Setting up, displaying and interpreting a regression model

1. Regression to the mean. Considering a DGP $t = x + \epsilon$, where $x \sim N(\mu, \sigma^2)$ and $\epsilon \sim N(0, \tau^2)$, we want to prove that for an observation $x = \mu + \sigma$, the predicted mean of t given this observation –that is, $\mathbb{E}(t \mid x = \mu + \sigma) = y(\mu + \sigma)$– is less that one standard deviation away from the mean of the marginal distribution of t.

We can get the mean and standard deviation of the marginal distribution of t considering that $\mathbb{E}(t) = \mathbb{E}(x) + \mathbb{E}(\epsilon) = \mu$ and $\text{var}(t) = \text{var}(x) + \text{var}(\epsilon) - \text{cov}(x, \epsilon)$. As $\text{cov}(x, \epsilon) = 0$ by construction, we know that $t \sim N(\mu, \sigma^2 + \tau^2)$.

Now, we want to prove that,

$y(\mu + \sigma) \leq \mathbb{E}(t) + \text{var}(t)$

$\mu + \sigma \leq \mu + \sqrt{\sigma^2 + \tau^2}$

$\sigma \leq \sqrt{\sigma^2 + \tau^2}$

$\sigma^2 \leq \sigma^2 + \tau^2$

Which will always hold in the presence of some noise with variance $\tau^2 \geq 0$.

```r
widths <- c(5, 1, 2, 3, 3, 1, 1, 2, 2, rep(1, 20), rep(2, 12), 1, 3, 1,
            rep(3, 10), 1, 2, 1, 1, 1, 1, 1, 2, 3, 3, 1, 3, rep(1, 19), 2, 1, 1,
            1, 3, 3, rep(1, 40), 2, 2, 2, 2, 1, 1, 1, 1, 1, 6, 2, 6, 2, 1)
cols <- c('ID', 'EMP', 'EMPOTH', 'OCC', 'IND', 'SELFEMP', 'ROUTINE', 'TASK1',
          'TASK2', 'ROUTINE2', 'GOOD', 'OTHGOOD', 'THANK', 'WENJOY', 'LEARN',
          'RECOG', 'WORKSAT', 'RECOM', 'JOBHOME', 'SUP', 'DECHOW', 'DECWHAT',
          'DISAG', 'PROMOTE', 'SUPERV', 'SUPERV2', 'GOALS', 'MANAG', 'MANLEV',
          'ADULTS', 'KIDS', 'AGEKID1', 'AGEKID2', 'AGEKID3', 'AGEKID4',
          'AGEKID5', 'AGEKID6', 'AGEKID7', 'AGEKID8', 'KIDCARE', 'KIDCOTH',
          'DIFCARE', 'MONCARE', 'STRNCARE', 'COOK', 'SHOP', 'CLEAN', 'LAUNDRY',
          'REPAIR', 'DISHES', 'BUDGET', 'PLANS', 'CHILDC', 'HSWORK', 'MARSTAT',
          'YRALONE', 'PARTNER', 'MARHAPPY', 'CHANGE', 'DIVTHOT', 'SPEMP',
          'SPEMPOTH', 'SPOCC', 'SPIND', 'SPFEEL', 'SPHSWORK', 'VAC', 'HOUSE',
          'MOVE', 'BUY', 'STRNMED', 'STRNFOOD', 'STRNBILL', 'WORRY', 'TENSE',
          'RESTLESS', 'AFRAID', 'FEAR', 'MAD', 'YELL', 'ANGRY', 'TRUST', 'SUSP',
          'AGAINST', 'HEALTH', 'WALK', 'FARWALK', 'EXER', 'DIET', 'HEIGHT',
          'WEIGHT', 'SMOKENOW', 'SMOKEV', 'STAIRS', 'KNEEL', 'CARRY', 'HAND',
          'SEE', 'HEAR', 'DIFWALK', 'PAIN', 'HEAD', 'WEAK', 'SLEEP', 'EFFORT',
          'GETGO', 'MIND', 'SAD', 'LONELY', 'BLUE', 'ENJOY', 'HOPE', 'HAPPY',
          'FATGOOD', 'FATHAPPY', 'RESPSUC', 'RESPANY', 'FATPROB', 'FATBAD',
          'RESPMIS', 'RESPFAIL', 'EMOT', 'SUPTURN', 'SUPTALK', 'USGOODL',
          'USACHIEV', 'USDES', 'USEFFORT', 'USBADL', 'USGREED', 'OWN', 'ED',
          'MOMED', 'FATHED', 'YEARBN', 'RACE', 'RACEOTH', 'HISP', 'REL',
          'RELOTH', 'EARN1', 'EARN2', 'FAMINC1', 'FAMINC2', 'SEX')

# Load data
wfw <- read.fwf('wfw90.txt', widths)

# Set names
colnames(wfw) <- cols

# subset for variables of interest
data <- subset(wfw, select=c(EARN1, EARN2, SEX, HEIGHT, WEIGHT))
```

```r
# excluding outliers
data <- subset(data, WEIGHT < 900)
# converting inches to centimeters, rescaling variable SEX
data[,4] <- as.numeric(substr(data$HEIGHT, 1,1))*30.48+as.numeric(substr(data$HEIGHT, 2,3))*2.54
data[,3] = data[,3] - 1

# Create dataset including observations with aproximate earnings
data_aprox <- data
data_aprox$EARN1[is.na(data_aprox$EARN1)] <- 0
data_aprox$EARN2[is.na(data_aprox$EARN2)] <- 0
data_aprox[,1] <- data_aprox$EARN1/1000 + data_aprox$EARN2
data_aprox <- data_aprox[,c(1,3,4,5)]

# subset data witout aproximate earnings

data <- subset(data, EARN1 != "NA" & EARN1 != 0)
data <- data[,c(1,3,4,5)]

model1 <- lm(EARN1 ~ HEIGHT, data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = EARN1 ~ HEIGHT, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31474 -11518  -3871   6345 369472
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68800.03   11018.91  -6.244 5.95e-10 ***
## HEIGHT         543.13      64.71   8.394  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21760 on 1180 degrees of freedom
## Multiple R-squared:  0.05634,    Adjusted R-squared:  0.05554
## F-statistic: 70.46 on 1 and 1180 DF,  p-value: < 2.2e-16
```

```r
model2 <- lm(EARN1 ~ scale(HEIGHT)[,1], data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = EARN1 ~ scale(HEIGHT)[, 1], data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31474 -11518  -3871   6345 369472
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         23537.6      632.8  37.195   <2e-16 ***
## scale(HEIGHT)[, 1]   5314.1      633.1   8.394   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21760 on 1180 degrees of freedom
## Multiple R-squared:  0.05634,    Adjusted R-squared:  0.05554
## F-statistic: 70.46 on 1 and 1180 DF,  p-value: < 2.2e-16
```

```r
# stdh <- (data_clean$HEIGHT - mean(data_clean$HEIGHT))/sd(data_clean$HEIGHT)
# model2 <- lm(EARN1 ~ stdh, data = data_clean)

model3 <- lm(EARN1 ~ SEX + HEIGHT + WEIGHT, data = data)
summary(model3)
```

```
##
## Call:
## lm(formula = EARN1 ~ SEX + HEIGHT + WEIGHT, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30992 -11312  -3440   6066 368431
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7049.552  16240.149  -0.434   0.6643
## SEX         -9136.017   1802.926  -5.067 4.68e-07 ***
## HEIGHT        204.650     96.665   2.117   0.0345 *
## WEIGHT          6.443     22.024   0.293   0.7699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21530 on 1178 degrees of freedom
## Multiple R-squared:  0.07734,    Adjusted R-squared:  0.07499
## F-statistic: 32.91 on 3 and 1178 DF,  p-value: < 2.2e-16
```

```r
model4 <- lm(log(EARN1) ~ SEX + HEIGHT + WEIGHT, data = data)
summary(model4)
```

```
##
## Call:
## lm(formula = log(EARN1) ~ SEX + HEIGHT + WEIGHT, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2428 -0.3780  0.1421  0.5571  2.8266
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.4510189  0.6676248  12.658  < 2e-16 ***
## SEX         -0.4161913  0.0741174  -5.615 2.45e-08 ***
## HEIGHT       0.0088050  0.0039738   2.216   0.0269 *
## WEIGHT       0.0000613  0.0009054   0.068   0.9460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8852 on 1178 degrees of freedom
## Multiple R-squared:  0.0881, Adjusted R-squared:  0.08578
## F-statistic: 37.94 on 3 and 1178 DF,  p-value: < 2.2e-16
```

```
model5 <- lm(scale(log(EARN1))[,1] ~ SEX + scale(HEIGHT)[,1] + scale(WEIGHT)[,1], data = data)
summary(model5)
```

```
##
## Call:
## lm(formula = scale(log(EARN1))[, 1] ~ SEX + scale(HEIGHT)[, 1] +
##     scale(WEIGHT)[, 1], data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5831 -0.4083  0.1535  0.6018  3.0533
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.257116   0.053573   4.799 1.80e-06 ***
## SEX                 -0.449573   0.080062  -5.615 2.45e-08 ***
## scale(HEIGHT)[, 1]  0.093059   0.041999   2.216   0.0269 *
## scale(WEIGHT)[, 1]  0.002309   0.034108   0.068   0.9460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9561 on 1178 degrees of freedom
## Multiple R-squared:  0.0881, Adjusted R-squared:  0.08578
## F-statistic: 37.94 on 3 and 1178 DF,  p-value: < 2.2e-16
```

```
model6 <- lm(log(EARN1) ~ SEX + HEIGHT + WEIGHT + SEX*HEIGHT + SEX*WEIGHT, data = data)
summary(model6)
```

```
##
## Call:
## lm(formula = log(EARN1) ~ SEX + HEIGHT + WEIGHT + SEX * HEIGHT +
##     SEX * WEIGHT, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3010 -0.3939  0.1460  0.5636  2.8308
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.941097   0.963630   9.279  < 2e-16 ***
## SEX         -0.861802   1.288954  -0.669 0.503878
## HEIGHT       0.001991   0.005929   0.336 0.737058
## WEIGHT       0.004126   0.001415   2.915 0.003625 **
## SEX:HEIGHT   0.009042   0.008011   1.129 0.259215
## SEX:WEIGHT  -0.006920   0.001841  -3.760 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8806 on 1176 degrees of freedom
## Multiple R-squared:  0.09901,    Adjusted R-squared:  0.09518
## F-statistic: 25.85 on 5 and 1176 DF,  p-value: < 2.2e-16
```

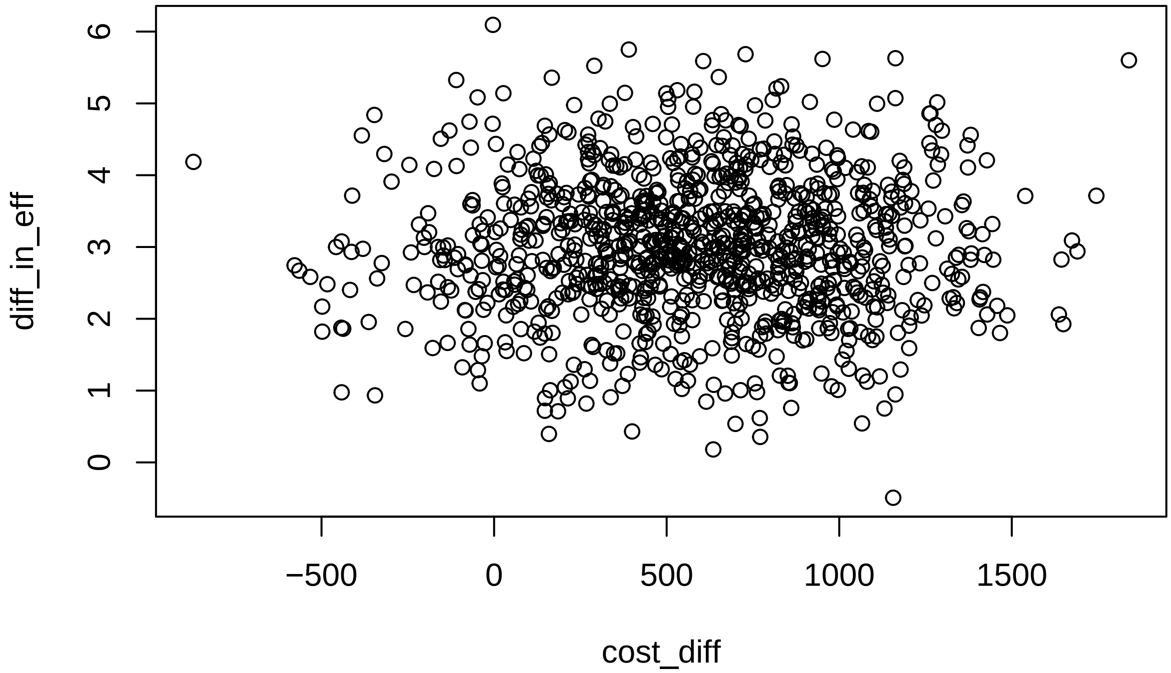
```
model7 <- lm(log(EARN1) ~ SEX + WEIGHT + SEX*HEIGHT + SEX*WEIGHT, data = data)
summary(model6)
```

```
##
## Call:
## lm(formula = log(EARN1) ~ SEX + HEIGHT + WEIGHT + SEX * HEIGHT +
##     SEX * WEIGHT, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3010 -0.3939  0.1460  0.5636  2.8308
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.941097   0.963630   9.279  < 2e-16 ***
## SEX         -0.861802   1.288954  -0.669 0.503878
## HEIGHT       0.001991   0.005929   0.336 0.737058
## WEIGHT       0.004126   0.001415   2.915 0.003625 **
## SEX:HEIGHT   0.009042   0.008011   1.129 0.259215
## SEX:WEIGHT  -0.006920   0.001841  -3.760 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8806 on 1176 degrees of freedom
## Multiple R-squared:  0.09901,    Adjusted R-squared:  0.09518
## F-statistic: 25.85 on 5 and 1176 DF,  p-value: < 2.2e-16
```

```
# Create 1000 simulation draws of the cost difference and the effectiveness difference, and make a scat
library(metRology)
```

```
##
## Attaching package: 'metRology'
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
cost_diff <- rt.scaled(n = 1000, mean = 600, sd = 400, df = 50)
diff_in_eff <- rt.scaled(n = 1000, mean = 3, sd = 1, df = 100)
plot(cost_diff, diff_in_eff)
```



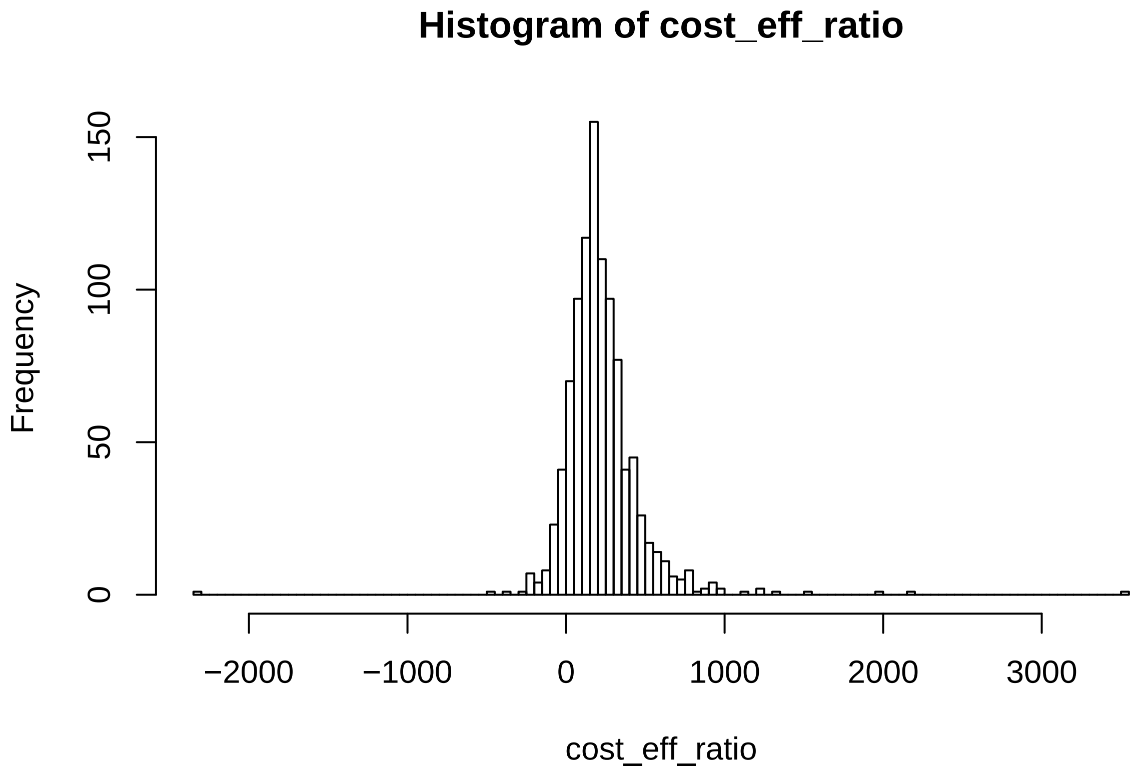
```
cost_eff_ratio = cost_diff / diff_in_eff

mean(cost_eff_ratio)
```

```
## [1] 224.8948
```

```
sd(cost_eff_ratio)
```

```
## [1] 254.0305
```

```
hist(cost_eff_ratio,100)
```


Histogram of cost_eff_ratio
Frequency
0
50
100
150
−2000
−1000
0
1000
2000
3000
cost_eff_ratio