Bayes Simple Linear Regression

Data and EDA

We have data that contains teacher salaries from 2009-2010 for teachers employed by the St. Louis Public School in Michigan. The dataset has been filtered to retain only teachers with at most 15 years of service and who have a full teaching load (FTE). This leaves in total 39 teachers. *I have modified the data slightly to make this analysis more interesting.*

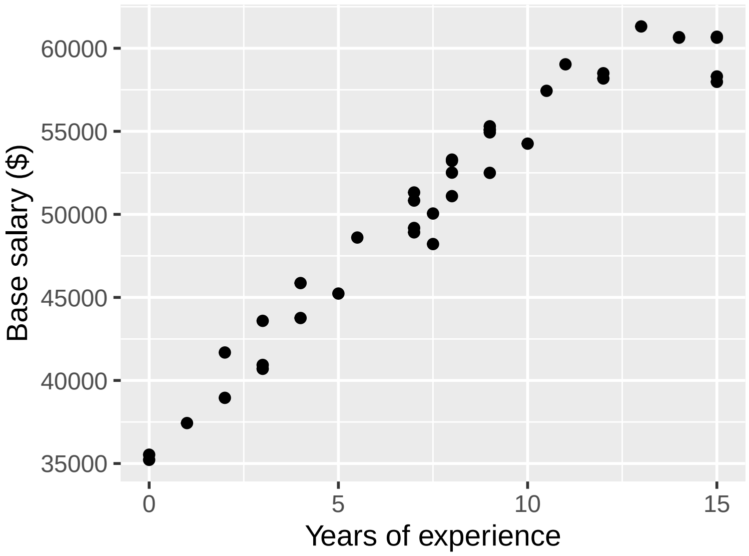
We have the following variables:

- `id`: Identification code for each teacher, assigned randomly
- `degree`: Highest educational degree attained: BA (bachelor's degree) or MA (master's degree)
- `years`: Number of years employed by the school district
- `base`: Base annual salary, in dollars.

We have the following snapshot of the data:

id	degree	fte	years	base
01	BA	1	5	45231.39
02	MA	1	15	60694.91
04	BA	1	10	54257.09
07	BA	1	12	58495.82
11	BA	1	12	58179.38

Suppose I'm a teacher who is moving to Michigan, and I'm interested in learning what kind of salary I could expect to earn. I know that salaries are related in some way to years of experience, so let's do some EDA to see the empirical relationship between years of experience and base salary at this public school:



What do we notice?

Fitting a SLR

I will fit a SLR model to these data, regressing base salary on years of experience (ignoring degree for now). Thus, in context, my SLR looks like the following:

$$
\begin{gathered}
\text{base}_i = \beta_0 + \beta_1 \text{years}_i + \epsilon_i \\
\epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \qquad\qquad i = 1, \dots, n = \text{\`{}}\ rnrow(teachers)
\end{gathered}
$$

Prior solicitation 1

My priors might be specified as follows:

$$
\begin{gathered}
\beta_0 \sim N(0, 100^2) \\
\beta_1 \sim N(0, 100^2) \\
\frac{1}{\sigma^2} \sim \text{Gamma}(1, 1)
\end{gathered}
$$

What's possibly wrong with this prior specification??

Prior solicitation 2

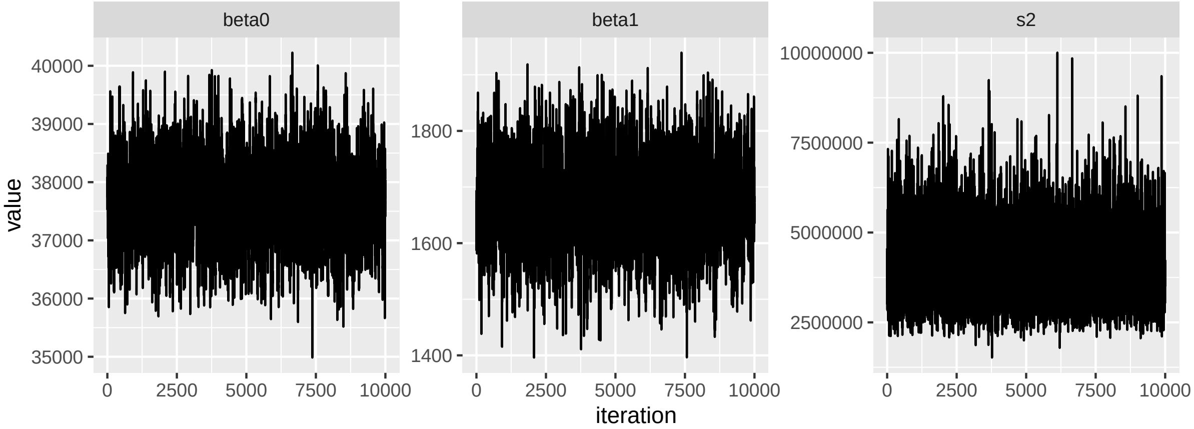
Let's choose the following new priors:

$$\beta_0 \sim N(50000, 10000^2)$$
$$\beta_1 \sim N(0, 10000^2)$$

To approximate the joint posterior distribution, I will run a Gibbs sampler for 20000 iterations, throwing the first half away to burn-in.

Diagnostics

Here are some traceplots:



The effective sample sizes of $\beta_0, \beta_1, \sigma^2$ from these 10000 iterations are: 1340.28, 1291.16, 8044.04. Why do we think some of these are so low?

Model assessment

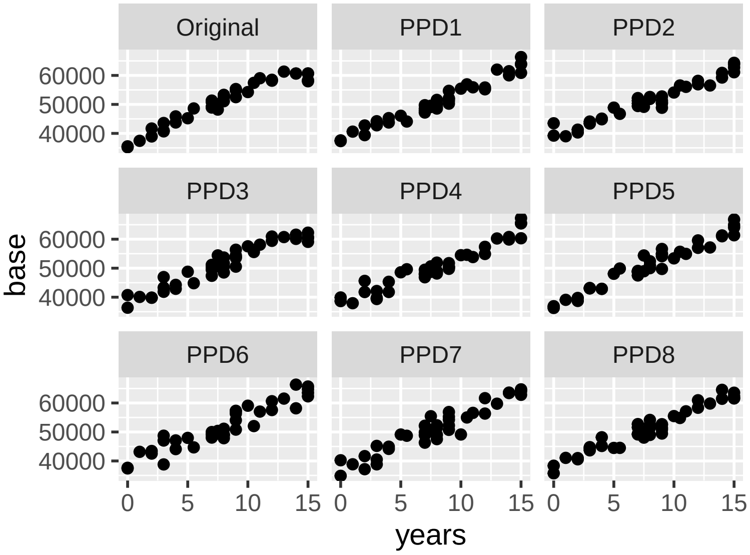
Posterior predictive check

To assess the fit of a SLR model, one approach is to generate PPDs and compare them (visually) to the original data. Once again, we want the observed response values to be consistent with predicted responses generated from the fitted model. So, as before, let's generate some PPDs. For $k \in 1 : K$:

1. Obtain/sample posterior values of $\beta_0, \beta_1, \sigma_2$, call these $\beta_0^*, \beta_1^*, \sigma^{2*}$.
2. Sample $\mathbf{y}^* = \{y_1^*, \ldots, y_n^*\}$ where n is the sample size $n = 39$ and $y_i^* \sim N(\beta_0^* + \beta_1^* x_i, \sigma^{2*})$. Note that these x_i are the same as in the original data, as they are assumed fixed!

```r
set.seed(1)
K <- 8
samp_ids <- sample(1:G, K) # G is number of iterations post burn-in
ppd_ls <- list()
# POSTS is matrix with columns [beta0, beta1, s2]
# y is vector of base salary, x is vector of years of experience
for (k in 1:K){
  ppd_y <- rnorm(n, POSTS[samp_ids[k],1] + POSTS[samp_ids[k],2] * x,
               sqrt(POSTS[samp_ids[k],3]))
  ppd_ls[[k]] <- data.frame(base = ppd_y,
                   years = x,
                   dist = paste0("PPD", k)) # variable for plotting
}
# append on original/true y
ppd_ls[[k+1]] <- data.frame(base = y, years = x, dist = "Original")

do.call(rbind, ppd_ls) |>
  ggplot(aes(x = years, y = base)) +
  geom_point() +
  facet_wrap(~ dist)
```



What do we notice?

Residuals

When fitting a linear regression model, a very common way of assessing model fit is to look at the *residuals*: $e_i = y_i - \hat{y}_i$ where $\hat{y}_i$ is the fitted/estimated value of the i-th response. We should ideally have residuals close to 0. Additionally, residuals shouldn't "look" very different from different values of x_i (i.e. we don't want to do better at predicting certain observations than others).

To do this, as before, for some number K times:

1. Obtain/sample parameter values from the posterior
2. Simulate y_i^* from its corresponding Normal distribution using the posterior values from 1.
3. Obtain $e_i = y_i - y_i^*$

Because the predicted values are random, so are the residuals e_i. *So, our residual plot will look different from the usual, frequentist residual plot.* We can plot credible intervals of the residuals for each observation i! Thus, in this diagnostic, we should choose K large (e.g. the total number of iterations after burn-in) so to get good approximations of the variablity.

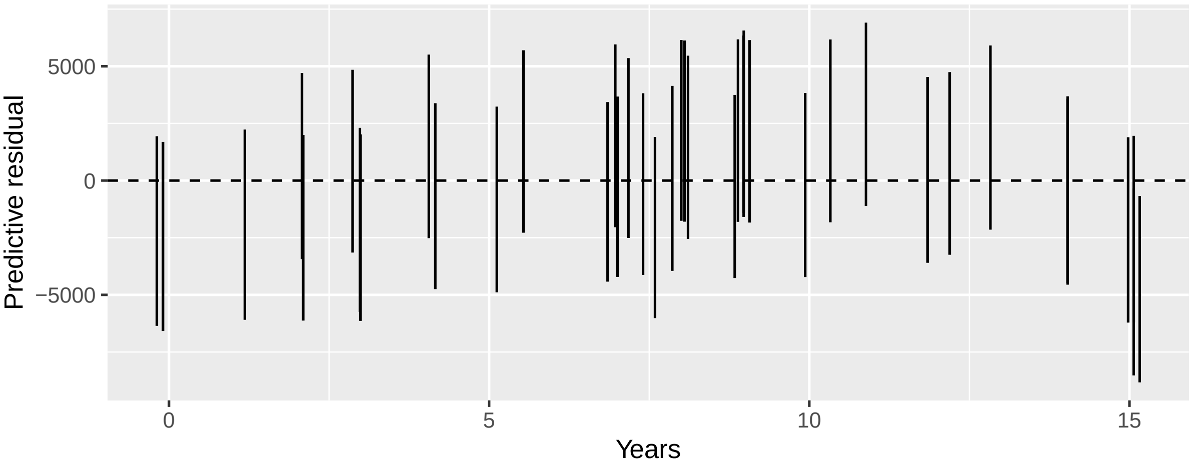
```r
set.seed(1)
resids_ls <- list()
for(i in 1:n){
  y_preds_i <- rnorm(G, POSTS[,1] + POSTS[,2] * x[i], sqrt(POSTS[,3]))
  resids <- y[i] - y_preds_i
  resids_ls[[i]] <- data.frame(resid = resids, obs = i, years = x[i])
}
resids_df_all <- do.call(rbind, resids_ls)
```

Snapshot of `resids_df_all`:

resid	obs	years
-254.1278	1	5
-1330.9272	1	5
1347.9132	1	5

⋮

resid	obs	years
2007.4218	39	5.5
731.1135	39	5.5
4374.1961	39	5.5



95% CI for residual

What do we notice here?

Posterior inference

Regardless of what we determined above about model fit, let's continue with doing some posterior inference.

Posterior summaries

Here are some posterior summaries of the three parameters:

	Posterior mean	2.5%	97.5%
β_0	37689.614	36393.632	38995.523
β_1	1673.672	1527.927	1817.890
σ	1982.952	1596.802	2492.883

Let's give some interpretation!

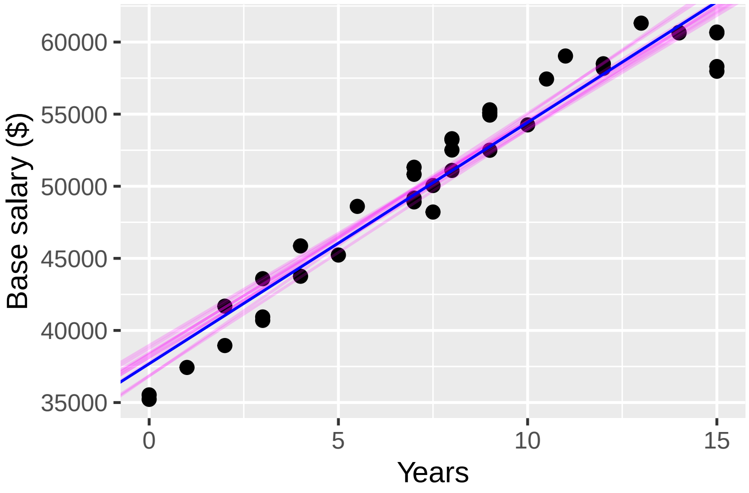
Line of best fit

Let's obtain *a* line of "best fit" (note that *a* is italicized because there is not a notion of a single best line). One might obtain such a line by taking the posterior mean estimates $(\hat{\beta}_0, \hat{\beta}_1)$ of (β_0, β_1), and use them to plot the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

If we use the posterior means as the estimates, then this "best line" represents the most likely value of the line $\beta_0 + \beta_1 x$ from the posterior distribution. To include some notion of uncertainty of this "best line", we might also plot some number of J line estimates where we take samples $(\beta_0^{(j)}, \beta_1^{(j)})$ from the posterior (i.e. Gibbs sampler) and plot $\beta_0^{(j)} + \beta_1^{(j)} x$.

```r
set.seed(1)
J <- 10
iter_samps <- sample(1:G, J)
post_means <- colMeans(POSTS)

ggplot(teachers, aes(x = years, y = base)) +
  geom_point(size=2) +
  geom_abline(data=post_df[iter_samps, ], aes(intercept=beta0, slope=beta1),
              alpha = 0.2, col = "magenta") +
  geom_abline(intercept = post_means[1],
              slope = post_means[2], col = "blue") +
  ylab("Base salary ($)") + xlab("Years") +
  labs(caption = "Blue line: a line of best fit \n Pink lines: posterior samples of line")
```



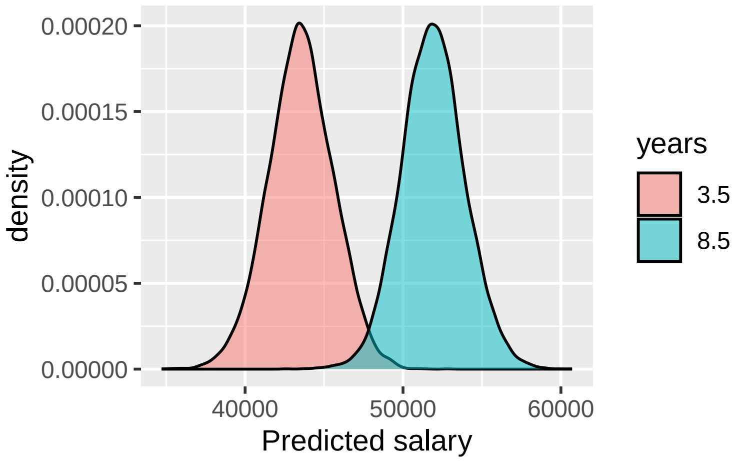
Blue line: a line of best fit
Pink lines: posterior samples of line

What do we notice?

Predicted responses

I have approximately 3.5 years of teaching experience and my friend has approximately 8.5 years. I might like to know what sort of ranges of base salaries we might expect to earn. To answer this question, why not obtain the posterior density of base salaries for each of $x^{(new)} = 3.5$ and $x^{(new)} = 8.5$? Do the following for $k = 1, \dots K$ (once again, with K large):

1. Use the posterior samples of β_0, β_1 to simulate the expected earnings at a given level $x^{(new)}$ of years of experience
2. Randomly sample the earnings at this mean using the Normal model and a posterior sample of σ^2


density
Predicted salary
years