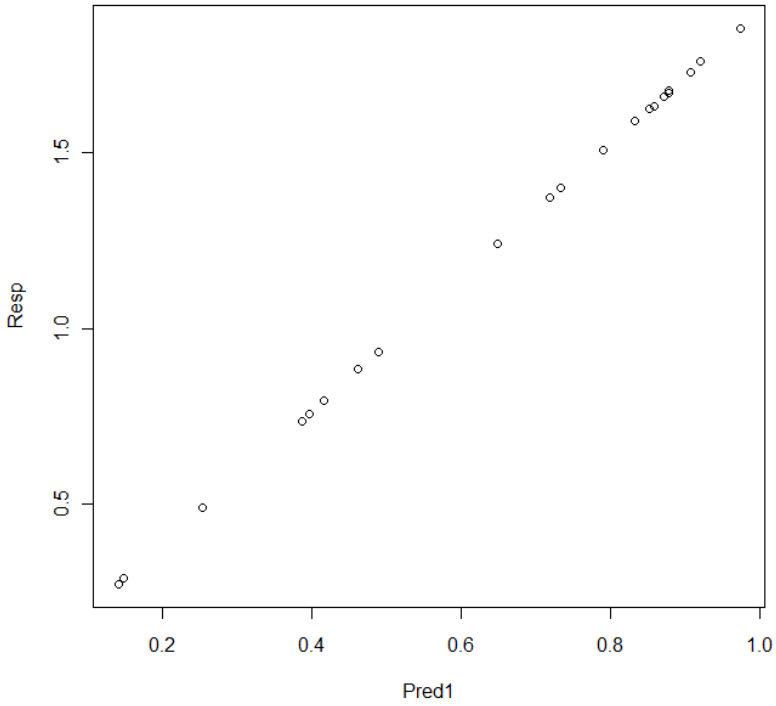
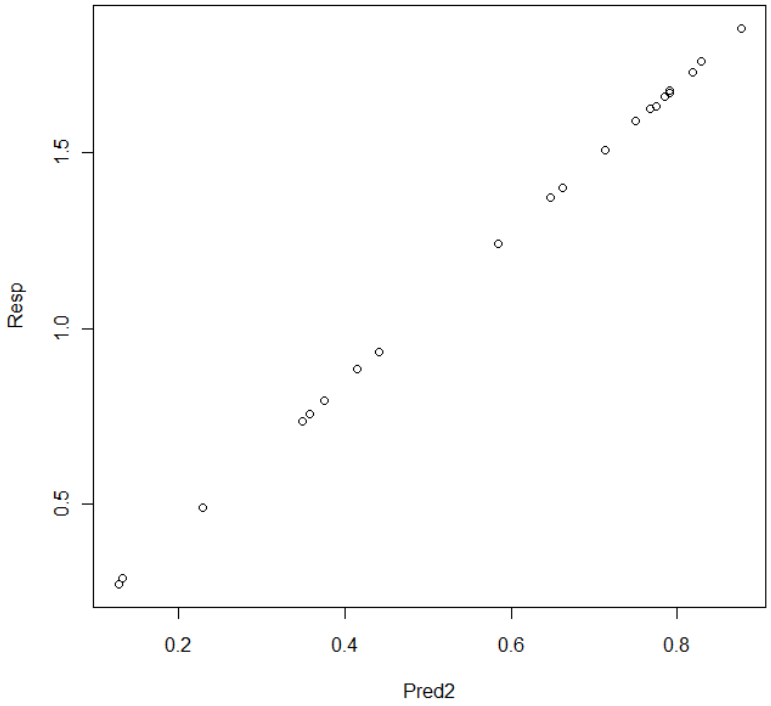
MULTICOLLINEARITY AND SAMPLE VS POPULATION EXAMPLE

multicollinearity.csv, see wattle

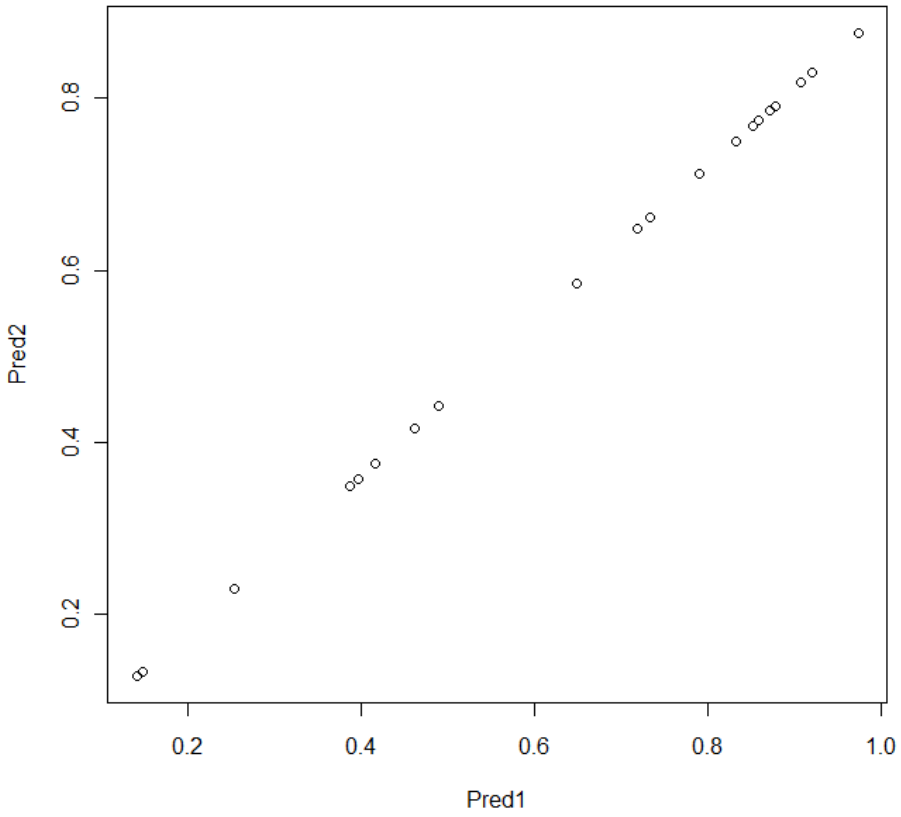
```
>  multi<-read.csv("multicollinearity.csv")
>  attach(multi)
The following object(s) are masked from 'multi (position 3)':

    Pred1, Pred2, Resp
>  names(multi)
[1] "Pred1" "Pred2" "Resp"
> plot(Pred1,Pred2)
> plot(Pred1,Resp)
> plot(Pred2,Resp)
```

Response against predictors





Predictor1 against Predictor2



Correlation and VIF's

```
> cor(cbind(Pred1,Pred2,Resp))
          Pred1     Pred2      Resp
Pred1 1.0000000 0.9999991 0.9999882
Pred2 0.9999991 1.0000000 0.9999866
Resp  0.9999882 0.9999866 1.0000000

> diag(solve(cor(cbind(Pred1,Pred2))))
   Pred1    Pred2
568178.8 568178.8
>
```

Order of fit doesn't matter – T, Coeff, P, Std.Err

```
> lsmulti<-lsfit(cbind(Pred1,Pred2), Resp)
> ls.print(lsmulti)
Residual Standard Error=0.0026
R-Square=1
F-statistic (df=2, 18)=386377
p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept   0.0043  0.0016  2.6599   0.0159
Pred1       2.7032  1.6308  1.6576   0.1147
Pred2      -0.8902  1.8115 -0.4914   0.6291

>
```

```
> lsmulti2<-lsfit(cbind(Pred2,Pred1),Resp)
> ls.print(lsmulti2)
Residual Standard Error=0.0026
R-Square=1
F-statistic (df=2, 18)=386377
p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept   0.0043  0.0016  2.6599   0.0159
Pred2      -0.8902  1.8115 -0.4914   0.6291
Pred1       2.7032  1.6308  1.6576   0.1147
```

Simple Linear Regression

```
> lsmulti3<-lsfit(Pred1,Resp)
> ls.print(lsmulti3)
Residual Standard Error=0.0026
R-Square=1
F-statistic (df=1, 19)=804886.6
p-value=0

          Estimate Std.Err  t-value Pr(>|t|)
Intercept   0.0040  0.0015   2.7188   0.0136
X           1.9018  0.0021 897.1547   0.0000
```

```
> lsmulti4<-lsfit(Pred2,Resp)
> ls.print(lsmulti4)
Residual Standard Error=0.0027
R-Square=1
F-statistic (df=1, 19)=707658
p-value=0

          Estimate Std.Err  t-value Pr(>|t|)
Intercept   0.0033  0.0016   2.1137    0.048
X           2.1126  0.0025 841.2241    0.000
```

Both predictors appear to be significant!

Sequential Sums of Squares

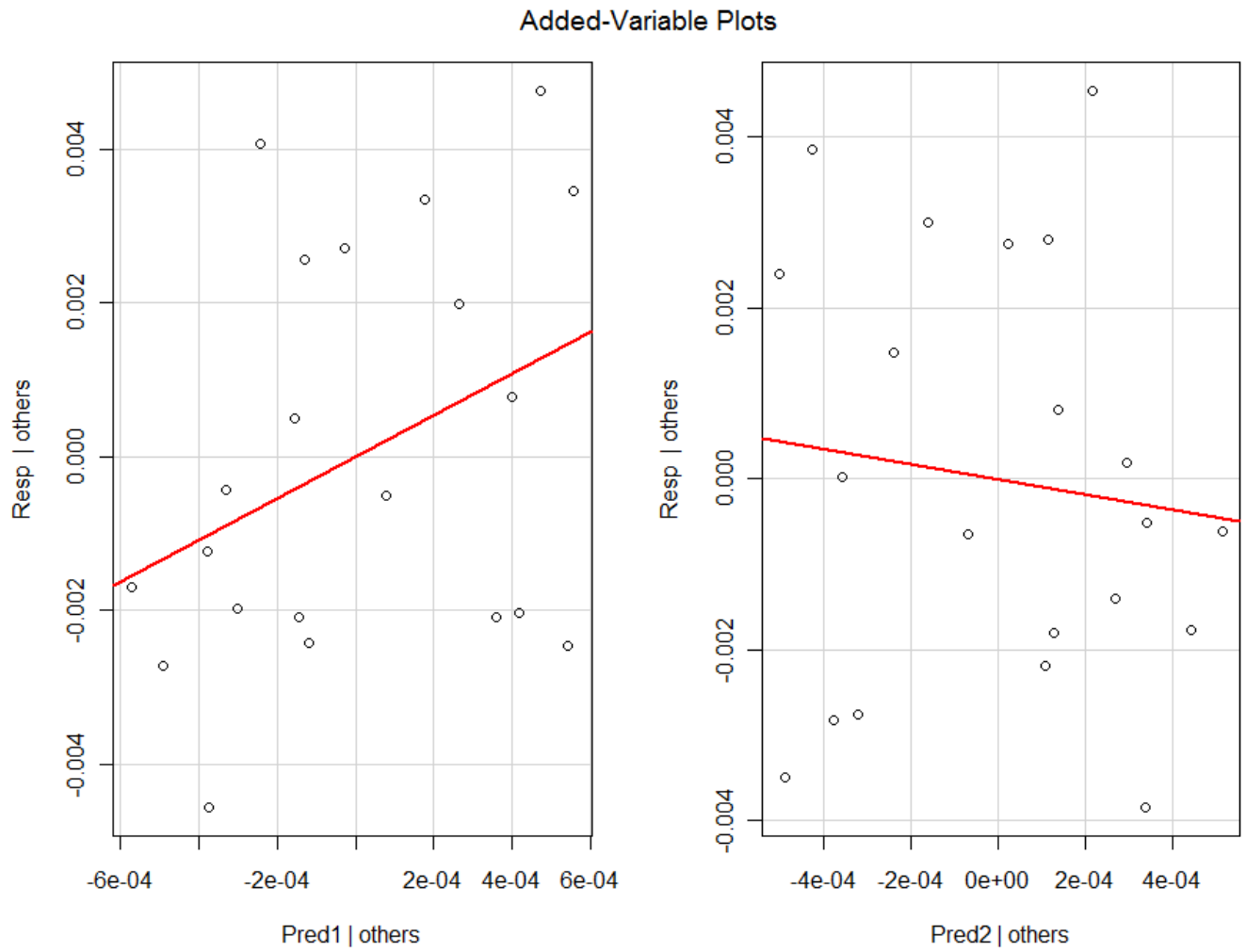
```
> lm1<-lm(Resp~Pred1+Pred2)
> anova(lm1)
Analysis of Variance Table

Response: Resp
          Df Sum Sq Mean Sq    F value Pr(>F)
Pred1      1 5.3056  5.3056 7.7275e+05 <2e-16 ***
Pred2      1 0.0000  0.0000 2.4150e-01 0.6291
Residuals 18 0.0001  0.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
    ‘.’ 0.1 ‘ ’ 1
```

```
> lm2<-lm(Resp~Pred2+Pred1)
> anova(lm2)
Analysis of Variance Table

Response: Resp
          Df Sum Sq Mean Sq    F value Pr(>F)
Pred2      1 5.3056  5.3056 7.7275e+05 <2e-16 ***
Pred1      1 0.0000  0.0000 2.7477e+00 0.1147
Residuals 18 0.0001  0.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
    ‘ ’ 1
```

Q1: Which predictor is significant? That is, calculate $SSR(\beta_1|\beta_0,\beta_2)$ and calculate $SSR(\beta_2|\beta_0,\beta_1)$ and then determine each partial F ratio. What do you notice?

Q2: Calculate the T stats for each predictor

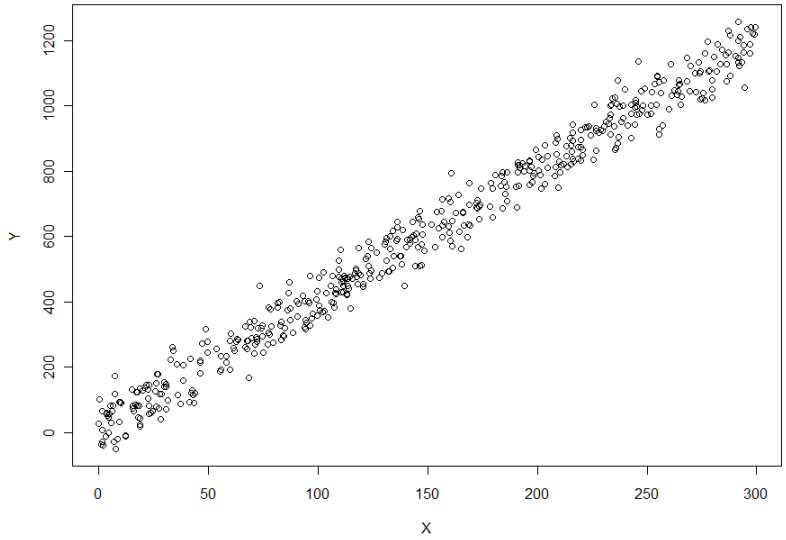
Homework. Construct added variable plots

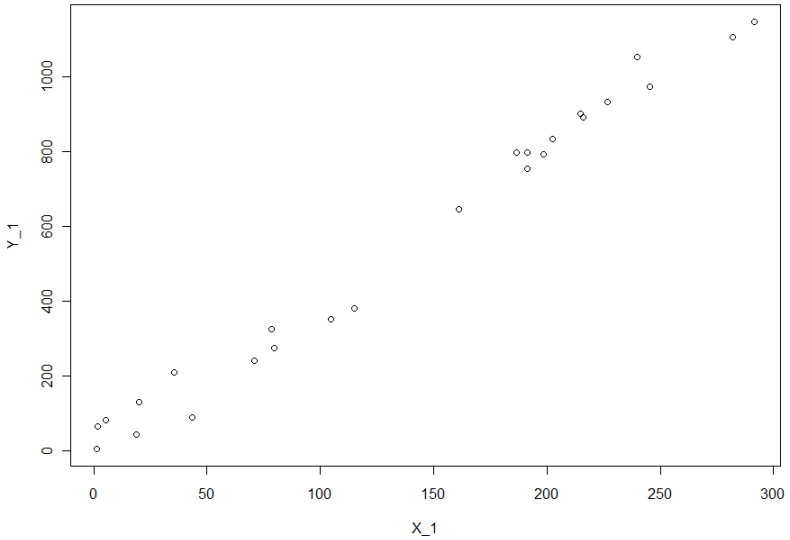


Population vs Sample

```
> svp<-read.csv("svp.csv")
>  attach(svp)
> names(svp)
[1] "X"   "Y"   "X.1" "X_1" "Y_1" "X.2" "X_2" "Y_2"
>
```

Plot Population vs Sample



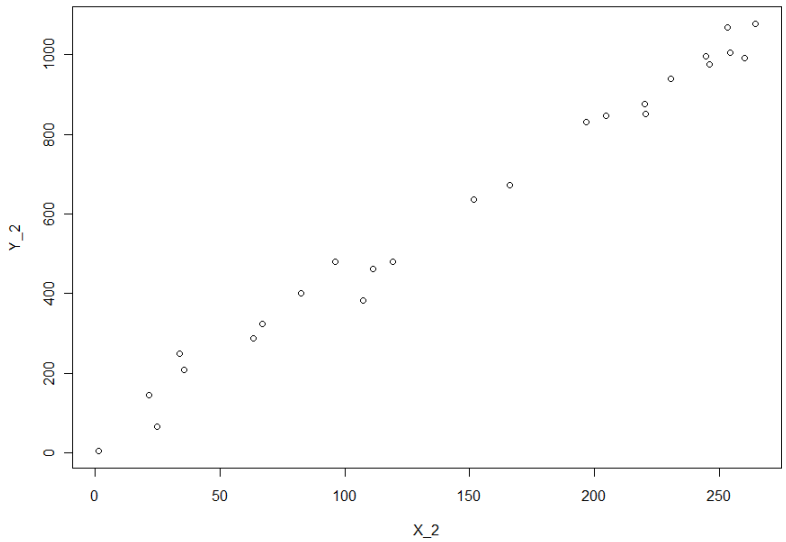


Fit of Population vs Sample

```
> poplm<-lm(Y~X)
> poplm

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
     14.588        3.964

> anova(poplm)
Analysis of Variance Table

Response: Y
           Df   Sum Sq  Mean Sq F value    Pr(>F)
X           1 59237383 59237383   25472 < 2.2e-16 ***
Residuals 497  1155827     2326
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
     0.1 ‘ ’ 1
```

```
> samlm<-lm(Y_1~X_1)
> samlm

Call:
lm(formula = Y_1 ~ X_1)

Coefficients:
(Intercept)          X_1
     0.5838       4.0338

> anova(samlm)
Analysis of Variance Table

Response: Y_1
          Df  Sum Sq Mean Sq F value    Pr(>F)
X_1        1 3542410 3542410  1533.8 < 2.2e-16 ***
Residuals 23   53119    2310
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
     0.1 ‘ ’ 1
```

Sample 2




```
> sam2lm<-lm(Y_2~X_2)
> sam2lm

Call:
lm(formula = Y_2 ~ X_2)

Coefficients:
(Intercept)          X_2
     47.965        3.821

> anova(sam2lm)
Analysis of Variance Table

Response: Y_2
          Df  Sum Sq Mean Sq F value    Pr(>F)
X_2        1 2827027 2827027  1796.6 < 2.2e-16 ***
Residuals 23   36192    1574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
    ‘.’ 0.1 ‘ ’ 1
```