STAT2402: Analysis of Observations

R. Nazim Khan
Department of Mathematics and Statistics
nazim.khan@uwa.edu.au

The University of Western Australia

Week 2

1. Linear Statistical Model—Continued

Contents

1.7 Multiple linear regression

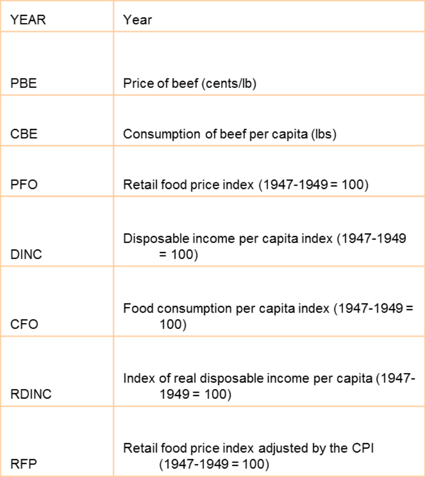
1.8 Including categorical variables

1.9 Transforming variables

1.10 Interactions

1.7 Multiple Linear Regression

Example: Multiple Linear Regression



YEAR	Year
PBE	Price of beef (cents/lb)
CBE	Consumption of beef per capita (lbs)
PFO	Retail food price index (1947-1949 = 100)
DINC	Disposable income per capita index (1947-1949 = 100)
CFO	Food consumption per capita index (1947-1949 = 100)
RDINC	Index of real disposable income per capita (1947-1949 = 100)
RFP	Retail food price index adjusted by the CPI (1947-1949 = 100)

Objective

- Data on Price of Beef
- PBE is the response variable – all others are explanatory variables.
- Determine which of the variables affect the price of beef.

The Model

Model Equation

$$
\begin{aligned}
\mathit{PBE} = {} & \beta_0 + \beta_1 \mathit{CBE} + \beta_2 \mathit{PFO} + \beta_3 \mathit{DINC} + \beta_4 \mathit{CFO} + \beta_5 \mathit{RDINC} \\
& + \beta_6 \mathit{RFP} + \mathit{error}
\end{aligned}
$$

Fitting the model

```
beef <- read.table("Data/Beef.txt", header = T, sep = "\t", stringsAsFactors = T)
beef.lm <- lm(PBE ~ CBE + PFO + DINC + CFO + RDINC + RFP, data = beef)
summary(beef.lm)

##
## Call:
## lm(formula = PBE ~ CBE + PFO + DINC + CFO + RDINC + RFP, data = beef)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2913 -0.7621  0.1914  1.1036  2.0770
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 213.36345   57.33459   3.721  0.00397 **
## CBE          -1.09641    0.14798  -7.409 2.29e-05 ***
## PFO           2.36201    1.58705   1.488  0.16751
## DINC         -3.97466    2.00628  -1.981  0.07573 .
## CFO          -1.59962    0.60274  -2.654  0.02415 *
## RDINC         2.62025    1.27235   2.059  0.06646 .
## RFP          -0.08535    0.10355  -0.824  0.42903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.672 on 10 degrees of freedom
## Multiple R-squared:  0.9635,Adjusted R-squared:  0.9416
## F-statistic: 43.99 on 6 and 10 DF,  p-value: 1.28e-06
```

Model Equation So Far

Only two variables are significant so far.

$$PBE = 213.36 - 1.10CBE - 1.60CFO$$

Model Fitting and Selection

- How do we know if we have the best model? That is, how do we know which variables to include in the model?
- The model should include only significant variables.
- Thus to find the best model we remove non-significant variables from the model **one at a time**, re-fitting the model at each stage, until the variables that remain are all significant.

Not so simple!

- Multiple regression can be very tricky! Sometimes removing/including one variable can make another variable significant, which was not previously significant!

- Sometimes removing/including one variable can make another variable, which was previously significant, non-significant!

- **General Advice**: Some experimentation is required with multiple regression! Often adding and removing variables to see their effect can give some insight into the data and the best model.

Significant terms in the Beef data

- P-values indicate which variables in the model are significant. Thus PFO, DINC, RDINC and RFP are not significant.
- The model should be re-fitted including only the significant variables. That is, drop the non-significant terms and re-fit the model. Some of the coefficients will change.

Refitting the Model

- Drop the non-significant variable with the largest p-value above 0.05.
- Refit the model with the remaining variables.
- The coefficients will change.
- Significances of the variables may also change. Some variables that were non-significant may now become significant.

Model Re-fitted without RFP

```
beef.lm1 <- update(beef.lm, . ~ . - RFP)
summary(beef.lm1)

##
## Call:
## lm(formula = PBE ~ CBE + PFO + DINC + CFO + RDINC, data = beef)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2961 -0.9603 -0.0510  1.0334  2.1738
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 194.1883    51.6330   3.761 0.003150 **
## CBE          -1.0995     0.1458  -7.543 1.14e-05 ***
## PFO           1.0998     0.4105   2.679 0.021431 *
## DINC         -2.4255     0.6916  -3.507 0.004908 **
## CFO          -1.4120     0.5499  -2.568 0.026157 *
## RDINC         1.6073     0.3245   4.953 0.000434 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.647 on 11 degrees of freedom
## Multiple R-squared:  0.961,Adjusted R-squared:  0.9433
## F-statistic: 54.23 on 5 and 11 DF,  p-value: 2.228e-07
```

Surprise!

- Now all the remaining variables are significant!
- The multiple R is almost unchanged from the full model (with all the variables in it).
- However, the Adjusted R Square has increased!

R Square or Coefficient of Determination

- R Square does not adjust for the number of variables in the model.
- However, if another variable is added to the model it will (almost) always increase the R Square.
- A simpler model, that is, one with fewer variables, is to be preferred over one with more variables.

Selecting better model?

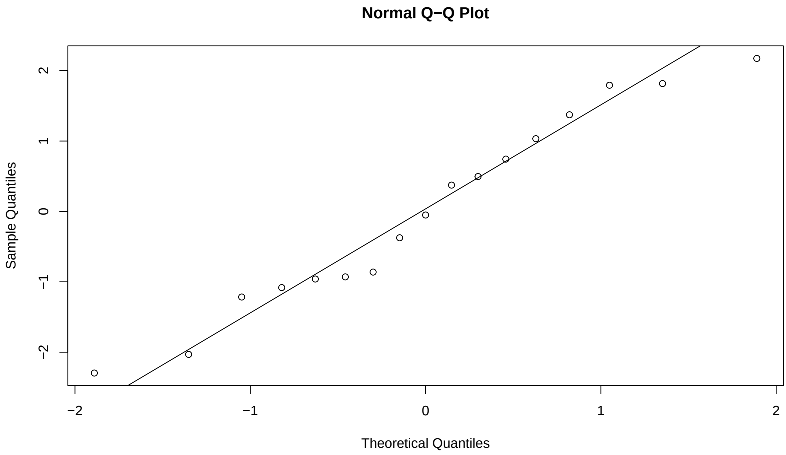
Use all of:

- Larger value of Multiple R or R Square.
- Adjusted R Square.
- Smaller standard error.

Model Diagnostics — Residual Analysis

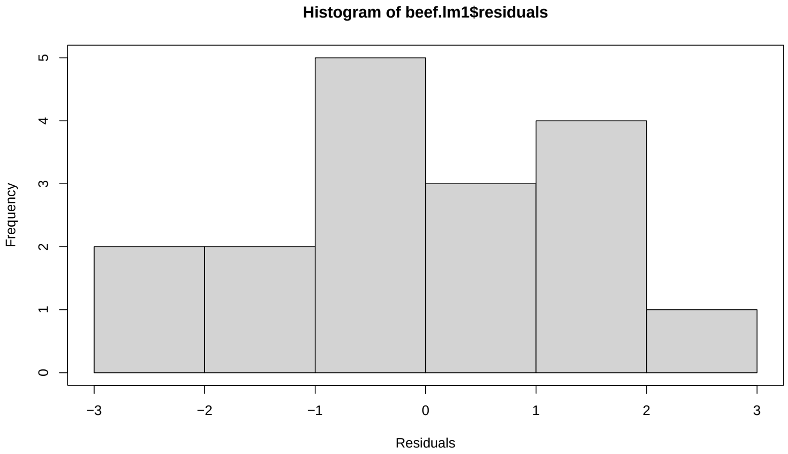
All diagnostics and model checking are as before.

```
qqnorm(beef.lm1$residuals)
qqline(beef.lm1$residuals)
```



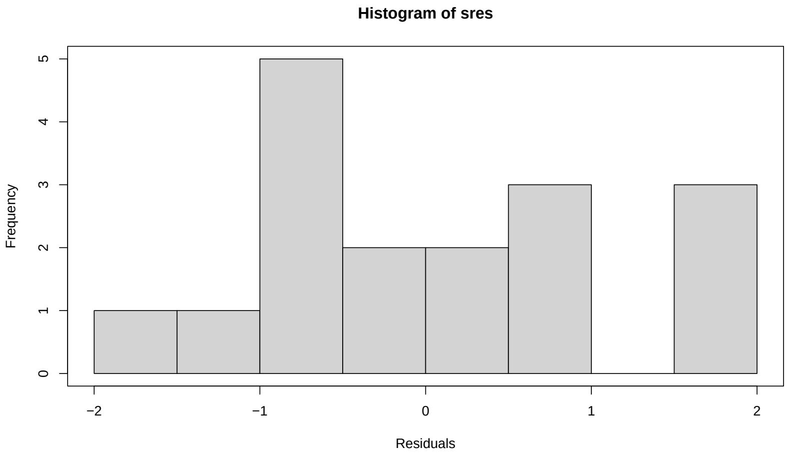
Model Diagnostics — Residual Analysis

```r
hist(beef.lm1$residuals, xlab = "Residuals")
box()
```



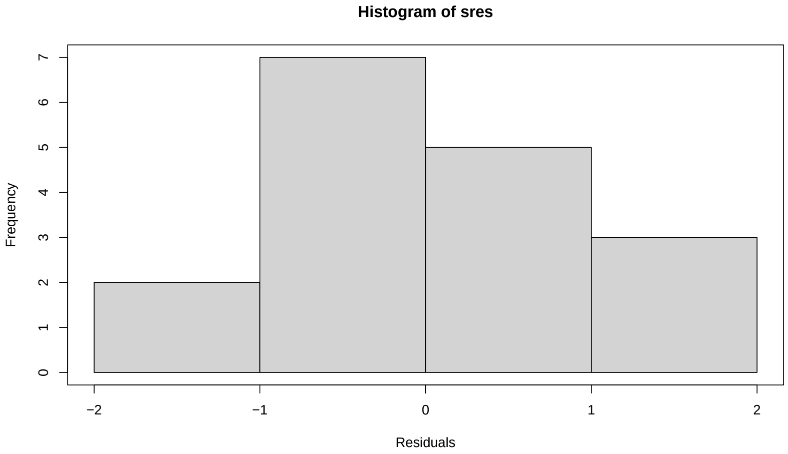
Standardised residuals

```
sres <- rstandard(beef.lm1)
hist(sres, xlab = "Residuals")
box()
```



Standardised residuals again

```r
sres <- rstandard(beef.lm1)
hist(sres, xlab = "Residuals", nclass = 4)
box()
```

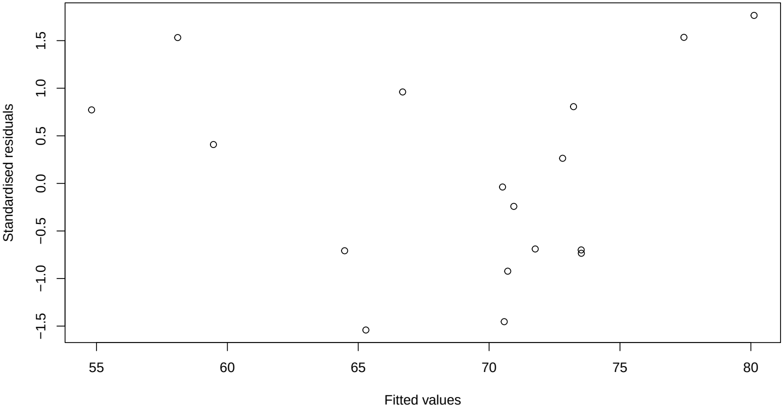


Normality assumption

- Histogram is not too different from that expected for a normal distribution.
- On the basis of the normal probability plot and the histogram of residuals, we conclude that the normality assumption is not violated.

Standardised Residuals against Predicted Values

```
plot(sres ~ beef.lm$fitted.values, xlab = "Fitted values", ylab = "Standardised residuals")
```



Conclusion

1. No clear pattern in the plot (so linear model is appropriate).
2. No outliers (none of the points are beyond 2 or -2).
3. No change in spread with predicted values (so no evidence against homogeneous variance or homoscedastic assumption).

1.8 Including categorical variables

- Often the covariates are categorical and non-continuous.
- For example, how does income depend on gender?
- In such cases care needs to be taken.

Simple Example

Model Monthly Income by Age and Gender.

Define Male $= 1$ for male workers and 0 otherwise, and

Thus Male $= 0$ indicates that the worker is Female.

Model Equation

$$\text{Income} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}$$

When Male is 1, the equation is

$$\text{Male:Income} = \beta_0 + \beta_1 \text{Age} + \beta_2$$

When Male is 0, the equation is

$$\text{Female:Income} = \beta_0 + \beta_1 \text{Age}$$

- Note that β_2 is the difference between the Male and Female incomes on average. If β_2 is significantly different from 0 then the incomes for Males and Females are different.

- In particular, if $\beta_2 > 0$ then Male employees are paid more then Female employees.

Fitting the model

```
Income <- read.table("Data/Income.txt", header = T, sep = "\t", stringsAsFactors = T)
head(Income)
```

```
##   MonthlyIncome Age.10years Male    Sex
## 1         1.548         3.2    1   male
## 2         1.629         3.8    1   male
## 3         1.011         2.7    0 female
## 4         1.229         3.4    0 female
## 5         1.746         3.6    1   male
## 6         1.528         4.1    1   male
```

Model output

```
Income.lm <- lm(MonthlyIncome ~ Age.10years + Male, data = Income)
summary(Income.lm)
```

```
##
## Call:
## lm(formula = MonthlyIncome ~ Age.10years + Male, data = Income)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.136697 -0.067380  0.001351  0.054888  0.154863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.73206    0.23558   3.107  0.00906 **
## Age.10years  0.11122    0.07208   1.543  0.14880
## Male         0.45868    0.05346   8.580 1.82e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09679 on 12 degrees of freedom
## Multiple R-squared:   0.89,Adjusted R-squared:  0.8717
## F-statistic: 48.54 on 2 and 12 DF,  p-value: 1.773e-06
```

Conclusion

1. Model equation is

$$\text{Income} = 0.7321 + 0.1112\text{Age} + 0.4587\text{Male}$$

2. The p-values indicate that Male is significant in the model.

3. Since the coefficient of Male > 0 this indicates that Males have a higher income.

4. The average difference in the incomes for Males and Females is 0.4587 in the appropriate units.

R and categorical variables

R can it regression models with categorical variables without requiring dummy variables to be defined. For a categorical variable with two or more levels, the first level in alphabetical order is taken as the base level, and all the other levels are compared with this. Any variable that contains characters is taken to be categorical. Numerical variables that are categorical need to be coerced to be categorical.

Example

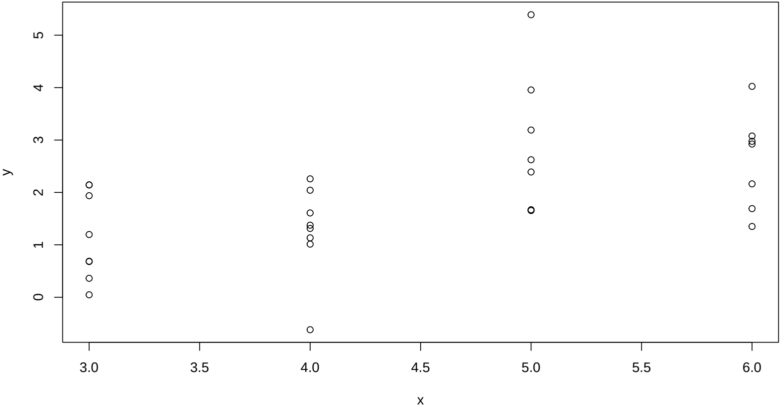
```
x <- rep(c(3:6), length = 30)
y <- rnorm(n = length(x), mean = log(x + 1), sd = 1)
eg.lm <- lm(y ~ x)
summary(eg.lm)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30170 -0.63406 -0.02618  0.52231  3.09836
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7551     0.7927  -0.953  0.34895
## x             0.6095     0.1734   3.515  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 28 degrees of freedom
## Multiple R-squared:  0.3062,Adjusted R-squared:  0.2814
## F-statistic: 12.35 on 1 and 28 DF,  p-value: 0.001516
```

Interpretation

The coefficient of x is 0.4187. This means that for a unit increase in x, the value of y increases by 0.4187.
Is this reasonable? Suppose x is the number of children and y is the expense on toys. The model indicates that, on average, the cost of toys for a family of two children is twice that for a family with one child. Not reasonable.

Better model

```
plot(y ~ x)
```



The plot shows that the cost is not linear with respect to the number of children.

Better model

```
x <- factor(x)
eg.lm1 <- lm(y ~ x)
summary(eg.lm1)
```

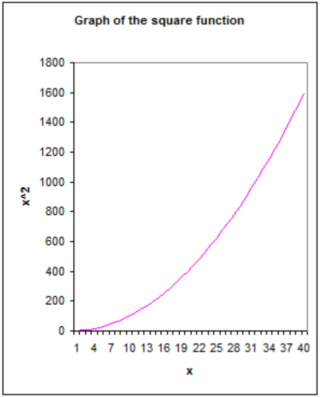
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88478 -0.56103  0.04706  0.70132  2.40851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1500     0.3539   3.249  0.00319 **
## x4            0.1158     0.5005   0.231  0.81884
## x5            1.8320     0.5181   3.536  0.00155 **
## x6            1.4506     0.5181   2.800  0.00951 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.001 on 26 degrees of freedom
## Multiple R-squared:  0.4253,Adjusted R-squared:  0.359
## F-statistic: 6.415 on 3 and 26 DF,  p-value: 0.002128
```

Interpretation

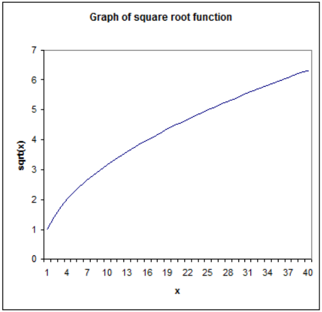
- The coefficients are the mean difference in y at each level of x compared with the *reference level*.
- The level is taken as the lowest value if the variable is numerical, or in alphabetical order if the variable is string.

1.9 Transforming variables

Graphs of common functions

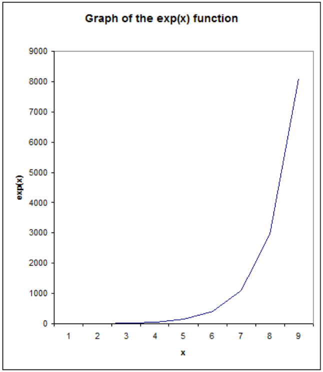


(a) Square function

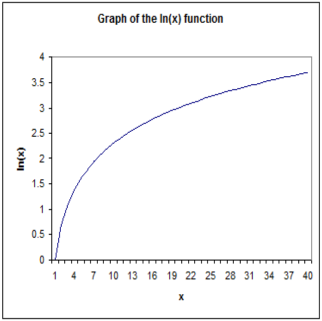


(b) Squareroot function

More Functions



(a) Exponential function



(b) Log function

Transforming variables

- Linear regression requires the residuals to be normal. Examination of residuals will indicate if this assumption holds.
- If the residuals are skewed then this can often be corrected by transforming the response variable.

Which Transformation?

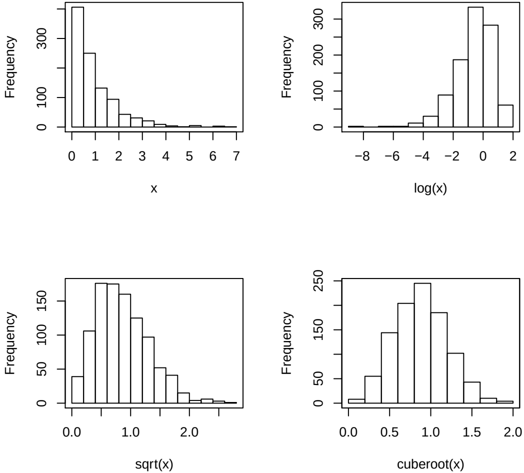


Figure: 1.Data is right skewed. 2.Ln transformation is too severe! 3.Square root is better. 4.Cube root is the best.

For Right Skewed Data

- Try square root, cube root, ...
- If none of these work, try Ln — this is the most severe.
- Idea is that these transformations pull together larger values more than small ones. Thus, for data 4, 9, 16, 25, the square root gives 2,3,4,5, which are much closer.

Left Skewed

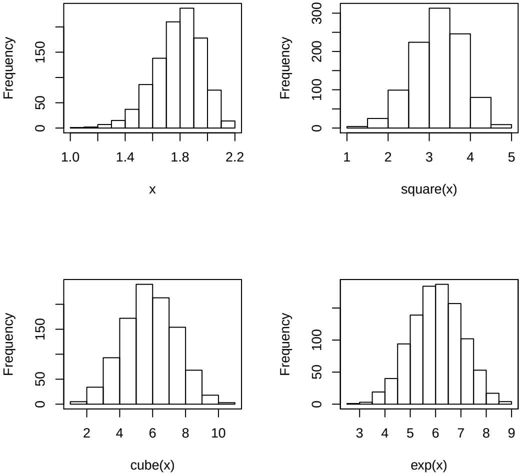


Figure: 1.Data is left skewed. 2.Square looks OK. 3.Cube looks OK. 4.Exp is the best.

For Left Skewed Data

- Try square, cube, ...
- If none of these work, try exp — this is the most severe.
- Idea is that these transformations expand out the right tail, making the data more symmetrical.

1.10 Interactions

- The response may not depend not just on the individual variables but a combination of variables.
- **Example** The salary (y) of bank workers depends on education and sex of employee. Linear model is

$$y_{ij} = \beta_0 + \beta_1 \times Sex_i + \beta_{2j} \times Educ_{ij} + \epsilon_i,$$

where for employee i, y_{ij} is the salary, Sex_i is the sex, j is the education level, and $Educ_i$ is the education level.

Bank data

For each employee the Bank Data as the following variables.

- EducLev: education level, a categorical variable with categories 1 (finished high school), 2 (finished some tertiary education), 3 (obtained a bachelor's degree), 4 (took some postgraduate courses), 5 (obtained a postgraduate degree).
- Job Grade: a categorical variable indicating the current job level, the possible levels being 1 (lowest) to 6 (highest).
- YrHired: year employee was hired.
- YrBorn: year employee was born.
- Gender: a categorical variable with values "Female" and "Male".
- YrsPrior: number of years of work experience at another bank prior to working at First National.
- PCJob: a categorical yes/no variable indicating whether the employees current job is PC related.
- Salary: current salary in thousands of dollar.

Bank data

```
bank <- read.table("Data/Bank.txt", sep = "\t", header = T, stringsAsFactors = T)
bank$EducLev <- factor(bank$EducLev)
table(bank$EducLev)
```

```
##
##  1  2  3  4  5
## 36 35 63  8 66
```

```
levels(bank$EducLev) <- c("HS", "Ter", "Bach", "PartPG", "PG")
table(bank$EducLev)
```

```
##
##     HS    Ter   Bach PartPG     PG
##     36     35     63      8     66
```

Model1

```
bank.lm <- lm(Salary ~ Gender + EducLev, data = bank)
summary(bank.lm)
```

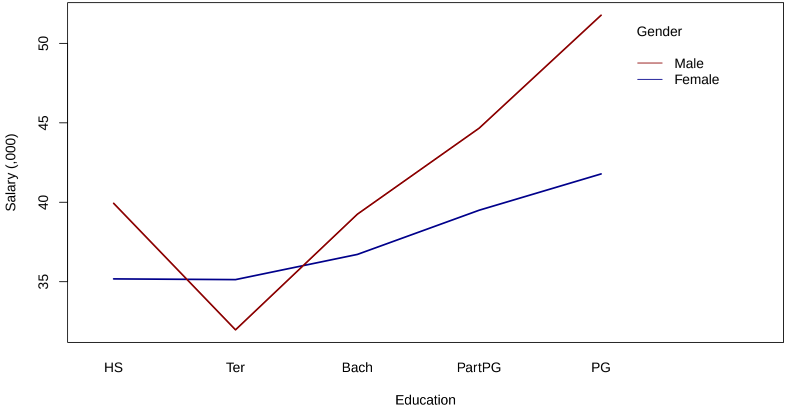
```
##
## Call:
## lm(formula = Salary ~ Gender + EducLev, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.585  -5.585  -1.631   3.515  53.947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.1411     1.6461  21.348  < 2e-16 ***
## GenderMale     5.1772     1.5643   3.310  0.00111 **
## EducLevTer    -1.4394     2.3412  -0.615  0.53936
## EducLevBach    0.7352     2.0894   0.352  0.72529
## EducLevPartPG  4.3550     3.8753   1.124  0.26244
## EducLevPG      9.2665     2.1642   4.282 2.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.846 on 202 degrees of freedom
## Multiple R-squared:  0.2534,Adjusted R-squared:  0.2349
## F-statistic: 13.71 on 5 and 202 DF,  p-value: 1.571e-11
```

Finding

Males are paid $5,177.20 more than females on average.

Data exploration

```
interaction.plot(x.factor = bank$EducLev, trace.factor = bank$Gender, response = bank$Salary,
    fun = mean, type = "l", ylab = "Salary (,000)", xlab = "Education", col = c("blue4",
        "red4"), lty = 1, lwd = 2, trace.label = "Gender", xpd = FALSE)
```



Observation

The mean salary for females is lower at all education levels except employees with some tertiary education.

Important notes

1. Interactions can change the significance of **main effects**, that is, the effect of the variables.
2. Models with interactions need care in interpretation.
3. **If the interaction term between two variables is significant, then the main effects need to be included in the model even if they are not significant.**

Example: Model with Interaction

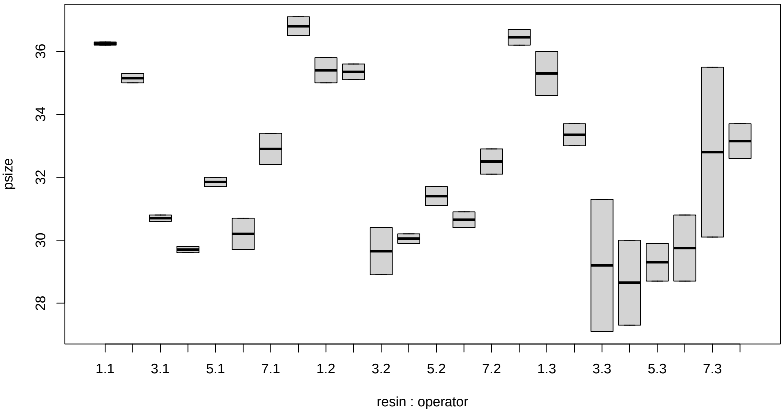
Morris and Watson (A comparison of the techniques used to evaluate the measurement process. *Quality Engineering*, 1998, **11**, 213–219) conducted an experiment to study the factors affecting the production of PVC. Three different operators used eight different devices (resin railcars). For each of the 24 combinations (crossed or factorial experiment), two samples were produced (replication). The response is the particle size of the product. The data are available in the library `faraway`.

```
library(faraway)
data(pvc)
summary(pvc)
```

```
##      psize        operator     resin   
##  Min.   :27.10   1:16     1      : 6  
##  1st Qu.:30.18   2:16     2      : 6  
##  Median :31.85   3:16     3      : 6  
##  Mean   :32.35            4      : 6  
##  3rd Qu.:35.02            5      : 6  
##  Max.   :37.10            6      : 6  
##                           (Other):12  
```

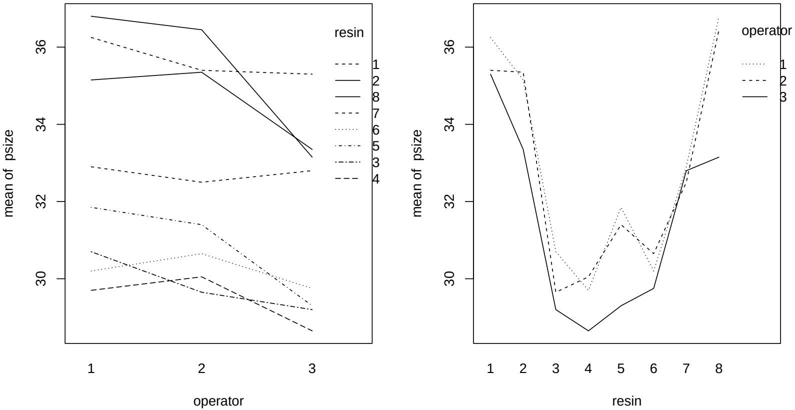
Box plot

```
library(car)
with(pvc, boxplot(psize ~ resin * operator))
```



Interaction plot

```
oldpar <- par(mfrow = c(1, 2))
with(pvc, interaction.plot(operator, resin, psize))
with(pvc, interaction.plot(resin, operator, psize))
par(oldpar)
```



Further data exploration

```
with(pvc, tapply(psize, operator, mean))
```

```
##        1        2        3
## 32.94375 32.68125 31.43750
```

```
with(pvc, tapply(psize, resin, mean))
```

```
##        1        2        3        4        5        6        7        8
## 35.65000 34.61667 29.85000 29.46667 30.85000 30.20000 32.73333 35.46667
```

```
with(pvc, tapply(psize, list(operator, resin), mean))
```

```
##       1     2     3     4     5     6    7     8
## 1 36.25 35.15 30.70 29.70 31.85 30.20 32.9 36.80
## 2 35.40 35.35 29.65 30.05 31.40 30.65 32.5 36.45
## 3 35.30 33.35 29.20 28.65 29.30 29.75 32.8 33.15
```

Observations

- Mean particle size depends on both the operator and resin railcar.
- The mean particle size for a given resin railcar is not the same for each operator.
- The mean particle size for a given operator is not the same for each resin railcar.
- Let us fit a main effects model.

Main effects model

```
psize.lm <- lm(psize ~ operator + resin, data = pvc)
sumary(psize.lm)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 36.23958    0.52260 69.3448 < 2.2e-16
## operator2   -0.26250    0.40480 -0.6485 0.5205861
## operator3   -1.50625    0.40480 -3.7209 0.0006401
## resin2      -1.03333    0.66104 -1.5632 0.1262993
## resin3      -5.80000    0.66104 -8.7740 1.135e-10
## resin4      -6.18333    0.66104 -9.3539 2.113e-11
## resin5      -4.80000    0.66104 -7.2613 1.093e-08
## resin6      -5.45000    0.66104 -8.2446 5.457e-10
## resin7      -2.91667    0.66104 -4.4122 8.164e-05
## resin8      -0.18333    0.66104 -0.2773 0.7830225
##
## n = 48, p = 10, Residual SE = 1.14496, R-Squared = 0.86
```

Model equation

$$psize_{ijk} = \beta_0 + \beta_{1i} \times Operator_i + \beta_{2j} \times Resin_j + \epsilon_{ijk}$$

where $i = 1, 2, 3$ is the operator, $j = 1, 2, \ldots, 8$ is the operator, $k = 1, 2$ is the replicate, and $psize_{ijk}$ is the particle size for operator i, resin j and replicate k.

Prediction is based on the model equation.

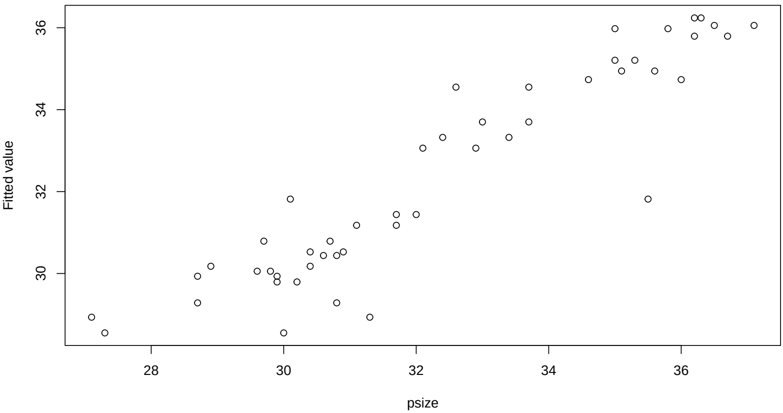
Model interpretation

- Operator 3 has a lower mean psize compared with the other two operators.

- Resins 1, 7 and 8 have a higher psize than the others.

Model fit

```r
plot(psize.lm$fitted.values ~ pvc$psize, xlab = "psize", ylab = "Fitted value")
```



Multiple comarisons: Operator

Many options. Using `emmeans`.

```
library(emmeans)
emmeans(psize.lm, pairwise ~ operator)
```

```
## $emmeans
##  operator emmean    SE df lower.CL upper.CL
##  1          32.9 0.286 38     32.4     33.5
##  2          32.7 0.286 38     32.1     33.3
##  3          31.4 0.286 38     30.9     32.0
##
## Results are averaged over the levels of: resin
## Confidence level used: 0.95
##
## $contrasts
##  contrast              estimate    SE df t.ratio p.value
##  operator1 - operator2    0.263 0.405 38   0.648  0.7944
##  operator1 - operator3    1.506 0.405 38   3.721  0.0018
##  operator2 - operator3    1.244 0.405 38   3.072  0.0107
##
## Results are averaged over the levels of: resin
## P value adjustment: tukey method for comparing a family of 3 estimates
```

Conclusion as before: Operator 3 has a lower mean psize compared with the other two.

Multiple comarisons: Operator

Using `multcomp`.

```r
library(multcomp)
post.hoc <- glht(psize.lm, linfct = mcp(operator = "Tukey"))
```

Multiple comarisons: Operator

```
# displaying the result table with summary()
summary(post.hoc)

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = psize ~ operator + resin, data = pvc)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0  -0.2625     0.4048  -0.648  0.79436
## 3 - 1 == 0  -1.5063     0.4048  -3.721  0.00176 **
## 3 - 2 == 0  -1.2437     0.4048  -3.072  0.01064 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Same as for `emmeans`.

Other options also exist.

Caution

- Need care when performing multiple comparisons.
- The p-value of 0.05 indicates that on average 5% of the time an incorrect conclusion will be made.
- **Type I Error** Rejecting a correct null hypothesis. We want to minimise this. The maximum probability of a Type I error is the significance level.
- If several comparisons are conducted the chances of Type I Error increases. See later when we cover binomial distribution.
- To compensate for this increase in chances of a Type I Error, some adjustment or correction needs to be made. Two common ones are Bonferroni, and Tukeys Honest Significance Difference (HSD),

Model with interactions 1

```
psize.lm1 <- lm(psize ~ operator * resin, data = pvc)
sumary(psize.lm1)
```

```
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      36.25000    0.85975 42.1635 < 2.2e-16
## operator2        -0.85000    1.21587 -0.6991 0.4912163
## operator3        -0.95000    1.21587 -0.7813 0.4422452
## resin2           -1.10000    1.21587 -0.9047 0.3746148
## resin3           -5.55000    1.21587 -4.5646 0.0001256
## resin4           -6.55000    1.21587 -5.3871 1.565e-05
## resin5           -4.40000    1.21587 -3.6188 0.0013718
## resin6           -6.05000    1.21587 -4.9759 4.419e-05
## resin7           -3.35000    1.21587 -2.7552 0.0110139
## resin8            0.55000    1.21587  0.4524 0.6550777
## operator2:resin2  1.05000    1.71950  0.6106 0.5471752
## operator3:resin2 -0.85000    1.71950 -0.4943 0.6255670
## operator2:resin3 -0.20000    1.71950 -0.1163 0.9083722
## operator3:resin3 -0.55000    1.71950 -0.3199 0.7518415
## operator2:resin4  1.20000    1.71950  0.6979 0.4919595
## operator3:resin4 -0.10000    1.71950 -0.0582 0.9541054
## operator2:resin5  0.40000    1.71950  0.2326 0.8180240
## operator3:resin5 -1.60000    1.71950 -0.9305 0.3613755
## operator2:resin6  1.30000    1.71950  0.7560 0.4569851
## operator3:resin6  0.50000    1.71950  0.2908 0.7737154
## operator2:resin7  0.45000    1.71950  0.2617 0.7957822
## operator3:resin7  0.85000    1.71950  0.4943 0.6255670
## operator2:resin8  0.50000    1.71950  0.2908 0.7737154
## operator3:resin8 -2.70000    1.71950 -1.5702 0.1294539
##
## n = 48, p = 24, Residual SE = 1.21587, R-Squared = 0.9
```

Model with interactions 2

```
psize.lm2 <- lm(psize ~ operator + resin + operator:resin, data = pvc)
sumary(psize.lm2)
```

```
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      36.25000    0.85975 42.1635 < 2.2e-16
## operator2        -0.85000    1.21587 -0.6991 0.4912163
## operator3        -0.95000    1.21587 -0.7813 0.4422452
## resin2           -1.10000    1.21587 -0.9047 0.3746148
## resin3           -5.55000    1.21587 -4.5646 0.0001256
## resin4           -6.55000    1.21587 -5.3871 1.565e-05
## resin5           -4.40000    1.21587 -3.6188 0.0013718
## resin6           -6.05000    1.21587 -4.9759 4.419e-05
## resin7           -3.35000    1.21587 -2.7552 0.0110139
## resin8            0.55000    1.21587  0.4524 0.6550777
## operator2:resin2  1.05000    1.71950  0.6106 0.5471752
## operator3:resin2 -0.85000    1.71950 -0.4943 0.6255670
## operator2:resin3 -0.20000    1.71950 -0.1163 0.9083722
## operator3:resin3 -0.55000    1.71950 -0.3199 0.7518415
## operator2:resin4  1.20000    1.71950  0.6979 0.4919595
## operator3:resin4 -0.10000    1.71950 -0.0582 0.9541054
## operator2:resin5  0.40000    1.71950  0.2326 0.8180240
## operator3:resin5 -1.60000    1.71950 -0.9305 0.3613755
## operator2:resin6  1.30000    1.71950  0.7560 0.4569851
## operator3:resin6  0.50000    1.71950  0.2908 0.7737154
## operator2:resin7  0.45000    1.71950  0.2617 0.7957822
## operator3:resin7  0.85000    1.71950  0.4943 0.6255670
## operator2:resin8  0.50000    1.71950  0.2908 0.7737154
## operator3:resin8 -2.70000    1.71950 -1.5702 0.1294539
##
## n = 48, p = 24, Residual SE = 1.21587, R-Squared = 0.9
```

Conclusion

1. In this case no significant interactions exist between the variables.
2. The model should therefore be refitted without the interaction terms.
3. Usually no more than two-way interactions are considered, as higher order interactions are difficult to interpret.
4. Note that interest is usually in the main effects, but interactions need to considered. The inclusion of interactions can change the significance of the main effects.

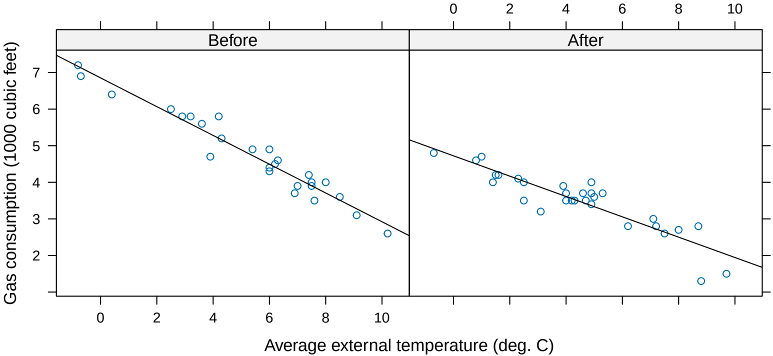
Example: Interactions

Data on weekly gas consumption and average external temperature in a house during two 'heating seasons', one before and one after cavity-insulation.

```
library(MASS)
data("whiteside")
summary(whiteside)
```

```
##     Insul         Temp              Gas       
##   Before:26   Min.   :-0.800   Min.   :1.300  
##   After :30   1st Qu.: 3.050   1st Qu.:3.500  
##               Median : 4.900   Median :3.950  
##               Mean   : 4.875   Mean   :4.071  
##               3rd Qu.: 7.125   3rd Qu.:4.625  
##               Max.   :10.200   Max.   :7.200  
```

Data exploration



Observations

- In the range of data a linear model seems appropriate.

- Gas consumption is reduced after insulation.

- However, the slope also seems to be affected, that is, the rate at which gas consumption increases as external temperature falls.

Model 1: No interactions

```
gas.lm1 <- lm(Gas ~ Temp + Insul, data = whiteside)
summary(gas.lm1)

##
## Call:
## lm(formula = Gas ~ Temp + Insul, data = whiteside)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74236 -0.22291  0.04338  0.24377  0.74314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.55133    0.11809   55.48   <2e-16 ***
## Temp        -0.33670    0.01776  -18.95   <2e-16 ***
## InsulAfter  -1.56520    0.09705  -16.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3574 on 53 degrees of freedom
## Multiple R-squared:  0.9097,Adjusted R-squared:  0.9063
## F-statistic: 267.1 on 2 and 53 DF,  p-value: < 2.2e-16
```

$$\hat{Gas} = 6.5513 - 0.3367 \times Temp - 1.5652 \times InsulAfter$$

Model 2: Interactions

```
gas.lm2 <- lm(Gas ~ Temp * Insul, data = whiteside)
summary(gas.lm2)
```

```
##
## Call:
## lm(formula = Gas ~ Temp * Insul, data = whiteside)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97802 -0.18011  0.03757  0.20930  0.63803
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.85383    0.13596  50.409  < 2e-16 ***
## Temp            -0.39324    0.02249 -17.487  < 2e-16 ***
## InsulAfter      -2.12998    0.18009 -11.827 2.32e-16 ***
## Temp:InsulAfter  0.11530    0.03211   3.591 0.000731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 52 degrees of freedom
## Multiple R-squared:  0.9277,Adjusted R-squared:  0.9235
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

Model 2: Equation

$$\begin{aligned}\hat{Gas} = 6.8538 - 0.3932 \times \mathit{Temp} - \\ 2.1300 \times \mathit{InsulAfter} + 0.1153\,\mathit{Temp} \times \mathit{InsulAfter}\end{aligned}$$

Prediction

Model 1: Simply use the fitted equation.

Model 2: A different equation for Before and After insulation.

$$\begin{aligned} \text{Before} \quad \hat{Gas} &= 6.8538 - 0.3932 \times \mathit{Temp} \\ \text{After} \quad \hat{Gas} &= 4.7238 - 0.2779 \times \mathit{Temp}. \end{aligned}$$

Lower intercept and lower slope for model with interaction. Also lower residual SE and higher adjusted r^2. Finally, lower AIC.

```
AIC(gas.lm1, gas.lm2)

##         df      AIC
## gas.lm1  4 48.60465
## gas.lm2  5 38.20098
```