Chapter 1. Introduction

MAST90139 Statistical Modelling for Data Science Slides

Guoqi Qian

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

Contents

§1.1 Overview (1)

- Linear (regression) model is arguably the most important statistical model used in almost all statistical applications.
- A linear model may be defined as the regression equation

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon, \quad \text{or} \\ Y &= \mu + \varepsilon \quad \text{with} \quad \mu = E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \end{aligned}$$

where

  - Y: response/outcome/dependent variable;
  - $X_1, \cdots, X_{p-1}$: covariate/explanatory/predictor/independent variables;
  - ε: random error having mean 0 and unknown constant variance σ^2;
  - $\beta_0, \beta_1, \cdots, \beta_{p-1}$: unknown regression parameters to be estimated based on the observed data.

§1.1 Overview (2)

- Given n observed data points of $(Y, X_1, \cdots, X_{p-1})$, the regression equation may be written in a vector/matrix form

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \text{or} \quad \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \text{ with } \boldsymbol{\mu} = E(\mathbf{y}) = X\boldsymbol{\beta},$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^\top$,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ \vdots & \cdots & \vdots & \\ 1 & x_{i1} & \cdots & x_{i,p-1} \\ \vdots & \cdots & \vdots & \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix}_{n \times p} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix}.$$

- In a linear model, $y_1, \cdots, y_n$ are assumed to be independent of each other, while X is assumed to be deterministic rather than random.

§1.1 Overview (3)

- In this subject we will extend the linear model from 4 prospects
  1. Distribution of Y is extended from normal to an exponential family member, leading to GLM;
  2. Dependence is assumed to exist in $y_1, \cdots, y_n$, leading to multivariate GLM;
  3. $\boldsymbol{\mu} = E(\mathbf{y})$ may be a nonparametric function of X and $\boldsymbol{\beta}$, leading to nonparametric regression models;
  4. Some effects of $X_1, \cdots, X_{p-1}$ on Y may be random, leading to random effects or mixed effects models.
- Statistics software `R` and the `R` package `faraway` are used in this subject.

Undercounted votes in Georgia in 2000 presidential election

- The 2000 US presidential election generated much controversy about the used voting machinery.
- Data on voting in Georgia is presented and analysed in Meyer (2002).

```
> library{faraway);data(gavote); help(gavote);dim(gavote); head(gavote)
[1] 159  10
          equip   econ perAA rural    atlanta gore bush other votes ballots
APPLING   LEVER   poor 0.182 rural notAtlanta 2093 3940    66  6099    6617
ATKINSON LEVER   poor 0.230 rural notAtlanta  821 1228    22  2071    2149
BACON     LEVER   poor 0.131 rural notAtlanta  956 2010    29  2995    3347
BAKER     OS-CC   poor 0.476 rural notAtlanta  893  615    11  1519    1607
BALDWIN   LEVER middle 0.359 rural notAtlanta 5893 6041   192 12126   12785
BANKS     LEVER middle 0.024 rural notAtlanta 1220 3202   111  4533    4773
```

- A ballot will be issued to a person going to the polling station if he or she has registered to vote.
- However, a vote is not recorded if he/she fails to vote for President, votes for more than one candidate or the equipment malfunctions.
- E.g., in Appling county, $6617 - 6099 = 518$ ballots did not result in votes for President. This is called the *undercount*.
- We aim to determine what factors affect the undercount by LM & R.

Description of the gavote data

- A data frame with 159 observations on the following 10 variables. Each case represents a county in Georgia.
`equip`:	voting equip. used: `LEVER`, `OS-CC`, `OS-PC`, `PAPER`, `PUNCH`
`econ`:	economic status of county: `middle poor rich`
`perAA`:	percent of African Americans in county
`rural`:	indicator of whether county is `rural` or `urban`
`atlanta`:	indicator of whether county is in `Atlanta` or `notAtlanta`
`gore`:	number of votes for Gore
`bush`:	number of votes for Bush
`other`:	number of votes for other candidates
`votes`:	number of votes
`ballots`:	number of ballots
- *Source*: Meyer M. (2002) Uncounted Votes: Does Voting Equipment Matter? Chance, 15(4), 33-38

§1.3 Initial Data Analysis (1)

- Numerical summary

```
> summary(gavote)
```

```
   equip          econ          perAA             rural          atlanta          gore        
 LEVER:74   middle:69   Min.   :0.0000   rural:117   Atlanta: 15   Min.   :   249  
 OS-CC:44   poor  :72   1st Qu.:0.1115   urban: 42   notAtl: 144   1st Qu.:  1386  
 OS-PC:22   rich  :18   Median :0.2330                             Median :  2326  
 PAPER: 2               Mean   :0.2430                             Mean   :  7020  
 PUNCH:17               3rd Qu.:0.3480                             3rd Qu.:  4430  
                        Max.   :0.7650                             Max.   :154509  
      bush             other              votes             ballots        
 Min.   :   271   Min.   :   5.0   Min.   :   832   Min.   :   881  
 1st Qu.:  1804   1st Qu.:  30.0   1st Qu.:  3506   1st Qu.:  3694  
 Median :  3597   Median :  86.0   Median :  6299   Median :  6712  
 Mean   :  8929   Mean   : 381.7   Mean   : 16331   Mean   : 16927  
 3rd Qu.:  7468   3rd Qu.: 210.0   3rd Qu.: 11846   3rd Qu.: 12251  
 Max.   :140494   Max.   :7920.0   Max.   :263211   Max.   :280975  
```

- Number of ballots cast ranges widely, suggesting the use of relative, rather than the absolute, undercount.

§1.3 Initial Data Analysis (2)

- Define the relative undercount variable and attach it to gavote

```
> gavote$undercount <- (gavote$ballots-gavote$votes)/gavote$ballots
> summary(gavote$undercount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.02779 0.03983 0.04379 0.05647 0.18812
> with(gavote, sum(ballots-votes)/sum(ballots))  #overall percent of undercount
[1] 0.03518021
```

- Graphic summary

```
> hist(gavote$undercount,main="Undercount",xlab="Percent Undercount")
> plot(density(gavote$undercount),main="Undercount"); rug(gavote$undercount)
> pie(table(gavote$equip),col=gray(0:4/4))
> barplot(sort(table(gavote$equip),decreasing=TRUE),las=2)
```

§1.3 Initial Data Analysis (3)

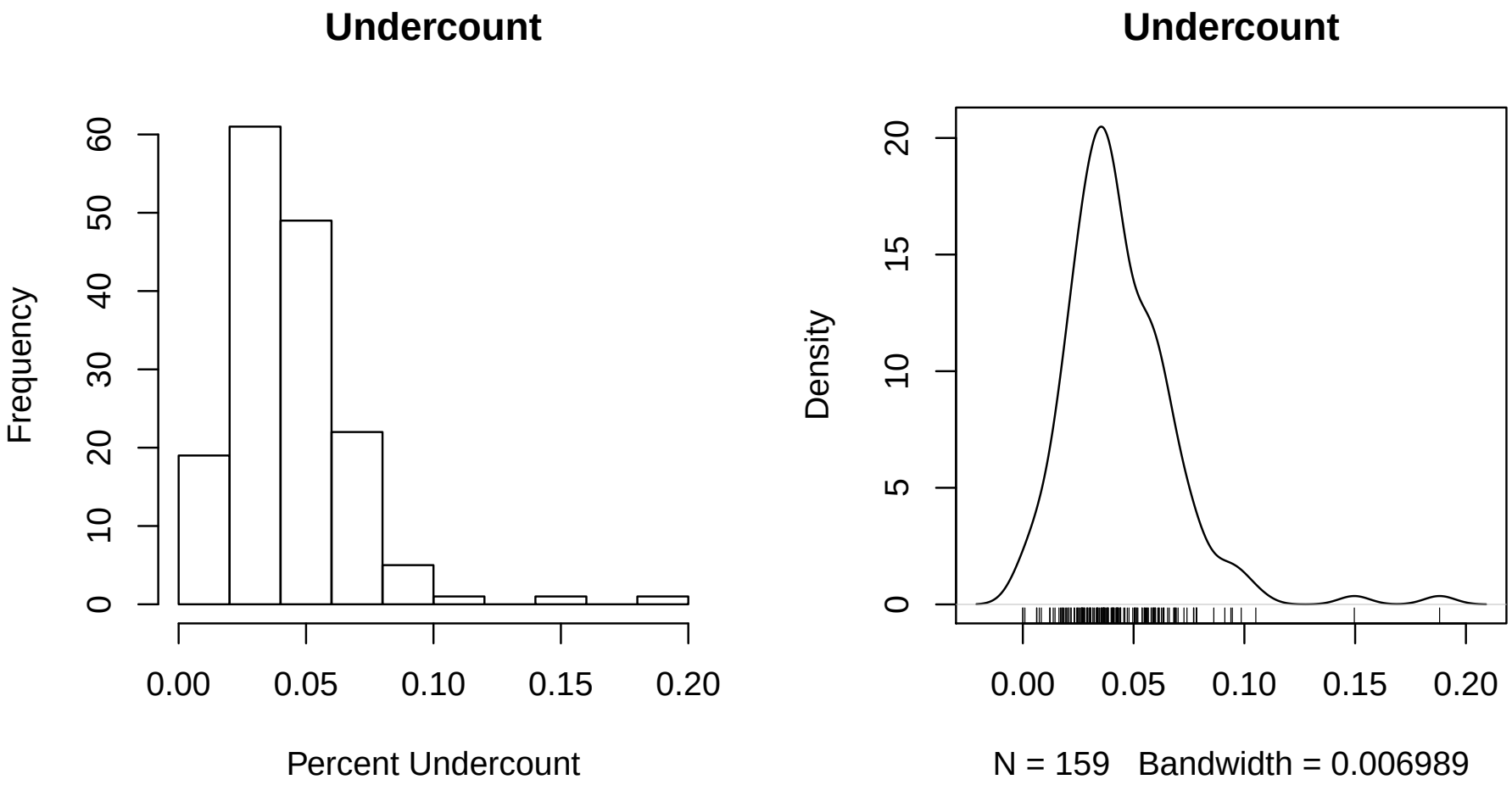

Figure 1.1: Histogram of the undercount is shown on the left and a density estimate with a data rug is shown on the right. Variable `undercount` is numerical.

§1.3 Initial Data Analysis (4)

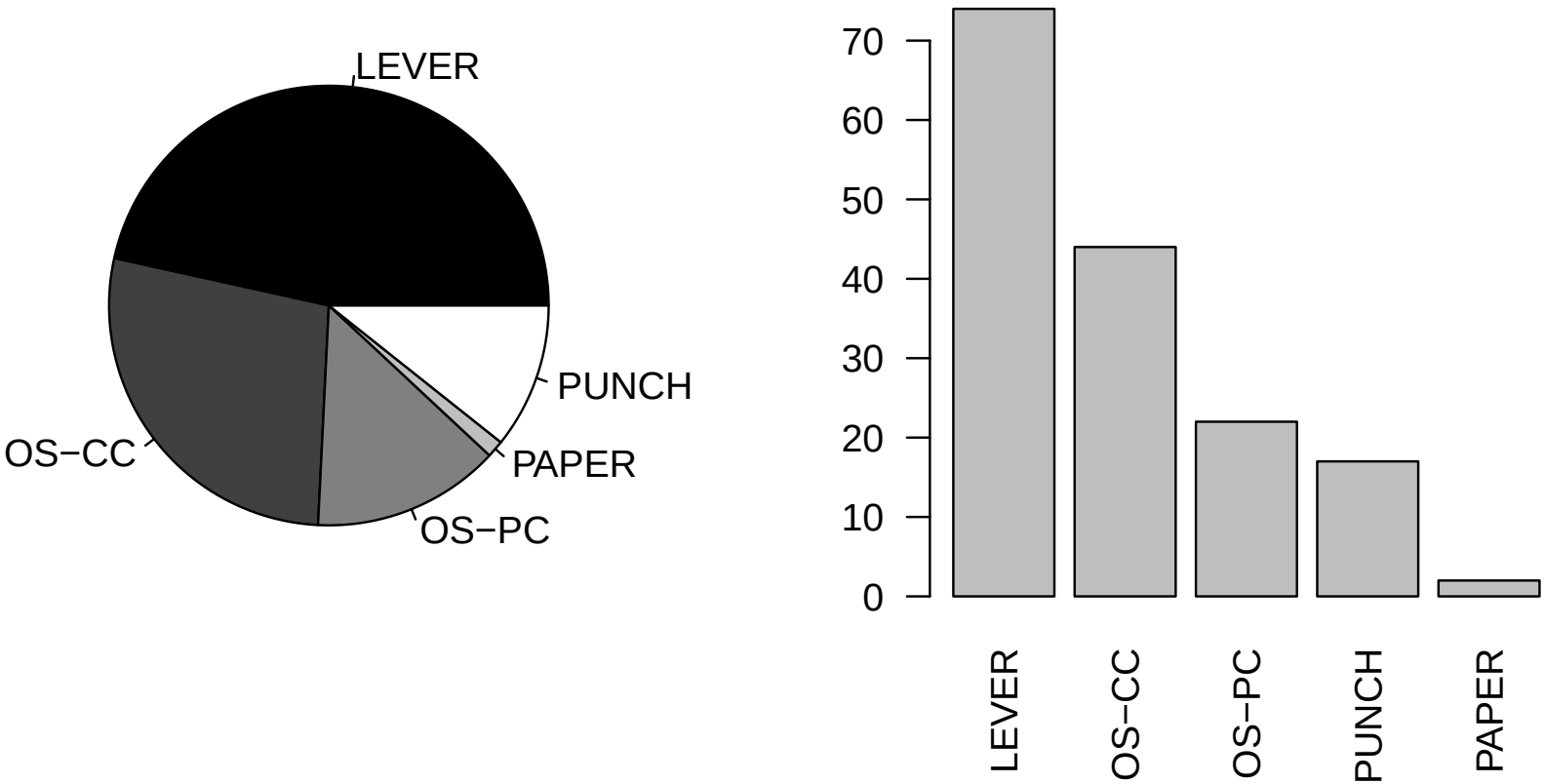

Figure 1.2: Pie chart of the voting equipment frequencies is shown on the left and a Pareto chart on the right. Variable `equip` is categorical.

§1.3 Initial Data Analysis (5)

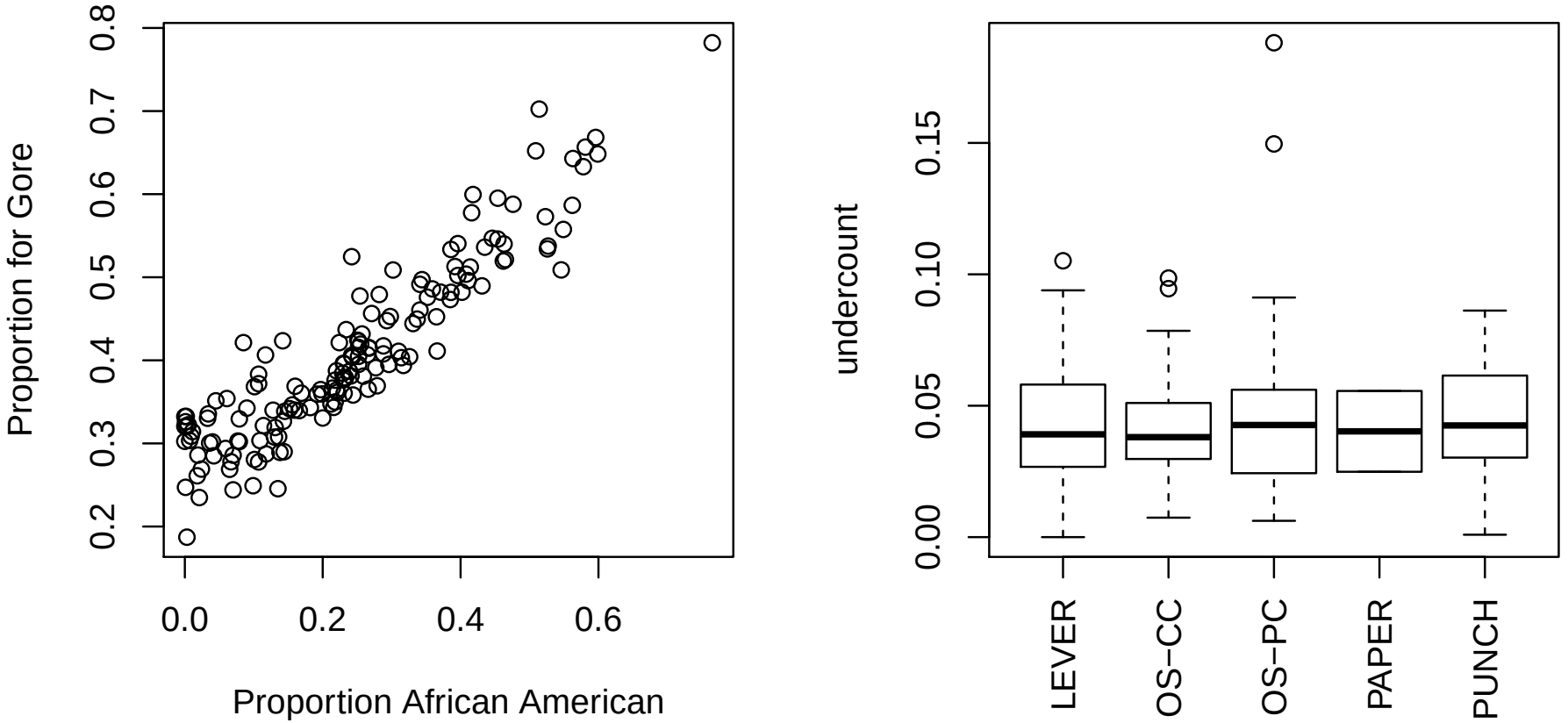

Figure 1.3: A scatterplot plot of proportions of Gore voters and African Americans by county is shown on the left. Boxplots showing the distribution of the undercount by voting equipment are shown on the right.

§1.3 Initial Data Analysis (6)

- Let's see how the proportion voting for Gore relates to the proportion of African Americans
- Side-by-side boxplots are one way of displaying the relationship between qualitative and quantitative variables (e.g. `equip` vs. `undercount`):

```
> gavote$pergore <- gavote$gore/gavote$votes
> plot(pergore ~ perAA, gavote, xlab="Proportion African American",
                                  ylab="Proportion for Gore")
> plot(undercount ~ equip, gavote, xlab="", las=3)
```

- The `xtabs` function is useful for cross-tabulations of qualitative variables.

```
> names(gavote); names(gavote)[4] <- "usage"
> xtabs(~ atlanta + usage, gavote)
```

```
            usage
atlanta      rural urban
  Atlanta        1    14
  notAtlanta   116    28
```

§1.3 Initial Data Analysis (7)

- Correlations are the standard way of numerically summarizing the relationship between quantitative variables.
- In the following we find the correlations between variables 3 (perAA), 10 (ballots), 11(undercount) and 12 (pergore) in gavote.

```
> nix <- c(3,10,11,12)
> cor(gavote[,nix])
```

	perAA	ballots	undercount	pergore
perAA	1.0000000	0.02773230	0.2296874	0.92165247
ballots	0.0277323	1.00000000	-0.1551724	0.09561688
undercount	0.2296874	-0.15517245	1.0000000	0.21876519
pergore	0.9216525	0.09561688	0.2187652	1.00000000

§1.4 Fitting a Linear Model to gavote (1)

- An LM $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ can be fitted to the data using the **least squares (LS) method**, from which the LS estimator of $\boldsymbol{\beta}$ can be shown to be

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}.$$

- Suppose we model the `undercount` as the response and the proportions of Gore voters and African Americans as predictors. The corresponding regression equation is

$$\texttt{undercount} = \beta_0 + \beta_1 \texttt{pergore} + \beta_2 \texttt{perAA} + \varepsilon.$$

```
> lmod <- lm(undercount ~ pergore + perAA, gavote); coef(lmod)
(Intercept)     pergore       perAA
 0.03237600  0.01097872  0.02853314
```

- Gauss–Markov theorem says that the LS estimator $\hat{\boldsymbol{\beta}}$ is the **best linear unbiased estimator (BLUE)**.
- If ε is assumed to be normal, the **maximum likelihood estimator (MLE)** of $\boldsymbol{\beta}$ can be shown to equal $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}$.

§1.4 Fitting a Linear Model to gavote (2)

- The predicted or **fitted values** are $\hat{\mathbf{y}} = X\hat{\beta}$

```
> predict(lmod)
 APPLING   ATKINSON     BACON     BAKER   BALDWIN     BANKS  ...
0.041337  0.043291  0.039618  0.052412  0.047955  0.036016   ...
```

- The **residuals** are $\hat{\varepsilon} = \mathbf{y} - X\hat{\beta} = \mathbf{y} - \hat{\mathbf{y}}$.

```
> residuals(lmod)
  APPLING    ATKINSON      BACON      BAKER     BALDWIN      ...
0.0369466  -0.0069949  0.0655506  0.0023484  0.0035899   ...
```

- The **residual sum of squares (RSS)**, also called **deviance**, is $\text{RSS} = \hat{\varepsilon}^\top \hat{\varepsilon}$, which measures how well the model fits the data.

```
> deviance(lmod)
[1] 0.09325
```

- The **degrees of freedom (df)** in the resuduals is $n - p$.

```
> df.residual(lmod)
[1] 156
> nrow(gavote)-length(coef(lmod))
[1] 156
```

§1.4 Fitting a Linear Model to gavote (3)

- Let $\sigma^2 = \text{Var}(\varepsilon)$. We estimate σ by the **residual standard error**

$$\hat{\sigma} = s = \sqrt{\frac{\text{RSS}}{\text{df}}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}}.$$

```
> sqrt(deviance(lmod)/df.residual(lmod))
[1] 0.024449
> lmodsum <- summary(lmod); lmodsum$sigma
[1] 0.024449
```

- The deviance measures how well the model fits in an absolute sense, but not in a relative sense. The relative goodness of fit of the model is measured by R^2 — the **coefficient of determination** or percentage of variance explained:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

§1.4 Fitting a Linear Model to gavote (4)

- R^2 also equals the (squared) correlation between $\hat{\mathbf{y}}$ and $\mathbf{y}$.

```
> summary(lmod)$r.squared
[1] 0.05308861
> cor(predict(lmod),gavote$undercount)^2
[1] 0.05308861
>  summary(lmod)$adj.r.squared
[1] 0.04094872
```

- The **adjusted** R^2, denoted as R^2_a, can be used as a criterion for model selection

$$R^2_a = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-p)}{\sum_{i=1}^n (y_i - \bar{y}_i)^2/(n-1)} = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)}.$$

§1.4 Fitting a Linear Model to gavote (5)

- More results from model fitting can be found from `summary(lmod)`, or `sumary(lmod)` in `faraway` package.

```
> lmodsum <- summary(lmod); attributes(lmodsum)
$`names`
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
> summary(lmod)
Residuals:
      Min        1Q    Median        3Q       Max
-0.046013 -0.014995 -0.003539  0.011784  0.142436

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03238    0.01276   2.537   0.0122 *
pergore      0.01098    0.04692   0.234   0.8153
perAA        0.02853    0.03074   0.928   0.3547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02445 on 156 degrees of freedom
Multiple R-squared:  0.05309,   Adjusted R-squared:  0.04095
F-statistic: 4.373 on 2 and 156 DF,  p-value: 0.01419
>  summary(lmod)$adj.r.squared
```

§1.5 Qualitative predictors or categorical factors in LM (1)

- Either **qualitative** (i.e. category) or **quantitative** (i.e. numerical) predictors can be used in LM. Qualitative predictors are **factors** in R.
- A qualitative predictor has to be coded into a set of numerical variables called **contrasts** when it is included into a linear model.
- There are many possible contrasts for coding a qualitative predictor. R uses 5 of them:

```
> contr.helmert(n, contrasts = TRUE, sparse = FALSE)
> contr.poly(n, scores = 1:n, contrasts = TRUE, sparse = FALSE)
> contr.sum(n, contrasts = TRUE, sparse = FALSE)
> contr.treatment(n, base = 1, contrasts = TRUE, sparse = FALSE)
> contr.SAS(n, contrasts = TRUE, sparse = FALSE)
> options()$contrasts
        unordered           ordered
"contr.treatment"      "contr.poly"
```

  where we see the default coding is `contr.treatment` for nominal factor, and `contr.poly` for ordinal factor.

- `contr.treatment` encodes a factor by **dummy variables**, being most interpret-able in practice.

§1.5 Qualitative predictors or categorical factors in LM (2)

- For a factor d having k levels that are displayed alphabetically, $k-1$ dummy variables $d_2, \cdots, d_k$ are used to code d by default in R, where

$$d_j = \begin{cases} 0 & \text{if } d \text{ is not at level } j \\ 1 & \text{if } d \text{ is at level } j \end{cases}, \quad j = 2, \cdots, k.$$

- Suppose a factor d has 3 levels, and we fit a linear model containing pergore, perAA and d by the following R command

```
> lmodI <- lm(undercount ~ pergore + perAA + d + perAA:d, gavote)
```

The corresponding regression equation is

$$\begin{aligned} \texttt{undercount} &= \beta_0 + \beta_1 \texttt{pergore} + \beta_2 \texttt{perAA} + \beta_3 d_2 + \beta_4 d_3 \\ &\quad + \beta_5 \texttt{perAA} \times d_2 + \beta_6 \texttt{perAA} \times d_3 + \varepsilon \end{aligned}$$

§1.6 Interpretation in LM (1)

- Let's fit an LM and see how its terms can be interpreted.
- Terms pergore and perAA are centered in the model for ease of interpretation, while usage (i.e. rural) and equip are factors.

```
> gavote$cpergore <- gavote$pergore - mean(gavote$pergore)
> gavote$cperAA <- gavote$perAA - mean(gavote$perAA)
> lmodi <- lm(undercount ~ cperAA+cpergore*usage+equip, gavote)
> sumary(lmodi)
                       Estimate Std. Error t value  Pr(>|t|)
(Intercept)           0.0432973  0.0028386 15.2529 < 2.2e-16
cperAA                0.0282641  0.0310921  0.9090  0.364786
cpergore              0.0082368  0.0511562  0.1610  0.872299
usageurban           -0.0186366  0.0046482 -4.0095 9.564e-05
equipOS-CC            0.0064825  0.0046799  1.3852  0.168060
equipOS-PC            0.0156396  0.0058274  2.6838  0.008097
equipPAPER           -0.0090920  0.0169263 -0.5372  0.591957
equipPUNCH            0.0141496  0.0067827  2.0861  0.038658
cpergore:usageurban -0.0087995  0.0387162 -0.2273  0.820515

n = 159, p = 9, Residual SE = 0.02335, R-Squared = 0.17
```

§1.6 Interpretation in LM (2)

- Mathematical form of the model `lmodi` is

$$\begin{aligned}\texttt{undercount} &= \beta_0 + \beta_1\texttt{cperAA} + \beta_2\texttt{cpergore} + \beta_3\texttt{usageurban} \\ &\quad + \beta_4\texttt{equipOSCC} + \beta_5\texttt{equipOSPC} + \beta_6\texttt{equipPAPER} + \beta_7\texttt{equipPUNCH} \\ &\quad + \beta_8\texttt{cpergore}\times\texttt{usageurban} + \varepsilon\end{aligned}$$

  where dummy variable `usaageurban=1` if `usage=urban`, and =0 if `usage!=urban`. Other dum-var. `equipOS-CC` etc. similarly defined.
- The estimate of the model `lmodi` is

$$\begin{aligned}\widehat{\texttt{undercount}} &= 0.043 + 0.028\texttt{cperAA} + 0.008\texttt{cpergore} - 0.019\texttt{usageurban} \\ &\quad + 0.006\texttt{equipOSCC} + 0.016\texttt{equipOSPC} - 0.009\texttt{equipPAPER} \\ &\quad + 0.141\texttt{equipPUNCH} - 0.0088\texttt{cpergore}\times\texttt{usageurban}\end{aligned}$$

- Standard errors of the β estimates can be found from `sumary(lmodi)`.
- If interaction `cpergore:equip` is included, there will be 4 extra terms `cpergore:equipOS-CC`, ..., `cpergore:equipPUNCH` in the model.

§1.6 Interpretation in LM (3)

- Consider a **rural** county that has an average proportion of Gore voters and an average proportion of African Americans where **lever** machines are used for voting.
- Because rural and lever are the reference levels for the two qualitative variables,there is no contribution to the predicted undercount from these terms. Furthermore, because we have centered the two quantitative variables at their mean values, these terms also do not enter into the prediction.
- Notice the worth of the centering because otherwise we would need to set these variables to zero to get them to drop out of the prediction equation; zero is not a typical value for these predictors.
- Given that all the other terms are dropped, the predicted undercount is just given by the intercept $\hat{\beta}_0$, which is 4.33%.

§1.6 Interpretation in LM (4)

- The interpretation of the coefficients can now be made relative to this baseline.
- We see that, with all other predictors unchanged, except using optical scan with precinct count (OS-PC), the predicted undercount increases by 1.56%. The other equipment methods can be similarly interpreted.
- Notice that we need to be cautious about the interpretation. Given two counties with the same values of the predictors, except having different voting equipment, we would predict the undercount to be 1.56% higher for the OS-PC county compared to the lever county. However, we cannot go so far as to say that if we went to a county with lever equipment and changed it to OS-PC that this would cause the undercount to increase by 1.56%.

§1.6 Interpretation in LM (5)

- With all other predictors held constant, we would predict the undercount to increase by 2.83% going from a county with no African Americans to all African American.
- Sometimes a one-unit change in a predictor is too large or too small, prompting a rescaling of the interpretation. For example, we might predict a 0.283% increase in the undercount for a 10% increase in the proportion of African Americans. Of course, this interpretation should *not be taken too literally*.
- Note the proportions of African Americans and Gore voters are strongly correlated so that an increase in the proportion of one would lead to an increase in the proportion of the other. This is the problem of *collinearity* that makes the interpretation of regression coefficients much more difficult.
- The proportion of African Americans is likely to be associated with other socioeconomic variables which might also be related to the undercount. This further hinders the possibility of a *causal conclusion*.

§1.6 Interpretation in LM (6)

- Interpretation of the `usage` and `pergore` cannot be done separately as an interaction term exists between these two variables in `lmodi`.
- For an average number of Gore voters, we would predict a 1.86%-lower undercount in an urban county compared to a rural county.
- In a rural county, we predict a 0.08% increase in the undercount as the proportion of Gore voters increases by 10%. In an urban county, we predict a $(0.00824 - 0.00880) \times 10 = -0.0056\%$ increase in the undercount as the proportion of Gore voters increases by 10%. A negative increase is actually a decrease.
- This illustrates the *potential pitfalls* in interpreting the effect of a predictor in the presence of an interaction. We cannot give a simple stand-alone interpretation of the effect of the proportion of Gore voters. The effect is to increase the undercount in rural counties and to decrease it, if only very slightly, in urban counties.

§1.7 Hypothesis testing in LM (1)

- Hypothesis testing may be used to determine the significance of any set of the predictors in a linear model.
- Assume the errors ε_i's are i.i.d. normal.
- Suppose we want to know whether a model Ω can be replaced by one of its sub-models, ω, without losing significant goodness of fit. This can be achieved by testing

$$H_0 : \boldsymbol{\beta}_{\Omega-\omega} = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_{\Omega-\omega} \neq \mathbf{0}$$

  where $\boldsymbol{\beta}_{\Omega-\omega}$ is the subset of $\boldsymbol{\beta}$ in Ω but not in ω.
- Suppose $\dim(\boldsymbol{\beta}_\Omega) = p$ and $\dim(\boldsymbol{\beta}_\omega) = q$. Then when ω is correct (i.e. H_0 is true), the F-statistic is

$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(p-q)}{\mathrm{RSS}_\Omega/(n-p)} \sim F_{p-q,n-p}$$

  Thus we would reject H_0 at significance level α if $F > F^{(\alpha)}_{p-q,n-p}$.

§1.7 Hypothesis testing in LM (2)

- For example, compare ω=`lmod` and Ω=`lmodi`.. ω has just `pergore` and `perAA` while Ω adds `usage` and `equip` along with their interaction. We compute the F-test as

```
> anova(lmod, lmodi)
Analysis of Variance Table

Model 1: undercount ~ pergore + perAA
Model 2: undercount ~ cperAA + cpergore * usage + equip
  Res.Df      RSS Df Sum of Sq      F   Pr(>F)
1    156 0.093249
2    150 0.081775  6  0.011474 3.5077 0.002823 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- It does not matter that the variables have been centered in Ω but not in ω, because the centering makes no difference to the RSS. The p-value here is small indicating the null hypothesis of preferring the smaller model should be rejected.

§1.7 Hypothesis testing in LM (3)

- Sometimes one wants to test the effect of a specific predictor on the response in the model.
- It is possible to use the general F-testing method: fit a model with the predictor and without the predictor and compute the F-statistic. It is however important to know what other predictors are also included in the models, and the results may differ if these are also changed.
- An alternative is to use a t-statistic for testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ regarding the effect of X_j:

$$t_j = \hat{\beta}_j / \text{se}(\hat{\beta}_j) \sim t_{n-p} \quad \text{under } H_0$$

- This approach produces exactly the same p-value as the F-testing method. For example, in `lmodi` the test for the significance of the proportion of African Americans gives a p-value of 0.3648. This indicates that this predictor is not statistically significant after adjusting for the effect of the other predictors on the response.

§1.7 Hypothesis testing in LM (4)

- We usually avoid using the t-tests for the levels of a qualitative predictor with more than 2 levels. A comparison of all models with one predictor less than the larger model may be obtained using F-test:

```
> drop1(lmodi, test="F")
Single term deletions
Model: undercount ~ cperAA + cpergore * usage + equip
               Df Sum of Sq      RSS     AIC F value  Pr(>F)
<none>                      0.081775 -1186.1
cperAA          1 0.0004505 0.082226 -1187.2  0.8264 0.36479
equip           4 0.0054438 0.087219 -1183.8  2.4964 0.04521 *
cpergore:usage  1 0.0000282 0.081804 -1188.0  0.0517 0.82051
```

- We see that the equipment is barely statistically significant in that the p-value is just less than the traditional 5% significance level.
- The main effects terms, `cpergore` and `usage` are not tested above. This demonstrates respect for the *hierarchy principle*: all lower-order terms corresponding to an interaction be retained in the model.
- Many difficulties with interpreting the hypothesis testing results. Avoid taking them too literally before understanding the problems.

§1.8 Confidence intervals in LM (1)

- A confidence interval can be obtained for each element of β by

$$\hat{\beta}_j \pm t_{n-p}^{(\alpha/2)} \mathrm{se}(\hat{\beta}_j)$$

where $t_{n-p}^{(\alpha/2)}$ is the upper $(\alpha/2)$th quantile of a t distribution with $n - p$ degrees of freedom.

```
> confint(lmodi)
```

	2.5 %	97.5 %
(Intercept)	0.0376884415	0.048906189
cperAA	-0.0331710614	0.089699222
cpergore	-0.0928429315	0.109316616
usageurban	-0.0278208965	-0.009452268
equipOS-CC	-0.0027646444	0.015729555
equipOS-PC	0.0041252334	0.027153973
equipPAPER	-0.0425368415	0.024352767
equipPUNCH	0.0007477196	0.027551488
cpergore:usageurban	-0.0852990903	0.067700182

§1.8 Confidence intervals in LM (2)

- Confidence intervals have a duality with the corresponding t-tests in that if the p-value is $> 5\%$, 0 will fall in the interval and vice versa.
- Confidence intervals give a range of plausible values for the parameter and are more useful for judging the size of the effect of the predictor than a p-value that merely indicates statistical significance, not necessarily practical significance.
- These intervals are individually correct, but there is not a 95% chance that the true parameter values fall in all the intervals. This problem of *multiple comparisons* is particularly acute for the voting equipment, where five levels leads to 10 possible pairwise comparisons, more than just the four shown here.

§1.8 Confidence intervals in LM (3)

- In general, a $100(1-\alpha)\%$ confidence interval (C.I.) for $\mathbf{a}^\top\boldsymbol{\beta}$, where $\mathbf{a}$ is a given p vector, is

$$\mathbf{a}^\top\hat{\boldsymbol{\beta}} \pm t_{n-p}^{(\alpha/2)}\text{se}(\mathbf{a}^\top\hat{\boldsymbol{\beta}})$$

  where

$$\text{se}(\mathbf{a}^\top\hat{\boldsymbol{\beta}}) = s\sqrt{\mathbf{a}^\top(X^\top X)^{-1}\mathbf{a}}$$

- This confidence interval has a duality with the corresponding t-test for $H_0 : \mathbf{a}^\top\boldsymbol{\beta} = 0$.
- The `R` function `esticon()` in the `doBy` package computes the confidence intervals for $\mathbf{a}^\top\boldsymbol{\beta}$, as well as the corresponding testing.

§1.9 Diagnostics in LM (1)

- Validity of the inference depends on the assumptions concerning the linear model.
- One assumption is that $E\mathbf{y} = X\beta$ is correct: it includes all the right variables and transforms and combines them correctly.
- Another assumption concerns ε: its elements have equal variance, are uncorrelated and have a normal distribution.
- We are also interested in detecting points, called **outliers**, that are unusual in that they do not fit the model that seems otherwise adequate for the rest of the data.
- Ideally, we would like each case to have an equal contribution to the fitted model; yet sometimes a few points have a much larger effect than others. Such points are called **influential**.

§1.9 Diagnostics in LM (2)

- Diagnostic tools are designed to detect discrepancies between the data and the fitted values; as well as discrepancies between a few data points and the rest.
- Most of these techniques are based on graphical presentations of *residuals*, the *hat matrix*, and *case deletion measures*.

§1.9 Diagnostics in LM (3)

Consider a linear model (LM):

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$$

where $X : n \times p$ matrix, and $\boldsymbol{\beta} : p \times 1$ vector parameter.

1. **Least square (LS) estimator** of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$.
2. **Residual** vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ where $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ is the **fitted value** vector. The residual vector shows which data points are ill-fitting.
3. **Hat matrix** $H = X(X^T X)^{-1} X^T$.

   Hence $\hat{\mathbf{y}} = H\mathbf{y}$; $\hat{\mathbf{y}}$ can be interpreted as the perpendicular *projection* of $\mathbf{y}$ into the p-dimensional subspace $\{X\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^p\}$.

   It means H is *symmetric* and *idempotent*, i.e. $H^2 = H$.

§1.9 Diagnostics in LM (4)

- $H = (h_{ij})$ is used to determine how much **influence** or **leverage** the data have on the fitted values $\hat{\mathbf{y}}$.
- In particular, h_{ij} shows the amount of leverage or influence exerted on $\hat{y}_i$ by y_j. The influence is due to X not $\mathbf{y}$. We may write h_i for h_{ii}.
- The most interesting influence is that of y_i on $\hat{y}_i$, which is reflected by h_{ii}, the i-th diagonal element of H.
- Since $\text{rank}(H) = \sum_{i=1}^{n} h_{ii} = p$ and $0 \leq h_{ii} \leq 1$, $\frac{p}{n}$ is the average size of a diagonal element.
- As a rule of thumb, an x-point for which $h_{ii} > \frac{2p}{n}$ holds is considered a **high-leverage point**.

§1.9 Diagnostics in LM (5)

- If $\mathbf{r} = (r_1, \cdots, r_n)^T$ is the basic residual vector, the standardization $r_i^* = \dfrac{r_i}{\sqrt{1-h_{ii}}}$ is called the *i*-**th scaled residual**.
- It can be shown that $\text{var}(r_i^*) = \sigma^2$ or $\text{var}(\dfrac{r_i^*}{\sigma}) = 1$, and σ^2 is consistently estimated by $\hat{\sigma}^2 = \dfrac{\mathbf{r}^T\mathbf{r}}{n-p} = \dfrac{\text{RSS}}{n-p}$. Now $r_i^{(t)} = \dfrac{r_i^*}{\hat{\sigma}} = \dfrac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ is called the *i*-**th standardized (or (internally) studentized) residual**.
- The role of residual vector $\mathbf{r}$ and hat matrix H may also be seen by looking at the effect of omitting single observations.
- If the observation i is omitted, the change in LS estimates is given by

$$\Delta_i\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = (X^TX)^{-1}\mathbf{x}_i r_i/(1-h_{ii}),$$

  $\hat{\boldsymbol{\beta}}_{(i)}$ being the LSE of $\boldsymbol{\beta}$ based on the data excluding observation i.
- $\Delta_i\hat{\boldsymbol{\beta}}$ increases with increasing residual r_i and increasing h_{ii}.

§1.9 Diagnostics in LM (6)

- Graphical diagnostic methods tend to be more versatile and informative, thus are preferred.
- It is virtually impossible to verify that a given model is exactly correct.
- The purpose of the diagnostics is more to check whether the model is not grossly wrong. Indeed, a successful data analyst should pay more attention to avoiding big mistakes than optimizing the fit.
- A collection of four useful diagnostics can be obtained with: `plot(lmodi)`, as can be seen in Figure 1.4.

§1.9 Diagnostics in LM (7)

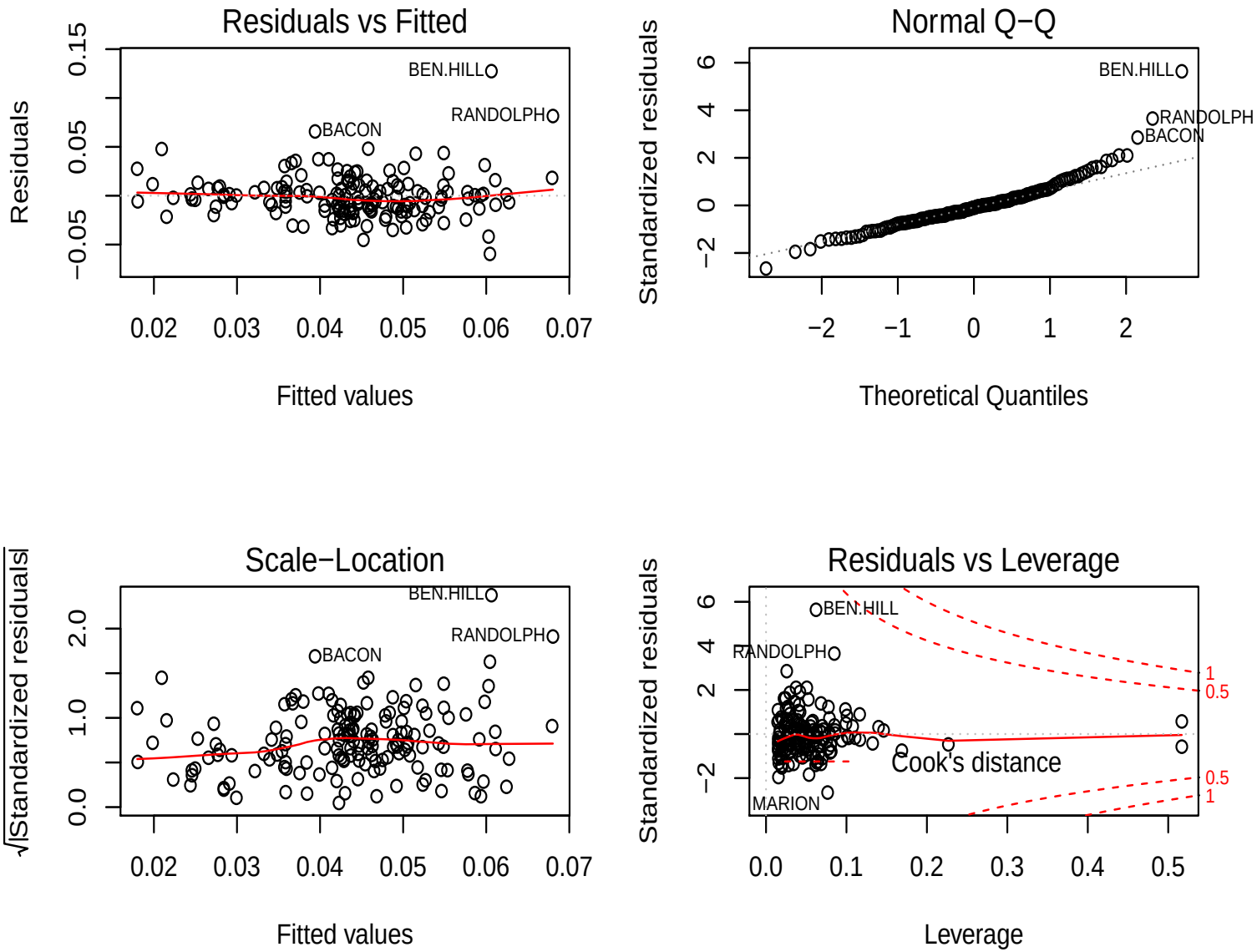

Figure 1.4: Diagnostics obtained from plotting the model object.

§1.9 Diagnostics in LM (8)

- The plot in the **upper-left panel** shows the residuals plotted against the fitted values. The plot can be used to detect lack of fit.
- If the residuals show some curvilinear trend, this is a sign that some change to the model is required, often a transformation of one of the variables. A smoothed curve has been added to the plot to aid in this assessment. In this instance, there is no sign of such a problem.
- The plot is also used to check the constant variance assumption on the errors. In this case, it seems the variance is roughly constant as the fitted values vary.
- Assuming symmetry of the errors, we can effectively double the resolution by plotting the absolute value of the residuals against the fitted values. As it happens $|\hat{\varepsilon}|$ tend to be rather skewed, it is better to use $\sqrt{|\hat{\varepsilon}|}$. Such a plot is shown in the **lower-left panel**, confirming what we have already observed about the constancy of the variance. Notice that a few larger residuals have been labeled.

§1.9 Diagnostics in LM (9)

- The residuals can be assessed for normality using a QQ plot.
- This compares the residuals to "ideal" normal observations. We plot the sorted residuals against $\Phi^{-1}(\frac{i}{n+1})$ for $i = 1, \cdots, n$. This can be seen in the **upper-right panel** of Figure 1.4.
- In this plot, the points follow a linear trend (except for one or two cases), indicating that normality is a reasonable assumption.
- If we observe a curve, this indicates skewness, suggesting a possible transformation of the response, while two tails of points diverging from linearity would indicate a long-tailed error, suggesting that we should consider robust fitting methods.
- Particularly for larger datasets, the normality assumption is not crucial, as the inference will be approximately correct in spite of the nonnormality. Only a clear deviation from normality should necessarily spur some action to change the model.

§1.9 Diagnostics in LM (10)

- Since $\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$, a large h_i tends to make $\text{Var}(\hat{\varepsilon}_i)$ small. The fit will be "forced" close to y_i. It is useful to examine the leverages to determine which cases have the power to be influential. Points on the boundary of the predictor space will have the most leverage.
- **Cook statistics** are a popular influence diagnostic as they reduce the information to a single value for each case. They are defined as:

$$D_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})^\top (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{p\hat{\sigma}^2} = \frac{\hat{\varepsilon}_i^2}{p\hat{\sigma}^2} \frac{h_i}{(1 - h_i)^2}$$

  They represent a scaled measure of the change in the fit if the single case is dropped from the data.
- Information about the leverages and Cook statistics for `lmodi` is given in the **lower-right panel** of Figure 1.4. A large residual combined with a large leverage will result in a larger Cook statistic. The plot shows two contour lines for the Cook statistics as these are a function of the standardized residuals and leverages.

§1.9 Diagnostics in LM (11)

- We can see that there are a couple of cases that stick out and we should investigate more closely the influence of these points. We can pick out the top two influential cases with:

```
> gavote[cooks.distance(lmodi) > 0.1,]
          equip econ perAA usage    atlanta gore bush other votes ballots
BEN.HILL  OS-PC poor 0.282 rural notAtlanta 2234 2381    46  4661    5741
RANDOLPH  OS-PC poor 0.527 rural notAtlanta 1381 1174    14  2569    3021
          undercount   pergore   cpergore     cperAA
BEN.HILL   0.1881205 0.4792963 0.07097452 0.03901887
RANDOLPH   0.1496193 0.5375633 0.12924148 0.28401887
```

- These are the same two counties that stuck out in the boxplots seen in Figure 1.3. These points are influential because they have much higher undercounts than would be expected. Their leverages are not high so they do not have unusual predictor values. The standardized residual for Ben Hill is over 5. Roughly speaking, standardized residuals exceeding 3.5 deserve closer attention so this case would attract some attention.

§1.9 Diagnostics in LM (12)

- A useful technique for judging whether some leverages are unusually extreme is the **half-normal plot**, where we plot the sorted leverage values vs. $\Phi^{-1}(\frac{n+i}{2n+1})$.
- Left panel of Figure 1.5 displays the half-normal plot for the leverages of `lmodi`, showing two points with much higher leverage than the rest. These points are:

```
> halfnorm(hatvalues(lmodi))
> gavote[hatvalues(lmodi)>0.3,]
           equip econ perAA usage    atlanta gore bush other votes ballots
MONTGOMERY PAPER poor 0.243 rural notAtlanta 1013 1465    31  2509    2573
TALIAFERRO PAPER poor 0.596 rural notAtlanta  556  271     5   832     881
           undercount   pergore     cpergore       cperAA
MONTGOMERY 0.02487369 0.4037465 -0.004575261 1.886792e-05
TALIAFERRO 0.05561862 0.6682692  0.259947458 3.530189e-01
```

- These are the only two counties that use a paper ballot, so are the only cases determining the coefficient for paper. This is sufficient to give them high leverage as the remaining predictor values are all unremarkable. Note these counties were not identified as influential — having high leverage alone is not necessarily enough to be influential.

§1.9 Diagnostics in LM (13)

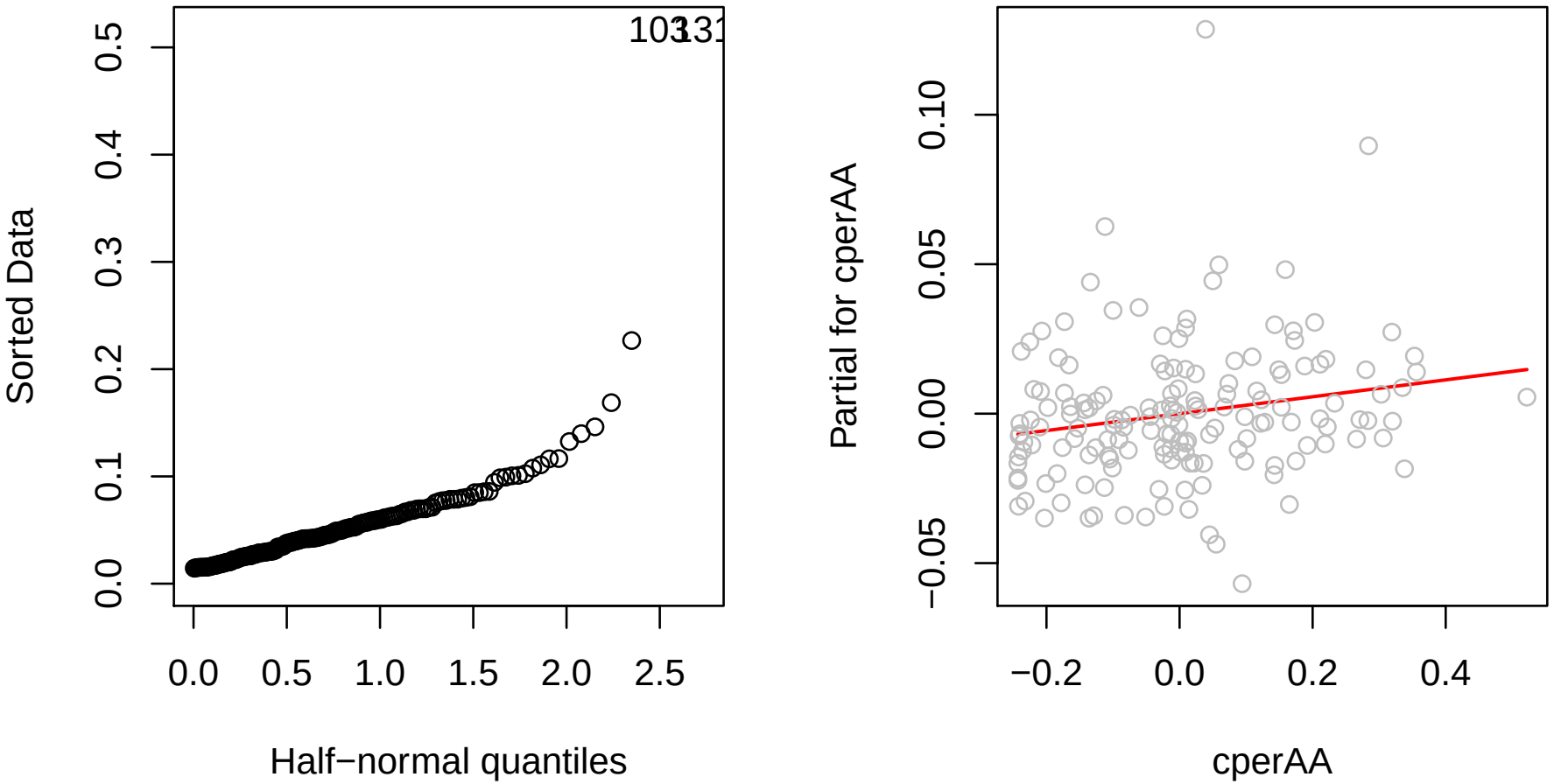

Figure 1.5: Half-normal plot of the leverages is shown on the left and a partial residual plot for the proportion of African Americans is shown on the right.

§1.9 Diagnostics in LM (14)

- **Partial residual plots** display $\hat{\varepsilon} + \hat{\beta}_i \tilde{\mathbf{x}}_i$ against $\tilde{\mathbf{x}}_i$, where$\tilde{\mathbf{x}}_i$ is the n observations of the predictor X_i.
- To see the motivation, look at the response with the predicted effect of the other X removed:

$$\mathbf{y} - \sum_{j \neq i} \tilde{\mathbf{x}}_j \hat{\beta}_j = \hat{\mathbf{y}} + \hat{\varepsilon} - \sum_{j \neq i} \tilde{\mathbf{x}}_j \hat{\beta}_j = \tilde{\mathbf{x}}_i \hat{\beta}_i + \hat{\varepsilon}$$

  Right panel of Figure 1.5 gives the partial residual plot for `cperAA`.

```
> termplot(lmodi,partial=TRUE,terms=1)
```

- The line is the LS fit to the data on this plot which is the same slope as the `cperAA` term in `lmodi`. This plot gives us a snapshot of the marginal relationship between this predictor and the response. In this case, we see a linear relationship indicating that it is not worthwhile seeking transformations. Furthermore, there is no sign that a few points are having undue influence on the relationship.

§1.10 Robust regression (1)

- Least squares performs poorly for long-tailed errors. Robust alternative to LS can be used instead.

```
> library(MASS)
> rlmodi <- rlm(undercount ~ cperAA+cpergore*usage+equip, gavote);
> summary(rlmodi)
Call: rlm(undercount ~ cperAA + cpergore * usage + equip,  data = gavote)
Residuals:
       Min          1Q      Median          3Q         Max
-6.026e-02 -1.165e-02 -6.587e-06  1.100e-02  1.379e-01
Coefficients:
                    Value   Std. Error t value
(Intercept)          0.0414  0.0023    17.8662
cperAA               0.0327  0.0254     1.2897
cpergore            -0.0082  0.0418    -0.1972
usageurban          -0.0167  0.0038    -4.4063
equipOS-CC           0.0069  0.0038     1.8019
equipOS-PC           0.0081  0.0048     1.6949
equipPAPER          -0.0059  0.0138    -0.4269
equipPUNCH           0.0170  0.0055     3.0720
cpergore:usageurban  0.0073  0.0316     0.2298

Residual standard error: 0.01722 on 150 degrees of freedom
```

§1.10 Robust regression (2)

- Inferential methods are more difficult to apply when robust estimation methods are used, hence there is less in the above output than for the corresponding `lmodi` output previously.
- The most interesting change is that the coefficient for `OS-PC` is now about half the size. Recall that, using the treatment coding, this represents the difference between `OS-PC` and the reference `lever` method.
- There is some fluctuation in the other coefficients, but not enough to change our impression of the important effects. The robust fit here has reduced the effect of the two outlying counties.

§1.11 Weighted least squares (1)

- The sizes of the counties in gavote vary greatly with the # of ballots cast in each county ranging from 881 to 280,975. The proportion of undercounted votes is expected to be more variable in smaller counties than in larger ones. Since the responses from the larger counties might be more precise, perhaps they should count for more in the fitting of the model.
- This effect can be achieved by the use of weighted least squares where we attempt to minimize $\sum w_i \varepsilon_i^2$. The appropriate choice for the weights w_i is to set them to be inversely proportional to $\text{Var}(y_i)$.
- Now $\text{Var}(y)$ for a binomial proportion is inversely proportional to the group size, in this case, the number of ballots. This suggests setting the weights proportional to the number of ballots:

```
> wlmodi <- lm(undercount~cperAA+cpergore*usage+equip, gavote,weights=ballots)
> sumary(wlmodi)
```

§1.11 Weighted least squares (2)

- This results in a fit that is substantially different from the unweighted fit. It is dominated by the data from a few large counties.

```
                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)          0.0436866  0.0034429 12.6889 < 2.2e-16
cperAA               0.0680927  0.0275381  2.4727  0.014527
cpergore            -0.0468847  0.0527468 -0.8889  0.375499
usageurban          -0.0179164  0.0037210 -4.8150 3.562e-06
equipOS-CC           0.0055822  0.0046484  1.2009  0.231682
equipOS-PC          -0.0058228  0.0046888 -1.2419  0.216229
equipPAPER          -0.0141543  0.0372938 -0.3795  0.704827
equipPUNCH           0.0156606  0.0053761  2.9130  0.004127
cpergore:usageurban  0.0119975  0.0356354  0.3367  0.736833

n = 159, p = 9, Residual SE = 2.17941, R-Squared = 0.41
```

- However, the variation in the response is likely to be caused by more than just binomial variation due to the number of ballots. There are likely to be other variables that affect the response in a way that is not proportional to ballot size.

§1.12 Transformation (1)

- Models may be improved by transforming the variables. Ideas for transformations can come from several sources.
- One method is to search through a family of possible transformations looking for the best fit. E.g. the Box–Cox transformation on the response variable.
- Alternatively, the diagnostic plots for the current model can suggest transformations that might improve the fit or ameliorate apparent violations of the assumptions.
- In other situations, transformations may be motivated by theories concerning the relationship between the variables or to aid the interpretation of the model.

§1.12 Transformation (2)

- For `gavote`, transformation of the response is problematic for both technical and interpretational reasons.
  - Still one may use the `boxcox` function in `MASS` to come up with a square root transformation on the response (add e.g. 0.005 to it to ensure positivity).
  - One may also use the log-response transformation.
  - Further, a variance stabilization transformation may be used.
- Transformations of the predictors are less problematic. Let's first consider the proportion of African Americans predictor in the current model. Polynomials provide a commonly used family of transformations.
- The use of **orthogonal polynomials** is recommended as these are more numerically stable and make it easier to select the correct degree.

§1.12 Transformation (3)

```
> plmodi <- lm(undercount ~ poly(cperAA,4)+cpergore*usage+equip, gavote)
> summary(plmodi)
Call: lm(formula=undercount ~ poly(cperAA, 4)+cpergore * usage+equip, data = gavote)
Residuals:
      Min        1Q    Median        3Q       Max
-0.058563 -0.012963 -0.001987  0.009230  0.127984
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.043460   0.002875  15.115  < 2e-16 ***
poly(cperAA, 4)1     0.052258   0.069391   0.753  0.45260
poly(cperAA, 4)2    -0.002988   0.026135  -0.114  0.90914
poly(cperAA, 4)3    -0.005363   0.024267  -0.221  0.82538
poly(cperAA, 4)4    -0.016513   0.024199  -0.682  0.49606
cpergore             0.013153   0.056930   0.231  0.81761
usageurban          -0.019129   0.004741  -4.035 8.76e-05 ***
equipOS-CC           0.006440   0.004720   1.364  0.17455
equipOS-PC           0.015587   0.005879   2.652  0.00889 **
equipPAPER          -0.010272   0.017204  -0.597  0.55137
equipPUNCH           0.014053   0.006866   2.047  0.04247 *
cpergore:usageurban -0.010538   0.041362  -0.255  0.79926
---
Residual standard error: 0.02354 on 147 degrees of freedom
Multiple R-squared:  0.1726,    Adjusted R-squared:  0.1107
F-statistic: 2.788 on 11 and 147 DF,  p-value: 0.002539
```

§1.12 Transformation (4)

- The advantage of the orthogonal polynomials is that the coefficients for the lower-order terms do not change as we change the maximum degree of the model.
- Here we see that all the terms of `cperAA` are not significant and all can be removed. Some insight into the relationship may be gained by plotting the fit on top of the partial residuals:

  ```
  > termplot(plmodi,partial=TRUE,terms=1)
  ```

- The plot, seen in the first panel of Figure 1.6, shows that the quartic polynomial is not so different from a constant fit, explaining the lack of significance.

§1.12 Transformation (5)

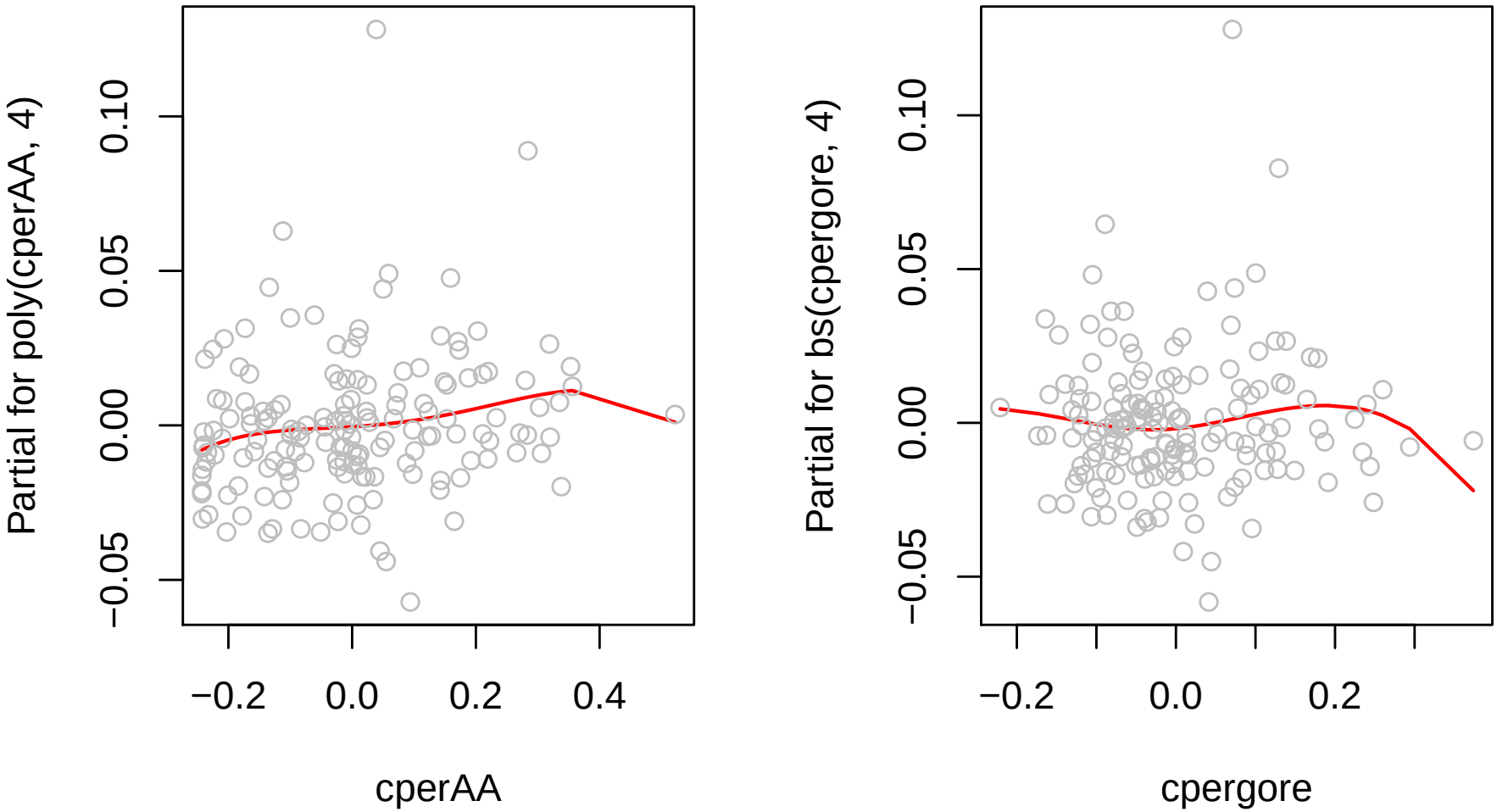

Figure 1.6: Partial fits using orthogonal polynomials for `cperAA` (shown on the left) and cubic B-splines for `cpergore` (shown on the right).

§1.12 Transformation (6)

- Polynomial fits become less attractive with higher-order terms. The fit is not local in the sense that a point in one part of the range of the variable affects the fit across the whole range. Furthermore, polynomials tend to have rather oscillatory fits and extrapolate poorly.
- A more stable fit can be obtained by using **splines**, which are piecewise polynomials, and have the local fit and stable extrapolation.
- We demonstrate the use of cubic B-splines here:

```
> library(splines)
> blmodi <- lm(undercount ~ cperAA+bs(cpergore,4)+usage+equip, gavote)
> termplot(blmodi,partial=TRUE,terms=2)
```

- Because the spline fit for `cperAA` was very similar to orthogonal polynomials, we consider `cpergore` here instead. The interaction with `usage` was removed for simplicity. The complexity of the B-spline fit may be controlled by specifying the degrees of freedom. The nature of the fit can be seen in the second panel of Figure 1.6, which we see is not much different from a constant

§1.12 Transformation (7)

```
> summary(blmodi)
Call: lm(formula = undercount~cperAA+ bs(cpergore, 4) +usage+equip, data=gavote)

Residuals:
      Min        1Q    Median        3Q       Max
-0.057902 -0.013422 -0.002504  0.008562  0.126729

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.048166   0.020962   2.298  0.02297 *
cperAA            0.027045   0.031168   0.868  0.38695
bs(cpergore, 4)1 -0.001337   0.026753  -0.050  0.96020
bs(cpergore, 4)2 -0.018839   0.023693  -0.795  0.42783
bs(cpergore, 4)3  0.019684   0.033185   0.593  0.55399
bs(cpergore, 4)4 -0.026591   0.034162  -0.778  0.43759
usageurban       -0.019426   0.004752  -4.088  7.1e-05 ***
equipOS-CC        0.006815   0.004691   1.453  0.14843
equipOS-PC        0.015527   0.005845   2.657  0.00876 **
equipPAPER       -0.008347   0.016939  -0.493  0.62292
equipPUNCH        0.012998   0.006926   1.877  0.06251 .

Residual standard error: 0.02333 on 148 degrees of freedom
Multiple R-squared:  0.1817,    Adjusted R-squared:  0.1264
F-statistic: 3.286 on 10 and 148 DF,  p-value: 0.0007199
```

§1.13 Variable selection (1)

- One theoretical view of the problem of variable selection is that one subset of the available variables represents the correct model for the data and that any method should be judged by its success in identifying this correct model.
- While this may be a tempting world in which to test competing variable selection methods, it seems unlikely to match with reality.
- Even if we believe that a correct model ever exists, it is more than likely that we will not have recorded all the relevant variables or not have chosen the correct transformations or functional form for the model amongst the set we choose to consider.
- We might then retreat from this ideal view and hope to identify the best model from the available set.
- Even then, we would need to define what is meant by best.

§1.13 Variable selection (2)

- Linear modeling serves two broad goals: *making predictons* and *understanding the relationship*.
- Re. prediction — we expect to observe new X and wish to predict y, along with measures of uncertainty in the prediction.
- Prediction is improved by removing variables contributing little or nothing to the model. We can define a criterion for prediction by which we search for the model optimizing that criterion. One such criterion is the adjusted R^2 mentioned before. The `regsubsets` function in the `leaps` package implements this search.
- For problems involving a moderate # of variables, it is possible to exhaustively search all possible models for the best. As the # of variables increases, exhaustive search becomes prohibitive and various stepwise or stochastic search methods must be used to search the model space.

§1.13 Variable selection (3)

- Another popular criterion is **AIC** defined as:

$$\text{AIC} = -2\text{maximum log likelihood} + 2p$$

  where p is the number of parameters. This criterion has the advantage of generality and can be applied far beyond normal linear models. AIC becomes BIC when $2p$ there is replaced with $p \log n$.
- The `step` command implements a stepwise search strategy through the space of possible models. It does allow qualitative variables and respects the hierarchy principle.
- We start by defining a rather large model, then use `step` to find a small model minimizing AIC.

§1.13 Variable selection (4)

```
> biglm <- lm(undercount ~ (equip+econ+usage+atlanta)^2 +
                           (equip+econ+usage+atlanta)*(perAA+pergore), gavote)
> smallm <- step(biglm,trace=FALSE); smallm

Call:  lm(formula = undercount ~ equip + econ + usage + perAA + equip:econ +
    equip:perAA + usage:perAA, data = gavote)

Coefficients:
        (Intercept)           equipOS-CC           equipOS-PC
          0.0435310           -0.0128784            0.0034922
         equipPAPER           equipPUNCH             econpoor
          -0.0578329           -0.0142618            0.0180113
            econrich           usageurban                perAA
          -0.0157358           -0.0006736           -0.0389879
equipOS-CC:econpoor  equipOS-PC:econpoor  equipPAPER:econpoor
          -0.0114503            0.0424178                   NA
equipPUNCH:econpoor  equipOS-CC:econrich  equipOS-PC:econrich
          -0.0160832            0.0047127           -0.0111987
equipPAPER:econrich  equipPUNCH:econrich     equipOS-CC:perAA
                   NA            0.0168340            0.1181524
   equipOS-PC:perAA     equipPAPER:perAA     equipPUNCH:perAA
          0.0321434            0.1260840            0.1243346
   usageurban:perAA
          -0.0472147
```

§1.13 Variable selection (5)

- Linear modeling is also used to try to understand the relationship between the variables — to develop an explanation for the data.
- For `gavote`, we are more interested in explanation than prediction. However, the two goals are not mutually exclusive and often the same methods are used for variable selection in both cases.
- Even so, when explanation is the goal, it may be unwise to rely on completely automated variable selection methods. **E.g.**, `pergore` was eliminated from the model by the AIC-based `step` method and yet we know `pergore` to be strongly correlated with `perAA` which is in the model. It would be rash to conclude that the latter variable is important and the former is not—the two are intertwined.
- Researchers interested in explaining the relationship may prefer a more manual variable selection approach that takes into account background information and is geared toward the substantive questions of interest.

§1.13 Variable selection (6)

- Another major class of variable selection methods is based on testing.
- E.g. F-tests are used to compare larger models with smaller nested models. A stepwise testing approach is then applied to select a model.
- The consensus view among statisticians is that this is an inferior method to variable selection compared to the criterion-based methods. Nevertheless, testing-based methods are still useful, particularly when under manual control. They have the advantage of applicability across a wide class of models where tests have been developed. They allow the user to respect restrictions of hierarchy and situations where certain variables must be included for explanatory purposes.

§1.13 Variable selection (7)

- Let's compare the AIC-selected models above to models with one fewer term:

```
> drop1(smallm,test="F")
Single term deletions

Model:
undercount ~ equip + econ + usage + perAA + equip:econ + equip:perAA +
    usage:perAA
             Df Sum of Sq      RSS     AIC F value   Pr(>F)
<none>                    0.053627 -1231.1
equip:econ    6 0.0075232 0.061150 -1222.3  3.2500 0.005084 **
equip:perAA   4 0.0068439 0.060471 -1220.0  4.4348 0.002101 **
usage:perAA   1 0.0010214 0.054649 -1230.1  2.6474 0.105984
```

- We see that the usage:perAA can be dropped. A subsequent test reveals that usage can also be removed. This gives us a final model of:

```
> finalm <- lm(undercount~equip +econ +perAA +equip:econ + equip:perAA, gavote)
> sumary(finalm)
```

§1.13 Variable selection (8)

```
> sumary(finalm)
Coefficients: (2 not defined because of singularities)
                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)          0.0418709  0.0050276  8.3282 6.497e-14
equipOS-CC          -0.0113268  0.0073726 -1.5363 0.1266999
equipOS-PC           0.0085750  0.0111781  0.7671 0.4442883
equipPAPER          -0.0584275  0.0370141 -1.5785 0.1166874
equipPUNCH          -0.0157513  0.0187454 -0.8403 0.4021751
econpoor             0.0202659  0.0055290  3.6654 0.0003489
econrich            -0.0169664  0.0123919 -1.3692 0.1731284
perAA               -0.0420403  0.0165935 -2.5335 0.0123853
equipOS-CC:econpoor -0.0109645  0.0098849 -1.1092 0.2692236
equipOS-PC:econpoor  0.0483848  0.0137954  3.5073 0.0006076
equipPUNCH:econpoor -0.0035601  0.0124266 -0.2865 0.7749211
equipOS-CC:econrich  0.0022777  0.0153780  0.1481 0.8824646
equipOS-PC:econrich -0.0133182  0.0170541 -0.7809 0.4361491
equipPUNCH:econrich  0.0200315  0.0219974  0.9106 0.3640450
equipOS-CC:perAA     0.1072494  0.0328551  3.2643 0.0013771
equipOS-PC:perAA    -0.0059062  0.0434140 -0.1360 0.8919805
equipPAPER:perAA     0.1291364  0.0818061  1.5786 0.1166763
equipPUNCH:perAA     0.0868490  0.0464997  1.8677 0.0638751

n = 159, p = 18, Residual SE = 0.02000, R-Squared = 0.43
```

Because there are only two paper-using counties, there is insufficient data to estimate the interaction terms involving paper. This model output is difficult to interpret because of the interaction terms.

§1.14 Conclusion (1)

- Let's attempt an interpretation of this final model. Certainly we should explore more models and check more diagnostics, so our conclusions can only be tentative.
- To interpret interactions, it is often helpful to construct predictions for all the levels of the variables involved. Here we generate all combinations of `equip` and `econ` for a median proportion of `perAA`:

```
> pdf <- data.frame(econ=rep(levels(gavote$econ), 5),
          equip=rep(levels(gavote$equip), rep(3,5)), perAA=0.233)
```

- We now compute the predicted undercount for all 15 combinations and display the result in a table:

```
> pp <- predict(finalm,new=pdf)
> xtabs(round(pp,3) ~ econ + equip, pdf)
        equip
econ      LEVER  OS-CC  OS-PC  PAPER  PUNCH
  middle  0.032  0.046  0.039  0.004  0.037
  poor    0.052  0.055  0.108  0.024  0.053
  rich    0.015  0.031  0.009 -0.013  0.040
```

§1.14 Conclusion (2)

- We can see that the undercount is lower in richer counties and higher in poorer counties. The amount of difference depends on the voting system.
- Of the three most commonly used voting methods, the LEVER method seems best. It is hard to separate the two optical scan methods, but there is clearly a problem with the precinct count in poorer counties, which is partly due to the two outliers we observed earlier.
- We notice one impossible prediction — a negative undercount in rich paper-using counties, but given the absence of such data (there were no such counties), we are not too disturbed.

§1.14 Conclusion (3)

- We use the same approach to investigate the relationship between perAA and the voting equipment. We set perAA at three levels — the first quartile, the median and the third quartile — and then compute the predicted undercount for all types of voting equipment. We set the econ variable to middle:

```
> pdf < -data.frame(econ=rep("middle",15), equip=rep(levels(gavote$equip),
                            rep(3,5)),  perAA=rep(c(.11,0.23,0.35),5))
> pp <- predict(finalm,new=pdf)
```

- We create a three-level factor for the three levels of perAA to aid the construction of the table:

```
> propAA <- gl(3,1,15,labels=c("low","medium","high"))
> xtabs(round(pp,3) ~ propAA + equip,pdf)
```

```
         equip
propAA     LEVER  OS-CC  OS-PC  PAPER  PUNCH
  low      0.037  0.038  0.045 -0.007  0.031
  medium   0.032  0.046  0.039  0.003  0.036
  high     0.027  0.053  0.034  0.014  0.042
```

§1.14 Conclusion (4)

- We see that the effect of the proportion of African Americans on the undercount is mixed. High proportions are associated with higher undercounts for OS-CC and PUNCH and associated with lower undercounts for LEVER and OS-PC.
- In summary, we have found that the economic status of a county is the clearest factor determining the proportion of undercounted votes, with richer counties having lower undercounts. The type of voting equipment and the proportion of African Americans do have some impact on the response, but the direction of the effects is not simply stated.
- We would like to emphasize again that this dataset deserves further analysis before any definitive conclusions are drawn.