TMA4267 Linear statistical models

Part 2: Linear regression

February 20, 2017

Normal equations

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } E(\boldsymbol{\varepsilon}) = \boldsymbol{0} \text{ and } Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$$

Which of the following are *the normal equations*?

A $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$

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

C $(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$

D $(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{Y} = \boldsymbol{X}^T\hat{\boldsymbol{\beta}}$

The hat matrix

Design matrix $\boldsymbol{X}$ has n rows and p linearly independent columns. $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called the hat-matrix.

Which of the following statements are NOT true?

A $\boldsymbol{H} = \boldsymbol{H}^T = \boldsymbol{H}^2$

B $\text{rank}(\boldsymbol{H}) = p$

C $\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}$

D $\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$

Estimator for σ^2

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } E(\boldsymbol{\varepsilon}) = \boldsymbol{0} \text{ and } Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I} \\ \boldsymbol{H} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \end{aligned}$$

An unbiased estimator for σ^2 is:

A SSE/n

B $\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}/(n-p)$

C $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Y}/(n-p)$

D $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\text{SSE}/n$

Inference about β

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \text{ where } \varepsilon \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

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

What are the properties of $\hat{\beta}$?

A Chi-squared distributed with $n - p$ degrees of freedom.

B Chi-squared distributed with p degrees of freedom.

C Multivariate normal with covariance matrix $(\boldsymbol{I} - \boldsymbol{H})\sigma^2$.

D Multivariate normal with covariance matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$.

Happiness=money+sex+love+work

	Estimate	Std. Error	t value	Pr(>|t|)
money	0.009578	0.005213	1.837	0.0749
sex	-0.149008	0.418525	-0.356	0.7240
love	1.919279	0.295451	6.496	1.97e-07
work	0.476079	0.199389	2.388	0.0227

Which of the regression coefficient estimates has the largest estimated variance?

A money

B sex

C love

D work

Happiness=money+sex+love+work

The R^2 for the happiness-regression model is 71%. What does that mean?

A The regression is significant for significance level 71%

B The regression explains 71% of the variability in the data

C The estimate for the variance σ^2 is 0.71

D The covariates have a correlation of 0.71

Happiness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.072081	0.852543	-0.085	0.9331
money	0.009578	0.005213	1.837	0.0749
sex	-0.149008	0.418525	-0.356	0.7240
love	1.919279	0.295451	6.496	1.97e-07
work	0.476079	0.199389	2.388	0.0227

For which β_j would we reject the null hypothesis $\beta_j = 0$ at significance level 1%?

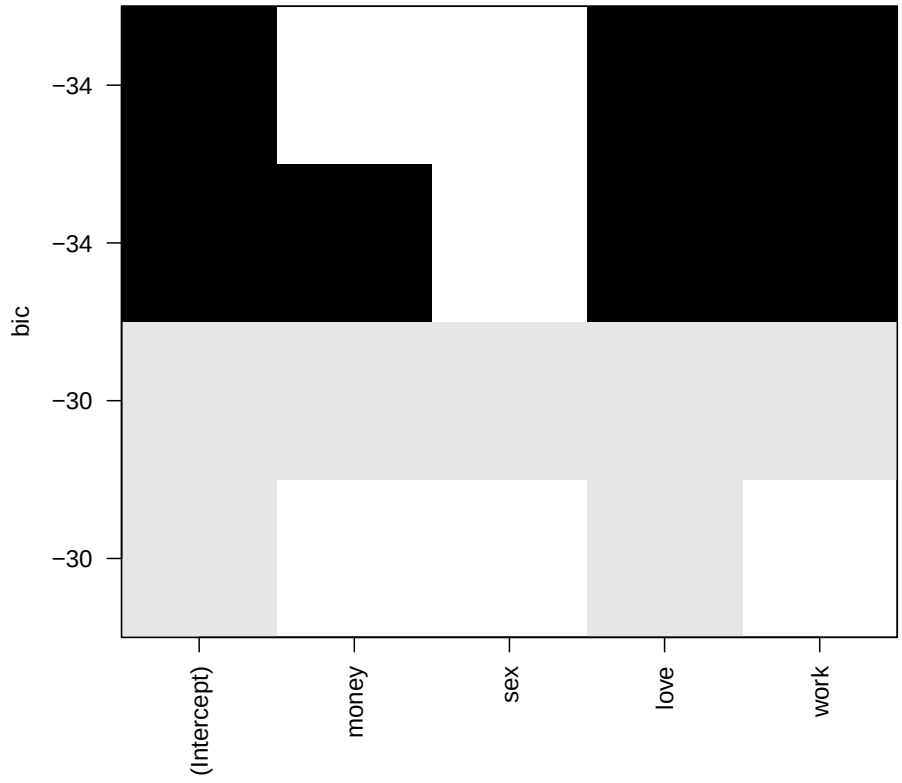
A money

B sex

C love

D work

Best model



Which model does the BIC criterion report to be the best?

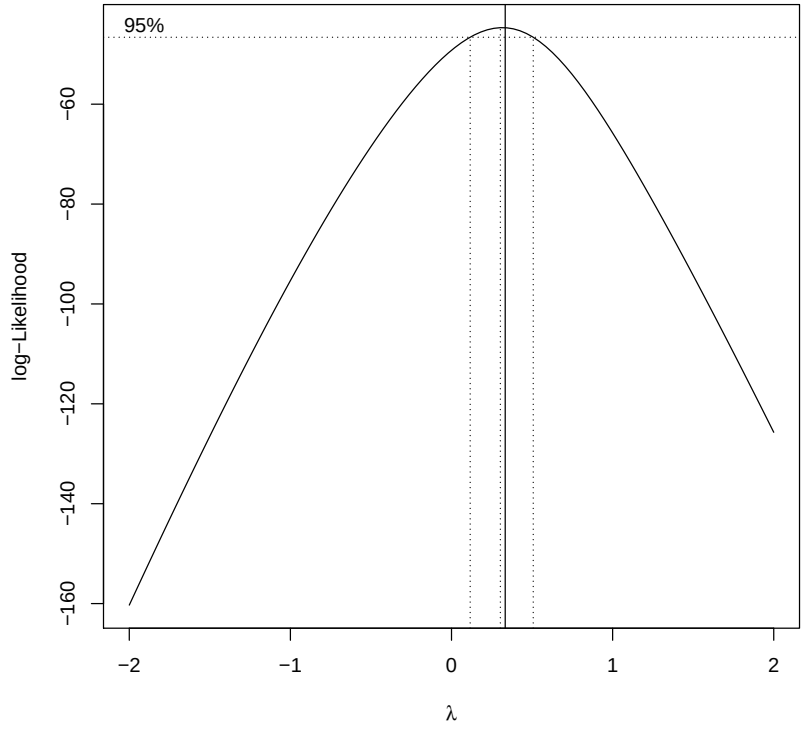
A love+work

B love

C money+love+work

D money+sex+love+work

What is this plot used for?



A Check residuals

B Assess normality of residuals

C Assess linearity

D Find transform of response

Correct?

Are you sure you want to read the correct answers? Maybe try first? The answers are explained on the next two slides.

Answers

1. C: The normal equation $(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$ is before you solve for $\hat{\boldsymbol{\beta}}$.
2. C: The hat matrix is symmetric and idempotent (so A is ok), and has rank p, but the reason for the name of the hat matrix is that is puts the hat on the $\boldsymbol{Y}$ so $\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{Y}}$. We know that for symmetric projection matrices the two matrices $\boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})$ are orthogonal so the product must be zero.

Answers

3. B: Since SSE has mean $(n-p)\sigma^2$, then SSE/(n-p) must be an unbiased estimator for σ^2. We know that $(\boldsymbol{I}-\boldsymbol{H})$ projects onto the space othogonal to the column space of the designmatrix, so that must have to do with SSE.
4. D: We know that linear combinations of multivariate normal random vectors are also multivariate normal (so the chisquare is not suitable). The residuals have $(\boldsymbol{I}-\boldsymbol{H})$ as part of their covariance matrix, but $\hat{\boldsymbol{\beta}}$ has not.

Answers

5. B: Sex has the largest estimated variance for regression estimate.
6. B: R^2 gives the percent of variability explained.
7. C: only love is significant on level 1%, since this is the only p-value below 0.01 (last column).
8. A: love+work has smallest BIC.
9. D: Box-Cox plot used to find transformation of response.