# BDA Final Project

This data set has entries with a person's years of experience, age, and salary. This is not my own data set; I obtained it from kaggle: https://www.kaggle.com/codebreaker619/salary-data-with-age-and-experience

This data set is used for machine learning, which I think is important in the data science field. However, I do wonder where the numbers come from, and how the source may have affected my results.

Let's start with making a model. I will also look at the outcome per every variable.

```
salary_data <- read_csv("Salary_Data.csv")
```

```
## Rows: 30 Columns: 3
## ── Column specification
────────────────────────────────────────────────────────
─
## Delimiter: ","
## dbl (3): YearsExperience, Age, Salary
```

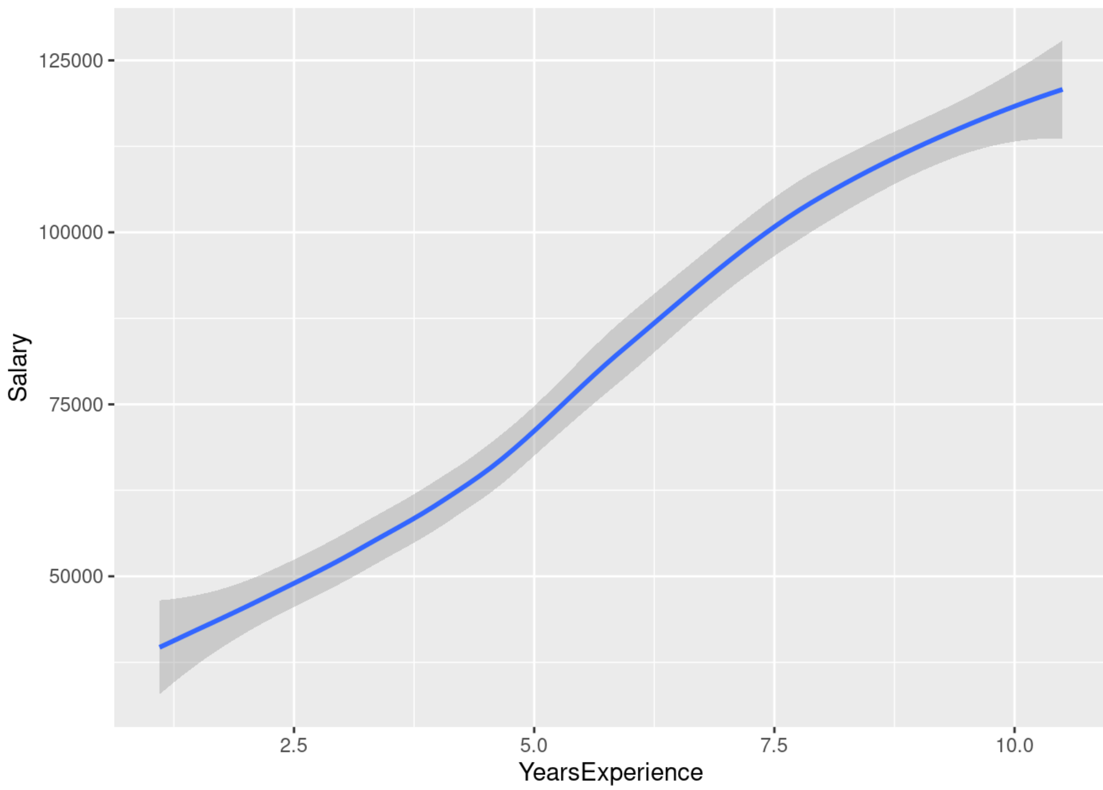
```
##
## ℹ Use `spec()` to retrieve the full column
specification for this data.
## ℹ Specify the column types or set `show_col_types =
FALSE` to quiet this message.
```

```
salary_data %>% glimpse()
```

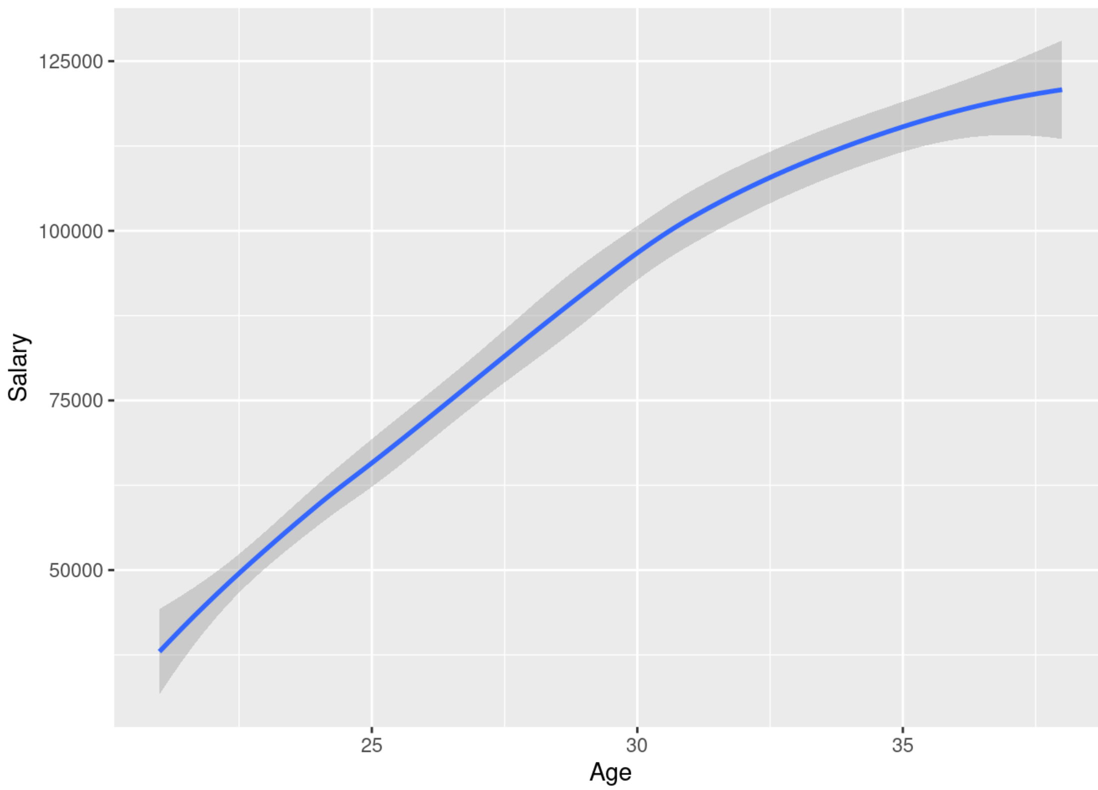
```
## Rows: 30
## Columns: 3
## $ YearsExperience <dbl> 1.1, 1.3, 1.5, 2.0, 2.2,
2.9, 3.0, 3.2, 3.2, 3.7, 3.9,…
## $ Age             <dbl> 21.0, 21.5, 21.7, 22.0,
22.2, 23.0, 23.0, 23.3, 23.3, …
## $ Salary          <dbl> 39343, 46205, 37731, 43525,
39891, 56642, 60150, 54445…
```

```
salary_data %>%
  ggplot(aes(YearsExperience, Salary)) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula
'y ~ x'
```




```
salary_data %>%
  ggplot(aes(Age, Salary)) +
  geom_smooth()
## `geom_smooth()` using method = 'loess' and formula
'y ~ x'
```

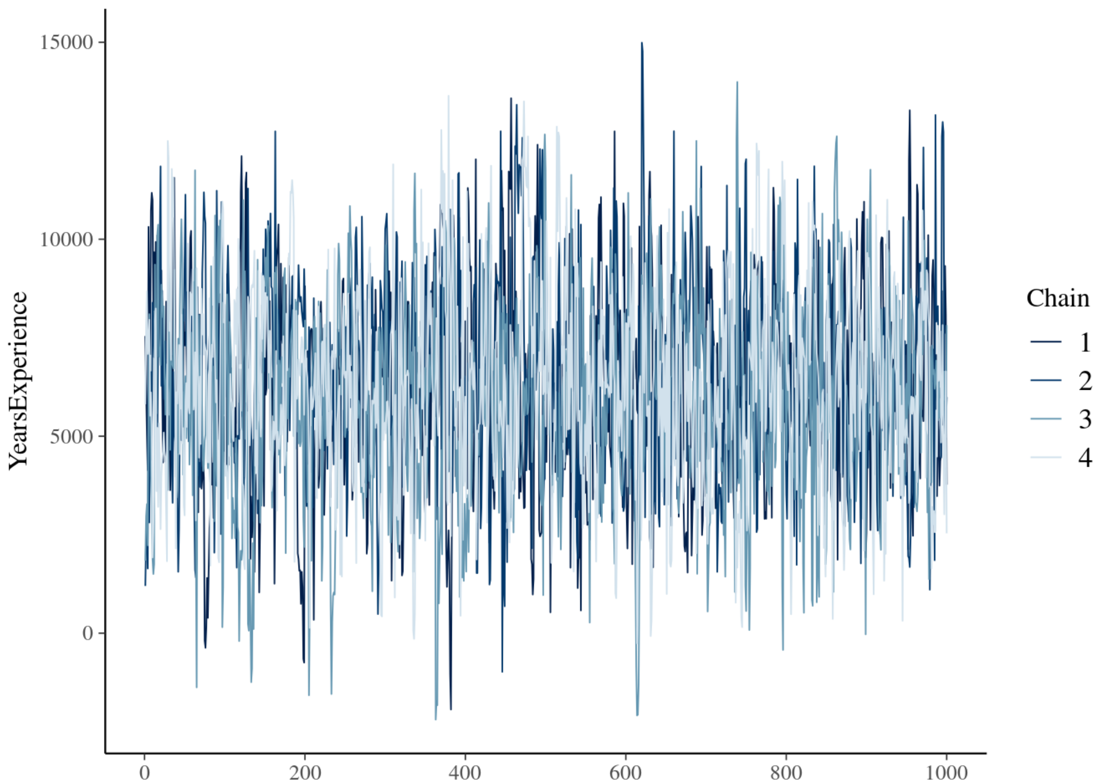



```
salary_stan <- stan_glm(Salary ~ YearsExperience + Age,
data = salary_data)
```

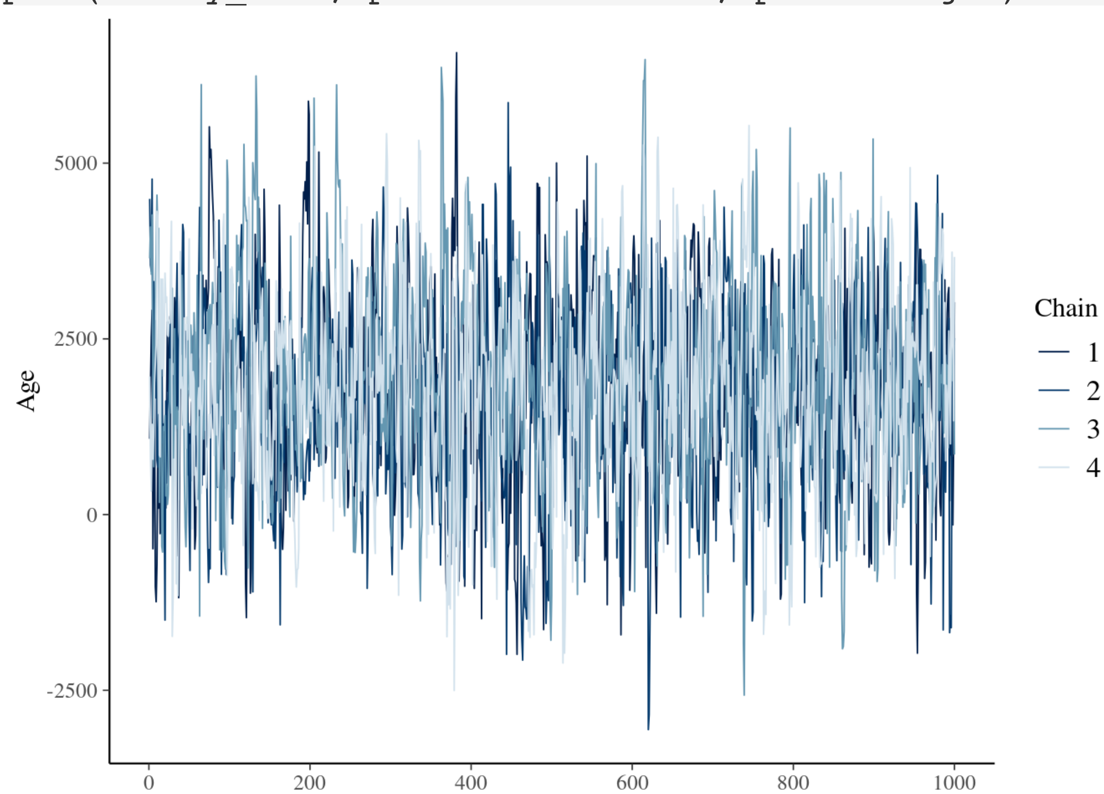
By the plots for every data point separated by the two predictors, we can see that Age and Salary and YearsExperience and Salary have a positive linear relationship.

Now, we can check to see the MCMC process worked well. Let's start with the trace plots.

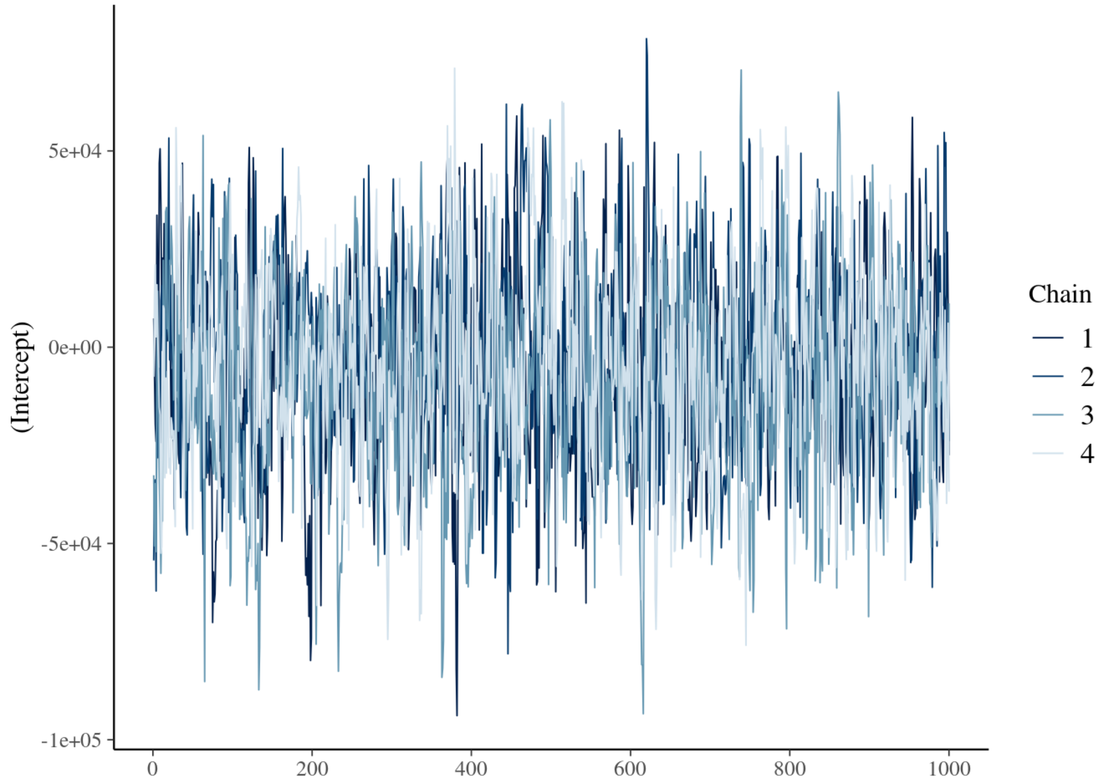
```
plot(salary_stan, plotfun = "trace", pars =
"YearsExperience")
```



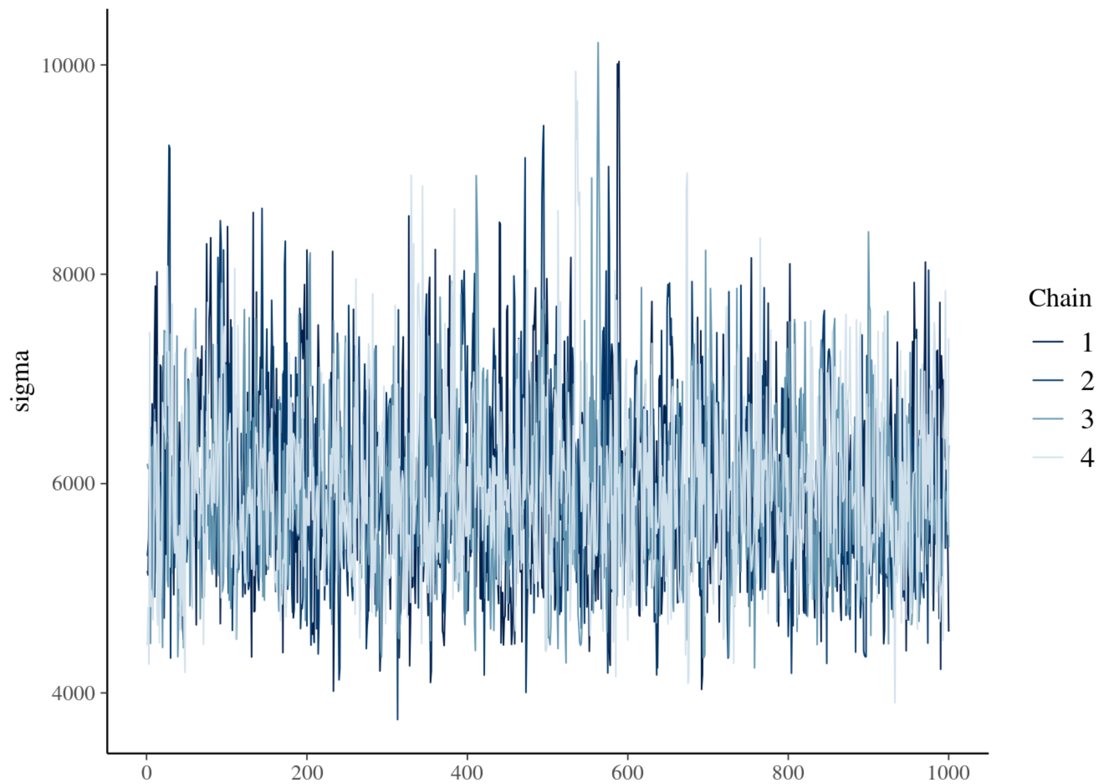

```
plot(salary_stan, plotfun = "trace", pars = "Age")
```



```
plot(salary_stan, plotfun = "trace", pars =
"(Intercept)")
```




```
plot(salary_stan, plotfun = "trace", pars = "sigma")
```




### The trace plots look pretty good. They are not stuck in one place, and go in one direction.

# Now, let’s look at the summary for salary_stan, from which we can see our rhat values.

```
summary(salary_stan, digits=4)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      Salary ~ YearsExperience + Age
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 30
##  predictors:   3
```

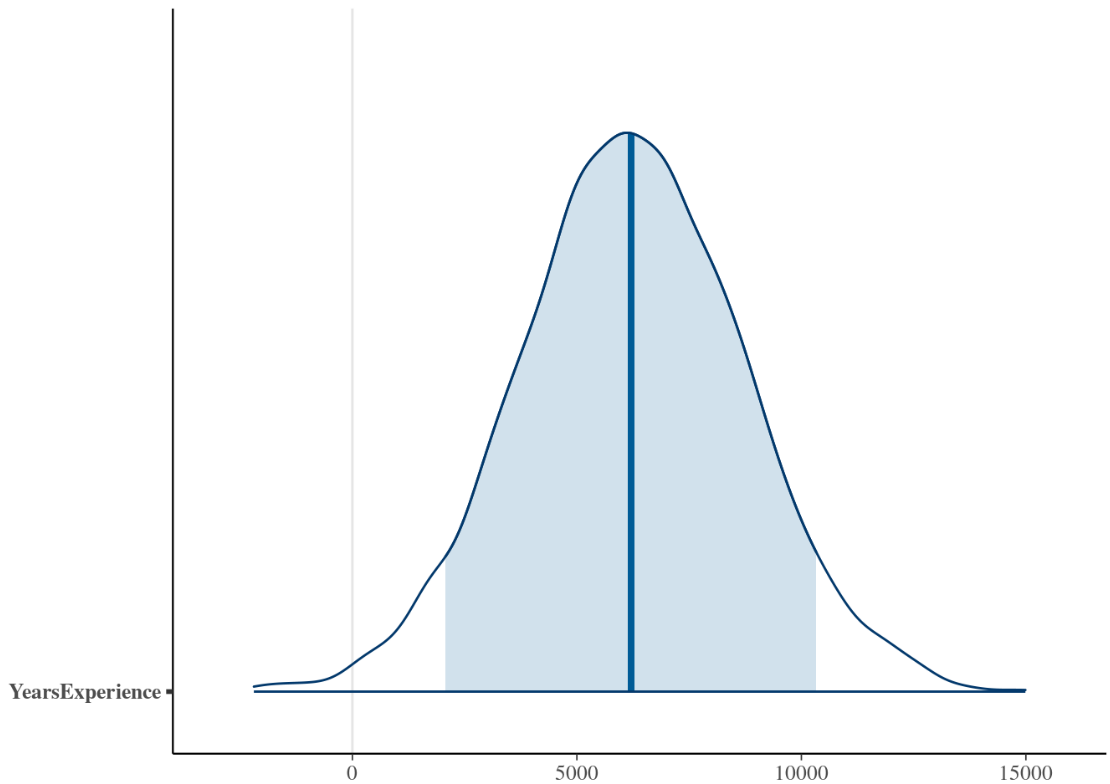
```
##
## Estimates:
##                    mean        sd          10%
50%         90%
## (Intercept)      -5897.7224  24363.5579 -37023.2778
-5560.6061  24846.3697
## YearsExperience   6224.4854   2490.5338   3063.8061
6215.5981   9381.8577
## Age               1793.8914   1371.2861     58.4145
1795.2391   3535.1514
## sigma             5939.6764    846.6408   4943.9079
5851.8679   7070.5199
##
## Fit Diagnostics:
##              mean       sd         10%        50%
90%
## mean_PPD 75981.9317  1546.8589 74009.4016 76020.5833
77915.8973
##
## The mean_ppd is the sample average posterior
predictive distribution of the outcome variable (for
details see help('summary.stanreg')).
##
## MCMC diagnostics
##                  mcse     Rhat     n_eff
## (Intercept)     697.6659   1.0042 1220
## YearsExperience  72.3018   1.0051 1187
## Age              39.7186   1.0047 1192
## sigma            19.8593   1.0008 1817
## mean_PPD         27.3187   1.0006 3206
## log-posterior     0.0440   1.0037 1237
##
## For each parameter, mcse is Monte Carlo standard
error, n_eff is a crude measure of effective sample
size, and Rhat is the potential scale reduction factor
```

```
on split chains (at convergence Rhat=1).
```

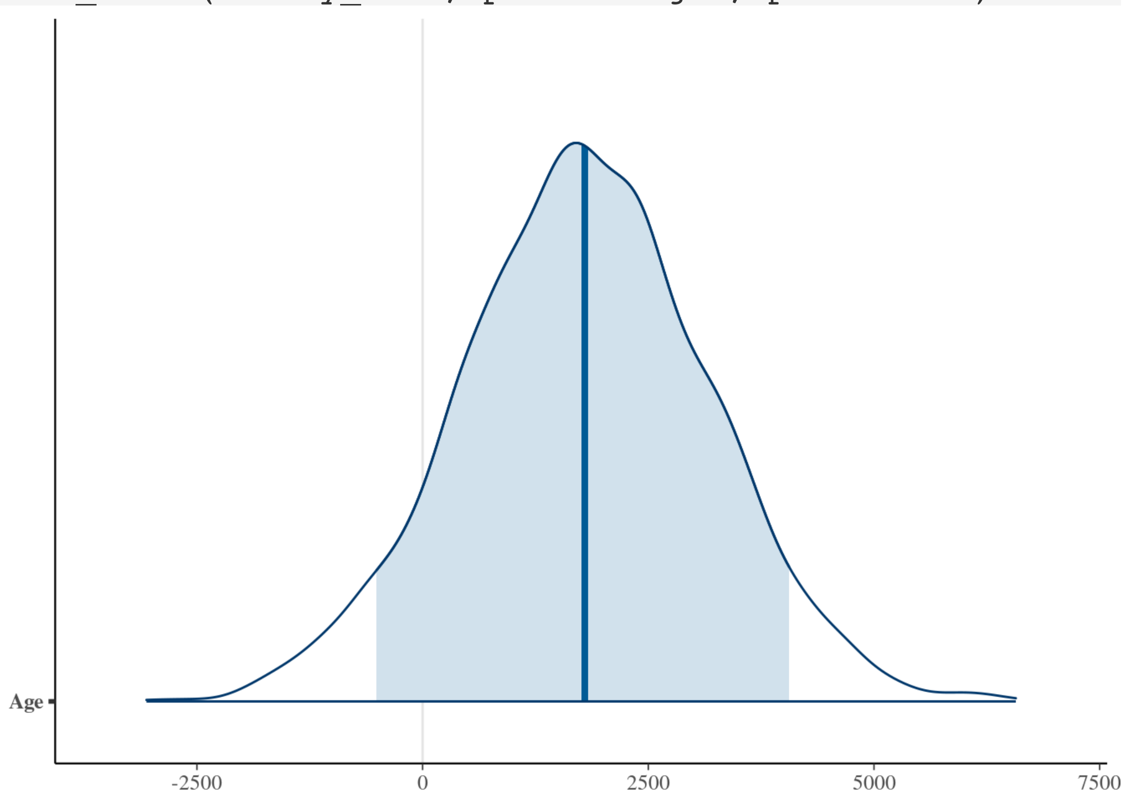
All of the rhat values are larger than one yet smaller than 1.01 and 1.05. This suggests that there is convergence, and we can keep going with our analysis.

Here is a summary of the posteriors, with 90% equal tails credible intervals. I also included a graph of the densities of he regression coefficients for YearsExperience and Age.

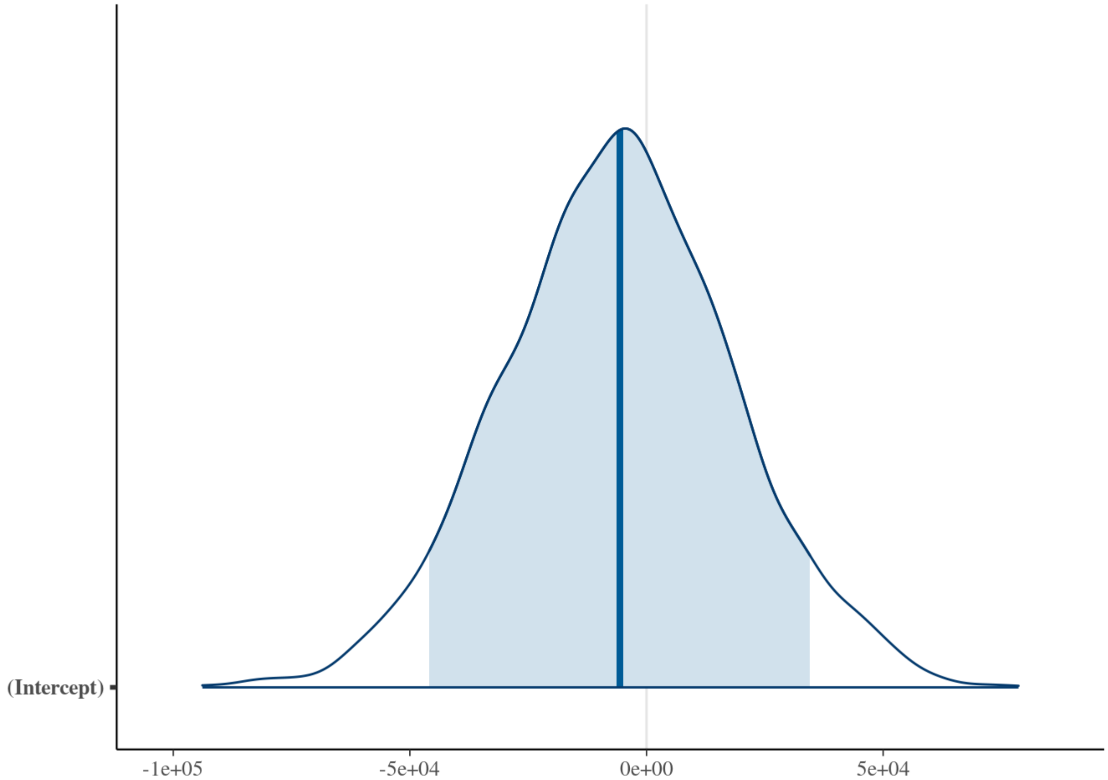
```
mcmc_areas(salary_stan, pars = "YearsExperience", prob
= 0.9)
```



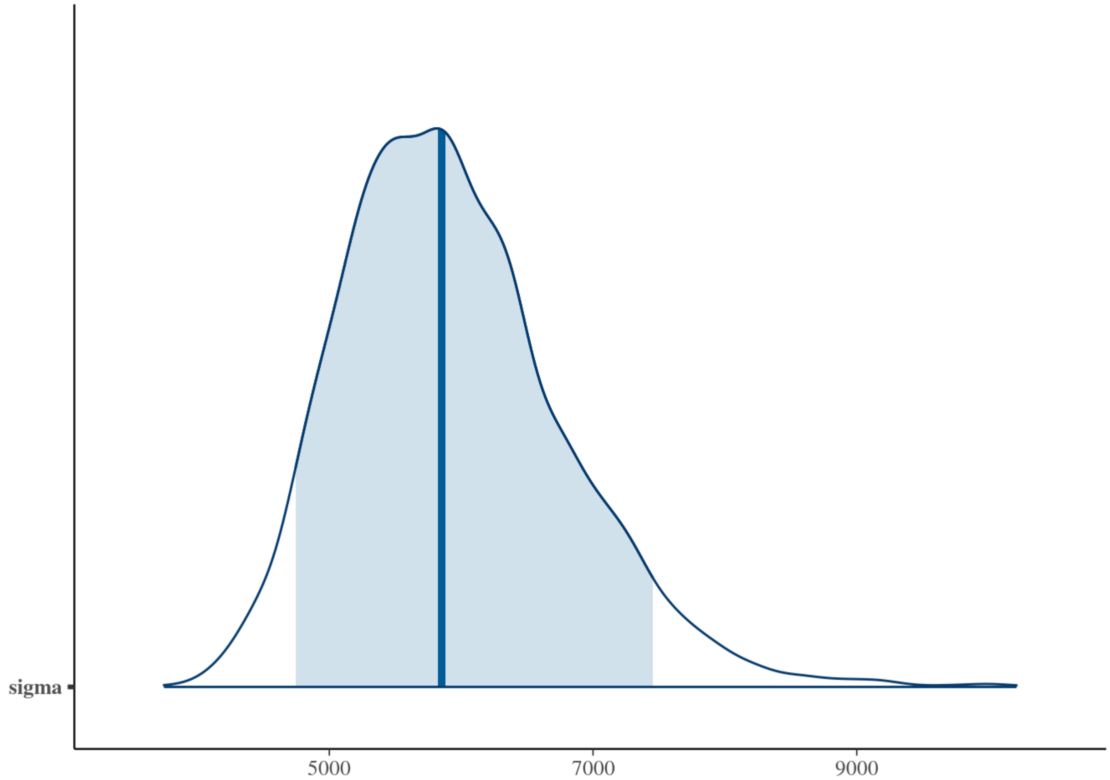

```
mcmc_areas(salary_stan, pars = "Age", prob = 0.9)
```



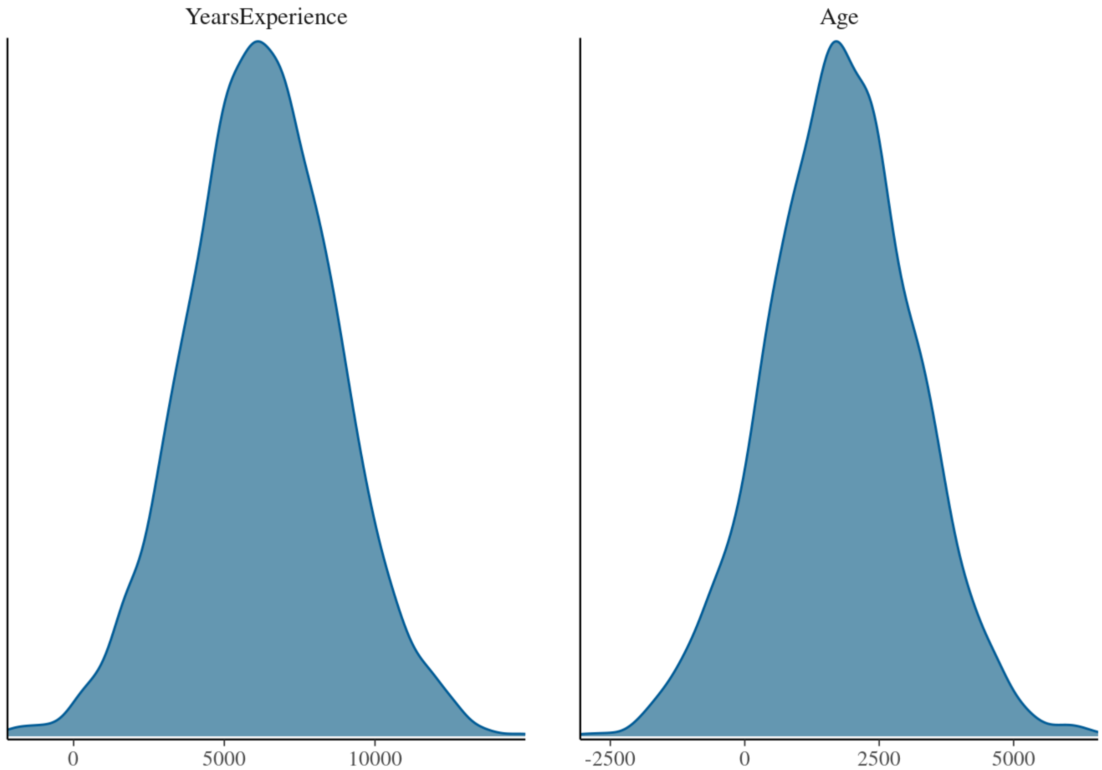
```
mcmc_areas(salary_stan, pars = "(Intercept)", prob =
0.9)
```




```
mcmc_areas(salary_stan, pars = "sigma", prob = 0.9)
```




```
plot(salary_stan, plotfun = "dens", pars =
c("YearsExperience", "Age"))
```

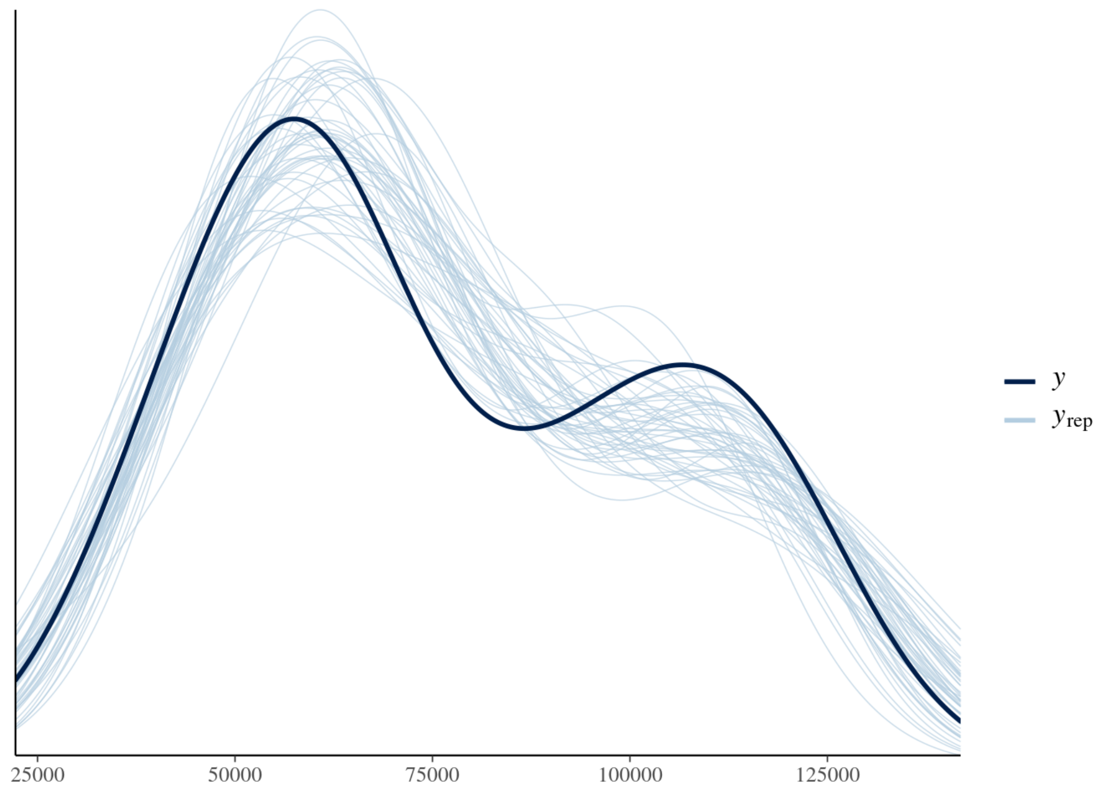



```
describe_posterior(salary_stan, par =
"YearsExperience", ci = .95, centrality = "mean")
## Summary of Posterior Distribution
##
## Parameter       |    Mean |                 95% CI |
pd |                  ROPE | % in ROPE |  Rhat |     ESS
##
-------------------------------------------------------
----------------------------------------------------
## YearsExperience | 6224.49 | [1524.79, 11311.47] |
99.38% | [-2741.44, 2741.44] |     5.18% | 1.005 |
1187.00
describe_posterior(salary_stan, par = "Age", ci = .95,
centrality = "mean")
## Summary of Posterior Distribution
##
## Parameter |    Mean |                 95% CI |     pd |
ROPE | % in ROPE |  Rhat |     ESS
```

```
##
---------------------------------------------------------------------------------------------
## Age          | 1793.89 | [-1000.22, 4445.26] | 90.55% | [-2741.44, 2741.44] |    77.58% | 1.005 | 1192.00
```

# Posterior Predictive Check

```
pp_check(salary_stan)
```




### The Posterior Predictive Check looks okay. The light blue lines seem to follow the shape of the dark blue line. However, I do find the distribution having two humps a bit strange. I wonder if that is due to the relatively small data set.

# Here, I will do some predictions. One prediction will be within the values of the dataset, another will try to go over

the reported values, and another will go below the reported values. Then, I will do a comparison with salary_prediction4 and salary_prediction5: which age will get a higher salary?

```
salary_prediction1 <- # within data set
  tibble(YearsExperience = 3.5, Age = 23.5)
salary_stan %>%
  posterior_predict(newdata = salary_prediction1) %>%
  colMeans()
```

```
##        1
## 58114.02
```

```
salary_prediction2 <- # over reported values of data
set
  tibble(YearsExperience = 11, Age = 40)
salary_stan %>%
  posterior_predict(newdata = salary_prediction2) %>%
  colMeans()
```

```
##        1
## 134286.6
```

```
salary_prediction3 <- # below reported values of data
set
  tibble(YearsExperience = 0.5, Age = 20)
salary_stan %>%
  posterior_predict(newdata = salary_prediction3) %>%
  colMeans()
```

```
##        1
## 33161.37
```

```
salary_prediction4 <- # more experience compared to age
  tibble(YearsExperience = 2, Age = 21)
```

```
salary_stan %>%
  posterior_predict(newdata = salary_prediction4) %>%
  colMeans()
```

```
##        1
## 44102.18
```

```
salary_prediction5 <- # less experience compared to age
  tibble(YearsExperience = 2, Age = 40)
salary_stan %>%
  posterior_predict(newdata = salary_prediction5) %>%
  colMeans()
```

```
##        1
## 78275.82
```

From these results, we can see that the more experience someone has, their salary will increase. Also, the older someone is, then the higher their salary is. When comparing people with the same amount of experience but different ages, the older person will get a higher salary. With that logic, of two people who are the same age, the one with more experience will get a higher salary.

Here, I will find the 90% predictive intervals for the new observations.

```
salary_stan %>%
  predictive_interval(newdata = salary_prediction1)
```

```
##          5%      95%
## 1 47960.99 68125.33
```

```
salary_stan %>%
  predictive_interval(newdata = salary_prediction2)
```

```
##       5%    95%
## 1 122168 146815
```

```
salary_stan %>%
  predictive_interval(newdata = salary_prediction3)
```

```
##        5%      95%
## 1 21923.8 44212.44
```

```
salary_stan %>%
  predictive_interval(newdata = salary_prediction4)
```

```
##          5%     95%
## 1 34191.57 54428.5
```

```
salary_stan %>%
  predictive_interval(newdata = salary_prediction5)
```

```
##          5%      95%
## 1 33883.59 121163.2
```

# I will also create a loo comparison of three models. I will create two new models with one predictor each, and then compare them.

```
salary_stan2 <- stan_glm(Salary ~ YearsExperience, data
= salary_data)
salary_stan3 <- stan_glm(Salary ~ Age, data =
salary_data)
salary_stan_loo <- loo(salary_stan, k_threshold = 0.7)
```

```
## All pareto_k estimates below user-specified
threshold of 0.7.
## Returning loo object.
```

```
salary_stan2_loo <- loo(salary_stan2, k_threshold =
0.7)
```

```
## All pareto_k estimates below user-specified
```

```
threshold of 0.7.
## Returning loo object.
```

```
salary_stan3_loo <- loo(salary_stan3, k_threshold =
0.7)
```

```
## All pareto_k estimates below user-specified
threshold of 0.7.
## Returning loo object.
```

```
loo_compare(salary_stan_loo, salary_stan2_loo,
salary_stan3_loo)
```

```
##              elpd_diff se_diff
## salary_stan2  0.0       0.0
## salary_stan  -0.1       1.3
## salary_stan3 -2.9       3.3
```

This comparison shows that the model using only YearsExperience as a predictor is a better model. However, the standard error and elpd are low for both salary_stan and salary_stan3, so it can be interpreted that all three models work well for the data.

From my analysis, I conclude that Age and YearsExperience have a big impact on Salary. The more experience someone has and/or the older they are, the higher their salary is. From the predictions I made, it seems that age has a bigger influence on salary due to the comparisons I made with salary_prediction4 and salary_prediction5. However, with the loo comparison, salary_stan2loo, which is based on only having YearsExperience as a predictor, was found to be the best model. I think that both can be true. YearsExperience and Age are both strong predictors for Salary. I think it may be difficult to see which may be better because of the connection the two predictors have. The older someone is, the more likely they are to have more experience.