Homework 07 Solutions

11/09/2018

1.

(a).

Summaries and diagnostic plots for both models are below. Since the variance of a Poisson random variable is equal to the mean, the raw residual plot and scale-location plot are not very informative, so we will focus on the other two diagnostic plots.

For model1, the model with no interaction term, the residuals appear to be approximately normally distributed (Figure 1). All the points have relatively high leverage, and points 5, 9, and 24 have been labeled as influential points that be worth investigating further. Many of these points have standardized Pearson residuals with absolute values between 2 and 4, so there are appear to be a lot of outliers. The binned residual plot doesn't give any particular cause for concern (Figure 2).

For model2, the model with the interaction term, the residuals also appear to be approximately normally distributed (Figure 3). Leverage is not shown directly on the leverage plot in this case, since R has chosen to plot the factors of the `Wool` variable on the x-axis instead, but there are no points that are marked as influential. There are still a lot of points that look like outliers, with very large standardized Pearson residual values. The binned residual plot now has values very close to 0, though all the values are slightly negative.

Summary of Model 1 with no interaction:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.692	0.04541	81.3	0
woolB	-0.206	0.05157	-3.994	0.0000649
tensionM	-0.3213	0.06027	-5.332	9.729e-08
tensionH	-0.5185	0.06396	-8.107	5.209e-16

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	297.4 on 53 degrees of freedom
Residual deviance:	210.4 on 50 degrees of freedom

Summary of Model 2 with interaction term:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.797	0.04994	76.03	0
woolB	-0.4566	0.08019	-5.694	1.24e-08
tensionM	-0.6187	0.0844	-7.33	2.295e-13
tensionH	-0.5958	0.08378	-7.112	1.146e-12
woolB:tensionM	0.6382	0.1222	5.224	1.747e-07
woolB:tensionH	0.1884	0.1299	1.45	0.147

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	297.4 on 53 degrees of freedom
Residual deviance:	182.3 on 48 degrees of freedom

(b) i.

The AICs and BICs are as follows:

	AIC	BIC
Model 1 (without interaction)	493	501
Model 2 (with interaction)	469	481

Both metrics are lower for model 2, so that is the model we would choose.

(b) ii.

The log likelihoods are -243 for model 1 without the interaction and -228 for model 2 with the interaction. The difference in degrees of freedom is 2. This comes from the fact that the variable `wool` has two levels, which corresponds to a single dummy variable, and the variable `tension` has three levels, which corresponds to two dummy variables. The interaction between them therefore corresponds to `2 x 1 = 2` dummy variables.

We therefore compare the statistic $-2(-243-(-228)) = 28$ to a $\chi^2_{(2)}$ distribution. The pvalue is 7.96×10^{-7}, so we reject the null hypothesis and decide to include the interaction. The likelihood ratio test therefore selects the same model as the AIC and BIC.

(c)

The Pearson residuals are

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$$

If the model is correctly specified, then $\sum_{i=1}^n r_i$ should have approximately a $\chi_{n-(p+1)^2}$ distribution.

We have $\sum_{i=1}^n r_i = 180.67$. Comparing this to a χ^2 distribution with 48 degrees of freedom yields a p-value of 2.93×10^{-17}. We conclude therefore that the residuals are overdispersed, meaning that the variability of the responses is larger than what we would expect if they really followed a conditional Poisson distribution.

(d)

Summary of Model 3, the quasipoisson model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.797	0.09688	39.19	4.193e-38
woolB	-0.4566	0.1556	-2.935	0.005105
tensionM	-0.6187	0.1637	-3.778	0.0004359
tensionH	-0.5958	0.1625	-3.666	0.000616
woolB:tensionM	0.6382	0.237	2.693	0.009727
woolB:tensionH	0.1884	0.252	0.7475	0.4584

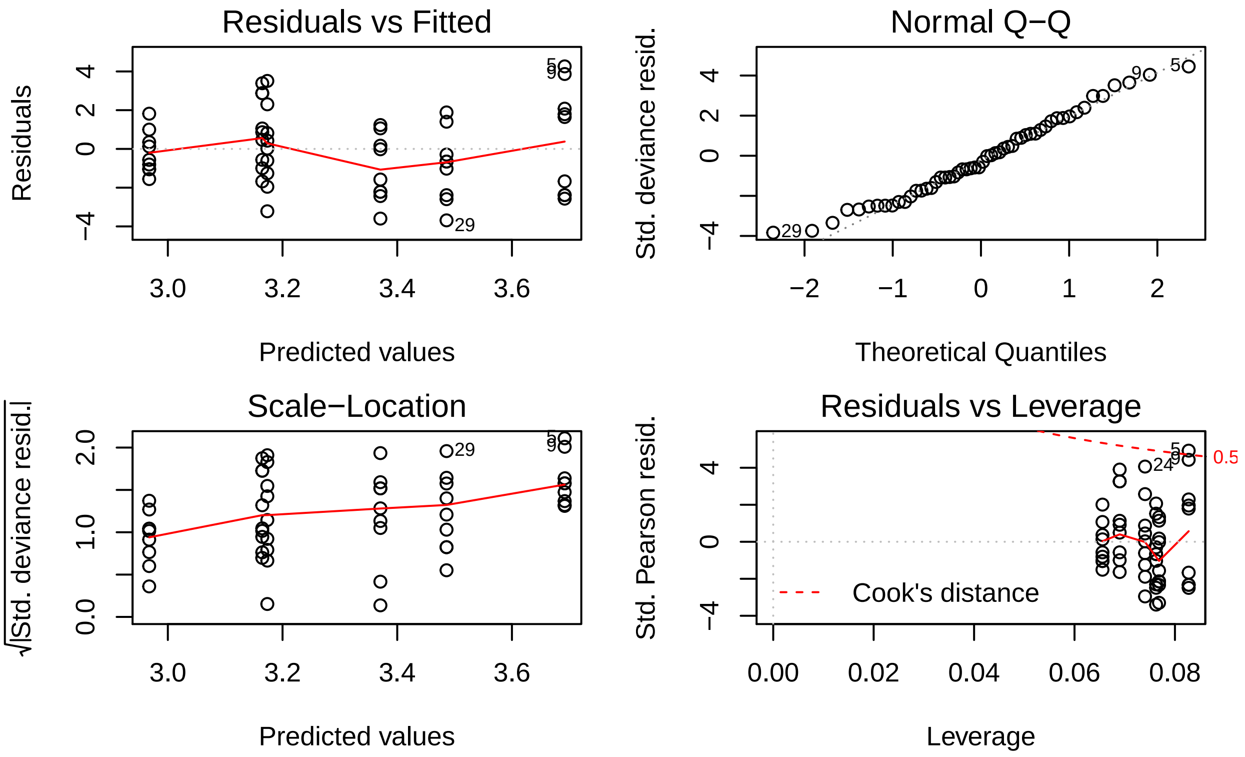

Figure 1: Problem 1(a): Diagnostic plots for model without interaction term.

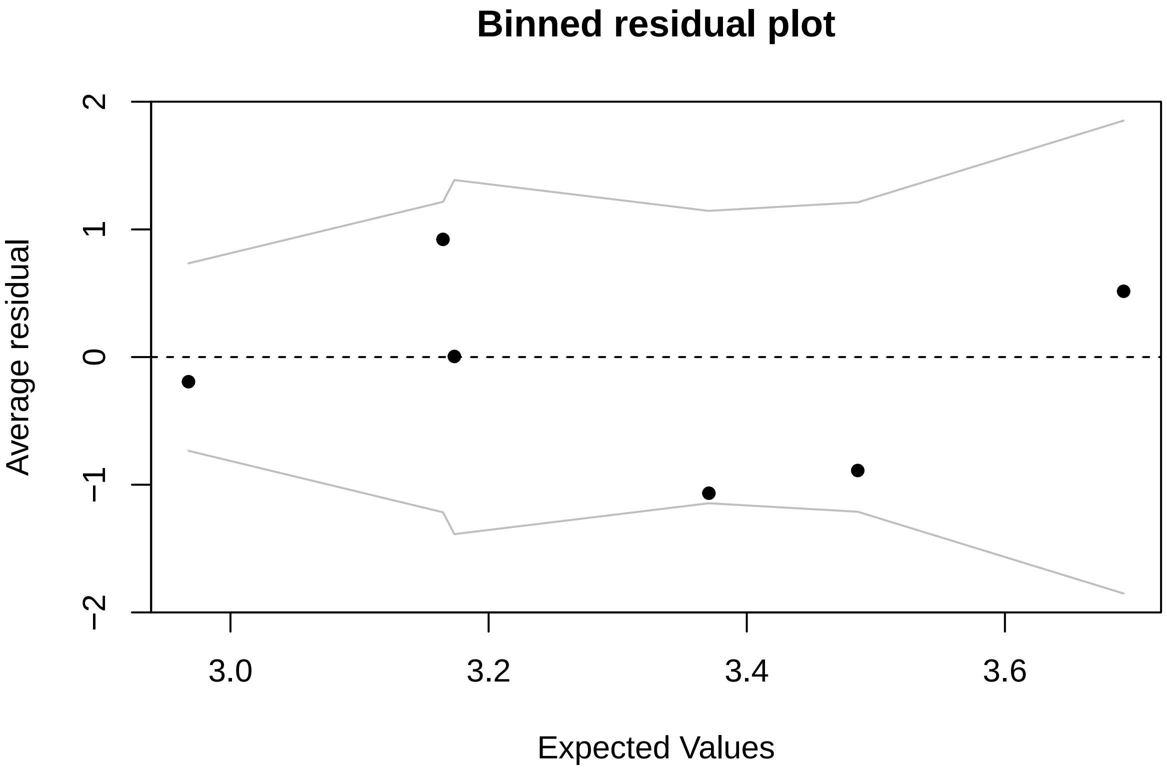

Figure 2: Problem 1(a): Binned plot of residuals vs. fitted values for model without interaction term.

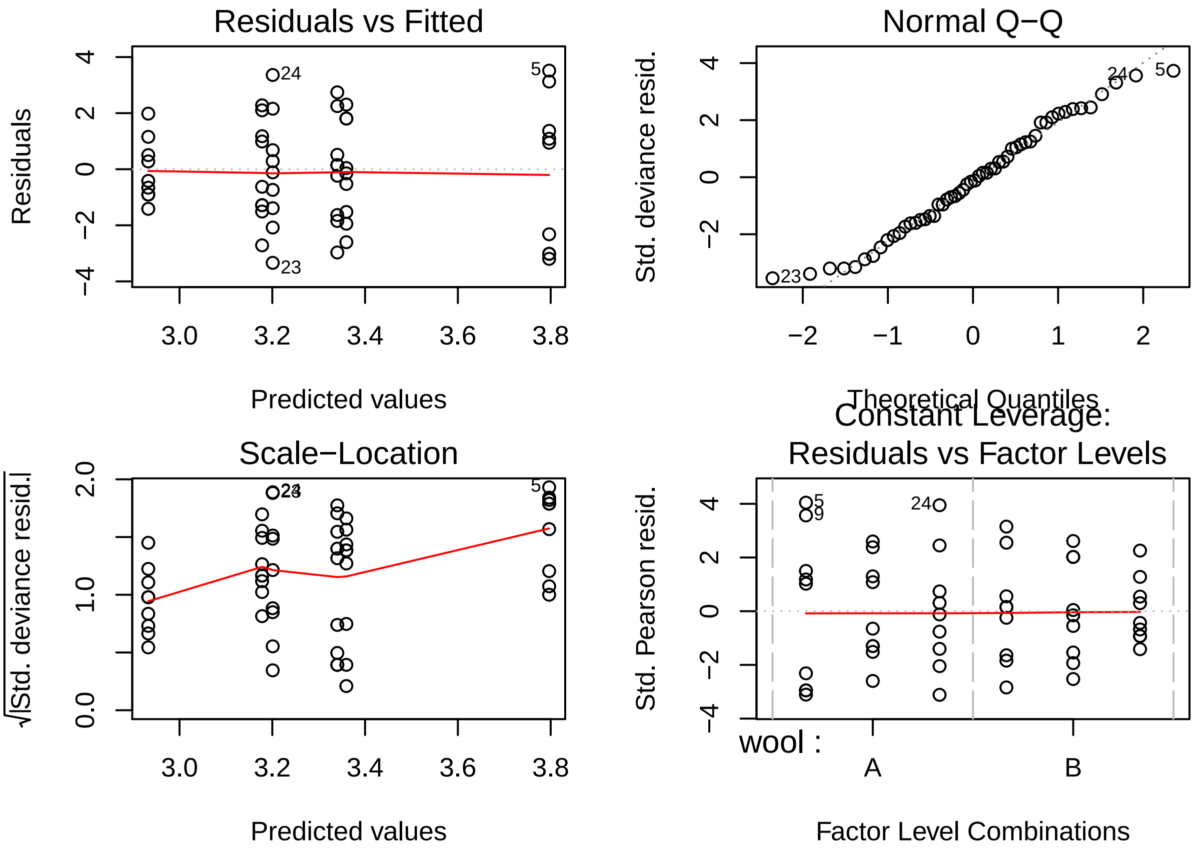

Figure 3: Problem 1(a): Diagnostic plots for model with interaction term.

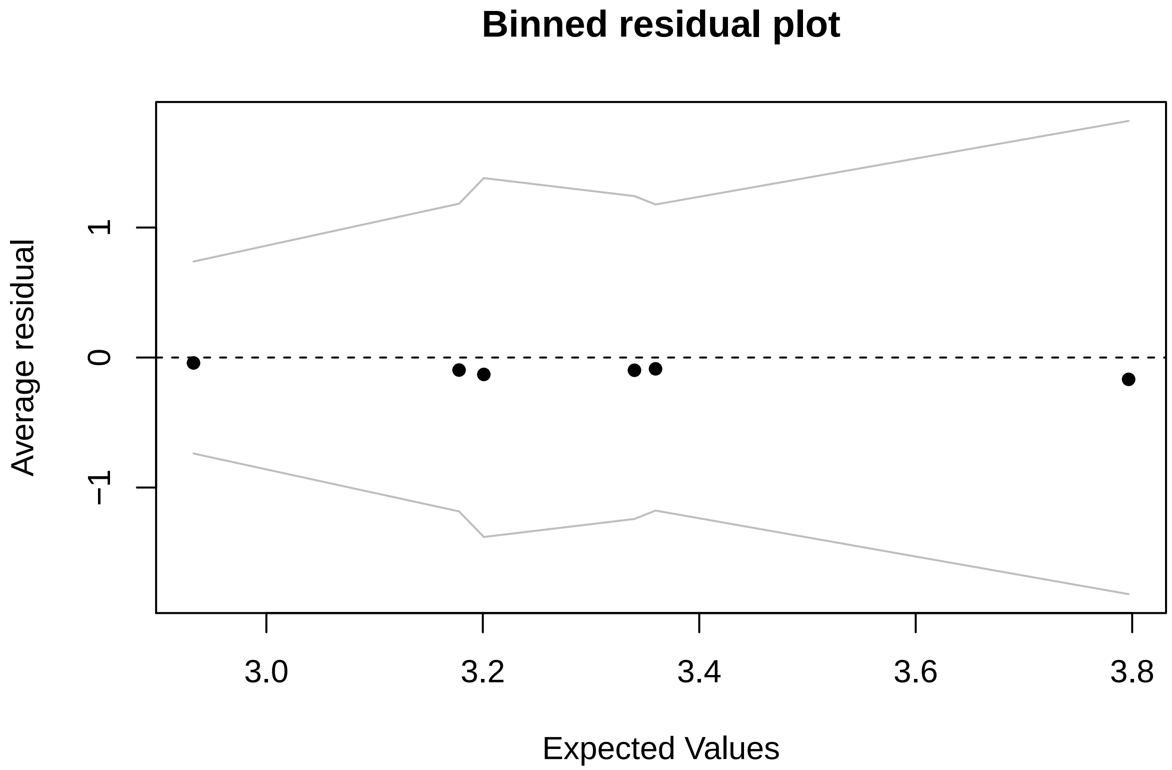

Figure 4: Problem 1(a): Binned plot of residuals vs. fitted values for model with interaction term.

(Dispersion parameter for quasipoisson family taken to be 3.8)

Null deviance:	297.4 on 53 degrees of freedom
Residual deviance:	182.3 on 48 degrees of freedom

The coefficient estimates are identical to those in Model 2, but the estimated standard errors are much larger, as expected. Consequently, the p-values are larger, though still significant at customary thresholds like 0.05 or 0.01. The dispersion parameter is estimated to be 3.8, whereas it was assumed to be 1 in Model 2.

2(a)

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

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

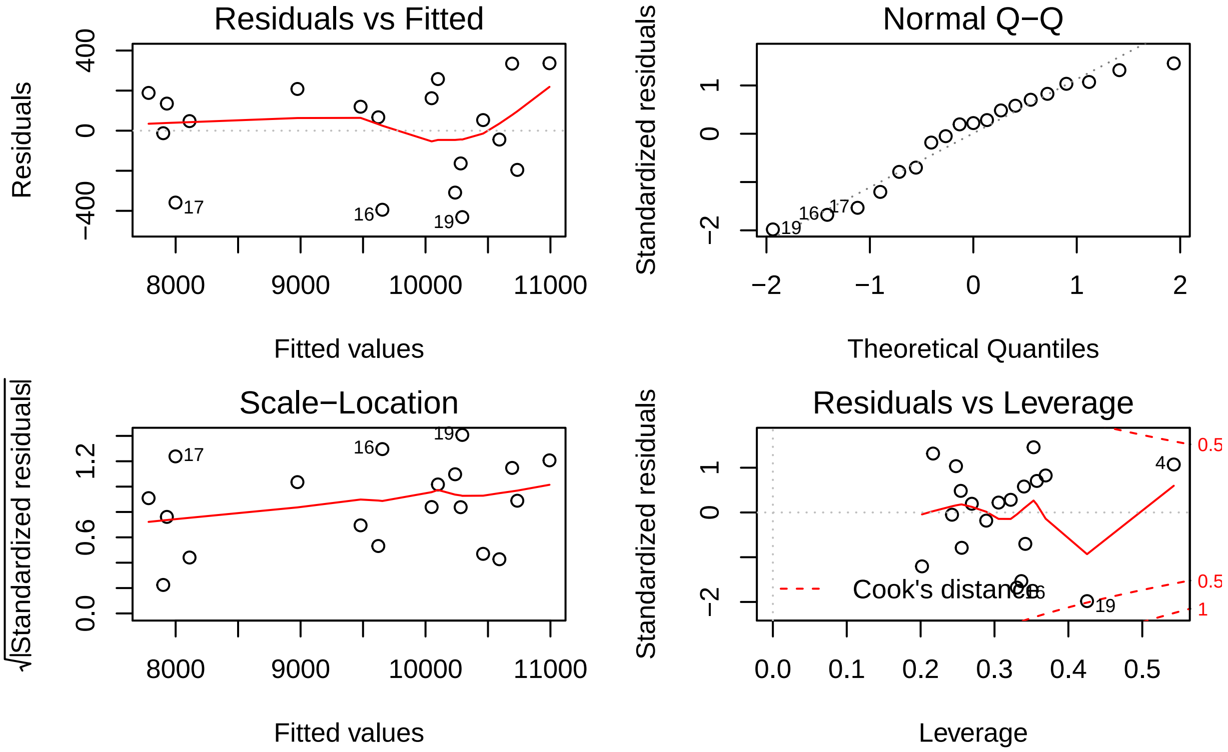

Figure 5: Problem 2a: Diagnostic plots for ordinary linear model.

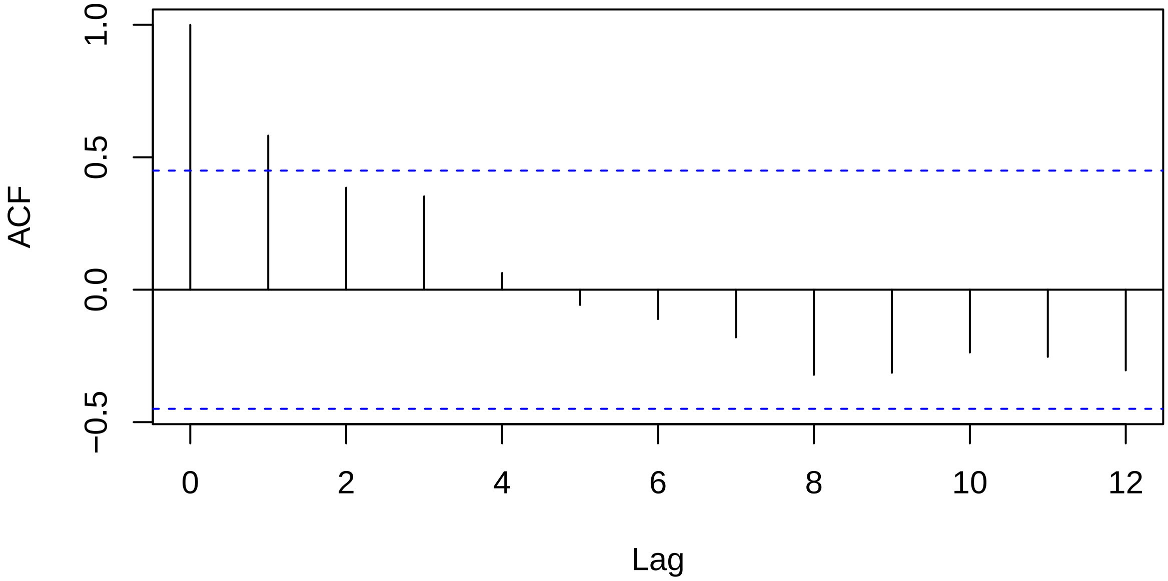

Figure 6: Problem 2a: acf plot for ordinary linear model residuals.

2(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:

```r
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 2b: 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
```

2(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.

2(d)

A summary of the model is below. Diagnostic plots are in Figure 7, and the autocorrelation plot of the residuals is in Figure 8.

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 2c. It appears to be more effective to estimate the autocorrelation simultaneously with estimating the model, as in 2c, 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 11: Problem 2d: Summary of ordinary linear model with transformed variables.

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

3.

(a) The treatment variable is no longer randomized in the way intended by the researcher, because individuals have selected their own level of treatment. This selection process is very likely confounded with the outcome variable, meaning that there is some unobserved variable that is correlated with both the treatment variable and the outcome.

For example, imagine that the therapy actually has no effect, but there is an “optimism” variable that determines both how many sessions a person attends and their emotional state at the end of the study. In other words, the people who engage the most with the treatment are those in the best state at the end of the study, but not because of the therapy. Then the number of sessions attended will strongly predict the outcome variable, but not because of an effect of the therapy.

An unobserved confounder could also obscure a real treatment effect. Suppose that instead of an optimism variable, there is a variable representing emotional wellbeing that is inversely correlated with the number of sessions attended. That is, those who start out the most depressed seek the most treatment, perhaps because

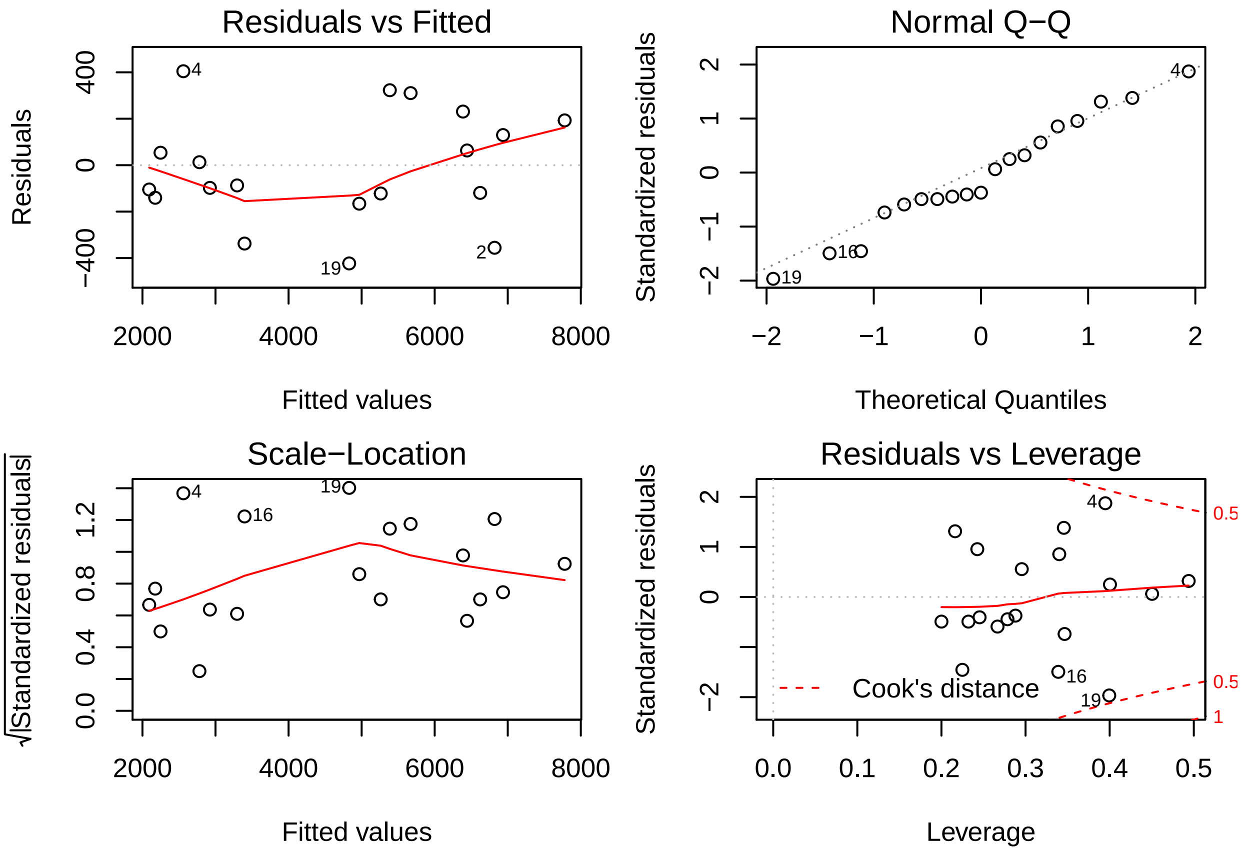

Figure 7: Problem 2d: diagnostic plots.

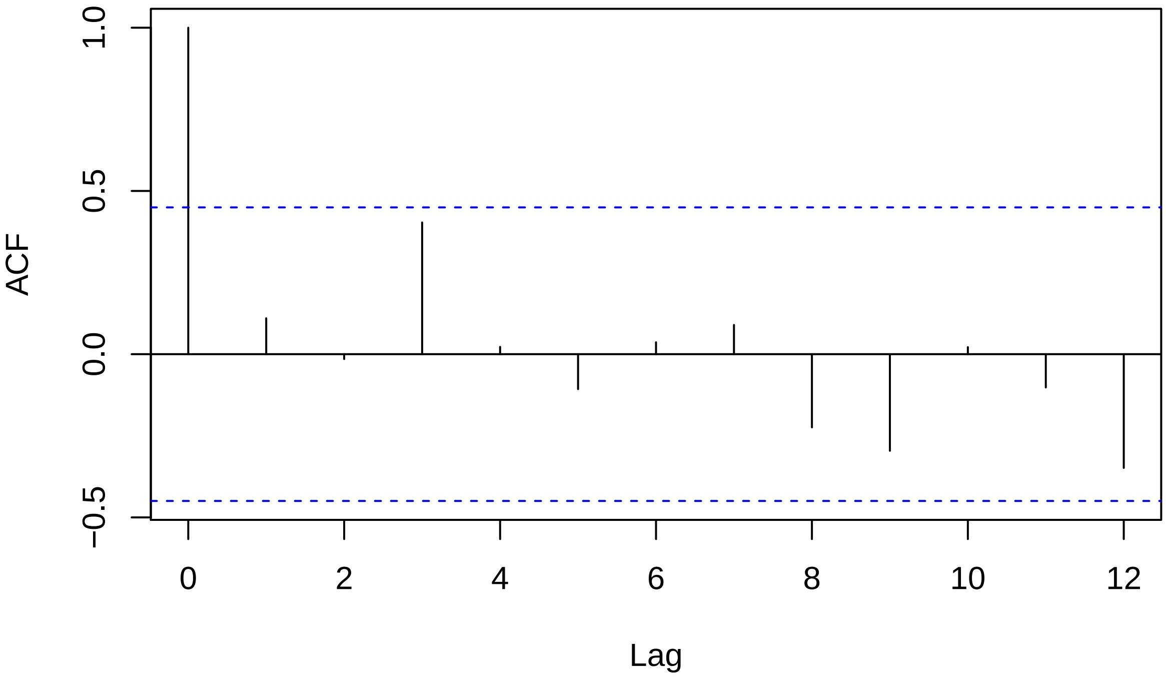

Figure 8: Problem 2d: acf plot for ordinary linear model residuals.

they feel that they have the most to gain. Suppose the treatment *does* have a beneficial effect that increases with the number of sessions attended, and suppose that everyone seeks out just enough treatment to get them to roughly the same level of emotional wellbeing at the end of the study. Then a coefficient for number of treatment sessions could end up being 0 even though there is a treatment effect!

In sum: whenever there is noncompliance in an experiment, there is potential for confounding due to selection effects. Including a variable for the amount of treatment received *does not* address this and can lead to highly misleading results.

3(b)

This is perfectly set up for an Instrumental Variable analysis. The assignment variable (assignment to treatment or control) is the instrument, and the number of sessions attended is the treatment variable. The assignment variable almost certainly affects the treatment variable, but it should have no effect on the outcome through other pathways. (That is, the mere fact of having been assigned to the treatment or control group shouldn't affect participants' emotional wellbeing at the end of the study.) An experimental design with noncompliance is almost always a good candidate for an IV analysis.

An analysis based on propensity score matching won't work here unless we have variables with which to construct the propensity score. The propensity score is the probability of receiving treatment as a function of some set of covariates X. That is, the propensity score is a function $\pi(x) := P(A = 1|X = x)$, where $A = 1$ means that someone receives treatment. We can define it more broadly as the probability of receiving a particular amount of treatment a: $\pi_a(x) = P(A = a|X = x)$. We want as rich a set of covariates X as possible in order to perform matching. If the investigator collected these prior to the start of the study, then this analysis may be possible.