Advanced Data Analysis

Haoyang Chen | hc2812 | Assignment 3

1. Consider the Pima.te dataset

a). Fit a multiple linear regression model:

The model: glu = 56.9314 – 0.8753npreg + 0.1039bp + 0.2626skin + 0.7958bmi + 0.7638age

> summary(LinearModel)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age, data = Pima.te)

Residuals:

Min 1Q Median 3Q Max

-61.285 -20.556 -4.356 17.370 76.509

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 56.8314 10.3090 5.513 7.19e-08 \*\*\*

npreg -0.8753 0.6475 -1.352 0.17735

bp 0.1039 0.1385 0.750 0.45353

skin 0.2626 0.2164 1.214 0.22575

bmi 0.7958 0.3020 2.636 0.00880 \*\*

age 0.7638 0.2068 3.693 0.00026 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

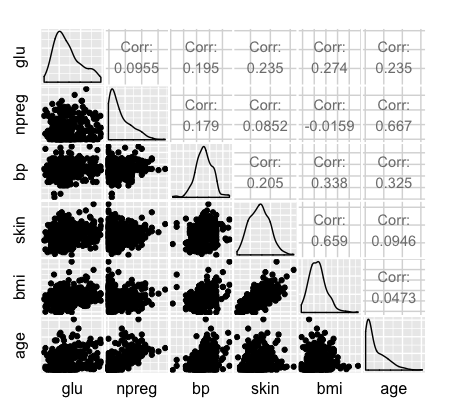
Residual standard error: 28.6 on 326 degrees of freedom

Multiple R-squared: 0.1338, Adjusted R-squared: 0.1205

F-statistic: 10.07 on 5 and 326 DF, p-value: 5.575e-09

b). State and assess the validity of the underlying assumptions:

* Linearity/functional form, including the need for any interaction terms:



From the scatter plot, we find that bp, skin, bmi, and age have linear relationship with glu, while npreg do not have linear relationship with glu. Thus the linearity form is not appropriate.

R-square is 0.1338, not very good.

For the interaction terms:

First, add all interactions into model:

> summary(LinearModelwithAllInteractions)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age + npreg \* bp +

npreg \* skin + npreg \* bmi + npreg \* age + bp \* skin + bp \*

bmi + bp \* age + skin \* bmi + skin \* age + bmi \* age, data = Pima.te)

Residuals:

Min 1Q Median 3Q Max

-63.424 -19.930 -4.356 19.575 75.418

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.679e+01 4.855e+01 -0.346 0.7298

npreg 2.116e+00 5.318e+00 0.398 0.6909

bp 7.623e-01 6.614e-01 1.153 0.2500

skin 1.504e+00 1.384e+00 1.087 0.2780

bmi 5.571e-01 1.702e+00 0.327 0.7437

age 2.924e+00 1.631e+00 1.793 0.0740 .

npreg:bp -2.026e-02 5.869e-02 -0.345 0.7302

npreg:skin -1.323e-02 9.023e-02 -0.147 0.8835

npreg:bmi -2.978e-02 1.130e-01 -0.263 0.7924

npreg:age 7.744e-04 5.998e-02 0.013 0.9897

bp:skin 8.502e-03 1.829e-02 0.465 0.6424

bp:bmi -3.677e-03 1.813e-02 -0.203 0.8394

bp:age -2.441e-02 1.714e-02 -1.424 0.1555

skin:bmi -1.341e-02 2.034e-02 -0.659 0.5102

skin:age -4.546e-02 2.583e-02 -1.760 0.0793 .

bmi:age 3.097e-02 3.639e-02 0.851 0.3954

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.58 on 316 degrees of freedom

Multiple R-squared: 0.1615, Adjusted R-squared: 0.1217

F-statistic: 4.058 on 15 and 316 DF, p-value: 8.889e-07

We find that bp\*age and skin\*age may have interactions, thus we construct a model with these two interactions.

> summary(LinearModelwithTwoInteractions)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age + bp \* age +

skin \* age, data = Pima.te)

Residuals:

Min 1Q Median 3Q Max

-61.52 -20.00 -4.30 18.24 75.77

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -24.73944 32.36098 -0.764 0.445136

npreg -0.79329 0.64643 -1.227 0.220643

bp 0.80056 0.40276 1.988 0.047690 \*

skin 1.33841 0.53688 2.493 0.013169 \*

bmi 0.74008 0.30027 2.465 0.014231 \*

age 3.61883 1.08370 3.339 0.000938 \*\*\*

bp:age -0.02360 0.01243 -1.899 0.058449 .

skin:age -0.03636 0.01583 -2.296 0.022293 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.31 on 324 degrees of freedom

Multiple R-squared: 0.1567, Adjusted R-squared: 0.1385

F-statistic: 8.604 on 7 and 324 DF, p-value: 1.104e-09

We find that only skin\*age have significantly difference. Thus we construct a model only contain one interaction.

> summary(LinearModelwithOneInteractions)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age + skin \* age,

data = Pima.te)

Residuals:

Min 1Q Median 3Q Max

-64.849 -20.820 -4.357 17.453 75.701

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 28.66509 16.07608 1.783 0.07550 .

npreg -0.70926 0.64750 -1.095 0.27416

bp 0.08158 0.13795 0.591 0.55466

skin 1.38493 0.53847 2.572 0.01056 \*

bmi 0.73039 0.30143 2.423 0.01593 \*

age 1.78530 0.49410 3.613 0.00035 \*\*\*

skin:age -0.03614 0.01590 -2.273 0.02366 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.42 on 325 degrees of freedom

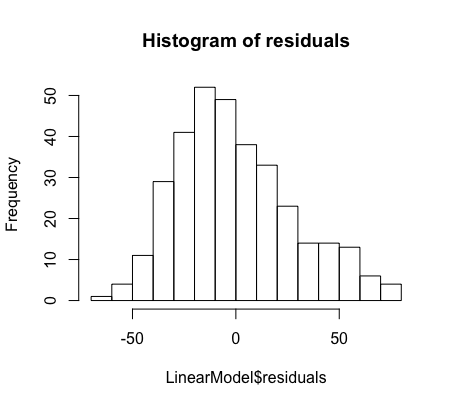
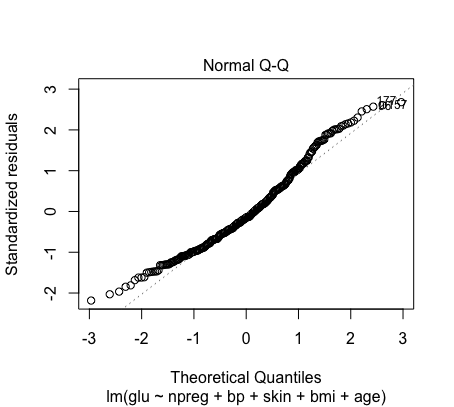
Multiple R-squared: 0.1474, Adjusted R-squared: 0.1316

F-statistic: 9.362 on 6 and 325 DF, p-value: 1.76e-09

Thus skin\*age is an important interaction to the glu.

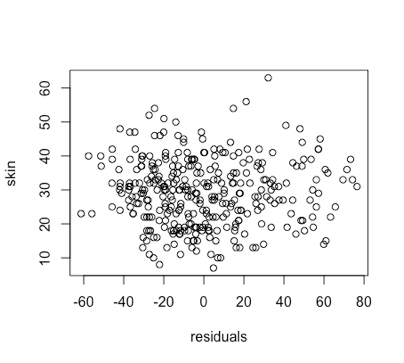
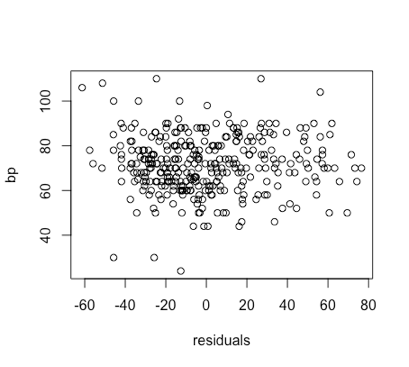
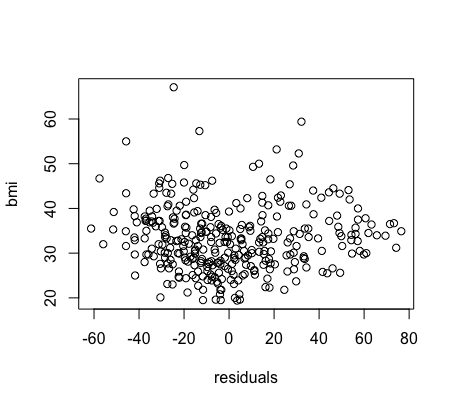
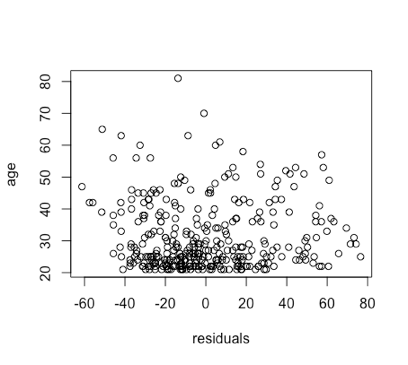
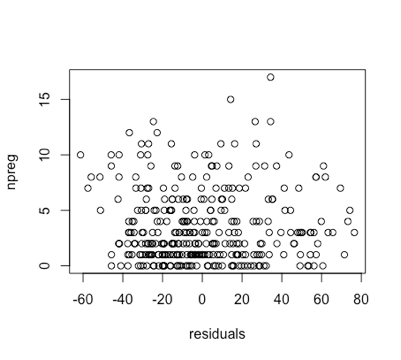
* Normality:

From the Q-Q plot is nearly a line, and the histogram is almost a normal distribution, thus the normality assumption is validated.



* Homoscedasticity:

The residuals vs predictors plots indicate that the variance is constant.



* Uncorrelated error:

The Durbin-Watson Test have a p-value = 0.558, which indicates that it rejects the null hypothesis and the observations are uncorrelated.

> durbinWatsonTest(LinearModel)

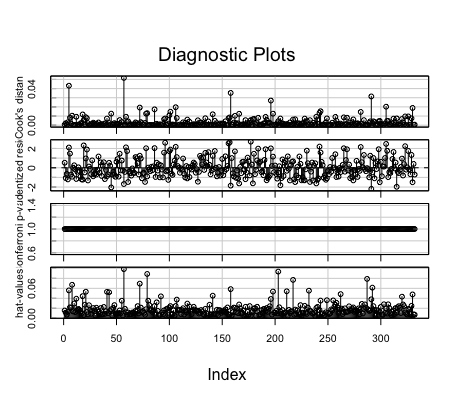
lag Autocorrelation D-W Statistic p-value

1 0.029888 1.937881 0.558

Alternative hypothesis: rho != 0

* Check for outliers and influential points:

Influential points can be detected by Cook’s distance. According to the plot below, there are 2 influential points: the first one is an outlier and a high leverage point, and the second one is a high leverage point.



c). Propose remedial measures in case of violations of any of the underlying assumptions

* Linearity:
* Transformations, basis functions: such as log, smoothing splines
* Non-linear models
* Other predictors
* Normality:
* Transformations, basis functions: such as log, smoothing splines
* Robust regression methods
* Homoscedasticity:
* Transformations, basis functions: such as log, smoothing splines
* Build variance structure into model: weighted least squares
* Uncorrelated error:
* Transformation: Cochrane-Orcutt Procedure
* Use models that incorporate the correlation structure: Generalized Estimating Equations
* Outliers and Influential Points:
* Delete the outliers
* Use robust regression methods: lease median of squares regression

d). Compare the Lease Median of Squares Regression and Linear Regression models:

The coefficients between linear regression and lease median of squares regression are quite different. This method is more robust for outliers and influential points. Compare to the result in a, the model has a higher breakdown point.

> coef(LeastMedianModel)

(Intercept) npreg bp skin bmi

47.7888316 1.9123228 0.3207494 1.2400925 -0.3783237

age

0.2133375

> coef(LinearModel)

(Intercept) npreg bp skin bmi

56.8313661 -0.8753016 0.1039174 0.2626200 0.7958464

age

0.7638058

Code:

library('MASS')

data("Pima.te")

Pima.te <- Pima.te[c('glu', 'npreg', 'bp', 'skin', 'bmi', 'age')]

# a) Fit a multiple linear regression model

LinearModel <- lm(glu ~ npreg + bp + skin + bmi + age, data = Pima.te)

summary(LinearModel)

# b) State and assess the validity of the underlying assumptions

# Linearity

library(GGally)

ggpairs(Pima.te)

# interaction

LinearModelwithAllInteractions = lm(glu ~ npreg + bp + skin + bmi + age + npreg\*bp + npreg\*skin + npreg\*bmi +

npreg\*age + bp\*skin + bp\*bmi+ bp\*age + skin\*bmi + skin\*age + bmi\*age, data = Pima.te)

summary(LinearModelwithAllInteractions)

LinearModelwithTwoInteractions = lm(glu ~ npreg + bp + skin + bmi + age + bp\*age + skin\*age, data = Pima.te)

summary(LinearModelwithTwoInteractions)

LinearModelwithOneInteractions = lm(glu ~ npreg + bp + skin + bmi + age + skin\*age, data = Pima.te)

summary(LinearModelwithOneInteractions)

# Non-normality

hist(LinearModel$residuals, main = 'Histogram of residuals')

qqnorm(LinearModel$residuals)

# Homoscedasticity

plot(LinearModel$residuals, Pima.te$npreg, xlab = 'residuals', ylab = 'npreg')

plot(LinearModel$residuals, Pima.te$bp, xlab = 'residuals', ylab = 'bp')

plot(LinearModel$residuals, Pima.te$skin, xlab = 'residuals', ylab = 'skin')

plot(LinearModel$residuals, Pima.te$bmi, xlab = 'residuals', ylab = 'bmi')

plot(LinearModel$residuals, Pima.te$age, xlab = 'residuals', ylab = 'age')

# Uncorrelated error

library(car)

durbinWatsonTest(LinearModel)

# Check for outliers and influential points

infIndexPlot(LinearModel)

# c Least Median of Squares Regression

LeastMedianModel <- lmsreg(glu ~ npreg + bp + skin + bmi + age, data = Pima.te)

summary(LeastMedianModel)

coef(LeastMedianModel)

coef(LinearModel)