Tutorial 11

Statistical Computation and Analysis

Spring 2025

Homework Submission Format

Please submit a **PDF report** & **Jupyter notebook** (.ipynb file)!

- The pdf report should include all plots and explanations.
  - Keep it organized and easy to read
  - Type it – don’t submit hand-written reports
    - If we can’t understand it – we can’t grade it…

- If you only submit one – we will not check the assignment
- If you submit the wrong format – we will deduct points (-5 now, next assignment -10)

- This is not new information… חבל על הציון שלכם

Tutorial Outline

- Categorical predictors
- Bayesian reporting

Categorical predictors

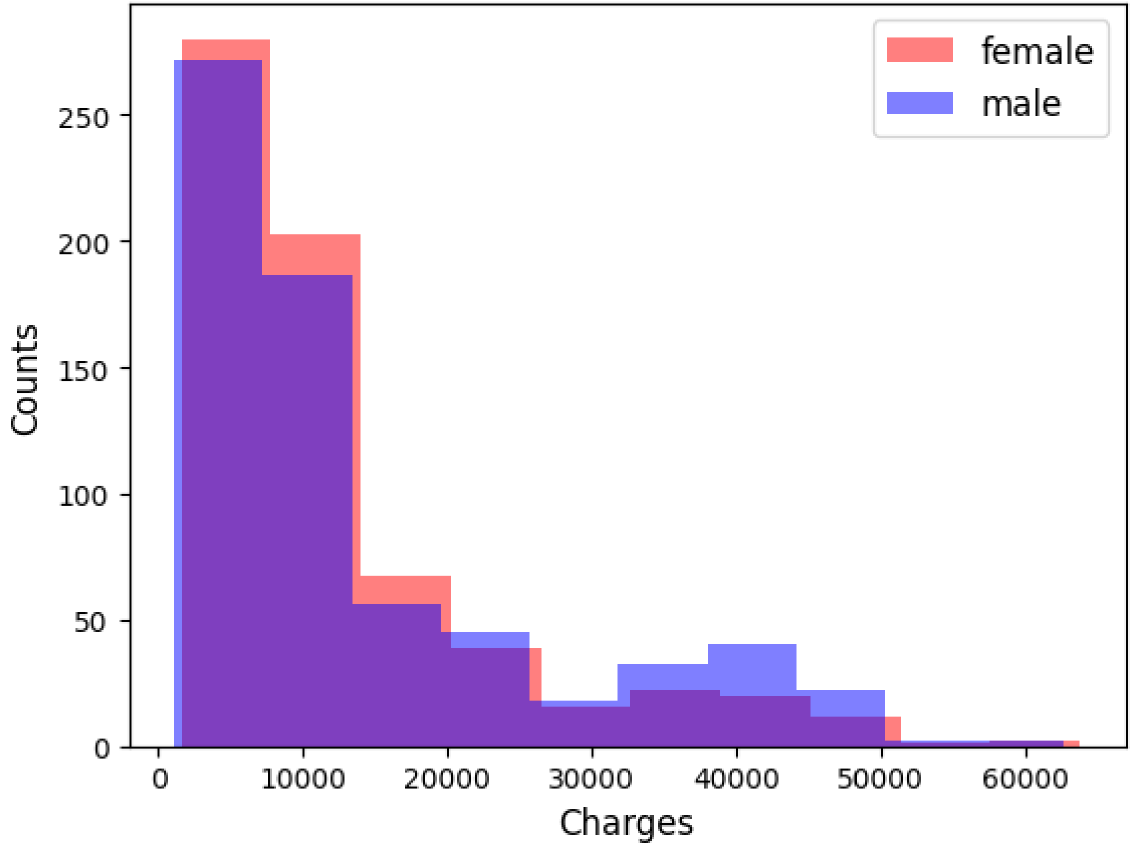
Categorical Predictors

- We mainly saw independent variables that were continuous.
- There are cases when they are categorical.
  - Male, female
  - Morning, afternoon, evening
- We can still do linear regression.
  - We need to encode the categorical variable as numbers.
  - Bambi

Categorical Predictors

- We have insurance costs for males and females.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030



Categorical Predictors

- Let's use bambi to assess the effect of gender on insurance charges.

```
model_t = bmb.Model("charges ~ sex", data)
idata_t = model_t.fit(1000, chains = 4)
```

- We don't need to tell bambi if a variable is categorical.
  - Bambi detects and handles them automatically.

Categorical Predictors

- We can examine the model:

```
       Formula: charges ~ sex
        Family: gaussian
          Link: mu = identity
  Observations: 1338
        Priors:
    target = mu
        Common-level effects
            Intercept ~ Normal(mu: 13270.4223, sigma: 43025.0381)
            sex ~ Normal(mu: 0.0, sigma: 60530.7385)

        Auxiliary parameters
            sigma ~ HalfStudentT(nu: 4.0, sigma: 12105.485)
```

- The model that we fit is: $charges = \beta_0 + b_1 \cdot sex$

Categorical Predictors

- And the inference data object:

arviz.InferenceData

▼ posterior

xarray.Dataset

► Dimensions: (**chain**: 4, **draw**: 1000, **sex_dim**: 1)

▼ Coordinates:

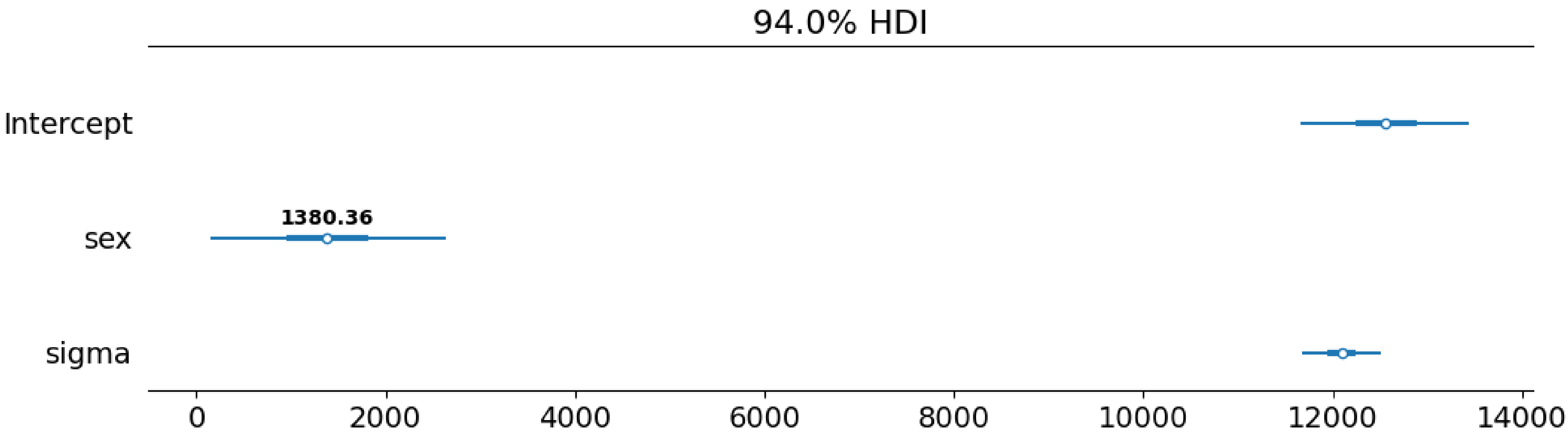
chain	(chain)	int64	0 1 2 3
draw	(draw)	int64	0 1 2 3 4 5 ... 995 996 997 998 999
sex_dim	(sex_dim)	<U4	'male'

▼ Data variables:

Intercept	(chain, draw)	float64	1.155e+04 1.233e+04 ... 1.22e+04
sex	(chain, draw, sex_dim)	float64	2.194e+03 591.1 ... 2.139e+03
sigma	(chain, draw)	float64	1.215e+04 1.206e+04 ... 1.161e+04

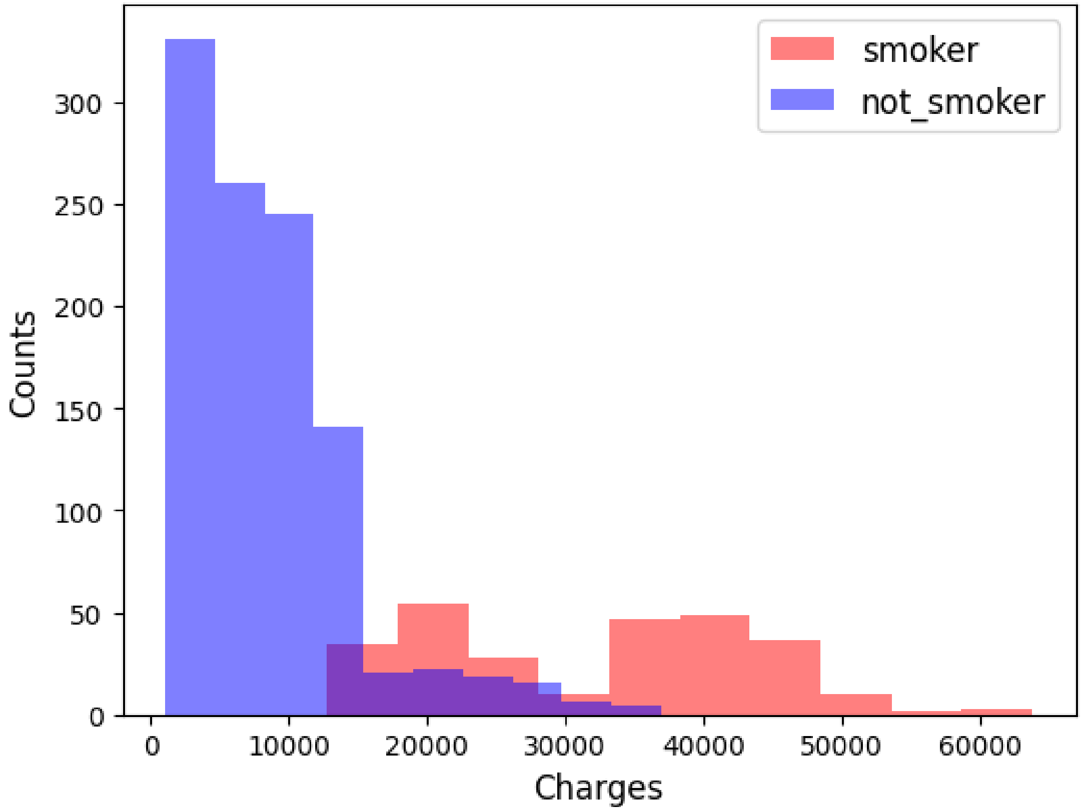
Categorical Predictors

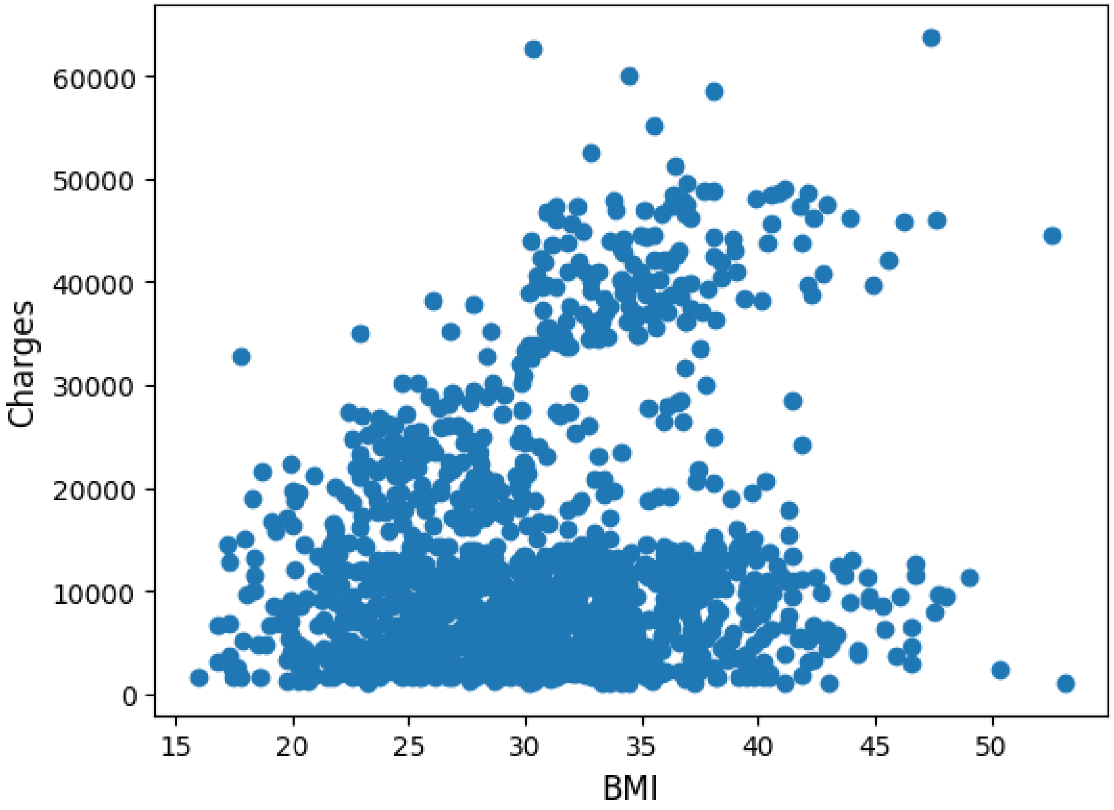
- We can see that the gender dimension is one, even though there are two genders.
- Bambi encodes categorical variables with N levels (2 genders) as N-1 dummy variables (1 gender).
- This means that on average the cost for males is 1380.36 more than for females.



Categorical Predictors

- Our dataset also contains other information, such as if the person smokes (categorical) and their BMI (continuous).





Categorical Predictors

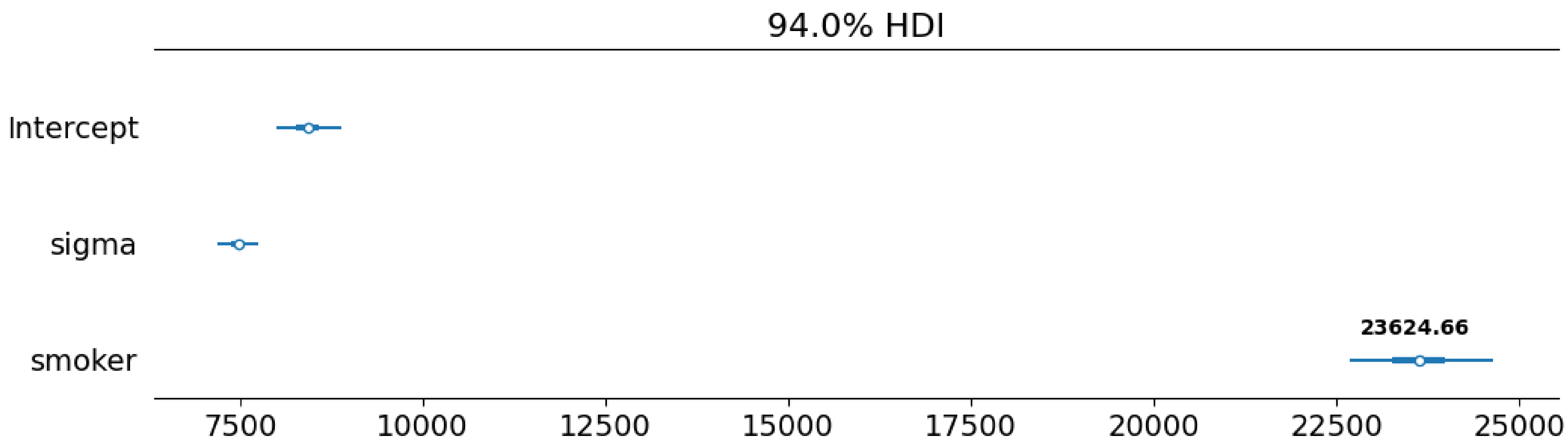
- We can fit a model to each independent variable.

```
model_s = bmb.Model("charges ~ smoker", data)
idata_s = model_s.fit(1000, chains = 4)
```

```
model_b = bmb.Model("charges ~ bmi", data)
idata_b = model_b.fit(1000, chains = 4)
```

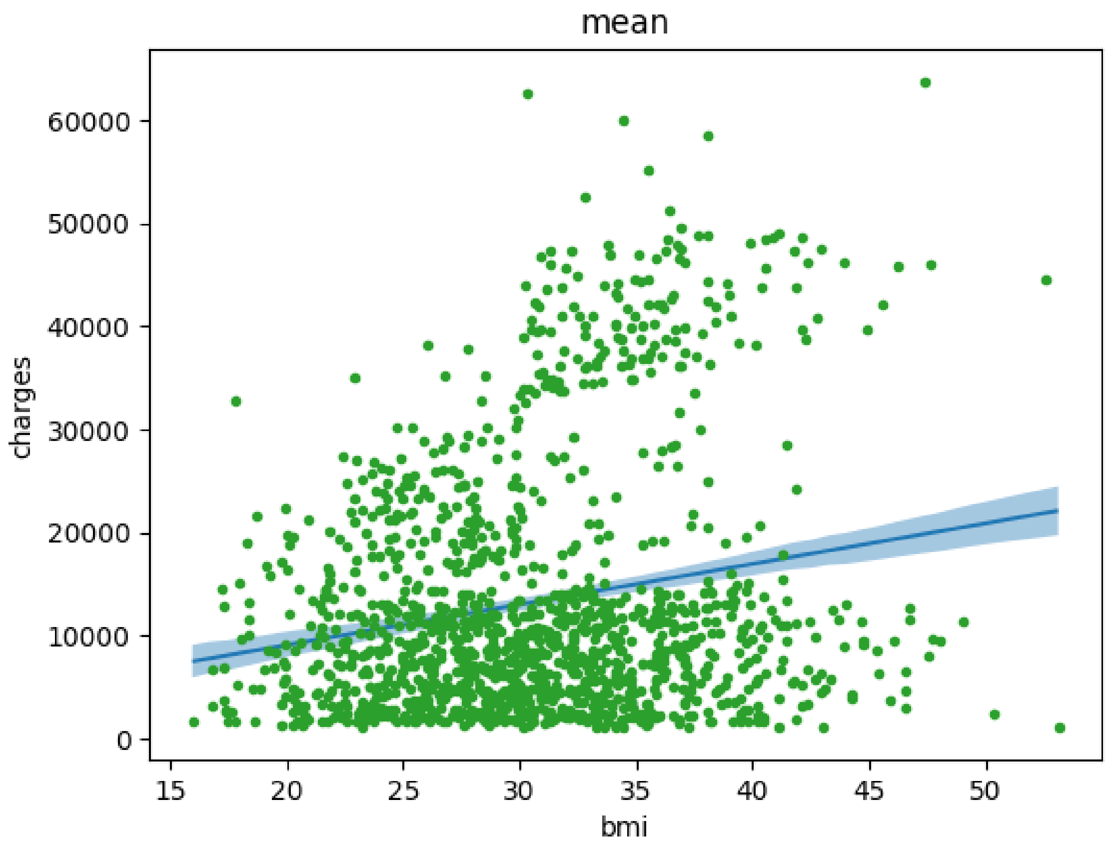
Categorical Predictors

- Results smoker:
  - We need to add 23624 to smoker charges relative to not smoker.



Categorical Predictors

- Results bmi:



Categorical Predictors

- How do we know which model is best?
  - We can compare using LOO.

```
az.compare({"gender": idata_g, "smoker": idata_s, "bmi": idata_b})
```

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
smoker	0	-13834.573669	4.661635	0.000000	0.969206	33.333494	0.000000	False	log
bmi	1	-14454.095370	3.646334	619.521701	0.030794	31.805255	31.998370	False	log
gender	2	-14478.792620	3.760558	644.218951	0.000000	34.641730	33.406934	False	log

  - And we can see that the smoker model is the best, as expected based on the data.

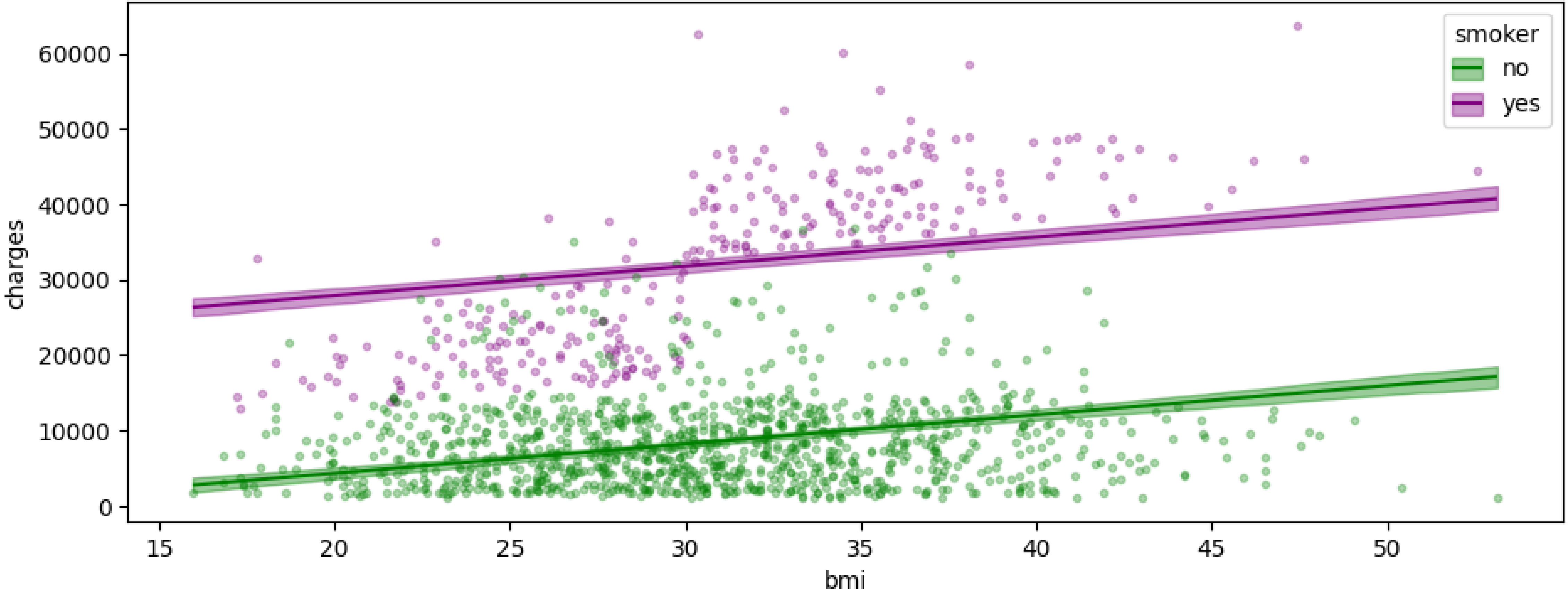
Multiple Regression

- Instead of using only one variable at a time, we can use multiple regression.
- For example, look at the regression of the insurance charges as a function of BMI (continuous) and smoker (categorical).

```
model_sb = bmb.Model("charges ~ smoker + bmi", data)
idata_sb = model_sb.fit(1000, chains = 4, idata_kwargs={"log_likelihood":True})
```

Multiple Regression

- Results:



Multiple Regression

- The results of this model show some relation between charges and bmi, and the difference we saw between smokers and non-smokers.
- The two lines are essentially parallel to each other, with the difference in height showing the amount needed to add to the smokers' charges.

Multiple Regression

- Let's compare the multiple regression model to each of the two simple models:

```
az.compare({"smoker": idata_s, "bmi": idata_b, "multiple": idata_sb})
```

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
multiple	0	-13764.681221	5.418968	0.000000	9.715566e-01	33.350731	0.000000	False	log
smoker	1	-13834.606650	4.683506	69.925429	2.542676e-12	33.346546	11.724257	False	log
bmi	2	-14454.094539	3.615807	689.413318	2.844339e-02	31.768267	33.660494	False	log

- We can see that the multiple regression model is better than the other two.

Interactions

- An interaction effect happens when the effect of an independent variable on the response changes depending on the value of another independent variable.
- For example, is the effect of bmi on charges different for smokers and not smokers?
  - This would be seen in the graph by not parallel regression lines after including the interaction term in the model.

Interactions

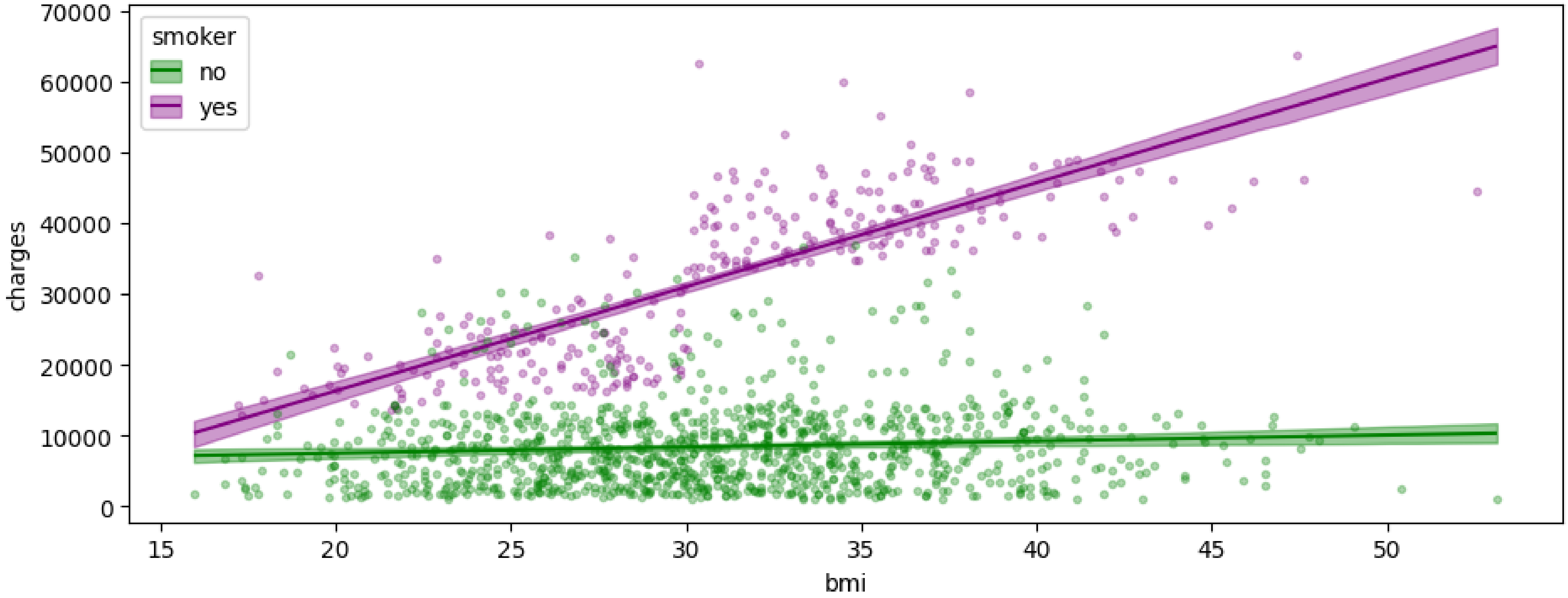
- How does the regression model look in this case?

$$charges = \beta_0 + \underbrace{\beta_1 \cdot BMI + \beta_2 \cdot smoker}_{\text{Main Terms}} + \underbrace{\beta_3 \cdot BMI \cdot smoker}_{\text{Interaction Term}}$$

```
model_sbi = bmb.Model("charges ~ smoker + bmi + smoker:bmi", data)
idata_sbi = model_sbi.fit(1000, chains = 4, idata_kwargs={"log_likelihood":True})
```

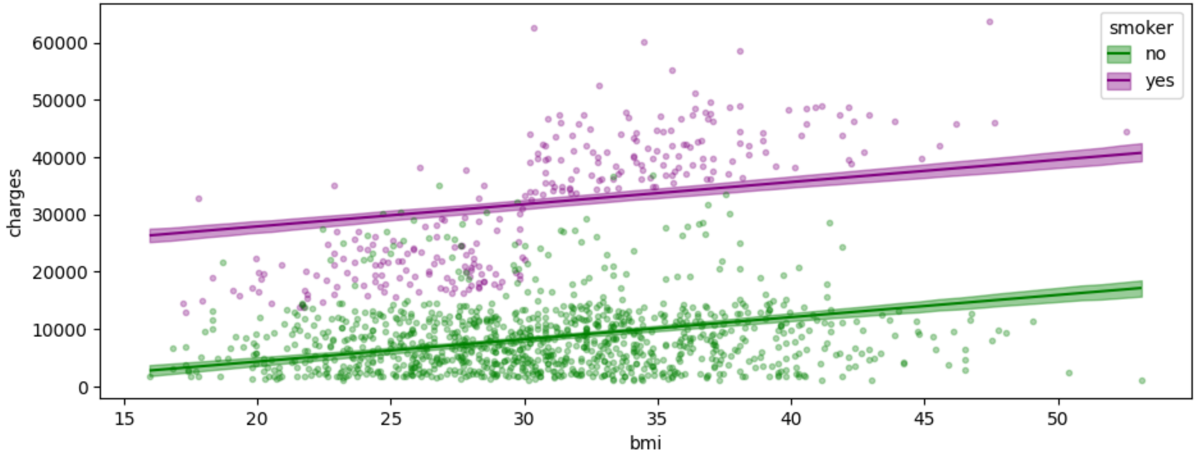
Interactions

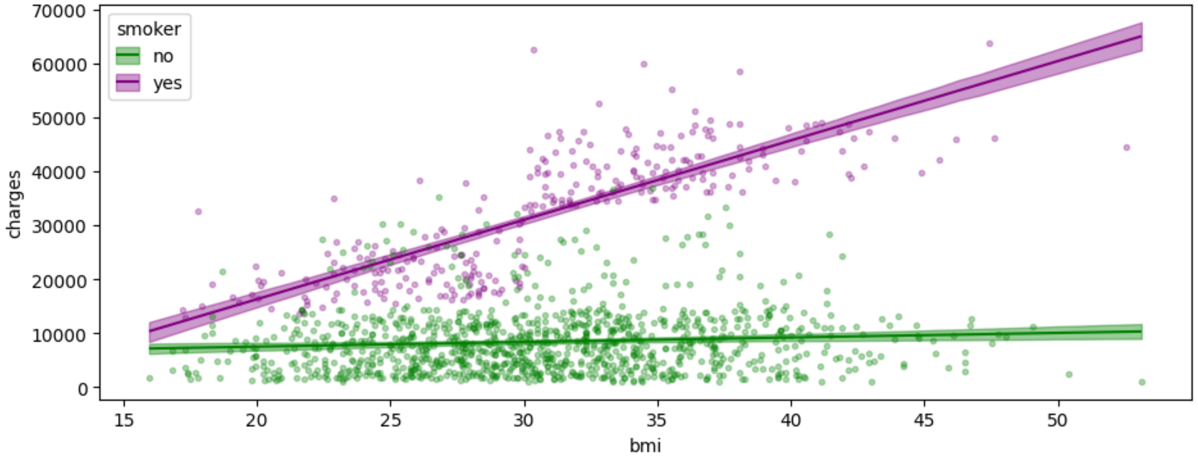
- Looking at the results of this model shows that now the two regression lines are not parallel.



Interactions

- Comparing the results visually of this model and the one without the interaction term show better fit to the data.





Interactions

- Using model comparison also shows that the model with the interaction term is better than the model without.

```
az.compare({"multiple": idata_sb, "interaction": idata_sbi})
```

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
interaction	0	-13577.998320	7.025457	0.000000	0.956279	39.883593	0.000000	False	log
multiple	1	-13764.681221	5.418968	186.682901	0.043721	33.350731	19.294858	False	log

Bayesian Reporting

When to Report Bayesian Statistics?

- Scientific papers
- Grant proposals & pre-registrations
- Ethics / Regulatory approvals
- Medical or engineering reports
- Patents or industrial R&D

Different Reporting Standards & Norms

- **Different publication has different norms:**
  - **Clinical trials** extremely strict (FDA guidance, etc.)
  - **Pre-registered experiments:** more flexible, but must include model-and-prior statements
  - **Grant/IRB applications:** require detailed description of priors, sampling algorithms, diagnostics.
- **Always check the reporting norm for your specific type of publication.**

Three Levels of Reporting

1. Main Document (Methods & Results)
2. Appendix / Supplementary Material
3. Public Repository

Three Levels of Reporting

1. Main Document (Methods & Results)

- Brief model description & equations
- Rationale for prior choice
- Diagnostics (e.g., convergence checks, $\hat{R}$, ESS)
- Methods for model comparison or hypothesis test
- Summary of results:
  - HDI for each important parameter
  - HDI, ROPE (if applicable), and effect size for important comparisons
  - Model comparison table if appropriate

Three Levels of Reporting

2. Appendix / Supplementary Material

- Full model specification: equations and diagram , all priors, hyperparameters, and sampling algorithms.
- Additional diagnostic plots (prior predictive checks; posterior predictive checks; trace & rank plots)
- Results: Table of parameters with the mean, HDI, ESS and MCSE,prior or prior HDI (if applicable).

Three Levels of Reporting

3. Public Repository

- Data (raw or preprocessed), plus instructions to regenerate any derived datasets
- Code: a Jupyter notebook or script that reproduces all steps:
    - Loads dataDefines and fits models (with explicit seed setting for reproducibility)
    - Produces all summary tables and diagnostic figures (saved as PNG/PDF)
    - Generates tables for main text and appendix
- Extra analyses that didn't make the paper (e.g., alternative models, sensitivity to priors)

Example (open attached PDF)

1. Volotsky, Svetlana, Opher Donchin, and Ronen Segev. “The Archerfish Uses Motor Adaptation in Shooting to Correct for Changing Physical Conditions.” Edited by Kunlin Wei and Tamar R Makin. *eLife* 12 (June 3, 2024): RP92909. https://doi.org/10.7554/eLife.92909.

2. Pech, Guillaume P, and Emilie A Caspar. “A Cross-Cultural EEG Study of How Obedience and Conformity Influence Reconciliation Intentions.” *Social Cognitive and Affective Neuroscience* 20, no. 1 (January 18, 2025): nsaf038. https://doi.org/10.1093/scan/nsaf038.