应用回归分析 HW3

邵智轩　　物理学院　　1400012141

3.2

在二元正态分布下，

$$y_i|x_i \sim \mathcal{N}\left(\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), \sigma_y^2(1-\rho_{xy}^2)\right)$$

那么 y 关于 x 的回归直线 $\mathrm{E}[y|x] = \beta_0 + \beta_1 x$ 中，截距 $\beta_0 = \mu_y - \beta_1\mu_x$，斜率 $\beta_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x}$。当 $\beta_1 \neq 0$，即 $\rho_{xy} \neq 0$ 时，反解出 $y = \beta_0 + \beta_1 x$，得到 $x = \frac{1}{\beta_1}(y - \beta_0)$

另一方面，

$$x_i|y_i \sim \mathcal{N}\left(\mu_x + \rho_{xy}\frac{\sigma_x}{\sigma_y}(y_i - \mu_y), \sigma_x^2(1-\rho_{xy}^2)\right)$$

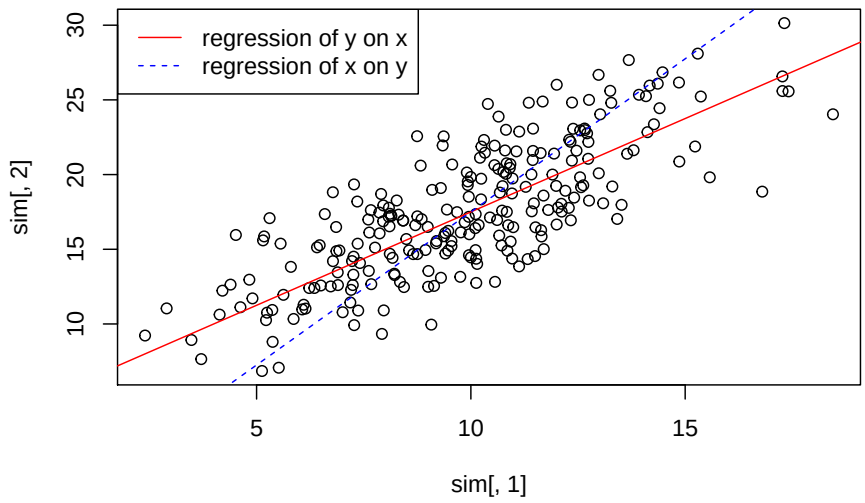
那么 x 关于 y 的回归直线 $\mathrm{E}[x|y] = \beta_0' + \beta_1' y$ 中，截距 $\beta_0' = \mu_x - \beta_1'\mu_y$，斜率 $\beta_1' = \rho_{xy}\frac{\sigma_x}{\sigma_y}$。

显然，一般 $\beta_1\beta_1' \neq 1$。若要 $\beta_1\beta_1' = 1$，显然要求 $\rho_{xy}^2 = 1$，$\rho_{xy} = \pm 1$。

如果 $\beta_1' = 1/\beta_1$，那么 y 关于 x 的回归直线反解得到的 x 轴截距：

$$-\frac{\beta_0}{\beta_1} = -\frac{\mu_y}{\beta_1} + \mu_x = \mu_x - \beta_1'\mu_y = \beta_0'$$

即两直线完全相同。



3.3

3.3.1

```
library(alr4)
attach(longley)
GNP <- GNP * 1e3
Unemployed <- Unemployed * 10
Armed.Forces <- Armed.Forces * 10
Population <- Population * 1e3
Employed <- Employed * 1e3
longley.lm <- lm(Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Population
summary(longley.lm)
```

```
##
## Call:
## lm(formula = Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces +
##     Population + Year)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -410.11 -157.67  -28.16  101.55  455.39
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+06  8.904e+05  -3.911 0.003560 **
## GNP.deflator  1.506e+01  8.491e+01   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e+00  4.884e-01  -4.136 0.002535 **
## Armed.Forces -1.033e+00  2.143e-01  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+03  4.555e+02   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 304.9 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

3.3.2 给出测量误差的估计下界

对于变量 X_2（GNP），单位为百万美元。若数据精确到个位，若认为舍入误差在 $(-0.5, 0.5)$ 上均匀分布，则通过均匀分布的方差 $(b-a)^2/12$ 得到 X_2 的误差估计 $s_2^2 = 1/12$（百万美元的平方）。

对其他变量误差的下界估计几乎完全类似。除了对 X_1 的误差估计 $s_1^2 = 0.01 \times 1/12$ 以外，其余自变量的误差下界估计都是 $s_i^2 = 1/12,\ i = 2, 3, 4, 5$，单位为相应的单位的平方。

对 X_6（Year）的处理要复杂一些，因为年份通常不是四舍五入，而是向下取整的，即误差范围为 $(0, 1)$。如果认为误差是 $(0, 1)$ 上的均匀分布，则 $\mathrm{E}[s_6] = 0.5 \neq 0$，误差估计的下界为

$$\mathrm{E}[(s_6 - 0)^2] = \int_0^1 (x^2 \cdot 1)\mathrm{d}x = \frac{1}{3}$$

，单位为 (Year^2)。非闰年为 365 天，闰年为 366 天，其影响基本可以忽略。

3.3.3 模拟试验

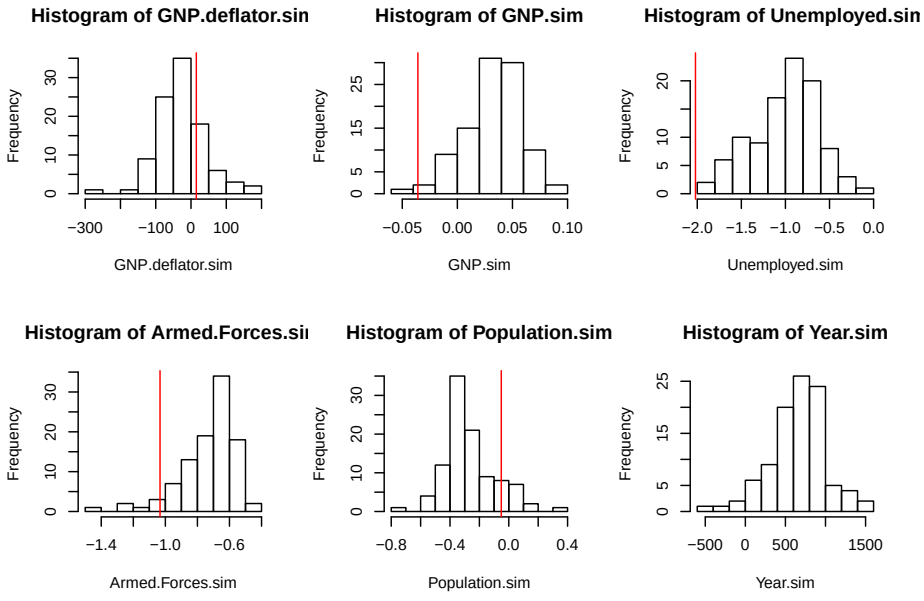
```
longley.sim <- function() {
    GNP.deflator.sim <- GNP.deflator + runif(16, -0.05, 0.05)
    GNP.sim <- GNP + runif(16, -0.5, 0.5)
    Unemployed.sim <- Unemployed + runif(16, -0.5, 0.5)
    Armed.Forces.sim <- Armed.Forces + runif(16, -0.5, 0.5)
    Population.sim <- Population + runif(16, -0.5, 0.5)
    Year.sim <- Year + runif(16, 0, 1)
    longley.sim.lm <- lm(Employed ~ GNP.deflator.sim + GNP.sim + Unemployed.sim + Armed
    longley.sim.lm$coef
}

N <- 100
record=c()
for (i in 1:N) {
    record <- rbind(record, longley.sim())
}
record<-data.frame(record)
summary(record)
```

```
##   X.Intercept.      GNP.deflator.sim      GNP.sim
##  Min.   :-2889037   Min.   :-272.043   Min.   :-0.04090
##  1st Qu.:-1542202   1st Qu.: -63.600   1st Qu.: 0.01760
##  Median :-1226985   Median : -30.473   Median : 0.03532
##  Mean   :-1197740   Mean   : -26.489   Mean   : 0.03261
##  3rd Qu.: -845829   3rd Qu.:   3.385   3rd Qu.: 0.04936
##  Max.   :  942528   Max.   : 164.716   Max.   : 0.09209
##  Unemployed.sim    Armed.Forces.sim  Population.sim       Year.sim
##  Min.   :-1.8414   Min.   :-1.4900   Min.   :-0.7666   Min.   :-438.1
##  1st Qu.:-1.2154   1st Qu.:-0.8133   1st Qu.:-0.3697   1st Qu.: 484.3
##  Median :-0.9431   Median :-0.6886   Median :-0.3041   Median : 671.5
```

```
##  Mean   :-0.9986   Mean   :-0.7347   Mean   :-0.2664   Mean   : 659.6
##  3rd Qu.:-0.7493   3rd Qu.:-0.6193   3rd Qu.:-0.1828   3rd Qu.: 841.7
##  Max.   :-0.1497   Max.   :-0.4718   Max.   : 0.3180   Max.   :1539.1
```

```r
with(record, {
    opar <- par(mfrow = c(2, 3))
    hist(GNP.deflator.sim)
    abline(v = longley.lm$coefficients['GNP.deflator'],col='red')
    hist(GNP.sim)
    abline(v = longley.lm$coefficients['GNP'], col = 'red')
    hist(Unemployed.sim)
    abline(v = longley.lm$coefficients['Unemployed'], col = 'red')
    hist(Armed.Forces.sim)
    abline(v = longley.lm$coefficients['Armed.Forces'], col = 'red')
    hist(Population.sim)
    abline(v = longley.lm$coefficients['Population'], col = 'red')
    hist(Year.sim)
    abline(v = longley.lm$coefficients['Year'], col = 'red')
     })
```



图中的红线代表原来拟合的斜率。可以看到，加入随机误差后，各变量斜率的变化普遍非常大，甚至正负号都有变化。由于我假设年份的误差分布为 $\mathcal{U}(0,1)$（有偏），年份原来的拟合系数甚至不在随机的 100 次的 range 中。

这一实验说明了这样的问题：尽管原来拟合的 R^2 很高，但这并不能说明拟合出的系数是可信的。这一模型对很小的误差（即使只是误差的理论下界）也非常敏感。