Tutorial 8

YANG YANG

The Australian National University

Week 8, 2017

Overview

- Download **Tutorial 3.pdf** and **teengamb.csv** from Wattle. Read in data and attach
  ⟶ `teengamb <- read.csv("teengamb.csv",header=T)`
  ⟶ `attach(teengamb)`
- Transform **gamble** and construct histograms of **gamble** and **trans.gamble**
  ⟶ `trans.gamble <- log(gamble + 1)`
  ⟶ `par(mfrow=c(1,2))` # *plot two graphs together*

  ```
  hist(gamble)
  hist(trans.gamble)
  ```

Added variable plot

- The residuals from regressing Y against all predictors other than $X_{interested}$ go on the vertical axis, while the residuals from regression $X_{interested}$ against all other predictors go on the horizontal axis.
- Since the mean residual from both of these regressions is zero, the mean point of ($X_{interested}$ given others, Y given others) will just be $(0, 0)$ which explains why the regression line in the added variable plot always goes through the origin.

Added variable plot

```
####################################added variable plots

# first, regress trans.gamble against all predictors other than income
# ---> residuals that go on the vertical axis of the plot
gamble.baselm <- lm(trans.gamble ~ sex + verbal + status)

# second, regress income against all other predictors
# ---> residuals that go on the horizontal axis of the plot
income.lm <- lm(income ~ sex + verbal + status)

# fit MLR with the interested predictor as the last independent variable
# ---> find slope of the partial regression line
gamble.fulllm <- lm(trans.gamble ~ sex + verbal + status + income)

# added variable plot
plot(residuals(income.lm), residuals(gamble.baselm),
     xlab="Residuals(income on sex, verbal, status)",
     ylab="Residuals (trans.gamble on sex, verbal, status)")
abline(0, coef(gamble.fulllm)[5])
title("Added variable plot for income")
```

Standardised residuals

- Residuals do not behave like the true errors: residuals do not all have the same variability, since $Var(e_i) = \sigma^2(1 - h_{ii})$, and the leverage values, h_{ii} are typically different for each data point.
- $Var(\frac{e_i}{\sigma\sqrt{1-h_{ii}}}) = \frac{Var(e_i)}{\sigma^2(1-h_{ii})} = \frac{\sigma^2(1-h_{ii})}{\sigma^2(1-h_{ii})} = 1$
- We don't know the true σ^2, need to use some estimated values.

Standardised residuals

- Internally studentised residuals:
  $r_i = \frac{e_i}{s_\epsilon\sqrt{1-h_{ii}}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$
- Externally studentised residuals:
  $t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$

Standardised residuals

Externally studentised residuals:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$$

- In the above equation, $e_{i,-i} = Y_i - \hat{Y}_{i,-i} = \frac{e_i}{1-h_{ii}}$ with $\hat{Y}_{i,-i}$ the predicted value at x_i from a regression with the i^{th} point removed.
- $e_{i,-i}$ measures how far the i^{th} response is from a prediction over which it has no influence.
- s_{-i} is the residual scale from the regression calculated without the i^{th} data point.

Diagnostic plots

```
plot(fitted(gamble.lm), rstudent(gamble.lm))
title("Externally studentised residuals vs fitted values",
      sub="lm(trans.gamble ~ sex + verbal + status + log_income)")
abline(0,0)
identify(fitted(gamble.lm), rstudent(gamble.lm))

plot(gamble.lm, which=2)

plot(gamble.lm, which=4)
```

In R, standardised residuals are internally studentised residuals. Need to use "**rstudent(model.name)**" to get externally studentised residuals when required.

Influence Statistics

- Get leverage using "**hatvalues(model.name)**"
- Get $DFFITS_i$ using "**dffits(model.name)**"
- Get $DEBETAS_i$ using "**dfbetas(model.name)**"
- Get $COVRATIO_i$ using "**covratio(model.name)**"

```
# influence statistics
sort(hatvalues(gamble.lm), decreasing=TRUE)[1:7]

dfbetas(gamble.lm)[c(5,6,23),]
dffits(gamble.lm)[c(5,6,23)]
covratio(gamble.lm)[c(5,6,23)]
```

Influence Statistics

- $DFFITS_i$: removal of the i^{th} data point affects the associated fitted value for this point $\longrightarrow |DFFITS_i| > 2\sqrt{p/n}$
- $DEBETAS_i$: each data point's influence on the estimated parameters $\longrightarrow |DEBETAS_i| > 2/\sqrt{n}$
- $COVRATIO_i$: the i^{th} data point influence overall performance of the model $\longrightarrow COVRATIO_i > 1 + 3p/n$ or $COVRATIO_i < 1 - 3p/n$

F-test and T-test

```
#overall F-test
anova(lm(trans.gamble ~cbind(sex, verbal, status, log_income)))
#sequential F test
anova(gamble.lm)
#T-tests
summary(gamble.lm)
```

- Solution does not include interpretation of the overall F test. Better provide answers to each part of your assignment question.
- T-tests are marginal tests $\longrightarrow$ p-values are the same even if we change the order of predictors.
- Interpret estimated coefficients.

Plot back-transformed models

```
# find the range of income
range(income)
# create a vector that covers the full range of income
incomes <- 1:160/10
incomes
log_incomes <- log(incomes)
log_incomes

newfemales <- data.frame(sex=1,verbal=mean(verbal),status=mean(status),log_income=log_incomes)
newfemales
newfemales.preds <- predict(gamble.lm, newdata=newfemales, interval="confidence")
newfemales.preds
newfemales.backtranspreds <- exp(newfemales.preds)-1
newfemales.backtranspreds
```

- Similar to commands used in Assignment 1
- Control values of **sex** (=1 **OR** =0), holding **verbal** and **status** (=mean) when making predictions
- Interval = “confidence”