Chapter 7

Case Study in Least Squares Fitting and Interpretation of a Linear Model

This chapter presents some of the stages of modeling, using a linear multiple regression model whose coefficients are estimated using ordinary least squares. The data are taken from the 1994 version of the *City and County Databook* compiled by the Geospatial and Statistical Data Center of the University of Virginia Library and available at `fisher.lib.virginia.edu/ccdb`. Most of the variables come from the U.S. Census[a]. Variables related to the 1992 U.S. presidential election were originally provided and copyrighted by the Elections Research Center and are taken from [365], with permission from the Copyright Clearance Center. The data extract analyzed here is available from this text's Web site (see Appendix). The data did not contain election results from the 25 counties of Alaska. In addition, two other counties had zero voters in 1992. For these the percent voting for each of the candidates was also set to missing. The 27 counties with missing percent votes were excluded when fitting the multivariable model.

The dependent variable is taken as the percentage of voters voting for the Democratic Party nominee for President of the U.S. in 1992, Bill Clinton, who received

---

[a]U.S. Bureau of the Census. 1990 Census of Population and Housing, Population and Housing Unit Counts, United States (CPH-2-1.), and Data for States and Counties, Population Division, July 1, 1992, Population Estimates for Counties, Including Components of Change, PPL-7.

43.0% of the vote according to this dataset. The Republican Party nominee George Bush received 37.4%, and the Independent candidate Ross Perot received 18.9% of the vote. Republican and Independent votes tended to positively correlate over the counties.

To properly answer questions about voting patterns of individuals, subject-level data are needed. Such data are difficult to obtain. Analyses presented here may shed light on individual tendencies but are formally a characterization of the 3141 counties (and selected other geographic regions) in the United States. As virtually all of these counties are represented in the analyses, the sample is in a sense the whole population so inferential statistics (test statistics and confidence bands) are not strictly required. These are presented anyway for illustration.

There are many aspects of least squares model fitting that are not considered in this chapter. These include assessment of *groups* of overly influential observations, and robust estimation. The reader should refer to one of the many excellent texts on linear models for more information on these and other methods dedicated to such models.

7.1 Descriptive Statistics

First we print basic descriptive statistics using the `Hmisc` library's `describe` function.[b]

```
> library(Hmisc,T);  library(Design,T)
> describe(counties[,-(1:4)])          # omit first 4 vars.
```

counties
15 Variables   3141 Observations

---

pop.density : 1992 pop per 1990 miles^2

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
3141	0	541	222.9	2	4	16	39	96	297	725

lowest :   0   1   2   3   4
highest: 15609 17834 28443 32428 52432

---

pop : 1990 population

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
3141	0	3078	79182	3206	5189	10332	22085	54753	149838	320167

lowest :   52   107   130   354   460
highest: 2410556 2498016 2818199 5105067 8863164

---

[b]For a continuous variable, `describe` stores frequencies for 100 bins of the variable. This information is shown in a histogram that is added to the text when the **latex** method is used on the object created by `describe`. The output produced here was created by `latex(describe(counties[,-(1:4)], descript='counties'))`.

pop.change : % population change 1980-1992

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	768	6.501	-16.7	-13.0	-6.0	2.7	13.3	29.6	43.7

lowest : -34.4 -32.2 -31.6 -31.3 -30.2
highest: 146.5 152.2 181.7 191.4 207.7

---

age6574 : % age 65-74, 1990

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	153	8.286	4.9	5.7	6.9	8.2	9.5	10.9	11.9

lowest : 0.6 0.9 1.8 1.9 2.0, highest: 19.8 20.0 20.6 20.9 21.1

---

age75 : % age ≥ 75, 1990

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	144	6.578	3.1	3.9	5.0	6.3	7.9	9.9	11.3

lowest : 0.0 0.3 0.5 0.8 0.9, highest: 14.9 15.2 15.4 15.5 15.9

---

crime : serious crimes per 100,000 1991

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	2339	3008	0	286	1308	2629	4243	6157	7518

lowest :   0   39   40   41   44
highest: 13229 13444 14016 16031 20179

---

college : % with bachelor's degree or higher of those age≥25

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	322	13.51	6.6	7.5	9.2	11.8	15.6	21.9	27.1

lowest : 0.0 3.7 4.0 4.1 4.2, highest: 49.8 49.9 52.3 52.8 53.4

---

income : median family income, 1989 dollars

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	2927	28476	19096	20904	23838	27361	31724	36931	41929

lowest : 10903 11110 11362 11502 12042
highest: 61988 62187 62255 62749 65201

---

farm : farm population, % of total, 1990

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	302	6.437	0.1	0.4	1.5	3.9	8.6	16.5	21.4

lowest : 0.0 0.1 0.2 0.3 0.4, highest: 50.9 54.6 55.0 65.8 67.6

---

democrat : % votes cast for democratic president

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3114	27	530	39.73	22.80	27.03	32.70	39.00	46.00	53.80	58.84

lowest : 6.8 9.5 12.9 13.0 13.6, highest: 79.2 79.4 79.6 82.8 84.6

---

republican : % votes cast for republican president

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3114	27	431	39.79	26.67	29.50	33.80	39.20	45.50	50.90	54.80

lowest : 9.1 12.9 13.1 13.6 13.9, highest: 68.0 68.1 69.1 72.2 75.0

Perot : % votes cast for Ross Perot

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3114	27	316	19.81	8.765	10.400	14.400	20.300	25.100	28.500	30.600

lowest : 3.2 3.3 3.4 3.6 3.7, highest: 37.7 39.0 39.8 40.4 46.9

white : % white, 1990

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3141	0	3133	87.11	54.37	64.44	80.43	94.14	98.42	99.32	99.54

lowest :   5.039   5.975   10.694   13.695   13.758
highest: 99.901   99.903   99.938   99.948 100.000

black : % black, 1990

	n	missing	unique	Mean	.05	.10	.25	.50	.75
	3141	0	3022	8.586	0.01813	0.04452	0.16031	1.49721	10.00701

.90	.95
30.72989	41.69317

lowest :   0.000000   0.007913   0.008597   0.009426   0.009799
highest: 79.445442 80.577171 82.145996 85.606544 86.235985

turnout : 1992 votes for president / 1990 pop x 100

	n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
	3116	25	3113	44.06	31.79	34.41	39.13	44.19	49.10	53.13	55.71

lowest :   0.000   7.075   14.968   16.230   16.673
highest: 72.899   75.027   80.466   89.720 101.927

Of note is the incredible skewness of population density across counties. This variable will cause problems when displaying trends graphically as well as possibly causing instability in fitting spline functions. Therefore we transform it by taking $\log_{10}$ after adding one to avoid taking the log of zero. We compute one other derived variable—the proportion of county residents with age of at least 65 years. Then the `datadist` function from the `Design` library is run to compute covariable ranges and settings for constructing graphs and estimating effects of predictors.

```
> older ← counties$age6574 + counties$age75
> label(older) ← '% age >= 65, 1990'
> pdensity ← logb(counties$pop.density+1, 10)
> label(pdensity) ← 'log 10 of 1992 pop per 1990 miles^2'

> dd ← datadist(counties)
> dd ← datadist(dd, older, pdensity)   # add 2 vars. not in data frame
> options(datadist='dd')
```

Next, examine how some of the key variables interrelate, using hierarchical variable clustering based on squared Spearman rank correlation coefficients as similarity measures.
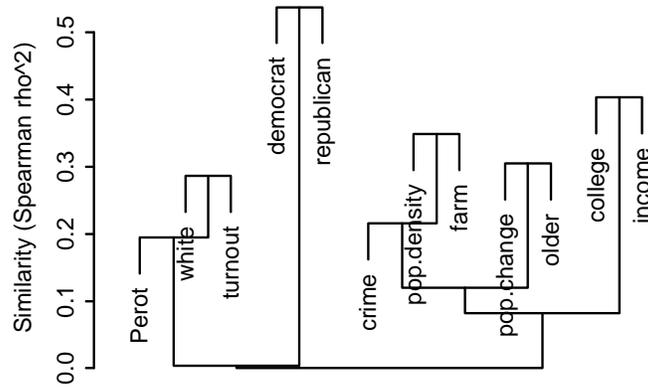
FIGURE 7.1: Variable clustering of some key variables in the `counties` dataset.

```
> v ← varclus(∼ pop.density + pop.change + older + crime + college +
+               income + farm + democrat + republican + Perot +
+               white + turnout, data=counties)
> plot(v)                           # Figure 7.1
```

The percentage of voters voting Democratic is strongly related to the percentage voting Republican because of the strong negative correlation between the two. The Spearman ρ^2 between percentage of residents at least 25 years old who are college educated and median family income in the county is about 0.4.

Next we examine descriptive associations with the dependent variable, by stratifying separately by key predictors, being careful not to use this information in formulating the model because of the phantom degrees of freedom problem.

```
> s ← summary(democrat ∼ pop.density + pop.change + older + crime +
+               college + income + farm + white + turnout,
+               data=counties)
> plot(s, cex.labels=.7)                    # Figure 7.2
```

There is apparently no "smoking gun" predictor of extraordinary strength although all variables except age and crime rate seem to have some predictive ability. The voter turnout (bottom variable) is a strong and apparently monotonic factor. Some of the variables appear to predict Democratic votes nonmonotonically (see especially population density). It will be interesting to test whether voter turnout is merely a reflection of the county demographics that, when adjusted for, negate the association between voter turnout and voter choice.
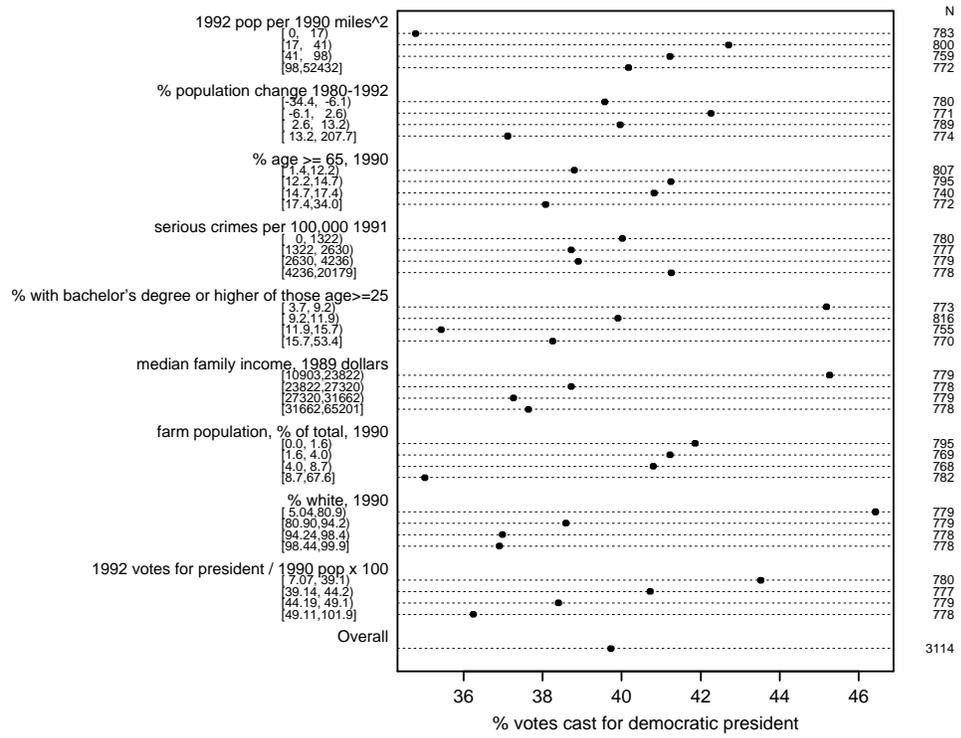
FIGURE 7.2: Percentage of votes cast for Bill Clinton stratified separately by quartiles of other variables. Sample sizes are shown in the right margin.

	N df
crime	3114 2
older	3114 2
pop.change	3114 2
turnout	3114 2
farm	3114 2
pop.density	3114 2
income	3114 2
college	3114 2
white	3114 2
black	3114 2

Adjusted rho^2
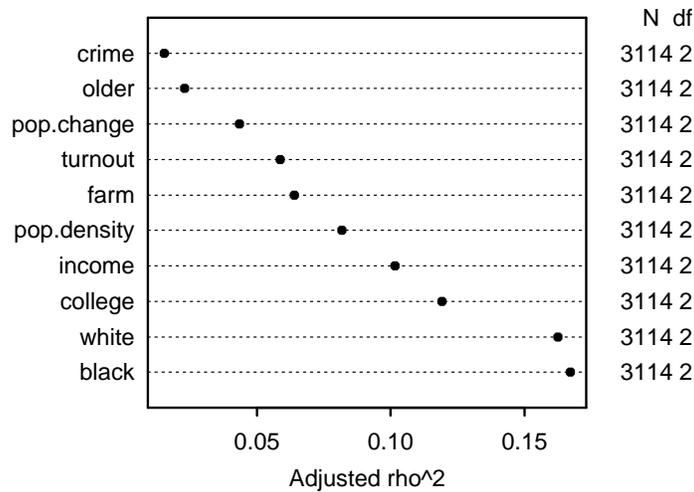
FIGURE 7.3: Strength of marginal relationships between predictors and response using generalized Spearman χ^2.

7.2   Spending Degrees of Freedom/Specifying Predictor Complexity

As described in Section 4.1, in the absence of subject matter insight we might spend degrees of freedom according to estimates of strengths of relationships without a severe "phantom d.f." problem, as long as our assessment is masked to the contributions of particular parameters in the model (e.g., linear vs. nonlinear effects). The following S-PLUS code computes and plots the nonmonotonic (quadratic in ranks) generalization of the Spearman rank correlation coefficient, separately for each of a series of prespecified predictor variables.

```
> s ← spearman2(democrat ∼ pop.density + pop.change + older + crime +
+               college + income + farm + black + white + turnout,
+               data=counties, p=2)
> plot(s)                        # Figure 7.3
```

From Figure 7.3 we guess that lack of fit will be more consequential (in descending order of importance) for racial makeup, college education, income, and population density.

7.3   Fitting the Model Using Least Squares

A major issue for continuous Y is always the choice of the Y-transformation. When the raw data are percentages that vary from 30 to 70% all the way to nearly 0 or 100%, a transformation that expands the tails of the Y distribution, such as the arcsine square root, logit, or probit, often results in a better fit with more normally distributed residuals. The percentage of a county's voters who participated is centered around the median of 39% and does not have a very large number of counties near 0 or 100%. Residual plots were no more normal with a standard transformation than that from untransformed Y. So we use untransformed percentages.

We use the linear model

$$E(Y|X) = X\beta, \tag{7.1}$$

where β is estimated using ordinary least squares, that is, by solving for $\hat{\beta}$ to minimize $\sum(Y_i - X\hat{\beta})^2$. If we want to compute P-values and confidence limits using parametric methods we would have to assume that $Y|X$ is normal with mean $X\beta$ and constant variance σ^2 (the latter assumption may be dispensed with if we use a robust Huber–White or bootstrap covariance matrix estimate—see Section 9.5). This assumption is equivalent to stating the model as conditional on X,

$$Y = X\beta + \epsilon, \tag{7.2}$$

where ϵ is normally distributed with mean zero, constant variance σ^2, and residuals $Y - E(Y|X)$ are independent across observations.

To not assume linearity the Xs above are expanded into restricted cubic spline functions, with the number of knots specified according the estimated "power" of each predictor. Let us assume that the most complex relationship could be fitted adequately using a restricted cubic spline function with five knots. `crime` is thought to be so weak that linearity is forced. Note that the term "linear model" is a bit misleading as we have just made the model as nonlinear in X as desired.

We prespecify one second-order interaction, between `income` and `college`. To save d.f. we fit a "nondoubly nonlinear" restricted interaction as described in Equation 2.38, using the `Design` library's `%ia%` function. Default knot locations, using quantiles of each predictor's distribution, are chosen as described in Section 2.4.5.

```
> f ← ols(democrat ∼ rcs(pdensity,4) + rcs(pop.change,3) +
+          rcs(older,3) + crime + rcs(college,5) + rcs(income,4) +
+          rcs(college,5) %ia% rcs(income,4) +
+          rcs(farm,3) + rcs(white,5) + rcs(turnout,3))
> f
```

Linear Regression Model

Frequencies of Missing Values Due to Each Variable
 democrat pdensity pop.change older crime college income farm white turnout
      27   0   0   0   0   0   0   0   0   25

    n Model L.R. d.f.   R2 Sigma
 3114   2210   29 0.5082 7.592

Residuals:
    Min   1Q Median   3Q   Max
 -30.43 -4.978 -0.299 4.76 31.99

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
Intercept	6.258e+01	9.479e+00	6.602144	4.753e-11
pdensity	1.339e+01	9.981e-01	13.412037	0.000e+00
pdensity'	-1.982e+01	2.790e+00	-7.103653	1.502e-12
pdensity''	7.637e+01	1.298e+01	5.882266	4.481e-09
pop.change	-2.323e-01	2.577e-02	-9.013698	0.000e+00
pop.change'	1.689e-01	2.862e-02	5.900727	4.012e-09
older	5.037e-01	1.042e-01	4.833013	1.411e-06
older'	-5.134e-01	1.104e-01	-4.649931	3.460e-06
crime	1.652e-05	8.224e-05	0.200837	8.408e-01
college	5.205e-01	1.184e+00	0.439539	6.603e-01
college'	-8.738e-01	2.243e+01	-0.038962	9.689e-01
college''	7.330e+01	6.608e+01	1.109281	2.674e-01
college'''	-1.246e+02	5.976e+01	-2.084648	3.718e-02
income	1.714e-05	4.041e-04	0.042410	9.662e-01
income'	-6.372e-03	1.490e-03	-4.275674	1.963e-05
income''	1.615e-02	4.182e-03	3.861556	1.150e-04
college * income	-8.525e-05	5.097e-05	-1.672504	9.453e-02
college * income'	7.729e-04	1.360e-04	5.684197	1.437e-08
college * income''	-1.972e-03	3.556e-04	-5.545263	3.183e-08
college' * income	-9.362e-05	8.968e-04	-0.104389	9.169e-01
college'' * income	-2.067e-03	2.562e-03	-0.806767	4.199e-01
college''' * income	3.934e-03	2.226e-03	1.767361	7.727e-02
farm	-5.305e-01	9.881e-02	-5.368650	8.521e-08
farm'	4.454e-01	1.838e-01	2.423328	1.544e-02
white	-3.533e-01	2.600e-02	-13.589860	0.000e+00
white'	2.340e-01	5.012e-02	4.668865	3.158e-06
white''	-1.597e+00	9.641e-01	-1.656138	9.780e-02
white'''	-1.740e+01	1.648e+01	-1.055580	2.912e-01
turnout	-7.522e-05	4.881e-02	-0.001541	9.988e-01
turnout'	1.692e-01	4.801e-02	3.524592	4.303e-04

```
Residual standard error: 7.592 on 3084 degrees of freedom
Adjusted R-Squared: 0.5036
```

The analysis discarded 27 observations (most of them from Alaska) having missing data, and used the remaining 3114 counties. The proportion of variation across counties explained by the model is $R^2 = 0.508$, with adjusted $R^2 = 0.504$. The estimate of σ (7.59%) is obtained from the unbiased estimate of σ^2. For the linear model the likelihood ratio statistic is $-n \log(1 - R^2)$, which here is $-3114 \log(1 - 0.5082) = 2210$ on 29 d.f. The ratio of observations to variables is 3114/29 or 107, so there is no issue with overfitting.[c]

In the above printout, primes after variable names indicate cubic spline components (see Section 2.4.4). The most compact algebraic form of the fitted model appears below, using Equation 2.26 to simplify restricted cubic spline terms.

```
> latex(f)
```

$$\mathrm{E(democrat)} = X\beta, \quad \text{where}$$

$X\hat{\beta} =$

62.57849

$+13.38714 \mathrm{pdensity} - 3.487746(\mathrm{pdensity} - 0.4771213)_+^3$

$+13.43985(\mathrm{pdensity} - 1.39794)_+^3 - 10.82831(\mathrm{pdensity} - 1.812913)_+^3$

$+0.8761998(\mathrm{pdensity} - 2.860937)_+^3$

$-0.2323114 \mathrm{pop.change} + 9.307077 \times 10^{-5}(\mathrm{pop.change} + 13)_+^3$

$-0.0001473909(\mathrm{pop.change} - 2.7)_+^3 + 5.432011 \times 10^{-5}(\mathrm{pop.change} - 29.6)_+^3$

$+0.5037175 \mathrm{older} - 0.004167098(\mathrm{older} - 9.6)_+^3 + 0.007460448(\mathrm{older} - 14.5)_+^3$

$-0.003293351(\mathrm{older} - 20.7)_+^3 + 1.651695 \times 10^{-5} \mathrm{crime}$

$+0.5205324 \mathrm{college} - 0.002079334(\mathrm{college} - 6.6)_+^3 + 0.17443(\mathrm{college} - 9.45)_+^3$

$-0.2964471(\mathrm{college} - 11.8)_+^3 + 0.123932(\mathrm{college} - 15)_+^3$

$+0.0001644174(\mathrm{college} - 27.1)_+^3$

$+1.71383 \times 10^{-5} \mathrm{income} - 1.222161 \times 10^{-11}(\mathrm{income} - 19096)_+^3$

$+3.097825 \times 10^{-11}(\mathrm{income} - 25437)_+^3 - 1.925238 \times 10^{-11}(\mathrm{income} - 29887)_+^3$

$+4.957448 \times 10^{-13}(\mathrm{income} - 41929)_+^3$

$+\mathrm{income}[-8.52499 \times 10^{-5} \mathrm{college} - 2.22771 \times 10^{-7}(\mathrm{college} - 6.6)_+^3$

$-4.919284 \times 10^{-6}(\mathrm{college} - 9.45)_+^3 + 9.360726 \times 10^{-6}(\mathrm{college} - 11.8)_+^3$

$-4.283218 \times 10^{-6}(\mathrm{college} - 15)_+^3 + 6.454693 \times 10^{-8}(\mathrm{college} - 27.1)_+^3]$

---

[c]This can also be assessed using the heuristic shrinkage estimate $(2210 - 29)/2210 = 0.987$, another version of which is proportional to the ratio of adjusted to ordinary R^2 as given on p. 64. The latter method yields $(3114 - 29 - 1)/(3114 - 1) \times 0.5036/0.5082 = 0.982$.

$$+\text{college}[1.482526\times10^{-12}(\text{income}-19096)_+^3 - 3.781803\times10^{-12}(\text{income}-25437)_+^3$$
$$+2.368292\times10^{-12}(\text{income}-29887)_+^3 - 6.901521\times10^{-14}(\text{income}-41929)_+^3]$$
$$-0.5304876\text{farm} + 0.00171818(\text{farm}-0.4)_+^3 - 0.002195452(\text{farm}-3.9)_+^3$$
$$+0.0004772722(\text{farm}-16.5)_+^3$$
$$-0.353288\text{white} + 0.0001147081(\text{white}-54.37108)_+^3$$
$$-0.0007826866(\text{white}-82.81484)_+^3 - 0.008527786(\text{white}-94.1359)_+^3$$
$$+0.03878391(\text{white}-98.14566)_+^3 - 0.02958815(\text{white}-99.53718)_+^3$$
$$-7.522335\times10^{-5}\text{ turnout} + 0.0004826373(\text{turnout}-34.40698)_+^3$$
$$-0.001010226(\text{turnout}-44.18553)_+^3 + 0.000527589(\text{turnout}-53.13093)_+^3$$

and $(x)_+ = x$ if $x > 0$, 0 otherwise.

Interpretation and testing of individual coefficients listed above is not recommended except for the coefficient of the one linear effect in the model (for `crime`) and for nonlinear effects when there is only one of them (i.e., for variables modeled with three knots). For `crime`, the two-tailed t-test of partial association resulted in $P = 0.8$. Other effects are better interpreted through predicted values as shown in Section 7.8.

7.4 Checking Distributional Assumptions

As mentioned above, *if* one wanted to use parametric inferential methods on the least squares parameter estimates, and to have confidence that the estimates are efficient, certain assumptions must be validated: (1) the residuals should have no systematic trend in central tendency against any predictor variable or against $\hat{Y}$; (2) the residuals should have the same dispersion for all levels of $\hat{Y}$ and of individual values of X; and (3) the residuals should have a normal distribution, both overall and for any subset in the X-space. Our first assessment addresses elements (1) and (2) by plotting the median and lower and upper quartiles of the residuals, stratified by intervals of $\hat{Y}$ containing 200 observations.[d]

```
> r ← resid(f)
> xYplot(r ∼ fitted(f), method='quantile', nx=200,
+        ylim=c(-10,10), xlim=c(20,60),
+        abline=list(h=0, lwd=.5, lty=2),
+        aspect='fill')                      # Figure 7.4
```

No trends of concern are apparent in Figure 7.4; variability appears constant. This same kind of graph should be done with respect to the predictors. Figure 7.5 shows

---

[d]The number of observations is too large for a scatterplot.
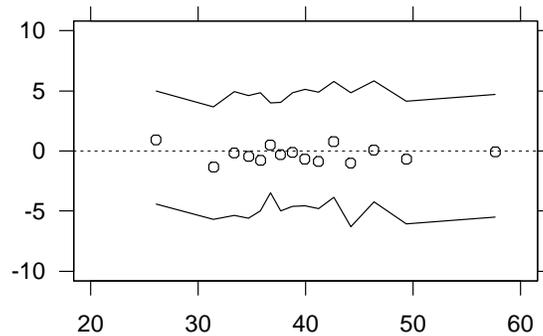
FIGURE 7.4: Quartiles of residuals from the linear model, stratifying $\hat{Y}$ into intervals containing 200 counties each. For each interval the x-coordinate is the mean predicted percentage voting Democratic over the counties in that interval. S-Plus trellis graphics are used through the `Hmisc` library `xYplot` function.

the results for two of the most important predictors. Again, no aspect of the graphs causes concern.

```
> p1 ← xYplot(r ∼ white, method='quantile', nx=200,
+             ylim=c(-10,10), xlim=c(40,100),
+             abline=list(h=0, lwd=.5, lty=2),
+             aspect='fill')

> p2 ← xYplot(r ∼ pdensity, method='quantile', nx=200,
+             ylim=c(-10,10), xlim=c(0,3.5),
+             abline=list(h=0, lwd=.5, lty=2),
+             aspect='fill')

> print(p1, split=c(1,1,1,2), more=T)        # 1 column, 2 rows
> print(p2, split=c(1,2,1,2))                # Figure 7.5
```

For the assessment of normality of residuals we use q–q plots which are straight lines if normality holds. Figure 7.6 shows q–q plots stratified by quartiles of population density.

```
> qqmath(∼r | cut2(pdensity,g=4))            # Figure 7.6
```

Each graph appears sufficiently linear to make us feel comfortable with the normality assumption should we need it to hold.
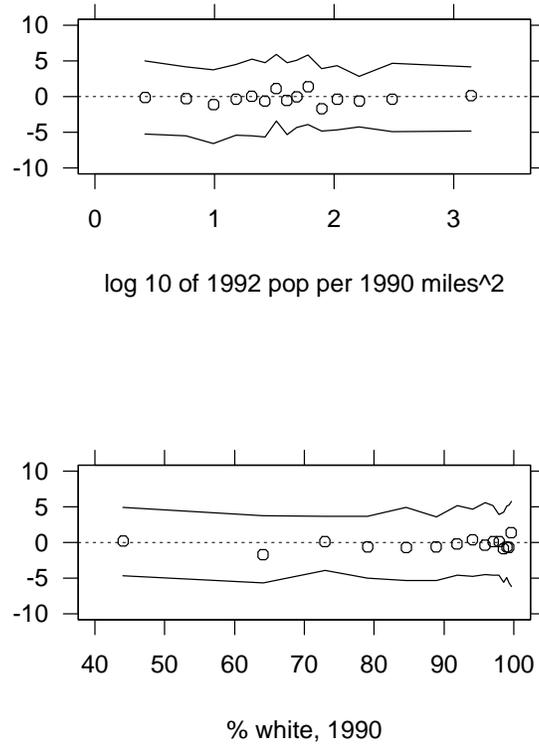
FIGURE 7.5: Quartiles of residuals against population density (top panel) and % white (bottom panel).
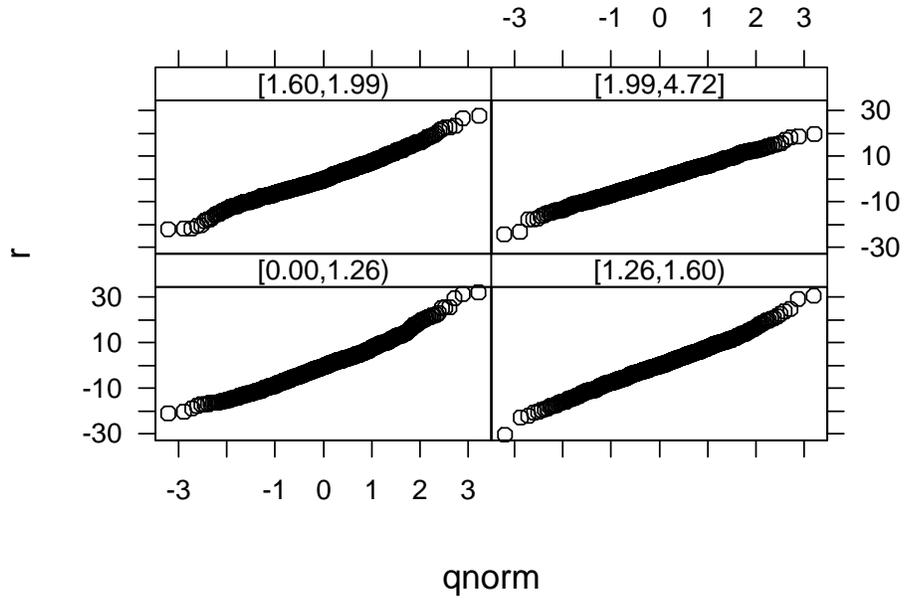
FIGURE 7.6:   Quantile–quantile plot for estimated residuals stratified by quartiles of population density.

7.5   Checking Goodness of Fit

Flexible specification of main effects (without assuming linearity) and selected interaction effects were built into the model. The principal lack of fit would be due to interactions that were not specified. To test the importance of all such (two-way, at least) interactions, including generalizing the `income` × `college` interaction, we can fit a linear model with all two-way interactions:

```
> f2 ← ols(democrat ~ (rcs(pdensity,4) + rcs(pop.change,3) +
+           rcs(older,3) + crime + rcs(college,5) + rcs(income,4) +
+           rcs(farm,3) + rcs(white,5) + rcs(turnout,3))^2)
> f2$stats

    n Model L.R.  d.f.     R2 Sigma
 3114        2974  254 0.6152 6.975
```

The F test for goodness of fit can be done using this model's R^2 and that of the original model ($R^2 = 0.5082$ on 29 d.f.). The F statistic for testing two nested models is

$$F_{k,n-p-1} = \frac{\frac{R^2 - R_*^2}{k}}{\frac{1-R^2}{n-p-1}}, \tag{7.3}$$

where R^2 is from the full model, R_*^2 is from the submodel, p is the number of regression coefficients in the full model (excluding the intercept, here 254), and k is the d.f. of the full model minus the d.f. of the submodel (here, $254 - 29$). Here $F_{225,2860} = 3.54, P < 0.0001$, so there is strong statistical evidence of a lack of fit from some two-way interaction term. Subject matter input should have been used to specify more interactions likely to be important. At this point, testing a multitude of two-way interactions without such guidance is inadvisable, and we stay with this imperfect model. To gauge the impact of this decision on a scale that is more relevant than that of statistical significance, the median absolute difference in predicted values between our model and the all-two-way-interaction model is 2.02%, with 369 of the counties having predicted values differing by more than 5%.

7.6   Overly Influential Observations

Below are observations that are overly influential when considered singly. An asterisk is placed next to a variable when any of the coefficients associated with that variable changed by more than 0.3 standard errors upon removal of that observation. DFFITS is also shown.

```
> g ← update(f, x=T)      # add X to fit to get influence stats
> w ← which.influence(g, 0.3)
```

```
> dffits ← resid(g, 'dffits')
> show.influence(w, data.frame(counties, pdensity, older, dffits),
+                 report=c('democrat','dffits'), id=county)
```

	Count	college	income	white	turnout	democrat	dffits
Jackson	4	* 5	*14767	100	38	17	-0.8
McCreary	4	* 5	*12223	99	40	31	-0.8
Taos	2	18	*20049	73	46	66	0.6
Duval	1	6	*15773	79	39	80	0.5
Loving	5	* 4	*30833	87	*90	21	-0.9
Starr	2	7	*10903	62	23	83	0.8
Menominee	5	* 4	*14801	* 11	30	60	-0.6

One can see, for example, that for Starr County, which has a very low median family income of \$10,903, at least one regression coefficient associated with `income` changes by more than 0.3 standard errors when that county is removed from the dataset. These influential observations appear to contain valid data and do not lead us to delete the data or change the model (other than to make a mental note to pay more attention to robust estimation in the future!).

7.7   Test Statistics and Partial R^2

Most of the partial F-statistics that one might desire are shown in Table 7.1.

```
> an ← anova(f)
> an
> plot(an, what='partial R2')                    # Figure 7.7
```
[e]

The 20 d.f. simultaneous test that no effects are nonlinear or interacting provides strong support for the need for complexity in the model. Every variable that was allowed to have a nonlinear effect on the percentage voting for Bill Clinton had a significant nonlinear effect. Even the nonlinear interaction terms are significant (the global test for linearity of interaction had $F_{5,3084} = 7.58$). `college` $\times$ `income` interaction is moderately strong. Note that voter turnout is still significantly associated with Democratic voting even after adjusting for county demographics ($F = 19.2$). Figure 7.7 is a good snapshot of the predictive power of all the predictors. It is very much in agreement with Figure 7.3; this is expected unless major confounding or collinearity is present.

---

[e]The output was actually produced using `latex(an, dec.ss=0, dec.ms=0, dec.F=1, scientific=c(-6,6))`.

TABLE 7.1: Analysis of Variance for `democrat`

	d.f.	PartialSS	MS	F	P
pdensity	3	18698	6233	108.1	< 0.0001
Nonlinear	2	4259	2130	36.9	< 0.0001
pop.change	2	8031	4016	69.7	< 0.0001
Nonlinear	1	2007	2007	34.8	< 0.0001
older	2	1387	694	12.0	< 0.0001
Nonlinear	1	1246	1246	21.6	< 0.0001
crime	1	2	2	0.0	0.8408
college (Factor+Higher Order Factors)	10	17166	1717	29.8	< 0.0001
All Interactions	6	2466	411	7.1	< 0.0001
Nonlinear (Factor+Higher Order Factors)	6	8461	1410	24.5	< 0.0001
income (Factor+Higher Order Factors)	9	12945	1438	25.0	< 0.0001
All Interactions	6	2466	411	7.1	< 0.0001
Nonlinear (Factor+Higher Order Factors)	4	3163	791	13.7	< 0.0001
college × income (Factor+Higher Order Factors)	6	2466	411	7.1	< 0.0001
Nonlinear	5	2183	437	7.6	< 0.0001
Nonlinear Interaction : f(A,B) vs. AB	5	2183	437	7.6	< 0.0001
Nonlinear Interaction in college vs. Af(B)	3	1306	435	7.6	< 0.0001
Nonlinear Interaction in income vs. Bg(A)	2	1864	932	16.2	< 0.0001
farm	2	7179	3590	62.3	< 0.0001
Nonlinear	1	339	339	5.9	0.0154
white	4	22243	5561	96.5	< 0.0001
Nonlinear	3	2508	836	14.5	< 0.0001
turnout	2	2209	1105	19.2	< 0.0001
Nonlinear	1	716	716	12.4	0.0004
TOTAL NONLINEAR	19	23231	1223	21.2	< 0.0001
TOTAL NONLINEAR + INTERACTION	20	37779	1889	32.8	< 0.0001
TOTAL	29	183694	6334	109.9	< 0.0001
ERROR	3084	177767	58		

7.8   Interpreting the Model

Our first task is to interpret the interaction surface relating education and income. This can be done with perspective plots (see Section 10.5) and image plots. Often it is easier to see patterns by making ordinary line graphs in which separate curves are drawn for levels of an interacting factor. No matter how interaction surfaces are drawn, it is advisable to suppress plotting regions where there are very few datapoints in the space of the two predictor variables, to avoid unwarranted extrapolation. The `plot` function for model fits created with the S-Plus `Design` library in effect makes it easy to display interactions in many different ways, and to suppress poorly supported points for any of them. In Figure 7.8 is shown the estimated relationship between percentage college educated in the county versus percentage
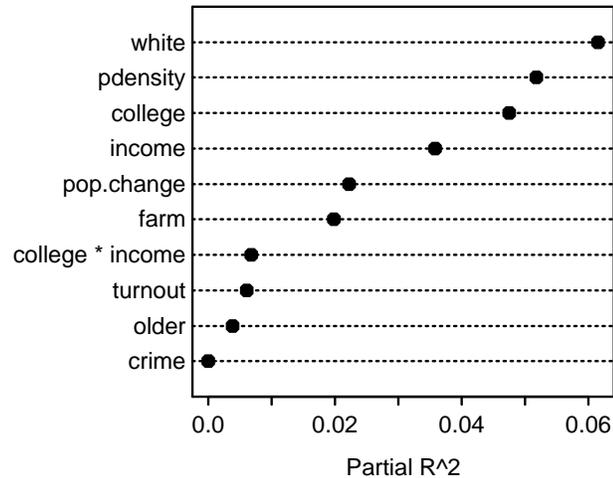
FIGURE 7.7: Partial R^2s for all of the predictors. For `college` and `income` partial R^2 includes the higher-order `college` $\times$ `income` interaction effect.

voting Democratic, with county median family income set to four equally spaced values between the 25th and 75th percentiles, and rounded. Curves are drawn for intervals of education in which there are at least 10 counties having median family income within \$1650 of the median income represented by that curve.

```
> incomes ← seq(22900, 32800, length=4)
> show.pts ← function(college.pts, income.pt) {
+    s ← abs(income - income.pt) < 1650
+ # Compute 10th smallest and 10th largest % college
+ # educated in counties with median family income within
+ # $1650 of the target income
+    x ← college[s]
+    x ← sort(x[!is.na(x)])
+    n ← length(x)
+    low ← x[10]; high ← x[n-9]
+    college.pts >= low & college.pts <= high
+    }

> plot(f, college=NA, income=incomes,          # Figure 7.8
+      conf.int=F, xlim=c(0,35), ylim=c(30,55),
+      lty=1, lwd=c(.25,1.5,3.5,6), col=c(1,1,2,2),
+      perim=show.pts)
```

The interaction between the two variables is evidenced by the lessened impact of low education when income increases.
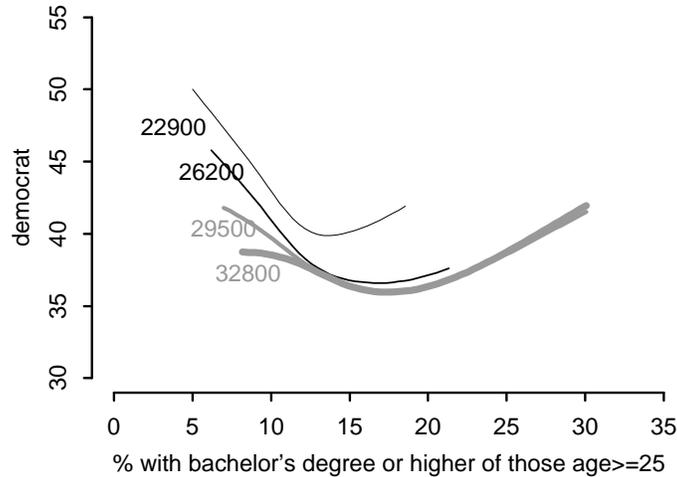
FIGURE 7.8: Predicted percentage voting Democratic as a function of college education (x-axis) and income (four levels used to label the curves) in the county. Other variables are set to overall medians.

Figure 7.9 shows the effects of all of the predictors, holding other predictors to their medians. All graphs are drawn on the same scale so that relative importance of predictors can be perceived. Nonlinearities are obvious.

```
> plot(f, ylim=c(20,70))                        # Figure 7.9
```

Another way to display effects of predictors is to use a device discussed in Section 5.3. We compute $\hat{Y}$ at the lower quartile of an X, holding all other Xs at their medians, then set the X of interest to its upper quartile and again compute $\hat{Y}$. By subtracting the two predicted values we obtain an estimate of the effects of predictors over the range containing one-half of the counties. The analyst should exercise more care than that used here in choosing settings for variables nonmonotonically related to Y.

```
> s ← summary(f)
> options(digits=4)
> plot(s)                                        # Figure 7.10
```

All predictor effects may be shown in a nomogram, which also allows predicted values to be computed. As two of the variables interact, it is difficult to use continuous axes for both, and the Design library's nomogram function does not allow this. We must specify the levels of one of the interacting factors so that separate scales can be drawn for each level.

```
> f <- Newlabels(f, list(turnout='voter turnout (%)'))
```
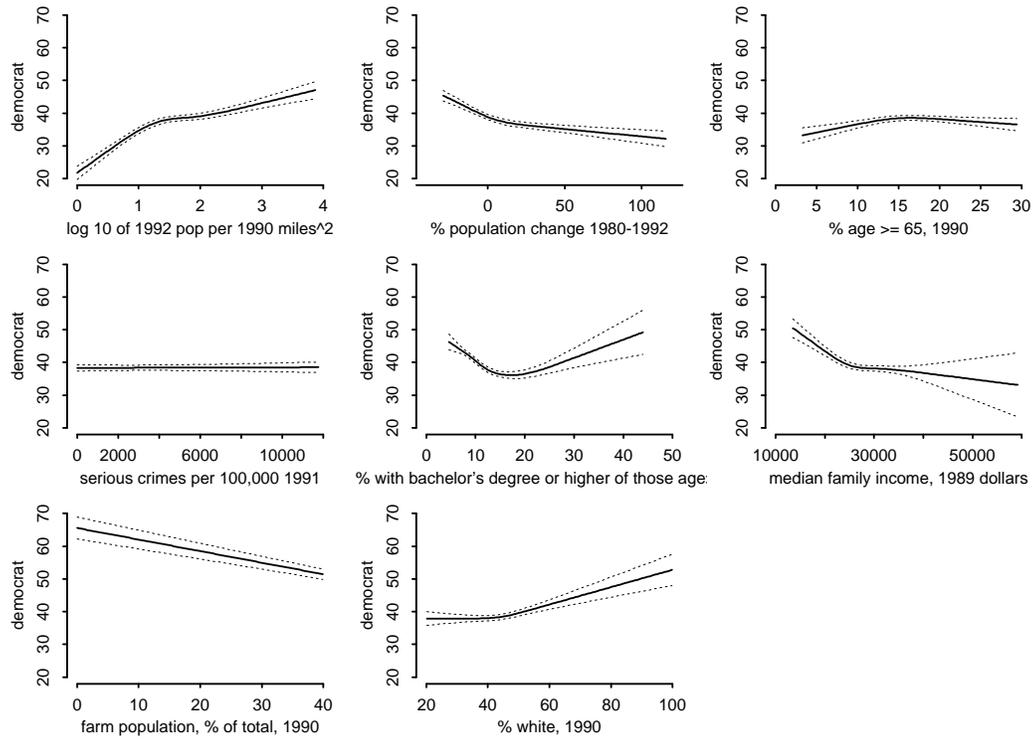
FIGURE 7.9: Partial effects of all county characteristics in the model.
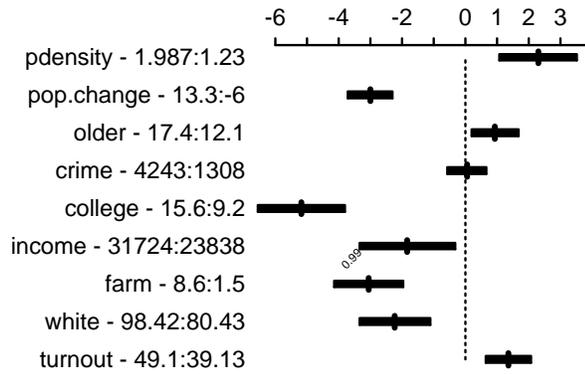
FIGURE 7.10: Summary of effects of predictors in the model using default ranges (interquartile). For variables that interact with other predictors, the settings of interacting factors are very important. For others, these settings are irrelevant for this graph. As an example, the effect of increasing population density from its first quartile (1.23) to its third quartile (1.987) is to add approximately an average of 2.3% voters voting Democratic. The 0.95 confidence interval for this mean effect is $[1.37, 3.23]$. This range of $1.987 - 1.23$ or 0.756 on the $\log_{10}$ population density scale corresponds to a $10^{0.756} = 5.7$-fold population increase.

TABLE 7.2

Characteristic	Points
Population density $10/\text{mile}^2$ ($log_{10} = 1$)	30
No population size change	27
Older age 5%	6
Median family income $29500	
and 40% college educated	27
Farm population 45%	37
White 90%	8
Voter turnout 40%	0

```
> nomogram(f, interact=list(income=incomes),
+          turnout=seq(30,100,by=10),
+          lplabel='estimated % voting Democratic',
+          cex.var=.8, cex.axis=.75)          # Figure 7.11
```

As an example, a county having the characteristics in Table 7.2 would derive the indicated approximate number of points. The total number of points is 135, for which we estimate a 38% vote for Bill Clinton. Note that the crime rate is irrelevant.

7.9   Problems

1. Picking up with the problems in Section 3.10 related to the SUPPORT study, begin to relate a set of predictors (age, sex, dzgroup, num.co, scoma, race, meanbp, pafi, alb) to total cost. Delete the observation having zero cost from all analyses.[f]

   (a) Compute mean and median cost stratified separately by all predictors (by quartiles of continuous ones; for S-Plus see the help file for the Hmisc summary.formula function). For categorical variables, compute P-values based on the Kruskal–Wallis test for group differences in costs.[g]

   (b) Decide whether to model costs or log costs. Whatever you decide, justify your conclusion and use that transformation in all later steps.

---

[f]In S-Plus issue the command attach(support[support$totcst > 0 | is.na(support$ totcst),]).

[g]You can use the Hmisc spearman2 function for this. If you use the built-in S-Plus function for the Kruskal–Wallis test note that you have to exclude any observations having missing values in the grouping variable. Note that the Kruskal–Wallis test and its two-sample special case, the Wilcoxon–Mann–Whitney test, tests in a general way whether the values in one group tend to be larger than values in another group.

Points

log 10 of 1992 pop per 1990 miles^2

% population change 1980-1992

% age >= 65, 1990

serious crimes per 100,000 1991

college (income=22900)

college (income=26200)

college (income=29500)

college (income=32800)

farm population, % of total, 1990

% white, 1990

voter turnout (%)

Total Points

estimated % voting Democratic
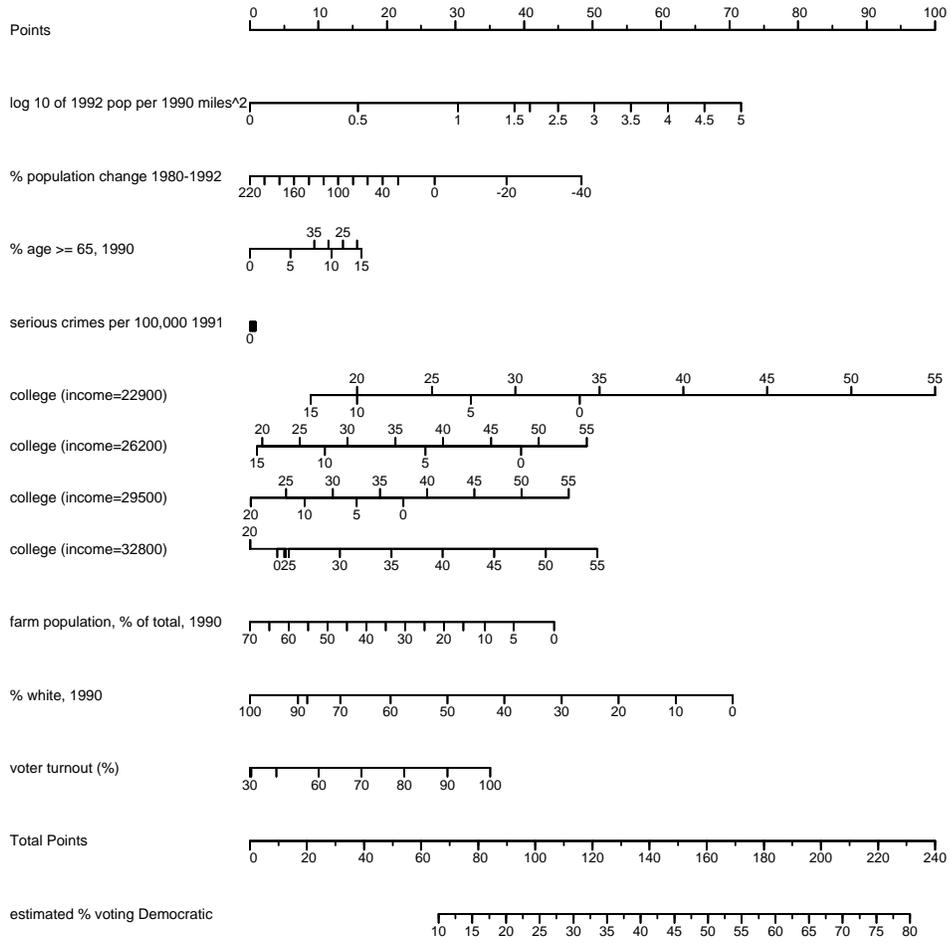
FIGURE 7.11: Nomogram for the full model for predicting the percentage of voters in a county who voted Democratic in the 1992 U.S. presidential election.

(c) Use all nonmissing data for each continuous predictor to make a plot showing the estimated relationship, superimposing nonparametric trend lines and restricted cubic spline fits (use five knots). If you used a log transformation, be sure to tell the nonparametric smoother to use the log of costs also. As the number of comorbidities and coma score have heavily tied values, splines may not work well unless knot locations are carefully chosen. For these two variables it may be better to use quadratic fits. You can define an S-PLUS function to help do all of this:

```
doplot ← function(predictor, type=c('spline','quadratic')) {
  type ← match.arg(type)
  r ← range(predictor, na.rm=T)
  xs ← seq(r[1], r[2], length=150)
  f ← switch(type,
             spline   = ols(log(totcst) ~ rcs(predictor, 5)),
             quadratic= ols(log(totcst) ~ pol(predictor, 2)))
  print(f)
  print(anova(f))
  plot(f, predictor=xs, xlab=label(predictor))
  plsmo(predictor, log(totcst), add=T, trim=0, col=3, lwd=3)
  scat1d(predictor)
  title(sub=paste('n=',f$stats['n']),adj=0)
  invisible()
}
doplot(pafi)
doplot(scoma, 'quadratic')
etc.
```

Note that the purpose of Parts (c) and (d) is to become more familiar with estimating trends without assuming linearity, and to compare parametric regression spline fits with nonparametric smoothers. These exercises should not be used in selecting the number of degrees of freedom to devote to each predictor in the upcoming multivariable model.

(d) For each continuous variable provide a test of association with costs and a test of nonlinearity, as well as adjusted R^2.

2. Develop a multivariable least squares regression model predicting the log of total hospital cost. For patients with missing costs but nonmissing charges, impute costs as you did in Problem 2b in Chapter 3. Consider the following predictors: `age, sex, dzgroup, num.co, scoma, race` (use all levels), `meanbp, hrt, temp, pafi, alb`.

(a) Graphically describe how the predictors interrelate, using squared Spearman correlation coefficients. Comment briefly on whether you think any of the predictors are redundant.

(b) Decide for which predictors you want to "spend" more than one degree of freedom, using subject–matter knowledge or by computing a measure (or generalized measure allowing nonmonotonic associations) of rank correlation between each predictor and the response. Note that rank correlations do not depend on how the variables are transformed (as long as transformations are monotonic).

(c) Depict whether and how the same patients tend to have missing values for the same groups of predictor and response variables.

(d) The dataset contains many laboratory measurements on patients. Measurements such as blood gases are not done on every patient. The PaO_2/FiO_2 ratio (variable `pafi`) is derived from the blood gas measurements. Using any method you wish, describe which types of patients are missing `pafi`, by considering other predictors that are almost never missing.

(e) Impute `race` using the most frequent category. Can you justify imputing a constant for `race` in this dataset?

(f) Physicians often decide not to order lab tests when they think it likely that the patient will have normal values for the test results. Previous analyses showed that this strategy worked well for `pafi` and `alb`. When these values are missing, impute them using "normal values," 333.3 and 3.5, respectively.

(g) Fit a model to predict cost (or a transformation of it) using all predictors. For continuous predictors assume a smooth relationship but allow it to be nonlinear. Choose the complexity to allow for each predictor's shape (i.e., degrees of freedom or knots) building upon your work in Part 2b. Quantify the ability of the model to discriminate costs. Do an overall test for whether any variables are associated with costs.

Here are some hints for using `Design` library functions effectively for this problem.

- Optionally `attach` the subset of the `support` data frame for which you will be able to get a nonmissing total hospital cost, that is, those observations for which either `totcst` or `charges` are not `NA`.

- Don't use new variable names when imputing `NA`s. You can always tell which observations have been imputed using the `is.imputed` function, assuming you use the `impute` function to do the imputations.

- Run `datadist` before doing imputations, so that quantiles of predictors are estimated on the basis of "real" data. You will need to update the `datadist` object only when variables are recoded (e.g., when categories are collapsed).

(h) Graphically assess the overall normality of residuals from the model. For the single most important predictor, assess whether there is a systematic trend in the residuals against this predictor.

(i) Compute partial tests of association for each predictor and a test of nonlinearity for continuous ones. Compute a global test of nonlinearity. Graphically display the ranking of importance of the predictors based on the partial tests.

(j) Display the shape of how each predictor relates to cost, setting other predictors to typical values (one value per predictor).

(k) For each predictor estimate (and either print or plot) how much $\hat{Y}$ changes when the predictor changes from its first to its third quartile, all other predictors held constant. For categorical predictors, compute differences in $\hat{Y}$ between all categories and the reference category. Antilog these differences to obtain estimated cost ratios.[h]

(l) Make a nomogram for the model, including a final axis that translates predictions to the original cost scale if needed (note that antilogging predictions from a regression model that assumes normality in log costs results in estimates of median cost). Use the nomogram to obtain a predicted value for a patient having values of all the predictors of your choosing. Compare this with the predicted value computed by either the `predict` or `Function` function in S-Plus.

(m) Use resampling to validate the R^2 and slope of predicted against observed response. Compare this estimate of R^2 to the adjusted R^2. Draw a validated calibration curve. Comment on the quality (potential "exportability") of the model.

(n) Refit the full model, excluding observations for which `pafi` was imputed. Plot the shape of the effect of `pafi` in this new model and comment on whether and how it differs from the shape of the `pafi` effect for the fit in which `pafi` was imputed.

Hints: Analyses (but not graph titles or interpretation) for Parts (a), (b), (c), (e), and (j) can be done using one S-Plus command each. Parts (f), (h), (i), (k), (l), and (n) can be done using two commands. Parts (d), (g), and (m) can be done using three commands. For part (h) you can use the `resid` and `qqnorm` functions or the pull-down 2-D graphics menu in Windows S-Plus. `plot.lm(fit object)` may also work, depending on how it handles `NA`s.

---

[h]There is an option on the pertinent S-Plus function to do that automatically when the differences are estimated.