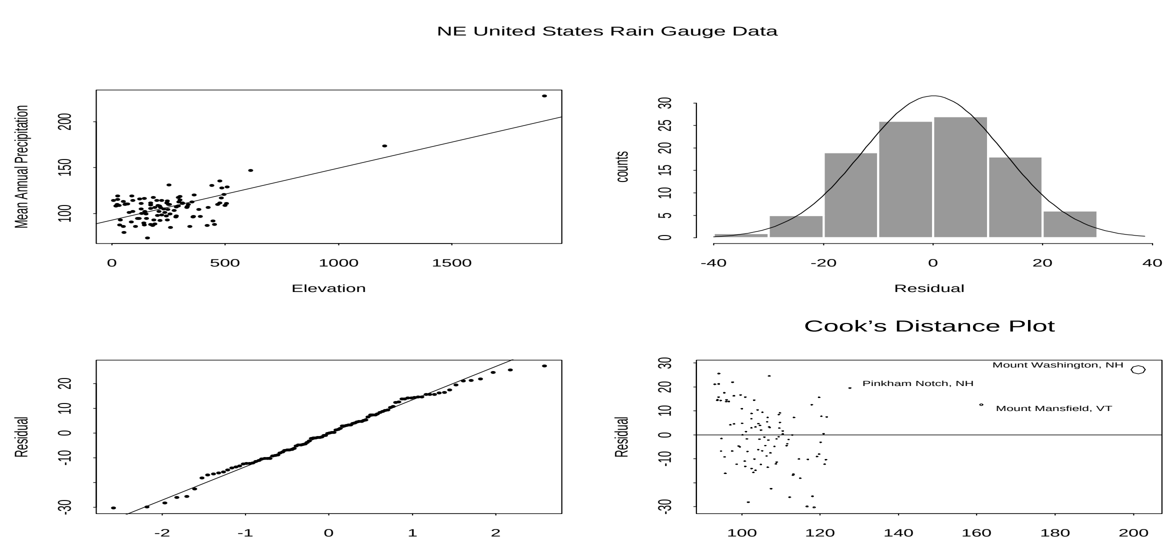
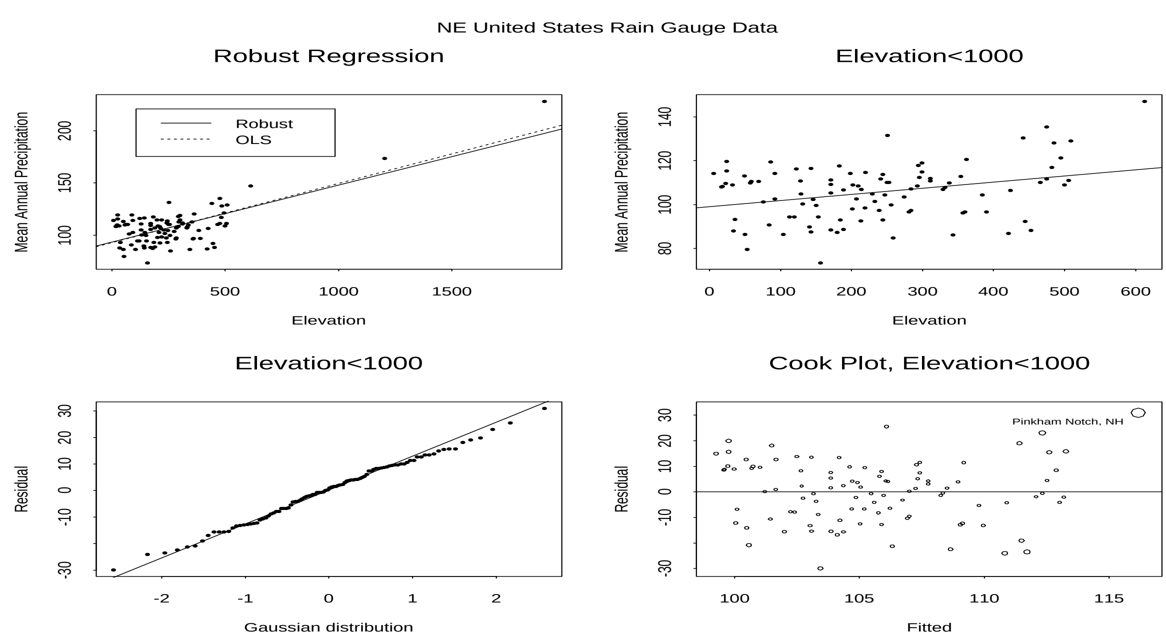
36-401 Class Notes and Exercise for 12/12/00
End-of-Term Review
Howard Seltman

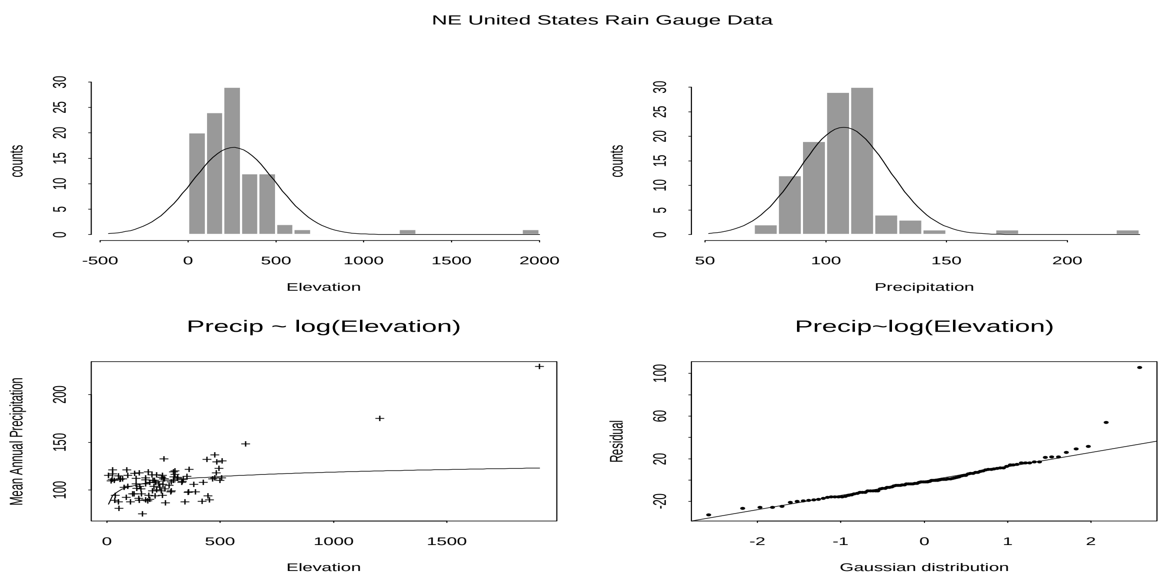
```
# Elevation in meters, Precipitation in cm (30 yr average)
rain_read.table("rain.dat", col.names=c("Site","Elevation","Precip"))
n_nrow(rain)  # 102
rain[1:3,]
#              Elevation Precip
#   Berlin, NH       284   97.5
#     Beth, NH       421   86.8
# Bradford, NH       296  112.4

par(mfrow=c(2,2), oma=c(1,1,2,1))
plot(rain$Elevation, rain$Precip, xlab="Elevation", ylab="Mean Annual Precipitation")
mtext("NE United States Rain Gauge Data", outer=T, cex=1.2)
rslt_lm(Precip~Elevation, rain); abline(rslt)
summary(rslt)
# Residuals:
#     Min     1Q   Median    3Q   Max
#  -30.28 -9.154 -0.02941 9.104 27.18
#
# Coefficients: Value Std. Error t value Pr(>|t|)
# (Intercept) 92.7587  1.8995    48.8344  0.0000
#   Elevation  0.0568  0.0054    10.4710  0.0000
#
# Residual standard error: 12.94 on 100 degrees of freedom
# Multiple R-Squared: 0.523
# F-statistic: 109.6 on 1 and 100 degrees of freedom, the p-value is 0
#
# Correlation of Coefficients:
#           (Intercept)
# Elevation -0.7384
ndhist(rslt$res, xlab="Residual")
qqn(rslt$res, main="", ylab="Residual")
CookPlot(rslt, main="Cook's Distance Plot", id=T, icex=0.7)
ps("rain1")
```

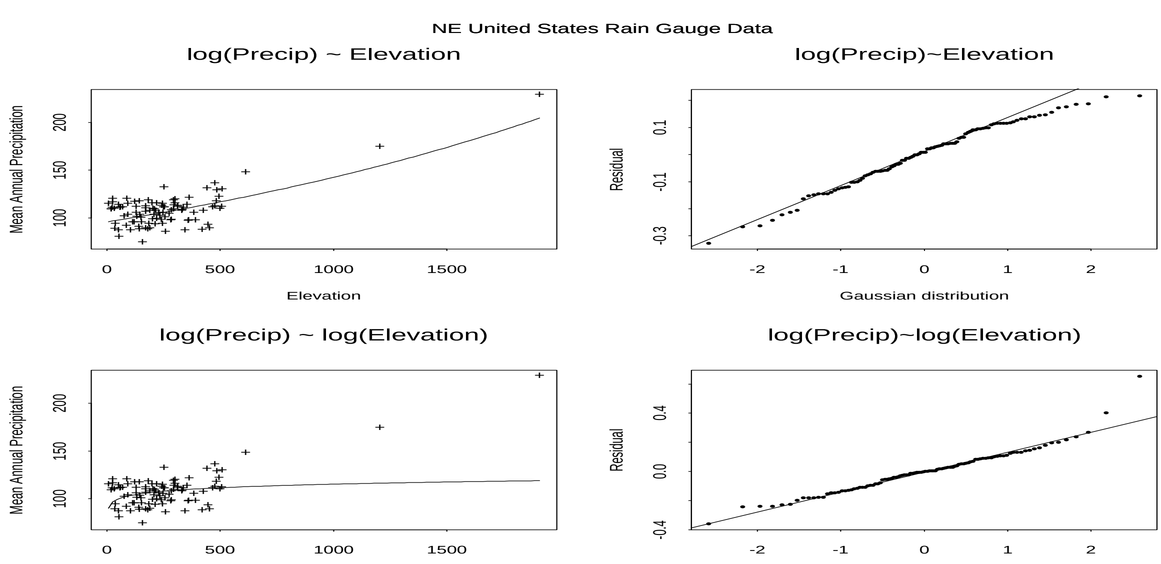


```
K_length(rslt$coef) # 2
inf_influence(rslt)
pot_inf$hat; names(pot)_row.names(rain)
cutoff_2*K/n  # 0.0392;  pot[pot>=cutoff]
# Mount Washington, NH Mount Mansfield, VT
#          0.4890035           0.1671871
dfb_inf$dfbeta;  cutoff_2/sqrt(n)  # 0.198
dfb[dfb[,2]>=cutoff,2]
# Mount Washington, NH Pinkham Notch, NH Mount Mansfield, VT
#             2.962498         0.2332548           0.4647625


library("MASS")
rrslt_rlm(Precip~Elevation, rain);  summary(rrslt)
# Coefficients: Value Std. Error t value
# (Intercept) 93.5332  1.9084    49.0124
#   Elevation  0.0545  0.0054     9.9971
#
# Residual standard error: 13.54 on 100 degrees of freedom
wt_rrslt$w; names(wt)_row.names(rain);  sort(round(wt[wt<1],2))
#  Mount Washington, NH Beth, NH Upper Dam, ME Peru, NY Montpelier, VT
#                  0.59     0.61          0.61     0.64            0.7
#  Middle Dam, ME Ipswich, MA Mays Mills, VT Gilman, VT Sanford, ME Rockport, MA
#            0.72        0.73           0.75        0.8        0.85         0.89
#  Pinkham Notch, NH Newburyport, MA
#                0.9             0.9

Low_rain$Elevation<1000
rsltX_lm(Precip~Elevation, rain, subset=Low);  summary(rsltX)
# Coefficients: Value Std. Error t value Pr(>|t|)
# (Intercept) 99.0877  2.3130    42.8400  0.0000
#   Elevation  0.0279  0.0085     3.2805  0.0014
#
# Residual standard error: 12.02 on 98 degrees of freedom
# Multiple R-Squared: 0.09895
```

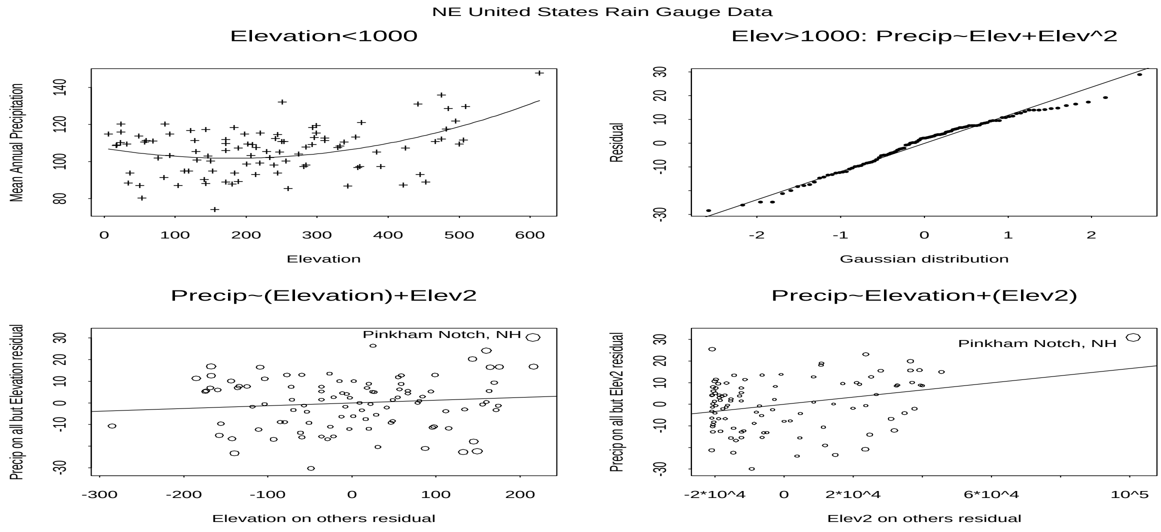


```
par(mfrow=c(2,2), oma=c(1,1,2,1))
ndhist(rain$Elevation, 15, xlab="Elevation")
ndhist(rain$Precip, 15, xlab="Precipitation")

plot(rain$Elevation, rain$Precip, xlab="Elevation", ylab="Mean Annual Precipitation",
  main="Precip ~ log(Elevation)", pch="+")
rain$LElev_log(rain$Elevation)
mtext("NE United States Rain Gauge Data", outer=T, cex=1.2)
x_seq(min(rain$Elevation), max(rain$Elevation), len=100)
cf_rsltL$coef
lines(x, cf[1]+cf[2]*log(x))
qqn(rsltL$res, main="Precip~log(Elevation)", ylab="Residual")
ps("rain3")
```

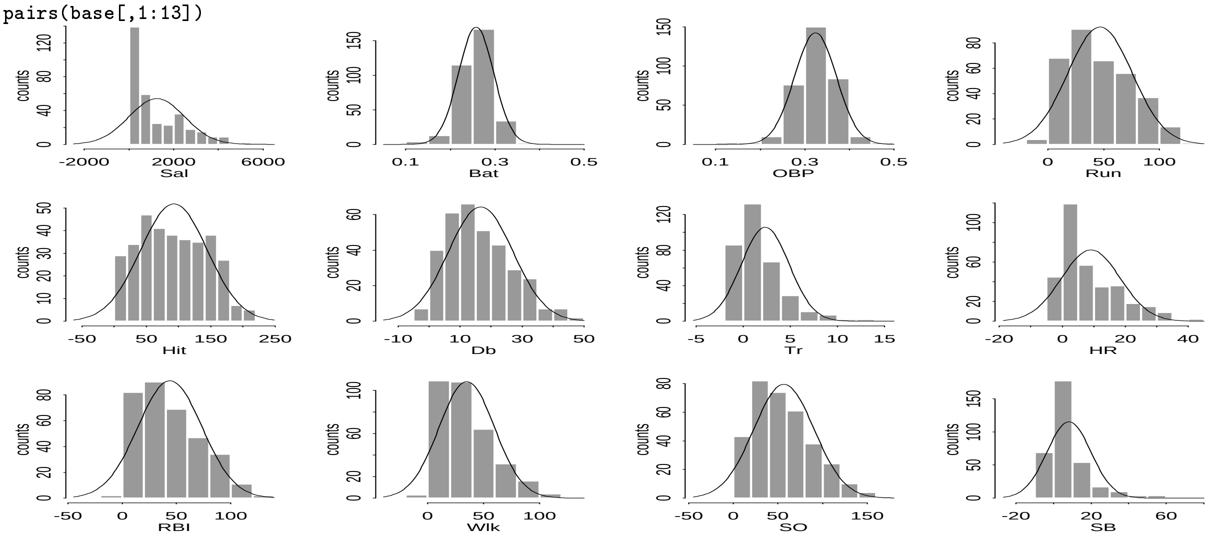


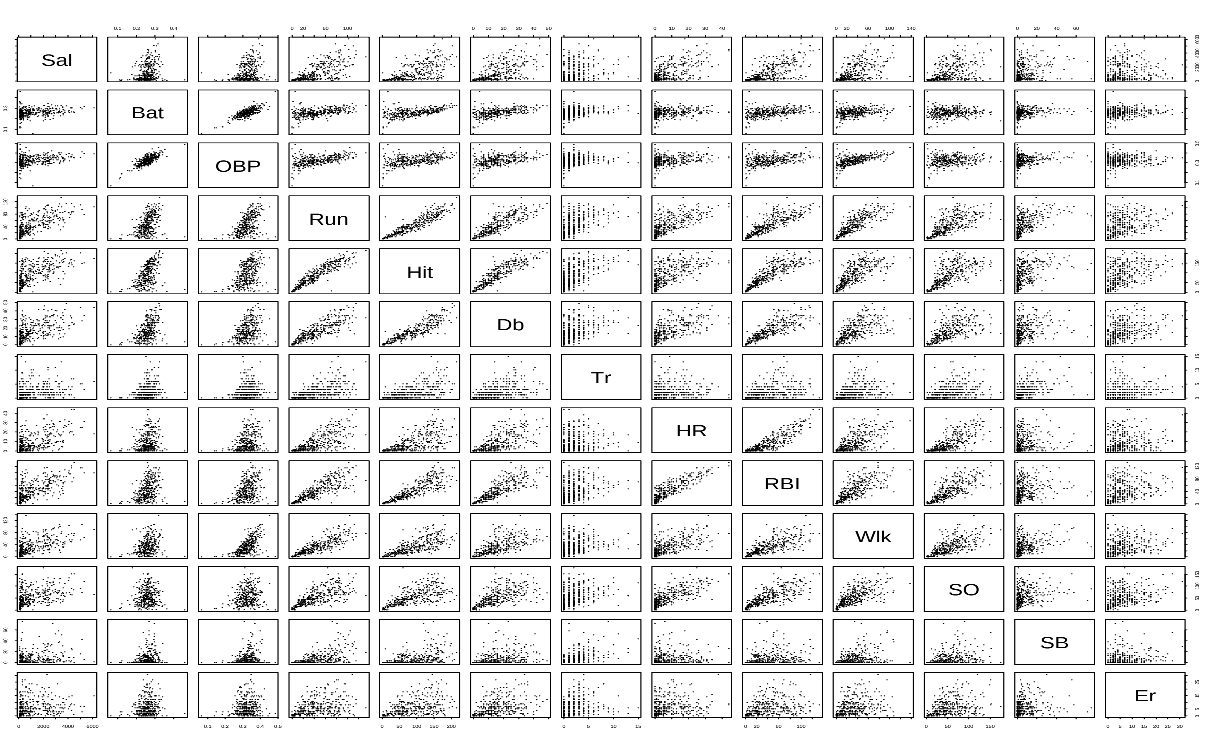


```
rain$Elev2_rain$Elevation^2
par(mfrow=c(2,2), oma=c(1,1,2,1))
plot(rain$Elevation[Low], rain$Precip[Low], xlab="Elevation",
  ylab="Mean Annual Precipitation", main="Elevation<1000", pch="+")
rsltQ_lm(Precip~Elevation+Elev2, rain, subset=Low)
mtext("NE United States Rain Gauge Data", outer=T, cex=1.2)
x_seq(min(rain$Elevation[Low]), max(rain$Elevation[Low]), len=100)
cf_rsltQ$coef
lines(x, cf[1]+cf[2]*x+cf[3]*x^2)
identify(rain$Elevation[Low], rain$Precip[low], row.names(rain)[Low], cex=0.7)
qqn(rsltQ$res, main="Elev>1000: Precip~Elev+Elev^2", ylab="Residual")
partreg(c("Elevation","Elev2"), "Precip", "Elevation", rain[Low], prop=T, id=T)
partreg(c("Elevation","Elev2"), "Precip", "Elev2", rain[Low,], prop=T, id=T)
ps("rain5")
```



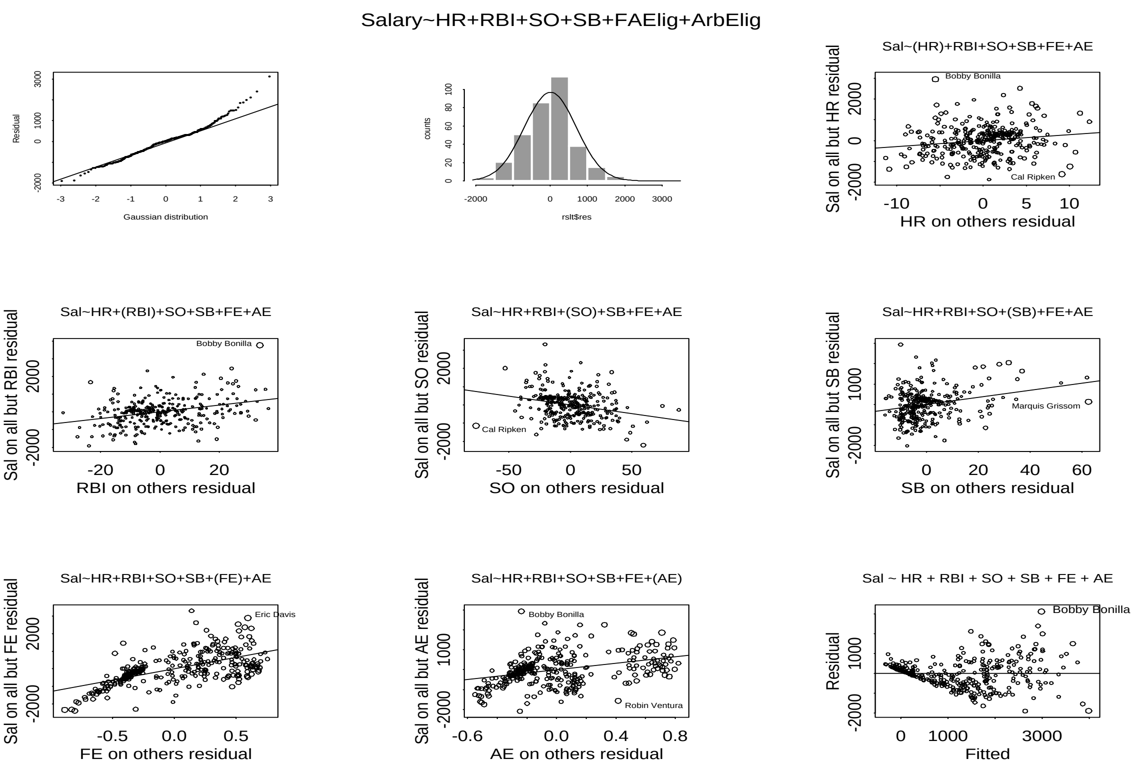
```
# http://www.amstat.org/publications/jse/archive.htm
# 1991 performance / 1992 salary: FA=free agent  FE=FA eligibility  A=arbitration  AE=A eligibility
base_read.table("baseball.dat",T,row.names="Name")
n_nrow(base)  # 337
#                Sal   Bat   OBP Run Hit Db Tr HR  RBI Wlk  SO SB Er FE FA AE A
# Andre Dawson   3300 0.272 0.302  69 153 21  4 31  104  22  80  4  3  1  0  0 0
# Steve Buchele  2600 0.269 0.335  58 111 17  2 18   66  39  69  0  3  1  1  0 0
# Kal Daniels    2500 0.249 0.337  54 115 15  1 17   73  63 116  6  5  1  0  0 0
par(mfrow=c(3,4), mai=c(0.5,0.5,0.1,0.1), oma=c(0.2,0.2,0.2,0.2))
for (i in 1:12) ndhist(base[,i], xlab=names(base)[i], cex=1.1)
pairs(base[,1:13])
```



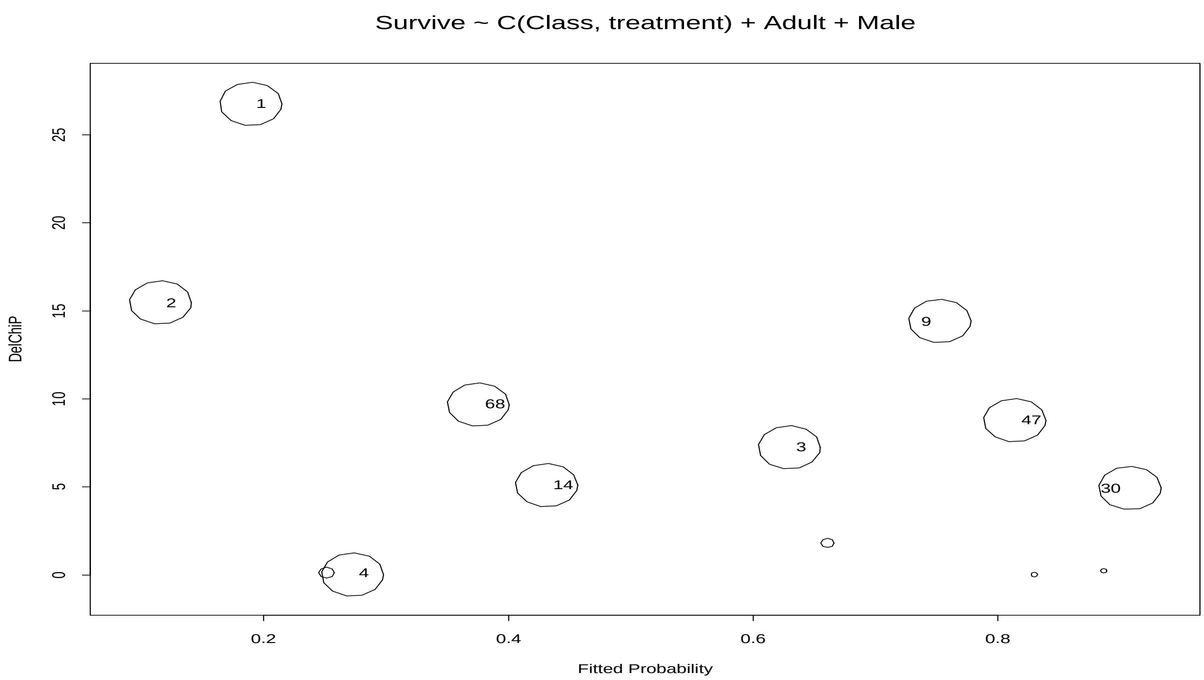


```
round(cor(base[,1:13]),2)
#      Sal  Bat  OBP  Run  Hit   Db   Tr   HR  RBI  Wlk   SO   SB   Er
# Sal 1.00 0.28 0.33 0.64 0.62 0.58 0.24 0.59 0.67 0.57 0.41 0.25 0.12
# Bat 0.28 1.00 0.81 0.44 0.51 0.45 0.26 0.21 0.37 0.28 0.08 0.20 0.15
# OBP 0.33 0.81 1.00 0.51 0.46 0.40 0.20 0.31 0.40 0.59 0.21 0.23 0.10
# Run 0.64 0.44 0.51 1.00 0.92 0.83 0.55 0.68 0.83 0.83 0.68 0.53 0.34
# Hit 0.62 0.51 0.46 0.92 1.00 0.89 0.54 0.61 0.85 0.72 0.65 0.43 0.42
#  Db 0.58 0.45 0.40 0.83 0.89 1.00 0.42 0.64 0.83 0.64 0.60 0.29 0.34
#  Tr 0.24 0.26 0.20 0.55 0.54 0.42 1.00 0.12 0.33 0.31 0.33 0.53 0.20
#  HR 0.59 0.21 0.31 0.68 0.61 0.64 0.12 1.00 0.88 0.62 0.75 0.08 0.15
# RBI 0.67 0.37 0.40 0.83 0.85 0.83 0.33 0.88 1.00 0.73 0.75 0.21 0.29
# Wlk 0.57 0.28 0.59 0.83 0.72 0.64 0.31 0.62 0.73 1.00 0.67 0.35 0.24
#  SO 0.41 0.08 0.21 0.68 0.65 0.60 0.33 0.75 0.75 0.67 1.00 0.28 0.30
#  SB 0.25 0.20 0.23 0.53 0.43 0.29 0.53 0.08 0.21 0.35 0.28 1.00 0.17
#  Er 0.12 0.15 0.10 0.34 0.42 0.34 0.20 0.15 0.29 0.24 0.30 0.17 1.00

collin(base[,2:17])
# Bat 0.875
# OBP 0.886
# Run 0.947
# Hit 0.951
# Db 0.823
# HR 0.895
# RBI 0.936
# Wlk 0.886
# SO 0.729

sw_stepwise(base[,2:17], base$Sal, plot=F, meth="exh", nbest=5)
sum.step(sw, base, "Sal", compact=T, top=5)
#           BIC   R2a Run HR RBI Wlk SO SB FE FA AE
# 6(#1) -2703.1 0.681   0  1   1   0  1  1  1  0  1
# 7(#1) -2703.2 0.685   0  1   1   0  1  1  1  1  1
# 6(#2) -2704.0 0.679   0  0   1   0  1  1  1  1  1
# 7(#2) -2704.4 0.683   1  1   1   0  1  1  1  0  1
# 7(#3) -2704.5 0.683   0  1   1   1  1  1  1  0  1

rslt_lm(Sal~HR+RBI+SO+SB+FE+AE, base)
summary(rslt)
# Residuals:     Min     1Q Median    3Q  Max
#              -1896 -455.6  24.14 329.1 3117
#
# Coefficients:    Value Std. Error    t value   Pr(>|t|)
# (Intercept)  -114.5452    88.4448    -1.2951     0.1962
#          HR    28.9651     9.4530     3.0641     0.0024
#         RBI    18.8733     2.9814     6.3304     0.0000
#          SO    -9.6720     1.8530    -5.2196     0.0000
#          SB    17.3368     3.5809     4.8415     0.0000
#          FE  1302.4209    92.6606    14.0558     0.0000
#          AE   810.6614   111.5695     7.2660     0.0000
#
# Residual standard error: 700.8 on 330 degrees of freedom
# Multiple R-Squared: 0.6863
#
# Correlation of Coefficients:
#     (Intercept)      HR     RBI      SO      SB      FE
#  HR  0.3993
# RBI -0.3185     -0.7222
#  SO -0.4831     -0.3448 -0.2161
#  SB  0.0351      0.2853 -0.1871 -0.2798
#  FE -0.2719     -0.0172 -0.2009  0.0608 -0.0432
#  AE -0.2216      0.0882 -0.2480  0.0784 -0.0756  0.4622

par(mfrow=c(3,3), oma=c(0.5,0.5,1.5,0.5), mai=c(0.7,0.7,0.5,0.5))
qqn(rslt$res, ylab="Residual", main="")
```

```
mtext("\nSalary~HR+RBI+SO+SB+FAElig+ArbElig", outer=T, cex=1.2)
ndhist(rslt$res, ylab="Residual")
partreg(c("HR","RBI","SO","SB","FE","AE"), "Sal", "HR", base, T, id=T, tcex=0.55)
partreg(c("HR","RBI","SO","SB","FE","AE"), "Sal", "RBI", base, T, id=T, tcex=0.55)
partreg(c("HR","RBI","SO","SB","FE","AE"), "Sal", "SO", base, T, id=T, tcex=0.55)
partreg(c("HR","RBI","SO","SB","FE","AE"), "Sal", "SB", base, T, id=T, tcex=0.55)
partreg(c("HR","RBI","SO","SB","FE","AE"), "Sal", "FE", base, T, id=T, tcex=0.55)
partreg(c("HR","RBI","SO","SB","FE","AE"), "Sal", "AE", base, T, id=T, tcex=0.55)
CookPlot(rslt, id=T, icex=0.7, tcex=0.55)
ps("base3")
```



```
# http://www.amstat.org/publications/jse/archive.htm
titan_read.table("titanic.dat",T)
nrow(titan) # 2201
titan[1:5,]
#   Class Adult Male Survive
# 1     1     1    1       1
# 2     1     1    1       1
# 3     1     1    1       1
# 4     1     1    1       1
# 5     1     1    1       1
levels(titan$Class)  # "1" "2" "3" "C"
jnk_table(titan$Class,titan$Survive)
100*round(jnk/apply(jnk,1,sum),3)
#      0    1
# 1 37.5 62.5
# 2 58.6 41.4
# 3 74.8 25.2
# C 76.0 24.0

table(titan$Adult,titan$Male,titan$Class)
, , 1       , , 2      , , 3       , , C
    0   1      0   1       0   1      0   1
0   1   5   0 13  11   0  31  48   0  0   0
1 144 175   1 93 168   1 165 462   1 23 862

rslt_glm(Survive~C(Class,treatment)+Adult+Male, data=titan, family="binomial")
summary(rslt)
# Coefficients:             Value Std. Error    t value
#          (Intercept)  3.1053779  0.2980741  10.418140
# C(Class, treatment)2 -1.0180945  0.1959554  -5.195543
# C(Class, treatment)3 -1.7777609  0.1715061 -10.365587
# C(Class, treatment)C -0.8576762  0.1573238  -5.451661
#                Adult -1.0615415  0.2439637  -4.351227
#                 Male -2.4200593  0.1403600 -17.241796

round(exp(rslt$coef[2:6]),2)
# C(Class,treatment)2 C(Class,treatment)3 C(Class,treatment)C  Adult Male
#                0.36                0.17                0.42   0.35 0.09

exp(rslt$coef[1])/(1+exp(rslt$coef[1]))  # 0.957

sum(titan$Class=="3"&titan$Adult==0&titan$Male==0&titan$Survive==1)/sum(titan$Class=="3"&titan$Adult==0&titan$Male==0)  # 0.45  (odds=0.81)

sum(titan$Class=="3"&titan$Adult==0&titan$Male==1&titan$Survive==1)/sum(titan$Class=="3"&titan$Adult==0&titan$Male==1)  # 0.27  (odds=0.37)


titanS_titan[sample(1:nrow(titan),1000),]
rsltS_glm(Survive~C(Class,treatment)+Adult+Male, data=titanS, family="binomial")
infS_logitinf(rsltS, titanS[,1:3], titanS$Survive)
infS$ChiP # 29.82
Bplot(rsltS, titanS[,1:3], titanS$Survive, inf=infS, id=T)
```




```
cbind(infS$coll[,1:5],round(infS$coll[,6:8],2),DelB=round(infS$DelB,2),
      DelChiP=round(infS$DelChiP,2))
# Pat Class Adult Male   Y   m  fit  obs  hat  DelB DelChiP
# 141     1     0    1   2   2 0.66 1.00 0.09  0.11    1.13
#  30     1     1    0  58  60 0.91 0.97 0.35  2.07    3.82
#  14     1     1    1  29  77 0.43 0.38 0.81 22.18    5.04
# 420     2     0    0   1   1 0.89 1.00 0.02  0.00    0.13
#  68     2     0    1   4   4 0.38 1.00 0.19  1.93    8.22
#   9     2     1    0  35  38 0.75 0.92 0.49 11.23   11.47
#   1     2     1    1   5  74 0.19 0.07 0.66 41.76   21.23
#  47     3     0    0   6  12 0.81 0.50 0.32  5.27   11.41
#  13     3     0    1   5  18 0.25 0.28 0.53  0.16    0.14
#   3     3     1    0  42  77 0.63 0.55 0.69 16.19    7.44
#   2     3     1    1  36 227 0.12 0.16 0.76 53.38   17.03
#  26     C     1    0  11  13 0.83 0.85 0.10  0.00    0.03
#   4     C     1    1 108 397 0.27 0.27 0.98  1.28    0.03
```

```
notcrew_titan[titan$Class!="C",]
notcrew$Class_factor(as.character(notcrew$Class))  # drop Crew as a "level"
rslt123_glm(Survive~C(Class,treatment)*Adult+C(Class,treatment)*Male+Adult*Male,
            data=notcrew, family="binomial")
summary(rslt123)
# Coefficients:                    Value Std. Error     t value
#               (Intercept)  10.52790654  8.7005865  1.21002263
#      C(Class, treatment)2  -0.82655263  9.5485134 -0.08656349
#      C(Class, treatment)3 -10.72206411  8.6931873 -1.23338699
#                     Adult  -6.97255855  8.6859626 -0.80273873
#                      Male  -3.59618720  0.7476558 -4.80995021
# C(Class, treatment)2Adult  -0.91171913  9.5308194 -0.09566010
# C(Class, treatment)3Adult   7.00881338  8.6771613  0.80773113
#  C(Class, treatment)2Male   0.06801075  0.6711139  0.10134010
#  C(Class, treatment)3Male   2.79994731  0.5686955  4.92345637
#                Adult:Male  -0.68679415  0.5252542 -1.30754616
#     Null Deviance: 1746.756 on 1315 degrees of freedom
# Residual Deviance: 1165.414 on 1306 degrees of freedom

rslt123B_glm(Survive~Adult+Adult:Male+C(Class,treatment)*Male,
            data=notcrew, family="binomial")
summary(rslt123B)
# Coefficients:(Intercept)  3.7414820  0.6178101  6.0560384
#                    Adult -0.1803451  0.3617844 -0.4984878
#     C(Class, treatment)2 -1.6143129  0.5842988 -2.7628205
#     C(Class, treatment)3 -3.7534231  0.5253327 -7.1448496
#                     Male -2.8919449  0.6920754 -4.1786558
#               Adult:Male -1.3581169  0.4550919 -2.9842697
# C(Class, treatment)2Male  0.3384202  0.6445041  0.5250863
# C(Class, treatment)3Male  2.6654869  0.5630967  4.7336218
#
#     Null Deviance: 1746.756 on 1315 degrees of freedom
# Residual Deviance: 1202.62 on 1308 degrees of freedom


Response: Survive

1165.41-1202.62  # 37.20
qchisq(0.95,2) # 5.99
```