Homework 08 Solutions

11/21/2019

2 (a)

We first fit a model regression Sales on all of the covariates except Advert and Lag1Advert.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	350.4	25.98	13.48	2.69e-22
Time	0.4298	0.2381	1.805	0.07485
Month_2	-18.4	31.81	-0.5785	0.5645
Month_3	77.54	31.8	2.438	0.01698
Month_4	101.9	31.8	3.203	0.001951
Month_5	93.43	31.79	2.939	0.004306
Month_6	89.75	31.79	2.823	0.005999
Month_7	24.82	31.79	0.7808	0.4372
Month_8	89.02	31.79	2.8	0.006403
Month_9	166.8	31.79	5.247	1.236e-06
Month_10	199.7	31.8	6.279	1.658e-08
Month_11	374.3	32.84	11.4	1.954e-18
Month_12	1151	32.83	35.05	2.62e-50

Table 2: Problem 2a: Summary of ordinary linear model.

Observations	Residual Std. Error	R^2	Adjusted R^2
93	61.43	0.9634	0.9579

The diagnostic plots for our first model are shown in Figure 1. Our standard set of diagnostic plots, the first four plots, don't look too bad although there are some deviations from normality in the QQ plot and maybe an upward trend in the Scale-Location plot. We see in the ACF plot that there is, statistically significant, auto-correlation in the lag1 term for the residuals. We also plot the residuals against the Time variable, which could conceivably be used to address the auto correlation but we do not observe any trends in this plot. We could estimate a model to account for the observed auto correlation but we will defer that until the next part since it is hypothesize that advertising in both the current and previous month may be important variables, and could conceivably account for the auto correlation. We do note that the current model explains a large portion of the observed variation and is straight forward to interpret relying only on the month of the year and the Time (cumulative count of months).

2 (b)

We try adding the Advert and Lag1Advert

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	392.4	40.81	9.614	6.994e-15
Advert	2.3	2.574	0.8934	0.3744

	Estimate	Std. Error	t value	Pr(>|t|)
Lag1Advert	-4.517	2.663	-1.697	0.09374
Time	0.4322	0.2377	1.818	0.07285
Month_2	-45.88	37.83	-1.213	0.2288
Month_3	26.69	40.59	0.6576	0.5127
Month_4	81.05	33.55	2.416	0.01803
Month_5	59.33	36.74	1.615	0.1104
Month_6	59.36	36.42	1.63	0.1072
Month_7	-12.4	37.42	-0.3314	0.7412
Month_8	60.02	35.91	1.671	0.09865
Month_9	129	37.24	3.464	0.0008671
Month_10	165.7	35.43	4.677	0.000012
Month_11	324.8	43.59	7.451	1.081e-10
Month_12	1149	39.48	29.11	1.209e-43

Table 4: Problem 2b: Summary of ordinary linear model.

Observations	Residual Std. Error	R^2	Adjusted R^2
93	60.55	0.9653	0.9591

We can see from the model outputs that neither Advert or Lag1Advert is statistically significant. Conducting an ANOVA test (results below) we see that we fail to reject the null hypothesis that the model without either Advert or Lag1Advert performs as well as the larger model. Examining the diagnostic plots for the larger model in Figure 2 we also the same issues as with the smaller model, namely: autocorrelation in the residuals and a slight trend in the scale-location plot. Given the diagnostics we would prefer the model from part a) but might choose several methods for improving it. We could try fitting a more general model to allow for the autocorrelation in the residuals. Alternatively, we could remove or transform the Time variable yielding a smaller. Since this variable is a just a linear trend it is possible that this could address the autocorrelation in the residuals observed in both models.

```
## Analysis of Variance Table
##
## Model 1: Sales ~ (Advert + Lag1Advert + Time + Month_2 + Month_3 + Month_4 +
##     Month_5 + Month_6 + Month_7 + Month_8 + Month_9 + Month_10 +
##     Month_11 + Month_12) - Advert - Lag1Advert
## Model 2: Sales ~ Advert + Lag1Advert + Time + Month_2 + Month_3 + Month_4 +
##     Month_5 + Month_6 + Month_7 + Month_8 + Month_9 + Month_10 +
##     Month_11 + Month_12
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     80 301843
## 2     78 285962  2     15881 2.17   0.12
```

3 (a)

A summary of the model is below. Diagnostic plots are in Figure 3, and the estimated autocorrelation between the residuals is plotted in Figure 4.

The residuals are very roughly normally distributed, but they appear to show an increasing trend relative to the fitted values. The scale-location plot also appears to show that the variance is increasing with fitted

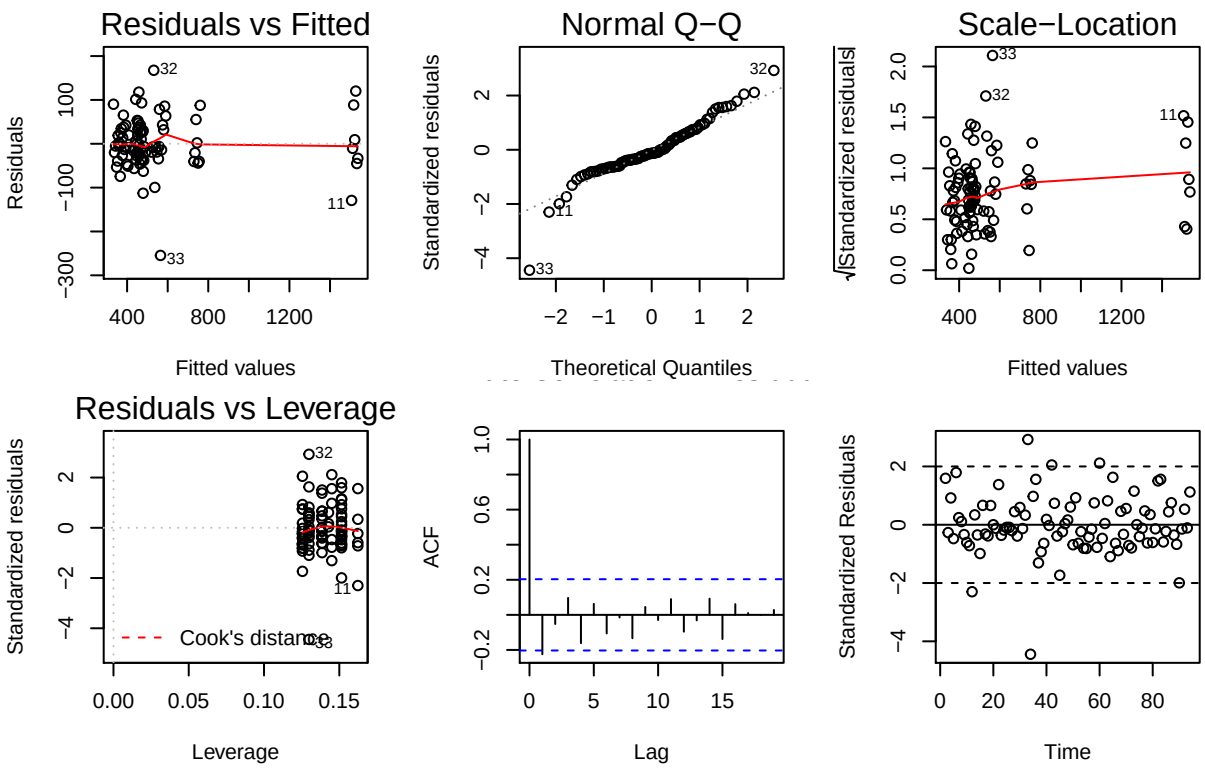

Figure 1: Problem 2a: Diagnostic plots for ordinary linear model.

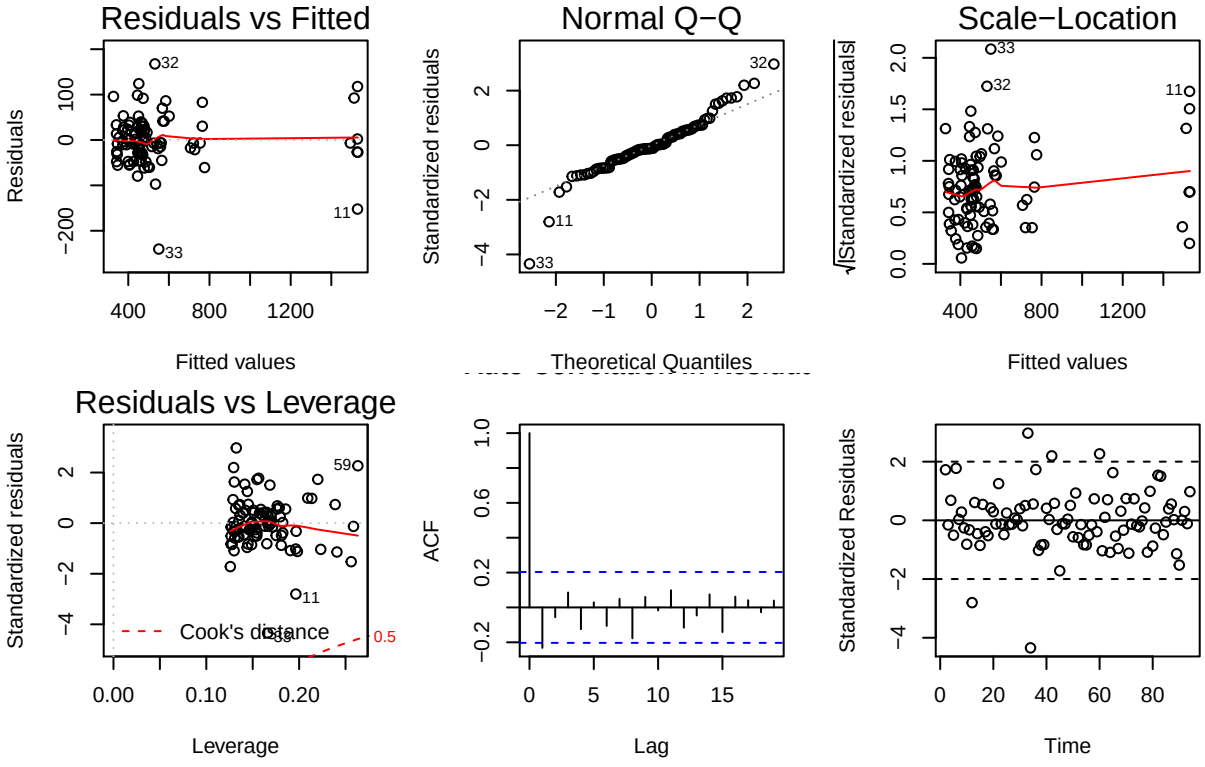

Figure 2: Problem 2b: Diagnostic plots for ordinary linear model.

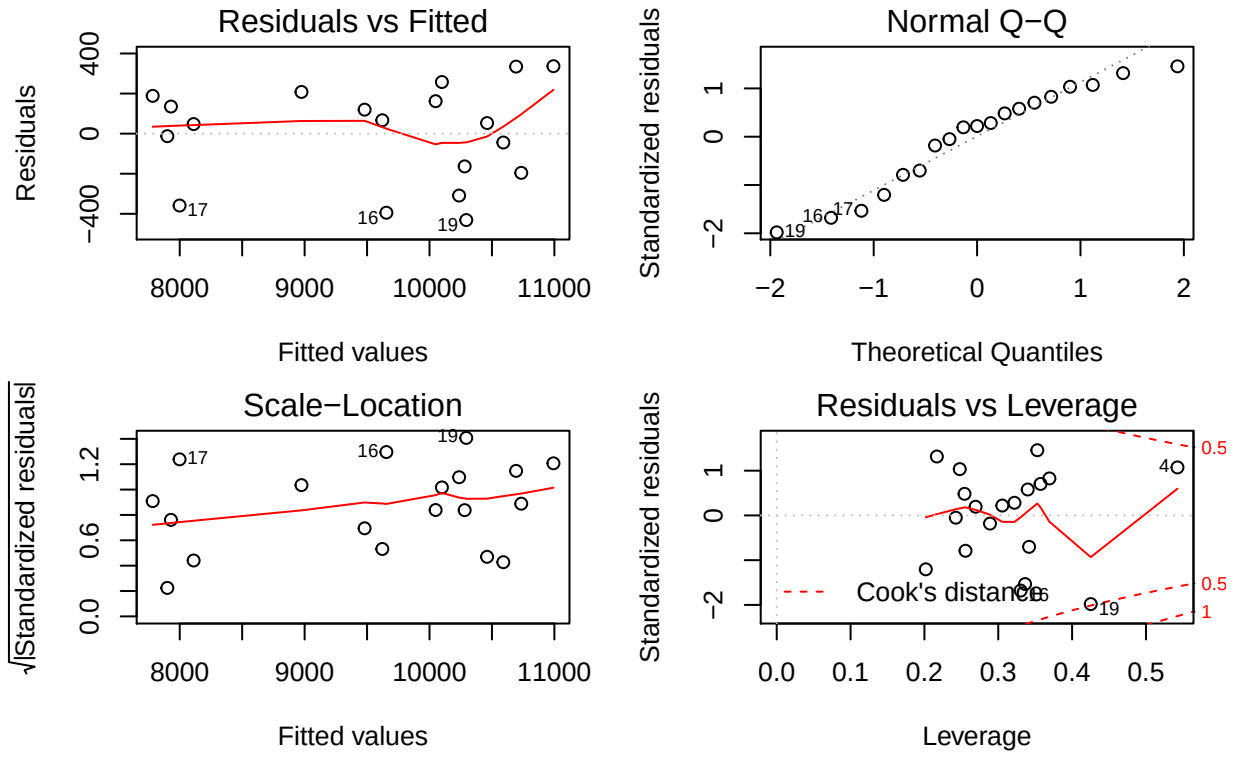

Figure 3: Problem 3a: Diagnostic plots for ordinary linear model.

values, and the autocorrelation plot shows large autocorrelations, all of which suggest that the linear model with assumed independence between the residuals is not appropriate.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4421	1235	3.58	0.00336
Temp	78.18	33.89	2.307	0.03817
Sun	2.585	1.556	1.661	0.1206
Q2	-879.7	1121	-0.7845	0.4468
Q3	-1552	1541	-1.007	0.3323
Q4	715.4	474.3	1.508	0.1554

Table 6: Problem 3a: Summary of ordinary linear model.

Observations	Residual Std. Error	R^2	Adjusted R^2
19	287.2	0.9537	0.9359

3 (b)

Note that the model can be estimated with either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML). I have used the default, which is REML, but either one is fine. They will produce different estimates. The code should look like one of the following:

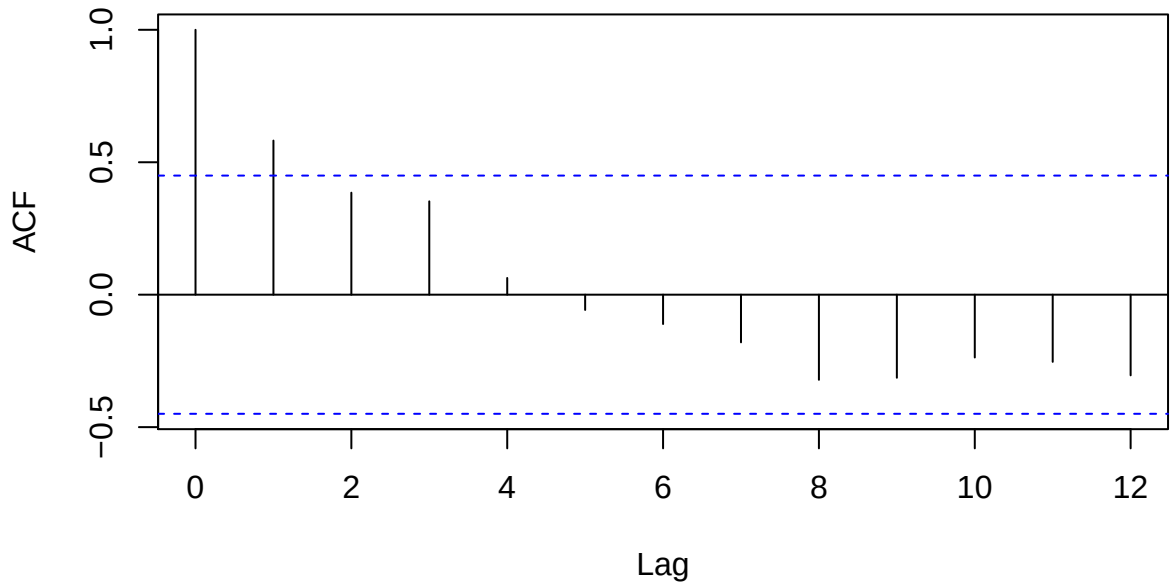

Figure 4: Problem 3a: acf plot for ordinary linear model residuals.

```
model <- gls(Sales ~ . -Case - Time, data=carlsen, correlation=corAR1())
model <- gls(Sales ~ . -Case - Time, data=carlsen, correlation=corAR1(),
             method = "ML")
```

i. A summary of this model is below. From this summary, the estimated value of ρ, the lag-1 autocorrelation parameter, is 0.87. (If you use ML for estimation, the estimate is 0.79).

ii. The coefficient values have changed in magnitude, though not in sign. The estimated standard errors in the AR1 model are much smaller than in the ordinary linear model, which is to be expected if the model is properly accounting for correlation among the residuals. As a result, both `Sun` and `Q4` are significant in the AR1 model but not in the ordinary linear model at conventional thresholds.

```
## Problem 3b: Summary of AR1 model:

## Generalized least squares fit by REML
##   Model: Sales ~ . - Case - Time
##   Data: carlsen
##   AIC BIC logLik
##   213 217    -98
##
## Correlation Structure: AR(1)
##  Formula: ~1
##  Parameter estimate(s):
##  Phi
## 0.87
##
## Coefficients:
##             Value Std.Error t-value p-value
## (Intercept)  5215       747     7.0  0.0000
## Temp           56        23     2.4  0.0300
## Sun             2         1     2.8  0.0148
## Q2            -36       663    -0.1  0.9572
## Q3           -371       942    -0.4  0.6999
## Q4            946       295     3.2  0.0069
```

```
##
##  Correlation:
##      (Intr) Temp   Sun    Q2     Q3
## Temp -0.870
## Sun   0.005 -0.357
## Q2    0.914 -0.938  0.033
## Q3    0.912 -0.960  0.100  0.992
## Q4    0.747 -0.937  0.485  0.840  0.876
##
## Standardized residuals:
##   Min    Q1   Med    Q3   Max
## -1.18 -0.29  0.50  0.57  1.13
##
## Residual standard error: 409
## Degrees of freedom: 19 total; 13 residual
```

3 (c)

The log likelihoods are -131 for the ordinary linear model and -98 for the AR1 mode. The difference in degrees of freedom is 1, so we compare the statistic $-2(-131-(-98)) = 66$ to a $\chi^2_{(1)}$ distribution. The pvalue is 8.08×10^{-16}, so we reject the null hypothesis and conclude that the autocorrelation is not 0.

3 (d)

A summary of the model is below. Diagnostic plots are in Figure 5, and the autocorrelation plot of the residuals is in Figure 6.

The diagnostic plots suggest problems with the fit. The residuals are not very close to normally distributed, and the scale-location plot appears to show a relationship between the fitted values and the variance. The autocorrelation plot, however, shows much reduced autocorrelation among the residuals, as hoped for.

The likelihood of the model is `-130`, as compared to `-98` for the model in 3c. It appears to be more effective to estimate the autocorrelation simultaneously with estimating the model, as in 3c, rather than following a two-stage estimation procedure.

	Estimate	Std. Error	t value	Pr(>|t|)
Xstar(Intercept)	5090	793.5	6.415	0.0000229
XstarTemp	61.01	24.64	2.476	0.02781
XstarSun	2.196	0.8827	2.488	0.02722
XstarQ2	-214.5	725.8	-0.2956	0.7722
XstarQ3	-631.1	1026	-0.6148	0.5493
XstarQ4	879.1	322.7	2.724	0.01738

Table 8: Problem 3d: Summary of ordinary linear model with transformed variables.

Observations	Residual Std. Error	R^2	Adjusted R^2
19	277.7	0.9979	0.9969

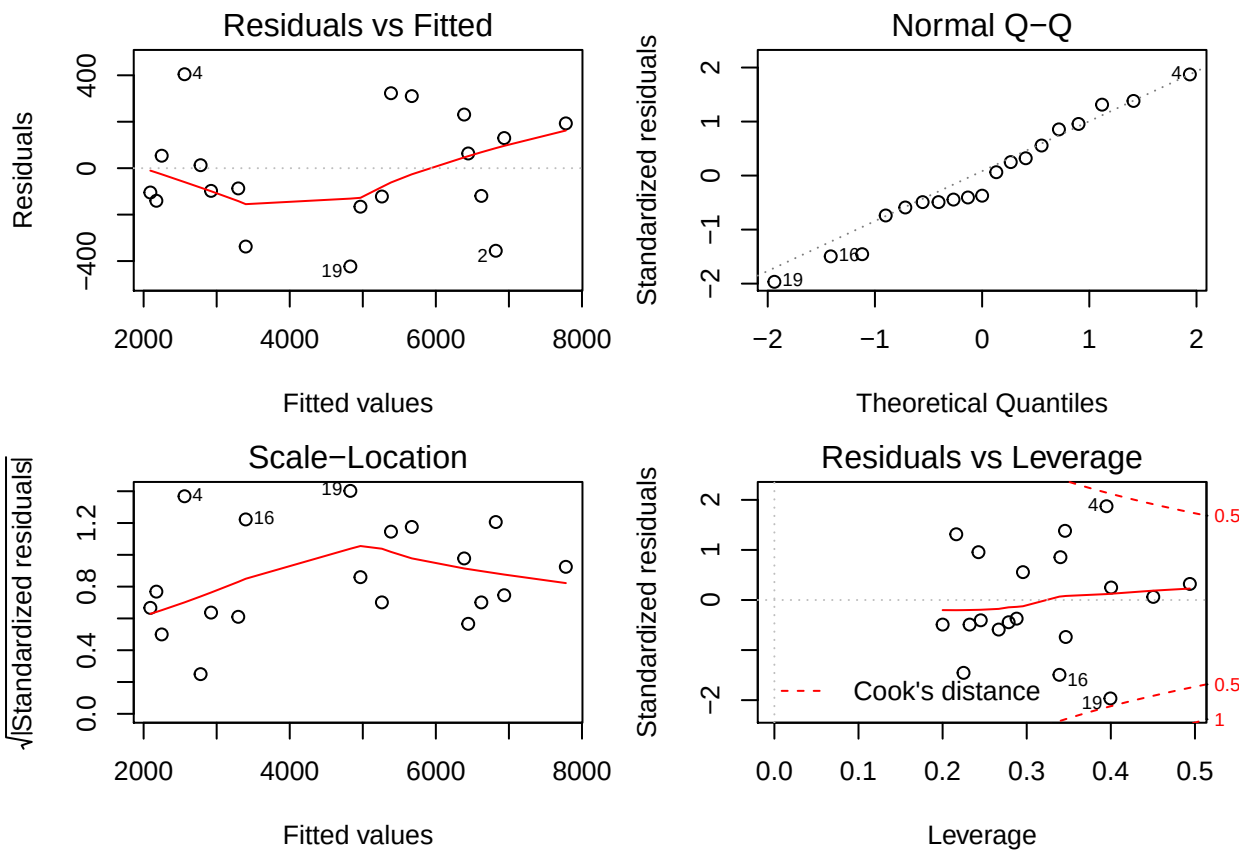

Figure 5: Problem 3d: diagnostic plots.

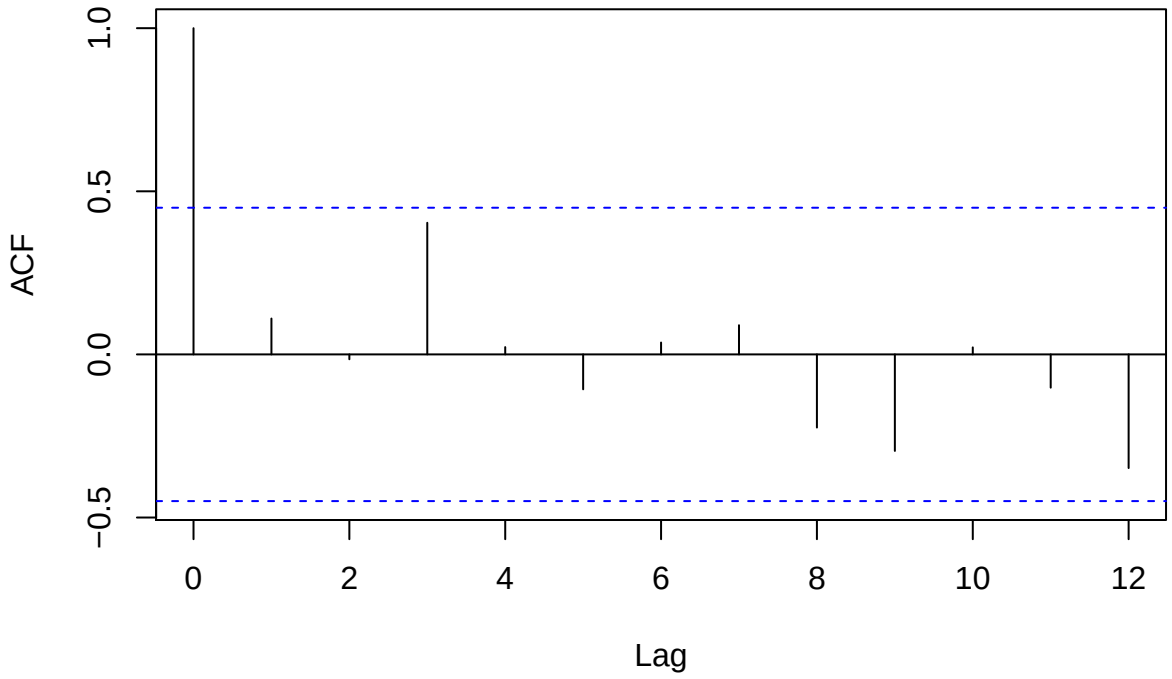

Figure 6: Problem 3d: acf plot for ordinary linear model residuals.

4 (a)

We see in Figure 7 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 7 is easy to interpret but is probably not the best way to convey the information it is displaying.

4 (b)

Figure 8 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 8 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.

4 (c)

Figure 9 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 8 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 7 and Figure 8.

4 (d)

We see in Figure 10 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 10 fails to convey any country level trends.

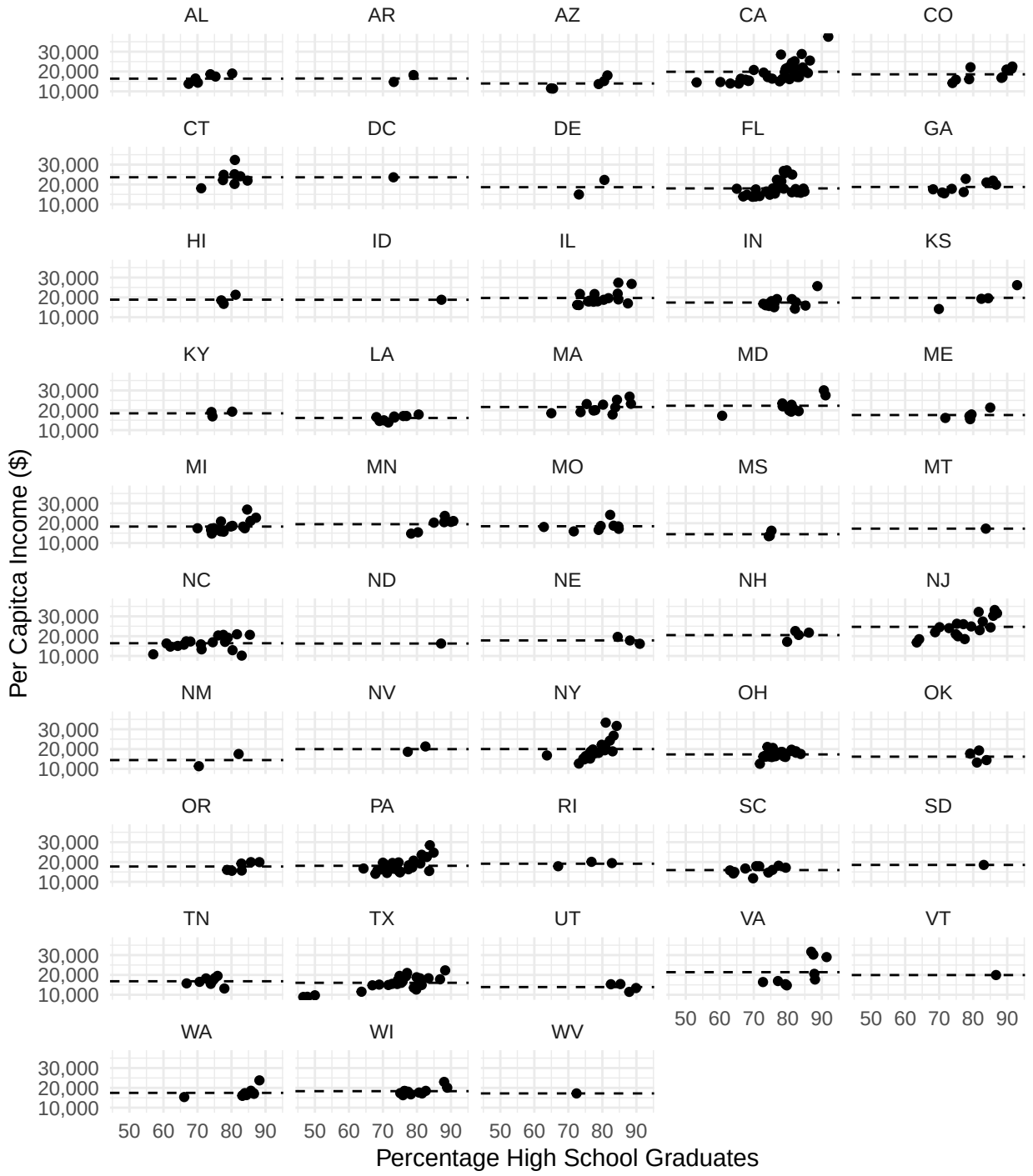

Figure 7: Problem 4a: HS graduation rate vs. county level per capita income with state level mean per capita income line.

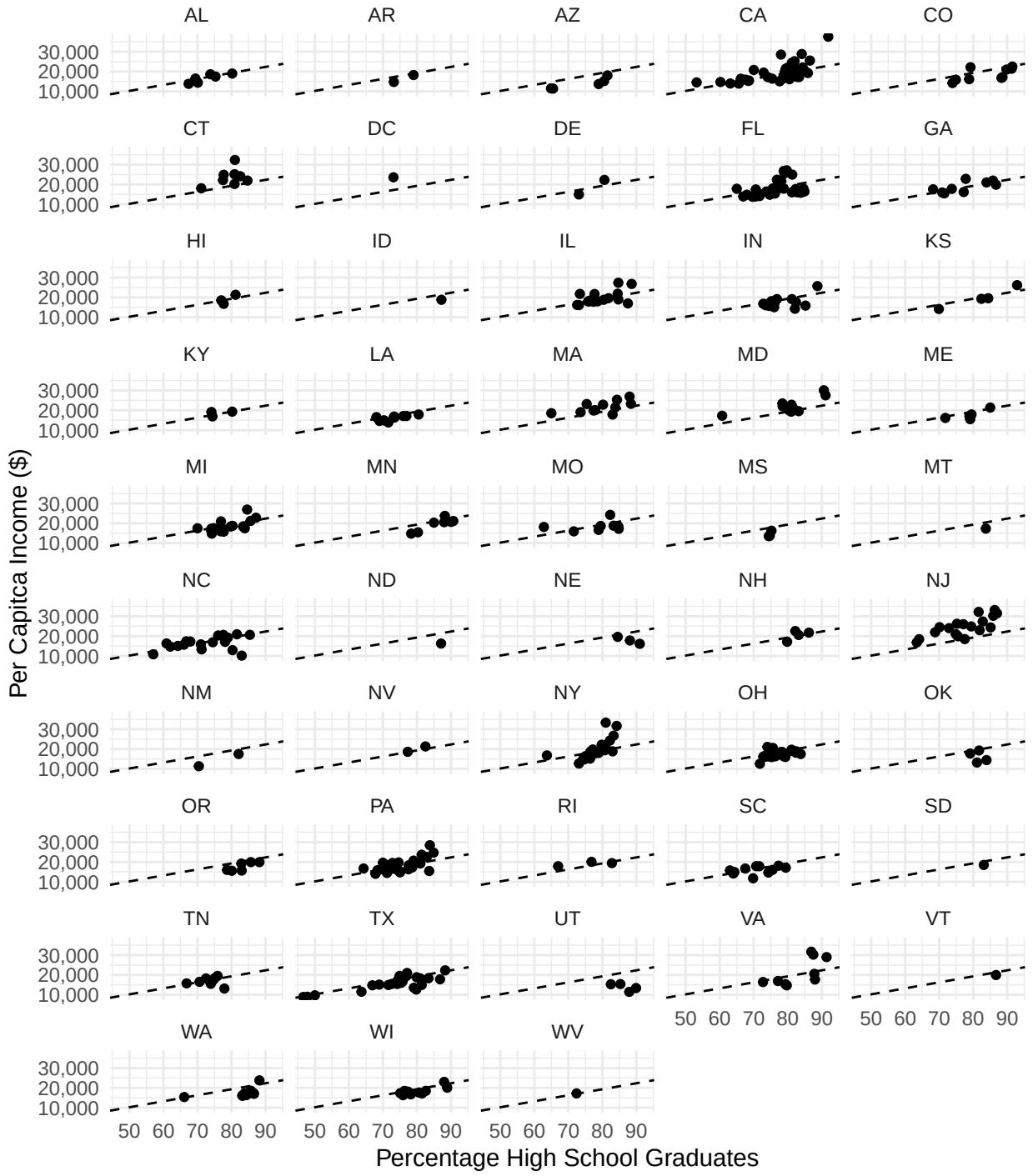

Figure 8: Problem 4b: HS graduation rate vs. county level per capita income with national level regression line.

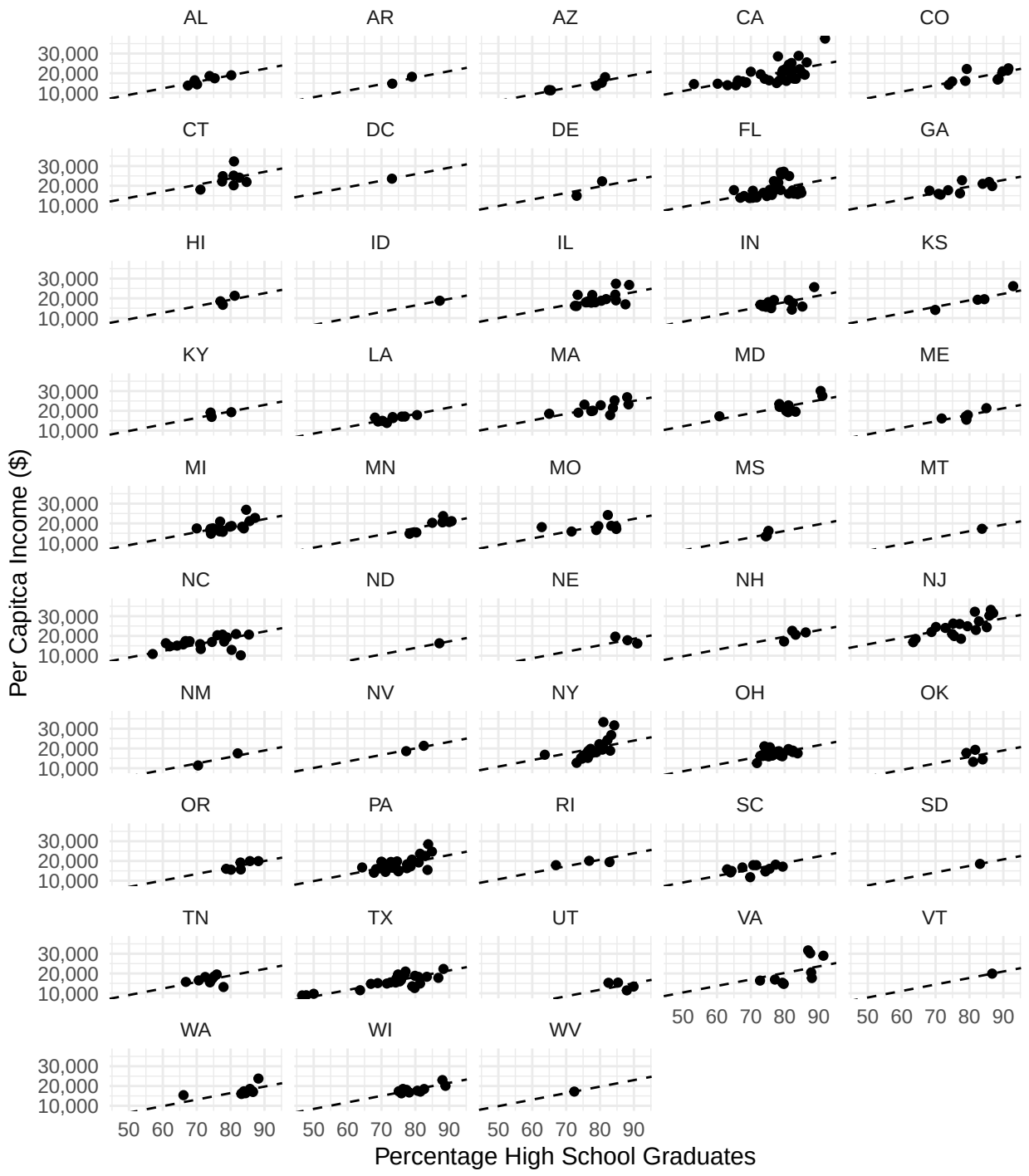

Figure 9: Problem 4c: HS graduation rate vs. county level per capita income with state level intercept and national slope line.

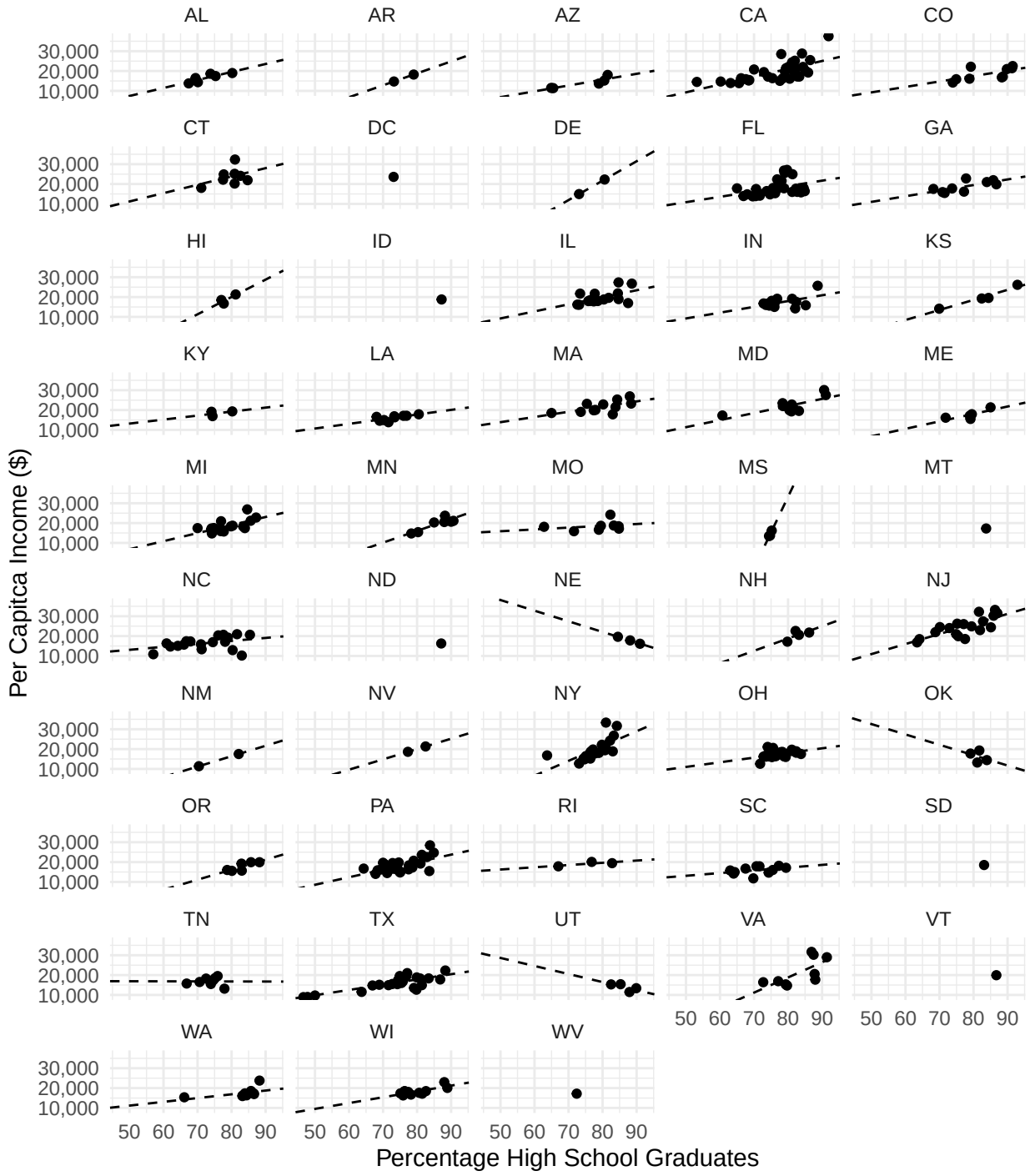

Figure 10: Problem 4d: HS graduation rate vs. county level per capita income with state level regression line.