STA135 homework9

Zhen Zhang

March 2, 2016

7.1

```
z <- c(10, 5, 7, 19, 11, 8)
y <- c(15, 9, 3, 25, 7, 13)
model1 <- lm(y ~ z)
# beta hat
model1$coefficients
```

```
## (Intercept)           z
##  -0.6666667   1.2666667
```

```
# fitted values
model1$fitted.values
```

```
##         1         2         3         4         5         6
## 12.000000  5.666667  8.200000 23.400000 13.266667  9.466667
```

```
# residuals
model1$residuals
```

```
##         1         2         3         4         5         6
##  3.000000  3.333333 -5.200000  1.600000 -6.266667  3.533333
```

```
# residual sum of squares
t(model1$residuals) %*% model1$residuals
```

```
##          [,1]
## [1,] 101.4667
```

7.9

```
z <- c(-2, -1, 0, 1, 2)
y <- matrix(c(5, 3, 4, 2, 1, -3, -1, -1, 2, 3), nrow = 5)
model2 <- lm(y ~ z)
# parameters
model2$coefficients
```

```
##             [,1]         [,2]
## (Intercept)  3.0 1.986027e-16
## z           -0.9 1.500000e+00
```

```
# fitted values
model2$fitted.values
```

```
##   [,1]          [,2]
## 1  4.8 -3.000000e+00
## 2  3.9 -1.500000e+00
## 3  3.0 -2.220446e-16
## 4  2.1  1.500000e+00
## 5  1.2  3.000000e+00
```

```
# residuals
model2$residuals
```

```
##   [,1]          [,2]
## 1  0.2  5.433906e-16
## 2 -0.9  5.000000e-01
## 3  1.0 -1.000000e+00
## 4 -0.1  5.000000e-01
## 5 -0.2 -5.412768e-17
```

```
# verification
(LRS = t(y) %*% y)
```

```
##      [,1] [,2]
## [1,]   55  -15
## [2,]  -15   24
```

```
(RHS = t(model2$fitted.values) %*% model2$fitted.values + t(model2$residuals) %*% model2$residuals)
```

```
##      [,1] [,2]
## [1,]   55  -15
## [2,]  -15   24
```

They are identical.

7.19

```
satellite <- read.table("T7-5.DAT")
names(satellite) <- c(paste0("Z", 1:5), "Y")
model3 <- lm(log(Y) ~ ., data = satellite)
summary(model3)
```

```
##
## Call:
## lm(formula = log(Y) ~ ., data = satellite)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0932 -0.6697  0.2702  0.7417  1.3986
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -63.68324   51.17800  -1.244 0.233802
## Z1           -0.45925    0.54926  -0.836 0.417124
## Z2           -0.32668    0.17615  -1.855 0.084834 .
## Z3           -0.01113    0.01699  -0.655 0.522932
## Z4            0.11577    0.02499   4.633 0.000387 ***
## Z5           33.81176   25.58798   1.321 0.207560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 14 degrees of freedom
## Multiple R-squared:  0.6633, Adjusted R-squared:  0.543
## F-statistic: 5.515 on 5 and 14 DF,  p-value: 0.005184
```

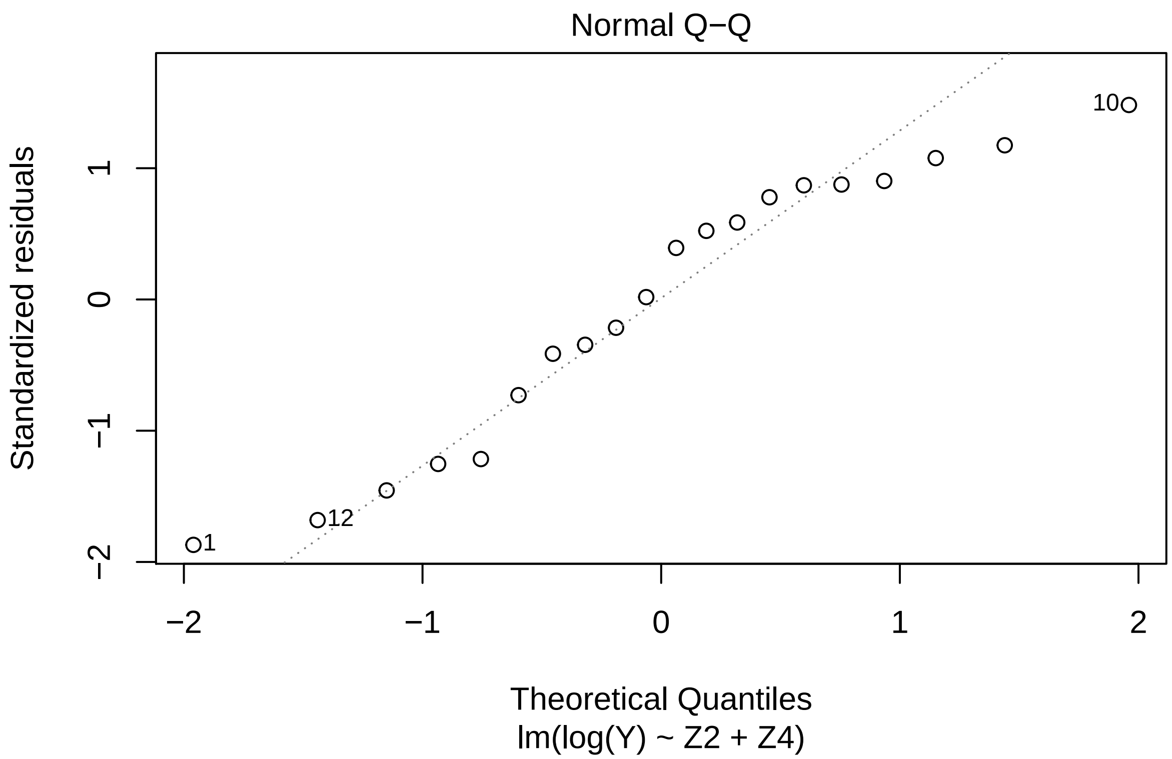
Based on the summary statistics, we should select Z_2 and Z_4.

```r
model4 <- lm(log(Y) ~ Z2 + Z4, data = satellite)
summary(model4)
```

```
##
## Call:
## lm(formula = log(Y) ~ Z2 + Z4, data = satellite)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6870 -0.8171  0.1999  0.8448  1.3938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.75648    0.74098   3.720 0.001702 **
## Z2          -0.32182    0.17043  -1.888 0.076180 .
## Z4           0.11382    0.02429   4.685 0.000213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 17 degrees of freedom
## Multiple R-squared:  0.6006, Adjusted R-squared:  0.5536
## F-statistic: 12.78 on 2 and 17 DF,  p-value: 0.000409
```

The model is $ln(Y) = 2.75648 - 0.32182 * Z_2 + 0.11382 * Z_4$

```
plot(model4, which = 2)
```



Although it is not perfectly normal distributed, due to its small sample, we should not count it as a problem.

8.3

```
sigma = diag(c(2, 4, 4))
```

The eigenvalues are 2, 4 and 4, and the eigenvectors are [1, 0, 0], [0, 1, 0] and [0, 0, 1] or [1, 0, 0], [0, 0, 1] and [0, 1, 0], since the eigenvalues are identical for the second and the third one.

8.11

```
census <- read.table("T8-5.DAT")
census$V5 <- 10 * census$V5
# covariance matrix
census_cov <- cov(census)
eigen(census_cov)
```

```
## $values
## [1] 108.271939  43.139674  31.267127   4.598098   2.347868
##
## $vectors
##              [,1]        [,2]        [,3]        [,4]        [,5]
## [1,]  0.03762881 -0.06230915 -0.03997936 -0.55553173  0.82733777
## [2,] -0.11892964 -0.24930105  0.26052476  0.76839232  0.51517455
## [3,]  0.47967291 -0.75967654 -0.43064872  0.02807896 -0.08098582
## [4,] -0.85891177 -0.31639989 -0.39364417 -0.06867379 -0.04989847
## [5,] -0.12893518 -0.50670427  0.76818907 -0.30895506 -0.20262977
```

```
# the first two component analysis
eigen(census_cov)$vectors[, 1:2]
```

```
##              [,1]        [,2]
## [1,]  0.03762881 -0.06230915
## [2,] -0.11892964 -0.24930105
## [3,]  0.47967291 -0.75967654
## [4,] -0.85891177 -0.31639989
## [5,] -0.12893518 -0.50670427
```

```
# total proportion explained
sum(eigen(census_cov)$values[1:2]) / sum(eigen(census_cov)$values)
```

```
## [1] 0.7984804
```

```
census_y1 <- t(eigen(census_cov)$vectors[, 1] %*% t(census))
census_y2 <- t(eigen(census_cov)$vectors[, 2] %*% t(census))
cor(cbind(census_y1, census_y2), census)
```

```
##              V1         V2         V3        V4         V5
## [1,]  0.2124404 -0.3978987  0.6692133 -0.946998 -0.2376782
## [2,] -0.2220491 -0.5264862 -0.6690038 -0.220200 -0.5895939
```

The correlation table indicates, the first component, is mostly dependent on the first variable, goverment employment, and second component are influenced by the second, third and fifth variable. Variable 5, median home value, has little impact in the first component, though large influence on the second component.