Mert Dogan
Z1910844

Assignment 4
10/09/2023

1. The editor of a major academic book publisher claims that a large part of the cost of books is the cost of paper. This implies that larger books will cost more money. As an experiment to analyze the claim, a university student visits the bookstore and records the number of pages and the selling price of 85 randomly selected books and finds a sample correlation of 0.32. At a significance of 5%, conduct a test for the population correlation.

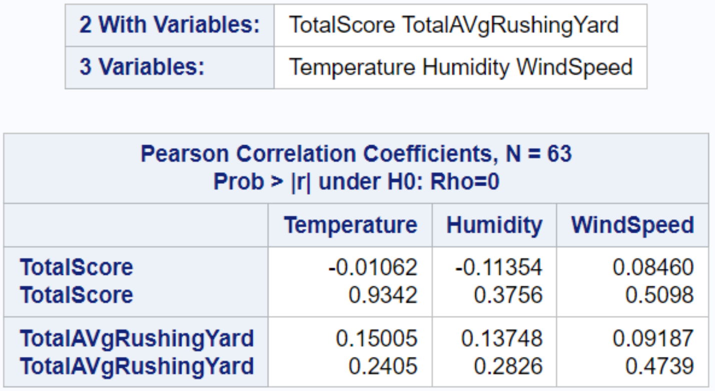
$H_0 : p = 0$
$H_1 : p \neq 0$

$$T = 0.32 * \sqrt{\frac{(85-2)}{1-(0.32)^2}} = 3.077$$

Since p-val < a (.05), we reject NULL. Thus, there is enough evidence to suggest that the number of pages and the selling price are related.

Two tailed P-value (df = 85 – 2 = 83) = .0028

2. Use the NFLData file for the following. In this question, we would like to explore 3 weather-related variables (temperature, wind speed, and humidity) for their correlation with total scores in a game and rushing yards.



2 With Variables:	TotalScore TotalAVgRushingYard
3 Variables:	Temperature Humidity WindSpeed

Pearson Correlation Coefficients, N = 63
Prob > \|r\| under H0: Rho=0

	Temperature	Humidity	WindSpeed
TotalScore **TotalScore**	-0.01062 0.9342	-0.11354 0.3756	0.08460 0.5098
TotalAVgRushingYard **TotalAVgRushingYard**	0.15005 0.2405	0.13748 0.2826	0.09187 0.4739

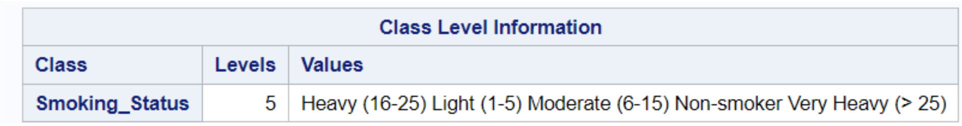
a. We would like to explore the correlation between the total score in a game and temperature, wind speed, and humidity. Conduct the three correlation tests and state your findings.

SAS provides us both the correlation coefficient and p-values (result of conducting a hypothesis test for the population correlation). For all three weather-related variables, we can state the hypothesis as:
$H_0 : p = 0$
$H_1 : p \neq 0$
We can observe that -0.01062 refers to the correlation(negative) between Temperature and TotalScore, -0.11354 refers to the correlation(negative) between Humidity and TotalScore, and 0.08460 refers to the correlation between WindSpeed and TotalScore. We can also observe that p-value > alpha (.05) for each case. Therefore, we fail to reject NULL for each variable and conclude that there is NOT a significant correlation at .05 level between the 3 weather-related variables (temperature, wind speed, and humidity) and total scores in a game out in the population.

b. Similarly, explore if total rushing yards is correlated with temperature, wind speed, and humidity.

Similarly, we can observe the r, and p-value for all three weather-related variables and state hypothesis as:
H_0 : p = 0
H_1: p ≠ 0
Then, we can observe that 0.15005 refers to the correlation between Temperature and TotalAVgRushingYard, 0.13748 refers to the correlation between Humidity and TotalAVgRushingYard, and 0.09187 refers to the correlation between WindSpeed and TotalAVgRushingYard. We can also observe that p-value > alpha (.05) for each case. Therefore, we fail to reject NULL for each variable and conclude that there is NOT a significant correlation at .05 level between the 3 weather-related variables (temperature, wind speed, and humidity) and total rushing yards in a game.

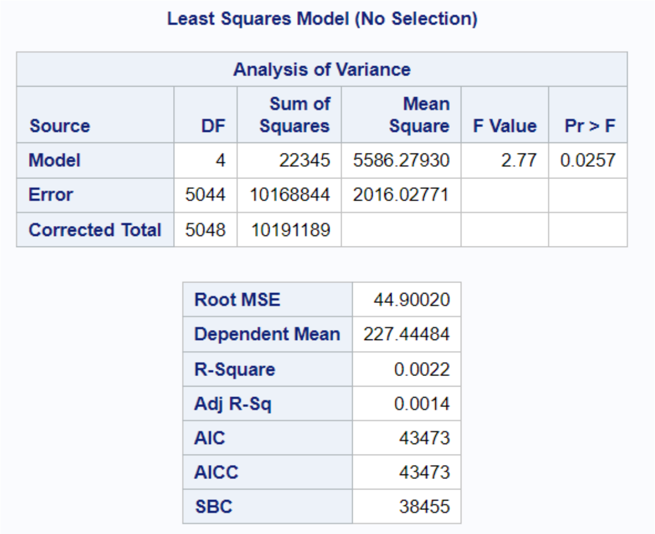
3. Using the Heart dataset in SASHELP directory, conduct a simple linear regression to examine if Smoking_Status (independent variable) impacts Cholesterol (dependent variable).



Class Level Information		
Class	Levels	Values
Smoking_Status	5	Heavy (16-25) Light (1-5) Moderate (6-15) Non-smoker Very Heavy (> 25)

Smoking_Status is a categorical variable with 5 variables, which means we will have 4 indicator variables, therefore 4 coefficients. We will interpret coefficients by comparing them to the comparison group

.



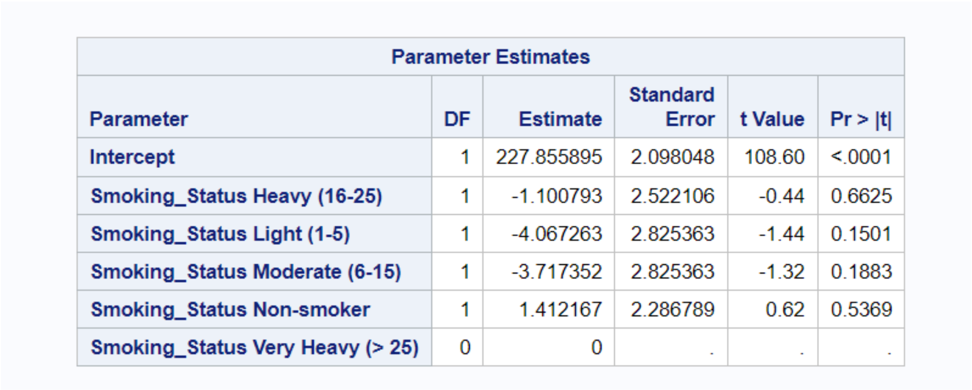
Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	22345	5586.27930	2.77	0.0257
Error	5044	10168844	2016.02771		
Corrected Total	5048	10191189			

Root MSE	44.90020
Dependent Mean	227.44484
R-Square	0.0022
Adj R-Sq	0.0014
AIC	43473
AICC	43473
SBC	38455

The overall model p-value (0.0257) of the F statistics being < alpha (.05) we can conclude that we have a significant model.

R-square is 0.0022, which means the smoking status explains 0.22% of the variance in Cholesterol.



Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	227.855895	2.098048	108.60	<.0001
Smoking_Status Heavy (16-25)	1	-1.100793	2.522106	-0.44	0.6625
Smoking_Status Light (1-5)	1	-4.067263	2.825363	-1.44	0.1501
Smoking_Status Moderate (6-15)	1	-3.717352	2.825363	-1.32	0.1883
Smoking_Status Non-smoker	1	1.412167	2.286789	0.62	0.5369
Smoking_Status Very Heavy (> 25)	0	0	.	.	.

We will interpret the coefficient as the comparison between the comparison group (very heavy smokers) and the other category of interests. We have an intercept of 227.855895, which is significant with an associated p-value (<.0001) being less than alpha. We can see that on average, heavy smokers have 1.100793 less, light smokers have 4.0677263 less, moderate smokers have 3.717352 less and non-smokers have 1.412167 greater cholesterol level than very heavy smokers. However, these numbers are not statistically significant due to their associated p-values being greater than nominal alpha. We can conclude that none of the groups are significantly different from the 'very heavy' smokers group.

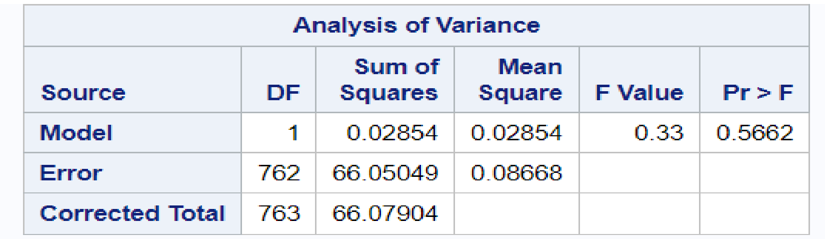
To conclude: The overall model is significant but the small coefficient of determination (0.22%) indicates that smoking status only explains .22% of the variance in Cholesterol. By looking at the parameter estimates table, we can see that the intercept is significant, which makes the overall model significant. However, it can be concluded that none of the coefficients are significant, which proves that smoking status does not have a significant impact on cholesterol. In other words, there is not a significant difference in cholesterol levels between non-smokers, moderate smokers, light smokers, heavy smokers, and very heavy smokers, therefore smoking status does not impact cholesterol.

4. For this question, use the Franchises data file. The file has data on several variables explained below. We would like to understand if the amount of competition in the same county (officesincounty) predicts financial growth in 2011. Run the required analysis and articulate the findings.

Variable	**offid_n**	**oyrsact**	**grth2011**	**grth2010**	**distancetoHQ**	**officesincounty**	**Ownmgrexp**
Explanation	Franchise ID number	How many years has the franchise been a part of the network	Financial growth in revenue from 2010 to 2011	Financial growth in revenue from 2009 to 2010	The distance from the franchise location to the headquarters	The number of other offices located in the same county (competition)	The total experience of the franchise's leadership

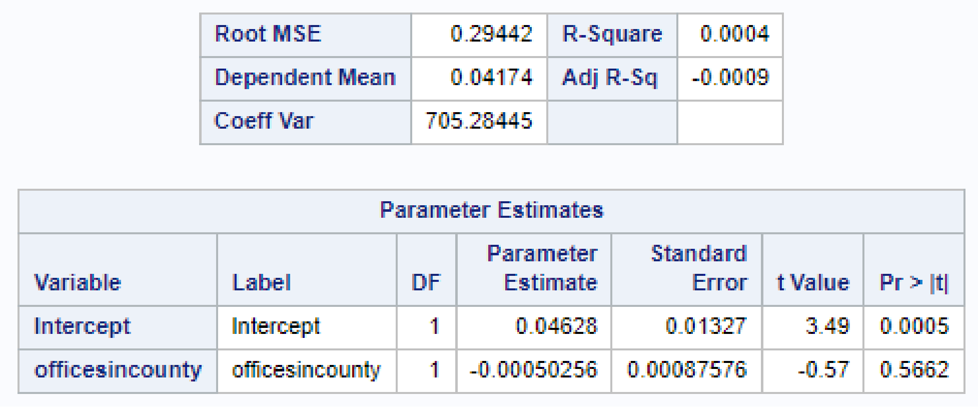
Hypothesis for intercept:
$H_0: \beta_0 = 0$
$H_1: \beta_0 \neq 0$

Hypothesis for slope:
$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.02854	0.02854	0.33	0.5662
Error	762	66.05049	0.08668		
Corrected Total	763	66.07904			

Looking at the analysis of the variance table, we can judge the fit of the overall model. F being very small, and the p-value associated with the f statistic being greater than alpha we fail to reject null and conclude the model is not useful. The independent variables in the model do not collectively have a significant relationship with the dependent variable.



Root MSE	0.29442	R-Square	0.0004
Dependent Mean	0.04174	Adj R-Sq	-0.0009
Coeff Var	705.28445		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.04628	0.01327	3.49	0.0005
officesincounty	officesincounty	1	-0.00050256	0.00087576	-0.57	0.5662

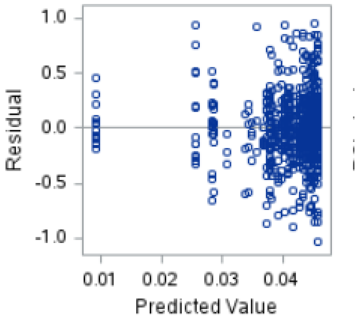
The R-square value of 0.0004 means .04% of the variance in financial growth is being explained by the number of other offices located in the same county.

Looking at the parameter estimates table, observe that the intercept is 0.04628. We can conclude that when there aren't other offices in the same county, the financial growth in revenue is 0.04628. The slope is -0.00050256, a negative small number. We do not judge the influence of officesincounty by how small the parameter estimate is, instead, we look at the t-value and associated p-value to judge if officesincounty influences financial growth or not. P-value being 0.5662 > a, we can conclude here that there is not an influence on the number of other offices in the county on financial growth in the revenue from 2010 to 2011.

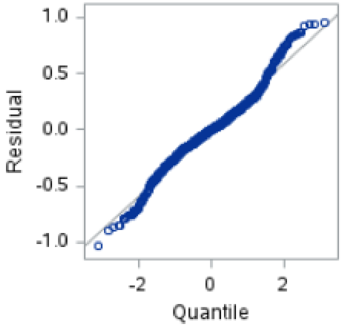
The regression line would look like:

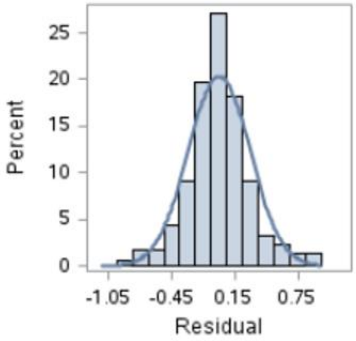
Financial growth = 0.04628 - 0.00050256 *officesincounty

Looking at the residuals vs predicted value graph, we can observe the significant megaphone effect.



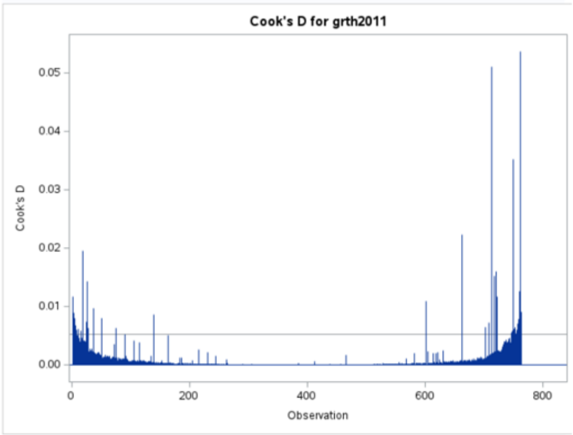
By looking at both residual vs percent and QQ plot, we can conclude our residuals are fairly normal.

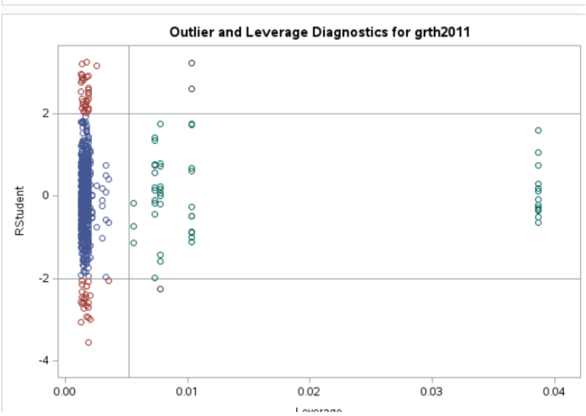




Autocorrelation is not an issue since this is not a time series data.

Furthermore, we can observe multiple outliers, leverage points, and influential observations with Rstudent and Cook 's D graphs that needs to be taken care of.


Cook's D for grth2011
Cook's D
Observation


Outlier and Leverage Diagnostics for grth2011
RStudent