Lab 2 Activity

PSYC 7804 - Spring 2026

Today we will be looking at the `insurance.csv` dataset. These data are data simulated according to the US census data and can be found here on this Kaggle page. To load the data run the following code:

```
dat <- rio::import("https://fabio-setti.netlify.app/data/insurance.csv")
```

Here is a descriptions of the variables in the data:

variable	description
age	Age of primary beneficiary
sex	Insurance contractor gender, female, male
bmi	Body mass index
children	Number of children covered by health insurance / Number of dependents
smoker	Whether the beneficiary is a smoker
region	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest
charges	Individual medical costs billed yearly by health insurance in $

1. Create a scatterplot with `bmi` on the x-axis and `charges` on the y-axis and draw a regression line. Does the regression line match your expectations? Do you see something a bit strange about the distribution of the dots around the regression line?

2. Create Mean-centered and Standardized versions of both `bmi` and `charges`.

- What is the mean, standard deviation, and range (you can use the `range()` function) of `bmi` and `charges` and their linearly transformed counterparts?

3. Run an unstandardized linear regression with `bmi` predicting `charges`. Write the regression equation and interpret the intercept and the slope.

- Do you notice anything off with the interpretation of the intercept?

4. Using any of either the original or linearly transformed variables, run a regression such that the intercept represents the expected value of `charges` when `bmi` is at the sample mean. According to these data, what are the expected annual `charges` in dollars for an individual with average BMI?

5. Using any of either the original or linearly transformed variables, find the correlation between `charges` and `bmi`. (only use the `lm()` and `summary()` function to answer this question)

6. Assuming that some variable is roughly normally distributed, standardization transforms the variable into Z-scores (see here). In the case of a Z-score, we know that a value of 0 generally represents the 50^{th} percentile, while a value of 1 (i.e., *1 standard deviation above the mean*) represents the 84^{th} percentile. The `pnorm()` function returns the corresponding percentile given any Z-score:

```
pnorm(0)
```

```
[1] 0.5
```

```
pnorm(1)
```

```
[1] 0.8413447
```

Assuming that `charges` is approximately normally distributed, in what percetile of `charges` are individuals who are 1.5 standard deviations above the mean `bmi` predicted to be?

- Think about the effect of `bmi` on `charges`. Assuming a roughly normal distribution, 1.5 corresponds to what `bmi` percentile? Given the magnitude of the change in `bmi` necessary for the predicted percentile of `charges`, how impactful do you think is the effect of `bmi` on `charges`?