Regression I

Professor: Hammou El Barmi
Columbia University

Introduction

Regression Analysis

- A statistical tool for studying the relationship between one variable (response variable) and other variables (predictor variables)
- Explain the effect of change in a predictor variable on response variable
- Predict the value of response variable based on the value(s) of predictor variable(s)

The response variable is called dependent variable
Predictor variables and called independent variables or explanatory variables

Example

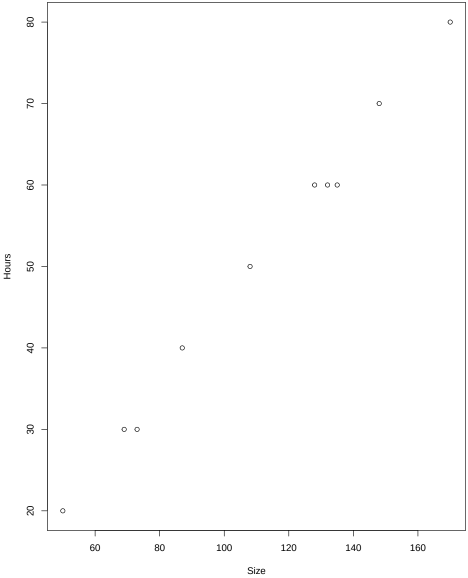
- A company manufactures standard wall clocks
- Wholesalers order the clocks in lot sizes
- The company wants to study relation between lot sizes and man-hours used for manufacture
- Data from a small sample are shown on the next slide

Lot size (X)	Man-hour (Y)
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	148
60	132

Example

```
> regdata<-read.table("/Users/HElbarmi/Desktop/EDA/Regressin/Lotsize.txt",heade
> regdata
     Size Hours
1    30   73
2    20   50
3    60  128
4    80  170
5    40   87
6    50  108
7    60  135
8    30   69
9    70  148
10   60  132
> Size<-regdata[,2]
> Hours<-regdata[,1]
> plot(Size, Hours, xlab="Hours", ylab="Size")
```

Example



Model

Model is

$$\begin{aligned} Y_{X=x} &= E(Y|X=x)+\epsilon \\ &= \beta_0+\beta_1 x+\epsilon \end{aligned}$$

That is, we assume that $E(Y|X=x)=\beta_0+\beta_1 x$

- $\beta_0 = E(Y|X=0), \beta_0$ is the mean of Y when $X=0$
- β_1 is the change in the mean of Y corresponding to a one unit in X.

Model

- Suppose our data is $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Therefore

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Estimates, b_0 and b_1 of β_0 and β_1 are solution to

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$$

i.e. they are the values of β_0 and β_1 that solve

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \epsilon_i^2$$

Model

The solution is

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ b_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = r\frac{S_y}{S_x} \end{aligned}$$

Here r is the correlation coefficient

Example Cont.

```
> lm(Hours~Size)
Call:
lm(formula = Hours ~Size)
Coefficients:
(Intercept)       Size
       10           2
```

This gives

$$b_0 = 10 \quad \text{and} \quad b_1 = 2$$

The estimated regression line is

$$\widehat{\text{Hours}} = 10 + 2\text{Size}$$

- Here b_0 has no meaningful interpretation
- $b_1 = 2$ means that if increase the size of the lot by one, the number of hours required to do the work will increase by about 2 hours.

```
> summary(lm(Hours˜Size))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.00000    2.50294   3.995  0.00398 **
Size         2.00000    0.04697  42.583 1.02e-10 ***


Residual standard error: 2.739 on 8 degrees of freedom
Multiple R-squared:  0.9956,Adjusted R-squared:  0.9951
F-statistic:  1813 on 1 and 8 DF,  p-value: 1.02e-10
```

Model Assumptions

A $100(1-\alpha)\%$ confidence interval for β_i is

$$b_i \pm t_{\alpha/2}(n-2)SE(b_i)$$

```
> confint(lm(Hours~Size))
                2.5 %    97.5 %
(Intercept) 4.228211 15.771789
Size        1.891694  2.108306
```

Interpretation: We are 95% confident that a one unit increase in lot size will increase on average the number of hours required to process the lot by a number between 1.89 hours and 2.11 hours.

Analysis of Variance Approach

- Total Sum of Squares (SST) = Total variation in the response
- Regression Sum of Squares (SSR) = Variation in the response explained by the explanatory (predictor) variable
- Error Sum of Squares (SSE) = Variation in the response not explained by the explanatory (predictor) variable
- $SST = SSR + SSE$ and

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad SST = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \quad \text{and} \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- The coefficient of determination is

$$R^2 = \frac{SSR}{SST}$$

As a percentage, this is the percentage variability in the response explained by the predictor variable.

Analysis of Variance Approach

- The ANOVA table is given by

Source	df	SS	MS	F
Model	1	SSR	MSR=SSR/1	MSR/MSE
Error	n-2	SSE	MSE=SSE/(n-2)	
Total	n-1	SST		

- The coefficient of determination is

$$R^2 = \frac{SSR}{SST}$$

- MSE is an estimate of σ^2
- To test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ we reject H_0 if $F > F(1-\alpha, 1, n-2)$ or if $p - value < \alpha$.

Analysis of Variance Approach

```
> aov(lm(Hours~Size))
Call:
   aov(formula = lm(Hours ~ Size))

Terms:
                Hours Residuals
Sum of Squares  13600        60
Deg. of Freedom     1         8

Residual standard error: 2.738613
Estimated effects may be unbalanced
```

- $SSR = 13600, SSE = 60$ and $SST = SSR + SSE = 13660$
- In addition $R^2 = 13600/13660 = 0.9956$.
- Interpretation: about 99.56% of the variability in the number of hours required to process a lot is explained by its size.

Analysis of Variance Approach

The ANOVA table is

Source	df	SS	MS	F
Model	1	13600	13600	1813.33
Error	8	60	7.5	
Total	9	13660		

```
> summary(lm(Hours~Size))
Residual standard error: 2.739 on 8 degrees of freedom
Multiple R-squared:  0.9956,Adjusted R-squared:  0.9951
F-statistic:  1813 on 1 and 8 DF,  p-value: 1.02e-10
```

An estimate of the error variance is $MSE = 7.5$
To test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ we reject H_0 since $p - value < 0.05$

Estimation of μ_x

- An estimator of μ_x is
$$\hat{y}_x = b_0 + b_1 x$$
- A $100(1-\alpha)\%$ confidence interval for μ_x is
$$\hat{y}_x \pm t_{n-2}(\alpha/2)\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Example

- Example: suppose we want to estimate the average number of hours it will take to process a lot of size equal to 45 using a 95% confidence interval
- In R we use

```
> fit<-lm(Hours~Size)
> predict(fit,newdata = data.frame(Size=45), interval="confidence")
  fit      lwr      upr
100 97.93082 102.0692
```

- The output shows that $\hat{y}_{45} = 100$ and a 95% confidence interval for the average number of hours it will take to process a lot of size 45 is $[97.93, 102.07]$.
- Interpretation: We are 95% confident that on average it will take between 97.93 hours and 102.07 hours to process a lot of size 45.

Prediction of y_x

- A predicted value y_x of the response when $X = x$

$$\hat{y}_x = b_0 + b_1 x$$

- A $100(1-\alpha)\%$ prediction interval for y_x is

$$\hat{y}_x \pm t_{n-2}(\alpha/2)\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Example

- Example: suppose we want to estimate the average number of hours it will take to process a lot of size equal to 45 using a 95% confidence interval
- In R we use

```
> fit<-lm(Hours~Size)
> predict(fit,newdata = data.frame(Size=45), interval="prediction")
  fit      lwr      upr
1 100 93.35441 106.6456
```

- The output shows that $\hat{y}_{45} = 100$ and a 95% confidence interval for the average number of hours it will take to process a lot of size 45 is $[93.35, 106.65$.
- Interpretation: We predict with 95% confident that it will take between 93.35 hours and 106.65 hours to process a lot of size 45.

Joint estimation

- Joint estimation of β_0 and β_1
- For the data where $X = 0$ is meaningful, β_0 should be estimated as well as β_1
- If seperate 95% confidence intervals for β_0 and β_1 are constructed, respectively, then
    - 5% of the samples would results in a confidence interval of β_0 that does contain it. Another 5% (possibly the same) will result in a confidence interval of β_1 that does not contain it.
    - Thus, as much as 10% of the samples will result in intervals for β_0 and β_1 that do not contains either or both parameters.
- Joint 95% CIs for $\beta_0 \& \beta_1$ can be constructed to ensure 95% correctness of the entire set of CIs
- Family confidence coefficient: The proportion of samples that are correct for every member of a family of CIs in repeated sampling and recalculation of each CI.

Joint estimation

Bonferroni joint CIs

- To achieve $1-\alpha$ family confidence, each of β_0 and β_1 is estimated with $1-\alpha/2$ confidence
- $1-\alpha$ Bonferroni joint CIs are given by

$$b_0 \pm Bs_{b_0} \quad \text{and} \quad b_1 \pm Bs_{b_1}$$

  where

$$B = t_{n-2}(\alpha/4)$$

- $1-\alpha$ is a lower bound on the true family confidence coefficient
- Family confidence coefficients are often specified at lower levels (say 90%)

```
> fit<-lm(Hours~Size)
> confint(fit, level=1-0.05/4)
               0.625 %  99.375 %
(Intercept) 1.975688 18.024312
 Size          1.849426   2.150574
```

Simulatenous estimation of mean response

Working Hotelling Method

- works for unlimited family of CIs
- as a result, yields a confidence band for the entire response line
- it is given by

$$\hat{y}_x \pm W\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

  where

$$W^2 = 2F(1-\alpha, 2, n-2)$$

- When the number of CIs is small, Bonferroni method gives smaller widths than Working-Hotelling method generally

Simulatenous estimation of mean response

```
ci.wh <- function(fit, newdata, alpha = 0.1){
  df    <- nrow(model.frame(fit)) - length(coef(fit))
                W     <- sqrt( 2 * qf(1 - alpha, length(coef(fit)), df) )
                ci    <- predict(fit, newdata, se.fit = TRUE)
                x <- cbind(
                  ’x’   = newdata,
                  ’s’   = ci$se.fit,
                  ’fit’ = ci$fit,
                  ’lwr’ = ci$fit - W * ci$se.fit,
                  ’upr’ = ci$fit + W * ci$se.fit)

                return(x)
}

newdata<-data.frame(Size=c(45,65,75))
ci.wh(fit,newdata)
```

Simulatenous estimation of mean response

```
  Size         s fit       lwr      upr
1   45 0.8972999 100  97.82992 102.1701
2   65 1.1163886 140 137.30006 142.6999
3   75 1.4589984 160 156.47148 163.5285
```

Diagnostics and Remedial Measures

Recall that the model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

and we assumed that

- ϵ_i is normally distributed with mean zero and variance σ^2
- ϵ_is are independent

If these assumptions do not hold, they invalidate our analysis

Diagnostics and Remedial Measures

Diagnostics:

- Examine appropriateness of model & and detect violations of model assumptions
- Typical violations
  - Regression function is not linear
  - Error terms do not have constant variance
  - Error terms are not independent
  - One or more observations are outliers
  - Error terms are not normally distributed
  - One or more important predictors have been omitted from the model

Diagnostics and Remedial Measures

We use the residuals to examine important departures from the simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon \quad \text{with independent and identically normally distributed errors}$$

- The ith residual $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$.
- The residuals are used to estimate the errors
- $\sum_{i=1}^{n} e_i = 0$.
- $var(e_i) = \sigma^2(1 - h_{ii})$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The residuals are not independent

Diagnostics and Remedial Measures

We use plots of the residuals to answer these questions. The plots that are commonly used are

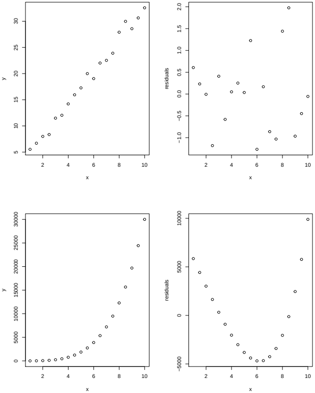
1. plot the residuals against the predictor variable
2. plot the residuals against the fitted values
3. plot the residuals against the time (important if data collected over time)
4. plot the residuals against omitted variables
5. box plot for the residuals
6. normal plot for the residuals

Diagnostics and Remedial Measures

We should check for:

1. Nonlinearity of the the regression function: this can be studied by a plot of the residuals against the predictor variable or equivalently by a plot of the residuals against the fitted values. The plot should not show any particular pattern.

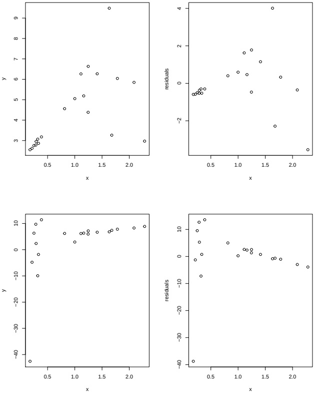
Figure: Linear in top not linear in the bottom



Diagnostics and Remedial Measures

2. Nonconstancy of the error variance: plot residuals versus the predictor variable or the fitted values

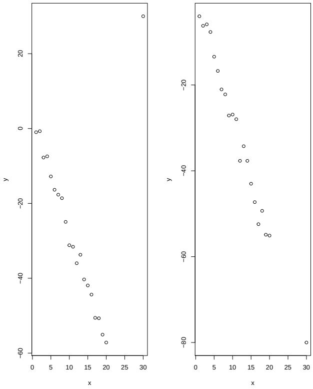
Figure: Noconstant Variance



Diagnostics and Remedial Measures

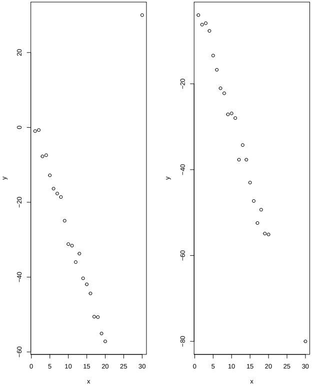
3. Presence of outliers: outliers are extreme observations. An outlier may dramatically change the regression line (when this is the case the outliers in an influential case). Outlier residuals can be identified from residual plots of residuals versus x or $\hat{y}$ as well as box plots.

Figure: Presence of outliers



4. Nonindependent errors: plot residuals versus time to see if there is any cyclical pattern. The residuals are always dependent but this dependency decreases with the sample size.

Figure: Presence of outliers



5. Nonnormality of error terms: make a box plot of the residuals