Problem 1

We consider a population with true model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, for which we fit the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

(a) For the fitted model, the estimated regression coefficients are given by the usual SLR formula:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

From this, the expected value of the regression coefficient estimates is

$$\begin{aligned}
\mathbb{E}\hat{\boldsymbol{\beta}} &= \mathbb{E}\Big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\Big] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}\Big[\boldsymbol{Y}\Big] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}\Big[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \varepsilon\Big] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Big(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}\Big) \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{\gamma}
\end{aligned}$$

(b)

$$\begin{aligned}
\mathbb{E}\hat{\boldsymbol{Y}} &= \mathbb{E}\Big[\boldsymbol{H}\boldsymbol{Y}\Big] = \boldsymbol{H}\mathbb{E}\boldsymbol{Y} \\
&= \boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}) \\
&= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{H}\boldsymbol{Z}\boldsymbol{\gamma} \\
&= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{Z}\boldsymbol{\gamma}
\end{aligned}$$

(c) For bias $\boldsymbol{\delta} = \mathbb{E}(\hat{\boldsymbol{y}}) - \mathbb{E}(\boldsymbol{Y})$, the expected bias is

$$\begin{aligned}
\mathbb{E}\boldsymbol{\delta} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{Z}\boldsymbol{\gamma} - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}) \\
&= \boldsymbol{H}\boldsymbol{Z}\boldsymbol{\gamma} - \boldsymbol{Z}\boldsymbol{\gamma} \\
&= (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{Z}\boldsymbol{\gamma}
\end{aligned}$$

Weisberg 8.2

(1) Trying a few different transformations, we find that the square root transformation on `Distance` linearizes the regression. The scatterplot of $\sqrt{\texttt{Distance}} \sim \texttt{Speed}$ is given in Figure 1, which shows a clear linear relationship. The residuals obtained from this model also confirm the linear relationship, with the residuals randomly distributed about the zero line (Figure 2). Comparing the R^2 value from this model to the model with untransformed `Speed`, we see an increase from 0.8777 to 0.9251, showing a significant improvement after transformation.

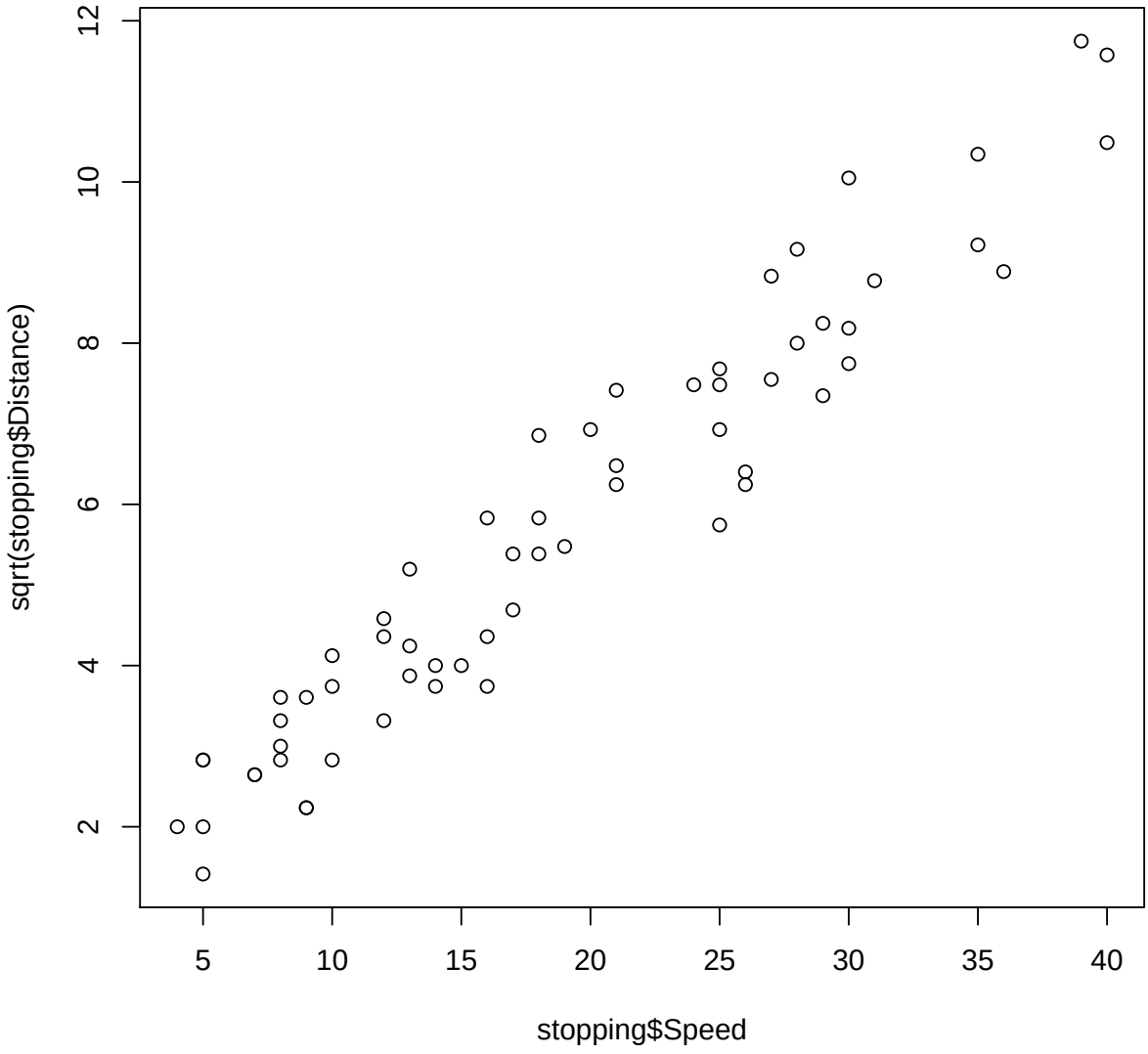


Figure 1: Plot of data with square root transformed response.

```
fit <- lm(sqrt(Distance) ~ Speed, data=stopping)
summary(fit)$r.squared

## [1] 0.9251019

raw.fit <- lm(Distance ~ Speed, data=stopping)
summary(raw.fit)$r.squared

## [1] 0.8777003
```

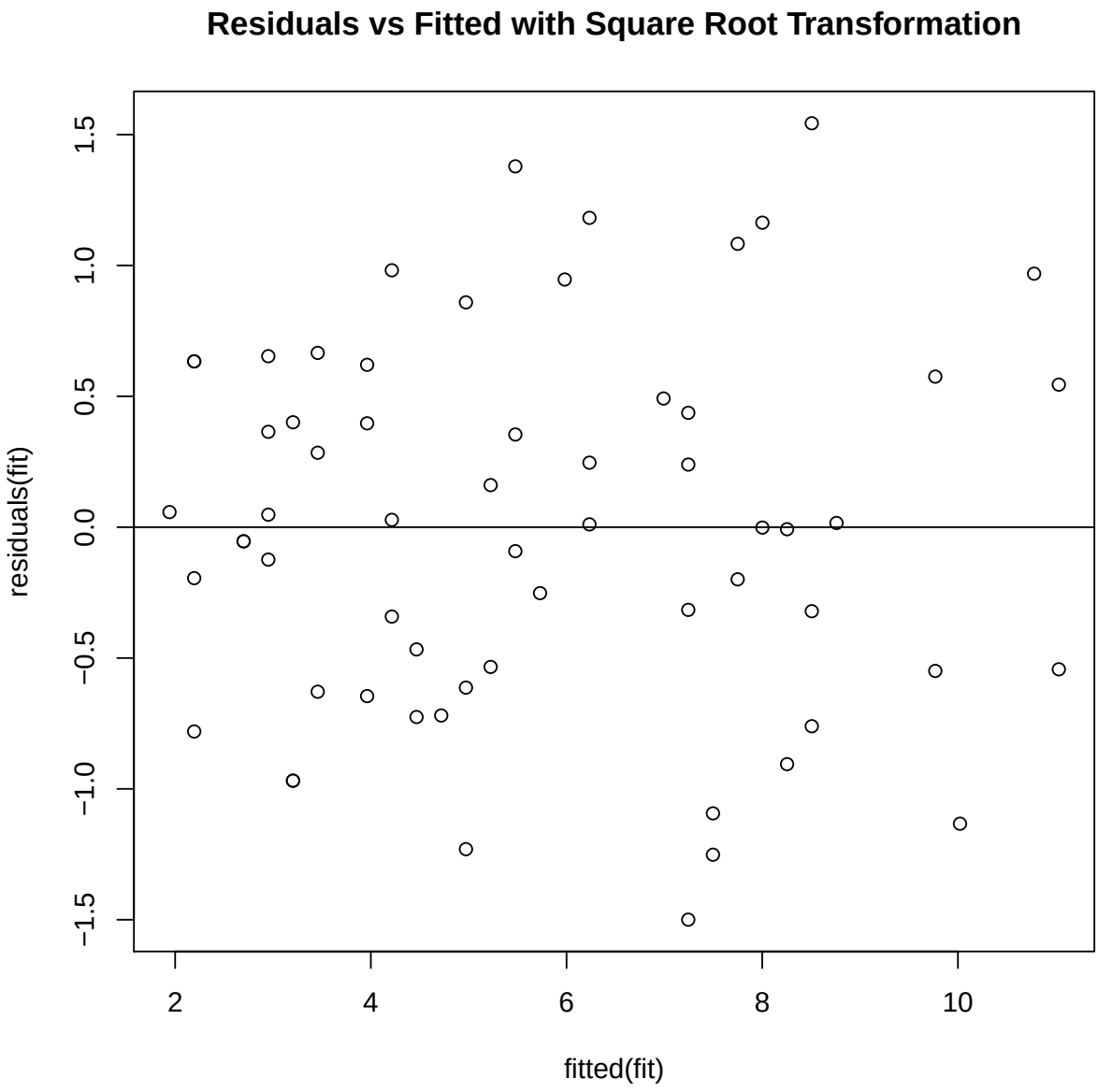


Figure 2: Plot of residuals against fitted values with $\sqrt{\texttt{Distance}}$ as response.

(2) The output below shows the results of fitting the models

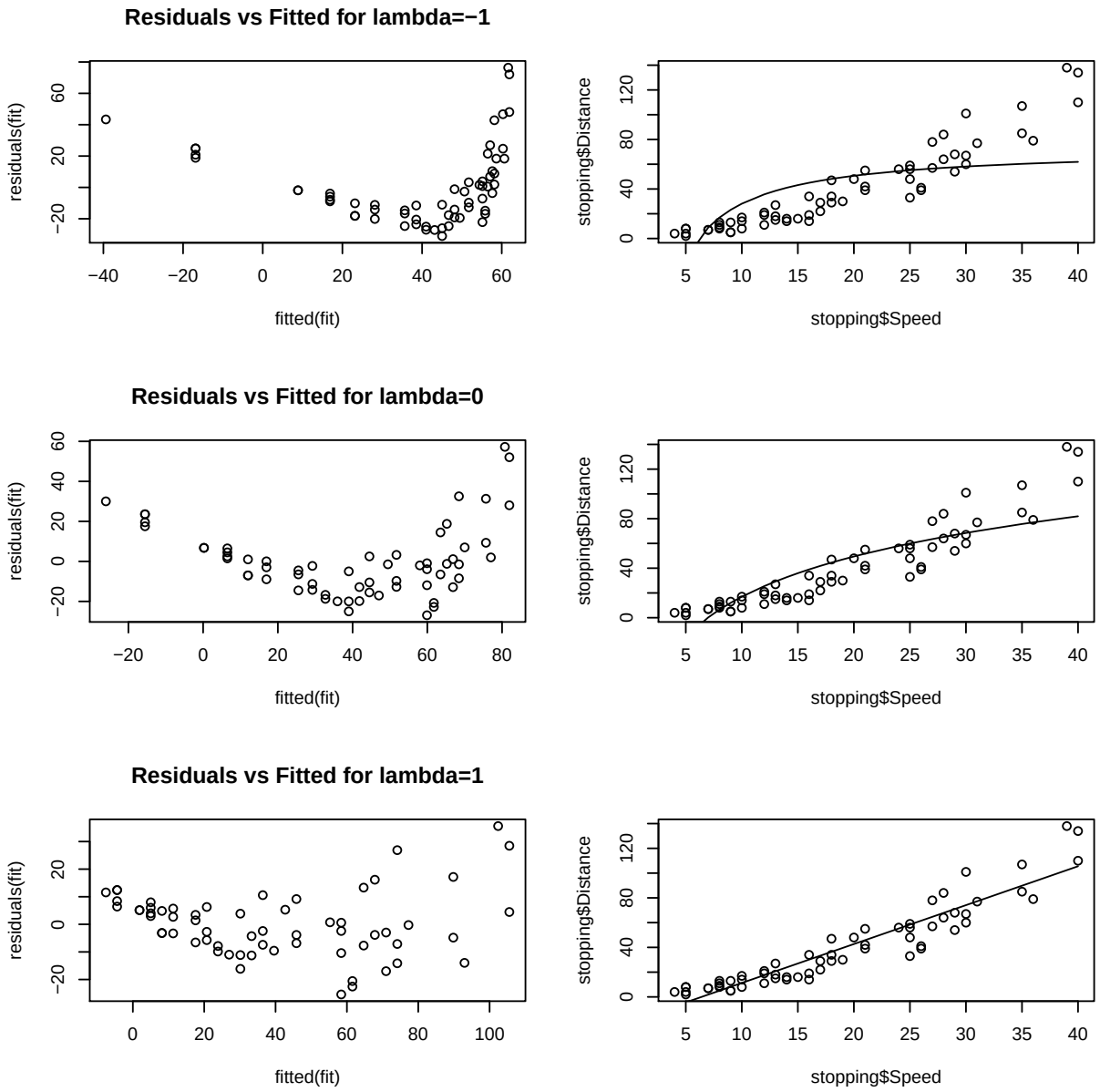
$$\texttt{Distance} \sim \texttt{Speed}^{\lambda},$$

for $\lambda = -1, 0, 1$. The R^2 of the models are

$$\begin{aligned} \lambda = -1 &: R^2 = 0.4856 \\ \lambda = 0 &: R^2 = 0.7227 \\ \lambda = 1 &: R^2 = 0.8777. \end{aligned}$$

We see that the transformations for $\lambda = -1, 0$ are detrimental to explaining `Distance`, with relatively low R^2 compared to the untransformed data with $\lambda = 1$, which has a decent R^2 of 0.8777. However, examining the residuals vs fitted plots for these transformations shows that there is a clear quadratic pattern in the residuals, indicating that none of these transformations are appropriate here.

```
par(mfrow=c(3,2))

# lambda=-1
stopping$Speed.inv <- 1/stopping$Speed
fit <- lm(Distance ~ Speed.inv, data=stopping)
plot(fitted(fit), residuals(fit), main="Residuals vs Fitted for lambda=-1")
plot(stopping$Speed, stopping$Distance)
lines(stopping$Speed, fitted(fit))
summary(fit)$r.squared

## [1] 0.4856287

# lambda=0
fit <- lm(Distance ~ log(Speed), data=stopping)
plot(fitted(fit), residuals(fit), main="Residuals vs Fitted for lambda=0")
plot(stopping$Speed, stopping$Distance)
lines(stopping$Speed, fitted(fit))
summary(fit)$r.squared

## [1] 0.7226725

# lambda=1
fit <- lm(Distance ~ Speed, data=stopping)
plot(fitted(fit), residuals(fit), main="Residuals vs Fitted for lambda=1")
plot(stopping$Speed, stopping$Distance)
lines(stopping$Speed, fitted(fit))
```

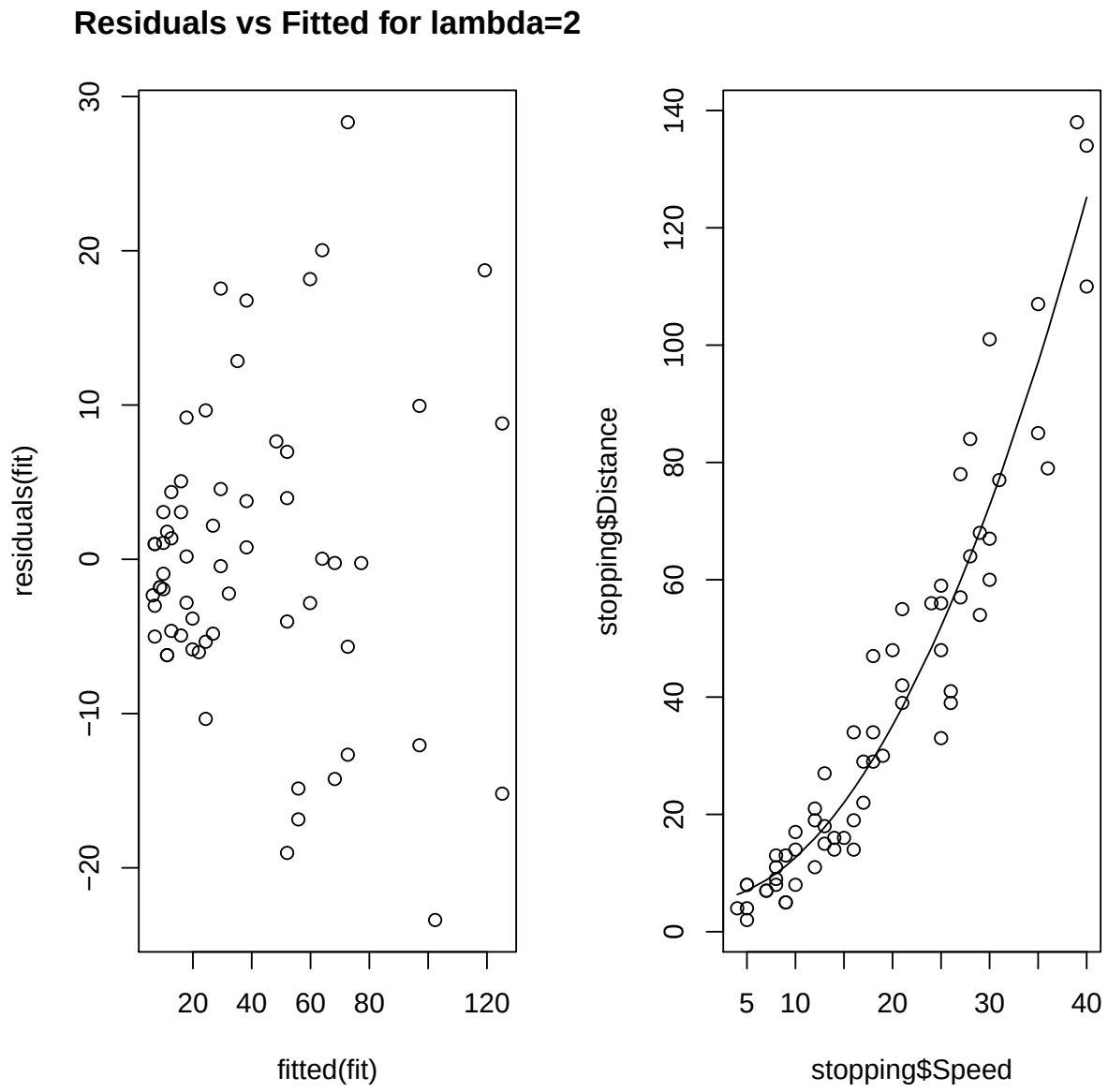



```
summary(fit)$r.squared
```

```
## [1] 0.8777003
```

(3) Now choosing $\lambda = 2$, the R^2 is 0.9136. The residuals vs fitted plot given below shows much better behavior in the residuals, with a fairly random pattern around the zero line, although there does seem to be a fan pattern indicating nonconstant variance. The fitted line shows a good fit of the data, suggesting that the quadratic transformation is best of the ones considered.

```
# lambda=2
par(mfrow=c(1,2))
stopping$Speed.sq <- stopping$Speed^2
fit <- lm(Distance ~ Speed.sq, data=stopping)
plot(fitted(fit), residuals(fit), main="Residuals vs Fitted for lambda=2")
plot(stopping$Speed, stopping$Distance)
lines(stopping$Speed, fitted(fit))
```




```
summary(fit)$r.squared

## [1] 0.9136232
```

Problem 3

(a) Let $x_c = \texttt{Speed} - \texttt{mean(Speed)}$, and $y = \texttt{Distance}$. We fit the model $y = \beta_0 + \beta_1 x_c + \beta_2 x_c^2$, which yields the fitted line shown in Figure 3.

```
stopping$Speed.c <- stopping$Speed - mean(stopping$Speed)
stopping$Speed.c.sq <- stopping$Speed.c^2
fit <- lm(Distance ~ Speed.c + Speed.c.sq, data=stopping)
```

(b) The estimate of the intercept is $\hat{\beta}_0 = 32.917$. In context, this estimates the stopping distance at $32.917ft$ for the average speed in the data set, $\bar{x} = 18.919mph$.

(c) The instantaneous rate of change for the fitted regression curve is $\hat{\beta}_1 + 2\hat{\beta}_2 x_c$. The value $\hat{\beta}_1$ then reprsents the approximate change in stopping distance for a small increase in speed when the speed is $\bar{x}$ (which corresponds to $x_c = 0$).

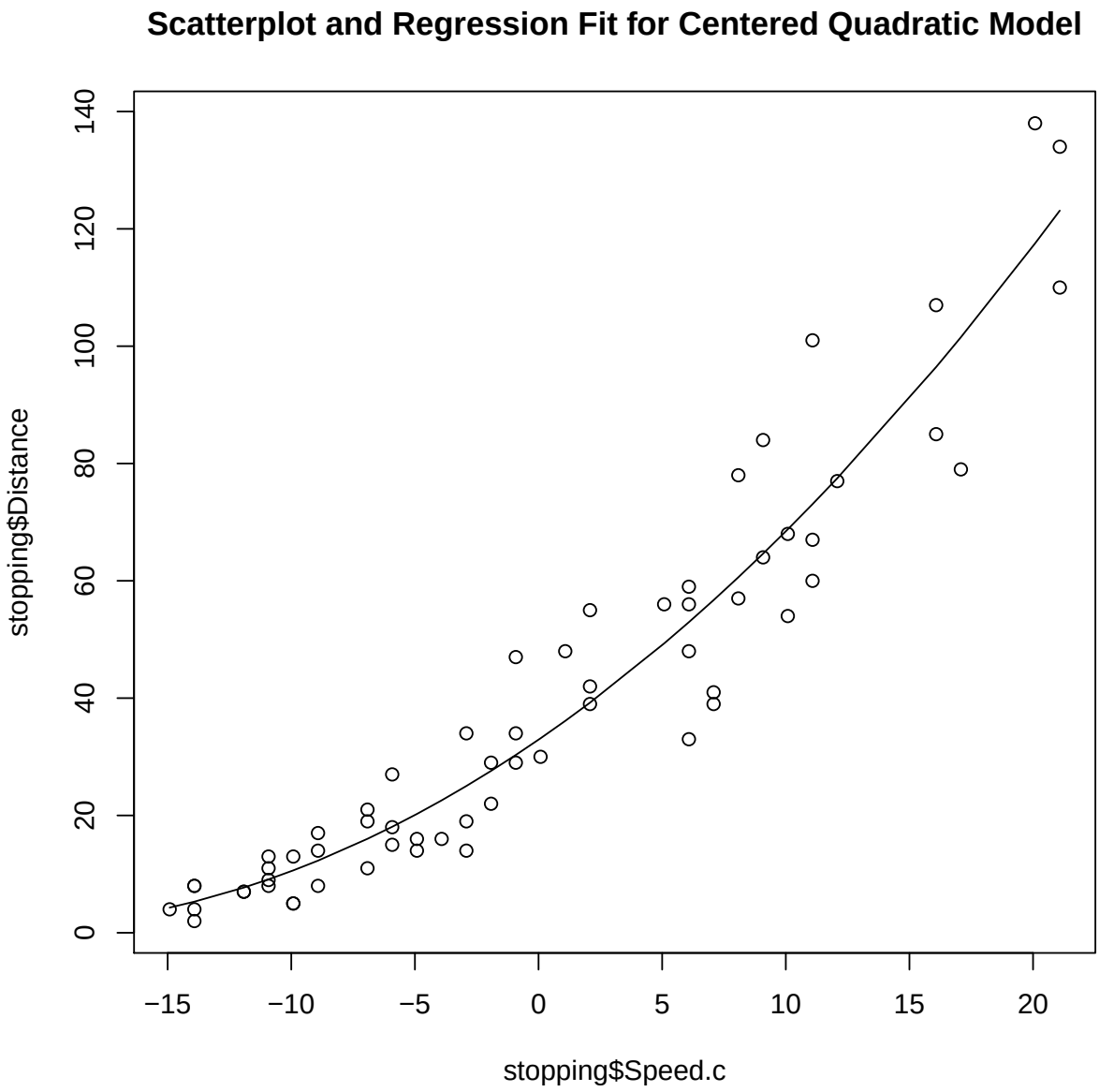


Figure 3: Plot of the data and fitted curve from the quadratic model.

Problem 5

(a) The correlation matrix is provided in Table 1. The majority of correlations are small, but there are a few variables that are strongly correlated. The predictor `Population` in particular is strongly correlated with `Crimes`, `Beds`, `TotalIncome`, and the response `Physicians`. As expected, the other pairs of predictors from this set also show high correlation, but it seems likely that the population size is the driving factor among these variables, although a causal relationship cannot be determined from the data alone. We also see some moderate correlation between `HS` and `Bachelor`, and negative correlation between these and `Poverty` and `Unemployment`.

	Area	Pop	18to34	65plus	Phys.	Beds	Crimes	HS	Bach.	Pov.	Unemp.	Inc.	TotInc.
Area	1.00	0.17	-0.05	0.01	0.08	0.07	0.13	-0.10	-0.14	0.17	0.20	-0.19	0.13
Pop	0.17	1.00	0.08	-0.03	0.94	0.92	0.89	-0.02	0.15	0.04	0.01	0.24	0.99
18to34	-0.05	0.08	1.00	-0.62	0.12	0.07	0.09	0.25	0.46	0.03	-0.28	-0.03	0.07
65plus	0.01	-0.03	-0.62	1.00	-0.00	0.05	-0.04	-0.27	-0.34	0.01	0.24	0.02	-0.02
Phys.	0.08	0.94	0.12	-0.00	1.00	0.95	0.82	-0.00	0.24	0.06	-0.05	0.32	0.95
Beds	0.07	0.92	0.07	0.05	0.95	1.00	0.86	-0.11	0.10	0.17	0.01	0.19	0.90
Crimes	0.13	0.89	0.09	-0.04	0.82	0.86	1.00	-0.11	0.08	0.16	0.04	0.12	0.84
HS	-0.10	-0.02	0.25	-0.27	-0.00	-0.11	-0.11	1.00	0.71	-0.69	-0.59	0.52	0.04
Bach.	-0.14	0.15	0.46	-0.34	0.24	0.10	0.08	0.71	1.00	-0.41	-0.54	0.70	0.22
Poverty	0.17	0.04	0.03	0.01	0.06	0.17	0.16	-0.69	-0.41	1.00	0.44	-0.60	-0.04
Unemp.	0.20	0.01	-0.28	0.24	-0.05	0.01	0.04	-0.59	-0.54	0.44	1.00	-0.32	-0.03
Income	-0.19	0.24	-0.03	0.02	0.32	0.19	0.12	0.52	0.70	-0.60	-0.32	1.00	0.35
TotInc.	0.13	0.99	0.07	-0.02	0.95	0.90	0.84	0.04	0.22	-0.04	-0.03	0.35	1.00

Table 1: Correlation matrix for CDI data.

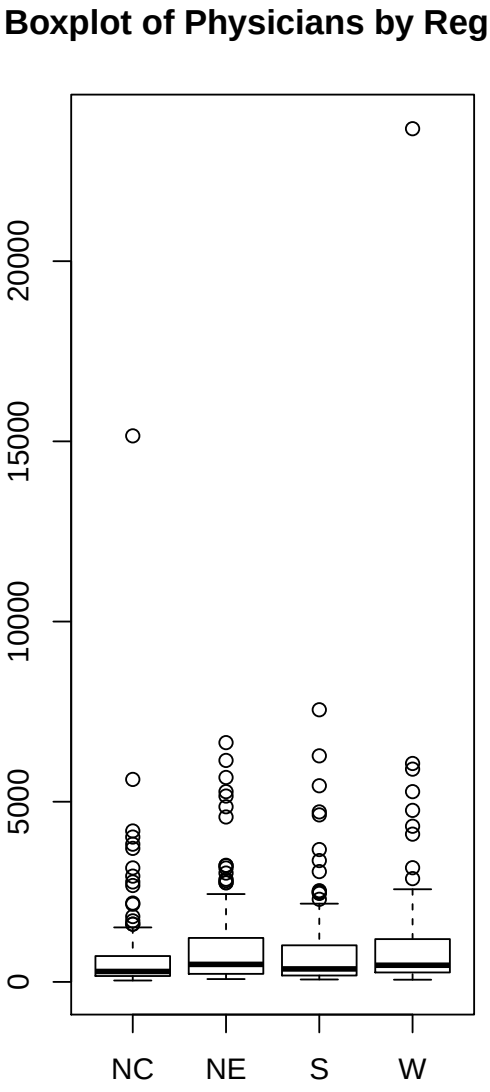
Problem 5 continued on next page...

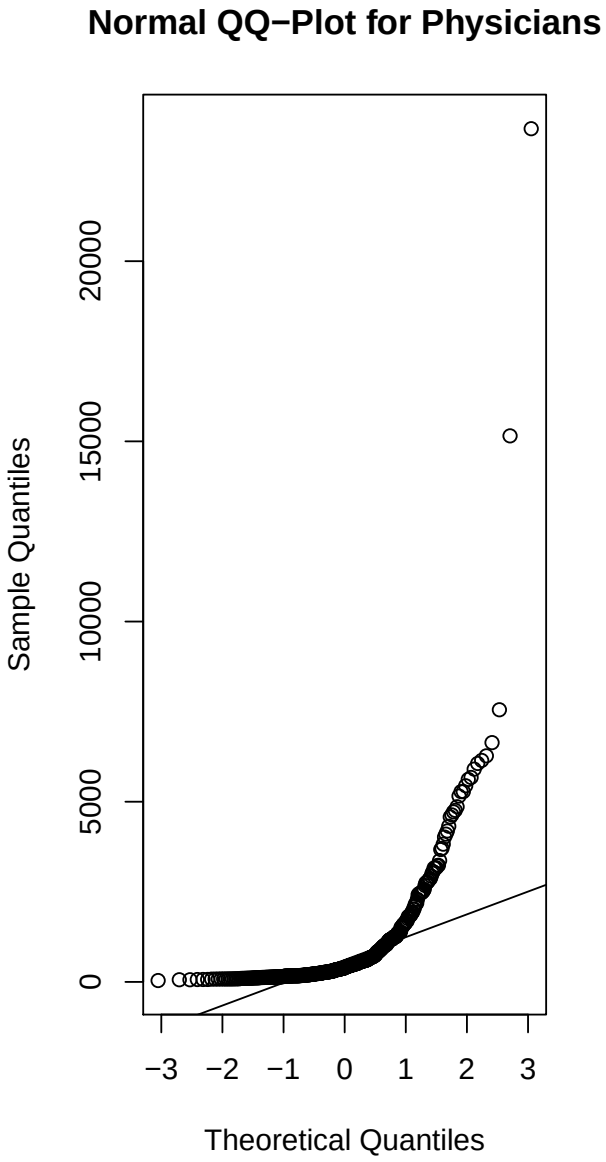
(b) We provide boxplots of the response `Physicians` stratified by `Region`, a QQ-normal plot of `Physicians` and a scatterplot matrix of some potential predictors of interest and the response.

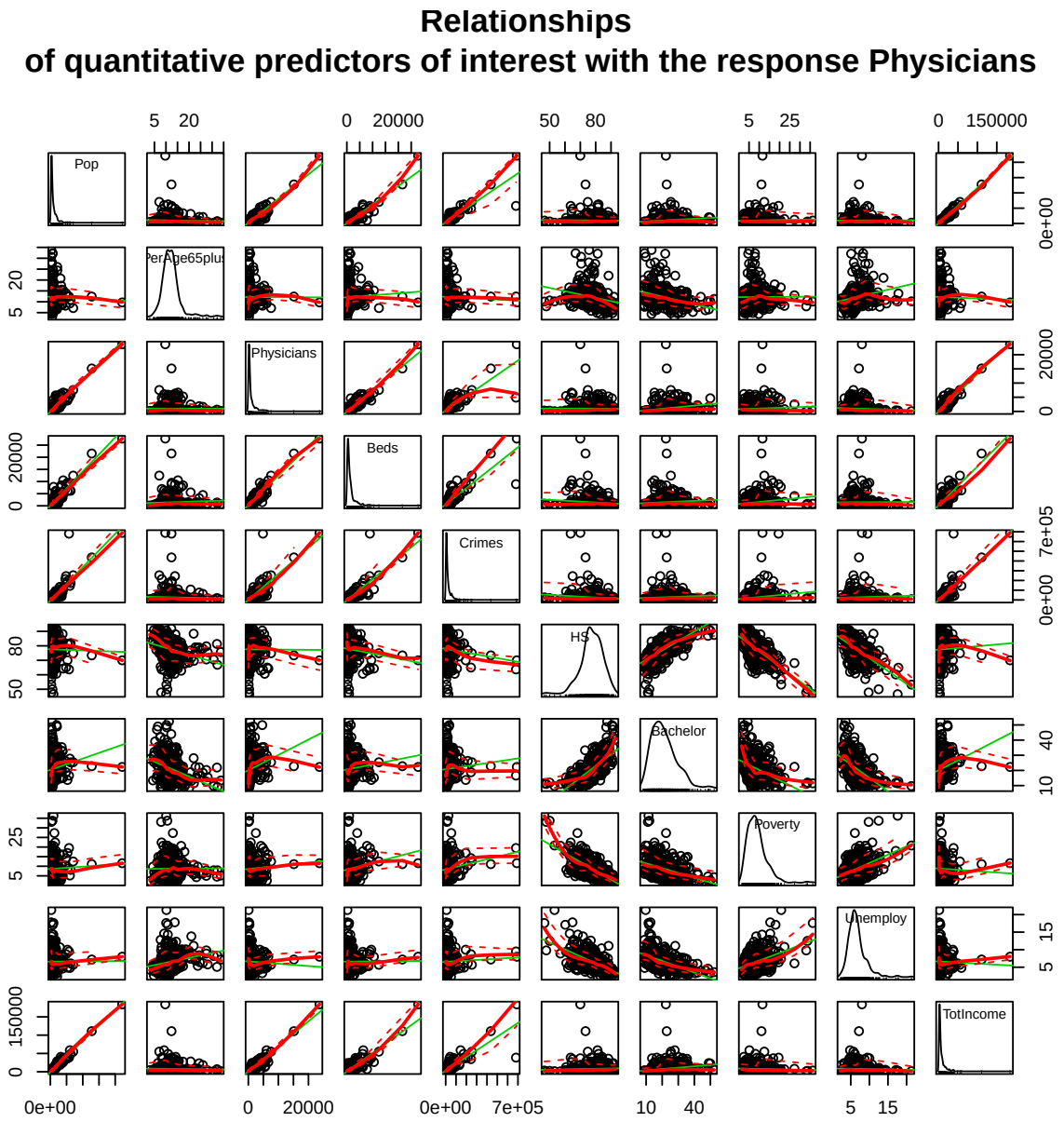
The boxplot of `Physicians` shows strong right-skewing for the response across all regions. This may indicate the need for a transformation to satisfy model assumptions of normality.

The QQ-plot confirms the indication of the boxplots, showing that `Physicians` is strongly right-skewed and nonnormal.

The scatterplot matrix illustrates the correlations of many of the predictors with `Physicians`, most notably `Beds`, `Crimes`, `Pop` and `TotIncome`. Based on this, we expect some or all of these predictors to be present in the chosen model.







(c) Considering the predictors in the data set, the predictors likely to have an effect on the number of physicians in a county are those most related to affluence and population in a county. These include: population, per capita income, total personal income, percent high school graduates and bachelor's degrees. Other predictors likely to be related to physicians in a county are number of hospital beds and percent of the population over 65, since these obviously correlate with need for physicians. We expect an inverse relationship with the level of crimes, unemployment, and percent population aged 18 to 34.

Possible interactions are `Crimes`$-$`Income`; `Income`$-$`Population`; `Crimes`$-$`Bachelor`. These are the variables that may have different effects on the number of physicians based on the level of the associated variable.

(d) Using the `regsubsets()` function to find the best subset of predictors at each number of predictors, the maximum adjusted R^2_{adj} is 0.961, which is achieved for 8 through 12 predictors. The output from the summary is given in Figure 2. Since the R^2_{adj} values are practically equivalent for 8 or more predictors, we

choose the model with 8 predictors for parsimony. The best model for 8 predictors is

$$\texttt{Physicians} \sim \texttt{Pop} + \texttt{Beds} + \texttt{Crimes} + \texttt{HS} + \texttt{Bach} + \texttt{Income} + \texttt{TotIncome} + \texttt{Region}.$$

	model	p	rsq	adjr2	cp	bic	stderr
1	Bd	2.000	0.903	0.903	646.522	-1016.105	556.949
2	Bd-T	3.000	0.948	0.947	153.373	-1279.029	410.724
3	Pp-Bd-T	4.000	0.955	0.955	69.143	-1342.800	379.815
4	Pp-PA1-Bd-T	5.000	0.958	0.957	41.984	-1363.090	369.023
5	Pp-Bd-Bc-I-T	6.000	0.960	0.959	22.031	-1377.960	360.753
6	Pp-Bd-Bc-I-T-RW	7.000	0.961	0.960	12.703	-1383.090	356.595
7	Pp-Bd-H-Bc-I-T-RW	8.000	0.962	0.961	6.491	-1385.318	353.650
8	Pp-Bd-C-H-Bc-I-T-RW	9.000	0.962	0.961	5.530	-1382.268	352.841
9	Pp-PA1-Bd-C-H-Bc-I-T-RW	10.000	0.962	0.961	5.832	-1377.932	352.549
10	Pp-PA1-Bd-C-H-Bc-I-T-RN-RW	11.000	0.962	0.961	6.738	-1372.978	352.505
11	Pp-PA1-Bd-C-H-Bc-U-I-T-RN-RW	12.000	0.962	0.961	8.321	-1367.323	352.744
12	A-Pp-PA1-Bd-C-H-Bc-U-I-T-RN-RW	13.000	0.962	0.961	10.205	-1361.356	353.108

Table 2: Summary of regsubsets() applied to the CDI data.

(e) Using Mallow's C_p as the selection criterion, we look for the model with C_p closest to p. For the CDI data, we again choose the model with 8 predictors selected in (d), which has $C_p = 6.49$.

(f) Using `stepAIC()` in `R` to perform stepwise (backward and forward) selection starting from the full model yields the final model

$$\texttt{Physicians} \sim \texttt{Pop} + \texttt{Beds} + \texttt{Crimes} + \texttt{HS} + \texttt{Bach} + \texttt{Income} + \texttt{TotIncome} + \texttt{Region},$$

which is the same model as selected above. The AIC for this model is 5173.6.

(g) Testing the following interactions in succession using the likelihood ratio test, we find that the interation terms `Crimes : Income`, `Crimes : Bachelor`, `Crimes : TotIncome` and `Beds : Region` are highly significant with $P < 10^{-6}$ for all LRTs. The interaction `Income : Pop` was not found to be significant, so we exclude it from the model.

The model we adopt from these tests is

$$\texttt{Physicians} \sim \texttt{Pop}+\texttt{Beds}+\texttt{Crimes}+\texttt{HS}+\texttt{Bach}+\texttt{Income}+\texttt{TotIncome}+\texttt{Region}+\texttt{Crimes} : \texttt{Income}+\texttt{Crimes} : \texttt{Bachelor}+\texttt{Crimes}$$

This model has $R^2 = 0.9715, R^2_{adj} = 0.97$, from which we conclude that the model is a very good fit, with the predictors explaining 97.15% of the variation in the number of physicians.

Problem 5 continued on next page...

```
fit.int1 <- lm(Physicians ~ Pop + Beds + Crimes + HS + Bachelor + Income +
               TotIncome + Region + Crimes:Income, data=CDI)
anova(base.fit, fit.int1)[2,"Pr(>F)"]

## [1] 0.01475392

fit.int2 <- lm(Physicians ~ Pop + Beds + Crimes + HS + Bachelor + Income +
               TotIncome + Region + Crimes:Income + Income:Pop, data=CDI)
anova(fit.int1, fit.int2)[2,"Pr(>F)"]

## [1] 0.347803

fit.int3 <- lm(Physicians ~ Pop + Beds + Crimes + HS + Bachelor + Income +
               TotIncome + Region + Crimes:Income + Crimes:Bachelor, data=CDI)
anova(fit.int1, fit.int3)[2,"Pr(>F)"]

## [1] 2.09852e-08

fit.int4 <-  lm(Physicians ~ Pop + Beds + Crimes + HS + Bachelor + Income +
               TotIncome + Region + Crimes:Income + Crimes:Bachelor +
                Crimes:TotIncome, data=CDI)
anova(fit.int3, fit.int4)[2,"Pr(>F)"]

## [1] 8.518686e-10

fit.int5 <-  lm(Physicians ~ Pop + Beds + Crimes + HS + Bachelor + Income +
               TotIncome + Region + Crimes:Income + Crimes:Bachelor +
                Crimes:TotIncome + Beds:Region, data=CDI)
anova(fit.int4, fit.int5)[2,"Pr(>F)"]

## [1] 1.504983e-08
```

(h) From the diagnostic plots given in Figure 4, the linear regression assumptions do not appear to met. In particular, the QQ-plot shows significant deviation from normality, and the residuals vs fitted plot shows an extreme right-skewing in the distribution of residuals. These violations are anticipated by the exploratory plots in (b), which shows extreme right-skewing in the response `Physicians`. There are also a number of outliers present in the Cook's Distance plot and standardized residuals plot. To alleviate these violations, we attempt a log transformation of the response.

With the square-root transformation, the response $\sqrt{\texttt{Physicians}}$ is symmetric and much more closely normally distributed. A comparison with the untransformed response is given in Figure 5. The diagnostic plots with the transformed response (Figure 6) show better behavior in the residuals, although there is still significant right-skewing and some violation of normality. The linear regression model is fairly robust to violation of normality, so the model results may be acceptable under this transformation. The R^2 for this model is $R^2 = 0.946$, which is lower from the untransformed model, but we may have more assurance in the validity of the model results since the model assumptions are more closely satisfied. We note that even after transformation, there are a few outliers identified in the diagnostic plots, most significantly observations 1, 3, 6, 48, and 418. Further analysis could look more closely at the effect of these outliers on the fitted model.

The log transformation was also tried for the reponse, which produced even better behavior in the residuals, but the R^2 of the resulting model was reduced to $R^2 = 0.85$, which is a large drop compared to the square

root transformation. Due to the significantly greater R^2, we adopt the square root transformation despite the minor violations in the model assumptions.

Taking into consideration all of the above analysis, the best model is

$$\begin{aligned}\sqrt{\texttt{Physicians}} \sim &\texttt{Pop} + \texttt{Beds} + \texttt{Crimes} + \texttt{HS} + \texttt{Bach} + \texttt{Income} + \texttt{TotIncome} \\ &+ \texttt{Region} + \texttt{Crimes} : \texttt{Income} + \texttt{Crimes} : \texttt{Bachelor} + \texttt{Crimes} : \texttt{TotIncome} + \texttt{Beds} : \texttt{Region}.\end{aligned}$$

However, we may wish to report this model with untransformed response since the interpretation is much clearer, and the R^2 is higher, including the caveat that the model assumptions may be violated. This model largely corroborates our intuition about what predictors will be significantly related to the number of physicians in a county. We see that population, education levels, region, and crime levels are the most important predictors in the number of physicians in a county. Model summary results are given in Table 3. A few things of note from the summary of the fitted model:

- The predictors with the largest effects are `HS`, `Bachelor`, `RegionNE`, `RegionW`.
- Even though `Crimes` : `Income`, `Crimes` : `Bachelor` and `Crimes` : `TotIncome` were found to be significant by the LRT, the effect estimates of these interaction terms is practically 0.
- The `Beds` : `Region` interaction terms have small but statistically significant effects.
- `RegionW` has the largest effect estimate, perhaps due to the population density in California, but the effect is not found to be statistically significant when adjusting for the other terms in the model. This may be because the effect is mainly accounted for by the `Beds` : `Region` interaction term.
- Positively correlated predictors with moderate to large effects are: `RegionW`, `Bachelor`, `Beds` and `Beds` : `RegionW`.
- Negatively correlated predictors with moderate to large effects are: `HS`, `RegionNE`.

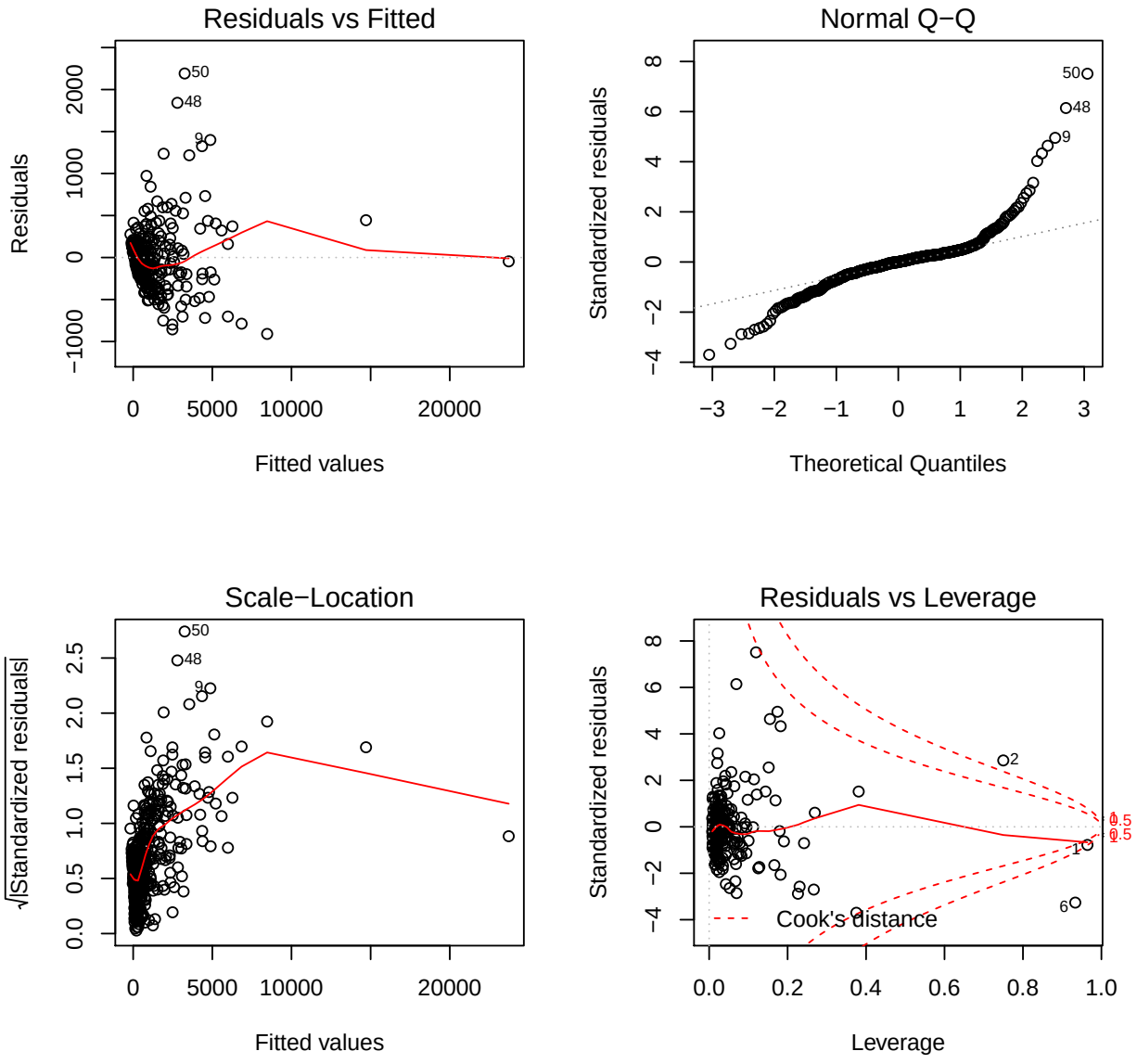


Figure 4: Diagnostic plots from selected model fit, no transformation on the response.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	590.1512	254.0297	2.32	0.0206
Pop	-0.0038	0.0004	-8.91	0.0000
Beds	0.5973	0.0252	23.66	0.0000
Crimes	0.0227	0.0050	4.57	0.0000
HS	-8.5812	3.5472	-2.42	0.0160
Bachelor	11.2316	4.2722	2.63	0.0089
Income	-0.0123	0.0081	-1.52	0.1296
TotIncome	0.2125	0.0200	10.60	0.0000
RegionNE	-39.3526	56.6530	-0.69	0.4877
RegionS	8.6167	52.2372	0.16	0.8691
RegionW	78.8651	57.1228	1.38	0.1681
Crimes:Income	-0.0000	0.0000	-6.54	0.0000
Crimes:Bachelor	0.0009	0.0001	6.94	0.0000
Crimes:TotIncome	0.0000	0.0000	1.59	0.1127
Beds:RegionNE	0.0576	0.0239	2.41	0.0166
Beds:RegionS	-0.0264	0.0213	-1.24	0.2175
Beds:RegionW	0.1267	0.0248	5.10	0.0000

Table 3: Summary of final model fit with untransformed response.

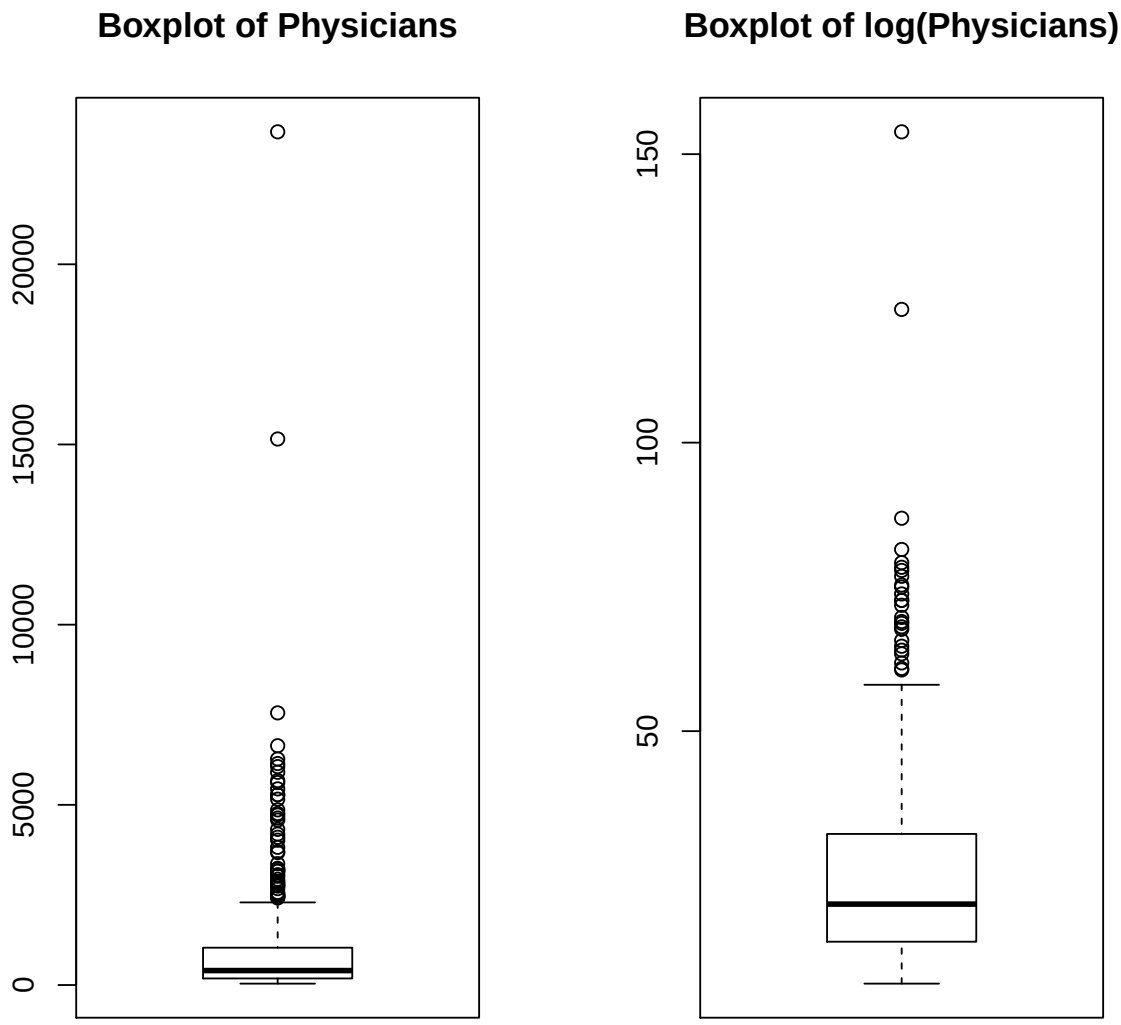


Figure 5: Boxplots comparing the response `Physicians` and $\sqrt{\texttt{Physicians}}$.

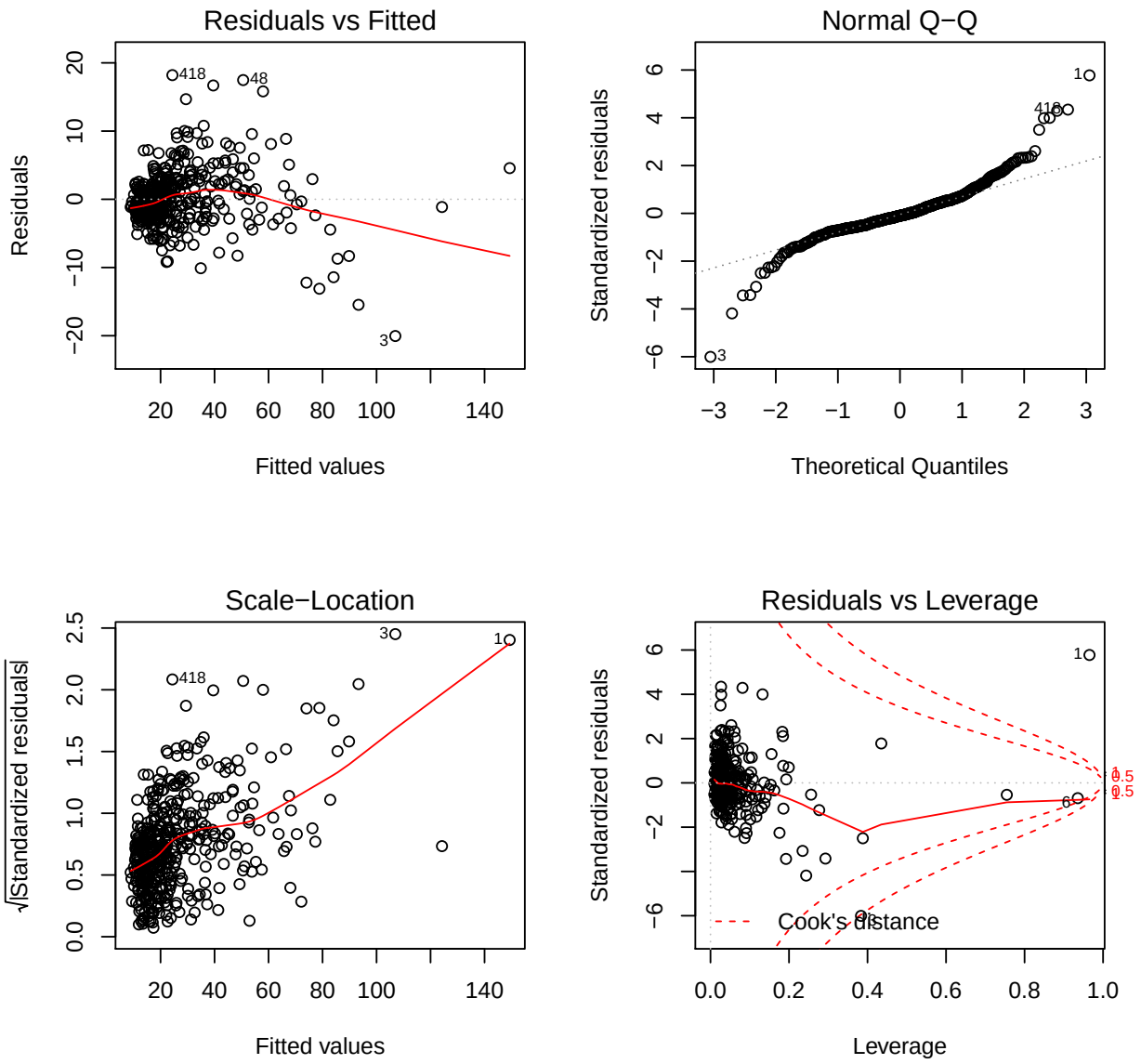


Figure 6: Diagnostic plots from the model with the selected predictors and square root transformation on the response.