Predicting academic job salary

Langyi Tian

March 2020

Summary

Data preparation:

- No dedup/imputation performed
- Construct new feature to avoid multicolinearity between yrs.since.phd and yrs.service
- 8:2 train/test split, 5-fold CV inside the training set for hyperparameter tuning
- Dummy code all factors; Add 2nd order polynomials on yrs.service; Normalize numeric features

Modeling:

- Salary (logged): a general penalized regression achieving 0.08 RMSE on test set.
- Binary salary indicator: a CART achieving 0.67 accuracy on test set.

Enhancement:

- Questions: causal relationship between gender and salary, interaction salary effect of gender on seniority, subgroup detection
- Additional attributes: citations, PhD at top schools, current academic instition rank, working hours, marriage status
- Sample: size >=45, oversampling high salary observations

Analysis

(1) What percentage of records are Assistant Professors with less than 5 years of experience?

```
## 15.87% of records are Assistant Professors with less than 5 years of experience.
```

(2) Is there a statistically significant difference between female and male salaries?

- Here we want to test the salary against two sex categories. Salary data does not follow a normal distribution in our sample, so we adopt the 2-sample Wilcoxon test.

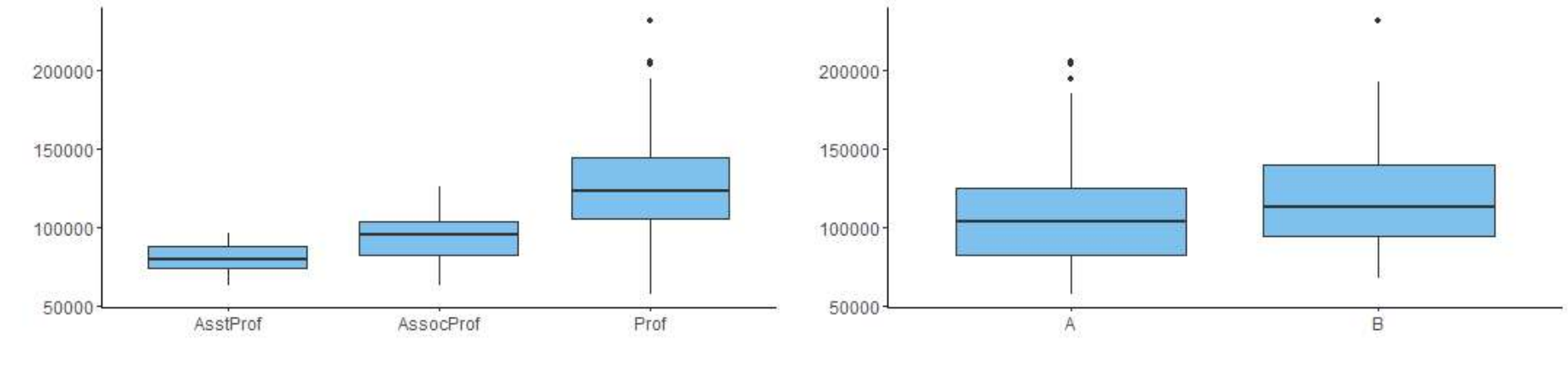
```
## # A tibble: 2 x 2
##   sex    `mean(salary)`
##   <chr>           <dbl>
## 1 Female        101002.
## 2 Male          115090.
```

```
##
##     Wilcoxon rank sum test with continuity correction
##
## data:  salary by sex
## W = 5182.5, p-value = 0.008237
## alternative hypothesis: true location shift is not equal
```

- The p-value is smaller than 0.01, strongly against the null hypothesis that male and female holds equal salary. In other words, female and male salaries are significantly different.

Analysis

(3) What is the distribution of salary by rank and discipline?



- We can see higher tank and discipline B (compared with A) is associated with higher salary.

(4) How would you recode discipline as a 0/1 binary indicator?

```
discipline_B<-(dta$discipline=="B")%>%as.numeric()#dummy code it to avoid redundancy/collinearity. will use dplyr::mu
```

```
## discipline_B
##   0   1
## 181 216
```

Modeling I: Data preparation

- Assume no duplicates due to no unique id column
- No need for missing value imputations as the data is complete
- Years since phd and years of service is strongly correlated.

```
##
##     Pearson's product-moment correlation
##
## data:  dta$yrs.since.phd and dta$yrs.service
## t = 43.524, df = 395, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8909977 0.9252353
## sample estimates:
##       cor
## 0.9096491
```

```
dta<-dta%>%mutate(gap=yrs.since.phd-yrs.service)#manually construct a variable measuring gap between year of phd grad
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -11.0     1.0     3.0     4.7     7.0    30.0
```

Modeling I: Preprocessing

- Use 80% records to train data since we don't have a big sample
- Apply 5-fold CV inside the training set for hyperparameter tuning (to make things faster)
- The preprocessing steps (recipe object in tidymodels framework):

```
## Data Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          5
##
## Operations:
##
## Log transformation on all_outcomes
## Factor variables from all_nominal, -, all_outcomes()
## Dummy variables from all_nominal, -, all_outcomes()
## Orthogonal polynomials on yrs.service
## Centering and scaling for all_predictors, -, all_nominal()
```

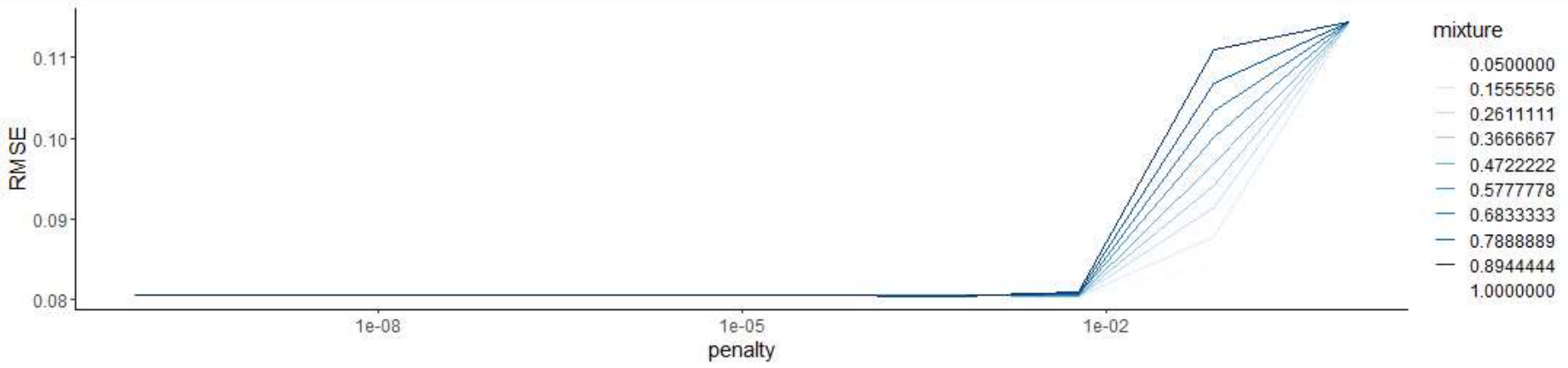
Modeling I: Modeling choice

- Start with a general penalized regression as baseline.
- It's quick, interpratable, and can usually serve as variable selection for subsequent modeling.
- The model definition:

```
## Linear Regression Model Specification (regression)
##
## Main Arguments:
##   penalty = tune()
##   mixture = tune()
##
## Computational engine: glmnet
```

Modeling I: Tuning the best model

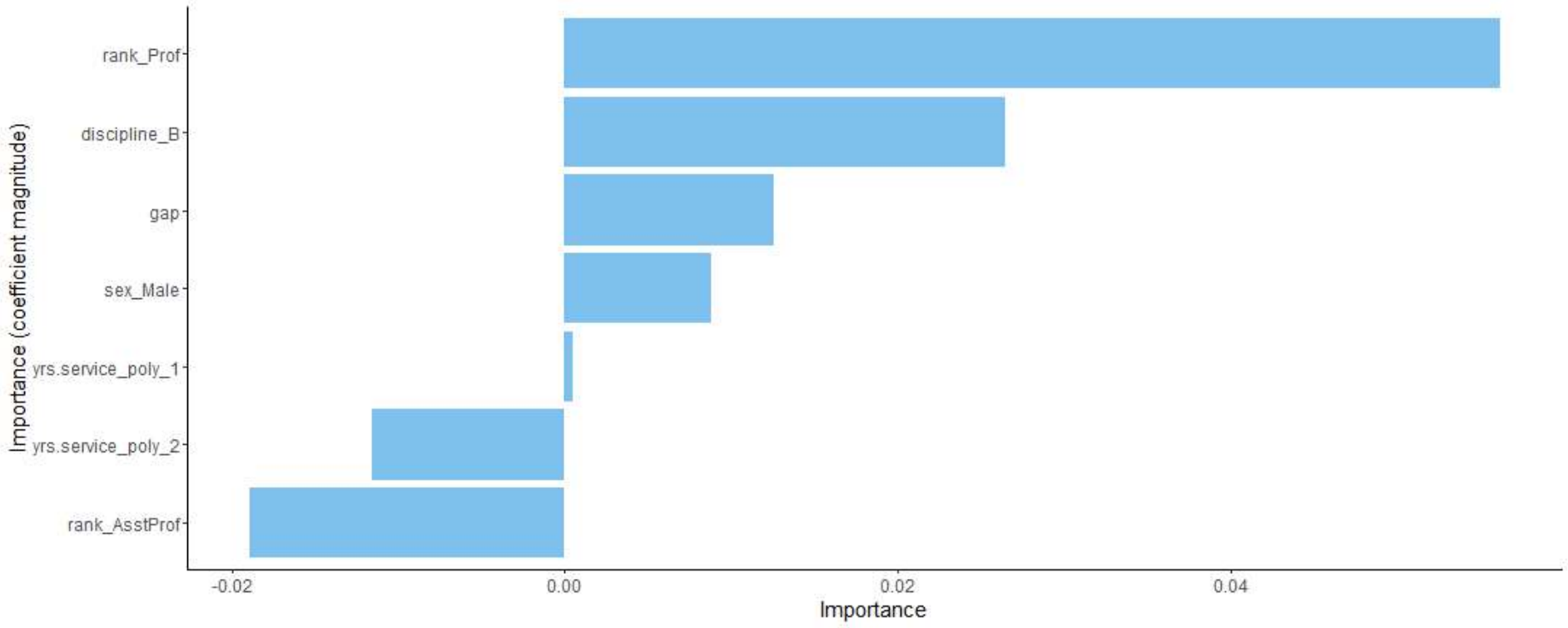
- We use a random grid searching 10 levels for each of mixture and penalty
- We use RMSE as the metric to evaluate performance to give a relatively high weight to large errors since it's usually difficulty to predict them in salary



- The best model:

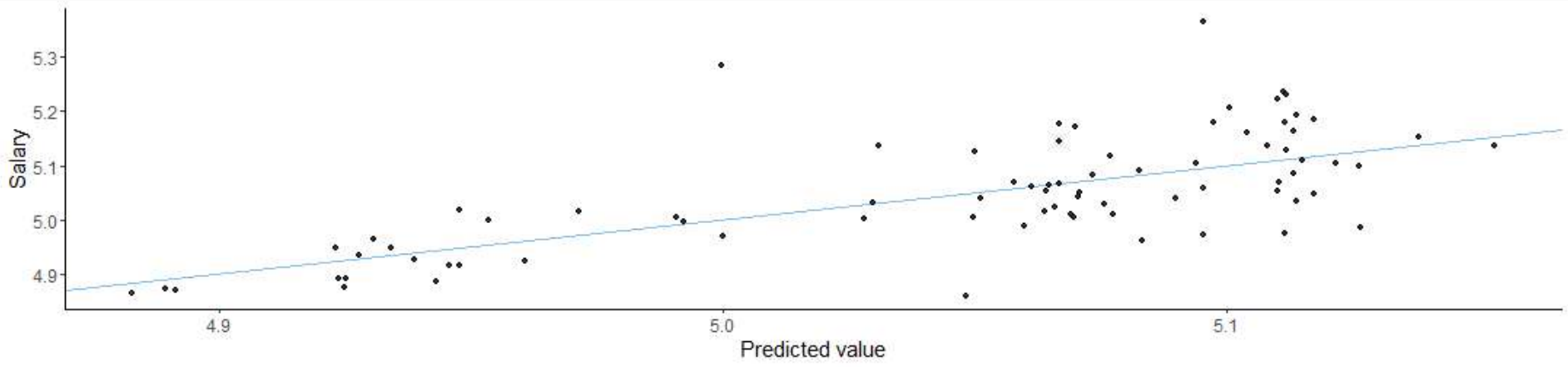
```
## # A tibble: 1 x 2
##   penalty mixture
##     <dbl>   <dbl>
## 1 0.00599   0.261
```

Modeling I: Variable importance



- Being a professor and in discipline B have good chance to predict higher salary
- Being an assistant professorl lead to lower salary
- Quadratic term of years of service has negative coefficient. Multicolinearity?

Modeling I: Residual analysis on test set




```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard      0.0767
```

- It's harder to predict higher salary well.
- In a larger workflow, I will use this as a baseline and try a few more models (e.g. tree algorithms which are less sensitive to extreme values).
- Since this is a showcase example, I'll stop here.

Modeling II: Preprocessing and modeling choice

```
dta <- dta %>% mutate(salary = (salary >= median(dta$salary)) %>% as.numeric()%>% as.factor())#Dummy code the salary
```

```
##
##   0   1
## 198 199
```

- No need for resampling since the response variable is quite balanced
- Same train/test split and preprocessing steps
- For classification I usually start with logistic regressions as baseline
- Since we used glmnet already, let's do a CART instead.

```
## Decision Tree Model Specification (classification)
##
## Main Arguments:
##   cost_complexity = tune()
##   min_n = tune()
##
## Computational engine: rpart
```

Modeling II: Tuning the best model

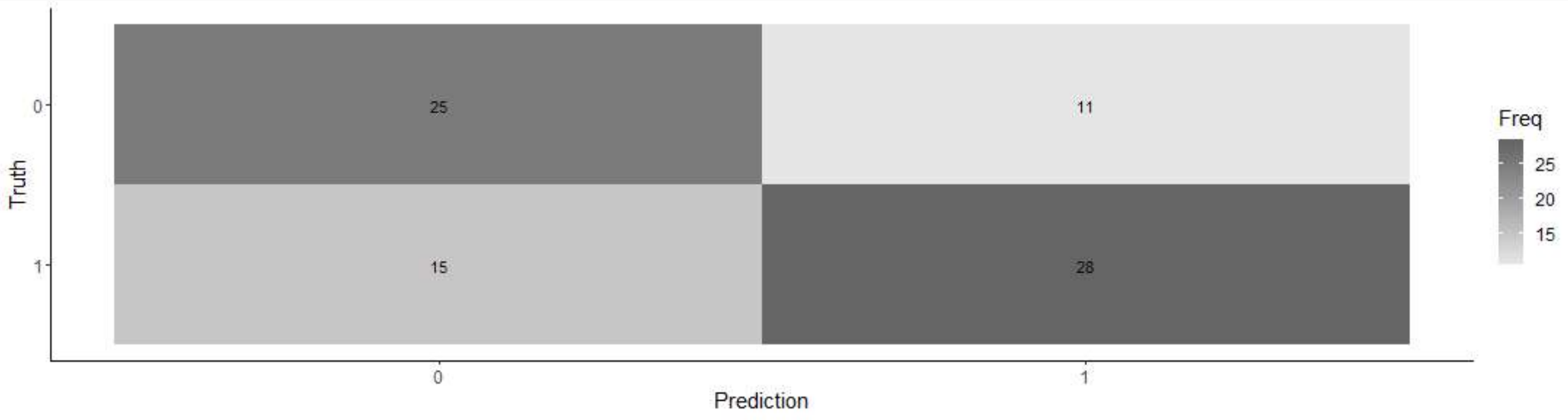
- We use a random grid search for cost complexity and minimum points in a node with 10 levels for each (to save time).
- Use accuracy as the evaluation metric since we don't have a particular weight among 0/1 label

- The best model:

```
## # A tibble: 1 x 2
##   cost_complexity min_n
##             <dbl> <int>
## 1    0.0000000001     2
```

- The n_min is small since our sample size is not big.

Modeling II: Performance on test set




```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.671
```

- The performance has quite a balance between FN and FP.
- However, CART model did worse on the test set which means it overfits. Next step would be to use bagged trees such as RF or GBM to reduce overfitting.

Data Set Enhancement: 3 further questions

1. Is there a causal relationship between gender and salaries among PhDs?

   - From the analysis part, we know that there's a significant salary gap among male and female. However, are female discriminated in the labor market due to their gender identity? In other words, is the gap induced by the gender identify itself rather than other difference (e.g. rank, discipline or seniority)? We can use a linear regression to observe the significance of the gender coefficient controlling other parameters.

2. Will gender identity influence the reward to seniority?

   - The mechanism for gender identity to impact labor market rewards is complicated. In addition to pure discrimination described by the previous case, it's also possible that women with seniority will receive less reward than men with seniority, due to an invisible salary ceiling or inability to work very hard (e.g. family obligations so cannot publish a lot). In other words, there might be an interaction effect between sex and ys.service.

3. What are the subgroups in the academic population?

   - Rather than the male vs. female or B vs. A difference, can we adopt a more holistic way to observe subgroups in the PhDs, who share similarity in their demographics and academic profile? We can achieve this by applying a clustering algorithm to the data (e.g. K-means or K-medoid since the sample size is small and data is mixed).

Data Set Enhancement: 5 additional attributes

As the previous questions targeted salary, it's better to have measurements of other the eterminants of salary to avoid omitted variable bias.

1. Number of publications/citations (e.g. 34 citations for SCI)
    - More publications usually lead to faster promotion and higher salary. If we add it we can determine whether the gender salary gap can rise from the publication difference. Data might be obtained with tools such as Elsevier API if real names are given.
2. PhD at a top school (e.g. Yes)
    - A degree from a top tier school such as Harvard always brings a premium, regardless of the curent place that someone teaches in, as another dimension of the academic profile.
3. Current academic institution rank (e.g. 102th on U.S. News)
    - Teaching in a top tier university versus a community college brings huge difference in career prospect, and we want to put in as a confounding variable to avoid introducing bias in other estimates.

Data Set Enhancement: 5 additional attributes

1. Weekly working hours (e.g. 30 hours)
    - This is a potential proxy for the mechanism of gender gap. If female PhDs are only able to work fewer hours than male, then the significance of the coefficient for gender will likely dissapear after adding working hours.
2. Marriage status (e.g. Unmarried)
    - For a similar reason, marriage might be a proxy for family obligations, since female usually have to spend more time at home after getting married. Adding that will be helpful to clarify the mechanism of gender influence.

Data Set Enhancement: Sample size

We use a power analysis to calculate a minimum requirement for sample size.

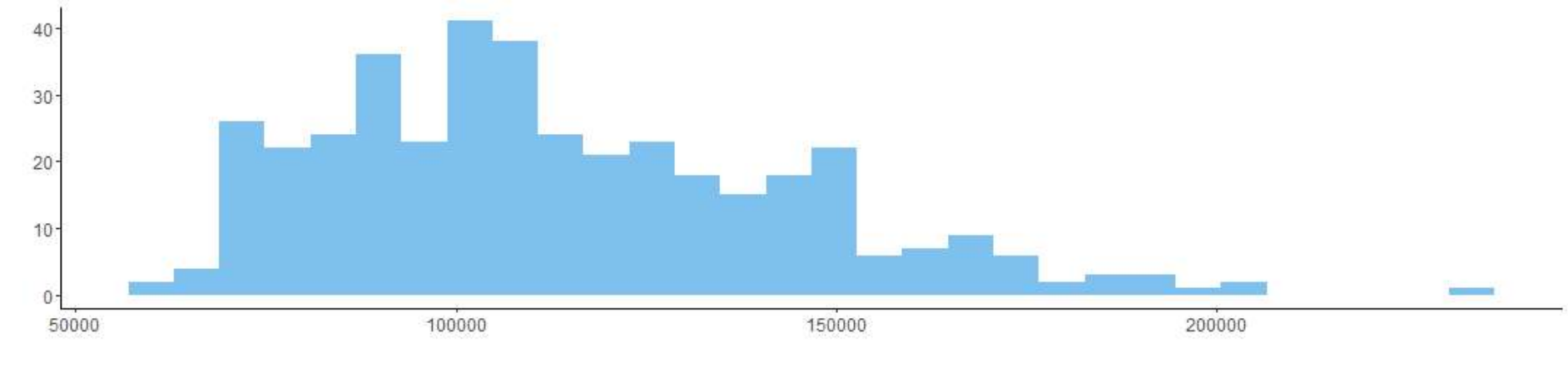
```
library(pwr)
pwr.f2.test(
  u = 10,#10 independent variables (less intercept)
  f2 = 0.5/(1-0.5),#Assume the R2 is 0.5
  sig.level = 0.001,#Need 0.001 significance level
  power = 0.8#Assuming 0.8 power
)
```

```
##
##      Multiple regression power calculation
##
##               u = 10
##               v = 34.05488
##              f2 = 1
##       sig.level = 0.001
##           power = 0.8
```

Therefore, the minimum sample size is v+u+1=45. We are well beyond this threhold.

Data Set Enhancement: Sampling design

However, it's also important to obtain a good estimate of population from the sample, which requires a representative sampling design.



Since most records are in the low to medium tier of the income, it would be the best if we can oversample the high salary group (e.g.>150,000) with techniques such as stratified sampling. Having more records in that group will help to form better estimates.

The case study ends here. Thank you very much for taking the time to read!