

The background features a dark blue gradient with abstract white and light blue geometric shapes. A line graph with four data points is visible on the left side. The data points are connected by white lines, and the values are 289.33, 289.33, 289.33, and 289.33. The title text is centered at the top in a large, bold, white font.

MEDICAL INSURANCE PREDICTOR USING MULTIPLE LINEAR REGRESSION

Semester Project

Group I

Purushottam Saha (BS2119)

Abdur Rahman (BS2146)

Amarjeet Kumar Bhanu (BS2045)

Our Target



ANALYSING DATA
AND COVARIATES



FITTING MODEL AND
BETTERMENTS



SIMULATION AND
CONCLUSION

OBJECTIVE

- Our aim is to do conduct an exploratory data analysis on Medical Charges in US, of different regions (northeast, southeast, southwest and northwest) , attempting to obtain suitable predictors and identify causal relations through fitting the data to a Multiple Linear Regression model (MLR) .
- After Fitting the data, we will be predicting Medical Charges for Insurees using simulation to average out the missing information.
- We have collected this dataset from the website [Kaggle](#). There are a total of 1338 subjects (Insurees) in the dataset.
- Along with the response variable Medical Charges , we have 7 other covariates as our predictor variables.

Data Description

■ The covariates of our Data are

- **Age** : Age of Primary Beneficiary
- **Sex** : Gender Of Insuree (Female=0, Male=1)
- **BMI** : Body Mass Index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9
- **Children** : Number of children covered by health insurance
- **Smoke** : Smoking state of Beneficiary (Non-Smoker=0 , Smoker=1)
- **Region** : Beneficiary's residential area in the US (Northeast = 0, Northwest = 1, Southeast = 2, Southwest = 3)
- **Insurance Claim** : Whether the Beneficiary had previously claimed the Insurance (No = 0 , Yes = 1)

And the Response Variable is

- **Charges** : Beneficiary medical costs billed by health insurance

Model Fitting

We have fitted an Multiple Linear Regression Model (MLR) to the data.

We have used the Ordinary Least Square technique assuming that errors follow Normal distribution.
Here is the summary.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-12357.63	966.21	-12.790	< 2e-16 ***
AGE	262.31	11.94	21.963	< 2e-16 ***
SEX	-134.69	331.62	-0.406	0.68469
BMI	377.69	30.91	12.218	< 2e-16 ***
CHILD	228.52	157.36	1.452	0.14668
SMOKER	24421.75	450.11	54.257	< 2e-16 ***
REGION	-374.40	151.52	-2.471	0.01360 *
CLAIM	-1458.74	448.61	-3.252	0.00118 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model Fitting

Residuals:

Min	1Q	Median	3Q	Max
-11613.2	-2803.1	-943.7	1236.6	29600.1

Residual standard error: 6038 on 1330 degrees of freedom

Multiple R-squared: 0.7527, Adjusted R-squared: 0.7514

There are six significant covariates and two insignificant covariates.

Three co-efficient are negative along with the intercept, and other four co-efficient are positive.

Correlation Matrix

	Age	Sex	BMI	Children	Smoker	Region	Insurance Claim	Charges
Age	1.000	-0.020	0.109	0.042	-0.025	0.002	0.113	0.299
Sex	-0.020	1.000	0.046	0.017	0.076	0.004	0.031	0.057
BMI	0.109	0.046	1.000	0.012	0.003	0.157	0.384	0.198
Children	0.042	0.017	0.012	1.000	0.007	0.016	-0.409	0.067
Smoker	-0.025	0.076	0.003	0.007	1.000	-0.002	0.333	0.787
Region	0.002	0.004	0.157	0.016	-0.002	1.000	0.020	-0.006
Insurance Claim	0.113	0.031	0.384	-0.409	0.333	0.020	1.000	0.309
Charges	0.299	0.057	0.198	0.067	0.787	-0.006	0.309	1.000

Observations from Correlation Matrix

- Most of the factors are not that much correlated, which agrees with our assumption that the covariates are mutually independent.
- The Response variable Charges is significantly correlated with the predictor variables, the notable ones being age, insurance claims and smoker (in increasing order). Smoker (i.e. if the insuree is smoker or not) has correlation of 0.787, which explains the such high value of the regression coefficient.
- Insurance claims and BMI pair, along with Insurance claims and Children pair are mainly the pairs that are somewhat highly correlated between the covariates. Most likely there are some latent variables for which both of them are correlated, like: health factors and family financial conditions, but more accurate reasons are up for further research,

Partial Correlation

	Age	Sex	BMI	Children	Smoker	Region	Insurance Claim	Charges
Age	1.000	-0.015	-0.131	0.064	-0.463	0.025	0.156	0.516
Sex	-0.015	1.000	0.045	0.013	0.048	-0.004	-0.004	-0.011
BMI	-0.131	0.045	1.000	0.199	-0.359	0.172	0.452	0.317
Children	0.064	0.013	0.199	1.000	0.082	-0.004	-0.484	0.039
Smoker	-0.463	0.048	-0.359	0.082	1.000	0.064	0.301	0.829
Region	0.025	-0.004	0.172	-0.004	0.064	1.000	-0.047	-0.067
Insurance Claim	0.156	-0.004	0.452	-0.484	0.301	-0.047	1.000	-0.088
Charges	0.516	-0.011	0.317	0.039	0.829	-0.067	-0.088	1.000

Observations from Partial Correlations

- Most of the relations of partial correlations for covariate pairs are somewhat unaltered, though there are some significant changes to be noticed, for example, correlation between sex and charges have flipped signs, explaining the negative regression coefficient.
- Also, the correlation between smoker and bmi has changed from 0.0037 to -0.3598 which is pretty interesting.
- The correlation between age and charges is approximately doubled, explaining the importance of the variable as a predictor, also the reason why partial correlations can be much more informative at times.
- The correlation between smoker and charges has also increased to 0.829, which is a pretty large value, supporting the p-value of the same being 0 by R, though being realistically absurd.

Semi Partial Correlation

	Age	Sex	BMI	Children	Smoker	Region	Insurance Claim	Charges
Age	1.000	-0.012	-0.111	0.054	-0.441	0.021	0.133	0.507
Sex	-0.015	1.000	0.045	0.0136	0.048	-0.004	-0.004	-0.011
BMI	-0.109	0.038	1.000	0.168	-0.320	0.145	0.421	0.278
Children	0.056	0.011	0.176	1.000	0.072	-0.003	-0.481	0.034
Smoker	-0.265	0.024	-0.195	0.041	1.000	0.0326	0.160	0.754
Region	0.025	-0.004	0.172	-0.004	0.063	1.000	-0.047	-0.066
Insurance Claim	0.117	-0.003	0.377	-0.411	0.235	-0.036	1.000	-0.066
Charges	0.299	-0.005	0.166	0.019	0.739	-0.033	-0.044	1.000

Observations from Semi Partial Correlation

- Though there are no significant changes in Semi Partial Correlations from Partial Correlations, but it is always interesting to notice that the p-value of correlation between smoker and charges being approximately 0.75 is of order e^{-231} .
- Also it is interesting to notice that the correlation between smoker and region changed from 0.064 to 0.0326, and that of smoker and Insurance claims changed from 0.301 to 0.160, simultaneously. It is also notable that semi partial correlation between smoker and Insurance claims (effects of others removed from Insurance claims) is 0.160, where the same when the effects of others are removed from smoker, the correlation becomes 0.235, in place of the partial correlation of 0.301. It says that smoker is not that much correlated with other variables which Insurance claims is.

Modified Model

- In the Coefficients Part for the MLR, we saw some higher p-values for some covariates, i.e. sex and children.
- So we fit another model by neglecting those two covariates. Here is the summary :

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-12375.26	957.28	-12.927	< 2e-16 ***
age	264.15	11.88	22.232	< 2e-16 ***
bmi	387.17	30.10	12.865	< 2e-16 ***
smoker	24544.58	439.15	55.891	< 2e-16 ***
prev_claims	-1776.41	391.01	-4.543	6.04e-06 ***
region	-375.78	151.53	-2.480	0.0133 *

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- The Multiple R-squared value is 0.7523, where the adjusted R-squared is 0.7514.
- The Multiple R-squared has very slightly decreased, but adjusted R-squared has not changed.
- And the Residuals are:

Min	1Q	Median	3Q	Max
-11503.0	-2803.5	-954.6	1346.3	29310.3

We have tried to remove the top 2 outliers from the response variable and got some betterment in the residuals, but that did not have much broader effects on other things.

Simulating to Predict Charges

With this models done, now comes the prediction part. Lets see the question:

- Question 1 : For an Male individual of age 24, if he does not smoke and has no children, then what would be the expected charges for him?
 - *From the first model, the answer comes out to be 3968.539 dollars*
 - *From second model, the answer comes out to be 4232.1 dollars, some increment from the previous case.*
 - *It's interesting that, if the only condition changed was that the gender was female, the predicted cost is 4103.228 dollars, a slight increase in price, supporting that the gender-neutral second model has also predicted a higher cost.*

- Question 2: For an Female individual of age in range 50-60 with a comparatively healthy body (BMI in range 23-27), who does smoke and already had a previous insurance claimed, what would be the expected charges for her?
 - *From the first model, the answer comes out to be 34159.99 dollars.*
 - *From second model, the answer comes out to be 34032.39 dollars, some decrement from the previous case.*
 - *If we specify that she has 0 child, then the predicted charges becomes 33922.72 and 34046.41 dollars respectively*
 - *If we more specify that she is from North-East, then her predicted cost becomes 34471.23 and 34596.67 respectively.*

So we can see, without the extra information added, the predicted value was also pretty close.

Another interesting fact is, if we simulated for the smoker predictor, as only 20% people in our data smokes, the charges drastically reduces down to 15185.54 and 15213.93 dollars respectively for two models, supporting such high coefficient of smoker.

Conclusion

- From the data it is evident that, sex is not a significant covariate, though there can be a difference in charges, as the coefficient is non-zero (-134.69).
- Having Children have a positive coefficient (228.52), though as it has not much of variance, its not that much of a significant covariate.
- The most important or can be said that the covariate that changes the cost drastically is smoking (coefficient 24421.75 and 24544.58 respectively in two models), which is pretty intuitive, as smoking brings a lot of diseases as possibilities.
- Previous Insurance claims are important while considering charges, which is intuitive, It is also interesting that removing sex and children covariates, the p-value of previous insurance claims becomes much more lower (from 0.00118 to 6.04e-06), which is up for future studies.

INTRODUCTION

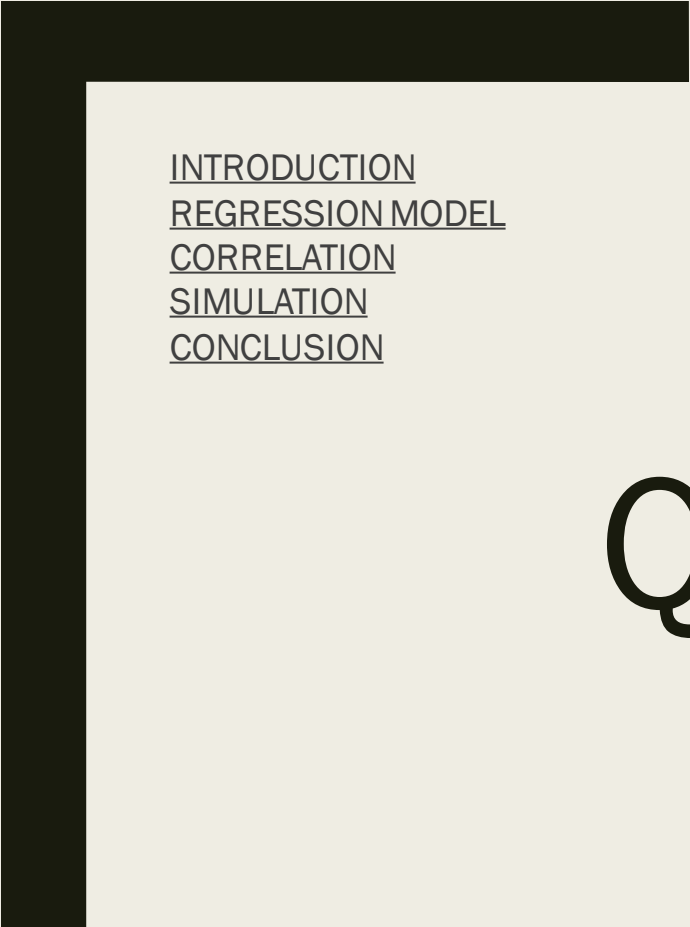
REGRESSION MODEL

CORRELATION

SIMULATION

CONCLUSION

- Decreasing two covariates in the better model, doesn't necessarily increase the R-squared value, in fact it decreases the R-squared value (not adjusted one). But the important thing is all the covariates are having significant effects on the response variable.
- Here we did not accurately use the regression techniques needed for a categorical data like region, where we had 4 categories. But for the more accurate models the interpretations were not that much clear and sound. But further analysis on the same is needed to say more about that.
- Also another further improvement for the data would be to consider other bad practices for health like drinking alcohols and having long term diseases i.e. Chronic Diseases like Heart Diseases, Cancer, Diabetes which has been the leading causes of death and disability in the United States; alleviating some pressure from smokers for high Medical Charges.

A thick black L-shaped bar in the top-left corner, consisting of a horizontal segment extending to the right and a vertical segment extending downwards.

INTRODUCTION
REGRESSION MODEL
CORRELATION
SIMULATION
CONCLUSION

QUESTIONS?

A thick black L-shaped bar in the bottom-right corner, consisting of a horizontal segment extending to the left and a vertical segment extending upwards.



THANK YOU