Homework 09 Solutions

11/10/2021

```r
library(knitr)
library(pander)
library(nlme)
library(ggplot2)
library(dplyr)
library(scales)
```

Problem 1

(a)

We see in Figure 1 that plotting the mean for each state is helpful in summarizing the overall income level within each stage but does not give an indication of the relationship between income and high school graduation rates. Relative to the other plots Figure 1 is easy to interpret but is probably not the best way to convey the information it is displaying.

```r
dat <- read.table("../cdi.dat")

mean.df <- group_by(dat, state) %>%
    summarise(per.cap.income = mean(per.cap.income))

ggplot(dat) +
    geom_point(aes(pct.hs.grad, per.cap.income)) +
    geom_hline(data = mean.df, aes(yintercept = per.cap.income),col="blue") +
    facet_wrap(~state) +
    scale_y_continuous(label = comma) +
    xlab("Percentage High School Graduates") + ylab("Per Capitca Income ($)") +
    theme_bw()
```

(b)

Figure 2 allows us to examine if the relationship between high school graduation rate and income are similar to what is observed at a national level but it is hard to see what the state level trends are. Relative to the others plots Figure 2 is best suited the answering the question of “Does state X look like nation as a whole?’ ’ instead of conveying state level results or tends.

```r
fit <- lm(per.cap.income ~ pct.hs.grad, data = dat)
ggplot(dat, aes(pct.hs.grad, per.cap.income)) +
    geom_point() +
    geom_abline(intercept = fit[["coefficients"]][1],
                slope = fit[["coefficients"]][2],
                col="blue") +
    facet_wrap(~state) +
    scale_y_continuous(label = comma) +
    xlab("Percentage High School Graduates") + ylab("Per Capitca Income ($)") +
```

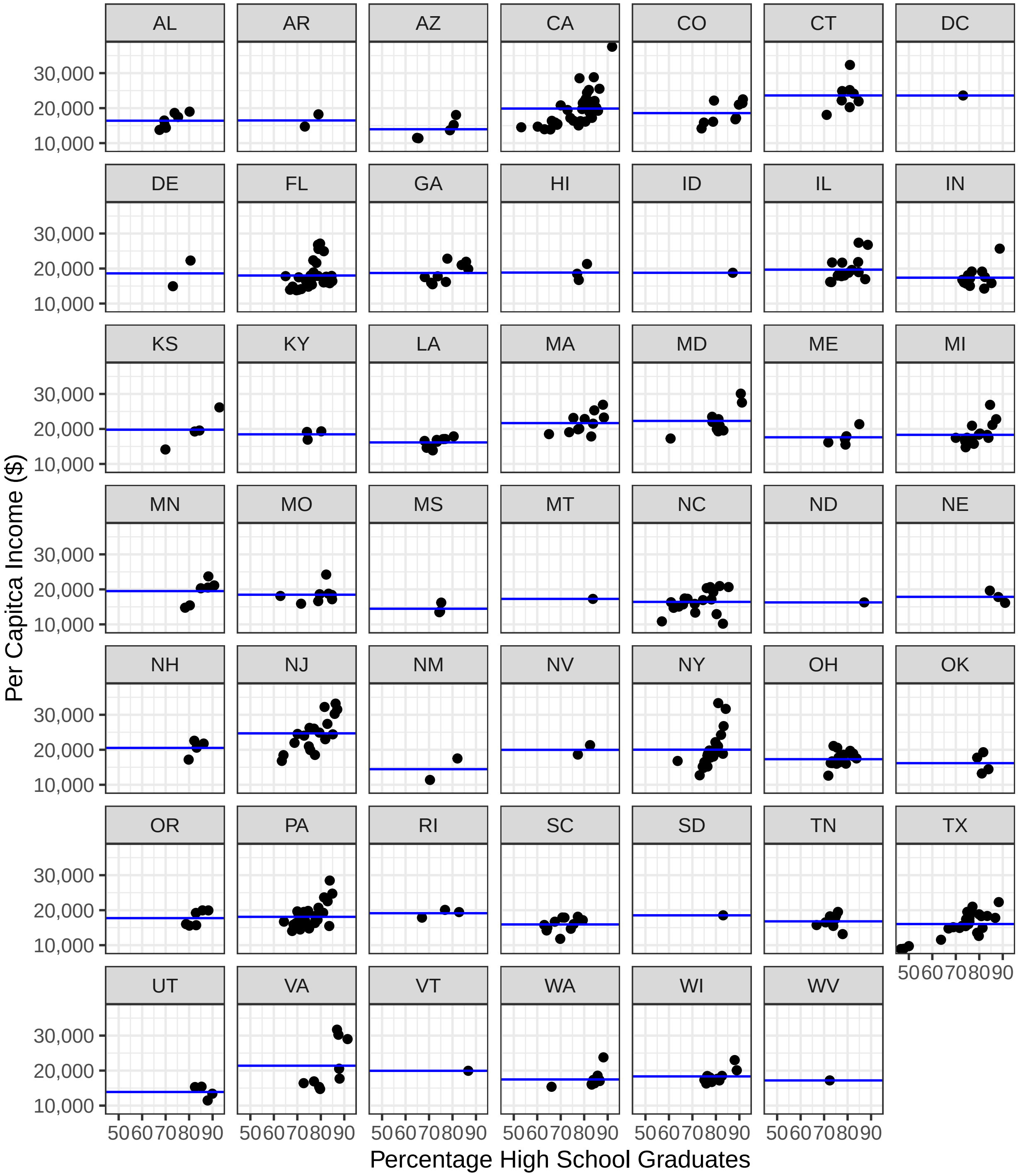

Figure 1: Problem 1a: HS graduation rate vs. county level per capita income with state level mean per capita income line.

```
    theme_bw()
```

(c)

Figure 3 allows us to examine if the relationship between high school graduation rate and income are similar to what is observed at a national level while adjusting for mean income in the state but makes it harder to see how the overall state income level compares to the nation as a whole. Relative to the others plots Figure 2 is best suited the answering the question of “Does the relationship between high gradation and income in state X look like nation as a whole?’’ but does not convey much about the mean income level as done in Figure 1 and Figure 2.

```
fit2 <- lm(per.cap.income ~ -1 + state + pct.hs.grad, data = dat)
slopes.df <- data.frame(state = substr(names(fit2[["coefficients"]][1:48]), 6, 7),
                       state.intercept = fit2[["coefficients"]][1:48],
                       slope = fit2[["coefficients"]][49])

ggplot(dat, aes(pct.hs.grad, per.cap.income)) +
    geom_point() +
    geom_abline(data = slopes.df, aes(intercept = state.intercept, slope = slope),
                col="blue") +
    facet_wrap(~state) +
    scale_y_continuous(label = comma) +
    xlab("Percentage High School Graduates") + ylab("Per Capitca Income ($)") +
    theme_bw()
```

(d)

We see in Figure 4 that the state level regression lines match the observed data, as closely as possible, but we are unable to fit them for all states since some states contain only a single county. In contrast to the other plots there is much greater variance in the fitted regression lines perhaps better describing each state but Figure 4 fails to convey any country level trends.

```
state_lm <- function(df){
    mod <- lm(per.cap.income ~ pct.hs.grad, data = df)
    return(data.frame(intercept = mod[["coefficients"]][1], slope = mod[["coefficients"]][2]))
}
lm.df <- group_by(dat, state) %>% do(., state_lm(.))

ggplot(dat, aes(pct.hs.grad, per.cap.income)) +
    geom_point() +
    geom_abline(data = lm.df, aes(intercept = intercept, slope = slope),
                col="blue") +
    facet_wrap(~state) +
    scale_y_continuous(label = comma) +
    xlab("Percentage High School Graduates") + ylab("Per Capitca Income ($)") +
    theme_bw()
```

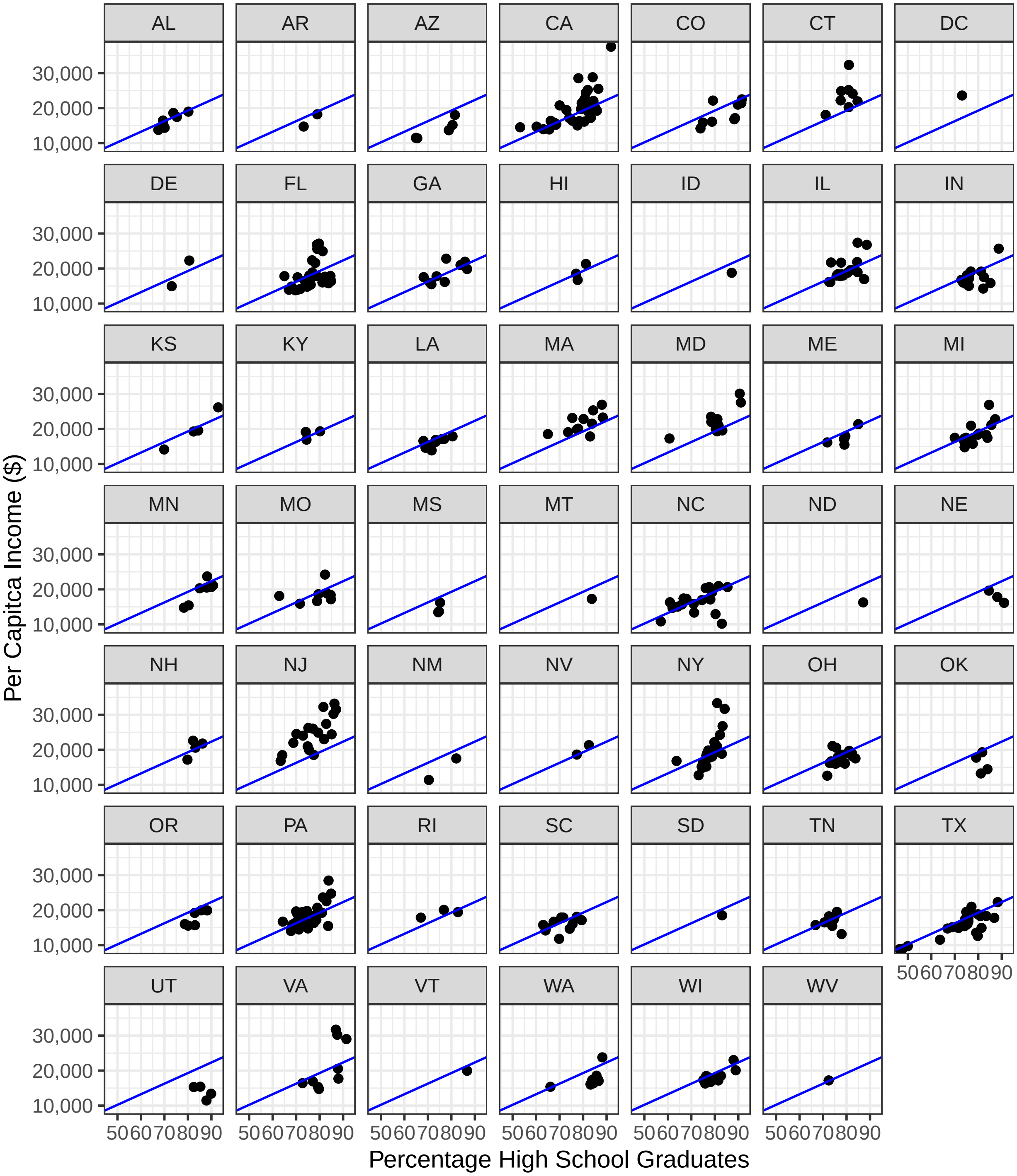

Figure 2: Problem 1b: HS graduation rate vs. county level per capita income with national level regression line.

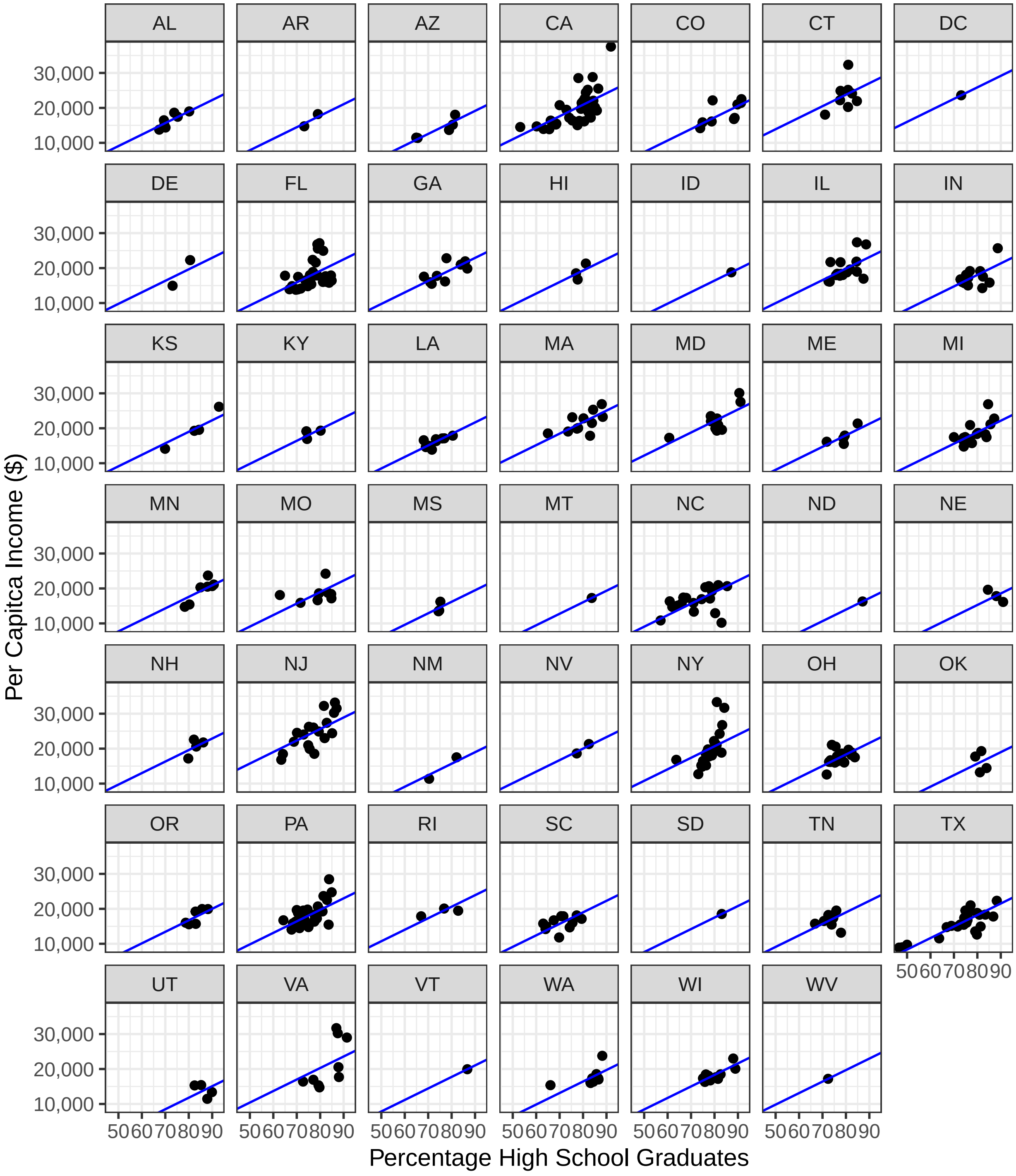

Figure 3: Problem 1c: HS graduation rate vs. county level per capita income with state level intercept and national slope line.

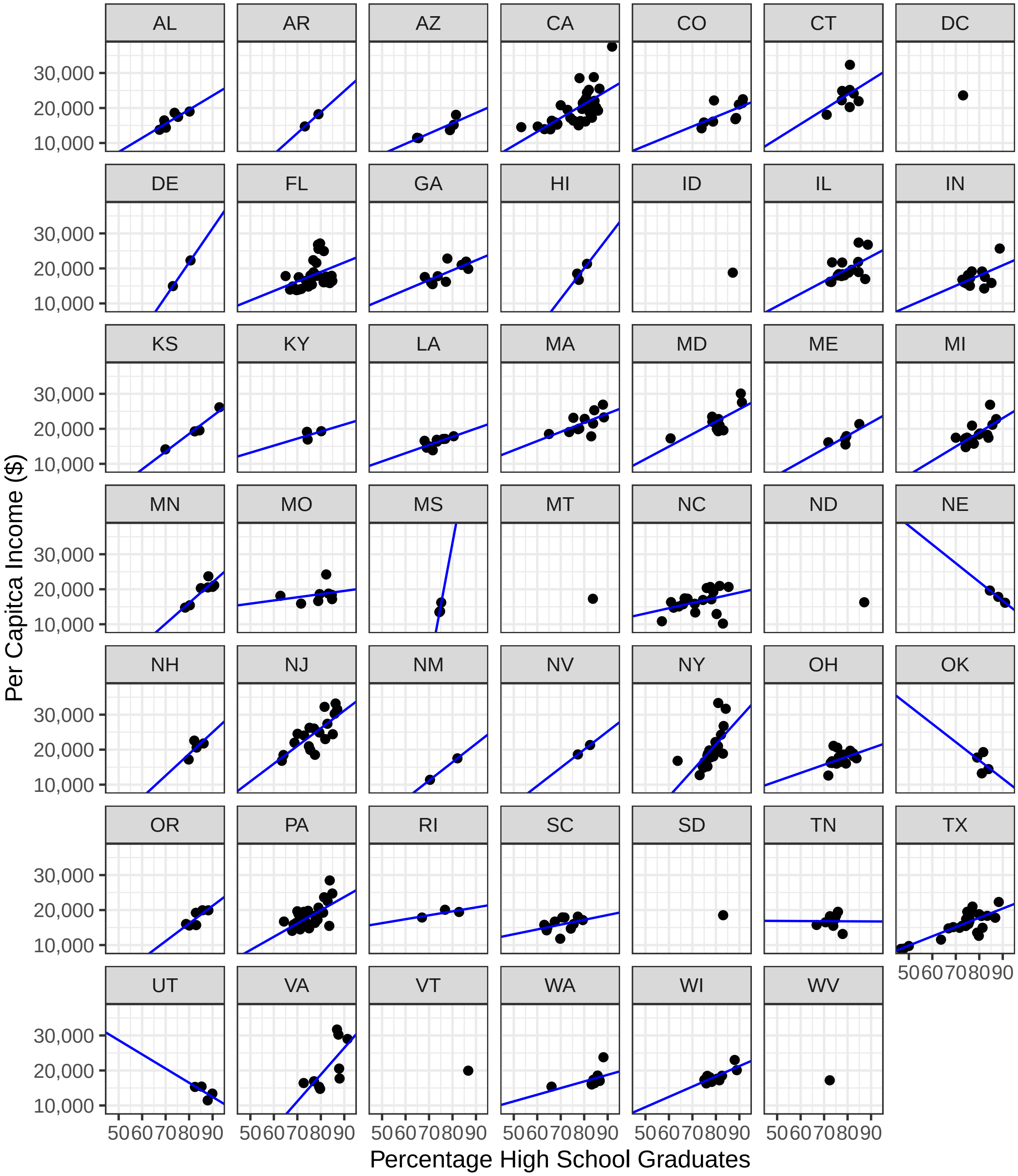

Figure 4: Problem 1d: HS graduation rate vs. county level per capita income with state level regression line.

Problem 2

We need to assume throughout this problem that the errors ϵ_i are independent of the quantities η_j. (Remember, ϵ_i represents the "unexplainable variation".) Also, since $\text{Cov}(a, X) = \text{Var}(a) = 0$ if a is a constant and X is a random variable, we will omit the constant β_0 from all of the calculations below.

(a)

Since, for two random variables A, B, we have that $\text{Corr}(A, B) = 0 \iff \text{Cov}(A, B) = 0$, let's just focus on the covariance. We have:

$$\begin{aligned}\text{Cov}(y_i, y_{i'}) &= \text{Cov}(\alpha_{j[i]} + \epsilon_i, \alpha_{j[i']} + \epsilon_{i'}) &(1)\\ &= \text{Cov}(\alpha_{j[i]}, \alpha_{j[i']}) + \text{Cov}(\alpha_{j[i]}, \epsilon_{i'}) + \text{Cov}(\epsilon_i, \alpha_{j[i']}) + \text{Cov}(\epsilon_i, \epsilon_{i'}) &(2)\end{aligned}$$

Recall that if two random variables A and B are independent, then $f(A)$ and $g(B)$ are independent for any functions f and g[1]. Since $\eta_{j[i]} \perp\!\!\!\perp \eta_{j[i']}$, and β_0 is just a constant, it follows that $\alpha_{j[i]} \perp\!\!\!\perp \alpha_{j[i']}$

Furthermore, we have the errors ϵ_i are independent of everything, by assumption.

Hence each pair of variables in (2) is independent, so the covariances are all 0.

(b)

Substituting $j[i]$ for $j[i']$ in (2), we have

$$\begin{aligned}\text{Cov}(y_i, y_{i'}) &= \text{Cov}(\alpha_{j[i]}, \alpha_{j[i]}) + \text{Cov}(\alpha_{j[i]}, \epsilon_{i'}) + \text{Cov}(\epsilon_i, \alpha_{j[i]}) + \text{Cov}(\epsilon_i, \epsilon_{i'}) &(3)\\ &= \text{Var}(\alpha_{j[i]}) + 0 + 0 + 0 &(4)\\ &= \tau^2 &(5)\end{aligned}$$

Now,

$$\text{Corr}(y_i, y_{i'}) = \frac{\text{Cov}(y_i, y_{i'})}{\sqrt{\sigma^2_{y_i}\sigma^2_{y_{i'}}}} \qquad (6)$$

Writing the model in variance components form, we have $y_i = \beta_0 + \eta_{j[i]} + \epsilon_i$, which means $y_i \sim N(\beta_0, \sigma^2 + \tau^2)$ (since the sum of two independent normally distributed random variables has variance equal to the sum of the variances). Substituting this into (6) yields

$$\begin{aligned}\text{Corr}(y_i, y_{i'}) &= \frac{\text{Cov}(y_i, y_{i'})}{\sqrt{(\sigma^2 + \tau^2)(\sigma^2 + \tau^2)}} &(7)\\ &= \frac{\tau^2}{\sigma^2 + \tau^2} &(8)\end{aligned}$$

(c)

$$\begin{aligned}\frac{1}{n_j}\sum Y_i &= \frac{1}{n_j}[\sum \alpha_j + \sum \epsilon_i] &(9)\\ &= \alpha_j + \frac{1}{n_j}\sum \epsilon_i &(10)\\ \implies \text{Var}(\bar{y}_j) &= \text{Var}(\alpha_j) + \frac{1}{n_j^2}\sum \text{Var}(\epsilon_i) &(11)\\ &= \tau^2 + \frac{\sigma^2}{n_j} &(12)\end{aligned}$$

[1] Any measurable functions f and g, that is, but you don't really need to worry about this technicality.

(d)

Notice that this is equivalent to sampling a group, sampling a bunch of data points from that group, splitting those data points into two sub-groups, and taking the means of each subgroup. We have:

$$\begin{aligned}
\mathrm{Cov}(\bar{y}_j, \bar{y}_j^*) &= \mathrm{Cov}\left(\frac{1}{n_j}\sum_i y_i, \frac{1}{n_j^*}\sum_k y_k^*\right) && (13)\\
&= \frac{1}{n_j^2}\sum_i\sum_k \mathrm{Cov}(y_i, y_k^*) && (14)\\
&= \frac{1}{n_j^2} n_j^2 \tau^2 && (15)\\
&= \tau^2 && (16)
\end{aligned}$$

where in line 2 we assumed that $n_j^* = n_j$, and in line 3 we used that $\mathrm{Cov}(y_i, y_k^*)$ is equivalent to $\mathrm{Cov}(y_i, y_{i'})$ from part (b) above. Now, using the result from part (c), we have

$$\begin{aligned}
\mathrm{Cov}(\bar{y}_j, \bar{y}_j^*) &= \frac{\tau^2}{\sqrt{\sigma_{\bar{y}}^2 \sigma_{\bar{y}^*}^2}} && (17)\\
&= \frac{\tau^2}{\tau^2 + \sigma^2/n_j} && (18)
\end{aligned}$$