Tasks

Objective: Predict expense

1. Download insurance dataset:

<https://drive.google.com/open?id=1QN5TSG3pJHo8v05gVsPKbvLwOyyaDy8t>

The insurance.csv file includes 1,338 examples of beneficiaries currently enrolled in the insurance plan, with features indicating characteristics of the patient as:

* **age:** An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
* **sex:** The policy holder's gender, either male or female.
* **bmi:** The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
* **children:** An integer indicating the number of children/dependents covered by the insurance plan.

• **smoker:** A yes or no categorical variable that indicates whether the

insured regularly smokes tobacco.

• **region:** The beneficiary's place of residence in the US, divided into four

geographic regions: northeast, southeast, southwest, or northwest.

1. Check data format (str()/summary())
2. See dataset is suitable for regression? Otherwise do the needful (sex, smoker, region is categorical variable, whereas regression works on only numerical variable)
3. Observe distribution of expense (Gaussian?)
4. Observe that features are independent each other?
5. Train a linear regression model on dataset.

**(A1)Expenses= B0 + B1age+B2sex+B3bmi+B4children+B5smoker**

**+B6region**

**Fill the blank:**

1. **every 1 year increase in the age of a person, the required expense goes up by………….**
2. **every 1 children (dependent) increase in family of a person, the required expense goes up by………….**
3. **How many independent (input) features in your model.**
4. **How many features are visible in your model when see the summary?**
5. **A=b?.......if not why…??**
6. **b is superset of a? if not why??**
7. **Refer Model A1, Observer the statistical properties of model:**
8. **Fill the following**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Coefficient** | **std. error of stm.** | **t-value** | **p-value** |  |  |
| **Age** |  |  |  |  |  |  |
| **Children** |  |  |  |  |  |  |
| **bmi** |  |  |  |  |  |  |
| **sex** |  |  |  |  |  |  |
| **smoker** |  |  |  |  |  |  |
| **region** |  |  |  |  |  |  |
| **R2** |  |  |  |  |  |  |

1. Improve the performance of model A1 as-
2. expenses ~ age2 + children + bmi +sex + smoker + region
3. expenses ~ age+ age2 + children + bmi +sex + smoker + region
4. optimize by converting **numeric variable to a binary indicator as:**

**if bmi more than 30 = 1, otherwise 0**

**give this new feature name bmi30**

1. expenses ~ age+age2 + children + bmi30 +

sex + smoker + region, data = insurance

1. expenses ~ age+age2 + children + bmi+ bmi30 +

sex + smoker + region

1. Observer R2 for all model in setp9.
2. All so observed in all model in setp9, which feature is imp?
3. Optimize by **adding interaction effects**

expenses ~ age+age2 + children +bmi + sex + bmi30\*smoker + region. Observe the value of R2.