Class 3

- Analysis of variance table
- General linear hypothesis test
- R^2
- Diagnostics for X

Analysis of Variance (ANOVA)

- A way to organize arithmetic
- (Total) variation in Y can be expressed as $\Sigma(Y_i - \bar{Y})^2$
- Partition this variation into two *sources*
  - Model (regression)
  - Error (residual)

ANOVA (Total)

- SST = $\Sigma(Y_i - \bar{Y})^2$
- dfT = n-1
- MST = SST/dfT

ANOVA (Model)

- SSM = $\Sigma(\hat{Y}_i - \bar{Y})^2$
- dfM = 1 (for the slope)
- MSM = SSM/dfM

ANOVA (Error)

- SSE = $\Sigma(Y_i - \hat{Y}_i)^2$
- dfE = n-2
- MSE = SSE/dfE
- MSE is an estimate of the variance of Y taking into account (or conditioning on) the explanatory variable(s))

ANOVA Table

Source	df	SS	MS
Model	1	$\Sigma(\hat{Y}_i - \bar{Y})^2$	SSM/dfM
Error	n-2	$\Sigma(Y_i - \hat{Y}_i)^2$	SSE/dfE
Total	n-1	$\Sigma(Y_i - \bar{Y})^2$	SST/dfT

ANOVA Table (2)

Source	df	SS	MS	F	P
Model	1	SSM	MSM	MSM/MSE	.nn
Error	n-2	SSE	MSE		
Total	n-1				

Expected Mean Squares

- **MSM, MSE are random variables**
- **$E(MSM) = \sigma^2 + \beta_1^2\Sigma(X_i - \bar{X})^2$**
- **$E(MSE) = \sigma^2$**
- **When H_0 is true, $\beta_1 = 0$, E(MSM) = E(MSE) and**

F test

- **F=MSM/MSE ~ F(dfM, dfE) = F(1, n-2)**
- **When H_0 is false, $\beta_1 \neq 0$ and MSM tends to be larger than MSE**
- **We reject H_0 when F is large:**
- **$F \geq F(1-\alpha, dfM, dfE) = F(.95, 1, n-2)$**
- **In practice we use P values**

F test (2)

- **When H_0 is false, F has a *noncentral* F distribution**
- **This can be used to calculate power**
- **Recall $t = b_1/s(b_1)$ tests H_0**
- **It can be shown that $t^2 = F$**
- **So the two approaches give the same P values**

```
time<-read.table('CH01TA01.txt',
col.names=c("size", "hours"));
reg1<-lm(hours~size, time);
anova(reg1)
summary(reg1)
```

```
Analysis of Variance Table

Response: hours
      Df Sum Sq Mean Sq  F value
size   1 252378  252378  105.88
Resid  23 54825    2384

Pr(>F)
4.449e-10 ***
            std  t-value p-value
Int 62.366 26.177 2.382 0.0259 *
size 3.570 0.347 10.290 4.45e-10
```

General linear test

- A different view of the same problem
- We want to compare two models
  - $Y_i = \beta_0 + \beta_1 X_i + \xi_i$ (*full* model)
  - $Y_i = \beta_0 + \xi_i$ (*reduced* model)
- Compare using SSEs: SSE(F), SSE(R)
- F=((SSE(R) - SSE(F))/(dfE(R) - dfE(F)))/MSE(F)

Simple Linear Regression

- $SSE(R)= \Sigma(Y_i-b_0)^2= \Sigma(Y_i- \bar{Y})^2=SST$
- SSE(F)=SSE
- dfE(R)=n-1, dfE(F)=n-2,
- dfE(R)-dfE(F)=1
- F=(SST-SSE)/MSE=SSM/MSE

R^2 , r^2

- r is the usual (Pearson) correlation
- It is a number between –1 and +1 and measures the strength of the linear relation between two variables
- r^2 = SSM/SST = 1 – SSE/SST
- Explained and unexplained variation

R^2 , r^2

- We use R^2 when the number of explanatory variables is arbitrary (simple and multiple regression)
- R^2 is often multiplied by 100 and thereby expressed as a percent

```
Response: hours
      Df Sum Sq Mean Sq  F value
size   1 252378  252378  105.88
Resid  23 54825    2384

Multiple R-squared: 0.8215
Adjusted R-squared: 0.8138
```

```
R-Square     0.8215 (R)
  = SSM/SST
  = 252378/307203

Adj R-Sq     0.8138 (R)
  =1-MSE/MST
  =1-2383/(307203/24)
```

Diagnostics and remedial measures

- Diagnostics: look at the data to diagnose situations where the assumptions of our model are violated
- Remedies: changes in analytic strategy to fix these problems

Look at the data

- Before trying to describe the relationship between a response variable (Y) and an explanatory variable (X), we should look at the distributions of these variables
- We should always look at X
- If Y depends on X, looking at Y alone may not be very informative

Diagnostics for X

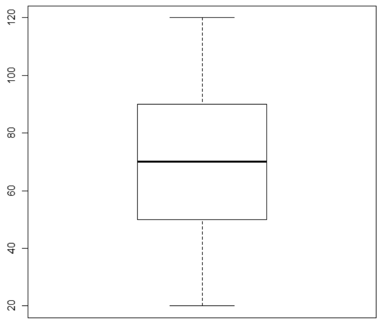
- summary(time$size)
- library(psych)
- describe(time$size)

Diagnostics for X (2)

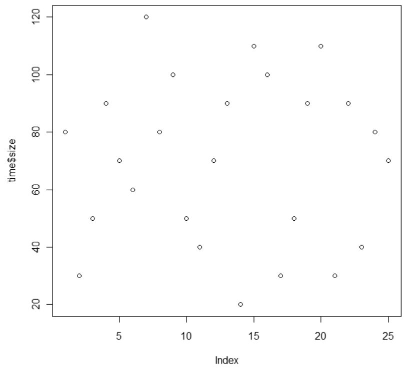
- Examine the distribution of X
  - Is it skewed?
  - Are there outliers?
- Do the values of X depend on time (order in which the data were collected)?

```
Min. 1st Qu. Median Mean
 20    50       70    70

3rd Qu.    Max.
90         120

 n mean sd med trm  mad  min max
25 70 28.72 70 70 29.65 20 120

range  skew kurtosis   se
100   -0.09  -1.25    5.74
```

```
stem(time$size, scale=2)
boxplot(time$size)

 2 | 0000
 4 | 00000
 6 | 0000
 8 | 0000000
10 | 0000
12 | 0
```




```
plot(time$size)
```



Normal distributions

- Our model does *not* state that X comes from a single normal population
- Same comment applies to Y
- In some cases, X and/or Y may be normal and it can be useful to know this

Normal quantile plots

- Consider n=5 observations iid N(0,1)
- From table of normal distribution, we find
  - $P(z \leq -.84) = .20$
  - $P(-.84 < z \leq -.25) = .20$
  - $P(-.25 < z \leq .25) = .20$
  - $P(.25 < z \leq .84) = .20$
  - $P(.84 < z) = .20$

Normal quantile plots (2)

- So we *expect*
  - One observation $\leq$ -.84
  - One observation in (-.84, -.25)
  - One observation in (-.25, .25)
  - One observation in (25, .84)
  - One observation > .84

Normal quantile plots (3)

- $Znorm_i = \Phi^{-1}((i-.375)/(n+.25))$, i=1 to n
- Plot the order statistics $X_{(i)}$ versus $Znorm_i$

Normal quantile plots (4)

- The standardized X variable is $z = (X - \mu)/\sigma$
- So, $X = \mu + \sigma z$
- If the data are approximately normal, the relationship will be approximately linear with slope close to σ and intercept close to μ.

```
qqnorm(time$size)
```