Metadata/ Data Dictionary

Variable	**Description**
Datetime	The date recorded
Season	Seasonality (1:winter, 2:spring, 3:summer, 4:fall)
Holiday	If the day is a holiday or not (1: Holiday 0: Not holiday)
Workingday	If day is neither weekend or holiday (1: Neither weekend or holiday 0: Otherwise)
Weather	1: Clear 2: Misty 3: Light Snow /Light Rain 4: Heavy Rain / Snow / Fog
Temp	Normalized temperature in Celsius
Atemp	Normalized feeling temperature in Celsius.
Humidity	Normalized humidity (max 100)
Windspeed	Normalized windspeed (max 67)
Casual	Bike count of casual riders
Registered	Bike count of registered riders
Count	Total Bike Count (Casual + Registered)

a. Find a multiple regression model for the data.

A multiple regression model can be derived off the following equation:

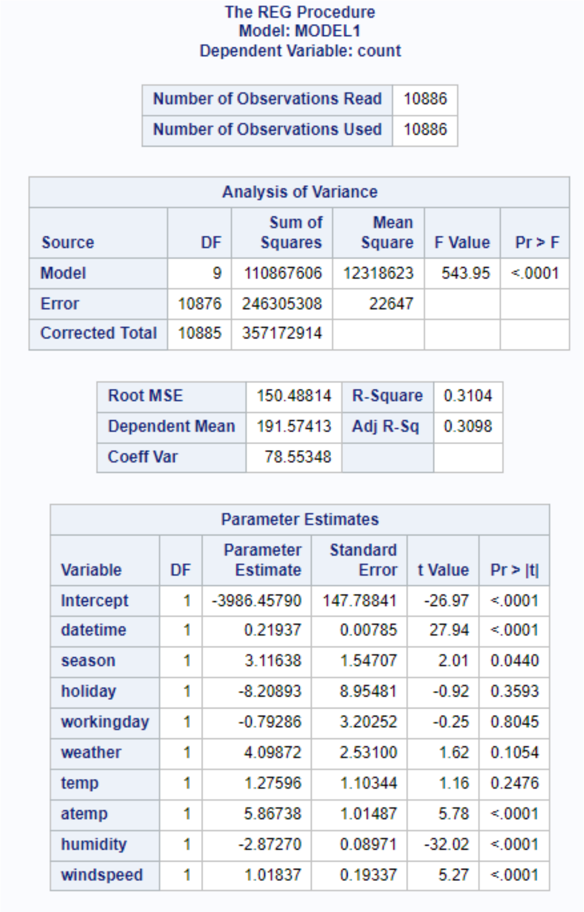
Count = β_0 + $\beta_1$datetime + β_2Season + β_3Holiday + β_4Workingday + β_5Weather + β_6Temp + β_7ATemp + β_8Humidity + β_9Windspeed

All variables are independent except for Casual, Registered and Count. As count is a sum of the casual and registered variable, they are thus a subset of the dependent variable and won't be used in the multiple regression model.

SAS Code:

```
proc reg data=work.bikes;

model count = datetime season holiday workingday weather temp atemp humidity
windspeed;

run;
```

Output:



The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	10886
Number of Observations Used	10886

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	110867606	12318623	543.95	<.0001
Error	10876	246305308	22647		
Corrected Total	10885	357172914			

Root MSE	150.48814	R-Square	0.3104
Dependent Mean	191.57413	Adj R-Sq	0.3098
Coeff Var	78.55348		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3986.45790	147.78841	-26.97	<.0001
datetime	1	0.21937	0.00785	27.94	<.0001
season	1	3.11638	1.54707	2.01	0.0440
holiday	1	-8.20893	8.95481	-0.92	0.3593
workingday	1	-0.79286	3.20252	-0.25	0.8045
weather	1	4.09872	2.53100	1.62	0.1054
temp	1	1.27596	1.10344	1.16	0.2476
atemp	1	5.86738	1.01487	5.78	<.0001
humidity	1	-2.87270	0.08971	-32.02	<.0001
windspeed	1	1.01837	0.19337	5.27	<.0001

b. Interpret the values of the coefficients in the model.

Interpreting the output of the SAS regression model:

The regression coefficient for datetime is 0.21937, there is a slightly positive relationship between datetime and count of bikes. For each increase in one unit of datetime, there is a 0.21937 increase in the count of bikes. This suggests that the bicycle service gains a small amount of popularity over the year.

The regression coefficient for season is 3.11638. There is a positive relationship between the season and count of bikes. For each increase in one unit of season, there is a 3.11638 increase in the count of bikes. As the season starts from winter and moves to spring, summer and fall, it can be interpreted that riders prefer the latter seasons and may not prefer winter due to the cold or snow.

The regression coefficient for holiday is -8.20893. There is a strong negative relationship between holiday and count of bikesFor each increase in one unit of holiday, there is a -8.20893 increase in the count of bikes. This suggests that people do not ride bikes during the holiday and may be with friends or family.

The regression coefficient for workingday is -0.79286. There is a slight negative relationship between workingday and count of bikes. For each increase in one unit of workingday, there is a -0.79286 increase in the count of bikes. This suggests that there may be slightly more riders on a weekend or holiday, compared to a normal working day.

The regression coefficient for weather is 4.09872. There is a positive relationship between the weather and count of bikes. For each increase in one unit of weather, there is a 4.09872 increase in the count of bikes. As the weather moves from a scale of 1 to 4, with 1 being clear and 4 being heavy rain, it can be interpreted as riders preferring to ride bikes in more difficult environments. This is possibly due to riders preferring to walk or jog in clear weather.

The regression coefficient for temp is 1.27596. There is a positive relationship between the temperature and count of bikes. For each increase in one unit of temperature, there is a 1.27596 increase in the count of bikes. This suggests riders prefer higher temperatures when riding bikes as they are warmer.

The regression coefficient for atemp is 5.86738. There is a strong positive relationship between the feeling temperature and count of bikes. For each increase in one unit of feeling temperature, there is a 5.86738 increase in the count of bikes. This similarly suggests that riders prefer a higher feeling temperature when riding bikes, as they feel warmer.

The regression coefficient for humidity is -2.87270. There is a negative relationship between humidity and count of bikes. For each increase in one unit of humidity, there is a -2.87270 increase in the count of bikes. It suggests that riders prefer less humid environments when biking due to possible slipperiness or fog.

The regression coefficient for windspeed is 1.01837. There is a positive relationship between the windspeed and count of bikes. For each increase in one unit of windspeed, there is a 1.01837 increase in the count of bikes. It suggests that riders prefer slightly higher windspeeds, possibly as they feel more comfort riding with the breeze.

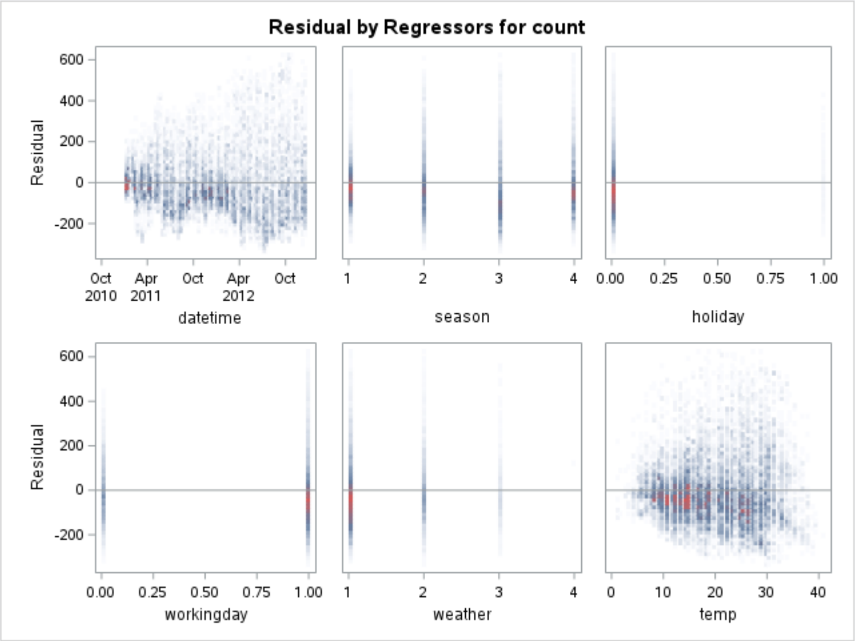
c. Test whether the model as a whole is significant. At the 0.05 level of significance, what is your conclusion?

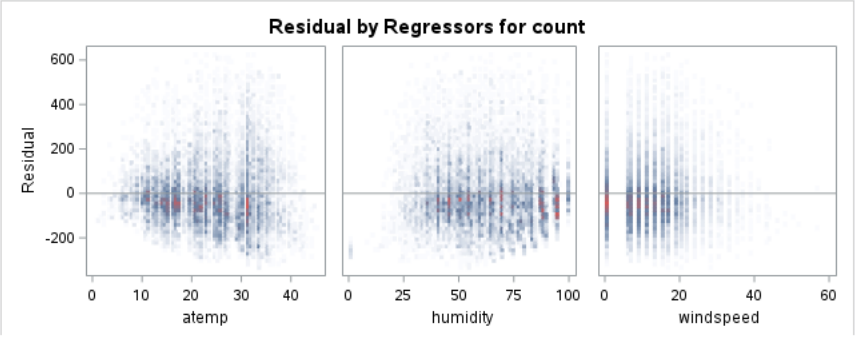
H0: All predictor variables are not significant in predicting total bikes count

H1: At least one of the predictor variables are significant in predicting total bikes count

We can see that the Pr>F value is <0.0001, which is lower than the significance level of 0.05. Thus we reject the null hypothesis and determine that the predictor variables are significant in predicting total count of bikes as a whole. This suggests that at least one of the predictor variables is significant in predicting the total bikes count as a whole.

d. Plot the residuals versus the actual values. Do you think that the model does a good job of predicting number of bikes? Why or why not?





Looking at the residual values for the numerical variables, we can see that some of them have a shape, for example, we can see a cone shape in windspeed, where the variability of windspeed decreases as the windspeed increases. This is similar for datetime and atemp, as the variability increases as the variable increases. As there should not be a shape or pattern in the residual plot, it can be determined that some of the variables do not do a good job at predicting the number of bikes.

Moreover, due to the selection of categorical variables, it may have negatively affected the ability of the model to predict the number of bikes, as the residual values are only distributed in a straight line.

e. Find and interpret the value of R2 for this model.

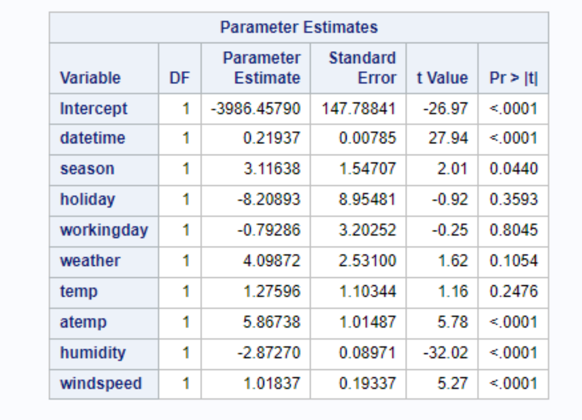
The value of R-squared for this model is 0.3104. This is a poor result for the R-squared coefficient, thus it can be interpreted that the model is not good at predicting the count of bikes variable, as it only explains 31% of the variation in the count of bikes variable. A R-squared coefficient above 80% is typically desired as it is the industry standard.

f. Do you think that this model will be useful in helping the planners? Why or why not?

I think the model will not be useful in helping the planners, as the R-squared coefficient is not above 80%. Rather, the model is at 31%, which is largely below the required level, and can only explain 31% of the variation in the count of bikes variable.

This means that the model is not very reliable in predicting the count of bikes on any given day, and thus planners that use this information may make poor decisions due to an inaccurate bike count.

g. Test the individual regression coefficients. At the 0.05 level of significance, what are your conclusions?



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3986.45790	147.78841	-26.97	<.0001
datetime	1	0.21937	0.00785	27.94	<.0001
season	1	3.11638	1.54707	2.01	0.0440
holiday	1	-8.20893	8.95481	-0.92	0.3593
workingday	1	-0.79286	3.20252	-0.25	0.8045
weather	1	4.09872	2.53100	1.62	0.1054
temp	1	1.27596	1.10344	1.16	0.2476
atemp	1	5.86738	1.01487	5.78	<.0001
humidity	1	-2.87270	0.08971	-32.02	<.0001
windspeed	1	1.01837	0.19337	5.27	<.0001

H0: The individual predictor variable has no effect on the count of bikes

H1: The individual predictor variable affects the count of bikes

Looking at the individual regression coefficients Pr>|t|, at the significance level 0.05, we can see that **the regression coefficients of datetime, season, atemp, humidity, and windspeed are statistically significant and below the 0.05 significance level**. This means we reject the null hypothesis and determine that the individual predictor variable affects the count of bikes.

While **the variables holiday, workingday, weather and temperature have a p-value above 0.05 and are not statistically significant**, thus we fail to reject the null hypothesis and determine that the individual predictor variable has no effect on the count of bikes.

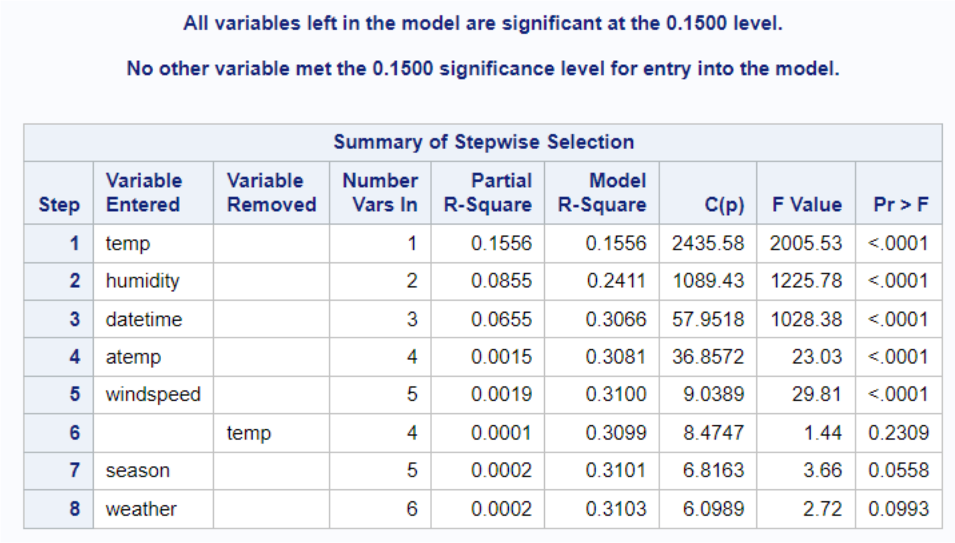
h. If you were going to drop just one variable from the model, which one would you choose? Why?

I would choose to drop workingday as the Pr> |t| variable is 0.8045. A p-value above 0.05 indicates that the variable cannot be concluded to affect the dependent variable. This makes it important to discard variables with a p-value above 0.05. Furthermore, workingday has the highest p-value, and this means that there is a high probability that the results have occurred due to chance, and doesn't have much effect on the count of bikes on a specific day, making it a poor predictor of the dependent variable.

i. Use stepwise regression to find the best model for the data.

SAS Code:

```
proc reg data=work.bikes;

model count = datetime season holiday workingday weather temp atemp humidity
windspeed / selection = stepwise;

run;
```



All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	temp		1	0.1556	0.1556	2435.58	2005.53	<.0001
2	humidity		2	0.0855	0.2411	1089.43	1225.78	<.0001
3	datetime		3	0.0655	0.3066	57.9518	1028.38	<.0001
4	atemp		4	0.0015	0.3081	36.8572	23.03	<.0001
5	windspeed		5	0.0019	0.3100	9.0389	29.81	<.0001
6		temp	4	0.0001	0.3099	8.4747	1.44	0.2309
7	season		5	0.0002	0.3101	6.8163	3.66	0.0558
8	weather		6	0.0002	0.3103	6.0989	2.72	0.0993

Using the stepwise regression in SAS, the variable temp is removed at Step 5, as the p-value of 0.2309 exceeded the 0.15 level as more variables were added into the model. This led to temp being removed from the regression model.

Ultimately, the model for the data is calculated using the following equation:

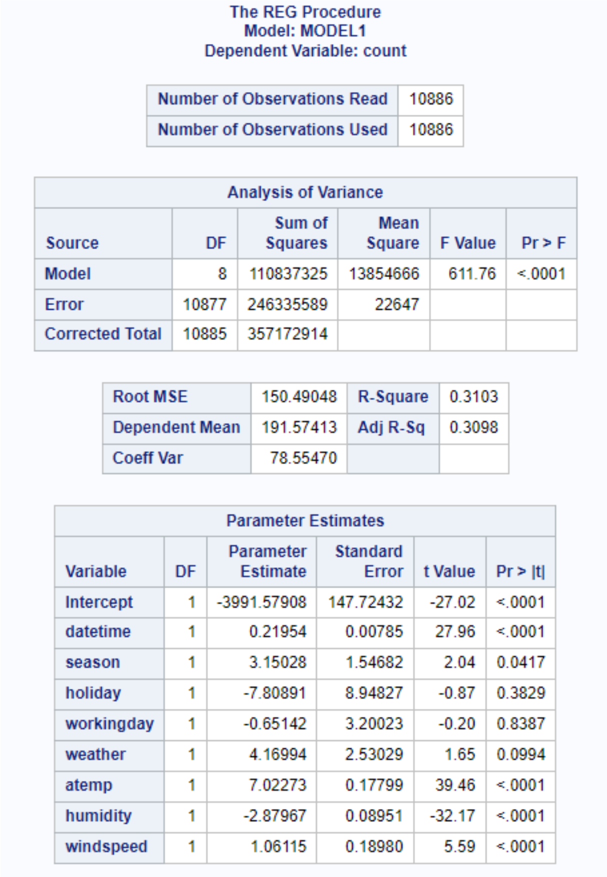
Count = β_0 + $\beta_1$datetime + β_2Season + β_3Holiday + β_4Workingday + β_5Weather + β_7ATemp + β_8Humidity + β_9Windspeed

Temp is removed. The beta is to be substituted by the individual parameter estimates mentioned in part (g) above.

j. Analyze the model you have identified to determine whether it has any problems.

SAS Code:

```
proc reg data=work.bikes;

model count = datetime season holiday workingday weather atemp humidity windspeed;

run;
```



The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	10886
Number of Observations Used	10886

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	110837325	13854666	611.76	<.0001
Error	10877	246335589	22647		
Corrected Total	10885	357172914			

Root MSE	150.49048	R-Square	0.3103
Dependent Mean	191.57413	Adj R-Sq	0.3098
Coeff Var	78.55470		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3991.57908	147.72432	-27.02	<.0001
datetime	1	0.21954	0.00785	27.96	<.0001
season	1	3.15028	1.54682	2.04	0.0417
holiday	1	-7.80891	8.94827	-0.87	0.3829
workingday	1	-0.65142	3.20023	-0.20	0.8387
weather	1	4.16994	2.53029	1.65	0.0994
atemp	1	7.02273	0.17799	39.46	<.0001
humidity	1	-2.87967	0.08951	-32.17	<.0001
windspeed	1	1.06115	0.18980	5.59	<.0001

After removing the variable Temp, we can see the model overall is relevant at the 0.05 level, as the Pr>F value is <0.0001, with at least one predictor variable having an effect on count of bikes.

However, the R-Squared value is 0.3103, which is a decrease of 0.0001 without Temp removed in the previous model (0.3104). As the R-Squared value should be above 0.80 which is typically desired by industry standards, and the R-Squared value is unchanged, it can be deduced that the model still has a problem in predicting values. To improve the model, more relevant variables must be selected and irrelevant variables must be discarded.

k. Write a memo reporting your findings to your boss. Identify the strengths and weaknesses of the model you have chosen

To: Boss

From: Keith

Date: 6/12/2023

Subject: Strengths and Weakness of Multiple Linear Regression Model

It has been found that the linear regression model was good at determining the relationships of the individual predictor variables, and whether variables were positively or negatively correlated with the count of bikes. This can allow us to understand how each of the variables affects the count of bikes on any given day, such as how days that are more humid are negatively correlated with the count of bike riders; these could be found in the regression coefficients for the individual predictor variables. Moreover, it could also be found that some of the individual predictor variables were not statistically significant and may not predict the count of bikes well, allowing for next steps to be conducted for the regression model, such that we can remove irrelevant variables.

However, the weakness of the model lies in its R-squared value of 0.3104, which is below the industry standard of 0.80 for a predictor model. This means that the model cannot accurately predict the count of bikes during this period. This is likely due to all variables being selected, and thus there is a need to remove certain variables to improve the R-squared value. Moreover, stepwise regression was not suitable in removing variables, as the R-squared value did not improve after it. This may add difficulty in choosing which variables to include or not include for the model to predict the count of bikes accurately.

Problem 2:

Metadata/ Data Dictionary

Variable	Description
Survived	If the passenger survived in the Titanic 0 = Not survived 1 = Survived
Age	Age in years

For this logistic regression analysis we will only be using two variables, age and survived.

a. Write the logistic regression equation relating Age and Survived.

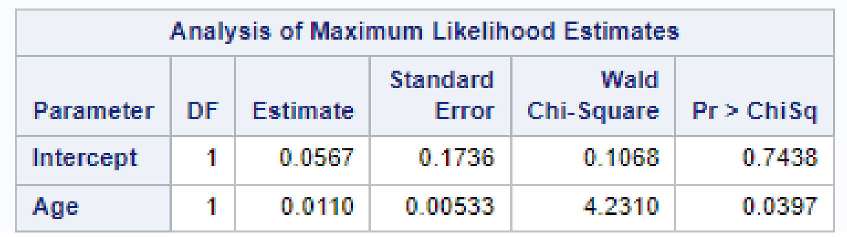
Equation predicting the probability of a passenger not surviving, based on their age:

Y = estimate of $p(y=0 \mid x_1, x_2, x_3 \ldots) = e^{\beta 0 + \beta 1 \text{ Age}} / (1 + e^{\beta 0 + \beta 1 \text{ Age}})$

b. For the Titanic data, use SAS to compute the estimated logistic regression equation.

SAS Code:

```
proc import datafile='/home/u63571133/Stats/Assignment 4/titanic.csv'

dbms=csv

out=titanic;

run;


Proc logistic data=work.titanic;model survived = age; run;
```



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0567	0.1736	0.1068	0.7438
Age	1	0.0110	0.00533	4.2310	0.0397

The intercept $\beta 0$ is 0.0567, while the regression coefficient for age $\beta 1$ is 0.0110.

Therefore, the probability of a passenger not surviving Y can be calculated with the logistic regression equation:

Y = estimate of $p(y=0 \mid x_1, x_2, x_3 \ldots) =$

$$e^{0.0567 + 0.0110 * \text{Age}} / (1 + e^{0.0567 + 0.0110 * \text{Age}})$$

c. Estimate the probability of surviving the passenger with the average Age 30.

e $^{0.0567\ +\ 0.0110*\ 30}$ / (1 + e $^{0.0567\ +\ 0.0110*30}$) = 0.59548803971

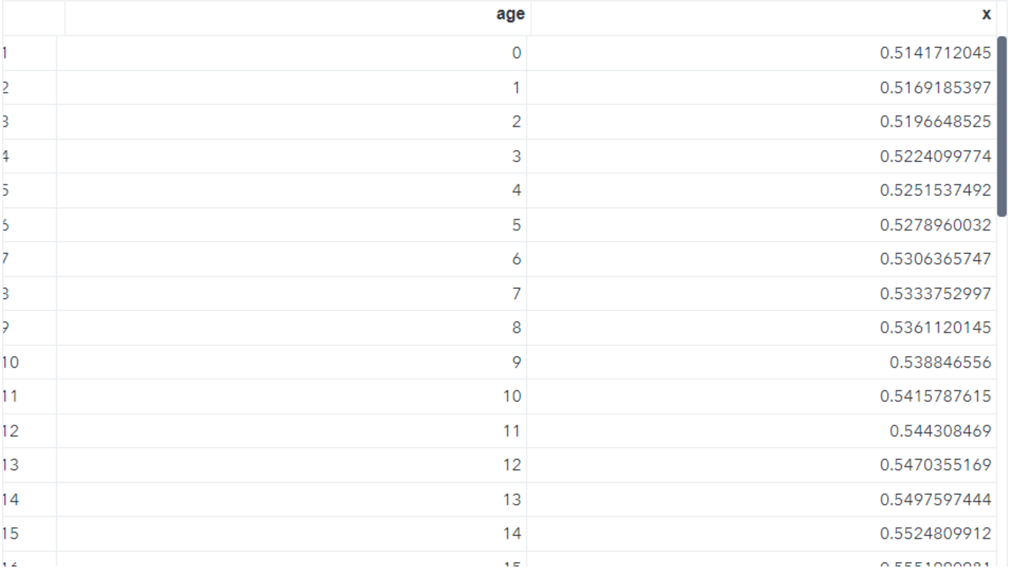
The probability of the passenger not surviving is around 59.5%.

Therefore, the probability of the passenger surviving is 100% -59.5% = 40.5%.

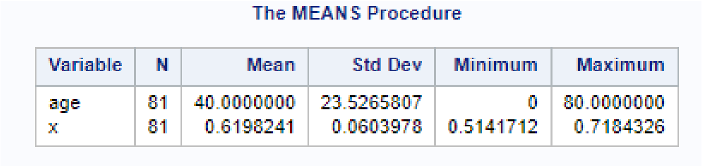
d. Suppose we want to check who have a 0.50 or higher probability of surviving. What is the average age to achieve this level of probability?

SAS Code:

```
data test;

do age = 0 to 80;

x = 2.718282**(0.0567 + 0.0110 * Age)  / (1 + 2.718282**(0.0567 + 0.0110* Age));

output;

end;

run;


proc means; run;
```



	age	x
1	0	0.5141712045
2	1	0.5169185397
3	2	0.5196648525
4	3	0.5224099774
5	4	0.5251537492
6	5	0.5278960032
7	6	0.5306365747
8	7	0.5333752997
9	8	0.5361120145
10	9	0.538846556
11	10	0.5415787615
12	11	0.544308469
13	12	0.5470355169
14	13	0.5497597444
15	14	0.5524809912

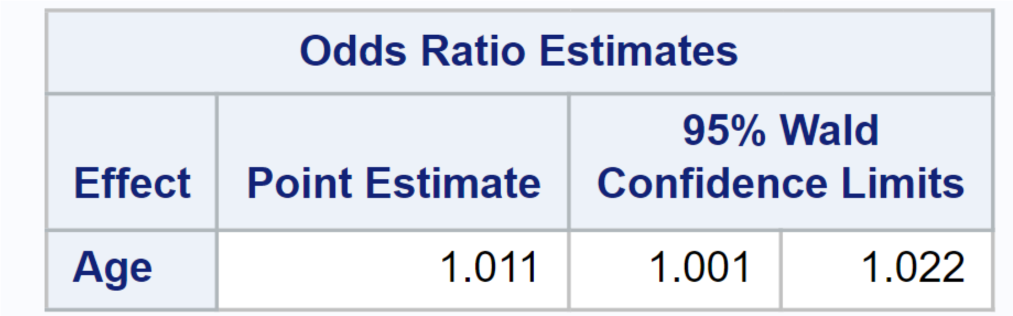


The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age x	81 81	40.0000000 0.6198241	23.5265807 0.0603978	0 0.5141712	80.0000000 0.7184326

As we can see from inputting the logistic regression function for ages 0 to 80, there is a probability of not surviving of at least 51.4%, even for age 0. This means that there is no age that has a probability of surviving above 50%, as 1-0.514 = 0.486, and there is a minimum survival chance of 48.6% even for babies with an age below 1.

e. What is the estimated odds ratio? What is the interpretation?



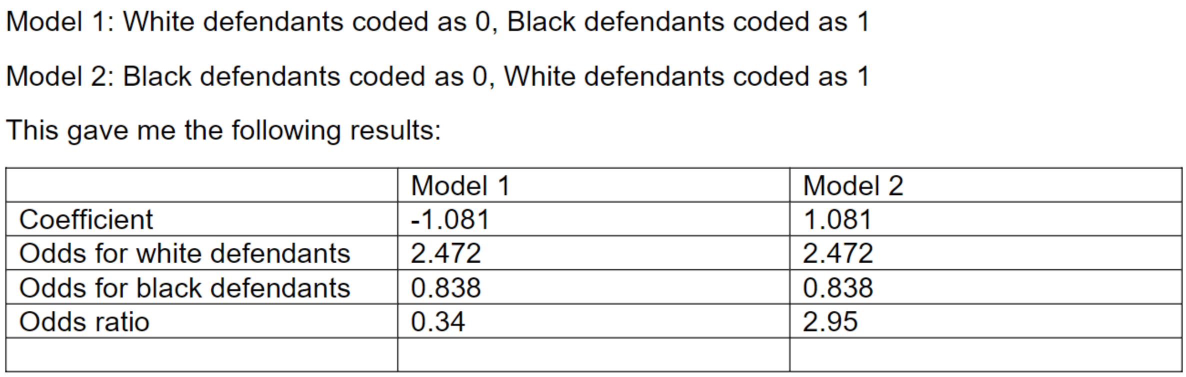
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.011	1.001	1.022

We can conclude that as age increases by one unit, the estimated odds of a person not surviving (dying) are 1.011 times greater. As the 95% Wald Confidence Limits do not contain 1 inside the limits, the result is significant and thus the results do not occur due to chance.

Ultimately, it can be interpreted that older individuals may not be able to survive due to weaker physical health, or perhaps they are unable to get on the lifeboat as lifeboats typically allow younger individuals on first.

Problem 3: Capital punishment



Model 1: White defendants coded as 0, Black defendants coded as 1

Model 2: Black defendants coded as 0, White defendants coded as 1

This gave me the following results:

	Model 1	Model 2
Coefficient	-1.081	1.081
Odds for white defendants	2.472	2.472
Odds for black defendants	0.838	0.838
Odds ratio	0.34	2.95

a. Why the odds ratios are different? Explain it

The odds ratio is different between model 1 and model 2, despite the odds for white and black defendants staying the same.

In Model 1, Black defendants are coded as 1 and White as 0. This means that it is modelling the event outcome for Black defendants. The odds ratio of 0.34 is less than one, therefore the odds of capital punishment are lower for Black defendants. This is supported by the odds for white defendants is 2.472 and the odds for black defendants is 0.838, where 0.838/2.472 is 0.34.

In Model 2, White defendants are coded as 1 and Black as 0. This means that it is modelling the event outcome for White defendants. The odds ratio of 2.95 is more than one, therefore the odds of capital punishment are higher for White defendants. This is supported by the odds for white defendants is 2.472 and the odds for black defendants is 0.838, where 2.472/0.838 is 2.95.

b. Show the relation between the odd ratios and coefficient

The relationship between the odds ratios and coefficient can be seen through an exponential function.

Odds Ratio = exp(β)

$0.34 = \exp(-1.081)$

$2.94 = \exp(1.081)$