For podcast: techical problems with audio, hopefully solved.

Recap

- introduction
- linear model
- deviance $D(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \| y - \hat{y} \|^2$

$$\hat{\sigma}^2 = s^2 = \frac{D(\hat{\beta})}{n-p} \qquad \begin{array}{l} n = \#\text{ observations} \\ p = \#\text{ variables} \end{array}$$

$$\widehat{Var}(\hat{\beta}) = s^2 (X^T X)^{-1}$$

H_0: $\beta = 0$

under assumption of normality $\dfrac{\hat{\beta} - 0}{\sqrt{s^2 (X^T X)^{-1}}} \overset{H_0}{\sim} \mathcal{N}(0;1)$ $\to$ t-value

Pr(obtaining a t-value "worse" than that actually obtained) = p-value

small p-value $\to$ evidence against H_0

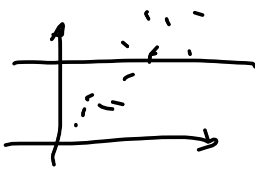
- polynomial
- dummy variables for categorical variables I_A

Today

- R^2 & graphical diagnostic
- variable transformations
- multivariate response
- computational tricks.

To evaluate the godness of fit, we need to compute the coefficient of determination

$$R^2 = \frac{\text{explained deviance}}{\text{total deviance}} = 1 - \frac{\text{residual deviance}}{\text{total deviance}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



$\hookrightarrow \bar{y} = \sum_{i=1}^{n} y_i$

R^2: - simple to interpret
(fraction of variability explained by the model)
- oversimplification
(everything is reduced to a number) $\longrightarrow$ graphical tools

$\Rightarrow$ graphical diagnosis, are based on the residuals

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \qquad i = 1, \dots, n \qquad \underline{Var(\varepsilon_i) = \sigma^2}$$

($\hat{\varepsilon}_i$ are surrogate of ε_i, that are not observable)

- figure 2.4(a) → Ascombe plot
  check violation of homoscedasticity
- figure 2.4(b) → quantile-quantile plot (qqplot) $\varepsilon_i \sim N(0, \sigma^2)$
  check violation of normality
  y-axis = (standardized) values of $\hat{\varepsilon}_i$
  x-axis = expected value under normality

<u>General conclusions</u>
- model is decent, especially for the "average" cars;
- simplicity;
- graphical diagnoses not totally satisfying
- model is bad for extrapolation;
- " " not realistic (cannot increase) the distance → for larger engine sizes

Variable transformations

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

linear model
↳ refers to the parameters (linear in the parameters)
we can use whatever transformation of the variables, as long as the parameters have a linear relationship

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \frac{1}{x_2} + \beta_3 x_1/x_2 + \beta_4 e^{x_2^2 + x_1} + \varepsilon$$

we can also transform y

E.g. transformation of the response

consumption $= \beta_0 + \beta_1$ (engine size) $+ \beta_2 I_A + \varepsilon$

distance covered

+ nice all
+ $R^2 = 0.64$
+ polynomial terms of higher orders are not necessary
↳ simplify the interpretation

going back to original scale:
- not totally satisfactory (left part of the plot including gas cars)
- $R^2 = 0.56$
- graphical diagnostics are unsatisfactory

Way more used transformation: logarithm
- especially good for variables with support $(0; +\infty)$, to transform into a variable with support $\mathbb{R}$ (more in line with normality assumptions)

$$\log(\text{distance covered}) = \beta_0 + \beta_1 \log(\text{engine size}) + \beta_2 I_A + \varepsilon$$

we cannot do much better with these variables, but we have a lot of additional information to try to explain the variability of y

→ add new variables to the model!

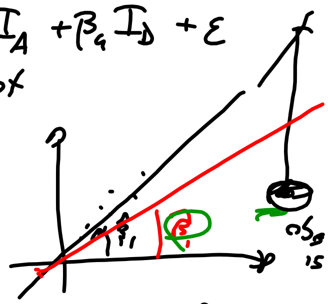
- weight, we know that can affects the distance per liter
- there are always points far from the others in the bottom left of the plots
  - → they belong to cars with engines with only 2 cylinders
    - → new dummy variables 2 cylinders / rest

$$I_j = \begin{cases} 1 & \text{if the engine has 2 cylinders} \\ 0 & \text{otherwise} \end{cases}$$

model:

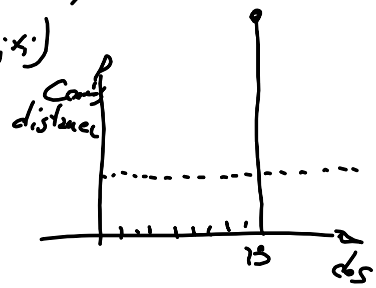
$$\log(\text{distance}) = \beta_0 + \beta_1 \log(\text{engine size}) + \beta_2 \log(\text{weight}) + \beta_3 I_A + \beta_4 I_D + \varepsilon$$

quite good result in Ascombe plot and quantile-quantile plot



Two additional graphical diagnostic tools:

- scatter plot of the residuals $\quad Cor(\varepsilon_i, \varepsilon_j) = 0 \;\; \forall i \neq j$
- Cook's distance
  - effect on $\hat{\beta}$ when removing a specific observation (y_j, x_j)
  - check if any point influence too much the estimates



Multivariate response

~~We have a~~ multivariable model → more than one explanatory variable

multivariate model → more than one response variable

$$\underline{y} = (y_1, y_2, \dots, y_q)$$

$$\underset{(n\times q)}{Y} = \underset{(n\times p)}{X}\,\underset{(p\times q)}{B} + \underset{(n\times q)}{E} \qquad B = \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pq} \end{pmatrix}$$

$$Var(E) = \underset{(q\times q)}{\Sigma}$$

$\sum_{i=1}^{n}$

Σ = capital sigma

$$\hat{B} = (X^T X)^{-1} X^T Y \qquad \hat{\Sigma} = \frac{1}{n-p} Y^T P Y$$

Computational aspects

→ very important in data analysis due to the size of the data (large n, large p)

→ $\hat{\beta} = \underbrace{(X^TX)^{-1}}X^Ty$

↳ for the least-squares estimator the most problematic task is the inversion of X^TX

→ Choleski factorization

A positive definite matrix

$\underset{\text{unique}}{A} = LL^{\cancel{*}T}$ where L is a lower triangular matrix
$L^{\cancel{*}T}$ is the ~~conjugate~~ transposed in $\mathbb{R}$

$$\underset{A}{\begin{pmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{pmatrix}} = \underset{L}{\begin{pmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{pmatrix}} \underset{L^T}{\begin{pmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{pmatrix}}$$

computational time $O(p^3 + np^2/2)$

→ Use the $\underbrace{\text{Gram-Schmidt process}}_{\text{orthogonalization}}$ to compute the QR decomposition

$X = QR$ such that Q $n \times p$ matrix, often $Q^TQ = I$
R $p \times p$ " , upper triangular

then $\hat{\beta} = R^{-1}Q^Ty$ → easy to invert because R is upper triangular

$\hat{y} = QQ^Ty$

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$
$$\overset{!}{=} (R^TQ^TQR)^{-1}Q^TR^Ty$$
$$\overset{!}{=} \cancel{(R^T)^{-1}R^T}R^{-1}Q^Ty \quad \text{to show}$$

n very large $\to$ storage problems

row by row procedure

$$\hat{\beta} = (\underbrace{X^T X}_{W = X^T X})^{-1} \underbrace{X^T y}_{U = X^T y}$$

W, U are the only terms needed to compute the OLS

$$X = \begin{pmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_n^T \end{pmatrix}$$ where $\tilde{x}_i^T$ is the i-th row of X

Then $W = \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T$, $U = \sum_{i=1}^{n} \tilde{x}_i y$

$$W_{(j)} = W_{(j-1)} + \tilde{x}_j \tilde{x}_j^T \quad , \quad U_{(j)} = U_{(j-1)} + \tilde{x}_j y_j \qquad j = 2, \dots, n$$

$$W_{(1)} = \tilde{x}_1 \tilde{x}_1^T \quad , \quad U_{(1)} = \tilde{x}_1 y_1$$

Recursive estimation

- we have W, we still need to invert it $\underset{p \times p}{W}$ ← problematic to invert, especially for large p

$$\hat{\beta}_{(n)} = V_{(n)} X_{(n)}^T y$$
$$\hookrightarrow W_{(n)}^{-1} = (X_{(n)}^T X_{(n)})^{-1}$$

update with a new observation $n+1$ $(\tilde{x}_{n+1}, y_{n+1})$

$$X_{(n+1)} = \begin{pmatrix} X_{(n)} \\ \tilde{x}_{n+1}^T \end{pmatrix} \Rightarrow W_{(n+1)} = W_{(n)} + \tilde{x}_{n+1} \tilde{x}_{n+1}^T$$
$$= X_{(n)}^T X_{(n)} + \tilde{x}_{n+1} \tilde{x}_{n+1}^T$$

SHERMAN-MORRISON FORMULA (see appendix A1, formula (A.2))

$$(A + bd^T)^{-1} = A^{-1} - \frac{1}{1 + d^T A^{-1} b} A^{-1} b d^T A^{-1}$$

$$V_{(n+1)} = V_{(n)} - \underbrace{\frac{1}{1 + \tilde{x}_{n+1}^T V_{(n)} \tilde{x}_{n+1}}}_{h} V_{(n)} \tilde{x}_{n+1} \tilde{x}_{n+1}^T V_{(n)}$$

$$\hat{\beta}_{(n+1)} = V_{(n+1)} \left(X_{(n)}^T y + \tilde{x}_{n+1} y_{n+1} \right)$$
$$= \left(V_{(n)} - \frac{1}{1 + \tilde{x}_{n+1}^T V_{(n)} \tilde{x}_{n+1}} V_{(n)} \tilde{x}_{n+1} \tilde{x}_{n+1}^T V_{(n)} \right) \left(X_{(n)}^T y + \tilde{x}_{n+1} y_{n+1} \right)$$

$$\hat{\beta}_{(n+1)} = V_{(n+1)}\left(X_{(n)}^T y + \tilde{x}_{n+1} y_{n+1}\right)$$

$$= \left(V_{(n)} - \overset{h}{\frac{1}{1+\tilde{x}_{n+1}^T V_{(n)} \tilde{x}_{n+1}}} V_{(n)} \tilde{x}_{n+1} \tilde{x}_{n+1}^T V_{(n)}\right)\left(X_{(n)}^T y + \tilde{x}_{n+1} y_{n+1}\right)$$

$$= \underbrace{V_{(n)} X_{(n)}^T y}_{\hat{\beta}_{(n)}} + V_{(n)} \tilde{x}_{n+1} y_{n+1} - h V_{(n)} \tilde{x}_{n+1} \tilde{x}_{n+1}^T V_{(n)} X_{(n)}^T y - h V_{(n)} \tilde{x}_{n+1} \tilde{x}_{n+1}^T V_{(n)} \tilde{x}_{n+1} y_{n+1}$$

$$= \hat{\beta}_{(n)} + h\Big(V_{(n)} \tilde{x}_{n+1} y_{n+1} + \tilde{x}_{n+1}^T \cancel{V_{(n)} \tilde{x}_{n+1} V_{(n)} \tilde{x}}_{n+1} y_{n+1} - V_{(n)} \tilde{x}_{n+1} \tilde{x}_{n+1}^T \underbrace{V_{(n)} X_{(n)}^T y}_{\hat{\beta}_{(n)}} - V_{(n)} \cancel{\tilde{x}_{n+1} \tilde{x}_{n+1}^T V_{(n)}} \tilde{x}_{n+1} y_{n+1}\Big)$$

$$= \hat{\beta}_{(n)} + \underbrace{h V_{(n)} \tilde{x}_{n+1}}_{K_n} \underbrace{\left(y_{n+1} - \tilde{x}_{n+1}^T \hat{\beta}_{(n)}\right)}_{e_{n+1}}$$

$\to$ prediction error of y_{n+1} based on the estimate obtained on the previous step n ($\hat{\beta}_{(n)}$)

$$= \hat{\beta}_{(n)} + K_n e_{n+1}$$

$\hat{\beta}_{(n+1)}$ and $V_{(n+1)}$ $\to$ $\beta_{(n+2)}$ and $V_{(n+2)}$ $\to$ ...