228.371 - Statistical Modelling for Engineers and Technologists

Week 4. Multiple Regression and ANOVA

Dr. Daniel Walsh
IIMS 3.07 x 41032
d.c.walsh@massey.ac.nz

Semester One - 2015

The General Linear Model

- The models dealt with so far only include one explanatory variable. We will now extend the model to include more than one explanatory variable.
- We will use a matrix notation because it simplifies the written form of the model considerably.

Matrix Notation for the Simple Linear Model

- The Simple Linear model can be written:

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad \text{where} \quad i = 1, \ldots, n.$$

- In matrix notation it is written:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- The residual sum of squares is:

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$$

Matrix Notation

- Including additional variables does not change the matrix representation.
- The model with two explanatory variables is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \text{where} \quad i = 1, \ldots, n.$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{(same as before)}$$

- The Least Squares Solution is:

$$\boldsymbol{b} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

Assumptions

- The x_i are recorded without error.
- The relationship between the variables is linear.
- The variance of the response is constant.
- The errors are uncorrelated.
- The distribution of the errors is normal.

Example: `trees`: Model Volume of Trees

The volume (v) of a cylinder of height h and radius r (girth $g = 2\pi r$) is:

$$v = h \times \pi r^2 = \frac{hg^2}{4\pi}$$

Taking logs gives us:

$$\begin{aligned}\log(v) &= -\log(4\pi) + \log(h) + 2\log(g)\\ \log(v_i) &= \beta_0 + \beta_1 \log(h_i) + \beta_2 \log(g_i) + \epsilon_i\end{aligned}$$

Example: trees: Model Volume of Trees

```
data(trees)
ltrees <- log(trees)
names (ltrees) <- c("logGirth", "logHeight","logVolume")
head (ltrees)
```

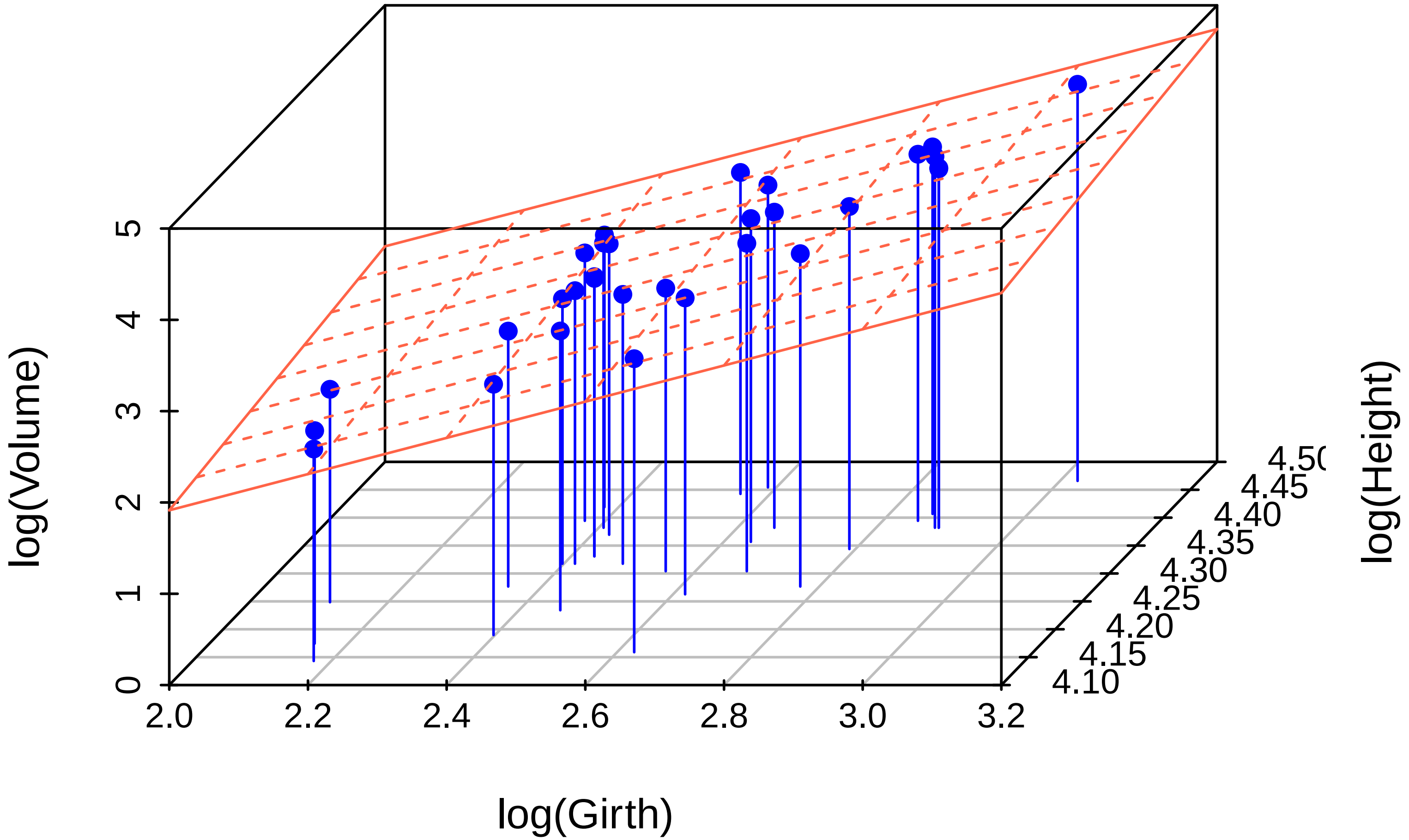
```
  logGirth logHeight logVolume
1 2.116256  4.248495  2.332144
2 2.151762  4.174387  2.332144
3 2.174752  4.143135  2.322388
4 2.351375  4.276666  2.797281
5 2.370244  4.394449  2.933857
6 2.379546  4.418841  2.980619
```

```
m1 <- lm (logVolume ~ logHeight + logGirth, data=ltrees)
cbind(coef(m1), confint (m1))
```

```
                          2.5 %    97.5 %
(Intercept) -6.631617 -8.269912 -4.993322
logHeight    1.117123  0.698353  1.535894
logGirth     1.982650  1.828998  2.136302
```

Example: `trees`: Visualising the Model

$$\log(\hat{v}_i) = -6.63 + 1.12\log(h_i) + 1.98\log(g_i)$$



Example: `trees`: Visualising the Model

- The regression model is a 2-d plane in 3-d space.
- It is defined by two simple linear regression lines in the $y{:}x_1$ ($v{:}h$) and $y{:}x_2$ ($v{:}g$) planes. In this case:

$$\log(\hat{v}_i) = -6.63 + 1.12 \log(h_i) + 1.98 \log(g_i)$$

is defined by the plane through:

$$\begin{aligned} \log(\hat{v}_i) &= -6.63 + 1.12 \log(h_i) \qquad (\text{when } \log(g_i) = 0) \\ \log(\hat{v}_i) &= -6.63 + 1.98 \log(g_i) \qquad (\text{when } \log(h_i) = 0) \end{aligned}$$

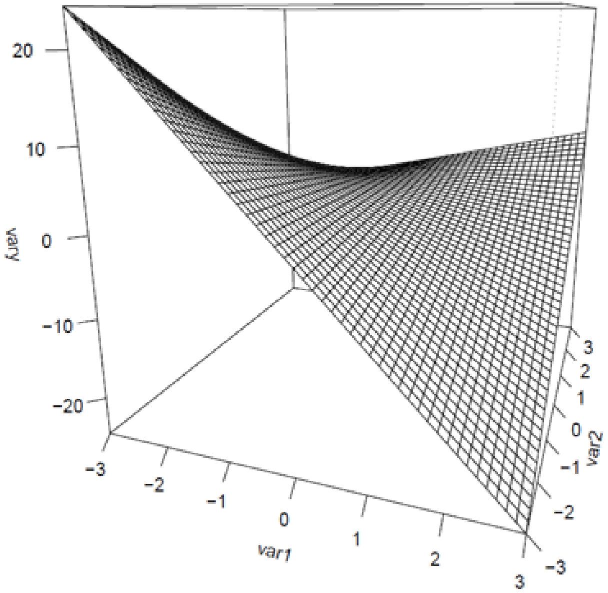
Example: `trees`: Interaction Term

We can make the regression model more flexible by adding a term to model interaction between the predictor variables.

Interaction terms like this (both variables continuous) are hard to interpret!

$$\log(v_i) = \beta_0 + \beta_1 \log(h_i) + \beta_2 \log(g_i) + \beta_{12}(\log(h_i) \times \log(g_i)) + \epsilon_i$$

β_0	Intercept
β_1, β_2	Main effects
β_{12}	Interaction term



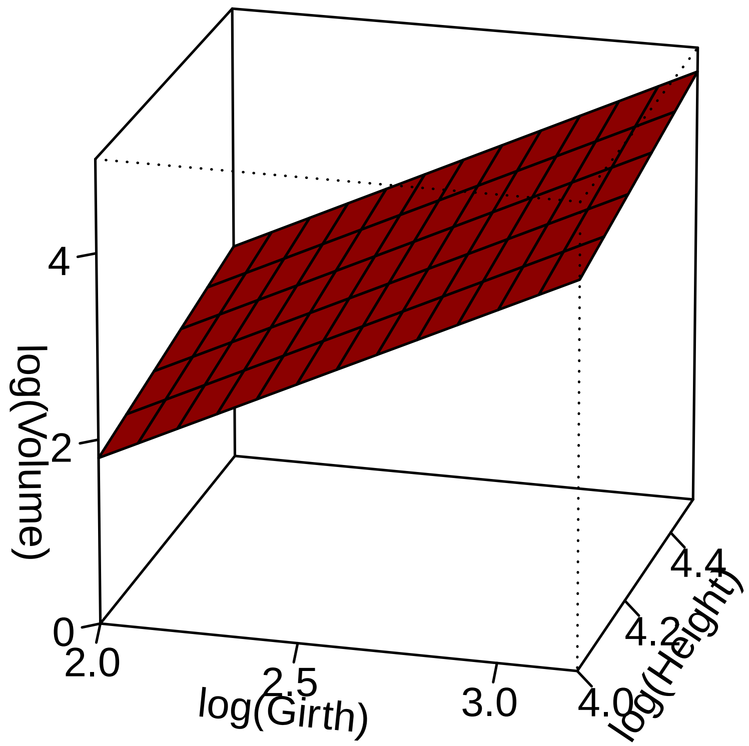
Example: trees: Interaction Term

We can fit the interaction term in different ways. These give the same results with only two predictors.

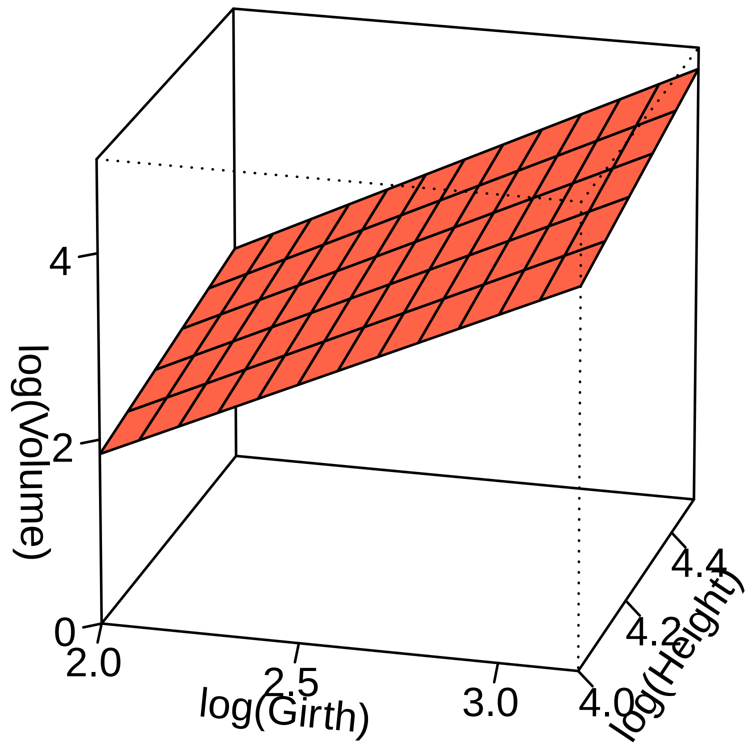
```
m2 <- lm (logVolume ~  logHeight + logGirth + logHeight:logGirth, data=ltrees)
m2 <- lm (logVolume ~  logHeight * logGirth,    data=ltrees)
m2 <- lm (logVolume ~ (logHeight + logGirth)^2, data=ltrees)
summary(m2)
```

```
Call:
lm(formula = logVolume ~ (logHeight + logGirth)^2, data = ltrees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.165941 -0.048613  0.006384  0.062204  0.132295

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.6869     7.6996  -0.479    0.636
logHeight            0.4377     1.7788   0.246    0.808
logGirth             0.7942     3.0910   0.257    0.799
logHeight:logGirth   0.2740     0.7124   0.385    0.704

Residual standard error: 0.08265 on 27 degrees of freedom
Multiple R-squared:  0.9778, Adjusted R-squared:  0.9753
F-statistic: 396.4 on 3 and 27 DF,  p-value: < 2.2e-16
```

Example: `trees`: Comparison of Surfaces



No interaction - flat plane



Interaction term - Curved surface

- These surfaces look the same - what is going on?

Example: `trees`: Model Summaries

```
 summary(m1)
Call:
lm(formula = logVolume ~ logHeight + logGirth, data = ltrees)

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 ***
logHeight    1.11712    0.20444   5.464 7.81e-06 ***
logGirth     1.98265    0.07501  26.432  < 2e-16 ***
---
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
```

```
 summary(m2)
Call:
lm(formula = logVolume ~ (logHeight + logGirth)^2, data = ltrees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.165941 -0.048613  0.006384  0.062204  0.132295

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.6869     7.6996  -0.479    0.636
logHeight            0.4377     1.7788   0.246    0.808
logGirth             0.7942     3.0910   0.257    0.799
logHeight:logGirth   0.2740     0.7124   0.385    0.704

Residual standard error: 0.08265 on 27 degrees of freedom
Multiple R-squared:  0.9778, Adjusted R-squared:  0.9753
F-statistic: 396.4 on 3 and 27 DF,  p-value: < 2.2e-16
```

- The interaction coefficient is not significant (p-value $= 0.704$).
- Adjusted-R^2 is lower for interaction model.
- None of the coefficients are significant in the interaction model - but the F-test is significant.
- Interaction model does not seem worth it.
- **Note:** *Do not judge p-values jointly but conditionally.*

Example: trees: Model Comparison

The ANOVA test is the same as testing if interaction coefficient is zero.

```
 anova (m1, m2)
```

```
Analysis of Variance Table

Model 1: logVolume ~ logHeight + logGirth
Model 2: logVolume ~ (logHeight + logGirth)^2
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     28 0.18546
2     27 0.18445  1 0.0010105 0.1479 0.7035
```

Variable selection

- Do not judge p-values **jointly** but **conditionally**.
- If a particular coefficients p-value is high (> 0.05) - we can say that this variable is not significant given the other variables in the model (whether they are significant or not).
- If an interaction term is in the model - we should include the corresponding main effects in the model.
- Therefore remove higher order interactions first before removing main effects.
- Adding variables to a model can change the significance of other variables already in the model.
- Scientific reasons can keep non-significant variables in a model.

Multiple regression

Examine relationship between predictor variables and response variable:

- Scatterplots using `pairs(`x_1`,`x_2`,...,`x_p`,y)`.
- Correlations using `cor.test(`x_i`,`x_j`)` and `cor.test(`x_i`,y)`.

Correlation matrix

```
horseshearts <- read.table("Data/horseshearts.txt", header=TRUE)
attach(horseshearts)
```

```
round(cor(horseshearts), 3)
```

```
         innersys innerdia outersys outerdia extsys extdia weight
innersys    1.000    0.909    0.825    0.756  0.807  0.807  0.778
innerdia    0.909    1.000    0.772    0.699  0.812  0.849  0.811
outersys    0.825    0.772    1.000    0.908  0.749  0.792  0.779
outerdia    0.756    0.699    0.908    1.000  0.724  0.783  0.686
extsys      0.807    0.812    0.749    0.724  1.000  0.961  0.681
extdia      0.807    0.849    0.792    0.783  0.961  1.000  0.759
weight      0.778    0.811    0.779    0.686  0.681  0.759  1.000
```

Multiple regression

All variables have similar sized correlations (0.7 - 0.9), and are significantly correlated with one another (from `cor.test()` output)

Suggests that **not all six predictor variables are needed**.

How do we choose a "good" set of explanatory variables?

R: Regression basics

- Full model.
  ```
  m1 <- lm(weight ~ ., data=horseshearts)
  ```
- Full model except for `outerdia`.
  ```
  m2 <- lm(weight ~ . - outerdia, data=horseshearts)
  ```
- Intercept only.
  ```
  m3 <- lm(weight ~ 1, data=horseshearts)
  ```
- Model includes `outerdia` and no intercept.
  ```
  m4 <- lm(weight ~ 0 + outerdia, data=horseshearts)
  ```
- All main effects and interactions up third order.
  ```
  m5 <- lm(weight ~ .^3, data=horseshearts)
  ```

Multiple regression

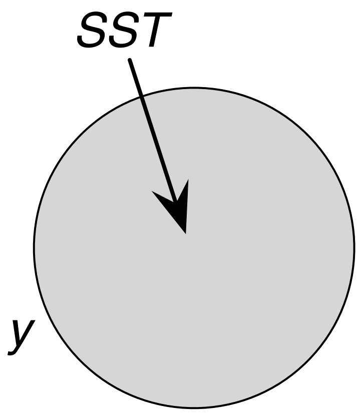
```
multiplereg <- lm(weight ~ extdia + extsys + innerdia + innersys + outerdia + outersys)
summary(multiplereg)
```

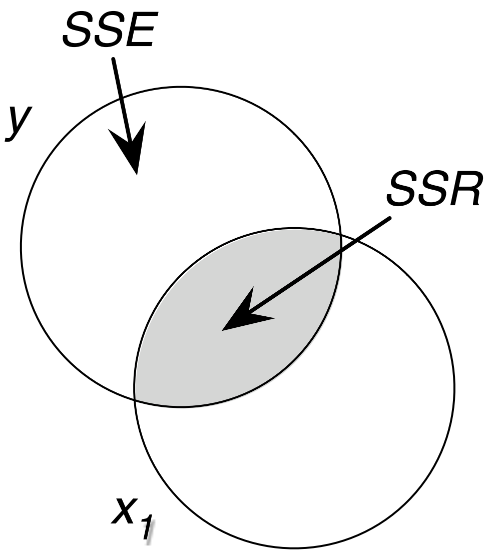
```
Call:
lm(formula = weight ~ extdia + extsys + innerdia + innersys +
    outerdia + outersys)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05051 -0.35313  0.01948  0.18674  2.09335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6311     0.4879  -3.343  0.00184 **
extdia        0.3387     0.1475   2.296  0.02716 *
extsys       -0.2996     0.1346  -2.227  0.03182 *
innerdia      0.5195     0.3954   1.314  0.19654
innersys      0.2321     0.3083   0.753  0.45617
outerdia     -0.5574     0.4510  -1.236  0.22386
outersys      0.7114     0.3288   2.164  0.03668 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6006 on 39 degrees of freedom
Multiple R-squared:  0.7525, Adjusted R-squared:  0.7145
F-statistic: 19.77 on 6 and 39 DF,  p-value: 1.922e-10
```

Not all variables significant!

Circle diagrams

Variation in a variable can be measured by its sum of squares (*SS*) and illustrated by a circle.





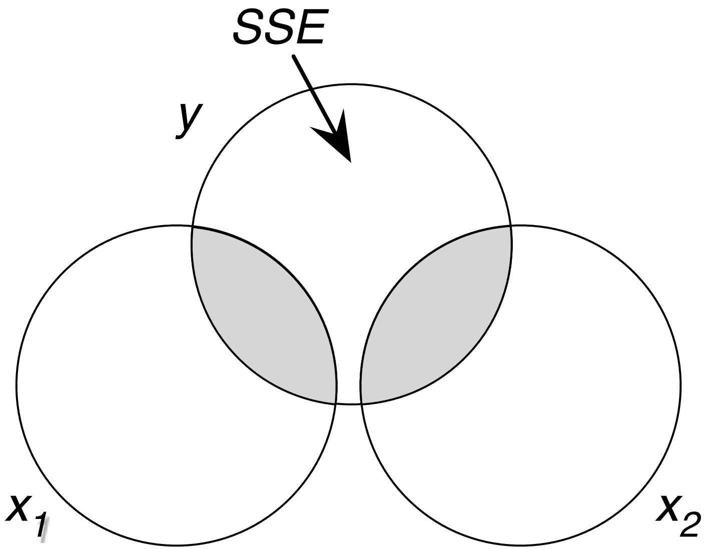
- The first single shaded circle represents *SST* (total).
- The shaded overlap of two circles represents the variation in Y explained by the variable X_1.

The variation in Y is separated into two parts: *SSR (regression)* $+$ *SSE (error or residual)* $=$ *SST*

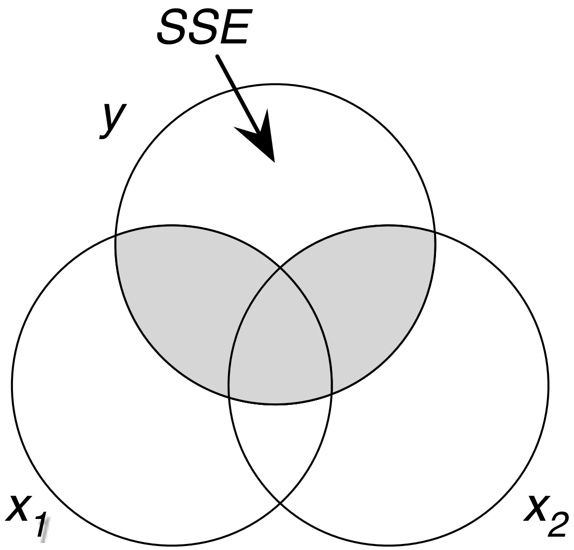
Circle diagrams and correlation

The total overlap of X_1 and X_2, and Y depends on:

- Relationship of Y with X_1 and X_2.
- Correlation between X_1 and X_2.



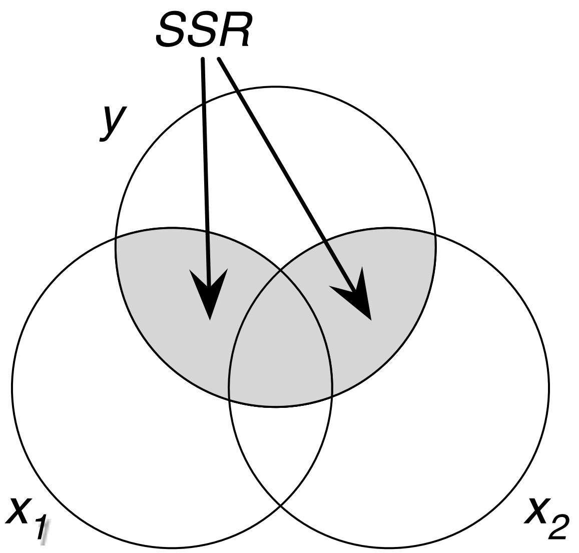
X_i's uncorrelated,

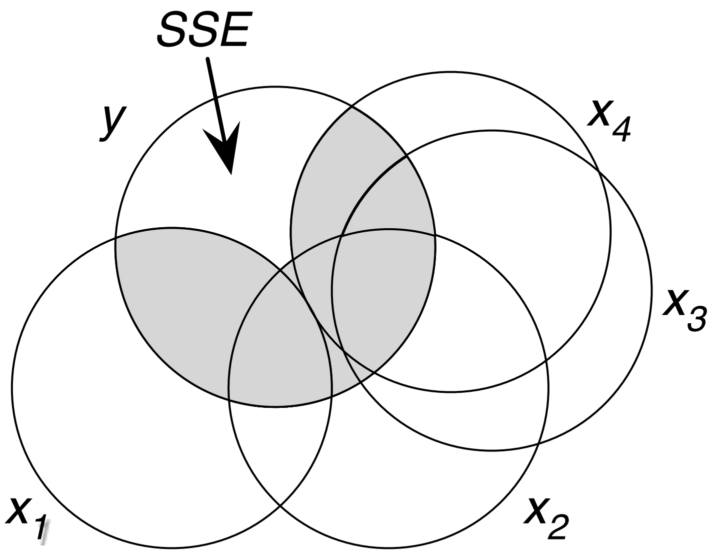


X_i's correlated

Sequential addition of predictors

Increase in number of predictors means **increase** *SSR* & R^2 and **decrease** in *SSE*.





However improvement “slows down” due to overlap (correlation)

Test for significance of increased SSR

Regress `weight` on `extdia` and `outerdia`.

```
edodreg <- lm(weight ~ extdia + outerdia)
anova(edodreg)
```

```
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
extdia     1 32.731  32.731 61.5468 7.863e-10 ***
outerdia   1  1.247   1.247  2.3453     0.133
Residuals 43 22.867   0.532
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

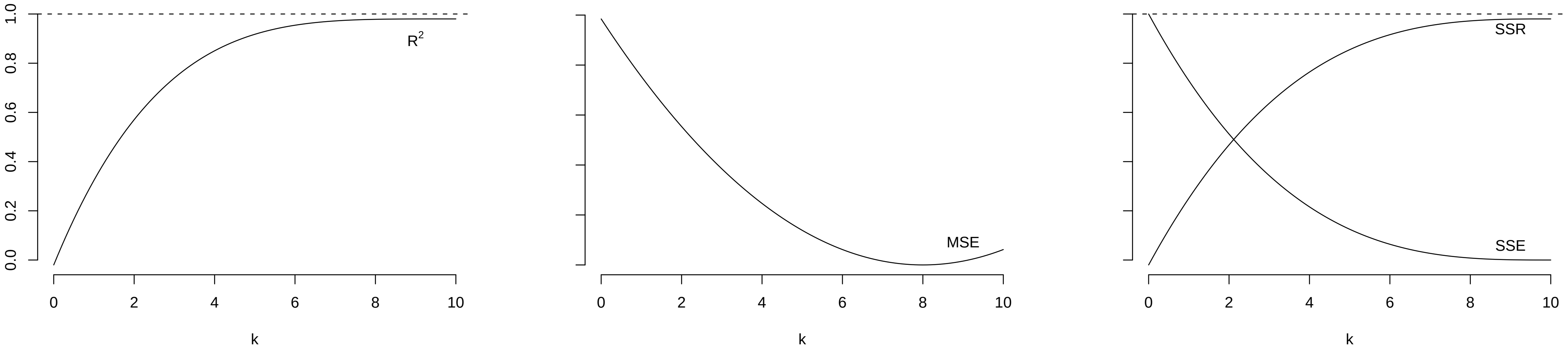
Adding `outerdia` to `extdia` does not produce significant improvement in *SSR* (32.7 to 34.0).

Test for significance of increased SSR

- Can use F test as above (`anova()`) or t test (`summary()`) because $t^2 = F$.
- If only one predictor required, then choose predictor that has **largest correlation with** Y ($r^2 = R^2$ for only one variable).
- Order of variables in model is important, so we cannot find "best" set of predictors by trial and error - **need procedure or algorithm for finding a "good" set of predictor variables**.

Effects of Adding Predictors

- Addition of variables **decreases** *SSE* and **increases** *SSR* and R^2.
- $s^2 = MSE = SSE/df$ **decreases to a minimum** and then **may increase** since addition of variable decreases *df* faster than *SSE*.



Adjusted R^2: $\bar{R}^2$

- R^2 measures the quality of the fit of a model **but** it always *increases* whenever any new variable is added.
- R^2 cannot be used as a meaningful comparison of models with different numbers of independent variables.
- Adjusted R^2 (sometimes written as $\bar{R}^2$) is a modification of R^2 that adjusts for the number of explanatory terms (p) in a model.

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\left(1 - R^2\right).$$

- $\bar{R}^2$ increases only if the new term improves the model more than would be expected by chance.
- $\bar{R}^2$ can be negative, and is always $\leqslant R^2$.

```
m1 = lm(weight ~ ., data=horsehearts)
summary(m1)$adj.r.squared
```

```
[1] 0.7144585
```

Best subset of explanatory variables

Consider “stepwise” procedures which either **add** or **delete** one variable at a time.

Forward selection:

- Start with **no** variables in the model.
- For each predictor not in the model, individually check its p-value if it is added to the model. Choose the one with the **lowest** p-value provided it is **less than** α.
- Continue until **no new predictors can be added**, i.e. **lowest** p-value $> \alpha$.

Backwards elimination:

- Start with **all** the predictors in the model.
- **Remove** the predictor with **highest** p-value greater than α.
- Refit the model and go to 2.
- Stop when all p-values are **less than** α.

Backwards elimination

Use `update()` in `R` to produce output of new model at each step.

```
bstepreg <- lm(weight~extdia+extsys+innerdia+innersys+outerdia+outersys)
summary(bstepreg)$coefficients
```

```
              Estimate Std. Error    t value   Pr(>|t|)
(Intercept) -1.6310876  0.4879347 -3.3428401 0.00183926
extdia       0.3386581  0.1475247  2.2956020 0.02716015
extsys      -0.2996260  0.1345664 -2.2266036 0.03181965
innerdia     0.5195055  0.3953815  1.3139349 0.19654470
innersys     0.2320751  0.3083393  0.7526614 0.45617269
outerdia    -0.5573826  0.4509658 -1.2359753 0.22386058
outersys     0.7114153  0.3288007  2.1636671 0.03667716
```

```
bstepreg <- update(bstepreg, . ~ . - innersys)
summary(bstepreg)$coefficients
```

```
              Estimate Std. Error   t value    Pr(>|t|)
(Intercept) -1.5303000  0.4666508 -3.279326 0.002159842
extdia       0.2972759  0.1361526  2.183403 0.034932338
extsys      -0.2598670  0.1230921 -2.111158 0.041056347
innerdia     0.7340914  0.2724511  2.694397 0.010258197
outerdia    -0.5076979  0.4436843 -1.144277 0.259310074
outersys     0.7716258  0.3171879  2.432709 0.019552503
```

Backwards elimination

```
bstepreg <- update(bstepreg, . ~ . - outerdia)
summary(bstepreg)$coefficients
```

```
              Estimate Std. Error   t value    Pr(>|t|)
(Intercept) -1.5120478  0.4681343 -3.229944 0.002441204
extdia       0.2500162  0.1302258  1.919867 0.061850908
extsys      -0.2360272  0.1217730 -1.938255 0.059498553
innerdia     0.7990585  0.2674732  2.987434 0.004733621
outersys     0.4931218  0.2041625  2.415340 0.020263541
```

```
bstepreg <- update(bstepreg, . ~ . - extdia)
summary(bstepreg)$coefficients
```

```
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -1.34067305 0.47401133 -2.8283565 0.0071411249
extsys      -0.03351747 0.06276285 -0.5340336 0.5961337480
innerdia     0.95751994 0.26242874  3.6486856 0.0007224999
outersys     0.59695512 0.20306598  2.9397101 0.0053207311
```

```
bstepreg <- update(bstepreg, . ~ . - extsys)
summary(bstepreg)$coefficients
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -1.4948399  0.3728253 -4.009492 0.0002380307
innerdia     0.8796763  0.2164017  4.065017 0.0002005695
outersys     0.5612408  0.1901373  2.951766 0.0051002805
```

All p-values below 0.05; `innerdia` and `outersys` selected.

Backwards elimination

However larger model with `extdia`, `extsys`, `innerdia` and `outersys` was only just rejected!

```
# larger model with predictors extdia, extsys, innerdia and
Residual standard error: 0.5996 on 41 degrees of freedom
Multiple R-Squared: 0.7407,     Adjusted R-squared: 0.7154
F-statistic: 29.28 on 4 and 41 DF,  $p$-value: 1.554e-11

# smaller model with predictors innerdia and outersys
Residual standard error: 0.6133 on 43 degrees of freedom
Multiple R-Squared: 0.7155,     Adjusted R-squared: 0.7023
F-statistic: 54.07 on 2 and 43 DF,  $p$-value: 1.833e-12
```

Perhaps larger model is better?

Overfitting

Stepwise procedures tend to find models that fit the given data set "too well" - may overfit and include too many variables.

Akaike Information Criterion (*AIC*) - balances **goodness of fit** with **model size**:

$$AIC = n \log\left(\frac{SSE}{n}\right) + 2k.$$

$n =$ number of data points, $k =$ number of variables selected.

AIC criterion - The step() procedure

step() function in R does stepwise selection by minimising AIC as selection criterion.

```
aicstepr <- lm(weight ~ ., data=horseshearts)
step(aicstepr)
```

```
 Start:  AIC=-40.5
weight ~ innersys + innerdia + outersys + outerdia + extsys +
    extdia

           Df Sum of Sq    RSS     AIC
- innersys  1     0.204 14.272 -41.836
- outerdia  1     0.551 14.618 -40.732
- innerdia  1     0.623 14.690 -40.507
<none>                  14.067 -40.500
- outersys  1     1.689 15.756 -37.285
- extsys    1     1.788 15.856 -36.995
- extdia    1     1.901 15.968 -36.670

Step:  AIC=-41.84
weight ~ innerdia + outersys + outerdia + extsys + extdia

           Df Sum of Sq    RSS     AIC
- outerdia  1     0.467 14.739 -42.355
<none>                  14.272 -41.836
- extsys    1     1.590 15.862 -38.977
- extdia    1     1.701 15.973 -38.657
- outersys  1     2.112 16.383 -37.489
- innerdia  1     2.590 16.862 -36.164
```

AIC criterion - The `step()` procedure

```
Step:  AIC=-42.35

weight ~ innerdia + outersys + extsys + extdia

           Df Sum of Sq     RSS     AIC
<none>                   14.739 -42.355
- extdia    1     1.325  16.064 -40.395
- extsys    1     1.351  16.090 -40.322
- outersys  1     2.097  16.836 -38.235
- innerdia  1     3.208  17.947 -35.295

Call:
lm(formula = weight ~ innerdia + outersys + extsys + extdia,     data = horseshearts)

Coefficients:
(Intercept)     innerdia     outersys       extsys       extdia
    -1.5120       0.7991       0.4931      -0.2360       0.2500
```

The procedure stops since removing any variable increases *AIC*.

AIC criterion - The step() procedure

Saving the final model.

```
ModelAIC <- step(aicstepr)
class(ModelAIC)
```

```
[1] "lm"
```

Specifying the direction, number of steps, and the scope of the models investigated.

```
m0 <- lm(weight ~ 1, data=horseshearts)
m1 <- lm(weight ~ ., data=horseshearts)
step(m1, scope=list(lower=m0,upper=m1), direction="both")
step(m0, scope=list(lower=m0,upper=m1), direction="forward",step=2)
step(m1, scope=list(lower=m0,upper=m1), direction="backward",step=1)
```

Choosing the “best” model

In general we use stepwise procedures to identify good models, but final choice should be made using **many criteria, not just one**, e.g:

- Compare residual plots (do I have outliers or influential points?);
- Some variables should be included because of context;
- Some variable choices may not make sense.
- Do model predictions make sense. Fit to part of data, predict remaining data and compare.

i.e. researcher chooses, not the computer!

Factors

- Factors are discrete valued variables. The values a factor takes are called levels. For example:
  - `Drug Treatment` (Control, Drug A, Drug B)
  - `Sex` (Male, Female)
  - `Month` (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)
  - `Water Temperature` (Cold, Warm, Hot)
- Factor levels can be ordered or nominal.
  - **Ordered:** `Water Temperature` (Cold $<$ Warm $<$ Hot)
  - **Nominal:** `Diet:` Corn, Grain, Soybean (no natural ordering of levels)
- Numerical variables with discrete levels can be modelled as factors.
- Common in experimental design because you can set the levels.
- We model factors by changes in the mean response for each level - not with regression lines.

One-way ANOVA

Used to analyse experiments with ONE factor.

Model:	fit $=$ overall effect $+$ treatment effect.	
Fit:	mean of response for treatment group	$\bar{y}_i$
Overall effect:	mean of all responses	$\bar{y}$
Treatment effect:	fit - overall effect	$\bar{y}_i - \bar{y}$

Let y_{ij} be the j^{th} replicate of treatment i. Then:

$$\begin{array}{lclclcl} y_{ij} & = & \bar{y} & + & (\bar{y}_i - \bar{y}) & + & (y_{ij} - \bar{y}_i) \\ \text{Observation} & = & \text{overall} & + & \text{treatment} & + & \text{residual} \\ & = & & \text{fit} & & + & \text{residual} \end{array}$$

One-way example: `fabrics.txt`

Flammability test on five samples from each of four different fabrics.

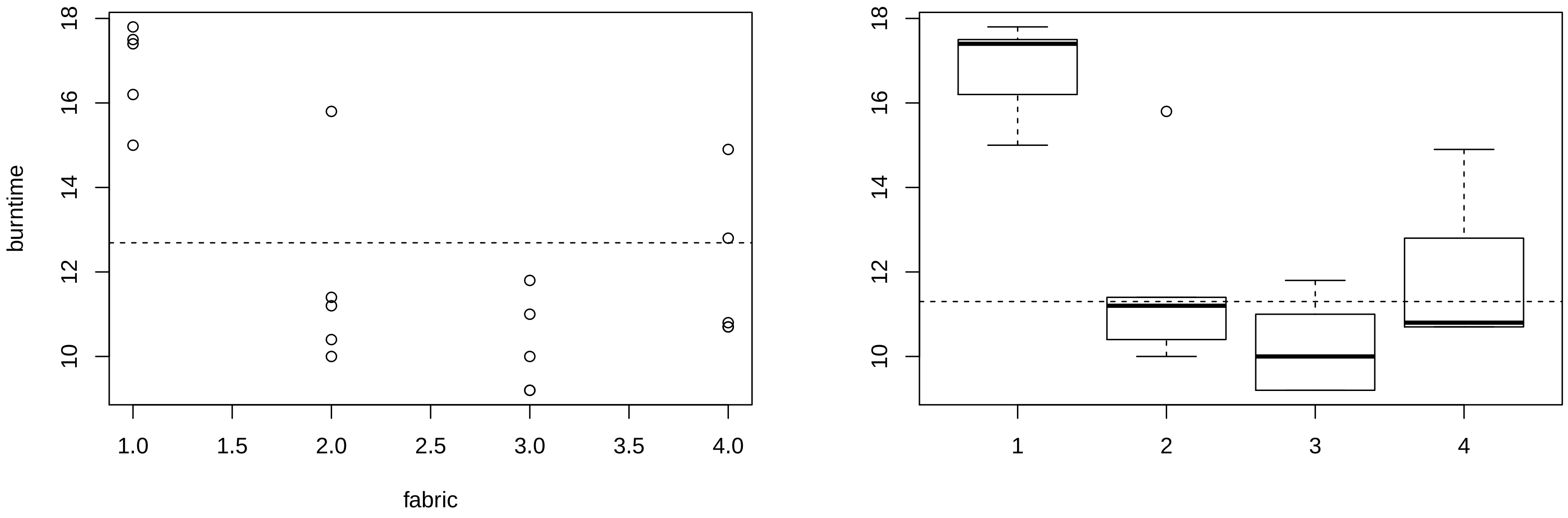
Fabric 1	Fabric 2	Fabric 3	Fabric 4
17.8	11.2	11.8	14.9
16.2	11.4	11.0	10.8
17.5	15.8	10.0	12.8
17.4	10.0	9.2	10.7
15.0	10.4	9.2	10.7
$\bar{y}_1 = 16.78$	$\bar{y}_2 = 11.76$	$\bar{y}_3 = 10.24$	$\bar{y}_4 = 11.98$
$\bar{y} = 12.69$	(mean of all 20 observations)		

e.g. Fabric 1 effect $= 16.78 - 12.69 = 4.09$
Residual for 5^{th} observation of Fabric 1
$y_{15} = 15.0 - 16.78 = -1.78.$

EDA - Plots: fabrics.txt

```
fabrics <- read.table("Data/fabrics.txt", header=TRUE)
attach(fabrics)
plot(fabric,burntime)          # Fabric is numeric, produces dot plots
abline(h=mean(burntime))       # Draw line at mean of all burntimes
plot(factor(fabric),burntime)  # Produces box plots
abline(h=median(burntime))     # Draw line at median of all burntimes
```

```
plot(factor(fabric),burntime)
abline(h=median(burntime), lty=2)
```



- Fabrics 2 and 4 burntimes are similar.
- Fabric 1 burntimes seem much higher.

Decompose variation

$$\sum\sum(y_{ij}-\bar{y})^2 \quad = \quad \sum\sum(\bar{y}_i-\bar{y})^2 \quad + \quad \sum\sum(y_{ij}-\bar{y}_i)^2$$

Total Sum Sq   Factor Sum Sq   Residual Sum Sq

i.e. **same as regression** but $\bar{y}_i$ replaces $\hat{y}$.

Let k =number of treatments (factor levels).
n =total number of observations (all groups).

Then *Total df* $= n-1$, *Factor df* $= k-1$, and *Residual df* $= n-k$.

MeanSq$=$ *SumSq/df* and F = *FactorMS/Residual MS*.

(i.e. both terms same as for regression)

One-way ANOVA in R

Let μ_i be the true mean of the i^{th} group $(i = 1, 2, \ldots, k)$.

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (no difference in means).

H_1 : not H_0 (i.e. at least one difference).

```
oneway <- lm(burntime ~ factor(fabric))
anova(oneway)
```

```
Analysis of Variance Table

Response: burntime
               Df Sum Sq Mean Sq F value    Pr(>F)
factor(fabric)  3 120.50  40.166  13.892 0.0001016 ***
Residuals      16  46.26   2.891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is close to 0 - reject H_0 and conclude that burntimes differ across fabrics.

One-way ANOVA in R

```
 summary(oneway)
Call:
lm(formula = burntime ~ factor(fabric))

Residuals:
   Min     1Q Median     3Q    Max
-1.780 -1.205 -0.460  0.775  4.040

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      16.7800     0.7604  22.067 2.09e-13 ***
factor(fabric)2  -5.0200     1.0754  -4.668 0.000257 ***
factor(fabric)3  -6.5400     1.0754  -6.081 1.59e-05 ***
factor(fabric)4  -4.8000     1.0754  -4.463 0.000392 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.7 on 16 degrees of freedom
Multiple R-squared:  0.7226, Adjusted R-squared:  0.6706
F-statistic: 13.89 on 3 and 16 DF,  p-value: 0.0001016
```

- Fabric type explains 72% of variation in burntimes.
- What does this mean?

One-way ANOVA in R

The R command lm() creates linear model for fit: equation takes **value of** i (fabric type) and outputs **the corresponding treatment mean**, $\bar{y}_i$.

How? By creating **indicator variables** for each level after the first level - e.g. for 2^{nd} fabric.

$$I_2 = \begin{cases} 1 & \text{if factor is at level 2} \\ 0 & \text{if factor is at other level} \end{cases}$$

So R output tells us that the regression equation is:

$$\text{Expected burn time} = 16.78 - 5.02 \times I_2 - 6.54 \times I_3 - 4.8 \times I_4.$$

One-way ANOVA in `R`

To explore differences between means (i.e. if H_0 is rejected), use `Tukey.HSD()` command - produces confidence intervals for every paired difference $(\mu_i - \mu_j)$.

Cannot use two-sample t-test because comparisons are **not all independent** of one another (same means), plus by chance we would expect 1 in 20 differences to be significant even if H_0 were true.

`Tukey.HSD()` gives joint **95% confidence for all intervals simultaneously** - note only works on `aov()`.

HSD $=$ **Honest Significant Difference** (doing separate t-tests would be “dishonest”!)

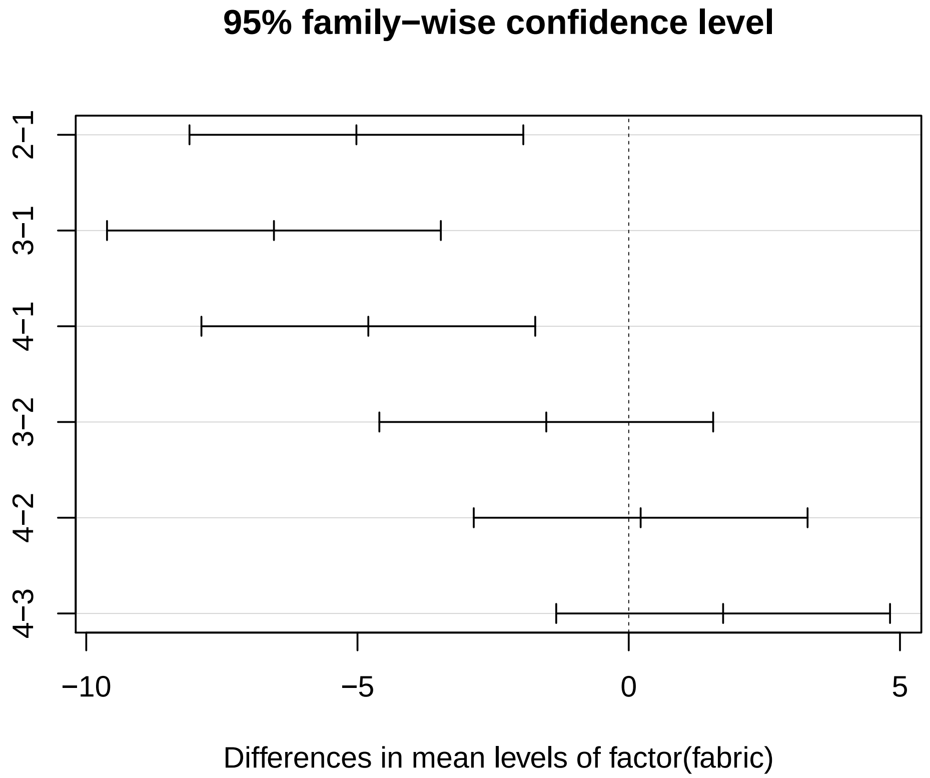
One-way ANOVA in R

```
MC <- TukeyHSD(aov(oneway)) #only works with aov()
## Note: Only works with aov()
MC

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = oneway)

$`factor(fabric)`
     diff      lwr      upr     p adj
2-1 -5.02 -8.09676 -1.94324 0.0013227
3-1 -6.54 -9.61676 -3.46324 0.0000851
4-1 -4.80 -7.87676 -1.72324 0.0019981
3-2 -1.52 -4.59676  1.55676 0.5094118
4-2  0.22 -2.85676  3.29676 0.9968426
4-3  1.74 -1.33676  4.81676 0.3968476
```

- Differences between Fabric 1 burntimes and other burntimes **are all significant**.
- Differences between other burntimes are **not** significant.

One-way ANOVA in R

`plot(MC)`



- Top three intervals DON'T include 0 ⇒ **differences** between means.
- Bottom three intervals DO include 0 ⇒ **NO differences** between means.

Important note to remember

- When performing a one-way (or two-way) ANOVA, ignore the "coefficients" part of the `summary()` output - always use `Tukey.HSD()` to explore significant differences between treatments.
- In other words, **summary()** is only used to output R^2.

ANOVA assumptions

Similar to LS regression assumptions:

- Responses are **normally distributed** about their means.
- **Variances are constant** across treatment groups.

Can examine residuals as before for normality.

Two tests for equal variances: **Bartlett's** (assumes normality) and **Levene's**.

If necessary, transform response variable to satisfy assumptions as before.

Tests for equal variances in R

```
 bartlett.test(burntime, fabric) # assumes normal data

Bartlett test of homogeneity of variances


data:  burntime and fabric
Bartlett's K-squared = 2.6606, df = 3, p-value = 0.447

 library("car")  ## Must have installed it first: install.packages("car")
 leveneTest(burntime ˜ factor(fabric), data=fabrics)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.1788 0.9092
      16
```

- Both tests have produced large p-values - not enough evidence to conclude variances differ.
- i.e. conclude that equal variances assumption is satisfied.

Two-way ANOVA

Used to analyse experiments with TWO factors - represent data in a table, **rows for factor 1 levels** and **columns for factor 2 levels**.

MODEL:
$\textit{fit} = \textit{overall effect} + \textit{row effect} + \textit{column effect}$.

Decomposition of variation and ANOVA table similar to one-way, but now there are `two` factor SumSqs giving **two** F tests and p-values.

FOUR possible test outcomes: neither significant, factor 1 significant, factor 2 significant, both factors significant

Two-way ANOVA: `concrete.txt`

Cement	1	2	3	Row Means	Row Effects
1	4	10	22	12.00	-2.42
2	5	8	28	13.67	-0.75
3	6	9	34	16.33	1.92
4	7	10	30	15.67	1.25
Col. Means	5.50	9.25	28.50	14.42	
Col. Effects	-8.92	-5.17	14.08		

- Overall mean $\bar{y} = 14.42$.
- Effect = (row or column) mean - overall mean:
  e.g. for first cell in table:
  fitted = 14.42 -2.42 - 8.92= 3.08
  Hence residual = 4 - 3.08 = 0.92

```
concrete <- read.table("Data/concrete.txt",header=TRUE)
attach(concrete)
```

```
rowmeans <- sapply(split(strength, cement),mean)
rowmeans
```

```
       1        2        3        4
12.00000 13.66667 16.33333 15.66667
```

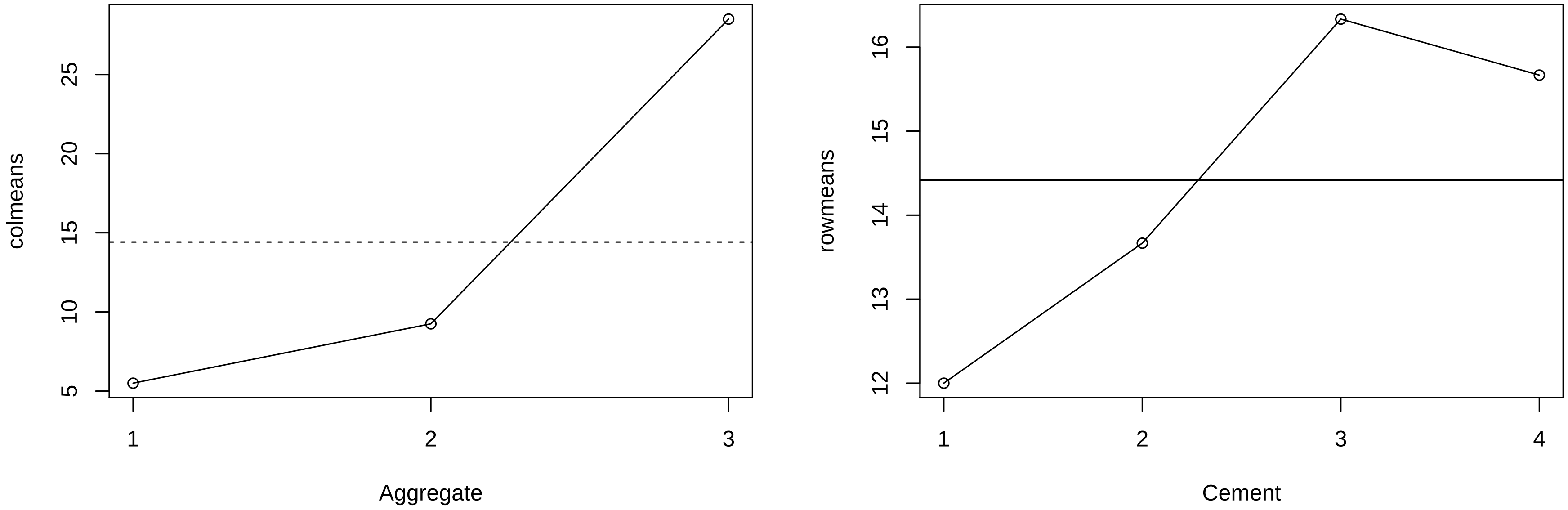
```
colmeans <- sapply(split(strength, aggregate),mean)
colmeans
```

```
   1     2     3
5.50  9.25 28.50
```

Main effects plot

Plot row means and plot column means to explore magnitude of differences due to each factor.

```
agg <- c(1,2,3)
cem <- c(1,2,3,4)
plot(agg, colmeans, type="o",xlab="Aggregate",xaxp= c(1,3,2))
abline(h = mean(strength), lty=2)

plot(cem, rowmeans, type="o", xlab="Cement", xaxp = c(1,4,3))
abline(h = mean(strength))
```



Note: Different scales on y axis.

Aggregate seems to have greater effect than Cement.

Two-way ANOVA in R

```
 twoway <- lm(strength ~ factor(aggregate) + factor(cement))
 anova(twoway)
Analysis of Variance Table

Response: strength
                  Df  Sum Sq Mean Sq F value    Pr(>F)
factor(aggregate)  2 1218.17  609.08 76.4007 5.394e-05 ***
factor(cement)     3   34.92   11.64  1.4599    0.3165
Residuals          6   47.83    7.97
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Aggregate is significant.
- Cement is not significant.
- To maximise concrete strength, choose Aggregate 3 (cement can be chosen using other reasons, e.g. cost, availability).

Two-way ANOVA with Interaction

Suppose we **hold one factor constant**, and **vary the levels of the other factor**.

If the changes are the same (similar) regardless of which first factor level was chosen, then there is no (little) interaction.

However if the response changes are not similar, then the two factors are **interacting with one another**.

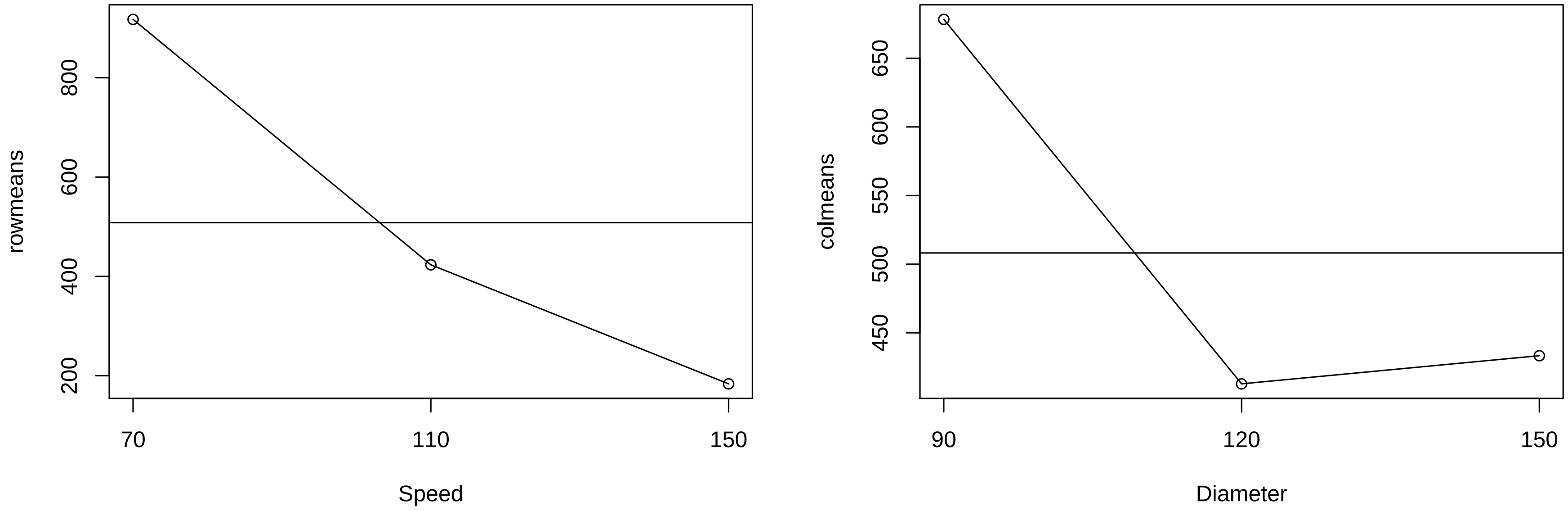
Need to investigate whether interaction is present, and if so incorporate it into the fitted model.

Note: Need at **least 2 replicates per cell**.

Example: `mixingtime.txt`

A large paddle is used to mix milk that has been collected and stored in large vats.

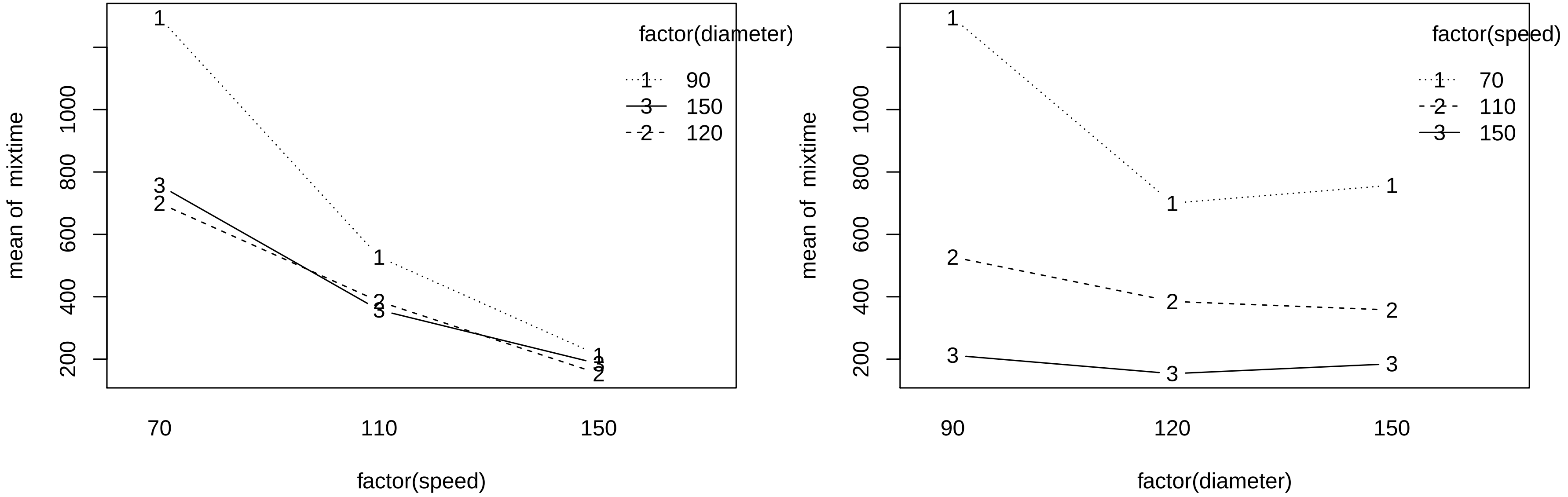
The optimal mixing time depends on both the `diameter` of the paddle and its rotation `speed` (3 levels for each). mixingtime.txt



The variable`speed` seems to have greater effect than `Diameter`. What about **interaction** between `speed` and `diameter`?

Interaction Plots: `mixingtime.txt`

```
interaction.plot(factor(speed), factor(diameter), mixtime, type = "b")
interaction.plot(factor(diameter), factor(speed), mixtime, type = "b")
```



- If **lines** (close to) **parallel** ⇒ **no** (little) **interaction present**.
- Plots suggest there may be some interaction.

Two-way model with interaction

$\textit{fit} = \textit{overall effect} + \textit{row effect} + \textit{column effect} + \textit{interaction effect}.$

Two factor SumSqs plus an interaction SumSq giving **three** F tests and p-values.

Interaction is represented using multiplication symbol ($\times$ or $*$) between factors.

If interaction term is not significant then can refit ANOVA model without interaction term.

Use `Tukey.HSD()` to explore differences as before.

R output

```
 twowayintact <- lm(mixtime ~ factor(diameter)*factor(speed))
 anova(twowayintact)
Analysis of Variance Table

Response: mixtime
                                Df  Sum Sq Mean Sq F value    Pr(>F)
factor(diameter)                 2  261871  130935 10.4688  0.004478 **
factor(speed)                    2 1680304  840152 67.1737 3.893e-06 ***
factor(diameter):factor(speed)   4  205749   51437  4.1126  0.036349 *
Residuals                        9  112565   12507
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diameter, speed and interaction term are all significant in determining optimal mixing time.

Note:

$$\begin{aligned}
\text{df(rowfactor)} &= \text{r - 1 (r= no. of levels/rows)}\\
\text{df(columnfactor)} &= \text{c - 1 (c= no. of levels/columns)}\\
\text{df(interaction)} &= \text{(r - 1)(c - 1), so}\\
\text{df(residuals)} &= \text{(n - 1) - (r - 1) -(c - 1) - (r - 1)(c -1)}
\end{aligned}$$

Analysis of covariance (ANCOVA)

Indicator variables allow us to explore the effect of a factor on a regression situation (quantitative explanatory variable and quantitative response variable) - the factor is called a **covariate**.

Analysis allows different regression lines (different slopes and/or intercepts) for different levels of the covariate by testing for significance using the linear model framework.

ANCOVA Example: `restaurant.txt`

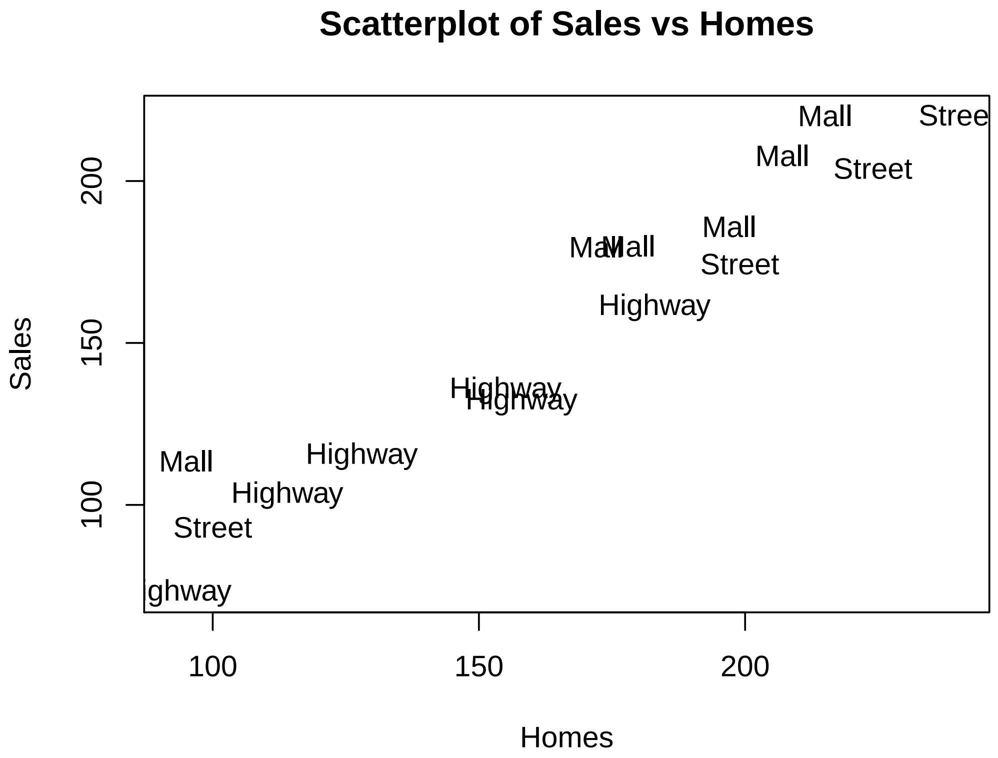
Restaurant sales (Y) depend on the number of households (X) and the restaurant location (factor).

Restaurant	Sales	Households	Location	I2	I3
1	135.27	155	Highway	0	0
2	72.74	93	Highway	0	0
3	114.95	128	Highway	0	0
4	102.93	114	Highway	0	0
5	131.77	158	Highway	0	0
6	160.91	183	Highway	0	0
7	179.86	178	Mall	1	0
8	220.14	215	Mall	1	0
9	179.64	172	Mall	1	0
10	185.92	197	Mall	1	0
11	207.82	207	Mall	1	0
12	113.51	95	Mall	1	0
13	203.98	224	Street	0	1
14	174.48	199	Street	0	1
15	220.43	240	Street	0	1
16	93.19	100	Street	0	1

ANCOVA Example: `restaurant.txt`

```
restaurant <- read.table("Data/restaurant.txt", header = TRUE)
attach(restaurant)
```

```
plot(homes,sales,type="n",xlab="Homes",ylab="Sales",
     main="Scatterplot of Sales vs Homes")
text(homes,sales,as.character(location))
```



- Looks like **two parallel lines are needed**, one for Mall and a common one for Highway and Street.
- i.e. **one** slope, **two** intercepts.

ANCOVA Model

$$Y = \beta_0 + \beta_1 I_2 + \beta_2 I_3 + (\beta_3 + \beta_4 I_2 + \beta_5 I_3) X \,.$$

Allows (up to) **three different lines:**

$$\begin{array}{lll}
\text{Highway:} & I_2 = I_3 = 0 & \Rightarrow Y = \beta_0 + \beta_3 X \\
\text{Mall:} & I_2 = 1, I_3 = 0 & \Rightarrow Y = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) X \\
\text{Street:} & I_2 = 0, I_3 = 1 & \Rightarrow Y = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) X
\end{array}$$

R Output - Interaction model with X and factor

```
anc <- lm(sales ~ homes*location)
summary(anc)
```

```
Call:
lm(formula = sales ~ homes * location)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.656  -2.794   1.128   3.013   7.528 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -6.20298   11.81820  -0.525   0.6111    
homes                 0.90949    0.08339  10.906 7.14e-07 ***
locationMall         39.22325   16.45140   2.384   0.0383 *  
locationStreet        8.03627   16.27339   0.494   0.6321    
homes:locationMall   -0.07418    0.10450  -0.710   0.4940    
homes:locationStreet -0.01205    0.10067  -0.120   0.9071    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.136 on 10 degrees of freedom
Multiple R-squared:  0.9889, Adjusted R-squared:  0.9833 
F-statistic: 177.8 on 5 and 10 DF,  p-value: 1.97e-09
```

Only X (Homes) and I_2 (Mall indicator) are significant (confirms EDA from plot - one slope, two intercepts)

R Output

Extract I_2 and then regress Y on X and I_2.

```
 I2 <- model.matrix(anc)[,3]
 final <- lm(sales ~ homes + I2)
 summary(final)
```

```
Call:
lm(formula = sales ~ homes + I2)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.040  -2.031   1.063   4.526   6.928 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.29802    5.81307  -0.567     0.58    
homes        0.90567    0.03429  26.411 1.11e-12 ***
I2          23.84174    3.28906   7.249 6.47e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.257 on 13 degrees of freedom
Multiple R-squared:  0.985, Adjusted R-squared:  0.9826 
F-statistic: 425.7 on 2 and 13 DF,  p-value: 1.418e-12
```

R Output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.29802    5.81307  -0.567     0.58
homes        0.90567    0.03429  26.411 1.11e-12 ***
I2          23.84174    3.28906   7.249 6.47e-06 ***
```

Highway and Street locations ($I_2 = 0$):

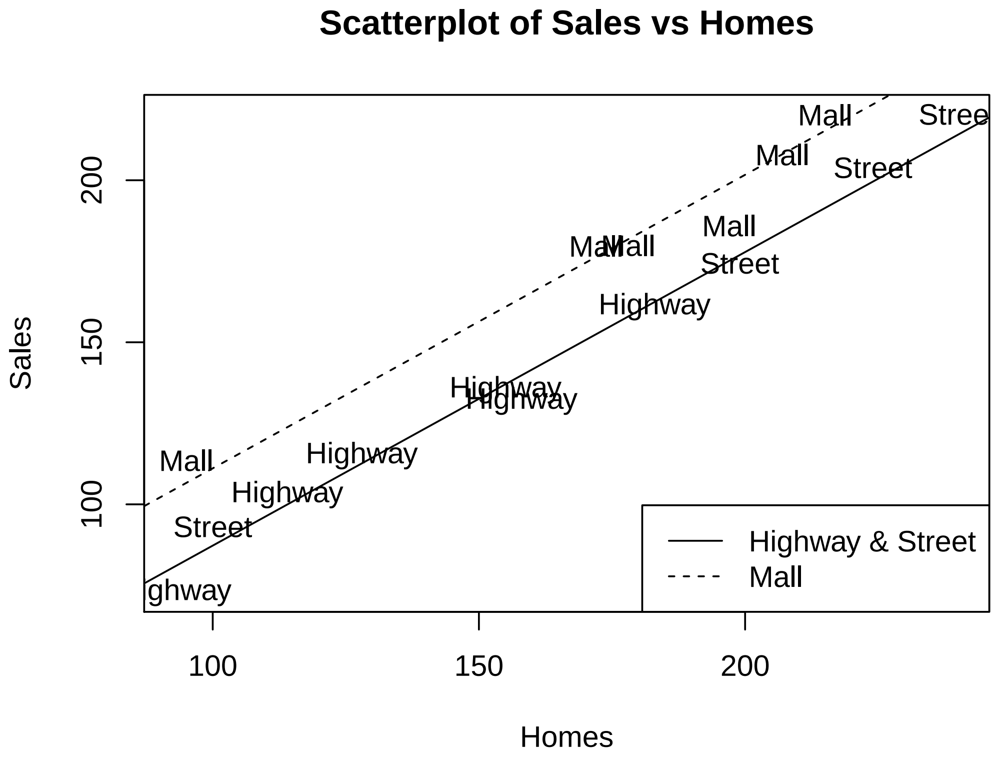
$$\hat{y} = -3.298 + 0.906x$$

Mall locations ($I_2 = 1$):

$$\begin{aligned}\hat{y} &= (-3.298 + 23.84) + 0.906x \\ &= 20.54 + 0.906x\end{aligned}$$

R Output

```
plot(homes,sales,type="n",xlab="Homes",ylab="Sales",
     main="Scatterplot of Sales vs Homes")
text(homes,sales,as.character(location))
abline(reg = final)
abline(coef = c(20.54,0.9057), lty = 2)
legend("bottomright", lty=1:2, c("Highway & Street", "Mall"))
plot(homes,sales,type="n",xlab="Homes",ylab="Sales")
text(homes,sales,as.character(location))
```


Scatterplot of Sales vs Homes
Sales
Homes
Mall
Stree
Mall
Street
MaMall
Mall
Street
Highway
Highway
Highway
Mall
Highway
Highway
Street
ghway
Highway & Street
Mall