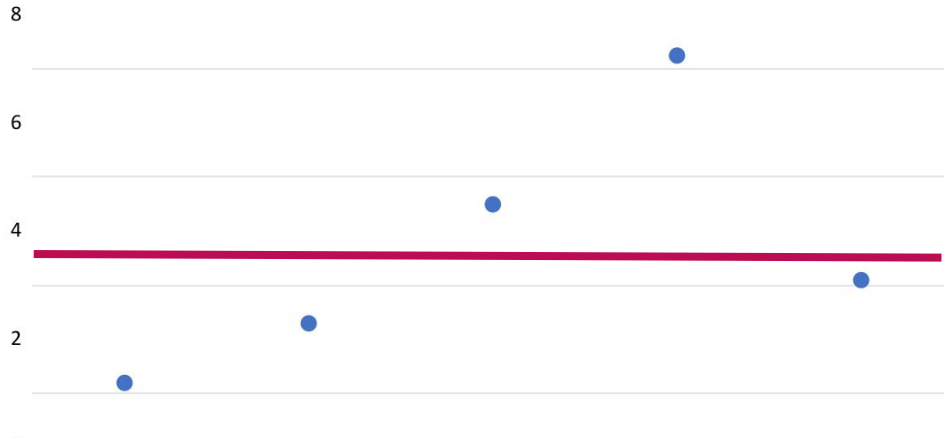
INDEPENDENT
VARIABLE

x

The factor we think
might influence the
dependent variable

How do we find the line of best fit?

For a dataset with only one variable, the best fit line is the mean value of the data points.

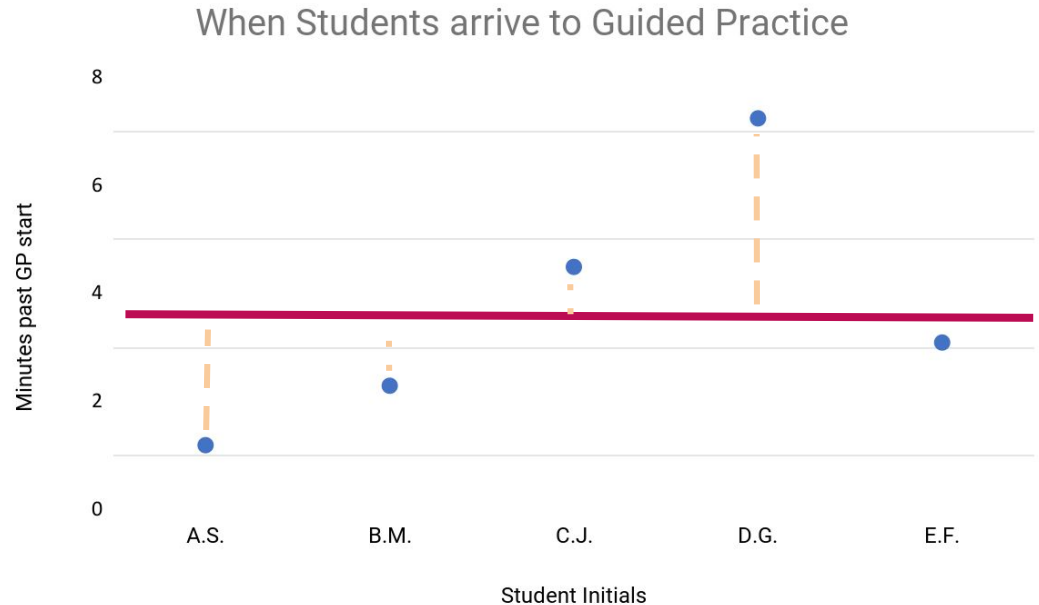


$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

Line of Best fit

For a dataset with only one variable, the best fit line is the mean value of the data points.

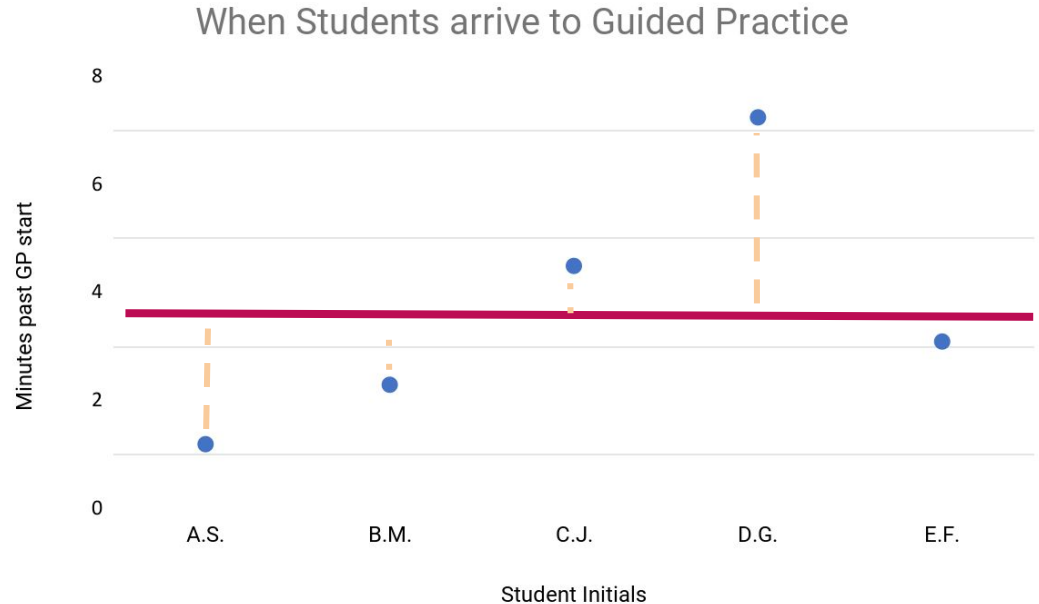
Student Initial	Minutes past start of GP
A.S.	1.2
B.M.	2.3
C.J.	4.5
D.G.	7.25
E.F.	3.1
Mean:	3.67



Line of Best fit

The sum of the distances (error) above the mean has the same absolute value as the distances (error) below the mean.

Student Initial	Minutes past start of GP	$(y - \bar{y})$
A.S.	1.2	-2.47
B.M.	2.3	-1.37
C.J.	4.5	0.83
D.G.	7.25	3.58
E.F.	3.1	-0.57
Mean:	3.67	

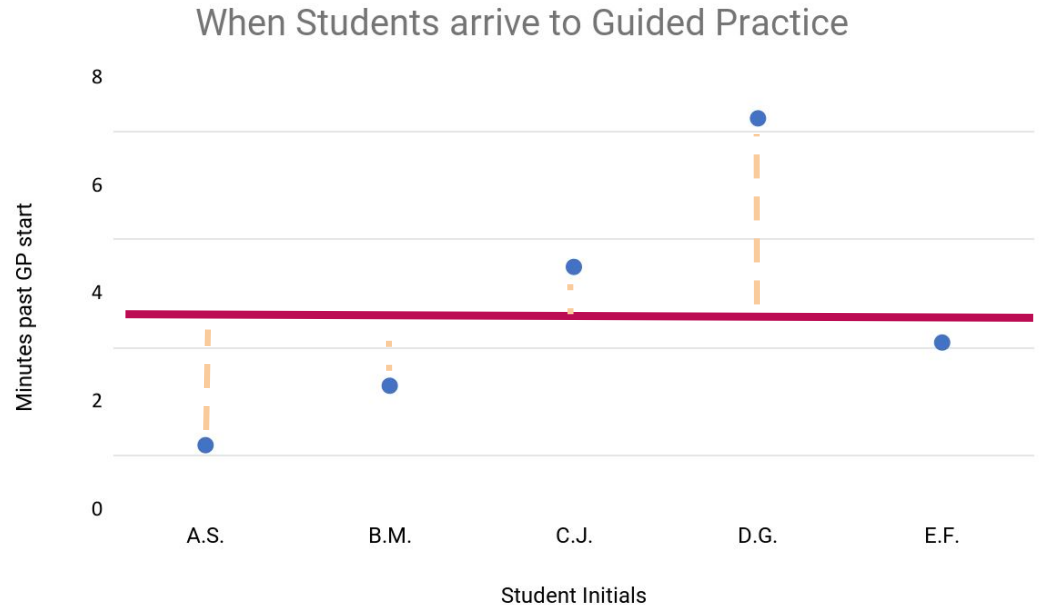


Distance = Residual = Error

Line of Best fit

The goal of the line of best fit is to minimize the Sum of Squares Error.

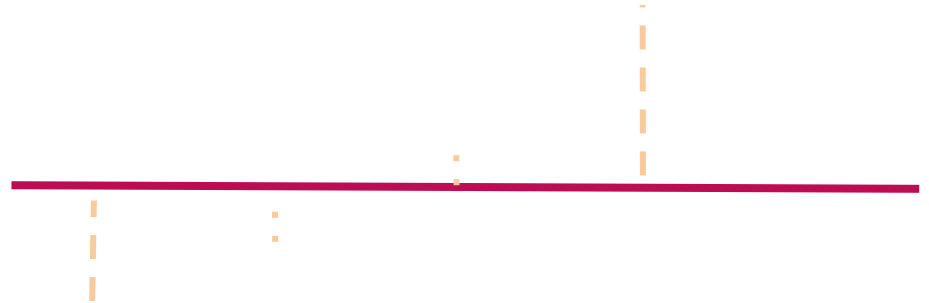
Error (E)	Square Error (SE)
-2.47	6.1009
-1.37	1.8769
0.83	0.6889
3.58	12.8164
-0.57	0.3249
-3.67	13.4689
Sum:	35.2769



Line of Best fit

Ordinary Least Square (OLS) regression utilizes the principle of least squares, i.e. minimize the sum of the squares of the differences.

Error (E)	Square Error (SE)
-2.47	6.1009
-1.37	1.8769
0.83	0.6889
3.58	12.8164
-0.57	0.3249
-3.67	13.4689
Sum:	35.2769





Open Ended Question

Ready? Enter your answer here.

How to Edit

Click [Edit This Slide](#) in the plugin to make changes.

Don't have the Nearpod add-on? Open the "Add-ons" menu in Google Slides to install.



Equation of a Line

$$y = mx + b$$

&

Linear Regression Equation

$$\hat{y} = b_0 + b_1x$$

Simple Linear Regression Equation

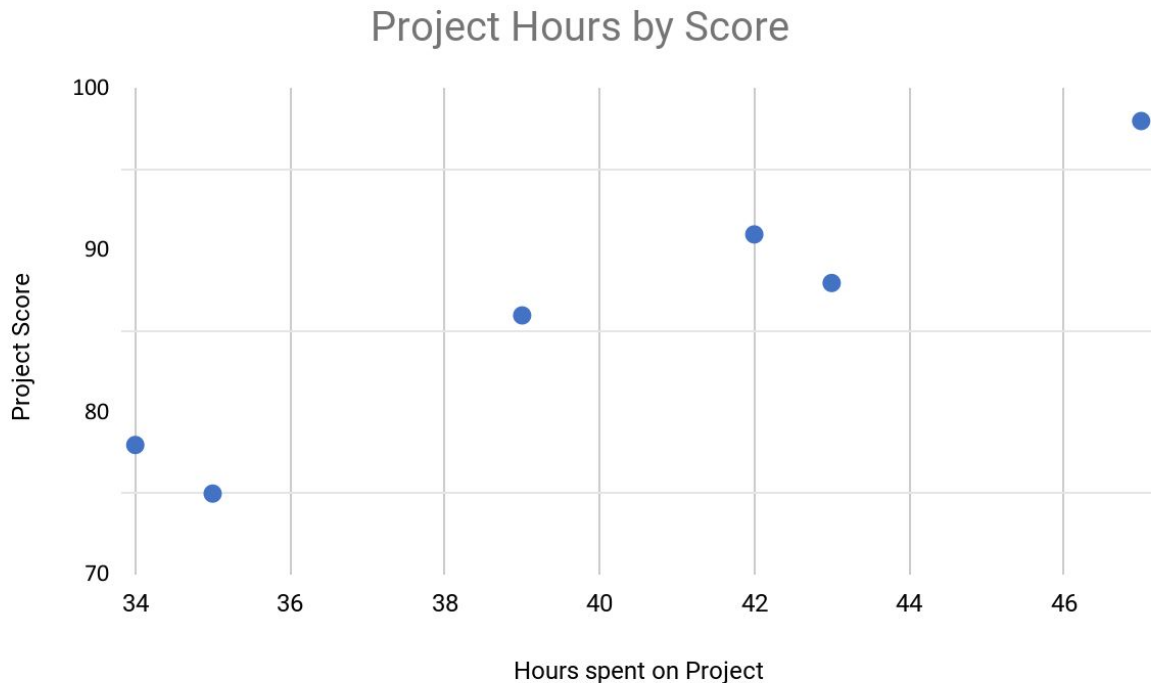
$$\hat{y} = b_0 + b_1x$$

- \hat{y} the predicted value of our dependent variable y
- b_0 the y -intercept when x is equal to 0
- b_1 slope of the simple linear regression
- x the independent variable

Linear Regression Example

Let's explore the relationship between number of hours spent on a project and the score a project was graded.

Project Hours (x)	Project Score (y)
34	78
35	75
39	86
42	91
43	88
47	98



Linear Regression Example

We need to run our calculations to determine our coefficients, i.e. **fit our model**.

$$\hat{y} = b_0 + b_1x$$
$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1\bar{x}$$

Project Hours (x)	Project Score (y)	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
34	78	-6	-8	48	36
35	75	-5	-11	55	25
39	86	-1	0	0	1
42	91	2	5	10	4
43	88	3	2	6	9
47	98	7	12	84	49
40	86		Sum:	203	124

Simple Linear Regression Equation

- $b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = 1.63$

- $b_0 = \bar{y} - b_1\bar{x} = 86 - (1.63 \times 40) = 20.8$

$$\hat{y} = b_0 + b_1x$$

$$\hat{y} = 20.8 + 1.63x$$

Simple Linear Regression Equation

$$\hat{y} = 20.8 + 1.63x$$

With our linear model we could predict the number of hours spent for a score of 75:

$$80 = 20.8 + 1.63x$$

$$x = 36.31 \text{ hours}$$

Simple Linear Regression Equation

$$\hat{y} = 20.8 + 1.63x$$

With our linear model we could predict the number of hours spent for a score of 80:

$$80 = 20.8 + 1.63x$$

$$x = 36.31 \text{ hours}$$

Anscombe's Quartet: The Importance of Visualization

```
import matplotlib.pyplot as plt
import numpy as np

x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
x4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

datasets = {
    'I': (x, y1),
    'II': (x, y2),
    'III': (x, y3),
    'IV': (x4, y4)
}

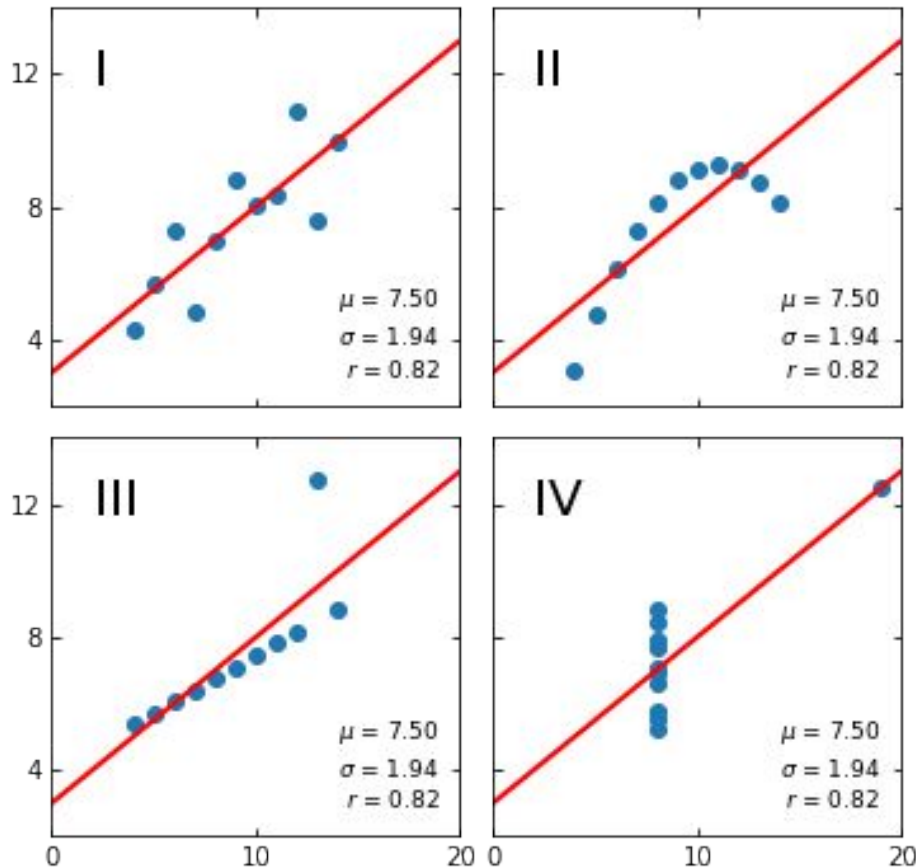
fig, axs = plt.subplots(2, 2, sharex=True, sharey=True, figsize=(6, 6),
                        gridspec_kw={'wspace': 0.08, 'hspace': 0.08})
axs[0, 0].set(xlim=(0, 20), ylim=(2, 14))
axs[0, 0].set(xticks=(0, 10, 20), yticks=(4, 8, 12))

for ax, (label, (x, y)) in zip(axs.flat, datasets.items()):
    ax.text(0.1, 0.9, label, fontsize=20, transform=ax.transAxes, va='top')
    ax.tick_params(direction='in', top=True, right=True)
    ax.plot(x, y, 'o')

    # linear regression
    p1, p0 = np.polyfit(x, y, deg=1) # slope, intercept
    ax.axline(xyl=(0, p0), slope=p1, color='r', lw=2)

    # add text box for the statistics
    stats = (f'$\mu$ = {np.mean(y):.2f}\n'
             f'$\sigma$ = {np.std(y):.2f}\n'
             f'$r$ = {np.corrcoef(x, y)[0][1]:.2f}')
    #bbox = dict(boxstyle='round', fc='blanchedalmond', ec='orange', alpha=0.5)
    ax.text(0.95, 0.07, stats, fontsize=9,
            transform=ax.transAxes, horizontalalignment='right')

plt.show()
```



Assumptions of Linearity_(after)



when should i check for assumptions of linear regression



[All](#)

[Videos](#)

[News](#)

[Images](#)

[Shopping](#)

[More](#)

[Tools](#)

About 310,000,000 results (0.66 seconds)

The very first step **after building a linear regression model** is to check whether your model meets the assumptions of linear regression. These assumptions are a vital part of assessing whether the model is correctly specified.

<https://www.godatadrive.com/blog/basic-guide-to-test-...>

[What Are the Assumptions of Linear Regression? - DataDrive](#)

[About featured snippets](#) • [Feedback](#)

People also ask

Why do we need to **check assumptions before we run regression?**

If the assumptions of regression analysis are met, then **the errors associated with one variable are not correlated with the errors of any other variables**. Independence of residuals can be examined via the Durban – Watson statistic which tests for correlations between errors.


https://www.researchgate.net/post/Why_is_it_importan...

[Why is it important to examine the assumption of linearity when using ...](#)


Search for: [Why do we need to check assumptions before we run regression?](#)


Let's open our notebook


https://github.com/mojo-flat/lin_reg_gp1

 [mojo-flat / lin_reg_gp1](#) Private

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Security](#) [Insights](#) [Settings](#)

 main [lin_reg_gp1 / lin_reg_gp1.ipynb](#)

 **mojo-flat** added presentation & solution Latest commit

 1 contributor

1356 lines (1356 sloc) | 60.2 KB <>

Linear Regression GP 1

GP 1 Goals

Performing simple linear regression and understanding evaluation metrics.

- Run a simple linear regression model in statsmodels or scikit-learn.
- Discuss the real-world implications of the model results.

LINEAR REGRESSION

The thing we want
to explain

DEPENDENT
VARIABLE

y

i.e 77% of the variance in y is
explained by x. Below c.30% means
they're hardly connected. Above 95%
and they're practically the same.

$$R^2 = 0.77$$

If you only had data on x, this line
provides your best estimate of y. If the
fit is strong and no major outliers, x could
be used as a surrogate or forecast of y.

LINE OF BEST FIT

DATA
POINT

95% CONFIDENCE BAND

If a data point falls outside these
lines, you're 95% sure there is
something special about it causing it
to do better or worse than others -
an 'outlier' worth understanding

OUTLIER

INDEPENDENT
VARIABLE

x

The factor we think
might influence the
dependent variable

Interpreting R Squared

- Percentage of variation in the dependent variable explained by the independent variable. Values between 0 & 1

$$SS_{residual} = \sum (y - \hat{y})^2$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

$$SS_{total} = \sum (y - \bar{y})^2$$

An R-squared value of 0.928 can be described conceptually as:

92.8% of the variations in dependent variable score are explained by the independent variables hours in our model.

Interpreting Model Coefficients

Understand the **Marginal Effect** of the independent variable on the dependent variable.

Given a one-unit change in the independent variable, how much is the mean of the dependent variable changed.

b_1 $x + 1$ = Increase in mean score of 1.637 points

b_0 $x = 0$, score is equal to 20.8 points

Hypothesis Testing

To determine if our independent variable has a statistically significant relationship with the dependent variable, we conduct a hypothesis test.

The null hypothesis should contain an equality ($=, \leq, \geq$):
Average NBA Player's Height = 2.0m (6ft 7in)

The alternate hypothesis should not have an equality ($\neq, <, >$):
Average NBA Player's Height \neq 2.0m (6ft 7in)

- $H_0 : \mu = 3.5$

- $H_1 : \mu \neq 3.5$

The null hypothesis should contain an equality ($=, \leq, \geq$):
old scores \geq new scores

The alternate hypothesis should not have an equality ($\neq, <, >$):
old scores $<$ new scores

$H_0 : \mu \geq \mu$

$H_1 : \mu < \mu$

- H_0 - The average NBA player's height is 2.0m tall.
- H_1 - The average NBA player's height is not 2.0m tall.
- H_0 - The old average of the scores is equal to or greater than the new average of the scores.
- H_1 - The old average of the scores is less than the new average of the scores.

p-values & Hypothesis Tests

From: *Interpreting Significance and p-values*

We reject or fail to reject a null hypothesis based on an associated significance level or p-value.

The p-value represents a probability of observing your results (or something more extreme) given that the null hypothesis is true

Applied to a regression model, p-values associated with coefficients indicate the probability of observing the associated coefficient given that the null-hypothesis is true. As a result, very small p-values indicate that coefficients are statistically significant. A very commonly used cut-off value for the p-value is 0.05.

Just like for statistical significance, rejecting the null hypothesis at an alpha level of 0.05 is the equivalent for having a 95% confidence interval around the coefficient that does not include zero. In short

The p-value represents the probability that the coefficient is actually zero.

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

Interpreting p-values

OLS Regression Results

Dep. Variable:	score	R-squared:	0.928
Model:	OLS	Adj. R-squared:	0.910
Method:	Least Squares	F-statistic:	51.79
Date:	Sun, 05 Jun 2022	Prob (F-statistic):	0.00198
Time:	20:10:26	Log-Likelihood:	-12.874
No. Observations:	6	AIC:	29.75
Df Residuals:	4	BIC:	29.33
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	20.5161	9.158	2.240	0.089	-4.911	45.944
hours	1.6371	0.227	7.196	0.002	1.005	2.269

Omnibus:	nan	Durbin-Watson:	2.912
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.929
Skew:	-0.588	Prob(JB):	0.628
Kurtosis:	1.473	Cond. No.	357.

- Prob(F-statistic) p-value: likelihood that we observe our score values by random chance if **linear model** had no statistically significant relationship.
- $P > |t|$ p-value: likelihood that we observe our score values by random chance if **hours** spent had no statistically significant relationship.

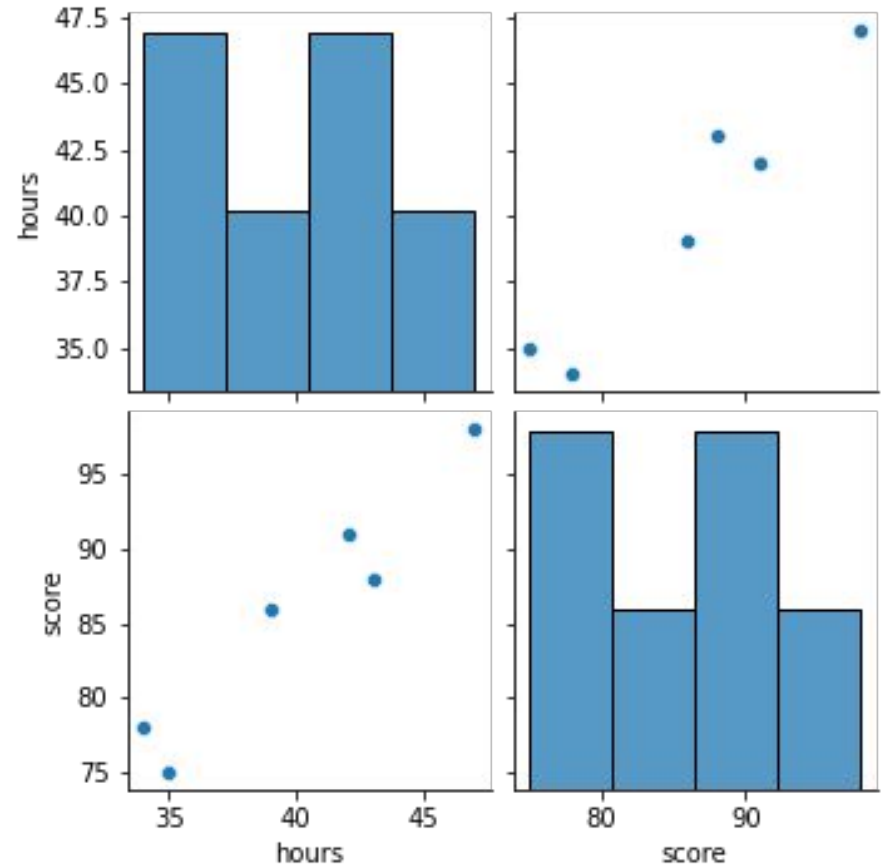
$$\alpha 0.05 > p$$

Assumptions of Linear Regression

- **Linearity**: there is a linear relationship between the independent and dependent variables
- **Normality**: residuals are normally distributed
- **Homoscedasticity**: the variance for the residual is the same for any value of x
- **Independence**: observations are independent of one another

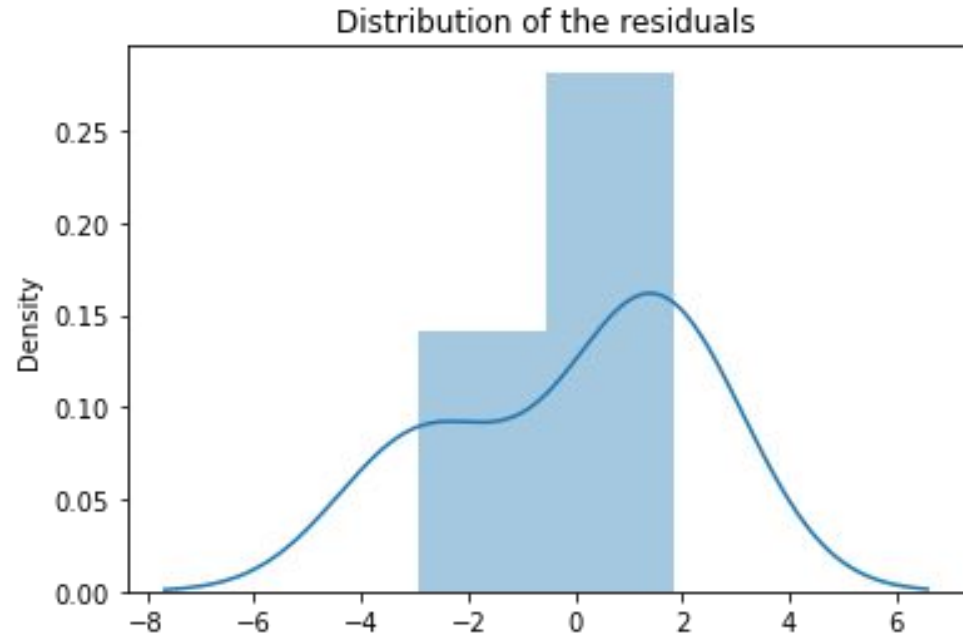
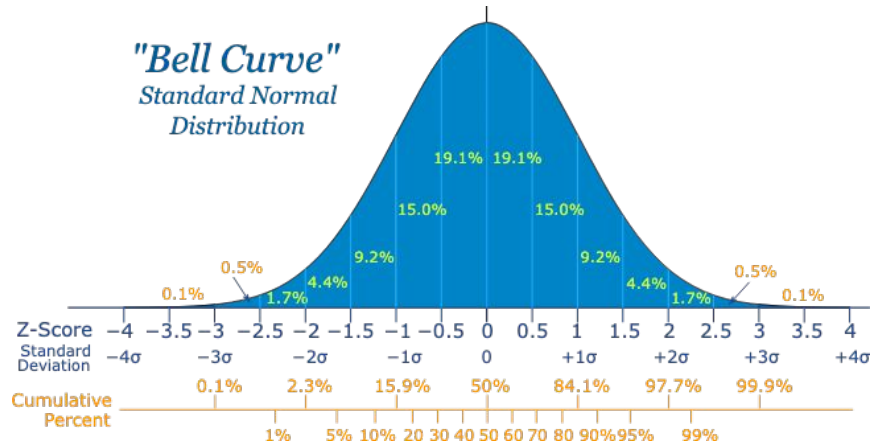
Assumption of Linearity

- **Linearity:** there is a linear relationship between the independent and dependent variables



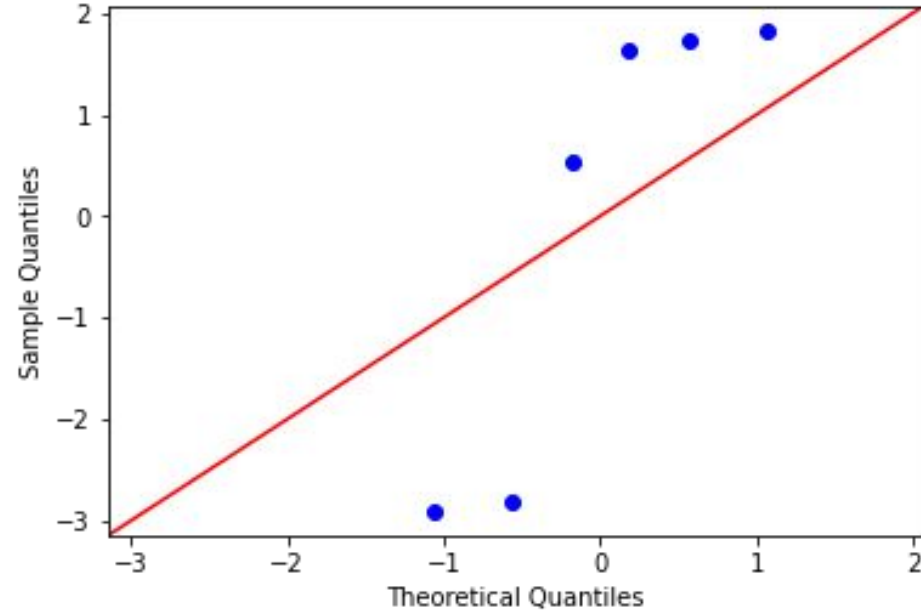
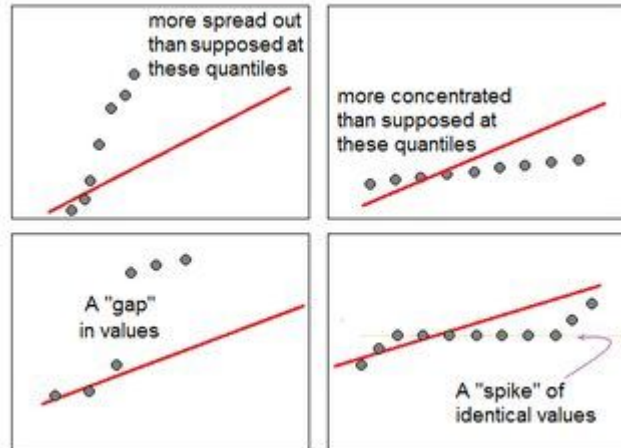
Assumption of Normality

- **Normality:** residuals are normally distributed (symmetric about the mean)



Normality (quantile-quantile plot)

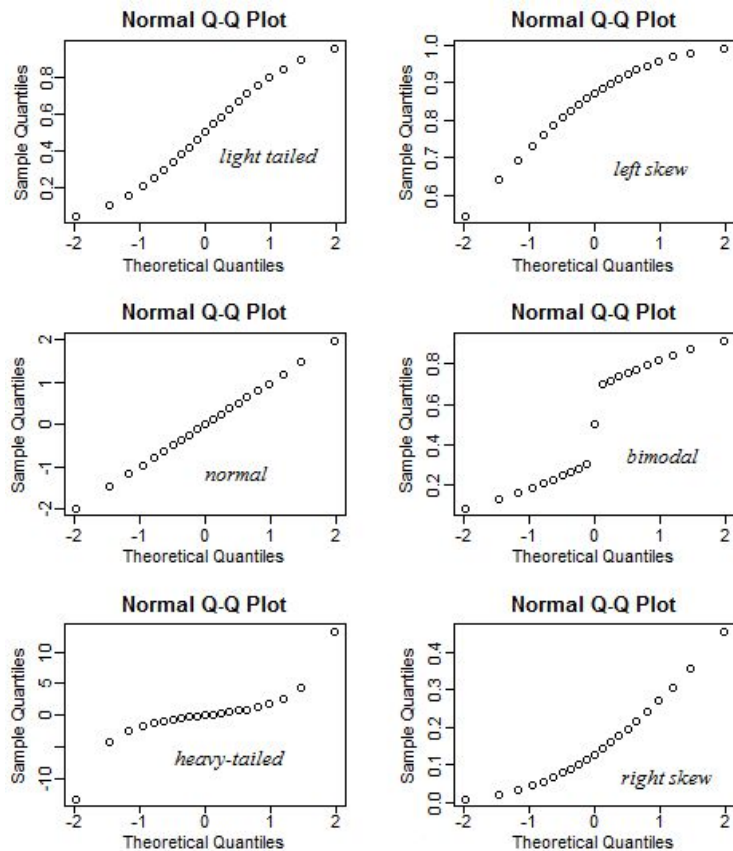
Plot the quantiles of the residuals to compare distributions, if points lie close to or along 45° line from *x-axis*, samples have similar distributions.



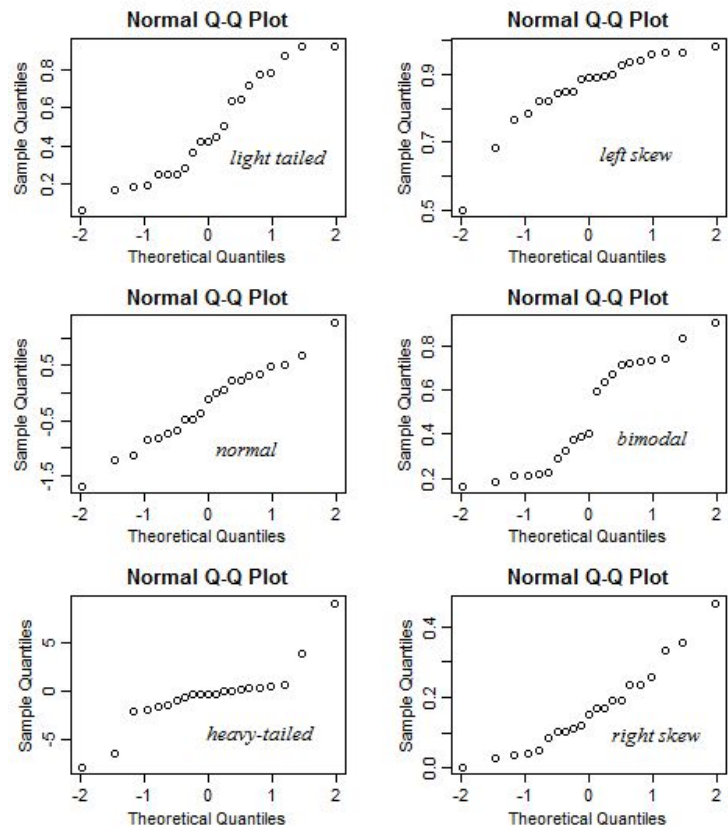
Our Q-Q plot indicates a gap where before it y quantiles are lower than the x quantiles and after which the y quantiles are higher.

Q-Q Plot

Without randomness

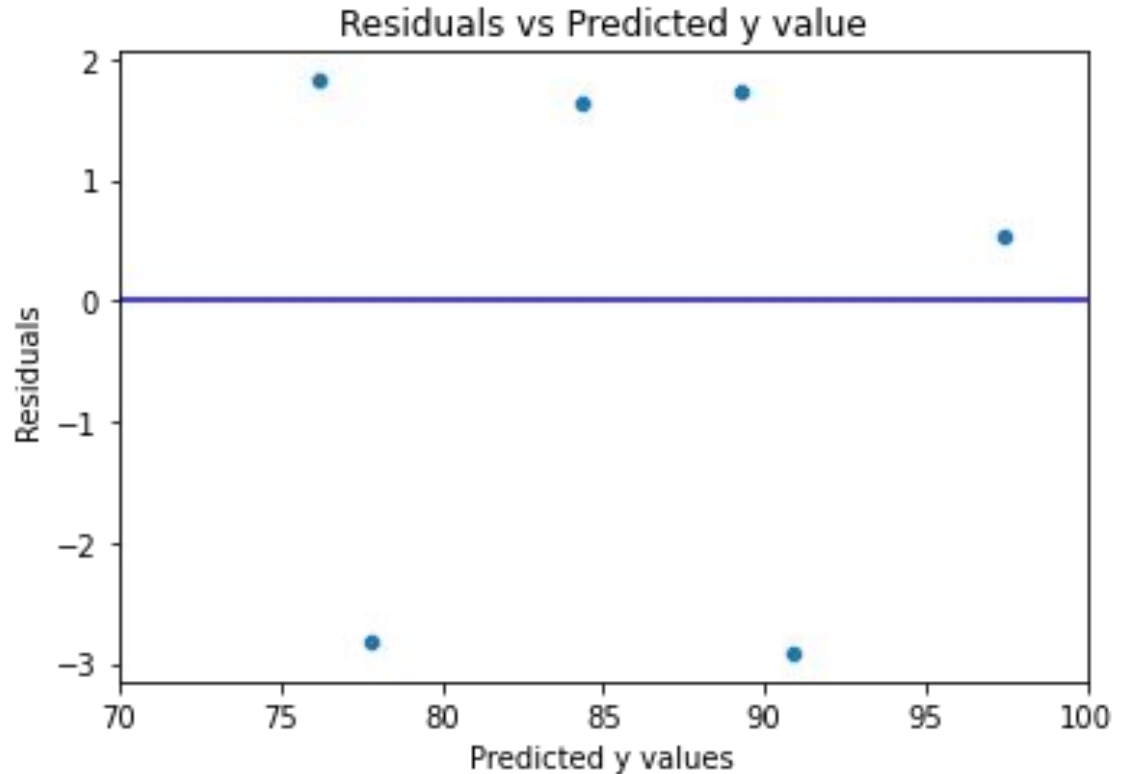


With randomness



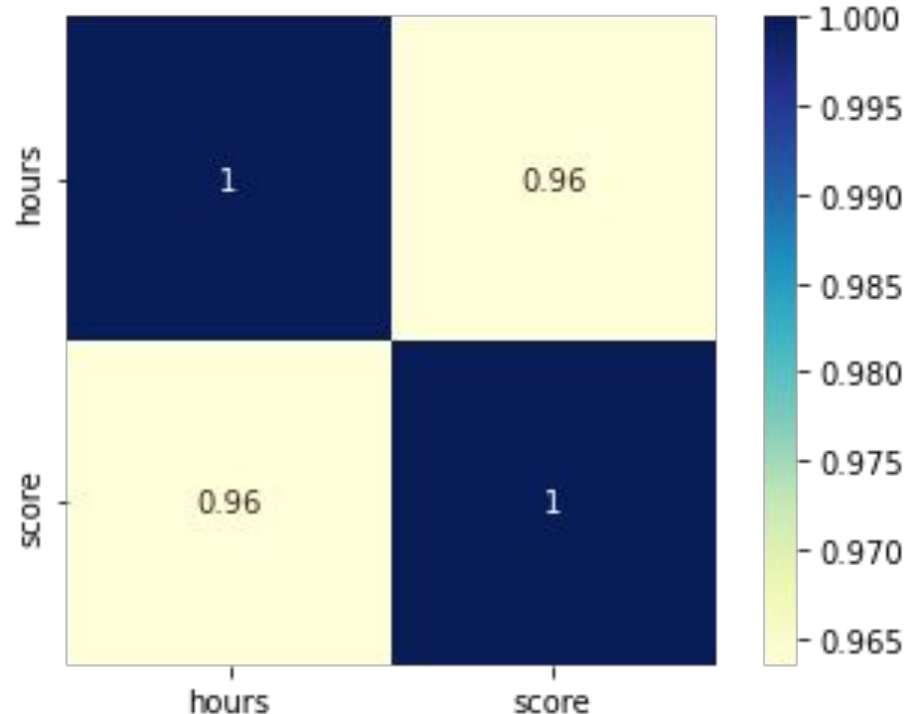
Assumption of Homoscedasticity

- **Homoscedasticity:**
the variance for the residual is the same for any value of x



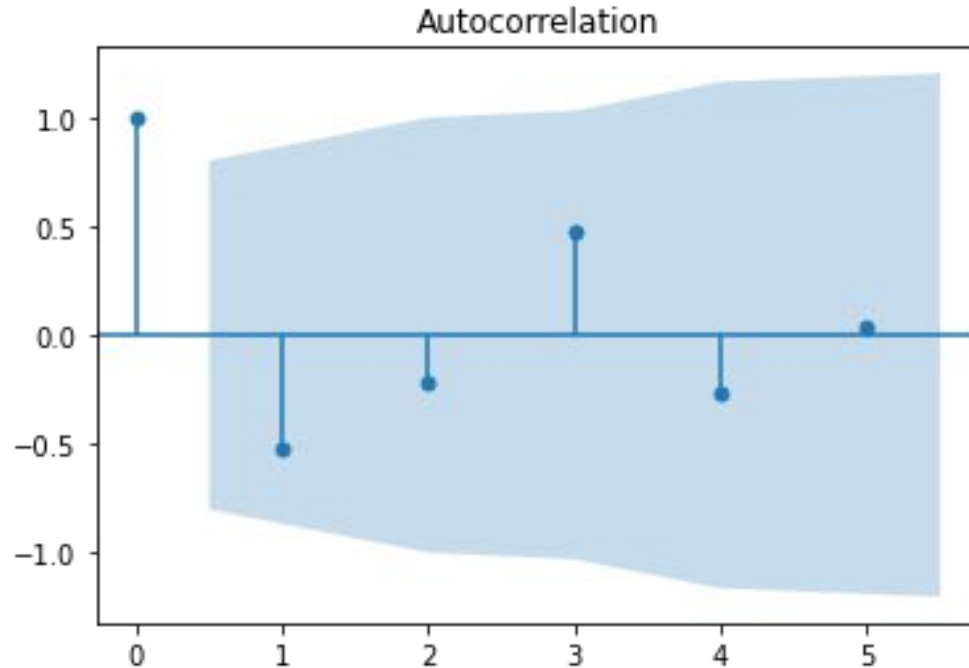
Assumption of Independence

- No perfect multicollinearity



Assumption of Independence

- **Autocorrelation:** correlation between the residuals. This would violate an assumption of independence



Linear Regression Q & A

The thing we want
to explain

DEPENDENT
VARIABLE

y

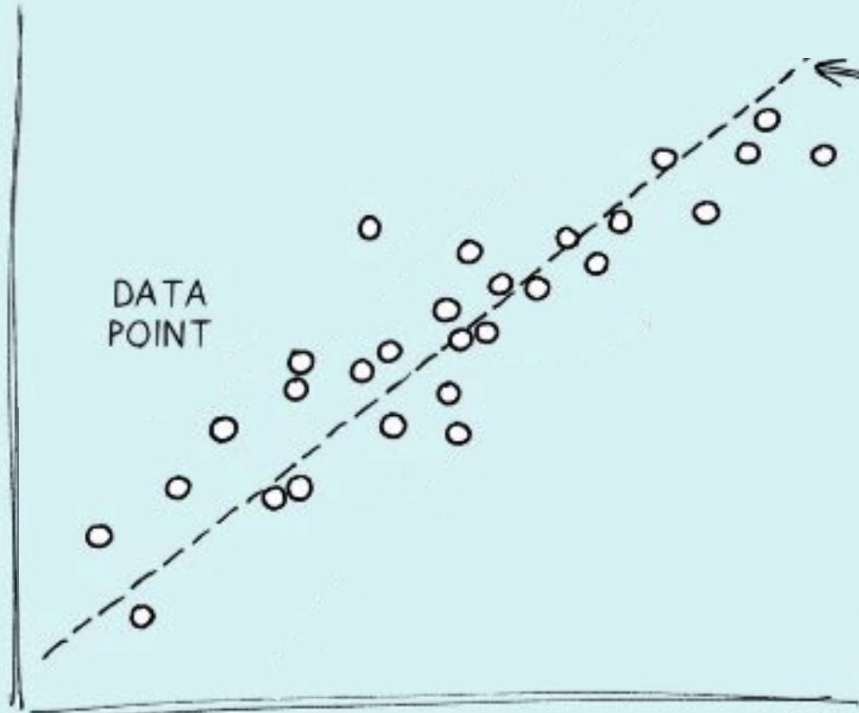
DATA
POINT

LINE OF BEST FIT

INDEPENDENT
VARIABLE

x

The factor we think
might influence the
dependent variable



Linear Regression Appendix

The thing we want
to explain

DEPENDENT
VARIABLE

y

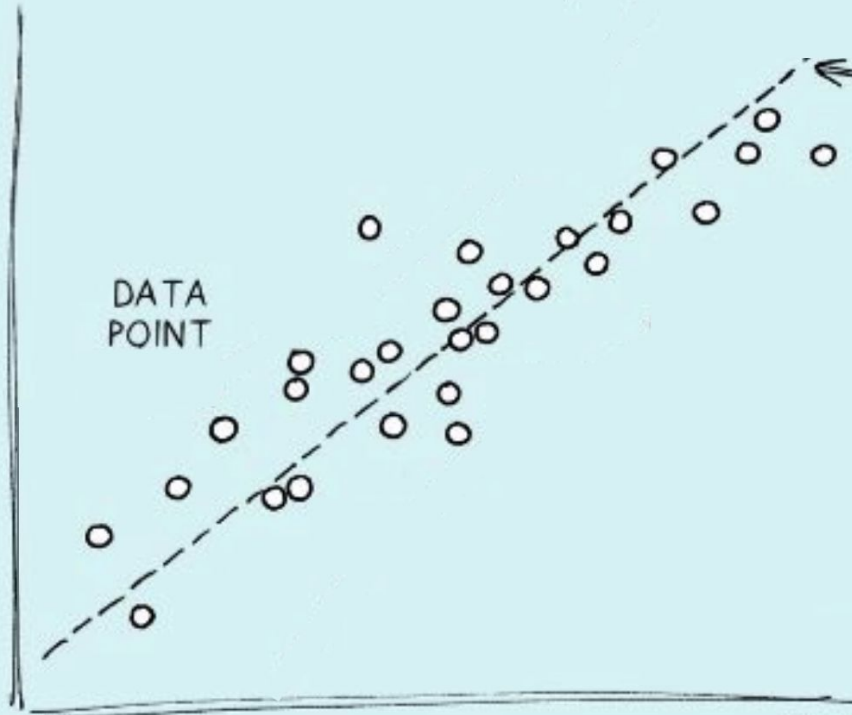
DATA
POINT

LINE OF BEST FIT

INDEPENDENT
VARIABLE

x

The factor we think
might influence the
dependent variable



Pearson Correlation Coefficient

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n} \sum (x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\sum (x - \bar{x})^2}{n}} \sqrt{\frac{\sum (y - \bar{y})^2}{n}}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

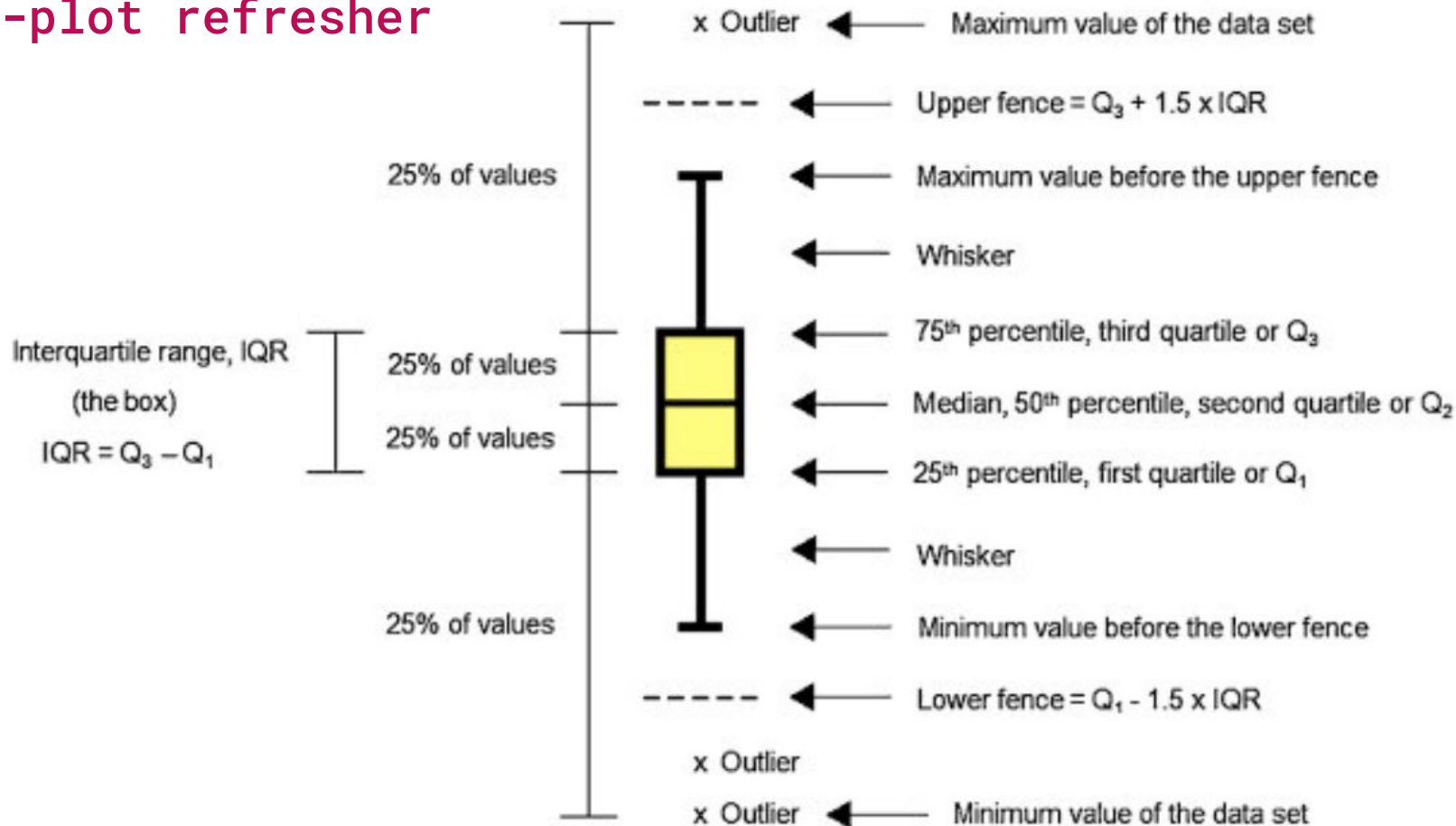
ρ = Greek letter “rho”

σ = standard deviation

cov = covariance

\bar{x} = mean of X

Box-plot refresher



Line of Best fit

The sum of the distances (error) above the mean has the same absolute value as the distances (error) below the mean.

Error (E)	Square Error (SE)
-2.47	6.1009
-1.37	1.8769
0.83	0.6889
3.58	12.8164
-0.57	0.3249
-3.67	13.4689
Sum:	35.2769

