Honor code

By submitting this assignment, I affirm the following:

1. All work presented in this assignment is my own. I have not collaborated with others or copied work from any unauthorized source.
2. If I used AI tools or large language models like ChatGPT, Co-Pilot, etc., I only sought guidance or clarification. Any generated content has been fully understood and appropriately modified to align with the assignment.
3. I understand the submitted code and can explain my work if asked.

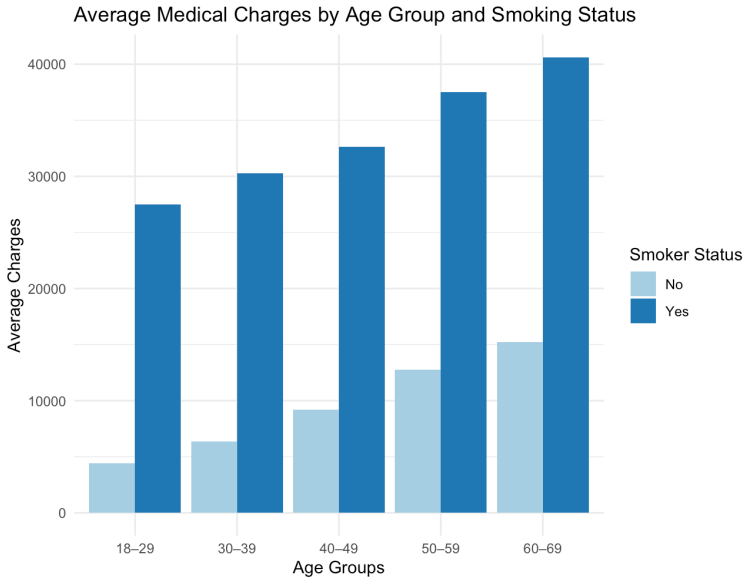
I declare that I have read, understood, and agree to abide by this honor code.

Name: Paola D'Incecco

Student number: 775206

Date: 21st November 2025

1. Age group with the largest difference in charges between smokers and non-smokers:
   Value of the maximum difference: **25,398**

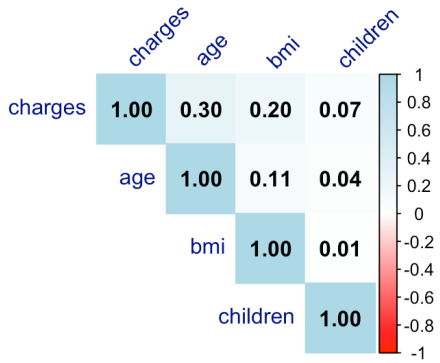


2. Correlation table:

	charges	age	bmi	children
charges	1.00	**0.30**	0.20	0.07
age	0.30	1.00	0.11	0.04
bmi	0.20	0.11	1.00	0.01
children	0.07	0.04	0.01	1.00

Variable that has strongest correlation with charges: **age**

Heat matrix:



Linear Regression Results

	Dependent variable:
	charges
age	256.856***
	(11.899)
gendermale	-131.314
	(332.945)
bmi	339.193***
	(28.599)
children	475.501***
	(137.804)
smokeryes	23,848.530***
	(413.153)
regionnorthwest	-352.964
	(476.276)
regionsoutheast	-1,035.022**
	(478.692)
regionsouthwest	-960.051**
	(477.933)
Constant	-11,938.540***
	(987.819)
Observations	1,338
R^2	0.751
Adjusted R^2	0.749
Residual Std. Error	6,062.102 (df = 1329)
F Statistic	500.811*** (df = 8; 1329)
Note:	*p<0.1; **p<0.05; ***p<0.01

3. Report variables that are statistically significant and the significance level: Statistically significant variables are those marked by *, **, *** which respectively indicate a significance level of 0.1, 0.05 and 0.01. Hence:

- Age, with significance level 0.01
- Bmi, with significance level 0.01
- Children, with significance level 0.01
- Smokeryes, with significance level 0.01
- Regionsoutheast, with significance level 0.05
- Regionsouthwest, with significance level 0.05

4. Interpretation of coefficients (1-2 sentences each):

Disclaimer: As approved by the Professor, I chose the reference category for the predictor variables gender, smoker and region based on R default use of alphabetical order. Consequently, the used reference categories respectively are genderfemale, smokerno and regionnortheast.

- age: The coefficient for age represents the marginal effect of a one-year increase in age on annual medical charges, after controlling for gender, BMI, number of children, smoking status, and region. Since the coefficient is positive, controlling for all other predictors, each additional year of age increases expected annual medical charges by 256.856 units, i.e. older individuals are expected to incur higher medical costs.

- gender: The coefficient for gendermale represents the difference in expected charges between males and the reference category (females), controlling for age, BMI, number of children, smoking status, and region. The estimate of −131.314 suggests that males have slightly lower predicted charges than females, although the large standard error (332.945) indicates this effect is not statistically significant.

- bmi: The BMI coefficient represents the marginal increase in expected medical charges associated with a one-unit rise in BMI, controlling for age, gender, number of children, smoking status, and region. The estimate of 339.193 indicates that individuals with higher BMI incur substantially higher predicted medical expenditures, with a small standard error (28.599) compared to the magnitude of the coefficient indicating that the estimate is relatively precise.

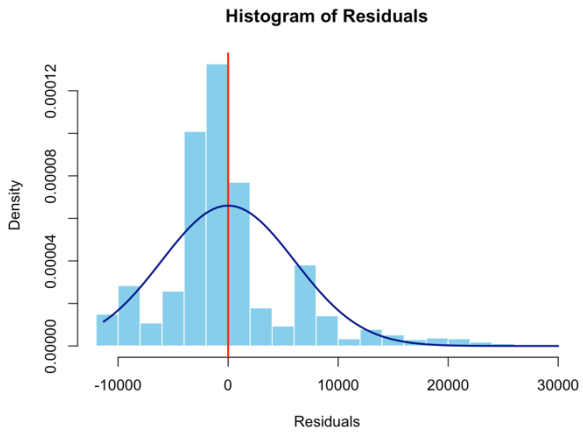
- region (change this accordingly):

    - regionnorthwest: Compared to the reference region (northeast), living in the northwest is associated with lower predicted charges (-352.964 lower), holding all other variables constant. However, the standard error (476.276) is

large relative to the coefficient, so the model does not estimate difference between regional areas certainly.

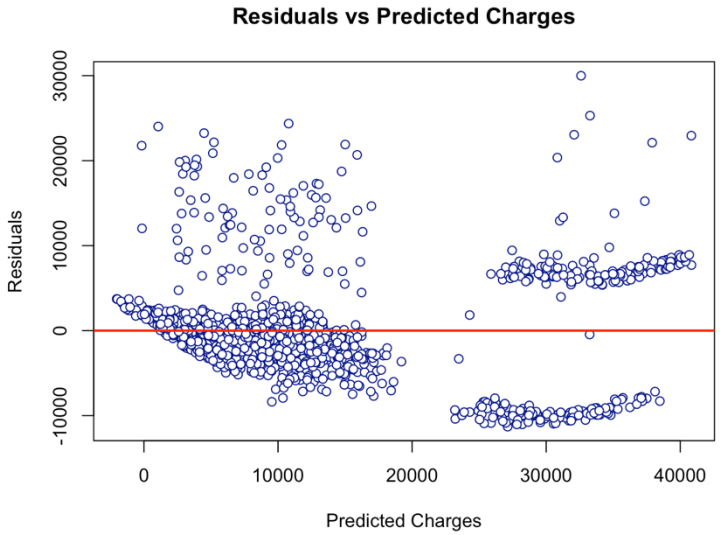
- regionsoutheast: Compared to the northeast, individuals in the southeast have lower predicted charges (-1035.022), controlling for all the other variables. Here, despite the large standard error (478.692) despite the large standard error suggests a cautious intepretation, the effect is still statistically significant.
- regionsouthwest: Compared to the northeast, living in the southwest is associated with lower predicted charges (-960.051), holding all other variables constant. Again, the standard error (477.933) is large, so this estimate should be interpreted with caution, still being statistically significant.

5. a) Histogram of residuals:



Comment on the histogram:
Residuals appear symmetric around zero for most values, but the distribution is slightly right skewed due to several large positive outliers. Therefore, consistently with the model's strong performance metrics (R^2 ≈ 0.75 and a highly significant F-statistic), the model fits the data well despite the normality assumption is somewhat violated in the righten residual tail.

b) Residuals vs. predicted charges



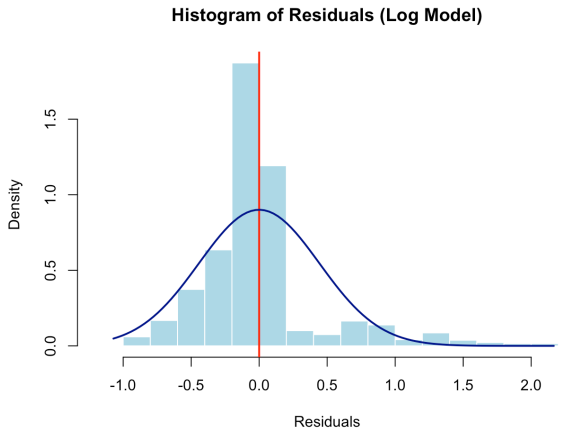
Comment on homoscedasticity assumption:
Homoscedasticity assumption is evidently violated, since residuals are not constantly spread across levels of predicted charges. Despite the plot doesn't display the typical heteroskedastic fan or cone shape, residuals' variance still forms distinct bands rather than remaining randomly and evenly scattered.

Log Regression Results

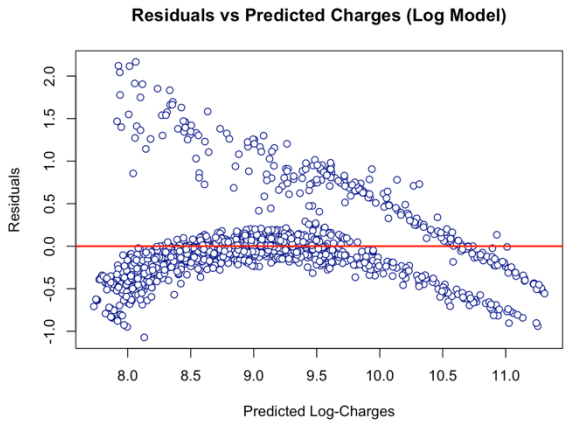
	Dependent variable:
	log_charges
age	0.035***
	(0.001)
gendermale	-0.075***
	(0.024)
bmi	0.013***
	(0.002)
children	0.102***
	(0.010)
smokeryes	1.554***
	(0.030)
regionnorthwest	-0.064*
	(0.035)
regionsoutheast	-0.157***
	(0.035)
regionsouthwest	-0.129***
	(0.035)
Constant	7.031***
	(0.072)
Observations	1,338
R^2	0.768
Adjusted R^2	0.767
Residual Std. Error	0.444 (df = 1329)
F Statistic	549.770*** (df = 8; 1329)
Note:	*p<0.1; **p<0.05; ***p<0.01

6. a) Interpretation of the coefficient of age: Since it is a log-linear model, a 0.035 coefficient indicates a 1-unit increase in age is associated with a 3.5% change in predicted charges, holding all other variables constant.

b) Histogram of residuals:



Comment on the histogram:
The histogram of residuals from the log-transformed model shows a pronounced right skew (even more than the one analysed before), with a long tail of larger positive residuals. Thus, as it can be seen from the theoretical normal curve overlay, normality assumption is not fully satisfied, even after applying the log transformation.



c) Residuals vs. predicted charges

Comment on homoscedasticity assumption:
Homoscedasticity assumption is violated, since the spread of residuals evidently changes across the range of predicted log-charges (wide at low values, tighter in the middle, finally patterned going downwards at higher values). The slightly curved, nonrandom, distribution demonstrates a nonlinear relationship.

7. R-squared and Adjusted R-squared value and explanation:
   R-squared: 0.768.
   Adjusted R-squared: 0.767
   The R-squared value of 0.768 indicates that 76.8% of variance in log-transformed predicted charges (the dependent variable) is explained by residuals predictors (independent variables) in the log regression model. The very similar Adjusted R-squared value of 0.767 proves that the additional predictor variables are not significantly improving the model's explanatory power. These results suggest that the log regression model fits the data well and that the variables included are meaningful.

8. Model prediction→add euro symbol to charges:

age	**gender**	**bmi**	**children**	**smoker**	**region**	**Charges (€)**
25	male	28.0	1	no	northeast	4,007.95
45	Female	35.2	3	yes	southeast	47,109.14
32	male	30.5	0	no	northwest	4,473.38
54	female	24.7	2	yes	southwest	51,917.39
29	female	22.8	1	yes	southeast	18,720.13

9. % of observations assigned to class 0: 68.61
   % of observations assigned to class 1: 31.39

Logistic Regression Results

	Dependent variable:
	binary_charges
age	0.071***
	(0.008)
gendermale	-0.275
	(0.189)
bmi	0.018
	(0.016)
children	0.118
	(0.076)
smokeryes	8.397***
	(1.025)
regionnorthwest	-0.158
	(0.257)
regionsoutheast	-0.203
	(0.263)
regionsouthwest	-0.665**
	(0.276)
Constant	-5.311***
	(0.639)
Observations	1,338
Log Likelihood	-378.402
Akaike Inf. Crit.	774.804
Note:	*p<0.1; **p<0.05; ***p<0.01

10. Report variables that are statistically significant and the significance level: Statistically significant variables are those marked by *, **, *** which respectively indicate a significance level of 0.1, 0.05 and 0.01. Hence:

a. Age: 0.071, with significance level 0.01).

b. Smokeryes: 8.397, with significance level 0.01).

c. Regionsouthwest: -0.665, with significance level 0.05.

11. Interpretation of coefficients (1-2 sentences each)→improve:

- age: Holding all other variables constant, each additional year of age increases the logistic odds of having high charges by ≈ 7.36% (since odds ratio of (1.0736−1)x100%≈7.36%).
As a consequence, older individuals are more likely to fall into the high-charges category.

- gender: Compared to females, being male slightly reduces the log-odds of belonging to the high-charges category. Indeed, the odds ratio (≈ 0.76) indicates a small decrease in odds, but the effect is small and not statistically significant.

- bmi: Keeping all other variables constant, a higher BMI is associated with a small increase in the log-odds of having high medical charges. The odds ratio (≈ 1.02) corresponds to a ~2% increase in odds per BMI point, but even in this case, the effect is weak and not statistically significant.

- region (change this accordingly):
    - regionnorthwest: Compared to people living in the northeast, people living in the northwest have slightly lower log-odds of high charges – the odds ratio (≈ 0.85) implies a 15% reduction in odds. Still, the effect is small and not statistically significant.
    - Regionsoutheast: Similarly tot he previous case, people living in the southeast have slightly lower log-odds of having high charges than people living in the northeast. The odds ratio (≈ 0.82) suggests an 18% decrease in odds, but this difference is not statistically significance.
    - Regionsouthwest: Living in the southwest significantly reduces the log-odds

of belonging to the high-charges class by 48.6% when compared to living in the northeast (since odds ratio 0.514−1x100%≈−48.6%).

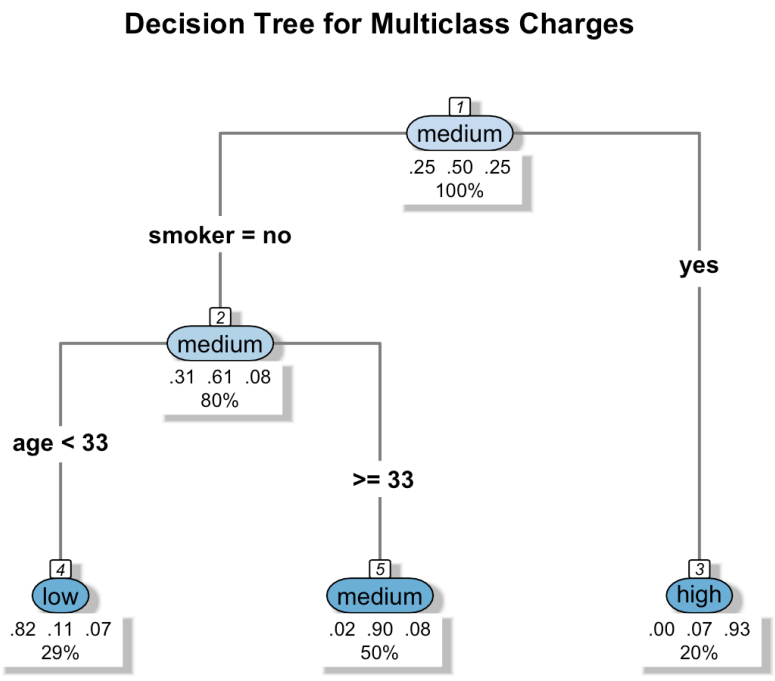
12. Model prediction:

age	gender	bmi	children	smoker	region	binary_charges
25	male	28.0	1	no	northeast	0
45	female	35.2	3	yes	southeast	1
32	male	30.5	0	no	northwest	0
54	female	24.7	2	yes	southwest	1
29	female	22.8	1	yes	southeast	1

13. % of observations assigned to class 'low': 25.04
    % of observations assigned to class 'medium': 49.93
    % of observations assigned to class 'high': 25.04

14. # of leaf nodes in the decision tree: 3

15. Decision tree plot:



Explain one of the paths (1-2 sentences):

Starting from the root node (1), if an individual is a non-smoker, the tree moves to a node where 61% of observations fall into the medium charges class. From there, if the individual is younger than 33 years old, the model assigns them to the low charges category (4), where 82% of cases in that leaf actually belong to the low class.

16. Model prediction:

age	gender	bmi	children	smoker	region	multiclass_charges
25	male	28.0	1	no	northeast	low
45	female	35.2	3	yes	southeast	high
32	male	30.5	0	no	northwest	low
54	female	24.7	2	yes	southwest	high
29	female	22.8	1	yes	southeast	high