Example of constructing an Added Variable Plot

- Used when adding a variable X_2 to a model which already contains one or more variables in $\boldsymbol{X}_1$.
- Uses a a sequential three step procedure.

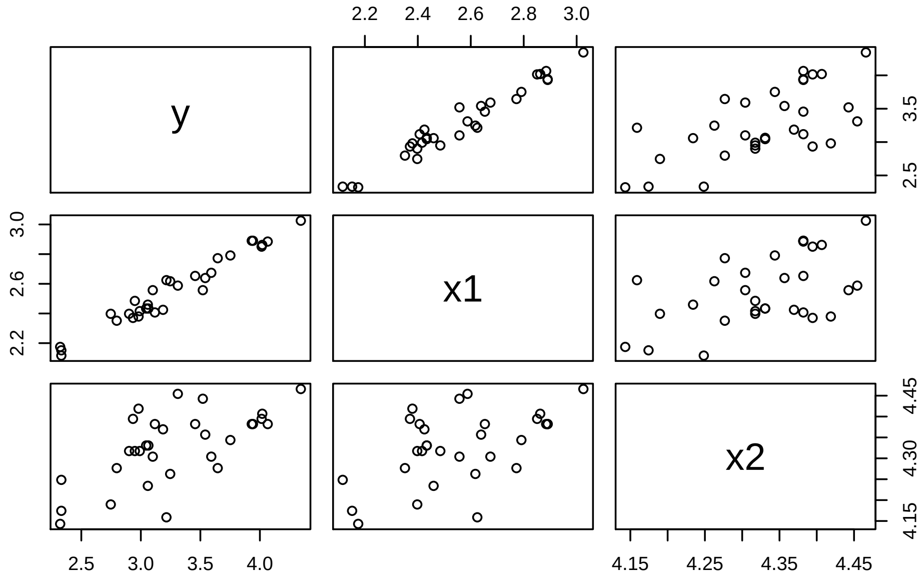
  1. Regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$ to get residuals $\boldsymbol{e}_1$.
  2. Regress X_2 on $\boldsymbol{X}_1$ to get residuals $\boldsymbol{e}_2$
  3. A third regression of $\boldsymbol{e}_1$ on $\boldsymbol{e}_2$ has intercept 0, estimated slope equal to the coefficient of β_2 in the regression $lm(y \sim x_1 + x_2)$, and p-value equal to that for β_2 in $anova(lm(y \sim x_1 + x_2))$

- The scatterplot of $\boldsymbol{e}_1$ vs $\boldsymbol{e}_2$ is called an added variable plot.
- It is used the help decide what function of X_2 should be added to the regression (ie linear, quadratic, etc), in the same way that a scatterplot of y vs X_2 would be used.
- it captures the marginal relationship between y and X_2 given that X_1 is already accounted for.

Example: trees data

- for a cylinder of diameter d, and height h, the volume equals $\pi(d/2)^2 h$, which motivates fitting a transformed model ($y \sim x_1 + x_2$) where $y = log(Volume)$, $x_1 = log(Girth)$ and $x_2 = log(Height)$.
- Regress y on x_1 to get residuals e_1.

```
>   attach(trees)
>   y=log(Volume)
>   x1=log(Girth)
>   x2=log(Height)
>   pairs(cbind(y,x1,x2))
>   lm.x1=lm(y~x1)
>   summary(lm.x1)

Call:
lm(formula = y ~ x1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.205999 -0.068702  0.001011  0.072585  0.247963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.35332    0.23066  -10.20 4.18e-11 ***
x1           2.19997    0.08983   24.49  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.115 on 29 degrees of freedom
Multiple R-squared:  0.9539,        Adjusted R-squared:  0.9523
F-statistic: 599.7 on 1 and 29 DF,  p-value: < 2.2e-16
```



- Note the coefficient of x_1 from the regression.

```
>   anova(lm.x1)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 7.9254  7.9254  599.72 < 2.2e-16 ***
Residuals 29 0.3832  0.0132
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Note the regression and error sum of squares for future comparison.

- should x_2 be added to the model, and if so, should a linear function of x_2 be used?
  - calculate the residuals e_1
  - calculate the residuals e_2
  - plot e_1 vs e_2 to deduce the form of the relationship

```
>   e1=residuals(lm.x1)
>   summary(lm.x1)

Call:
lm(formula = y ~ x1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.205999 -0.068702  0.001011  0.072585  0.247963 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.35332    0.23066  -10.20 4.18e-11 ***
x1           2.19997    0.08983   24.49  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.115 on 29 degrees of freedom
Multiple R-squared:  0.9539,      Adjusted R-squared:  0.9523 
F-statistic: 599.7 on 1 and 29 DF,  p-value: < 2.2e-16

>   lm.x2=lm(x2~x1)
>   summary(lm.x2)

Call:
lm(formula = x2 ~ x1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.181448 -0.037403  0.007068  0.031853  0.126194 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.82974    0.14833  25.819  < 2e-16 ***
x1           0.19454    0.05777   3.367  0.00216 ** 
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07393 on 29 degrees of freedom
Multiple R-squared:  0.2811,       Adjusted R-squared:  0.2563
F-statistic: 11.34 on 1 and 29 DF,  p-value: 0.002155
```

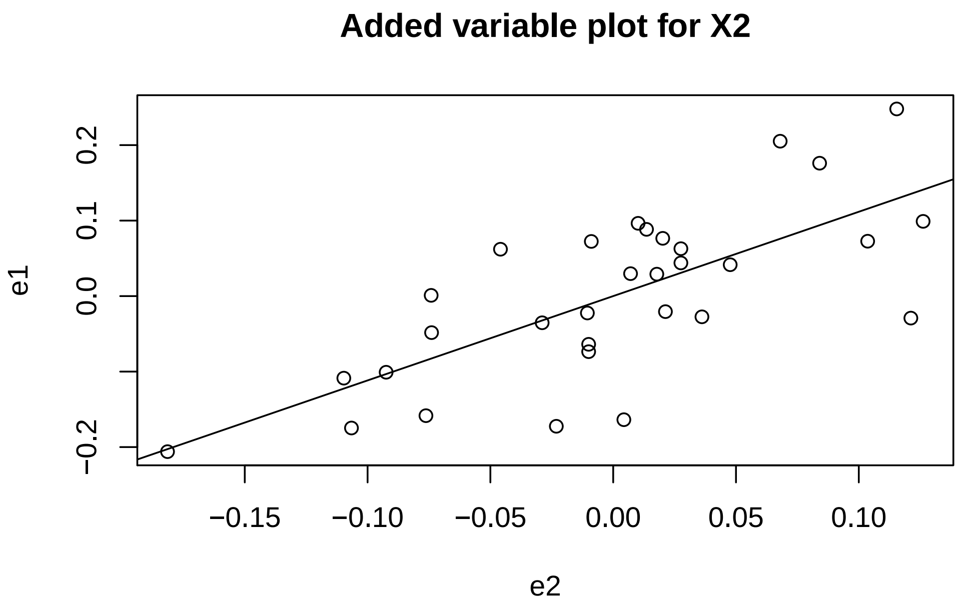
```
>   e2=residuals(lm.x2)
>   lm.e1e2=lm(e1~e2-1)
>   summary(lm.e1e2)
```

```
Call:
lm(formula = e1 ~ e2 - 1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
e2   1.1171     0.1975   5.656 3.66e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07863 on 30 degrees of freedom
Multiple R-squared:  0.5161,       Adjusted R-squared:  0.4999
F-statistic: 31.99 on 1 and 30 DF,  p-value: 3.658e-06
```

```
>   anova(lm.e1e2)
```

```
Analysis of Variance Table

Response: e1
          Df  Sum Sq  Mean Sq F value    Pr(>F)
e2         1 0.19778 0.197780  31.992 3.658e-06 ***
Residuals 30 0.18546 0.006182
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
>   plot(e2,e1,main="Added variable plot for X2")
>   abline(lm.e1e2)
```



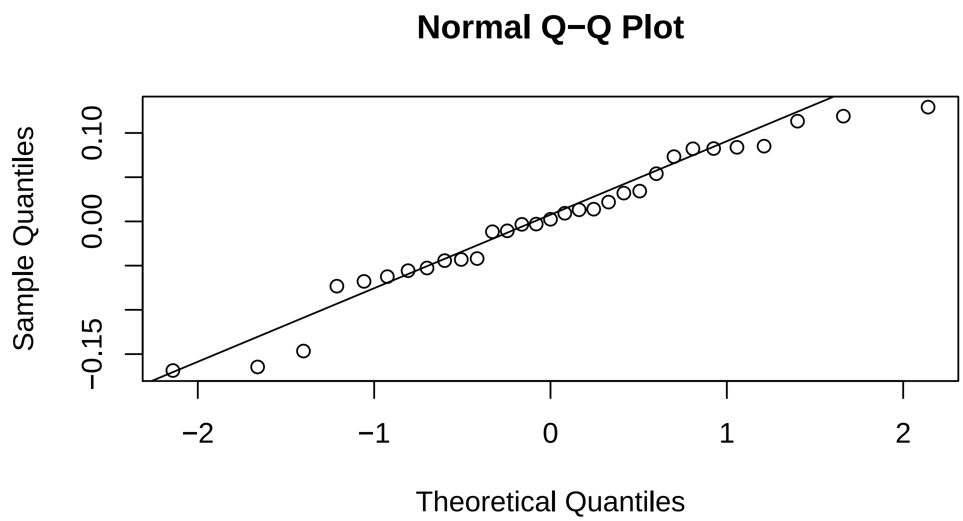
- From the added variable plot, a linear term in X_2 seems appropriate. If the added variable plot showed a quadratic trend, that would suggest adding a predictor X_2^2 in addition to the predictor X_1.
- Note the coefficient of X_2 in the regression of e_1 on e_2.
- Verify that if we substitute for e_1 and e_2 in $e_1 = \hat{\alpha} e_2$ then we recover the least squares estimates for the full model including both x_1 and x_2.

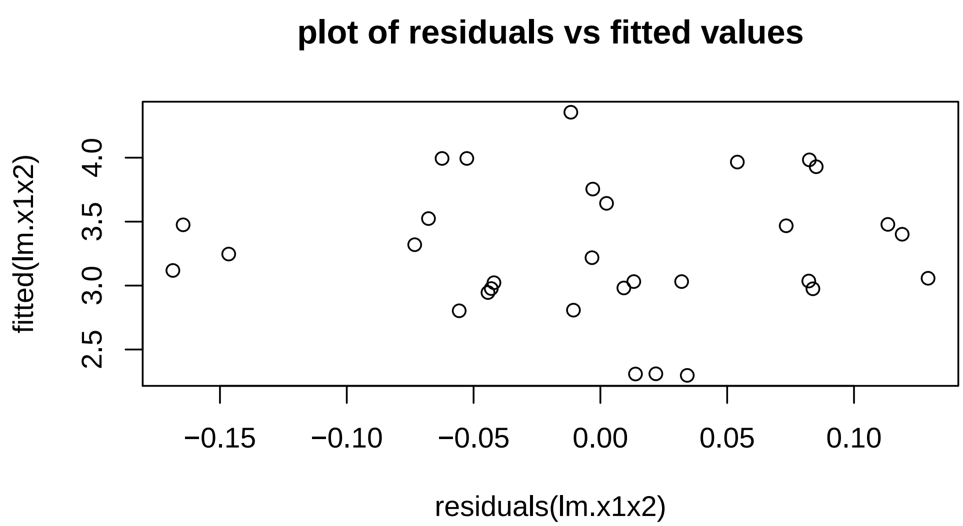
```
>   lm.x1x2=lm(y~x1+x2)
>   summary(lm.x1x2)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
x1           1.98265    0.07501  26.432  < 2e-16 ***
x2           1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,        Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16

>   anova(lm.x1x2)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 7.9254  7.9254 1196.53 < 2.2e-16 ***
x2         1 0.1978  0.1978   29.86 7.805e-06 ***
Residuals 28 0.1855  0.0066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The error sum of squares and degrees of freedom for the single variable regression have been partitioned into a new error sum of squares (with one less degree of freedom) and a sequential sum of squares $S(\beta_2|\beta_1)$ for X_2 given that X_1 is already in the model, this having one degree of freedom.

– The added variable plot suggested a linear function of X_2 to be included, but we still need to assess overall model adequacy.

```
>   par(mfrow=c(2,1))
>   qqnorm(residuals(lm.x1x2))
>   qqline(residuals(lm.x1x2))
>   plot(residuals(lm.x1x2),fitted(lm.x1x2),main="plot of residu
```





– residuals appear normally distributed

– no evidence of a trend in plot of residuals vs fitted values

– no suggestion from he latter plot that variance changes with the mean

CI for the mean of y using "predict"

For the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, find a 95% CI for the mean of y when $x_1 = log(10)$ and $x_2 = log(75)$

```
> predict.out=predict(lm.x1x2,
+        newdata=data.frame(x1=log(10), x2=log(75)),
+        interval="confidence")
> predict.out
```

```
      fit      lwr      upr
1 2.75677 2.709063 2.804477
```