R Club Sewer Project
Sewer and Surface Temperature Regression

Josh Nightingale, Christian Gunninng and Mark Holstad

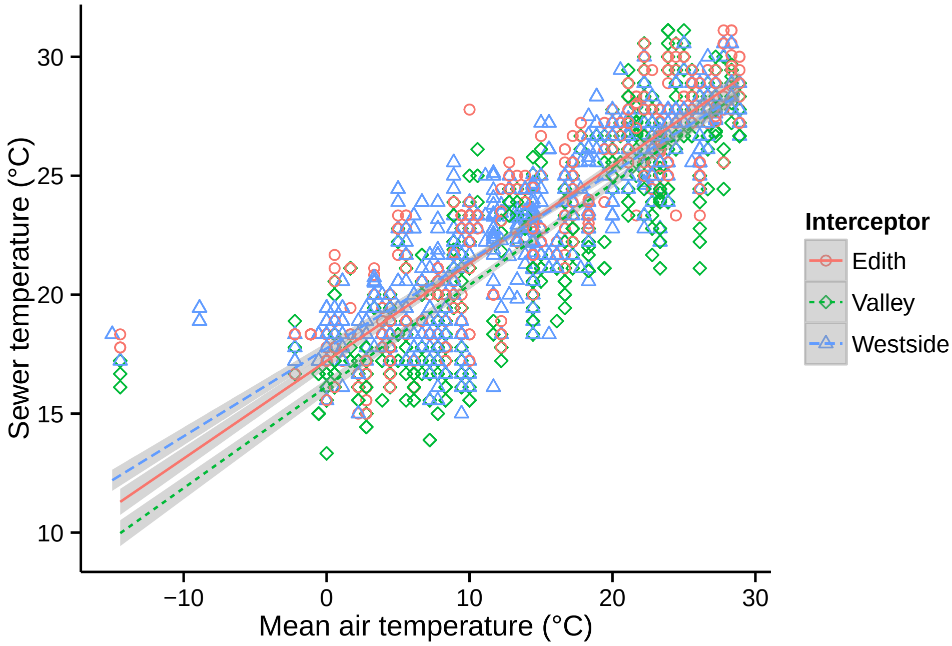
September 14, 2014

```r
## Define column classes to read data there are text comments in line
## with data force measurement cols to read as numeric
## Interceptor,Manhole,Date,Time,Temp,ph,Tot. Sulfide,Dis. Sulfide,Tot.
## Iron,Ferrous Fe,,
.colClasses <- c(Interceptor = "factor", Manhole = "factor", Date = "character",
    Time = "character", Temp = "numeric", ph = "NULL", Tot.Sulfide = "NULL",
    Dis.Sulfide = "NULL", Tot.Iron = "NULL", Ferrous.Fe = "NULL")
## read grab-data path relative to current dir
sewtemp <- read.table("allgrabdata_datefix.csv", sep = ",", header = T,
    comment.char = "#", colClasses = .colClasses)
## xian - posixct gives a full date spec, can't use it *just* for time
## we're not really using this though do this *before* date col
sewtemp$DateTime <- with(sewtemp, as.POSIXct(paste(Date, Time), format = "%d-%m-%y %H:%M"))
sewtemp$Date <- as.POSIXct(sewtemp$Date, format = "%d-%m-%y")  # fix dates

# some Temperatures have been entered as Celsius; most are Fahrenheit
# above freezing
.F.rows <- which(sewtemp$Temp > 32)
sewtemp$Temp[.F.rows] <- fahrenheit.to.celsius(sewtemp$Temp[.F.rows])
sewtemp <- unique(sewtemp)  # remove duplicate entries
# sewtemp£ph[sewtemp£ph > 14] <- NA # remove erroneous entries
str(sewtemp)  # inspect

## 'data.frame': 1998 obs. of  6 variables:
##  $ Interceptor: Factor w/ 3 levels "Edith","Valley",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Manhole    : Factor w/ 15 levels "12stove","blakeco",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Date       : POSIXct, format: "2005-12-14" ...
##  $ Time       : chr  "14:15" "11:15" "09:50" "11:51" ...
##  $ Temp       : num  21.1 20.6 25.6 26.1 26.1 ...
##  $ DateTime   : POSIXct, format: "2005-12-14 14:15:00" ...
```

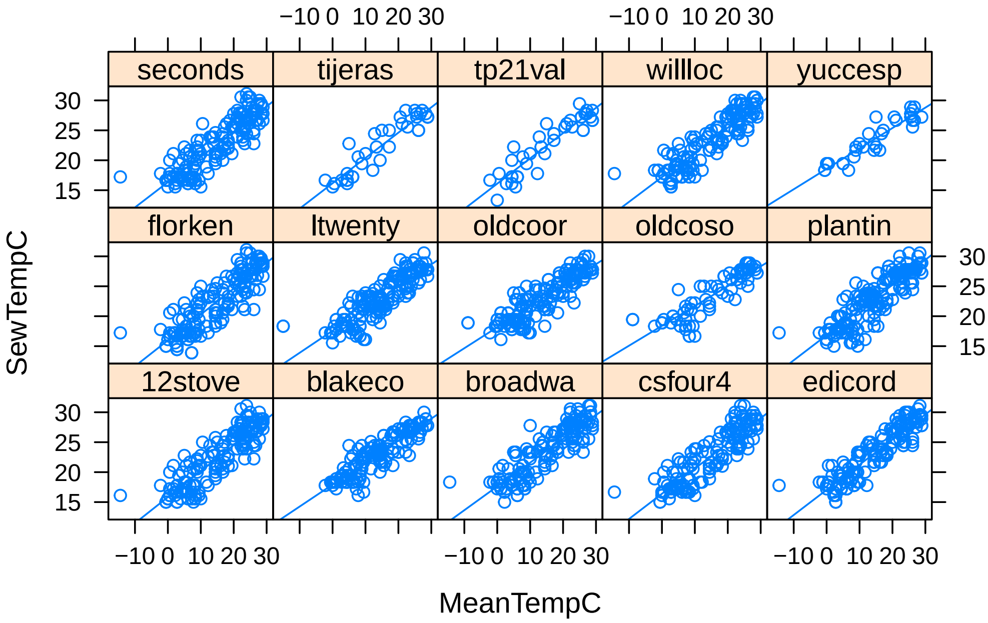
```
## read weather
weather <- read.csv("http://unm-r-programming.googlecode.com/git/sewer/abq-temps-2005-2014.cs
## shorten colnames for convenience
colnames(weather) <- gsub(".Temperature", "Temp", colnames(weather))

# weather <-
# read.csv('http://unm-r-programming.googlecode.com/files/kabq-2009-2013.csv')
# Turn factor into date
weather$Date <- as.POSIXct(weather$MST, format = "%Y-%m-%d")
# Convert Fahrenheit into Celsius find cols containing temp
.wcols <- grep("TempF", colnames(weather))
weather[, .wcols] <- fahrenheit.to.celsius(weather[, .wcols])
## update colnames to reflect C
colnames(weather) <- gsub("TempF", "TempC", colnames(weather))
# weather <- rename(weather, c(MST = Inspect
str(weather)

## 'data.frame': 3550 obs. of  5 variables:
##  $ MST      : Factor w/ 3288 levels "2005-10-1","2005-10-10",..: 32 73 115 118 119 120 121
##  $ MaxTempC : num  12.22 7.22 8.89 10 5.56 ...
##  $ MeanTempC: num  7.22 5 5.56 5.56 2.78 -1.11 0 6.67 8.89 10 ...
##  $ MinTempC : num  2.22 2.78 2.78 0.56 -3.33 -6.11 -4.44 1.11 3.89 3.89 ...
##  $ Date     : POSIXct, format: "2005-01-01" ...

## join to sewer temperatures intersect(colnames(weather),
## colnames(sewtemp)) # both contain 'Date
sewer.weather <- join(sewtemp, weather)

## Joining by:  Date

summary(sewer.weather)

##    Interceptor     Manhole          Date
##  Edith    :522   ltwenty:228   Min.   :2005-09-28 00:00:00
##  Valley   :682   plantin:220   1st Qu.:2007-08-07 00:00:00
##  Westside:940   blakeco:195   Median :2009-11-18 00:00:00
##                 oldcoor:187   Mean   :2009-07-31 22:00:16
##                 broadwa:168   3rd Qu.:2011-07-21 00:00:00
##                 12stove:162   Max.   :2012-12-19 00:00:00
##                 (Other):984   NA's   :22
##      Time                Temp          DateTime
##  Length:2144        Min.   :13.3   Min.   :2005-09-28 11:45:00
##  Class :character   1st Qu.:18.9   1st Qu.:2007-06-20 11:35:00
##  Mode  :character   Median :22.8   Median :2010-01-05 11:25:00
##                     Mean   :22.8   Mean   :2009-08-10 13:02:04
```

```
##                       3rd Qu.:26.7   3rd Qu.:2011-08-02 14:00:00
##                       Max.   :31.1   Max.   :2012-12-19 14:45:00
##                       NA's   :115    NA's   :163
##          MST             MaxTempC        MeanTempC          MinTempC
##   2008-1-22:  28   Min.   :-12.8   Min.   :-15.00   Min.   :-21.67
##   2009-1-21:  24   1st Qu.: 14.0   1st Qu.:  7.22   1st Qu.:  0.56
##   2007-1-23:  22   Median : 22.2   Median : 14.44   Median :  7.22
##   2010-1-19:  22   Mean   : 21.3   Mean   : 14.60   Mean   :  7.63
##   2012-1-17:  22   3rd Qu.: 30.0   3rd Qu.: 22.78   3rd Qu.: 16.11
##   (Other)  :2004   Max.   : 36.7   Max.   : 28.89   Max.   : 22.22
##   NA's     :  22   NA's   :22      NA's   :22       NA's   :22

## inspect, explicitly remove NAs
sewer.weather <- na.omit(sewer.weather)
## rename sewer temp col
sewer.weather <- rename(sewer.weather, c(Temp = "SewTempC"))

## xian - changed to merge, added suffixes head(sewer.weather)
```

```
# plot with ggplot
p <- ggplot(sewer.weather)
p <- p + geom_point(aes(x = MeanTempC, y = SewTempC, colour = Interceptor,
    shape = Interceptor))
p <- p + scale_shape_manual(values = c(21, 23, 24))
p <- p + geom_smooth(aes(x = MeanTempC, y = SewTempC, colour = Interceptor,
    linetype = Interceptor), method = "lm")
p <- p + theme_classic()
p <- p + xlab("Mean air temperature (C)") + ylab("Sewer temperature (C)")
print(p)
```

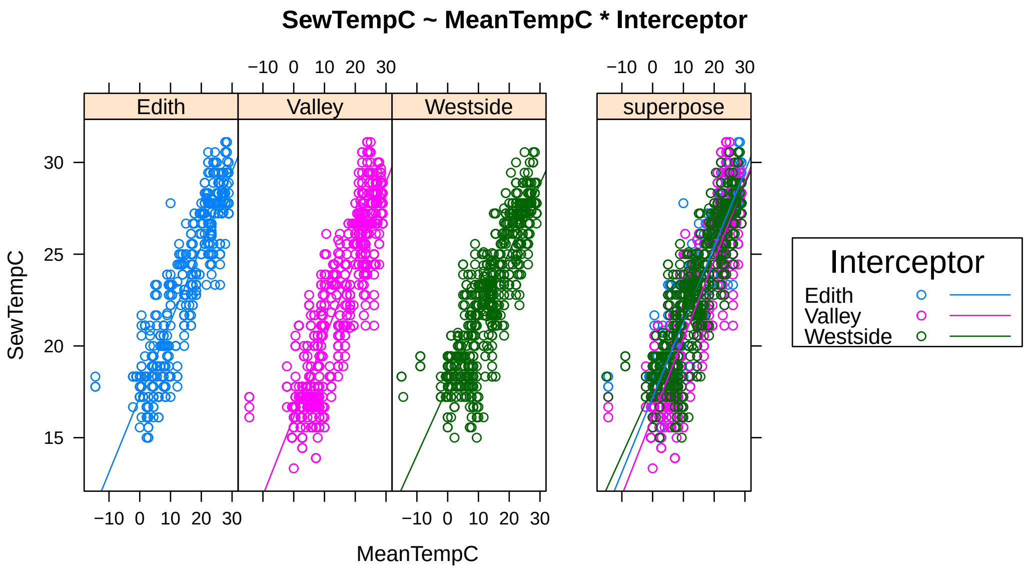



```
xyplot(SewTempC ~ MeanTempC | Manhole, sewer.weather, type = c("p", "r"))
```




```
## First, basic anova shows effect of Interceptor but not Manhole
anova(lm(SewTempC ~ Manhole, sewer.weather))

## Analysis of Variance Table
##
## Response: SewTempC
##              Df Sum Sq Mean Sq F value Pr(>F)
## Manhole      14    275    19.6    1.12   0.34
## Residuals 1900  33410    17.6

##
anova(lm(SewTempC ~ Interceptor, sewer.weather))

## Analysis of Variance Table
##
## Response: SewTempC
##                Df Sum Sq Mean Sq F value Pr(>F)
## Interceptor     2    215   107.3    6.13 0.0022 **
## Residuals    1912  33470    17.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ancova(SewTempC ~ MeanTempC * Interceptor, data = sewer.weather)
```

```
## Analysis of Variance Table
##
## Response: SewTempC
##                         Df Sum Sq Mean Sq F value  Pr(>F)
## MeanTempC                1  25828   25828  6598.2 < 2e-16 ***
## Interceptor              2    283     142    36.2 3.8e-16 ***
## MeanTempC:Interceptor    2    100      50    12.8 3.0e-06 ***
## Residuals             1909   7473       4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
## build all possible models in named list
## y = mx + b
temp.lin.models <- list(
    null=lm(SewTempC ~ MeanTempC, data=sewer.weather),
    ## including min & max temp - signif but doesn't help much
    #all.temp=lm(SewTempC ~ MeanTempC + Max, data=sewer.weather),
    b.by.interceptor=lm(SewTempC ~ MeanTempC + Interceptor, data=sewer.weather),
    b.by.manhole=lm(SewTempC ~ MeanTempC + Manhole, data=sewer.weather),
    m.by.interceptor=lm(SewTempC ~ MeanTempC : Interceptor, data=sewer.weather),
    m.by.manhole=lm(SewTempC ~ MeanTempC : Manhole, data=sewer.weather),
    mb.by.interceptor=lm(SewTempC ~ MeanTempC * Interceptor, data=sewer.weather),
    mb.by.manhole=lm(SewTempC ~ MeanTempC * Manhole, data=sewer.weather)
)
## xian - shared intercept, different slopes
```

```
## best model??
## use maximum likelihood (REML=F) so results are comparable w/anova
temp.mix.models <- list(
    rand_both=lmer(SewTempC ~ MeanTempC + (1|Interceptor:Manhole), data=sewer.weather, REML=F
    rand_both_1=lmer(SewTempC ~ MeanTempC + (1|Interceptor/Manhole), data=sewer.weather, REML
    rand_interceptor=lmer(SewTempC ~ MeanTempC + (1|Interceptor), data=sewer.weather, REML=F)
    rand_manhole=lmer(SewTempC ~ MeanTempC + (1|Manhole), data=sewer.weather, REML=F),
    fixed_b_by_interceptor.rand_manhole=lmer(SewTempC ~ MeanTempC+Interceptor  + (1|Manhole),
    fixed_m_by_interceptor.rand_manhole=lmer(SewTempC ~ MeanTempC:Interceptor  + (1|Manhole),
    fixed_mb_by_interceptor.rand_manhole=lmer(SewTempC ~ MeanTempC*Interceptor  + (1|Manhole)
)

## compare linear models
##
## convenience function
## function returns the list elements with the n best scores
.best.n <- function(.list, .scores, n=2) {
    ## order list
   .list <- .list[ order(unlist(.scores)) ]
    ## only return the first n elements
    ret <- .list[ 1:n ]
    ret
}
## show BIC of each model
## smaller is better
.lin.bic <- llply(temp.lin.models, function(x) BIC(x))
.lin.bic

## $null
## [1] 8160
##
## $b.by.interceptor
## [1] 8105
##
## $b.by.manhole
## [1] 8177
##
## $m.by.interceptor
## [1] 8147
##
## $m.by.manhole
## [1] 8229
##
## $mb.by.interceptor
## [1] 8095
```

```
##
## $mb.by.manhole
## [1] 8243

## pull out best 2
.lin.best <- .best.n(temp.lin.models, .lin.bic)
## compare with anova
anova(.lin.best[[2]], .lin.best[[1]])

## Analysis of Variance Table
##
## Model 1: SewTempC ~ MeanTempC + Interceptor
## Model 2: SewTempC ~ MeanTempC * Interceptor
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   1911 7573
## 2   1909 7473  2       100 12.8  3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## same for mix models
## show BIC of each model
.mix.bic <- llply(temp.mix.models, function(x) BIC(x))
.mix.bic

## $rand_both
## [1] 8120
##
## $rand_both_1
## [1] 8117
##
## $rand_interceptor
## [1] 8110
##
## $rand_manhole
## [1] 8120
##
## $fixed_b_by_interceptor.rand_manhole
## [1] 8113
##
## $fixed_m_by_interceptor.rand_manhole
## [1] 8123
##
## $fixed_mb_by_interceptor.rand_manhole
## [1] 8102
```

```
## best 2
.mix.best <- .best.n(temp.mix.models, .mix.bic)
## compare with anova
anova(.mix.best[[2]], .mix.best[[1]])

## Data: sewer.weather
## Models:
## .mix.best[[2]]: SewTempC ~ MeanTempC + (1 | Interceptor)
## .mix.best[[1]]: SewTempC ~ MeanTempC * Interceptor + (1 | Manhole)
##                Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## .mix.best[[2]]  4 8088 8110  -4040     8080
## .mix.best[[1]]  8 8058 8102  -4021     8042    38      4    1.1e-07
##
## .mix.best[[2]]
## .mix.best[[1]] ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## From Bolker's lmm page:
## http://glmm.wikidot.com/faq
.pseudo.r.sq1 <- function(m) {
    1-var(residuals(m))/(var(model.response(model.frame(m))))
}
.pseudo.r.sq2 <- function(m) {
   lmfit <-  lm(model.response(model.frame(m)) ~ fitted(m))
   summary(lmfit)$r.squared
}
## xian - I'm *not* sure the BIC numbers above are directly comparable
## between linear models and mixed models
## I *think* they are??
## In any case, the simple mb.by.interceptor model is good
## the best mixed model might satisfy model assumptions a little better...
##
## note that the model also fails badly in the lower tail -
## e.g. nonlinear at low temps

## pseudo-r-sq of both are approx equivalent to each other
## and to r-sq of best linear model
.pseudo.r.sq1(.mix.best[[1]])

## [1] 0.7784

.pseudo.r.sq2(.mix.best[[1]])

## [1] 0.7784
```

```
## show summary of best linear and mixed model;
summary(.lin.best[[1]])

##
## Call:
## lm(formula = SewTempC ~ MeanTempC * Interceptor, data = sewer.weather)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.233 -1.412  0.094  1.272  7.246
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    17.2139     0.1631  105.51  < 2e-16
## MeanTempC                       0.4103     0.0094   43.66  < 2e-16
## InterceptorValley              -1.0761     0.2168   -4.96  7.5e-07
## InterceptorWestside             0.5294     0.2112    2.51   0.0122
## MeanTempC:InterceptorValley     0.0165     0.0125    1.32   0.1864
## MeanTempC:InterceptorWestside  -0.0406     0.0124   -3.27   0.0011
##
## (Intercept)                   ***
## MeanTempC                     ***
## InterceptorValley             ***
## InterceptorWestside           *
## MeanTempC:InterceptorValley
## MeanTempC:InterceptorWestside **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.98 on 1909 degrees of freedom
## Multiple R-squared:  0.778,Adjusted R-squared:  0.778
## F-statistic: 1.34e+03 on 5 and 1909 DF,  p-value: <2e-16

summary(.mix.best[[1]])

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: SewTempC ~ MeanTempC * Interceptor + (1 | Manhole)
##    Data: sewer.weather
##
##      AIC      BIC   logLik deviance df.resid
##     8058     8102    -4021     8042     1907
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.149 -0.711  0.049  0.645  3.672
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Manhole  (Intercept) 0.00193  0.0439
##  Residual             3.90027  1.9749
## Number of obs: 1915, groups: Manhole, 15
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                    17.21001    0.16460   104.6
## MeanTempC                       0.41033    0.00938    43.7
## InterceptorValley              -1.07273    0.21870    -4.9
## InterceptorWestside             0.53512    0.21302     2.5
## MeanTempC:InterceptorValley     0.01655    0.01249     1.3
## MeanTempC:InterceptorWestside -0.04059    0.01241    -3.3
##
## Correlation of Fixed Effects:
##             (Intr) MnTmpC IntrcV IntrcW MTC:IV
## MeanTempC   -0.828
## IntrcptrVll -0.753  0.623
## IntrcptrWst -0.773  0.639  0.582
## MnTmpC:IntV  0.622 -0.751 -0.829 -0.480
## MnTmpC:IntW  0.626 -0.756 -0.471 -0.832  0.568
```