Midterm Cheat Sheet

Linear Regression

$y = \beta_0 + \beta_1 x + \varepsilon;$

$\varepsilon \sim N(0,\sigma^2) \Rightarrow y \sim N(\beta_0 + \beta_1 x; \sigma^2)$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}} = r_{XY}\sqrt{\frac{S_{YY}}{S_{XX}}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 \pm t_{\text{(crit, two-tailed}\,\alpha\text{, n-2 df)}} \times se(\hat{\beta}_1)$$

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} E(y_i) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(\beta_0 + \beta_1 x_i) = \beta_1;$$

$\mathrm{E}(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = \beta_0$

$E(\hat{\sigma}^2) = \sigma^2$

$\widehat{\sigma}^2$ = RSS/(n-2) = $\sum(y_i - \hat{y}_i)^2$ /(n − 2)

$$Var(\widehat{\beta_1}|\mathrm{X}) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}$$

$$Var(\widehat{\beta_0}|\mathrm{X}) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x - x_i)^2}\right)$$

$$se(\beta_0) = \sqrt{\frac{(y_i - \bar{y})^2}{n/2} * \frac{(\frac{1}{n} + \bar{x}^2)}{(\sum(x_i - (\bar{x}))^2)}}$$

$$se(\beta_1) = \sqrt{((1/(n-2)) * (\frac{((\sum(y_i - (\bar{y}))^2))}{\sum(x_i - (\bar{x}))^2}))}$$

$$se(y) = \hat{\sigma} * \sqrt{(\frac{1}{n} + \frac{(x - (\bar{x}))^2}{S_{xx}})}$$

Predicting a CI n

$$\sigma^2(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}})$$

□□0 + □□2x +/- t*

Multiple Linear Regression and Estimation

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$\boldsymbol{X}_{n\times(p+1)} = \begin{pmatrix} x_1{}' \\ \vdots \\ x_n{}' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$\hat{\beta}_k \pm t(1 - \alpha/2; \text{n-p-1}) * SE(\hat{\beta}_k)$$

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

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

$$\hat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

$Var(\hat{\boldsymbol{e}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$ if $Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$

$$s^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

$$= \frac{1}{n-p-1}(\mathbf{Y}\text{-}\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y}\text{-}\mathbf{X}\widehat{\boldsymbol{\beta}}) = \frac{RSS}{n-p-1}$$

$$R^2 = 1 - RSS/SYY = \frac{Y^I Y - \beta^I X^I Y}{Y^I Y - n\bar{Y}^2}$$

$$R_a^2 = 1 - \frac{RSS/(n-p-1)}{SYY/(n-1)} = \frac{(n-1)R^2 - p}{n-p-1}$$

□□0 : □□1 = □□2 = □□3 = ⋯ = □□□□ = 0

v.s. □□1 : not al' ... , □□

$$F^* = \frac{MSReg}{MSE} \quad t^* = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$$

rejection rule of □□ >= t(1 − alpha/2; □□ − □□ − 1)

$SS_{err} = RSS = Y^I Y - \beta^I X^I Y$

$SS_{yy} = Y^I Y - n\bar{Y}^2$

$SS_{reg} = \beta^I X^I Y - n\bar{Y}^2$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathrm{e}'\mathrm{e} = (\mathrm{Y} - \mathrm{X}\hat{\beta})'(\mathrm{Y} - \mathrm{X}\hat{\beta})$$

$$= \mathrm{Y}'\mathrm{Y} - \hat{\beta}'X'Y = Y'(I - H)Y$$

Model Fitting: Inference

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

df_{Ω} = n - p, and df_{ω} = n – q

$$F = \frac{(RSS_{\omega} - RSS_{\Omega})/(df_{\omega} - df_{\Omega})}{RSS_{\Omega}/df_{\Omega}} = \frac{SSReg/df_{Reg}}{\widehat{\sigma^2}}$$

Reject the null hypothesis if F > Fα p - q, n – p

$$t_i = \hat{\beta}_i / \mathrm{se}(\hat{\beta}_i)$$

$$RSS_{AH} = \sum(y - X\hat{\beta})'(y - X\hat{\beta}) = e'e$$

$$RSS_{NH} = \sum(y - \bar{y})'(y - \bar{y}) = SYY$$

$$F = \frac{(SYY - RSS)/((n-1)-(n-p-1))}{RSS/(n-p-1)} = \frac{(SYY - RSS)/p}{RSS/(n-p-1)}$$

$$F \sim F_{p,n-p-1}$$

Dummy Variables and Analysis of Covariance

Consider a Xi2 for which is 0 for – and 1 for +:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

An interaction between Xi1 and Xi2:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

A model with multiple categorical variables:

Tool Model	X_0	X_1	X_2	X_3	X_4
M1	1	X_{i1}	1	0	0
M2	1	X_{i1}	0	1	0
M3	1	X_{i1}	0	0	1
M4	1	X_{i1}	0	0	0

Model 4: $E[Y] = \beta_0 + \beta_1 X_1$

Model 1: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2$

Model 2: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_3$

Model 3: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_4$

Regression Diagnostics

Assumptions:

- Error: ~ N(0, SD2I);
  - Independent
  - Equal Variance
  - Normally Distributed
- Model: E[y] = Xβ is correct
- Unusual observations

Leverage Points: data point with unusual x-value

- ✓ $h_{ij} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_j = \boldsymbol{x}_j'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i = h_{ji}$
- ✓ $\sum_{i=1}^{n} h_{ii} = p' = \textit{number of parameters}$
- ✓ $\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$ if an intercept is included
- ✓ $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j$

The Hat Matrix – n*n matrix

h_{ii} is the leverage of the i^{th} case

leverage > 2p'/n should be looked at closely

Outliers: Unusual observation on x or y axis

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = r_i \left(\frac{n - p' - 1}{n - p' - r_i^2} \right)^{1/2} \sim t(n - p' - 1)$$

Calculate the t-test and compare abs with limit:

abs(qt(.05/(n*2), df = n - pprime - 1, lower.tail = T))

Influential Points: causes changes to regression

Difference in Fits:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}^2{}_{(i)} h_{ii}}}$$

with a threshold of

$$2 * \sqrt{\frac{p'+1}{n-p'-1}},$$

Where p' is the number of parameters

Cook's Distance:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta}_i)'(X'X)(\hat{\beta}_{(i)} - \hat{\beta}_i)}{p'\hat{\sigma}^2}$$

$$= \frac{(\hat{Y}_{(i)} - \hat{Y}_i)'(\hat{Y}_{(i)} - \hat{Y}_i)}{p'\hat{\sigma}^2} = \frac{1}{p'} r_i^2 \frac{h_{ii}}{1 - h_{ii}} \sim F(p', n - p')$$

with a threshold of

Di > 4/n should be looked at

Di > .5 possible influence

Di >= 1 very influential

Error: a plot of e_hat should

- have constant variance
- have no clear pattern
- H0: residuals are normal

Shapiro-Wilk normality test

H0: Residuals are normally distributed

Bonferroni Correction: Divide alpha by n

Variable Selection

Backwards Elimination:

1. Start model with all the predictors
2. Remove the predictor with highest p-value greater than alpha
3. Refit the model
4. Remove the remaining least significant predictor provided its p-value is greater than alpha
5. Repeat 3 and 4 until all "non-significant" predictors are removed

Cutoff p significance can be 15-20% for testing

Forward Selection:

1. Start model with no predictors
2. For predictors not in the model, check the p-value if they are added to the model. We choose the one with lowest p-value less than alpha
3. Continue until no new predictors can be added

Stepwise regression: A combination of the two

Selection Criteria:

Akaike Information Criterion (AIC):

- -2 max log-likelihood + 2p'
- n*log(RSS/n) + 2p'

Bayes Information Criterion (BIC):

- -2 max log-likelihood + p' log(n)
- n*log(RSS/n) + log(n) * p'

Adjusted R2:

R^2 = 1 – RSS/SSY

$$R_a^2 = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{SSY}{n-1}} = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2) = 1 - \frac{\hat{\sigma}^2_{Model}}{\hat{\sigma}^2_{Null}}$$

Mallow's C_p Statistic: Avg MSE of prediction

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

If a p-predictor fits then:

$$E\left[RSS_p\right] = (n-p)\sigma^2 \text{ and } E(C_p) \approx p$$

We desire models with small p and Cp around or less than p

R Code Snippets

```
# Model with only beta_0
sr_lm0 <- lm(y ~ 1, data=sr)
# Full model
sr_lm1 <- lm(y ~ ., data=sr)
sr_syy <- sum((savings$sr -
mean(savings$sr))^2)
sr_rss <- deviance(sr_lm1)
# F = ((SYY -RSS)/((n-1) - (n-2))) /
(RSS / (n - 1))
sr_num <- (sr_syy -
sr_rss)/(df.residual(sr_lm0) -
df.residual(sr_lm1))
sr_den <- sr_rss / df.residual(sr_lm1)
sr_f <- sr_num / sr_den
# df? = n - p, and df? = n - q
pf(sr_f, df.residual(sr_lm0) -
df.residual(sr_lm1),
df.residual(sr_lm1), lower.tail = F)

# ?=(XI X)?1 XIY
beta <- solve(t(x)%*%x)%*%(t(x)%*%y)
# Pearson's
cor(lin_reg$fitted.values,
lin_reg$residuals, method="pearson")

# Stratify variables by a factor
by(depress, depress$publicassist,
summary)
# Welsh's Two Sample T-test
# For difference in means
t.test(assist$cesd, noassist$cesd) # or
t.test(data.y ~ factor)
# CI of LS means based on covariates
library(lsmeans)
lsmeans(reg, ~Type)
# Apply a mean function to an array
# split on a factor
tapply(assist$cesd, assist$assist,
mean)
# When a regression factor has
# more than two categories
reg <- lm(Pulse1 ~ Height + Sex +
Smokes + as.factor(Exercise))
```

```
# Cook's Distance
cook <- cooks.distance(reg)
cook[cook > 4/n]
# Shapiro Test for normallity
shapiro.test(reg$residuals)
# Studentized residuals
stud <- rstudent(reg)
# Threshold for lower tail of
# studentized resids with correction
lim = abs(qt(.05/(n*2), df = n - pprime
- 1, lower.tail = T))
stud[which(abs(stud) > lim)]
# Hat values
hat <- hatvalues(reg)
lev <- 2 * pprime / n
hat[hat > lev]

# Forward selection
forward <- ~ year + unemployed + femlab
+ marriage + birth + military
m0 <- lm(divorce ~ 1, data = usa)
reg.forward.AIC <- step(m0, scope =
forward, direction = "forward", k = 2)
n <- nrow(usa)

# AIC = n*log(RSS/n) + 2p'
n*log(162.1228/n)+2*6
extractAIC(reg.forward.AIC, k=2)
# BIC
reg.forward.BIC <- step(m0, scope =
forward, direction = "forward", k =
log(n))
extractAIC(reg.forward,k=log(n))
# BIC = n*log(RSS/n) + p'*log*n)
n*log(162.1228/n)+6*log(n)

library(leaps)
leaps <- regsubsets(divorce ~ .)
rs <- summary(leaps)
par(mfrow=c(1,2))
plot(2:7, rs$cp, xlab="No. of
parameters", ylab="Cp Statistic")
abline(0,1)
```

Source of Variation	Sum of Square	DF	Mean of Square
Regression (model)	SSreg	p	MSReg= SSreg/p
Error	RSS	n - (p+1)	MSE= RSS/(n - p - 1)

ANOVA For Simple Linear Regression

TABLE 2.3 The Analysis of Variance Table for Simple Regression

Source	df	SS	MS	F	p-value
Regression	1	$SSreg$	$SSreg/1$	$MSreg/\hat{\sigma}^2$	
Residual	$n-2$	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	SYY			

Global Null Hypothesis

$\boldsymbol{H_0}$: **Model fit not significant** $(\boldsymbol{SS_{reg} = 0})$; **or** $\boldsymbol{E(Y) = \beta_0}$

$\boldsymbol{H_1}$: **Model fit significant** $(\boldsymbol{SS_{reg} > 0})$;

which is equivalent to $\boldsymbol{H_0: \beta_1 = 0}$

$\boldsymbol{H_1: \beta_1 \neq 0}$

$$\frac{SS_{reg}}{\widehat{\sigma^2}} = \left(\frac{\widehat{\beta_1}}{se(\widehat{\beta_1})}\right)^2$$

TABLE 2.1 Definitions of Symbols[a]

Quantity	Definition	Description
$\overline{x}$	$\sum x_i/n$	Sample average of x
$\overline{y}$	$\sum y_i/n$	Sample average of y
SXX	$\sum(x_i - \overline{x})^2 = \sum(x_i - \overline{x})x_i$	Sum of squares for the x's
SD_x^2	$SXX/(n-1)$	Sample variance of the x's
SD_x	$\sqrt{SXX/(n-1)}$	Sample standard deviation of the x's
SYY	$\sum(y_i - \overline{y})^2 = \sum(y_i - \overline{y})y_i$	Sum of squares for the y's
SD_y^2	$SYY/(n-1)$	Sample variance of the y's
SD_y	$\sqrt{SYY/(n-1)}$	Sample standard deviation of the y's
SXY	$\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum(x_i - \overline{x})y_i$	Sum of cross-products
s_{xy}	$SXY/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/(\mathrm{SD}_x \mathrm{SD}_y)$	Sample correlation

[a] In each equation, the symbol $\sum$ means to add over all the n values or pairs of values in the data.

Revision #14
Created 18 October 2022 14:42:08 by Elkip
Updated 20 October 2022 18:32:02 by Elkip