北京航空航天大學
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

Generalized Linear Models

Lecture 2: Linear models with R

Outline

Linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

- Response: y
- Predictors: $x_1, \ldots, x_{p-1}$
- Error: ε

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{p-1,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{p-1,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & x_{2,n} & \ldots & x_{p-1,n} \end{bmatrix}$$

(the model matrix).

Linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

- Response: y
- Predictors: $x_1, \ldots, x_{p-1}$
- Error: ε

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p-1,n} \end{bmatrix}$$

(the model matrix).

Then

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Matrix formulation

Least squares estimation

Minimize: $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$

Matrix formulation

Least squares estimation

Minimize: $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$

Differentiate wrt β gives

The "normal" equations

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)\right)$$

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)\right)$$

which is maximized when $(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)$ is minimized.

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)\right)$$

which is maximized when $(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)$ is minimized.

So **MLE = OLS**.

R modelling notation

```
fit <- lm(response ~ x1 + x2 + x3,
  data=tibble)
```

$$\texttt{response} = \beta_0 + \beta_1 \texttt{x1} + \beta_2 \texttt{x2} + \beta_3 \texttt{x3} + \varepsilon$$

Outline

Have a look at the data

```
library(faraway)
data(gavote)
gavote
head(gavote)
?gavote
str(gavote)
```

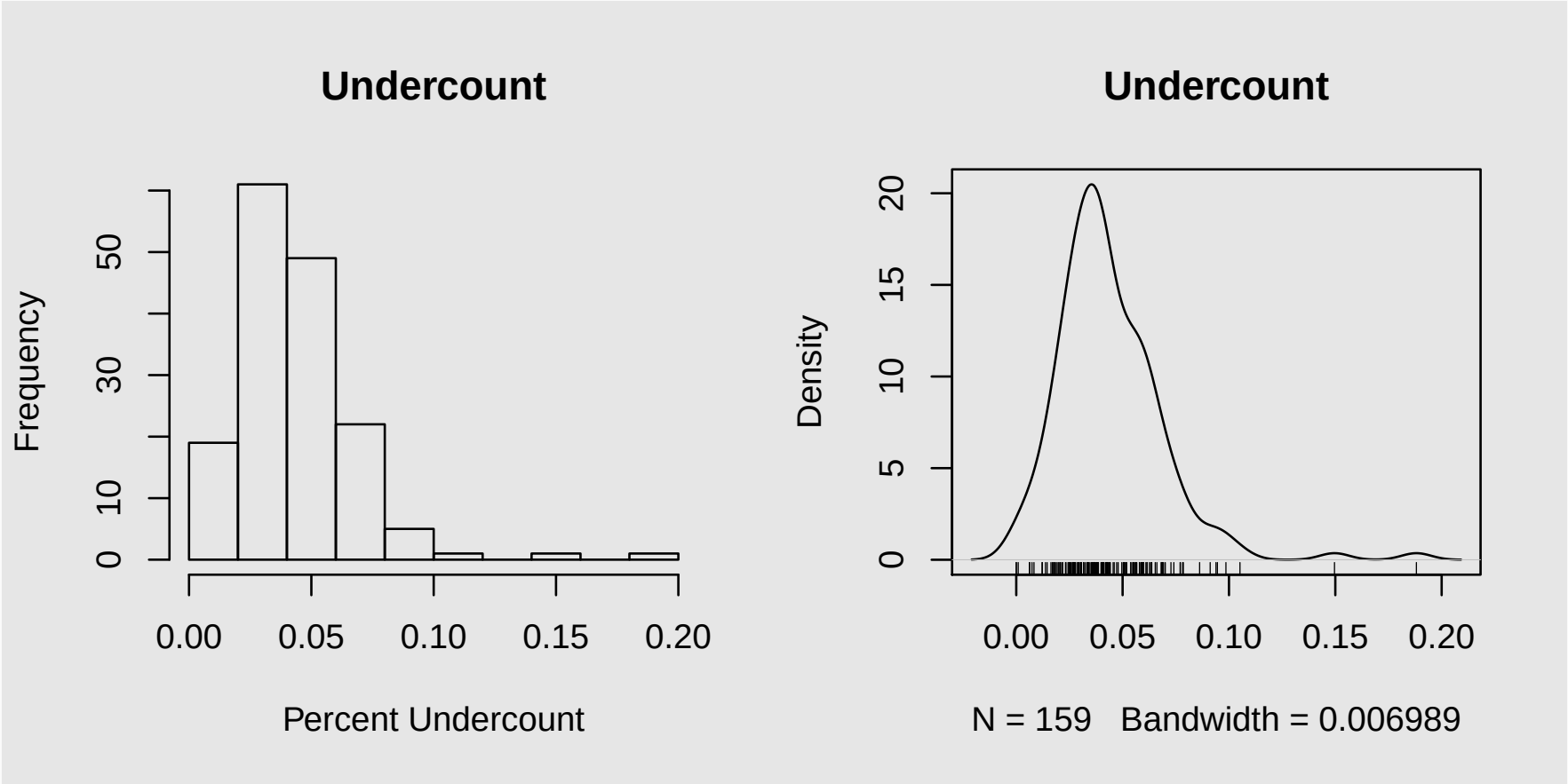
Initial data analysis

```
summary(gavote)
gavote$undercount <-
  (gavote$ballots-gavote$votes)/gavote$ballots
summary(gavote$undercount)
with(gavote,
     sum(ballots-votes)/sum(ballots))
```

Initial data analysis - visualization

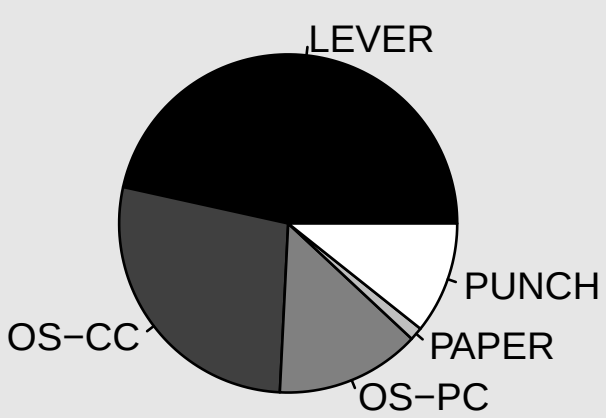
```
par(mfrow = c(1,2))
hist(gavote$undercount,main="Undercount",
     xlab="Percent Undercount")
plot(density(gavote$undercount),main="Undercount")
rug(gavote$undercount)
```

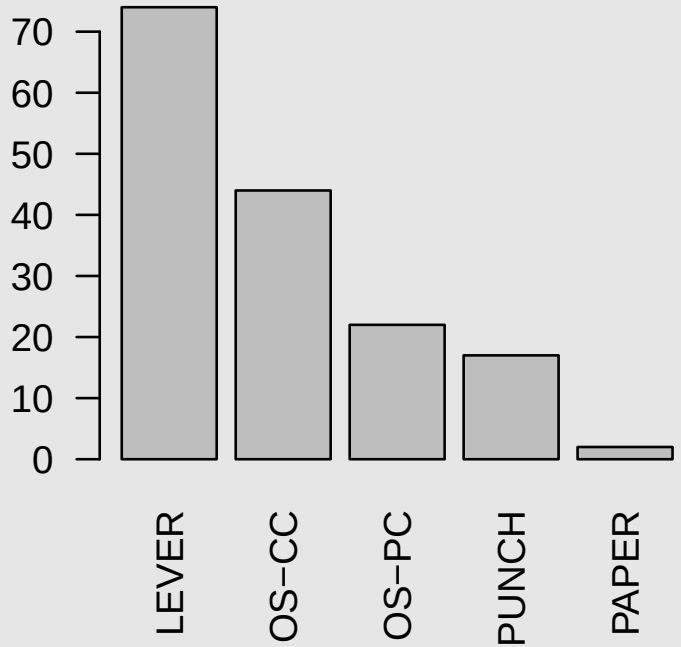
Initial data analysis - visualization



Initial data analysis - visualization

```
par(mfrow = c(1,2))
pie(table(gavote$equip),col=gray(0:4/4))
barplot(sort(table(gavote$equip),decreasing=TRUE),las=2)
```

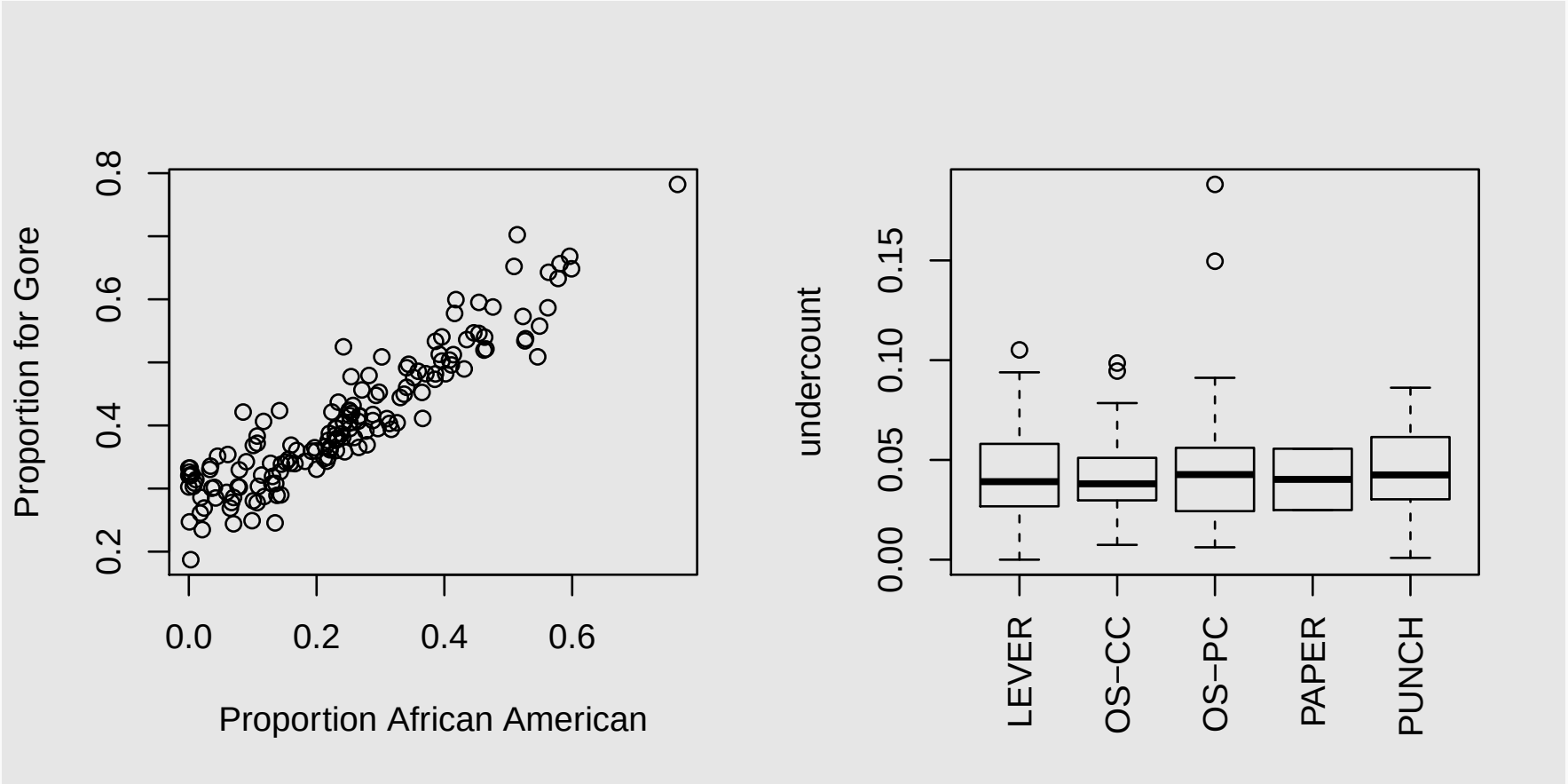




Initial data analysis - visualization

```
par(mfrow = c(1,2))
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)
```

Initial data analysis - visualization



Correlations

```
names(gavote)
names(gavote)[4] <- "usage"
nix <- c(3,10,11,12)
cor(gavote[,nix])
```

Outline

Fitting a linear model

$$\text{undercount} = \beta_0 + \beta_1 \text{pergore} + \beta_2 \text{perAA} + \epsilon$$

```
lmod <- lm(undercount ~ pergore + perAA, gavote)
```

Useful helper functions

- `coef`: $\hat{\beta}$
- `fitted`: $\hat{y}$
- `predict`: predicted values based on linear model
- `residuals`: $y - \hat{y}$
- `deviance`: RSS = $\hat{\epsilon}^T \hat{\epsilon}$
- `df.residual`: degrees of freedom = $n - p$
- `summary`: a summary of the model

Fitting a linear model

```
predict(lmod)
residuals(lmod)
deviance(lmod)
df.residual(lmod)
nrow(gavote)-length(coef(lmod))
lmodsum <- summary(lmod)
lmodsum$sigma
```

Fitting a linear model

- Estimation of σ

$$\hat{\sigma} = \sqrt{\text{RSS}/\text{df}}$$

- R^2: coefficient of determination

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} = r^2 = \text{cor}^2(\hat{y}, y)$$

- Adjusted R^2

$$R^2_a = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)}$$

```
sqrt(deviance(lmod)/df.residual(lmod))
lmodsum$r.squared
cor(predict(lmod),gavote$undercount)^2
lmodsum$adj.r.squared
```

Qualitative predictors

- R will create the required dummy variables from a categorical *factor*.
- The first level is used as the reference category.
- Use `relevel` to change the reference category

$$\text{undercount} = \beta_0 + \beta_1 \text{pergore} + \beta_2 \text{perAA} + \beta_3 d + \epsilon$$

$$d = \begin{cases} 0 & \text{rural} \\ 1 & \text{urban} \end{cases}$$

Interactions

Interactions:

- Interactions are obtained by multiplying the relevant columns of the model matrix.
- Use `a:b` for the interaction between `a` and `b`.
- Use `a*b` to mean `a + b + a:b`

Limited order interactions:

- Interactions up to 2nd order can be specified using the `^` operator.
- `(a+b+c)^2` is identical to `(a+b+c)*(a+b+c)`

Nested factors:

- `a + b %in% a` expands to `a + a:b`.

Interpretation

- Each coefficient gives the effect of a one unit increase of the predictor on the response variable, *holding all other variables constant*.
- Be careful with interactions: you cannot interpret the main effects when there is an interaction between them.

Interpretation

```
gavote$cpergore <- gavote$pergore - mean(gavote$pergore)
gavote$cperAA <- gavote$perAA - mean(gavote$perAA)
lmodi <- lm(undercount ~ cperAA+cpergore*usage+equip,
            gavote)
summary(lmodi)
```

Hypothesis testing

- Use F-tests between models:
  - Model 1: p_1 parameters.
  - Model 2 (nested within Model 1): p_2 parameters.
  - H_0 : the smaller model is correct.

$$F = \frac{(\text{RSS}_2 - \text{RSS}_1)/(p_1 - p_2)}{\text{RSS}_1/(n - p_1)} \sim F_{p_1 - p_2, n - p_1}$$

Hypothesis Testing

```
anova(lmod,lmodi)
drop1(lmodi,test="F")
```

- Hierarchy principal: all lower-order terms corresponding to an interaction should be retained in the model.

Outline

Regression diagnostics

- Residuals vs Fitted: Check heteroskedasticity
- Scale-Location: standardized residuals vs fitted values.
- Normal QQ plot: Check for non-normality
- Residuals vs Leverage: Check for influential points

Produced using `ggfortify::autoplot` or `plot`

LOO Residuals

Fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the “hat matrix”.

LOO Residuals

Fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat matrix".

Theorem: Leave-one-out residuals

Let $h_1, \ldots, h_n$ be the diagonal values of $\boldsymbol{H}$. Then $e_{(i)} = e_i/(1 - h_i)$ is the prediction error that would be obtained if the ith observation was omitted.

LOO Residuals

Fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat matrix".

Theorem: Leave-one-out residuals

Let $h_1, \ldots, h_n$ be the diagonal values of $\boldsymbol{H}$. Then $e_{(i)} = e_i/(1-h_i)$ is the prediction error that would be obtained if the ith observation was omitted.

Leverage values

h_i is called the "leverage" of observation i.

In R: `hatvalues(fit)`

LOO Residuals (Optional)

Let $\boldsymbol{X}_{[i]}$ and $\boldsymbol{Y}_{[i]}$ be similar to $\boldsymbol{X}$ and $\boldsymbol{Y}$ but with the ith row deleted in each case. Let $\boldsymbol{x}_i'$ be the ith row of $\boldsymbol{X}$ and let

$$\hat{\beta}_{[i]} = (\boldsymbol{X}_{[i]}'\boldsymbol{X}_{[i]})^{-1}\boldsymbol{X}_{[i]}'\boldsymbol{Y}_{[i]}$$

be the estimate of β without the ith case. Then $e_{[i]} = y_i - \boldsymbol{x}_i'\hat{\beta}_{[i]}$.

Now $\boldsymbol{X}_{[i]}'\boldsymbol{X}_{[i]} = (\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i')$ and $\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i = h_i$.

LOO Residuals (Optional)

Sherman-Morrison-Woodbury formula

Suppose $\boldsymbol{A}$ is a square matrix, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are column vectors of the same dimension. Then

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}')^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}'\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}'\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

So by SMW,

$$(\boldsymbol{X}'_{[i]}\boldsymbol{X}_{[i]})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1 - h_i}.$$

Also note that $\boldsymbol{X}'_{[i]}\boldsymbol{Y}_{[i]} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{x}y_i$.

LOO Residuals (Optional)

Therefore

$$\hat{\beta}_{[i]} = \left[(X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1 - h_i} \right] (X'Y - x_i y_i)$$

$$= \hat{\beta} - \left[\frac{(X'X)^{-1}x_i}{1 - h_i} \right] \left[y_i(1 - h_i) - x_i'\hat{\beta} + h_i y_i \right]$$

$$= \hat{\beta} - (X'X)^{-1}x_i e_i/(1 - h_i)$$

LOO Residuals (Optional)

Thus

$$
\begin{aligned}
e_{[i]} &= y_i - \boldsymbol{x}_i' \hat{\beta}_{[i]} \\
&= y_i - \boldsymbol{x}_i' \left[\hat{\beta} - (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_i e_i / (1 - h_i) \right] \\
&= e_i + h_i e_i / (1 - h_i) \\
&= e_i / (1 - h_i),
\end{aligned}
$$

LOO Residuals

Cross-validation statistic

$$\text{CV} = \frac{1}{T}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

- Measures MSE of out-of-sample prediction
- Asymptotically equivalent to AIC (up to monotonic transformation)

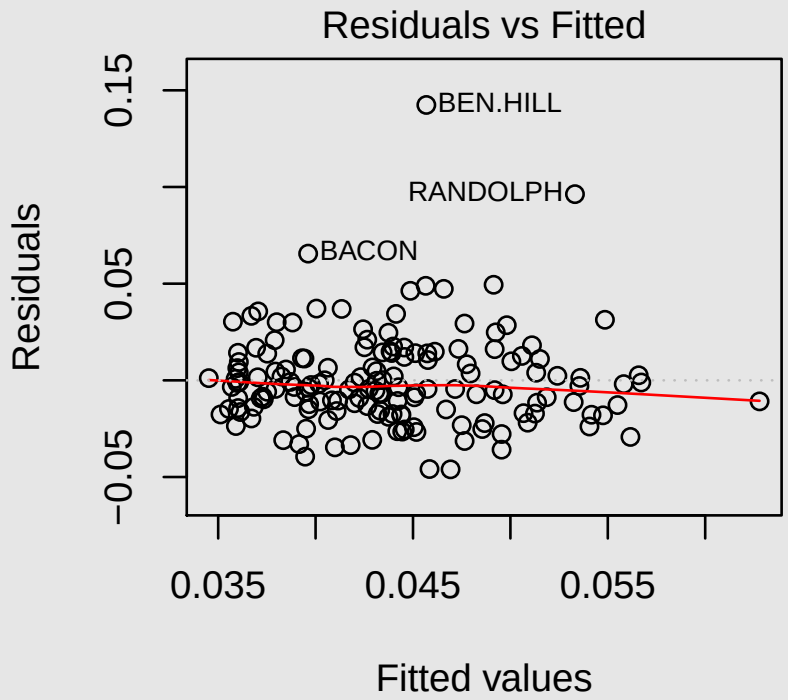
Cook distances

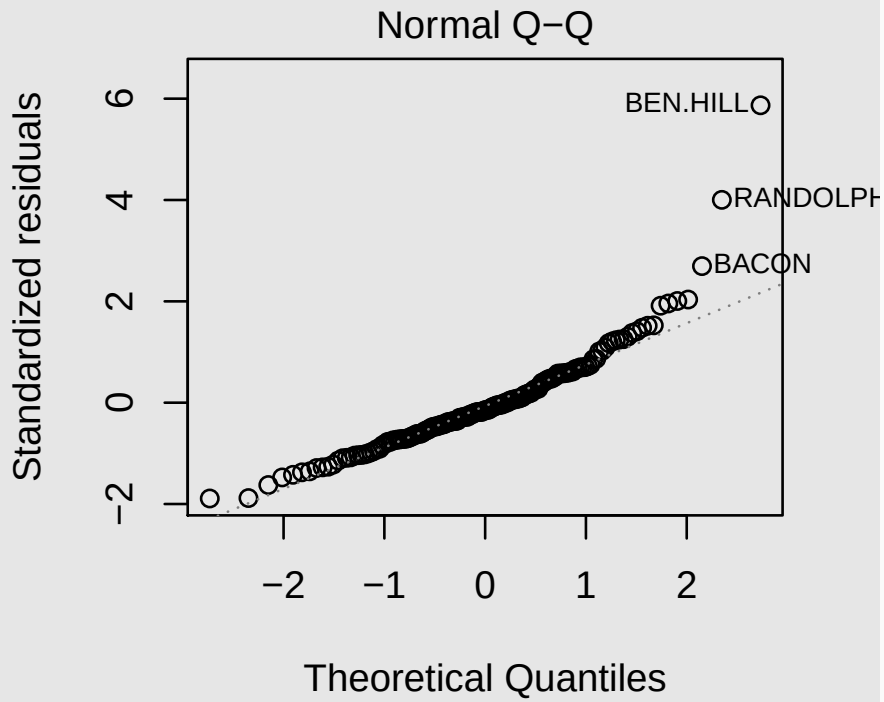
$$D_i = \frac{e_i^2 h_i}{\hat{\sigma}^2 p(1-h_i)}$$

- Measures change in fit if observation *i* dropped.

Regression diagnostics

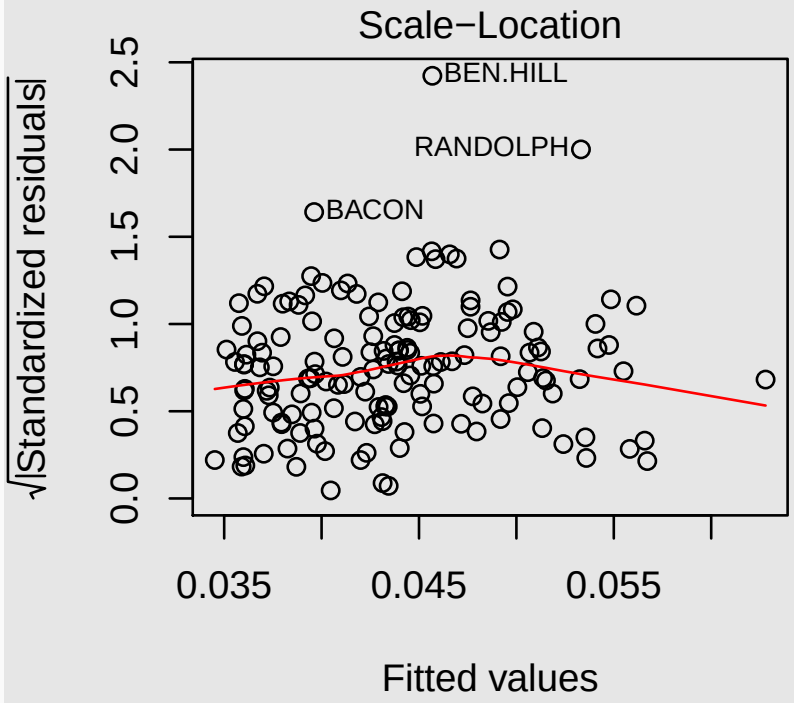
```
par(mfrow = c(1,2))
plot(lmod, which = c(1:2))
```

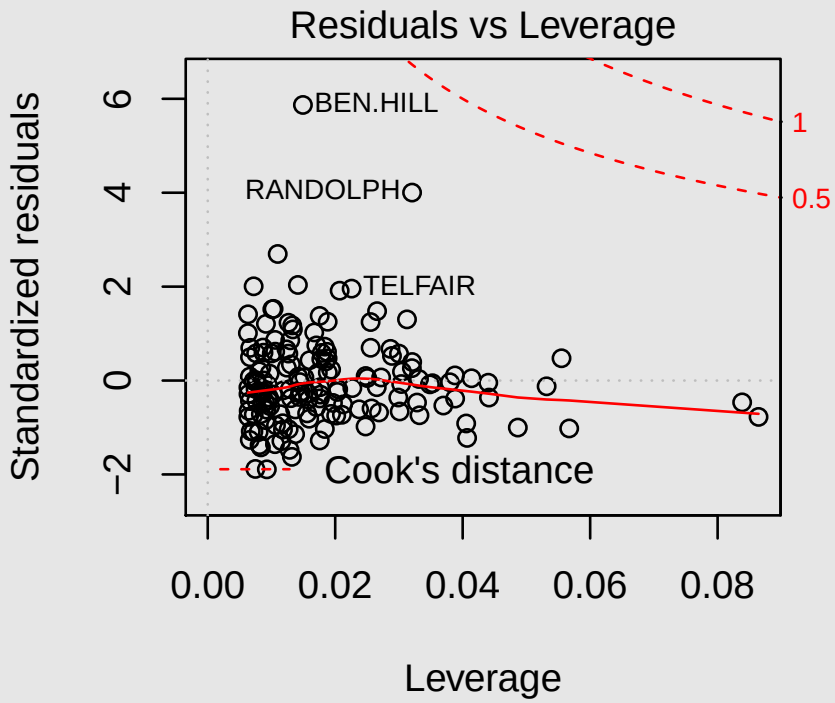




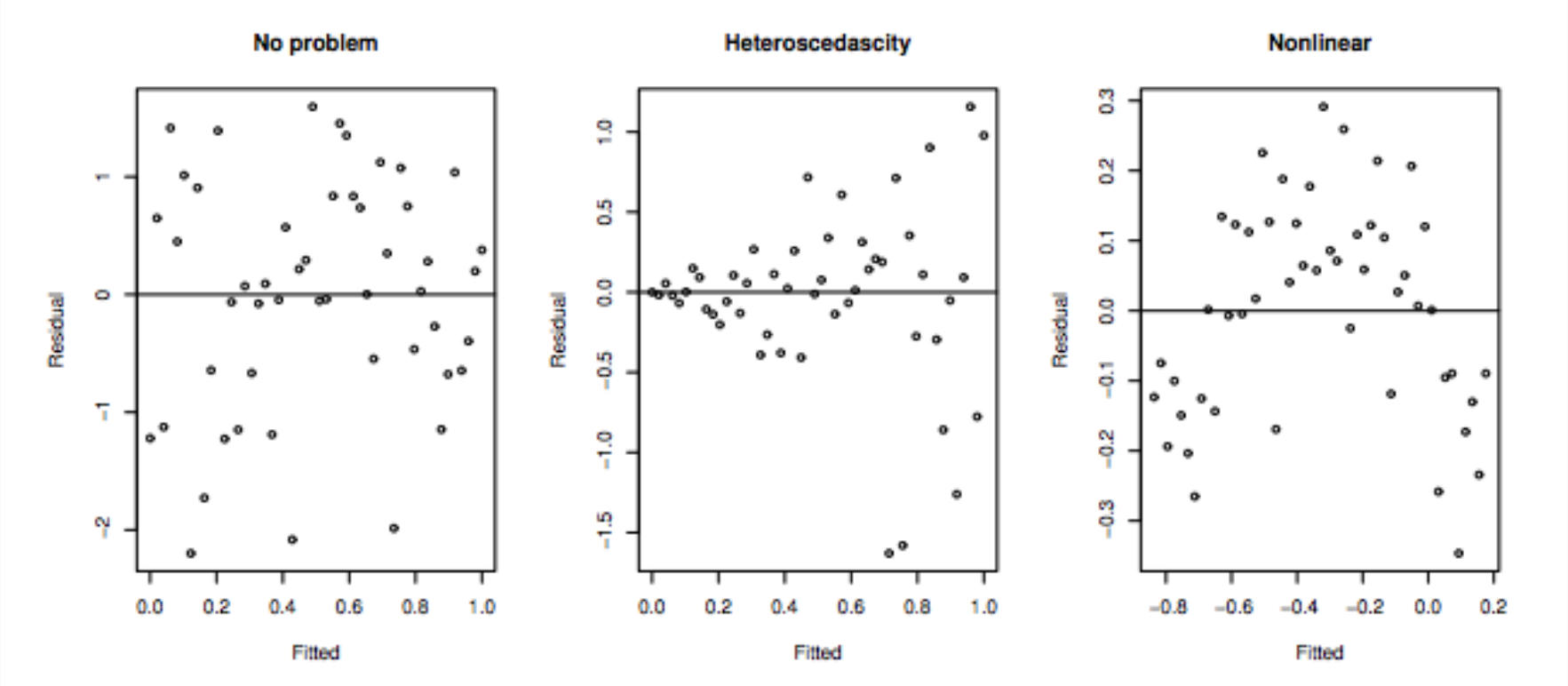
Regression diagnostics

```
par(mfrow = c(1,2))
plot(lmod, which = c(3,5))
```





Which one is not expected?



Robust regression

- OLS works well for normal errors
- How about when we have outliers?
  1. outliers are incorrect observations: drop them
  2. outliers are real observations: robust regression
- Robust regression can downweight the effects of large errors
- `rlm` in R

Polynomial regression

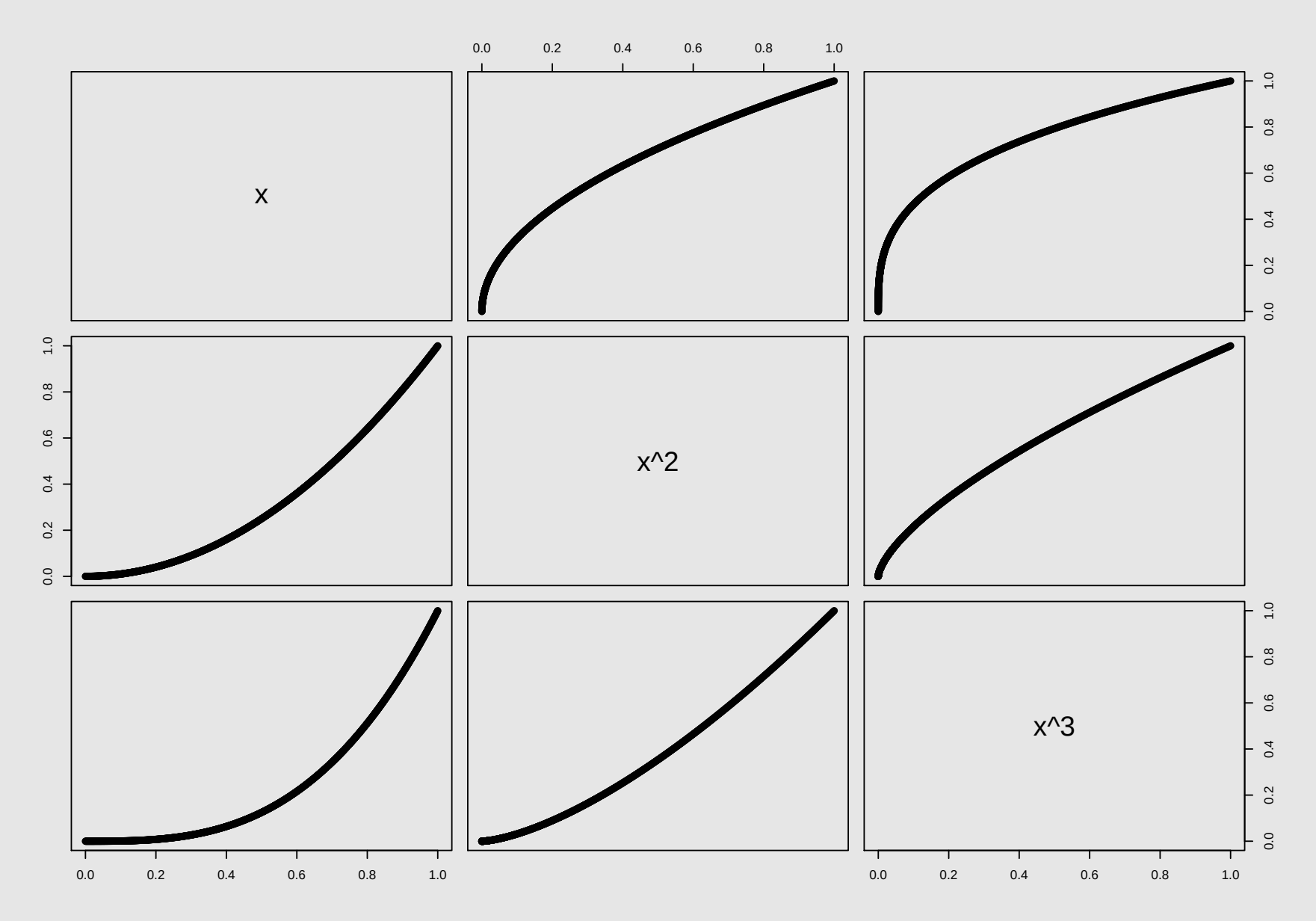
Simple approach: Add x, x^2, x^3, ... to model.

```
fit <- lm(y ~ x + I(x^2) + I(x^3))
```

Problems:

- predictors are highly correlated, so difficult to separate effects of terms.
- numerical instability of coefficients

Polynomial regression



Orthogonal polynomials

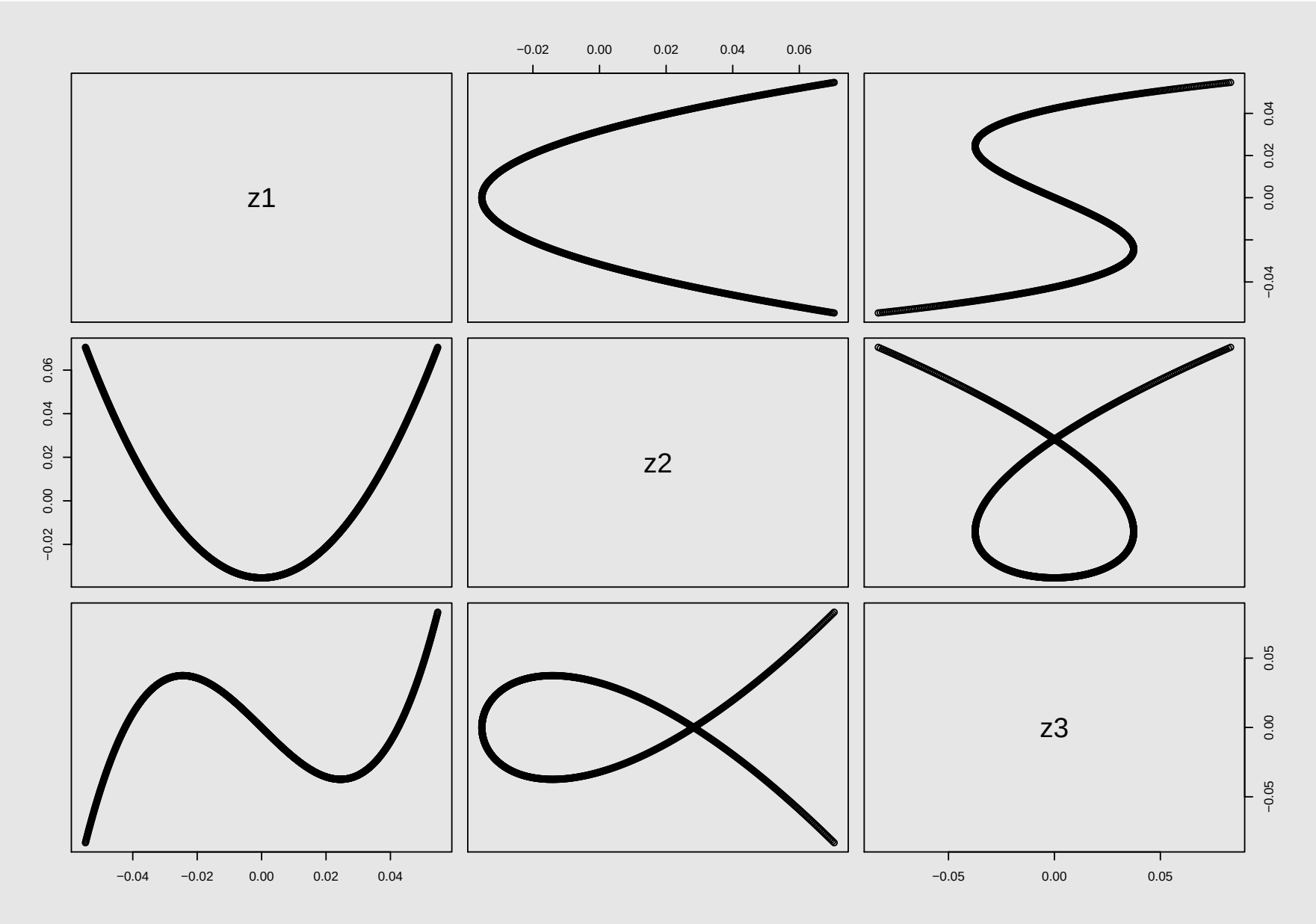
$$z_1 = a_1 + b_1 x$$
$$z_2 = a_2 + b_2 x + c_2 x^2$$
$$z_3 = a_3 + b_3 x + c_3 x^2 + d_3 x^3$$

where coefficients such that $z_i' z_j = 0$ when $i \neq j$.

```
fit <- lm(y ~ poly(x,3))
```

- Because z_i involve constants $a_1, a_2, \ldots$, the intercept will be affected.
- Coefficients of lower order terms are unchanged when higher order terms are added.

Orthogonal polynomials



Polynomial regression

```
plmodi <- lm(undercount ~ poly(cperAA,4)+
                cpergore*usage+equip, gavote)
summary(plmodi)
```

Outline

Variable selection

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

- `step` will minimize AIC using backwards selection
- `R330::allpossregs` will fit all possible regressions
- Do not use coefficient t-tests for variable selection
- Hierarchy principle: don't eliminate lower terms while retaining higher order terms.
- If explanation is the goal, you may not rely on completely automated variable selection methods.

Variable selection

```
biglm <- lm(undercount ~
              (equip+econ+usage+atlanta)^2+
              (equip+econ+usage+atlanta)*(perAA+pergore),
            gavote)
smallm <- step(biglm,trace=FALSE)
drop1(smallm,test="F")
finalm <- lm(undercount~equip + econ  + perAA +
               equip:econ + equip:perAA, gavote)
summary(finalm)
```

Outline

What is your data analysis like?

- An initial data analysis that explores the numerical and graphical characteristics of the data.
- Variable selection to choose the best model.
- An exploration of transformations to improve the fit of the model.
- Diagnostics to check the assumptions of your model.
- Some predictions of future observations for interesting values of the predictors.
- An interpretation of the meaning of the model with respect to the particular area of application.

Outline

Assignment One

The exercise on Page 24 of our textbook.

- Choose any one of the data sets and perform linear data analysis.
- Your analysis should at least consist of:
  - an initial data analysis
  - fitting of a linear model
  - interpretation
  - testing
  - diagnostics
  - variable selection
  - conclusion