The *STUDENTDATA* dataset contains information on marks secured by students in various subjects corresponding to different social attributes. The data contains 548 entries of different students. In particular, the data include the following variables:

- **Gender:** represents the gender of the student. It can take the values "male" and "female".
- **Race/Ethnicity:** This column contains 5 groups that represent the race and ethnicity of the student. For simplicity, we chose the four most frequent groups, Group B, C, D and E, and we changed the name of the column to "raceEthnicity".
- **Parental level of education:** this variable represents the level of education of the parents of each student. Initially, this variable had 6 levels but for simplicity we only considered 4 levels: "master's degree", "bachelor's degree", "associate's degree" and "high school". This variable was renamed to "prntEduLvl"
- **Lunch:** some students pay the complete lunch fees whereas others pay reduced fees for lunch.
- **Test preparation course:** a preparation course was offered for students. "complete" means that the student completed the course, and "none" means the opposite. This variable was then renamed as "prepCourse".
- **Math score:** the score of a student on a math test.
- **Reading score:** the score of a student on a reading test.
- **Writing score:** the score of a student on a writing test.

This dataset is publicly available. For this project the dataset is downloaded from https://www.kaggle.com/spscientist/students-performance-in-exams

Q1) a) Fit a linear regression model for average score using indicator variables gender, Test preparation course and Parental level of education taking male, none, masters degree respectively as reference level. Provide the fitted model and interpret the regression parameters of the model.

Solution –

To calculate the average scores, the formula used is:

Avg score = (math.score + writing.score + reading.score)/3 and rounded to nearest integer. SAS code is for this in appendix. Below table present the estimates for the model:

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	69.57313390	B	1.94312185	35.80	<.0001	65.75616154	73.39010626
gender female	3.59806081	B	1.13412512	3.17	0.0016	1.37024157	5.82588005
gender male	0.00000000	B	.	.	.	.	.
prepCourse completed	6.62852170	B	1.19501957	5.55	<.0001	4.28108441	8.97595899
prepCourse none	0.00000000	B	.	.	.	.	.
prntEduLvl associate's degree	-3.98747457	B	1.99097875	-2.00	0.0457	-7.89845466	-0.07649449
prntEduLvl bachelor's degree	-1.83050711	B	2.18350099	-0.84	0.4022	-6.11966836	2.45865415
prntEduLvl high school	-9.85307128	B	2.03101478	-4.85	<.0001	-13.84269616	-5.86344641
prntEduLvl master's degree	0.00000000	B	.	.	.	.	.

We are interested in fitting a simple linear regression model:

$$Y = \beta_0 + \beta_1 * gender\_female + \beta_2 * prepCourse\_completed + \beta_3 * prntEduLvl\_associate + \beta_4 * prntEduLvl\_bachelors + \beta_5 * prntEduLvl\_high\_school$$

Where Y denotes the average score of students.

From the SAS output, the fitted model is:

$$\widehat{E}(Y|X = x) = 69.5731 + 3.5980 * gender\_female - 6.6285 * prepCourse\_completed - 9.8530 * prntEduLvl\_associate - 3.9874 * prntEduLvl\_bachelors - 1.8305 * prntEduLvl\_high\_school$$

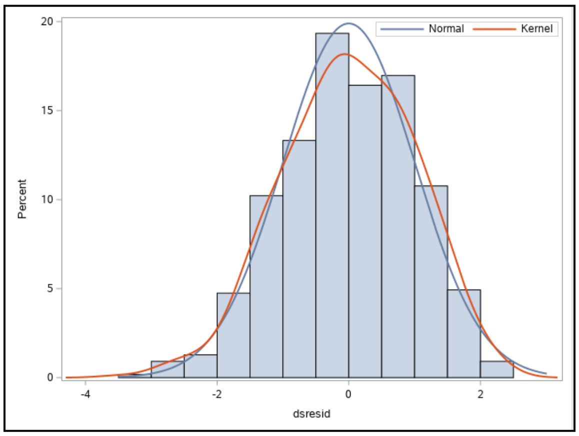
<u>Interpretation of regression parameters:</u>

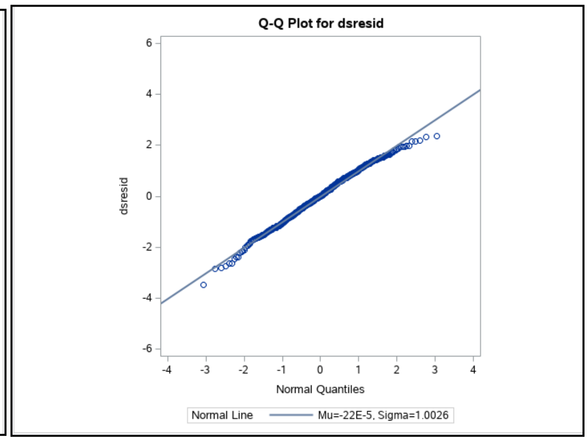
- $\beta_0 = E(Y|gender\_female, prepCourse\_completed, prntEduLvl\_associate, prntEduLvl\_bachelors, prntEduLvl\_high\_school = 0)$ represents the average marks scored by a male student. who have completed the preparation course and whose parents has master's level education, which is estimated as $\widehat{\beta}_0 = 76.2\ marks$
- $\beta_1 =$ $E\left(Y\middle|\begin{matrix} gender\_female = 1, prepCourse\_completed, prntEduLvl\_associate, prntEduLvl\_bachelors, \\ prntEduLvl\_high\_school = 0 \end{matrix}\right) -$ $E(Y|gender\_female, prepCourse\_completed, prntEduLvl\_associate, prntEduLvl\_bachelors, prntEduLvl\_high\_school = 0)$ represents the mean difference in student average marks when gender is male vs female, which is estimated at $\widehat{\beta}_1 = 3.598\ marks$
- $\beta_2 =$ $E\left(Y\middle|\begin{matrix} gender\_female = 0, prepCourse\_completed = 1, prntEduLvl\_associate, prntEduLvl\_bachelors, \\ prntEduLvl\_high\_school = 0 \end{matrix}\right) -$ $E(Y|gender\_female, prepCourse\_completed, prntEduLvl\_associate, prntEduLvl\_bachelors, prntEduLvl\_high\_school = 0)$ represents the mean difference in student's marks when one has completed prep course vs. when one hasn't taken, which is estimated as $\widehat{\beta}_2 = -6.628\ marks$
- $\beta_3 =$ $E\left(Y\middle|\begin{matrix} prntEduLvl\_associate = 1, gender\_female, prepCourse\_completed, prntEduLvl\_highschool, \\ prntEduLvl\_bachelors = 0 \end{matrix}\right) -$ $E(Y|prntEduLvl\_associate, prepCourse\_completed, gender\_female, prntEduLvl\_high\_school, prntEduLvl\_bachelors = 0)$ represents the mean difference in student's marks when their parent's education level is associate vs master's degree, which is estimated as $\widehat{\beta}_3 = -9.853\ marks$
- $\beta_4 =$ $E\left(Y\middle|\begin{matrix} prntEduLvl\_bachelors = 1, gender\_female, prepCourse\_completed, prntEduLvl\_highschool, \\ prntEduLvl\_associate = 0 \end{matrix}\right) -$ $E(Y|prntEduLvl\_associate, prepCourse\_completed, gender\_female, prntEduLvl\_high\_school, prntEduLvl\_bachelors = 0$ represents the mean difference in student's marks when their parent's education level is bachelor degree vs master's degree, which is estimated as $\widehat{\beta}_4 = -3.987\ marks$
- $\beta_5 =$ $E\left(Y\middle|\begin{matrix} prntEduLvl\_highschool = 1, gender\_female, prepCourse\_completed, prntEduLvl\_bachelors, \\ prntEduLvl\_associate = 0 \end{matrix}\right) -$ $E(Y|prntEduLvl\_associate, prepCourse\_completed, gender\_female, prntEduLvl\_high\_school, prntEduLvl\_bachelors = 0$ represents the mean difference in student's marks when their parent's education level is high school vs master's degree, which is estimated as $\widehat{\beta}_5 = -1.83\ marks$

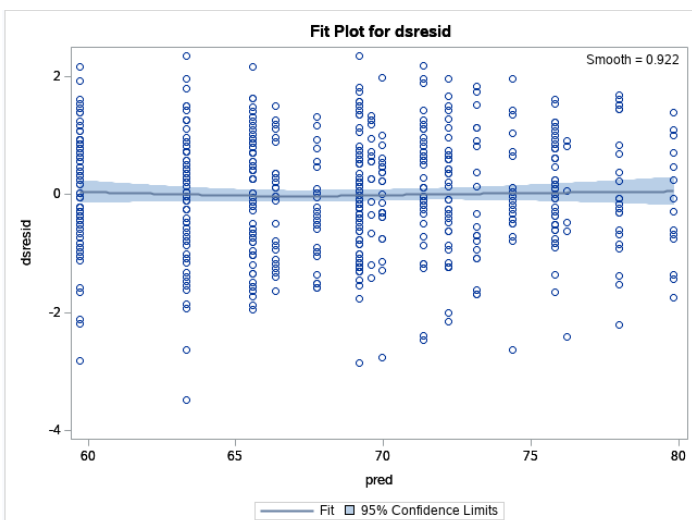
b) Using a residual analysis for the model considered in part a), assess the underlying model assumptions. Comment on the results.

Solution –

SAS output







- Independence: we will assume that the observations are independent
- Normality: the histogram shows a roughly symmetric, bell-shaped curve and for the QQ-plot, most points fall close to the diagonal line. Both plots indicate that the assumption of normality is reasonable here.
- Mean correctly specified: the plots of the residuals approximately make a straight line with dsresid = 0, the assumption that the errors have mean 0 seems reasonable here.

- Constant variance: since the data is categorical, we see vertical line formations. As such there is no clear pattern in variance in the plot, the assumption of constant variance (homoscedasticity) seems reasonable here.

c) Using the same model as above, test difference in reference levels to find the greatest difference in parameter estimates between each parental level of education. Provide the output of the estimates of the largest difference, interpret it and comment.

Solution-

Testing all parental level of education, we have found that the greatest difference between a level of education and its reference level was between a parent holding a **master's degree** compared to a parent holding a **high school degree**. Indeed, the estimated mean difference of the average scores tested at 16 will increased by 9.717 points when the parent has a master's degree compared to a parent with a high school degree, holding other variables constant. Vice versa, using the parent holding a master's degree, we find that the average test scores at 16 will decrease on average by 9.717 point when the parent holds a high school degree, holding other variables constant.

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à95%	
Constante	66.26258997	B	1.76830856	37.47	<.0001	62.78895439	69.73622556
gender female	3.53522591	B	1.07942998	3.28	0.0011	1.41481181	5.65564001
gender male	0.00000000	B	.	.	.	.	.
race group B	-4.96535700	B	1.79848286	-2.76	0.0060	-8.49826649	-1.43244752
race group C	-4.61028954	B	1.61217237	-2.86	0.0044	-7.77721383	-1.44336525
race group D	-2.83023039	B	1.71968717	-1.65	0.1004	-6.20835494	0.54789415
race group E	0.00000000	B	.	.	.	.	.
parent_edu associate's degree	5.38912771	B	1.28956384	4.18	<.0001	2.85593019	7.92232522
parent_edu bachelor's degree	8.06559375	B	1.54998493	5.20	<.0001	5.02082942	11.11035807
parent_edu master's degree	9.71718311	B	1.94534136	5.00	<.0001	5.89578727	13.53857896
parent_edu high school	0.00000000	B	.	.	.	.	.
lunch free/reduced	-8.18724047	B	1.12023690	-7.31	<.0001	-10.38781499	-5.98666595
lunch standard	0.00000000	B	.	.	.	.	.
prep_test completed	6.92686294	B	1.14360971	6.06	<.0001	4.68037526	9.17335062
prep_test none	0.00000000	B	.	.	.	.	.

d) Similar to the question above, which difference in education levels has the smallest impact of average test scores at 16?

Solution-

We found that the smallest estimated difference between education levels in parents is between those holding a master's degree compared to those holding a bachelor's degree. Estimated difference is 1.65 higher scores at 16 on average for parents holding a master's degree compared to those with a bachelor's degree holding other variables constant. Vice and versa, if we take parents holding a master's degree as reference, we find that the estimated average score at 16 decreases on average by 1.65 points for parents holding a bachelor's degree, holding all other variable constant. Although, using an $\alpha = 0{,}05$, this estimate is not significant as it's p-value is 0.4273 > 0.05.

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à95%	
Constante	74.32818372	B	1.95369789	38.04	<.0001	70.49037245	78.16599499
gender female	3.53522591	B	1.07942998	3.28	0.0011	1.41481181	5.65564001
gender male	0.00000000	B	.	.	.	.	.
race group B	-4.96535700	B	1.79848286	-2.76	0.0060	-8.49826649	-1.43244752
race group C	-4.61028954	B	1.61217237	-2.86	0.0044	-7.77721383	-1.44336525
race group D	-2.83023039	B	1.71968717	-1.65	0.1004	-6.20835494	0.54789415
race group E	0.00000000	B	.	.	.	.	.
parent_edu associate's degree	-2.67646604	B	1.49762193	-1.79	0.0745	-5.61836937	0.26543728
parent_edu high school	-8.06559375	B	1.54998493	-5.20	<.0001	-11.11035807	-5.02082942
parent_edu master's degree	1.65158937	B	2.07890874	0.79	0.4273	-2.43218400	5.73536273
parent_edu bachelor's degree	0.00000000	B	.	.	.	.	.
lunch free/reduced	-8.18724047	B	1.12023690	-7.31	<.0001	-10.38781499	-5.98666595
lunch standard	0.00000000	B	.	.	.	.	.
prep_test completed	6.92686294	B	1.14360971	6.06	<.0001	4.68037526	9.17335062
prep_test none	0.00000000	B	.	.	.	.	.

Used for reference

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à95%	
Constante	75.97977309	B	2.30522861	32.96	<.0001	71.45142079	80.50812539
gender female	3.53522591	B	1.07942998	3.28	0.0011	1.41481181	5.65564001
gender male	0.00000000	B	.	.	.	.	.
race group B	-4.96535700	B	1.79848286	-2.76	0.0060	-8.49826649	-1.43244752
race group C	-4.61028954	B	1.61217237	-2.86	0.0044	-7.77721383	-1.44336525
race group D	-2.83023039	B	1.71968717	-1.65	0.1004	-6.20835494	0.54789415
race group E	0.00000000	B	.	.	.	.	.
parent_edu associate's degree	-4.32805541	B	1.90208379	-2.28	0.0233	-8.06447681	-0.59163400
parent_edu bachelor's degree	-1.65158937	B	2.07890874	-0.79	0.4273	-5.73536273	2.43218400
parent_edu high school	-9.71718311	B	1.94534136	-5.00	<.0001	-13.53857896	-5.89578727
parent_edu master's degree	0.00000000	B	.	.	.	.	.
lunch free/reduced	-8.18724047	B	1.12023690	-7.31	<.0001	-10.38781499	-5.98666595
lunch standard	0.00000000	B	.	.	.	.	.
prep_test completed	6.92686294	B	1.14360971	6.06	<.0001	4.68037526	9.17335062
prep_test none	0.00000000	B	.	.	.	.	.

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à95%	
Constante	71.65171768	B	1.70173996	42.10	<.0001	68.30884833	74.99458703
gender female	3.53522591	B	1.07942998	3.28	0.0011	1.41481181	5.65564001
gender male	0.00000000	B	.	.	.	.	.
race group B	-4.96535700	B	1.79848286	-2.76	0.0060	-8.49826649	-1.43244752
race group C	-4.61028954	B	1.61217237	-2.86	0.0044	-7.77721383	-1.44336525
race group D	-2.83023039	B	1.71968717	-1.65	0.1004	-6.20835494	0.54789415
race group E	0.00000000	B	.	.	.	.	.
parent_edu bachelor's degree	2.67646604	B	1.49762193	1.79	0.0745	-0.26543728	5.61836937
parent_edu high school	-5.38912771	B	1.28956384	-4.18	<.0001	-7.92232522	-2.85593019
parent_edu master's degree	4.32805541	B	1.90208379	2.28	0.0233	0.59163400	8.06447681
parent_edu associate's degree	0.00000000	B	.	.	.	.	.
lunch free/reduced	-8.18724047	B	1.12023690	-7.31	<.0001	-10.38781499	-5.98666595
lunch standard	0.00000000	B	.	.	.	.	.
prep_test completed	6.92686294	B	1.14360971	6.06	<.0001	4.68037526	9.17335062
prep_test none	0.00000000	B	.	.	.	.	.

e) Fit a linear regression model for average score using the indicator variables for lunch, with level 0 (standard) as the reference level. Provide the fitted model and compare the difference in the average scores of students within two population groups. Population (A): group of students with “standard” lunch and population (B): group of students with “free/reduced” lunch. Interpret the results.

Solution –

The model has the form:

$$E(Y|lunch0) = \beta_0 + \beta_{lunch0} * lunch0$$

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	71.47008547	B	0.73225367	97.60	<.0001	70.03170620	72.90846474
lunch 0	-7.73404486	B	1.22128978	-6.33	<.0001	-10.13304670	-5.33504302
lunch 1	0.00000000	B	.	.	.	.	.

From the SAS output the fitted model is,

$$\hat{y} = 71.47 - 7.734 lunch0$$

Here we have considered two populations:

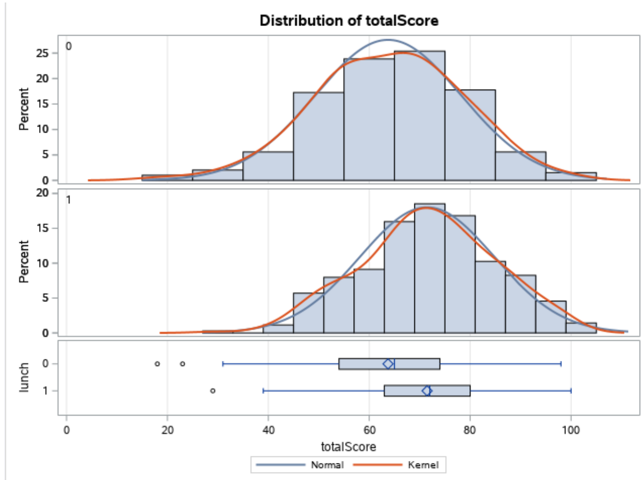
a) Population A) where lunch consumed by students is “standard”
b) Population B) where lunch consumed by students is “free/reduced”

The difference in the average scores of students in population A) & B) is given by β_{lunch0}. Thus, we are interested in testing $H_0 : \beta_{lunch0} = 0 \; vs \; H_1 : \beta_{lunch0} \neq 0$.

SAS Output of t-test:

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	546	-6.33	<.0001
Satterthwaite	Unequal	378.49	-6.19	<.0001

Checking the variance of both populations in the image below, we see that the variance of data-set for both populations is approximately the same. Therefore, we can go for Pooled method.



The corresponding test statistic is -6.33 with a p-value of $p < 0.0001$. Here $p < \alpha$, where α =0.05. We therefore reject the H_0 . We can thus conclude that there is sufficient evidence to suggest that on average, the average score is same for students in population A) & B) at 5% confidence level.

Q2 a) What is the effect of the preparation course on the odds of passing a math test?

Solution -

To answer this question, we fitted a logistic regression on the log-odds scale with “prepCourse” variable as the only explanatory variable and the binary variable “passed” as the predicted variable.

Parameter Estimates						
Effect	prepCourse	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		0.5465	0.1094	546	5.00	<.0001
prepCourse	completed	0.4687	0.1980	546	2.37	0.0183
prepCourse	none	0	.	.	.	.

Figure 1 parameter estimates

The fitted model is:

$$ln(odds) = \hat{\beta}_0 + \hat{\beta}_1 * prepCourse = 0.5465 + 0.4687 * prepCourse$$

from the fitted model above, we can say that on average, and a 5% level of significance, the odds of passing the math test for people who have completed the preparation course are $e^{0.4687} = 1.598$ times the odds of passing the exam for people who have not completed the preparation course.

b) Parameter interpretation of the fitted logistic model on the log-odds scale

Solution –

Parameter Estimates										
Effect	gender	raceEthnicity	prntEduLvl	lunch	prepCourse	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept						2.4804	0.4638	538	5.35	<.0001
gender	female					-0.7246	0.2006	538	-3.61	0.0003
gender	male					0	.	.	.	.
raceEthnicity		group B				-0.9358	0.3520	538	-2.66	0.0081
raceEthnicity		group C				-0.9049	0.3273	538	-2.76	0.0059
raceEthnicity		group D				-0.5589	0.3465	538	-1.61	0.1073
raceEthnicity		group E				0	.	.	.	.
prntEduLvl			associate's degree			-0.4009	0.3551	538	-1.13	0.2594
prntEduLvl			bachelor's degree			0.3461	0.4035	538	0.86	0.3914
prntEduLvl			high school			-0.7928	0.3600	538	-2.20	0.0281
prntEduLvl			master's degree			0	.	.	.	.
lunch				free/reduced		-1.2085	0.2026	538	-5.97	<.0001
lunch				standard		0	.	.	.	.
prepCourse					completed	0.5098	0.2147	538	2.38	0.0179
prepCourse					none	0	.	.	.	.

Fitted model:

$$ln(odds) = \hat{\beta}_0 + \hat{\beta}_1 * female + \hat{\beta}_2 * groupB + \hat{\beta}_3 * groupC + \hat{\beta}_4 * groupD + \hat{\beta}_5 * AD + \hat{\beta}_6 * BD + \hat{\beta}_7 * HS + \hat{\beta}_8 * free/reduced + \hat{\beta}_9 * completed$$

$$ln(odds) = 2.4804 - 0.7246 * female - 0.9358 * groupB - 0.9049 * groupC - 0.5589 * groupD - 0.4009 * AD + 0.3461 * BD - 0.7928 * HS - 1.2085 * free/reduced + 0.5098 * completed$$

where:

- “female”: represents the gender of the individual
- “groupB”, “groupC”, “groupD” represent groups of the “raceEthnicity” variable
- AD: associate’s degree, BD: bachelor’s degree, HS: high school, are levels of the “prntEduLvl”
- “Completed” represents whether the preparation course was completed or not

Interpretation of the model parameters on the log-odds scale:

- $\hat{\beta}_1 = -0.7246$ ➔ $e^{\widehat{\beta}_1} = 0.4845 < 1$ ➔according to this dataset, on average, the odds of males passing a math exam are 0.4845 times the odds of females passing a math exam holding all other variables constant.
- $\hat{\beta}_2 = -0.9358$ ➔ $e^{\widehat{\beta}_2} = 0.3923 < 1$ ➔ On average, people who belong to group B have odds of passing a math exam equal to 0.3923 times the odds of people from group E, holding all other variables constant.
- $\hat{\beta}_3 = -0.9049$ ➔ $e^{\widehat{\beta}_3} = 0.4046 < 1$ ➔On average, the odds of people who belong to group C to pass a math exam are 0.4046 times the odds of people from group E, holding all other variables constant.
- $\hat{\beta}_4 = -0.5589$ ➔ $e^{\widehat{\beta}_4} = 0.5718 < 1$ ➔ On average, the odds of people who belong to group C to pass a math exam are 0.5718 times the odds of people from group E, holding all other variables constant.

- $\hat{\beta}_5 = -0.4009$ ➔ $e^{\hat{\beta}_5} = 0.6697 < 1$ ➔ On average, the odds of students whose parents have an associate's degree to pass a math exam are 0.6697 the odds of students whose parents have a master's degree, holding all other variables constant.
- $\hat{\beta}_6 = 0.3461$ ➔ $e^{\hat{\beta}_6} = 1.4135 > 1$ ➔ On average, the odds of students whose parent's have a bachelor's degree to pass a math test are 1.4135 times the odds of students whose parents have a master's degree, holding all other variables constant.
- $\hat{\beta}_7 = -0.7928$ ➔ $e^{\hat{\beta}_7} = e^{-0.7928} = 0.4526 < 1$ ➔ On average, the odds of students whose parents have only a high school degree are 0.4526 times the odds of passing a math test compared to the odds of students whose parents have a master's degree, holding all other variables constant.
- $\hat{\beta}_8 = -1.2085$ ➔ $e^{\hat{\beta}_8} = e^{-1.2085} = 0.2986 < 1$ ➔ On average, the odds of passing a math test for students having lunch for free or at a reduced priced are 0.2986 times the odds of students who pay the standard fee for their lunch, holding all other variables constant.
- $\hat{\beta}_9 = 0.5098$ ➔ $e^{\hat{\beta}_9} = e^{0.5098} = 1.665 > 1$ ➔ On average, the odds of passing the math test for people who did the preparation course are 1.665 times the odds of students who did not do the preparation course, holding all other variables constant.

c) Are the mean odds of passing a math test for the four groups of race/Ethnicity the same? Formally test this.

Solution –

To answer this question, we will do an ANOVA test on the mean odds of passing a math test for the four groups. The hypotheses for this test are:

$$H_0: \mu_B = \mu_C = \mu_D = \mu_E$$

$$H_1: at\ least\ two\ means\ are\ different\ from\ one\ another$$

With $\mu_B, \mu_C, \mu_D, \mu_E$ are the mean odds of passing a math test for people from race/Ethnicity groups B, C, D and E respectively.

The fitted model that corresponds to this ANOVA test is a logistic regression model with the variable "raceEthnicity" as the only explanatory variable. The model that we will fit is:

$$\eta = \beta_0 + \beta_B * GroupB + \beta_C * GroupC + \beta_D * GroupD$$

With:

η: linear combination of the covariates
β_0: The model intercept
β_B: The model parameter corresponding to "GroupB"
β_C: The model parameter corresponding to "GroupC"
β_D: The model parameter corresponding to "GroupD"

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
raceEthnicity	3	544	3.90	0.0090

From the figure above, we notice that the p-value for the F-test is 0.009 <<0.05 ➔ reject the null hypothesis stating that the average odds of passing a math exam for all the race/Ethnicity groups are the same at 5% significance level ➔ on average, there is significant difference in the odds of passing a math test of the 4 race/Ethnicity group at 5% significance level.

d) Fit a logistic regression model on the log-odds scale taking into consideration interactions between students taking the preparation course and the parental education level.

Solution –

In this part we aim to fit a logistic regression on a log-odds scale to predict the odds of passing certain test for a certain individual. However, we are concerned that parental level of education might affect if the students do the preparatory course or not. Therefore, we will add an interaction term between these two variables to account for this interaction.

Parameter Estimates

Effect	gender	raceEthnicity	prntEduLvl	lunch	prepCourse	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept						2.3983	0.5003	535	4.79	<.0001
gender	female					-0.7292	0.2012	535	-3.62	0.0003
gender	male					0	.	.	.	.
raceEthnicity		group B				-0.9455	0.3554	535	-2.66	0.0080
raceEthnicity		group C				-0.9278	0.3304	535	-2.81	0.0052
raceEthnicity		group D				-0.5626	0.3484	535	-1.61	0.1070
raceEthnicity		group E				0	.	.	.	.
prntEduLvl			associate's degree			-0.4166	0.4368	535	-0.95	0.3406
prntEduLvl			bachelor's degree			0.4753	0.4996	535	0.95	0.3419
prntEduLvl			high school			-0.5691	0.4385	535	-1.30	0.1949
prntEduLvl			master's degree			0	.	.	.	.
lunch				free/reduced		-1.2158	0.2041	535	-5.96	<.0001
lunch				standard		0	.	.	.	.
prepCourse					completed	0.7877	0.6711	535	1.17	0.2411
prepCourse					none	0	.	.	.	.
prntEduLv*prepCourse			associate's degree		completed	0.08574	0.7522	535	0.11	0.9093
prntEduLv*prepCourse			associate's degree		none	0	.	.	.	.
prntEduLv*prepCourse			bachelor's degree		completed	-0.3623	0.8572	535	-0.42	0.6727
prntEduLv*prepCourse			bachelor's degree		none	0	.	.	.	.
prntEduLv*prepCourse			high school		completed	-0.7453	0.7629	535	-0.98	0.3291
prntEduLv*prepCourse			high school		none	0	.	.	.	.
prntEduLv*prepCourse			master's degree		completed	0	.	.	.	.
prntEduLv*prepCourse			master's degree		none	0	.	.	.	.

e) Using the likelihood ratio test, is the nested model with interactions better than the model in the first part?

Solution –

Likelihood ratio test:

Fit Statistics	
-2 Log Likelihood	618.27
AIC (smaller is better)	638.27
AICC (smaller is better)	638.68
BIC (smaller is better)	681.33
CAIC (smaller is better)	691.33
HQIC (smaller is better)	655.10
Pearson Chi-Square	536.23
Pearson Chi-Square / DF	1.00

Fit Statistics	
-2 Log Likelihood	615.30
AIC (smaller is better)	641.30
AICC (smaller is better)	641.98
BIC (smaller is better)	697.28
CAIC (smaller is better)	710.28
HQIC (smaller is better)	663.18
Pearson Chi-Square	536.32
Pearson Chi-Square / DF	1.00

The test statistic for the likelihood ratio test is:

$$D = [-2LL(\hat{\theta}_{mle})_{reduced}] - [-2LL(\hat{\theta}_{mle})_{complete}] = 618.27 - 615.3 = 2.97$$

Under the null hypothesis: the complete and reduced models are not different in terms of goodness of fit, the test statistic D follows a Chi-squared distribution with 13-10=3 degrees of freedom. The obtained p-value is 0.39, which is greater than 0.05 (using $\alpha = 0.05$ as significance level), thus we fail to reject the null hypothesis stating that the complete and reduced models are not different in terms of goodness of fit. Therefore, we can say that at a 5% confidence level, on average, the complete and reduced models are not different in terms of goodness of fit.