#1 (Total 30 Points)

The top management of TechCo Inc. is concerned about the increasing cyber incidences occurring in their industry. In a weekly strategy meeting, the managers discuss the problem of remote logins and possible intrusions into the IT infrastructure of the company. The COVID-19 situation forced the organization to allow their members to work from home while the cybersecurity department could not timely implement adequate security measures. The Head of Cyber Risk Management should develop an analytical concept to detect outliers in the remote login data. She starts with a data set containing the number of remote logins during one day of the Research & Development department.

Time	# remote logins
8:00 – 9:00 o'clock	24
9:00 – 10:00 o'clock	21
10:00 – 11:00 o'clock	25
11:00 – 12:00 o'clock	29
12:00 – 13:00 o'clock	30
14:00 – 15:00 o'clock	22
16:00 – 17:00 o'clock	27
17:00 – 18:00 o'clock	26

(1) Develop a 95% confidence interval around the mean and determine graphically whether outliers exist in the data or not. The t-value is given with $t_{(0,05/2)}^{[8-1]} = 2{,}37$.
(2) Check whether the data are (from a practical point of view) normally distributed or not. Use standardized data and construct a histogram. Develop adequate bins according to the following rule: $Lower\ Level \leq Value < Upper\ Level$.
(3) Why is it necessary to check the normality of the data?

#2 (Total 45 Points)

John Durbin and Bill Watson are cybersec managers at the TechnoCorp Inc.. Both are specialized in time series regression topics. The production runs per week showed several defective products. Therefore, the two managers collect data on the number ('000) of unsuccessful remote logins per week. The following data are available

# ('000) of uns. remote logins	Week
933	1
826	2
748	3
908	4
983	5
1009	6
1101	7
1149	8
1207	9
1255	10

The results of the linear regression analysis are

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,91
R Square	0,82
Adjusted R Square	0,80
Standard Error	73,69
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	202.517,05	202.517,05	37,29	0,00
Residual	8	43.445,85	5.430,73		
Total	9	245.962,90			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	739,40	50,34	14,69	0,00	623,31	855,49
Week	49,55	8,11	6,11	0,00	30,84	68,25

(1) Calculate the residuals $e_t = y_t - \hat{y}_t$ and the squared residuals e_t^2.
(2) Compute the standardized residuals using the formula $e_t^z = \frac{e_t}{\sqrt{\frac{1}{n-1} * \sum_{t=1}^{n} e_t^2}}$
(3) Check whether positive or negative autocorrelation exists in the data.

$$d_{L;0,05} = 0{,}879 \text{ and } d_{U;0,05} = 1{,}320$$

#3 (Total 15 Points)

A large cybersecurity firm manages a central server that holds various sensitive information about its clients. Security analysts from offices in six different regions use secure internet connections to access this central database for their investigative tasks. Currently, the server allows three users to log in and access the database simultaneously. Analysts attempting to log in when the server is at capacity are denied access, with no waiting queue. Management recognizes that, with the growing number of clients, more access requests will be made to the server. Denying access to analysts is inefficient and frustrating. Access requests follow a Poisson probability distribution, with an average of 42 login attempts per hour. The service rate for each connection is 20 logins per hour.

(1) What is the probability that 0, 1, 2 and 3 access lines will be in use?
(2) What is the probability that a user will be denied access to the system?
(3) What is the average number of access lines in use?
(4) In planning for the future, management wants to be able to handle λ=50 database requests per hour; in addition, the probability that a user will be denied access to the system should be no greater than the value computed in part (2). How many access lines should this system have?

#4 (Total 30 Points)

Navindra Kalpakutta is a new business analyst in the Security Operations Center of the India Hightech Corporation. On his first day, the IT-Systems together with all applications (including the data analytics software) were not available due to a serious breakdown of the company's IT infrastructure. The multiple regression analysis regarding the cost of data breaches (in T€), downtime in operations A (in minutes) and downtime in operations B (in minutes) was interrupted. The following information are available

$$(\boldsymbol{X}'\boldsymbol{X}) = \begin{bmatrix} 6 & 8 & 9 \\ 10 & 5 & 7 \\ 8 & 5 & 6 \end{bmatrix} \text{ and } \boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 120 \\ 150 \\ 160 \end{bmatrix}$$

Support Navindra to solve the following problems.

[1] Compute the determinant of $(\boldsymbol{X}'\boldsymbol{X})$.
[2] Develop the Cofactor matrix $\boldsymbol{C}$.
[3] Calculate the inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
[4] Compute the matrix of the $\boldsymbol{b}$-extimator.

#5 (Total 30 Points)

Let's assume that an organization experienced the following number of phishing attempts over a 12-month period:

Month	Phishing attempts
January	10
February	15
March	18
April	20
May	25
June	30
July	35
August	40
September	45
October	50
November	55
December	60

(1) Compute a simple exponential smoothing model with $\alpha = 0{,}07$ and y_0 as the mean value of all given empirical values.
(2) Compute the standard error s.
(3) Compute a 95% prediction interval for t = 13 with $z_{[0,025]} = 1{,}96$.

#6 (Total 45 Points)

The Biotech Company produces an input factor for a new pharmaceutical product. Last year the entire manufacturing landscape was transformed, and a new cyber-physical process is now in place. Due to several incidences that caused downtimes, the controlling department is concerned about the costs induced by these incidents. Therefore, a data analyst evaluates the following data.

Total Loss in T€	Downtime in minutes
10	8,1
14	9,6
5	5,7
8	6,8
9	8,7
12	10,9
4	4,3
7	4,8
11	8,3
13	7,6
6	7,3

The result of the regression analysis, where the downtime is the independent and the total loss is the dependent variable, is given with

$$\hat{y}_i = -1{,}00 + 1{,}34 * x_i$$

The residual analysis shows the following values

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*	*Standard Residuals*
1	9,85	0,15	0,07
2	11,86	2,14	1,09
3	6,64	-1,64	-0,83
4	8,11	-0,11	-0,06
5	10,66	-1,66	-0,84
6	13,60	-1,60	-0,82
7	4,76	-0,76	-0,39
8	5,43	1,57	0,80
9	10,12	0,88	0,45
10	9,18	3,82	1,94
11	8,78	-2,78	-1,41

The following analyses are necessary to get a better insight into the data.

[1] Develop a histogram to assess whether the results of the regression analysis are normally distributed or not. Which data should be used: the residuals or the standardized residuals? Develop adequate bins according to the following rule: $Lower\ Level \leq number\ of\ values < Upper\ Level$.

[2] Perform a hypothesis test and assess the statistical significance of the slope parameter and the y-intercept. The sum of squared residuals is given with 38,72. The t-statistics of the y-intercept is given with $t_{b_0} = -0{,}39$. The analyst is slightly confused which value from the t-distribution should be used. Three values are given: $t^{(11)}_{[0{,}025]} = 2{,}20, t^{(10)}_{[0{,}025]} = 2{,}23, t^{(9)}_{[0{,}025]} = 2{,}26$.

[3] Construct the 95% confidence intervals for the y-intercept and the slope parameter. The standard error for b_0 is given with 2,54.

#7 (Total 17 Points)

City Cab Inc. uses two dispatchers to handle requests for service and to dispatch the cabs. The telephone calls that are made to City Cab use a common telephone number. When both dispatchers are busy, the caller hears a busy signal: no waiting is allowed. Callers who receive a busy signal can call back later or call another cab service. Assume that the arrival of calls follows a Poisson probability distribution with a mean of 40 calls per hour and that each dispatcher can handle a mean of 30 calls per hour.

[1] What percentage of time are both dispatchers idle?
[2] What percentage of time are both dispatchers busy?
[3] What is the probability that callers will receive a busy signal if two, three, or four dispatchers are used?
[4] If management wants no more than 12% of the callers to receive a busy signal, how many dispatchers should be used?

#8 (Total 40 Points)

A company wants to predict the likelihood of a cyber-attack on their network using a single regression analysis. They collect data on the number of past attacks, the number of employees, and the number of network vulnerabilities.Let's assume that the company has collected data on the number of past cyber-attacks, the number of employees, and the number of network vulnerabilities for a period of 10 years. The collected data is as follows:

Year	Attacks	Employees	Vulnerabilities
1	10	100	20
2	15	120	25
3	18	150	30
4	20	180	35
5	25	210	40
6	30	240	45
7	35	270	50
8	40	300	55
9	45	330	60
10	50	360	65

[1] Run a single regression analysis using the number of past attacks as the dependent variable and the number of network vulnerabilities as the independent variables.
[2] Examine the regression coefficients and determine the relationship between the independent and dependent variables.
[3] Now, the analysts look at the second model. The regression output shows the relationship between the number of attacks (dependent variable) and the number of employees (independent variable). Assess whether the regression coefficients are statistically significant or not. Perform a hypotheses test for this model using the following information: $b_0 = -5{,}15, b_1 = 0{,}15, s_{b_0} = 1{,}15, s_{b_1} = 0{,}005, t^8_{0{,}025} = 2{,}31$.
[4] Computer the confidence intervals for both regression parameters of the second model.

#9 (Total 25 Points)

The following data show the number of remote logins of a research and development department. Support the management to assess the problem of outliers in the data.

Day	# remote logins
Monday	20
Tuesday	41
Wednesday	35
Thursday	29
Friday	22
Saturday	45
Sunday	27

(1) Check whether the data are (from a practical point of view) normally distributed or not. Use standardized data and construct a histogram. Develop adequate bins according to the following rule: $Lower\ Level \leq \#\ of\ values < Upper\ Level$.
(2) Develop a 95% confidence interval around the mean and determine graphically whether outliers exist in the data or not. The t-value is given with $t_{(0,05/2)}^{[7-1]} = 2{,}45$.
(3) Why is it necessary to check the normality of the data?

#10 (Total 25 Points)

Let's assume that an organization experienced the following number of phishing attempts over a 12-month period:

Month	Phishing attempts
January	10
February	15
March	18
April	20
May	25
June	30
July	35
August	40
September	45
October	50
November	55
December	60

(1) Compute a simple exponential smoothing model with $\alpha = 0{,}07$ and y_0 as the mean value of all given empirical values.
(2) Compute the standard error s.
(3) Compute a 95% prediction interval for t = 13 with $z_{[0,025]} = 1{,}96$.

#11 (Total 30 Points)

In a manufacturing process the assembly line speed (feet per minute) was thought to affect the number of defective parts found during the inspection process. To test this theory, managers devised a situation in which the same batch of parts was inspected visually at a variety of line speeds. They collected the following data:

Line Speed (ft. per min.)	No. of defective parts found
20	21
20	19
40	15
30	16
60	14
40	17

(1) Develop an adequate chart. What does the chart indicate about the relationship between line speed and the number of defective parts found? Formulate the theoretical regression model and determine the model that should be estimated!
(2) Use the data to develop an estimated regression equation that could be used to predict the number of defective parts found, given the line speed. What is the estimated regression model?
(3) Compute the coefficient of determination. What is the difference between the coefficient of determination and the correlation coefficient?

#12 (Total 22 Points)

The production activities of PharmResearch Inc. are highly sensible to interruptions. The Head of Operations is concerned about possible cyber attacks and asks Jimmy Chow, the Cyber Data Analyst, to develop a regression model with which the costs of production discontinuities could be measured. The following data table displays the cost of an interruption in thousand € (y) and the corresponding time (x) in minutes.

y	x
28	2,3
29	2,4
26	2,2
30	2,5
31	2,6
27	2,3
29	2,4
30	2,4

Jimmy is very experienced in data analysis and starts directly with an advanced mathematical model: $y = \beta_0 * x^{\beta_1} * e^{\varepsilon}$. Linearize the model and estimate the corresponding regression equation. Do not forget to formulate it! Consider what it means, when you transform the equation!

#13 (Total 24 Points)

The Smart Data Company experienced several attacks (total number) during the last eight months. These data are recorded in the following data table.

Month	January	February	March	April	May	June	July	August
Attacks	242	265	283	312	312	340	335	323

Perform a Durbin-Watson Test to evaluate whether positive or negative autocorrelation exists. The corresponding values of the Durbin-Watson Statistic are $d_{L,(0,05)} = 0{,}763$ and $d_{U,(0,05)} = 1{,}332$. The estimation of the time series regression equation is given with $\hat{y}_t = 243{,}21 + 12{,}95 * t$.

Evaluate using a histogram and the given standardized residuals whether the data are (approximately) normally distributed.

Observation #	1	2	3	4	5	6	7	8
Stand. Residuals	-0,98	-0,29	0,06	1,18	0,28	1,32	0,08	-1,65

Choose the bins according to the form $Lower\ Level \leq Stand.Residual < Upper\ Level$. Consequently, the bin includes the lower, but not the upper level value.

#14 (Total 23 Points)

The risk management department of the Telco Company is concerned about the increasing customer complaints due to delays in its call center. Consequently, and after a discussion with top management, a risk analyst should assess the operational characteristics of the current operations. In normal operation, an average of 2,5 customers arrive each hour. One call agent is available per pre-defined region to answer customer questions and make product recommendations. The agent averages 10 minutes with each customer.

[1] Compute the operating characteristics of the customer waiting line, assuming Poisson arrivals and exponential service times.
[2] Service goals dictate that an arriving customer should not wait for service more than an average of 5 minutes. Is this goal being met? Transform the value in hours into a value in minutes. If not, what action do you recommend?
[3] If the consultant can reduce the average time spent per customer to 8 minutes, what is the mean service rate? Will the service goal be met?

#15 (Total 10 Points)

The following additive Holt-Winters time series model is given. Complete the model by calculating the missing values (1) up to (10). Note: L = 4.

	n	alpha	gamma	delta		SSE	ssquare	s			Coefficients
	16	0,2	0,1	0,1		20,0810	*(9)*	*(10)*		Intercept	19,8500
										t	0,7450

Year	Quarter	t	Thousand y(t)	Level	Growth Rate	Seasonal Factor	Forecast	FC Error	Sq(FCE)	Regression	Detrended	Average
		-3				-11,5650						
		-2				9,4400						
		-1				21,6950						
		0		19,8500	0,7450	-7,5500						
1	1	1	10	**(6)**	**(7)**	**(8)**	**(3)**	**(4)**	**(5)**	***(1)***	**(2)**	-11,5650
	2	2	31	21,5547	0,7645	9,4405	30,9934	0,0066	0,0000	21,3400	9,6600	9,4400
	3	3	43	22,1164	0,7442	21,6139	44,0143	-1,0143	1,0287	22,0850	20,9150	21,6950
	4	4	16	22,9985	0,7580	-7,4949	15,3106	0,6894	0,4752	22,8300	-6,8300	-7,5500
											-12,5750	
											8,6800	
											19,9350	
						(...)					-8,8100	
											-12,5550	
											8,7000	
											21,9550	
											-7,7900	
4	1	13	19	29,2417	0,7479	-11,4660	17,2995	1,7005	2,8918	29,5350	-10,5350	
	2	14	41	30,3108	0,7800	9,5326	39,3936	1,6064	2,5804	30,2800	10,7200	
	3	15	55	31,5507	0,8260	21,7941	52,7010	2,2990	5,2855	31,0250	23,9750	
	4	16	25	32,4046	0,8288	-7,5054	24,8601	0,1399	0,0196	31,7700	-6,7700	
									20,0810			