STAT 8051 HW 2

Jingxiang Li

September 27, 2014

Problem 3.2

Added-variable plots (Data file: `UN11`) This problem uses the United Nations example in Section 3.1 to demonstrate many of the properties of added-variable plots. This problem is based on the mean function `fertility` $\sim$ `log(ppgdp) + pctUrban`. There is nothing special about a two-predictor regression mean function, but we are using this case for simplicity.

Problem 3.2.1

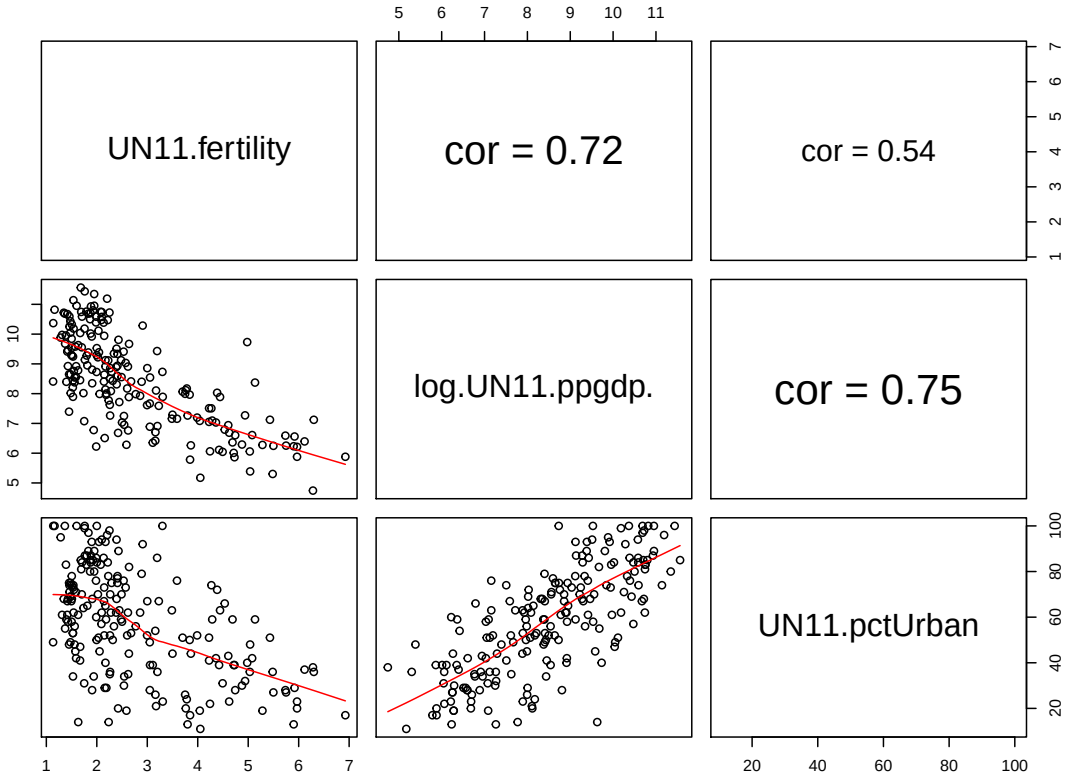
Examine the scatterplot matrix for (`fertility`, `log(ppgdp)`, `pctUrban`), and comment on the marginal relationships.

Solution

```
require(package = "alr4")

data <- data.frame(UN11$fertility, log(UN11$ppgdp), UN11$pctUrban)
panel.cor <- function(x, y, digits = 2, prefix = "cor = ", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(data, lower.panel = panel.smooth, upper.panel = panel.cor)
```



This is the Scatterplot Matrix for (`fertility`, `log(ppgdp)`, `pctUrban`). First, this graph shows decreasing tendencies of the response over two predictors respectively, suggesting that linear regression of `fertility` over these two predictors dose make sense. Plus, the correlation coefficients between response and two predictors also illustrate that there exists linear relationship among these variables. However, the bad news is that this graph also illustrates strong collinearity between two predictors, which means that the regression model we are going to establish may not give valid results about any individual predictor.

Problem 3.2.2

Fit the two simple regressions for `fertility` $\sim$ `log(ppgdp)` and for `fertility` $\sim$ `pctUrban`, and verify that the slope coefficients are significantly different from 0 at any conventional level of significance.

Solution

```
m1 <- lm(fertility ~ log(ppgdp), data = UN11)
m2 <- lm(fertility ~ pctUrban, data = UN11)
summary(m1)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp), data = UN11)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1631 -0.6451 -0.0659  0.6248  3.0052
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0097     0.3653    21.9   <2e-16 ***
## log(ppgdp)   -0.6201     0.0424   -14.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.93 on 197 degrees of freedom
## Multiple R-squared:  0.52,Adjusted R-squared:  0.518
## F-statistic:  213 on 1 and 197 DF,  p-value: <2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = fertility ~ pctUrban, data = UN11)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.493 -0.779 -0.147  0.652  2.903
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.55982    0.21368   21.34   <2e-16 ***
## pctUrban    -0.03105    0.00342   -9.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.13 on 197 degrees of freedom
## Multiple R-squared:  0.295,Adjusted R-squared:  0.291
## F-statistic: 82.4 on 1 and 197 DF,  p-value: <2e-16
```

m1 is the simple regression model for `fertility` $\sim$ `log(ppgdp)`, m2 is for `fertility` $\sim$ `pctUrban`. To see whether the slope coefficients are zero, we only need refer to the p-values of F-test given by two models. Note that the p-values given by two models are all $< 2e-16$, suggesting that both two slope coefficients are significantly different from 0, at any conventional level of significance.

Problem 3.2.3

Obtain the added-variable plots for both predictors. Based on the added-variable plots, is `log(ppgdp)` useful after adjusting for `pctUrban`, and similarly, is `pctUrban` useful after adjusting for `log(ppgdp)`? Compute the estimated mean function with both predictors included as regressors, and verify the findings of the added-variable plots.

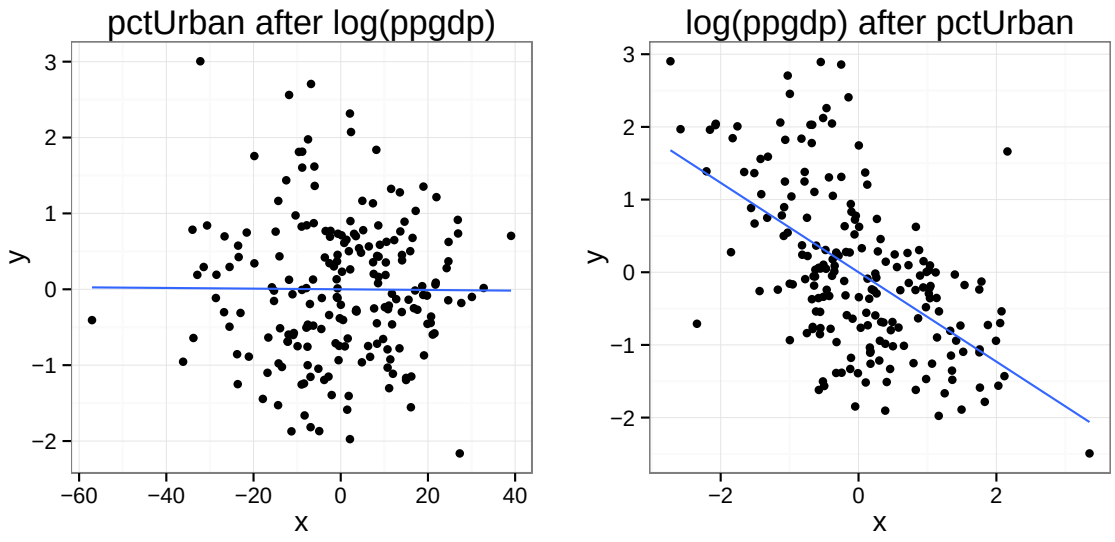
Solution

```
require(package = "ggplot2")
require(package = "gridExtra")

e_f2lp <- m1$residuals
e_f2pU <- m2$residuals
e_lp2pU <- lm(log(ppgdp) ~ pctUrban, data = UN11)$residuals
e_pU2lp <- lm(pctUrban ~ log(ppgdp), data = UN11)$residuals

p1 <- ggplot(data.frame(x = e_pU2lp, y = e_f2lp), aes(x = x, y = y))
p1 <- p1 + geom_point()
p1 <- p1 + geom_smooth(formula = y ~ x, method = "lm", se = FALSE)
p1 <- p1 + theme_bw() + theme(text = element_text(size = 14))
p1 <- p1 + ggtitle("pctUrban after log(ppgdp)")

p2 <- ggplot(data.frame(x = e_lp2pU, y = e_f2pU), aes(x = x, y = y))
p2 <- p2 + geom_point()
p2 <- p2 + geom_smooth(formula = y ~ x, method = "lm", se = FALSE)
p2 <- p2 + theme_bw() + theme(text = element_text(size = 14))
p2 <- p2 + ggtitle("log(ppgdp) after pctUrban")
grid.arrange(p1, p2, ncol=2)
```



Graph on the left side is the added-variable plot for `pctUrban` after `log(ppgdp)`, similarly, the one on the right side is the plot for `log(ppgdp)` after `pctUrban`. From these two graphs

we see that `log(ppgdp)` is still useful to explain the variance of response after adjusting for `pctUrban`. However, it seems that after adjusting for `log(ppgdp)`, `pctUrban` loses its power of explaining the response term, suggesting that the coefficient of `pctUrban` in the two predictor regression model should be insignificant. Let's see.

```
summary(lm(fertility ~ log(ppgdp) + pctUrban, data = UN11))
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + pctUrban, data = UN11)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.151 -0.649 -0.066  0.632  2.991
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.993270   0.399337   20.02   <2e-16 ***
## log(ppgdp)  -0.615142   0.064156   -9.59   <2e-16 ***
## pctUrban    -0.000439   0.004266   -0.10     0.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.933 on 196 degrees of freedom
## Multiple R-squared:  0.52,Adjusted R-squared:  0.515
## F-statistic:  106 on 2 and 196 DF,  p-value: <2e-16
```

Note that in this two predictor regression model, as what we expected, the p-value of `log(ppgdp)` is negligible, which means `log(ppgdp)` is significant in this model. On the other side, the p-value of `pctUrban` is 0.918! which strongly suggests that `pctUrban` is not significant in this two predictor model.

Problem 3.2.4

Show that the estimated coefficient for `log(ppgdp)` is the same as the estimated slope in the added-variable plot for `log(ppgdp)` after `pctUrban`. This correctly suggests that all the estimates in a multiple linear regression model are adjusted for all the other regressors in the mean function.

Solution

```
m1 <- coef(lm(fertility ~ log(ppgdp) + pctUrban, data = UN11))
m2 <- coef(lm(e_f2pU ~ e_lp2pU))
m1
```

```
## (Intercept)  log(ppgdp)    pctUrban
##   7.9932699  -0.6151425  -0.0004393
```

```
m2
```

```
## (Intercept)     e_lp2pU
##  -1.162e-16  -6.151e-01
```

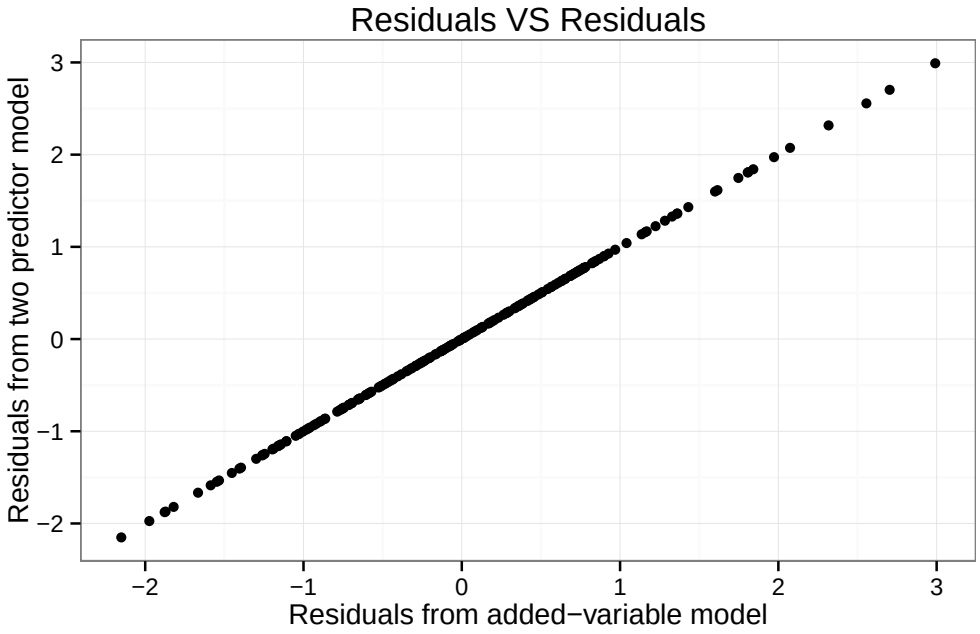
In the code chunk above, `m1` is the two predictor model and `m2` the added-variable model for `log(ppgdp)` after `pctUrban`. We see that the coefficient value of `log(ppgdp)` in two predictor model is the same as the slope coefficient in the added-variable model, suggesting that all the estimates in a multiple linear regression model are adjusted for all the other regressors in the mean function.

Problem 3.2.5

Show that the residuals in the added-variable plot are identical to the residuals from the mean function with both predictors.

Solution

```
m1 <- lm(fertility ~ log(ppgdp) + pctUrban, data = UN11)
m2 <- lm(e_f2pU ~ e_lp2pU)
p <- ggplot(data.frame(m1$residuals, m2$residuals),
            aes(x = m1$residuals, y = m2$residuals))
p <- p + geom_point()
p <- p + theme_bw() + theme(text = element_text(size = 12))
p <- p + ggtitle("Residuals VS Residuals")
p <- p + xlab("Residuals from added-variable model") +
    ylab("Residuals from two predictor model")
p
```



This graph is means to show that residuals from two models are the same. the x axis is the "residuals from added-variable model", the y axis is the "residuals from two predictor model". It shows that all points, representing residual pairs from two models for each case in this data set, are precisely on the line $y = x$, suggesting that the residuals in the added-variable plot are identical to the residuals from the mean function with both predictors.

Problem 3.2.6

Show that the t-test for the coefficient for `log(ppgdp)` is not quite the same from the added-variable plot and from the regression with both regressors, and explain why they are slightly different.

Solution

```
summary(m1)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + pctUrban, data = UN11)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.151 -0.649 -0.066  0.632  2.991
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.993270   0.399337   20.02   <2e-16 ***
## log(ppgdp)  -0.615142   0.064156   -9.59   <2e-16 ***
```

```
## pctUrban    -0.000439   0.004266   -0.10     0.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.933 on 196 degrees of freedom
## Multiple R-squared:  0.52,Adjusted R-squared:  0.515
## F-statistic:  106 on 2 and 196 DF,  p-value: <2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = e_f2pU ~ e_lp2pU)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.151 -0.649 -0.066  0.632  2.991
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.16e-16   6.60e-02    0.00        1
## e_lp2pU     -6.15e-01   6.40e-02   -9.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.93 on 197 degrees of freedom
## Multiple R-squared:  0.319,Adjusted R-squared:  0.316
## F-statistic: 92.4 on 1 and 197 DF,  p-value: <2e-16
```

In the code chunk above, `m1` is the two predictor model and `m2` the added-variable model for `log(ppgdp)` after `pctUrban`. It's easy to see that two t-values are different, suggesting that the t-test for the coefficient for `log(ppgdp)` is not quite the same from the added-variable plot and from the regression with both regressors. But it's only a matter of degree of freedom. In `m1`, the degree of freedom is $n - 3$, but in `m2`, the degree of freedom is $n - 2$. In fact, after adjusting the degree of freedom, the t-test of these two models are essentially the same.

Problem 4.2

(Data file: `Transact`) The data in this example consists of a sample of branches of a large Australian bank (Cunningham and Heathcote, 1989). Each branch makes transactions of two types, and for each of the branches we have recorded the number `t1` of type 1 transactions and the number `t2` of type 2 transactions. The response is `time`, the total minutes of labor used by the branch. Define $a = (\texttt{t2} + \texttt{t1})/2$ to be the average transaction time, and $d = \texttt{t1} - \texttt{t2}$, and fit the following four mean functions

$$\begin{aligned} M1 : E(time|t1, t2) &= \beta_{0,1} + \beta_{1,1}t1 + \beta_{2,1}t2 \\ M2 : E(time|t1, t2) &= \beta_{0,2} + \beta_{3,2}a + \beta_{4,2}d \\ M3 : E(time|t1, t2) &= \beta_{0,3} + \beta_{2,3}t2 + \beta_{4,3}d \\ M4 : E(time|t1, t2) &= \beta_{0,4} + \beta_{1,4}t1 + \beta_{2,4}t2 + \beta_{3,4}a + \beta_{4,4}d \end{aligned}$$

Solution

```
Transact$a <- (Transact$t1 + Transact$t2) / 2
Transact$d <- (Transact$t1 - Transact$t2)
m1 <- lm(time ~ t1 + t2, data = Transact)
m2 <- lm(time ~ a + d, data = Transact)
m3 <- lm(time ~ t2 + d, data = Transact)
m4 <- lm(time ~ t1 + t2 + a + d, data = Transact)
summary(m1)
```

```
##
## Call:
## lm(formula = time ~ t1 + t2, data = Transact)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4652   -601      2    456   5607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441    0.85      0.4
## t1            5.4621     0.4333   12.61   <2e-16 ***
## t2            2.0345     0.0943   21.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 258 degrees of freedom
## Multiple R-squared:  0.909,Adjusted R-squared:  0.908
## F-statistic: 1.29e+03 on 2 and 258 DF,  p-value: <2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = time ~ a + d, data = Transact)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4652   -601      2    456   5607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  144.369    170.544    0.85      0.4
## a              7.497      0.365   20.51  < 2e-16 ***
## d              1.714      0.255    6.73  1.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 258 degrees of freedom
## Multiple R-squared:  0.909,Adjusted R-squared:  0.908
## F-statistic: 1.29e+03 on 2 and 258 DF,  p-value: <2e-16
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = time ~ t2 + d, data = Transact)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4652   -601      2    456   5607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  144.369    170.544    0.85      0.4
## t2             7.497      0.365   20.51   <2e-16 ***
## d              5.462      0.433   12.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 258 degrees of freedom
## Multiple R-squared:  0.909,Adjusted R-squared:  0.908
## F-statistic: 1.29e+03 on 2 and 258 DF,  p-value: <2e-16
```

```
summary(m4)
```

```
##
```

```
## Call:
## lm(formula = time ~ t1 + t2 + a + d, data = Transact)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4652   -601      2    456   5607
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441    0.85      0.4
## t1            5.4621     0.4333   12.61   <2e-16 ***
## t2            2.0345     0.0943   21.57   <2e-16 ***
## a                 NA         NA      NA       NA
## d                 NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 258 degrees of freedom
## Multiple R-squared:  0.909,Adjusted R-squared:  0.908
## F-statistic: 1.29e+03 on 2 and 258 DF,  p-value: <2e-16
```

Problem 4.2.1

In the fit of M4, some of the coefficients estimates are labeled as aliased or else they are simply omitted. Explain what this means and why this happens.

Solution

It's easy to prove that

$$\begin{pmatrix} \texttt{a} \\ \texttt{d} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \texttt{t1} \\ \texttt{t2} \end{pmatrix}$$

Hence there exists perfect multicollinearity among these four predictors, suggesting that linear regression model with these four predictors (`m4`) is not identifiable. Mathematically, the truth is that $rank(X'X) = 3 < 5$, which means it does not have a inverse matrix.

Problem 4.2.2

What aspects of the fitted regressions are the same? What aspects are different?

Solution

- What aspects of the fitted regressions are the same
    1. Residuals

2. Coefficient of determination (R^2)
3. Degree of freedom
4. Adjusted R^2
5. F-test
6. Number of valid Predictors
7. Intercept term

- What aspects are different
  1. Coefficients of Predictors
  2. t-test of predictors' coefficients

Problem 4.2.3

Why is the estimate for `t2` different in M1 and M3?

Solution

Since the correlation between `t2`, `t1` and the correlation between `t2`, `d` are different. As long as predictors are correlated, interpretation of the effect of a predictor depends not only on the other predictors in a model but also upon which linear transformation of those variables is used.

Problem 4.10

Suppose you are given random variables x and y such that

$$x \sim N(\mu_x, \sigma_x^2)$$

$$y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

so you have the marginal distribution of x and the conditional distribution of y given x. The joint distribution of (x, y) is bivariate normal. Find the 5 parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{x,y})$ of the bivariate normal.

Solution

Given μ_x, σ_x^2, β_0, β_1, σ^2, and given the joint distribution of $(x,\ y)$ is bivariate normal, we have

$$\begin{cases} \beta_0 &= \mu_y - \rho\sigma_x\sigma_y\frac{1}{\sigma x^2}\mu_x \\ \beta_1 &= \rho\sigma_x\sigma_y\frac{1}{\sigma_x^2} \\ \sigma^2 &= \sigma_y^2 - \rho\sigma_x\sigma_y\frac{1}{\sigma x^2}\rho\sigma_x\sigma_y \end{cases}$$

$$\Rightarrow \begin{cases} \beta_0 &= \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x \\ \beta_1 &= \rho\frac{\sigma_y}{\sigma_x} \\ \sigma^2 &= (1-\rho^2)\sigma_y^2 \end{cases}$$

$$\Rightarrow \begin{cases} \mu_y = \beta_1\mu_x + \beta_0 \\ \sigma_y^2 = \sigma^2 + \beta_1^2\sigma_x^2 \\ \rho = \sqrt{\frac{\beta_1^2\sigma_x^2}{\sigma^2+\beta_1^2\sigma_x^2}} \end{cases}$$

Problem 4.12

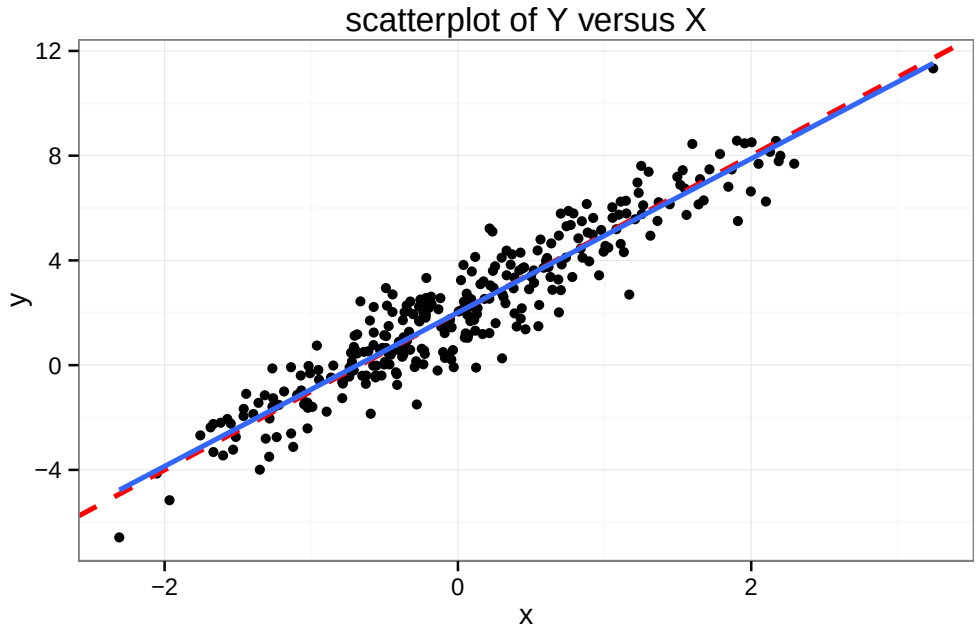
This problem is for you to see what two-dimensional plots of data will look like when the data are sampled from a variety of distributions. For this problem you will need a computer program that allows you to generate random numbers from given distributions. In each of the cases below, set the number of observations $n = 300$, and draw the indicated graphs. Few programs have easy-to-use functions to generate bivariate random numbers, so in this problem you will generate first the predictor X, then the response Y given X.

Problem 4.12.1

Generate X and e to be independent standard normal random vectors of length n. Compute $Y = 2 + 3X + \sigma e$, where in this problem we take $\sigma = 1$. Draw the scatterplot of Y versus X, add the true regression line $Y = 2 + 3X$, and the ols regression line. Verify that the scatter of points is approximately elliptical, and the regression line is similar to, but not exactly the same as, the major axis of the ellipse.

Solution

```
n <- 300
set.seed(123)
x <- rnorm(n = n)
sigma <- 1
y <- 2 + 3 * x + rnorm(n = n, sd = sigma)
f412 = function(x, y)
{
p <- ggplot(data.frame(x, y), aes(x = x, y = y))
p <- p + geom_point()
p <- p + geom_abline(intercept = 2, slope = 3, linetype = 2, size = 1, col = "#ff0000")
p <- p + geom_smooth(formula = y ~ x, method = "lm", se = FALSE, size = 1)
p <- p + theme_bw() + theme(text = element_text(size = 12))
return(p)
}
f412(x, y) + ggtitle("scatterplot of Y versus X")
```



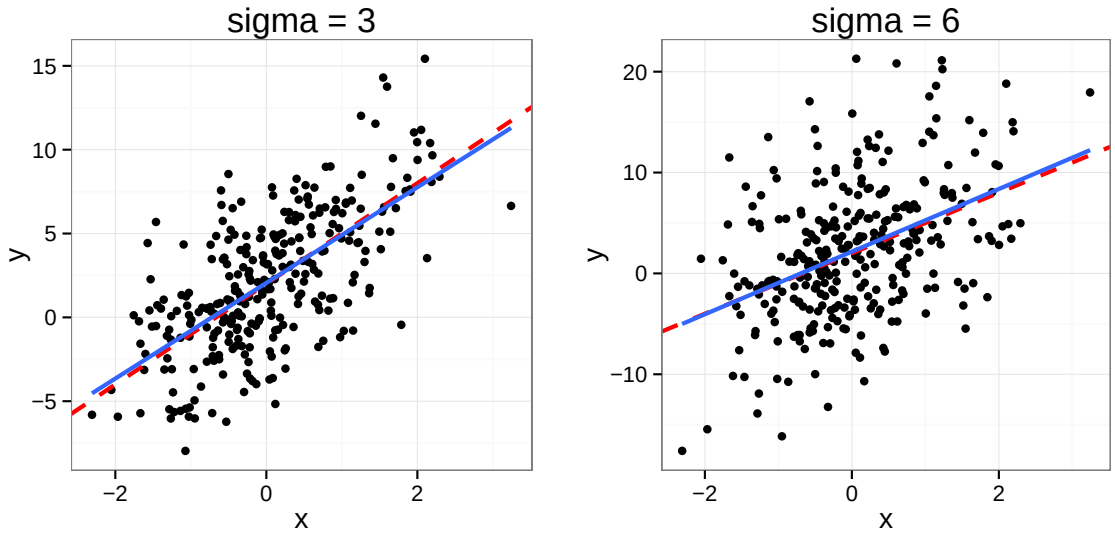
It's easy to see that the scatter of points is approximately elliptical. The solid line is the true regression line, the same as major axis of the ellipse, and the dashed line is the fitted regression line. Note that the fitted regression line is similar to, but not exactly the same as, the major axis of the ellipse.

Problem 4.12.2

Repeat Problem 4.12.1 twice, first set $\sigma = 3$ and then repeat again with $\sigma = 6$. How does the scatter of points change as σ changes?

Solution

```
p1 <- f412(x, y = 2 + 3 * x + rnorm(n = n, sd = 3))
p1 <- p1 + ggtitle("sigma = 3") + theme(text = element_text(size = 14))
p2 <- f412(x, y = 2 + 3 * x + rnorm(n = n, sd = 6))
p2 <- p2 + ggtitle("sigma = 6") + theme(text = element_text(size = 14))
grid.arrange(p1, p2, ncol = 2)
```



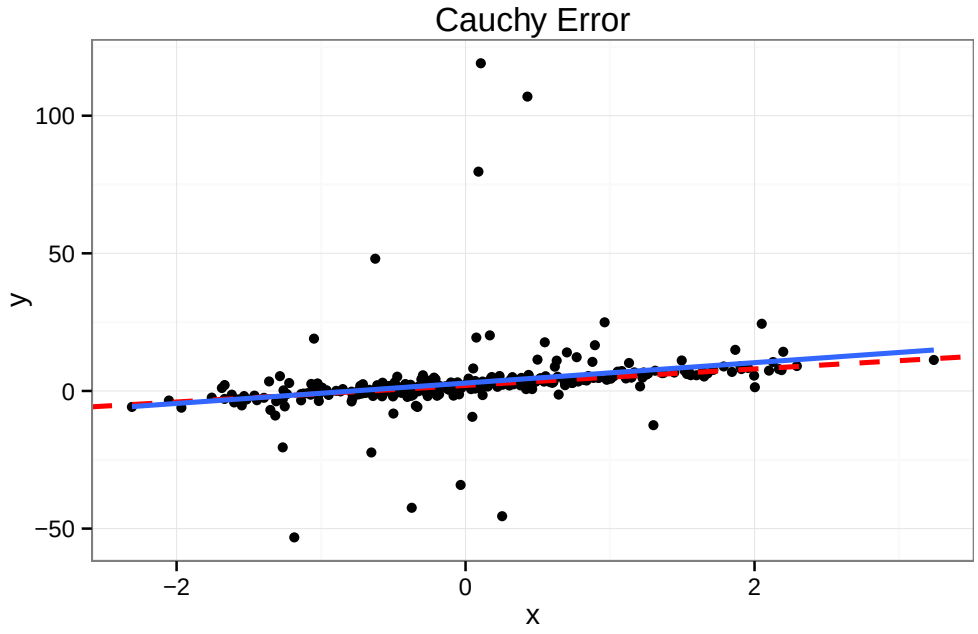
As the σ increase, the conditional variance of y given x also increase.

Problem 4.12.3

Repeat Problem 4.12.1, but this time set X to have a standard normal distribution and e to have a Cauchy distribution (set $\sigma = 1$). The easy way to generate a Cauchy is to generate two vectors V_1 and V_2 of standard normal random numbers, and then set $e = V_1/V_2$. With this setup, the values you generate are not bivariate normal because the Cauchy does not have a population mean or variance.

Solution

```
e <- rnorm(n = n, sd = 1) / rnorm(n = n, sd = 1)
p3 <- f412(x, y = 2 + 3 * x + e)
p3 <- p3 + ggtitle("Cauchy Error") + theme(text = element_text(size = 12))
p3
```



It seems that the scatter of points is not as elliptical as plot with normal error, and there exits many of extreme outliers.

Problem 5.8

Cake data (Data file: `cakes`)

Problem 5.8.1

Fit (5.12) and verify that the significance levels for the quadratic terms and the interaction are all less than 0.005. When fitting polynomials, tests concerning main effects in models that include a quadratic are generally not of much interest.

Solution

```
m <- lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1*X2), data = cakes)
summary(m)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2), data = cakes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.491 -0.308  0.020  0.266  0.545
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.20e+03   2.42e+02   -9.13  1.7e-05 ***
## X1           2.59e+01   4.66e+00    5.56  0.00053 ***
## X2           9.92e+00   1.17e+00    8.50  2.8e-05 ***
## I(X1^2)     -1.57e-01   3.94e-02   -3.98  0.00408 **
## I(X2^2)     -1.20e-02   1.58e-03   -7.57  6.5e-05 ***
## I(X1 * X2)  -4.16e-02   1.07e-02   -3.88  0.00465 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.429 on 8 degrees of freedom
## Multiple R-squared:  0.949,Adjusted R-squared:  0.917
## F-statistic: 29.6 on 5 and 8 DF,  p-value: 5.86e-05
```

Note that the significance levels (p-value) for the quadratic terms and the interaction are all less than 0.005.

Problem 5.8.2

The cake experiment was carried out in two blocks of seven observations each. It is possible that the response might differ by block. For example, if the blocks were different days, then differences in air temperature or humidity when the cakes were mixed might have some effect on Y. We can allow for block effects by adding a factor for block to the mean function and possibly allowing for block by regressor interactions. Add block effects to the mean function fit in Section 5.3.1 and summarize results. The blocking is indicated by the variable `Block` in the data file.

Solution

```
cakes$block <- as.numeric(as.character(cakes$block))
m1 <- lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1*X2) +
             block + I(X1 * block) + I(X2 * block) +
             I(X1^2 * block) + I(X2^2 * block) +
             I(X1*X2 * block), data = cakes)
m2 <- step(object = m1, direction = "both", trace = FALSE)
summary(m2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2) + I(X2 *
##     block) + I(X1^2 * block), data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3883 -0.0159  0.0001  0.0159  0.3417
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.20e+03   1.60e+02  -13.78  9.1e-06 ***
## X1               2.60e+01   3.08e+00    8.45  0.00015 ***
## X2               9.89e+00   7.71e-01   12.83  1.4e-05 ***
## I(X1^2)         -1.60e-01   2.61e-02   -6.15  0.00085 ***
## I(X2^2)         -1.19e-02   1.04e-03  -11.41  2.7e-05 ***
## I(X1 * X2)      -4.16e-02   7.08e-03   -5.88  0.00108 **
## I(X2 * block)   -1.64e-02   4.88e-03   -3.35  0.01536 *
## I(X1^2 * block)  4.76e-03   1.39e-03    3.43  0.01394 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.283 on 6 degrees of freedom
## Multiple R-squared:  0.983,Adjusted R-squared:  0.964
## F-statistic: 50.2 on 7 and 6 DF,  p-value: 6.63e-05
```

In this problem we fit regression model with regressors in problem 5.8.2, the main effect of `block` and interaction terms between `block` and all other predictors. We apply stepwise algorithm in this regression model to make sure that all effects contained in this model are of highly significance.

Note that the main effect of `block` is negligible, given other variable fixed. However, the interaction effect of `block` and `X2`, and `block` and $\texttt{X1}^2$ are significant, suggesting that the influence of `block` over the response is indirect.

Problem 5.14

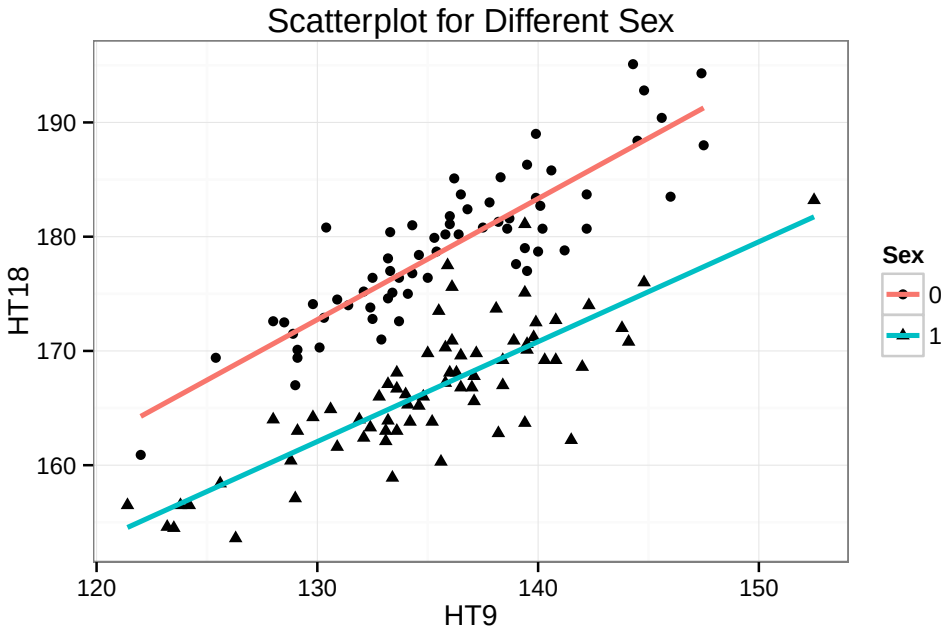
(Data file: BGSall) Refer to the Berkeley Guidance study described in Problem 3.3. Using the data file BGSall, consider the regression of HT18 on HT9 and the grouping factor Sex.

Problem 5.14.1

Draw the scatterplot of HT18 versus HT9, using a different symbol for males and females. Comment on the information in the graph about an appropriate mean function for these data.

Solution

```
BGSall$Sex <- factor(BGSall$Sex)
p <- ggplot(data = BGSall, aes(x = HT9, y = HT18))
p <- p + geom_point(aes(shape = Sex))
p <- p + theme_bw() + theme(text = element_text(size = 12))
p <- p + geom_smooth(formula = y ~ x, method = "lm", se = FALSE, size = 1, aes(colour =
p <- p + ggtitle("Scatterplot for Different Sex")
p
```



Note that the two slope coefficients of two mean functions are almost the same, but the intercepts are different, suggesting that given the same HT9, there exists significant difference of HT18 for two levels of Sex.

Problem 5.14.2

Obtain the appropriate test for a parallel regression model.

Solution

```
m <- lm(HT18 ~ HT9 + Sex, data = BGSall)
summary(m)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + Sex, data = BGSall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.469  -2.095  -0.014   1.710  10.447
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.5173     7.3339    6.62  8.3e-10 ***
## HT9           0.9601     0.0539   17.82  < 2e-16 ***
## Sex1        -11.6958     0.5904  -19.81  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.43 on 133 degrees of freedom
## Multiple R-squared:  0.852,Adjusted R-squared:  0.849
## F-statistic:  382 on 2 and 133 DF,  p-value: <2e-16
```

We can add the group factor `Sex` into the original regression model, and see the t-test result of the coefficient of `Sex`. As the result, we see that the significance level of `Sex` is very low, suggesting that it should be a parallel regression model.

Problem 5.14.3

Assuming the parallel regression model is adequate, estimate a 95% confidence interval for the difference between males and females. For the parallel regression model, this is the difference in the intercepts of the two groups.

Solution

```
confint(object = m, level = .95)
```

```
##                2.5 %  97.5 %
## (Intercept)  34.0112  63.023
## HT9           0.8535   1.067
## Sex1        -12.8635 -10.528
```

To estimate a 95% confidence interval for the difference between males and females, we only need obtain the 95% confidence interval for the coefficient of `Sex`. As we can see, the confidence interval is $[-12.8635, \ -10.528]$