AMS 572 Data Analysis I
Simple Linear Regression

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Prediction

- Want prediction interval (PI) for future observation given $X = x$ (denoted Y_x)

$$\hat{Y}_x = \hat{\alpha} + \hat{\beta}x$$

- Note: Y_x is a random variable, so we consider the random variable $Y_x - \hat{Y}_x$

$$E(Y_x - \hat{Y}_x) = \alpha + \beta x - (\alpha + \beta x) = 0$$

$$Var(Y_x - \hat{Y}_x) = Var(Y_x) + Var(\hat{Y}_x) - 2Cov(Y_x, \hat{Y}_x)$$

Prediction

- Since the ϵ's are normally distributed, it follows

$$Y_x - \hat{Y}_x \sim N\left(0, \sigma^2 \left\{1 + \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right\}\right)$$

- If σ^2 is not known,

$$\frac{Y_x - \hat{Y}_x}{\sqrt{\text{MSE}\left(1 + \frac{1}{N} + \frac{(x-\bar{X})^2}{\sum (X_i - \bar{X})^2}\right)}} \sim t_{N-2}$$

Prediction

- $(1-\alpha)100\%$ PI for future observation at $X = x$

$$\hat{Y}_x \pm t_{N-2,\alpha/2}\sqrt{\text{MSE}\left(1 + \frac{1}{N} + \frac{(x-\bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$$

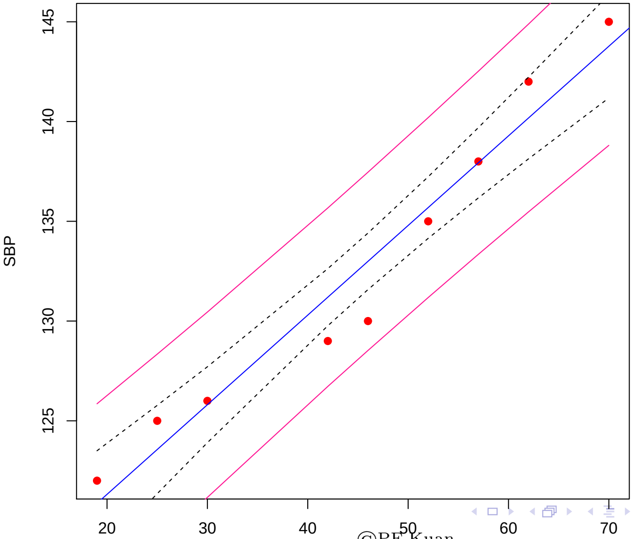
Prediction

- Suppose we want a 95% PI for an individual who is 40 years old:
- Point estimate: $\hat{Y}_{40} = 130.3$
- PI:

$$130.3 \pm 2.365(1.79)\sqrt{1 + \frac{1}{9} + \frac{(40 - 44.8)^2}{2417.59}}$$

$$(125.8, 134.8)$$

Example: SBP and Age



SAS Code

```
data sbpdat;
input age sbp @@;
datalines;
19 122 25 125 30 126 42 129 46 130 52 135 57 138 62 142 70 145 40 NA
;
run;

proc reg data=sbpdat;
model sbp = age;
output out=foo lcl=LCL lclm=LCLM p=P uclm=UCLM ucl=UCL;

proc print data=foo;
run;
```

SAS Output

```
                         The REG Procedure
                           Model: MODEL1
                     Dependent Variable: sbp

         Number of Observations Read                        10
         Number of Observations Used                         9
         Number of Observations with Missing Values          1


                        Analysis of Variance

                                  Sum of          Mean
Source                  DF       Squares        Square    F Value    Pr > F

Model                    1     487.74667     487.74667     151.91    <.0001
Error                    7      22.47555       3.21079
Corrected Total          8     510.22222


            Root MSE              1.79187    R-Square     0.9559
            Dependent Mean      132.44444    Adj R-Sq     0.9497
            Coeff Var             1.35292


                        Parameter Estimates

                          Parameter       Standard
     Variable      DF      Estimate          Error    t Value    Pr > |t|

     Intercept      1     112.33169        1.73773      64.64      <.0001
     age            1       0.44917        0.03644      12.33      <.0001
```

SAS Output

The SAS System 23:41 Saturday, September 5, 2015 4

Obs	age	sbp	P	LCLM	UCLM	LCL	UCL
1	19	122	120.866	118.234	123.498	115.878	125.854
2	25	125	123.561	121.347	125.774	118.780	128.341
3	30	126	125.807	123.905	127.708	121.162	130.451
4	42	129	131.197	129.764	132.629	126.724	135.669
5	46	130	132.993	131.577	134.410	128.526	137.461
6	52	135	135.688	134.145	137.232	131.179	140.198
7	57	138	137.934	136.172	139.696	133.345	142.523
8	62	142	140.180	138.131	142.229	135.474	144.887
9	70	145	143.773	141.181	146.366	138.806	148.741
10	40	.	130.298	128.827	131.770	125.813	134.784

R Code and Output

```
> fit <- lm(sbp~age)
> predict(fit,data.frame(age=40),interval='confidence')
       fit      lwr      upr
1 130.2984 128.8273 131.7696
> predict(fit,data.frame(age=40),interval='prediction')
       fit      lwr      upr
1 130.2984 125.8132 134.7836
```

Sum of Squares Decomposition

- Can decompose total sum of squares

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$
$$SST = SSR + SSE$$

- Total sample variance of the Y's:

$$s_y^2 = \frac{SST}{N-1} = \frac{\sum_i (Y_i - \bar{Y})^2}{N-1}$$

(Unadjusted) r^2

- The unadjusted r^2 is given by

$$r^2 = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

  where $S_{xy} = \sum(X_i - \bar{X})(Y_i - \bar{Y})$, $S_{yy} = SST$,
  $S_{xx} = \sum(X_i - \bar{X})^2$
- *Coefficient of determination*
- Proportion of total variation attributable to regression
- SBP Example

$$r^2 = \frac{487.75}{510.22} = 0.9559$$

Adjusted r^2

- Note that the sample variance of the Y's is $s_y^2 = 63.78$ while MSE $= 3.21$
- Thus X "explains"

$$\frac{63.78 - 3.21}{63.78} = 0.9497$$

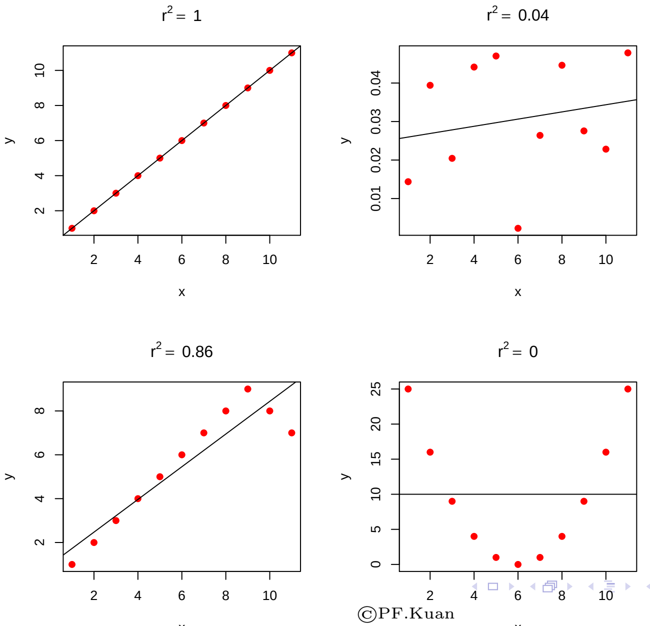
  proportion of the variance in Y.
- This quantity is called the *adjusted* r^2

$$r_a^2 = \frac{s_y^2 - \text{MSE}}{s_y^2} = 1 - \frac{SSE/(N-2)}{SST/(N-1)}$$

Unadjusted r^2

- Proportion of total variation attributable to regression
- Degree of linear association
- Ranges between 0 and 1
- $r^2 = 0 \rightarrow$
- $r^2 = 1 \rightarrow$

Examples of r^2



Linear Regression and 2 Sample t-test

- Suppose we have 2 groups of observations: Y_{1i} for $i = 1, \ldots, n_1$ and Y_{2i} for $i = 2, \ldots, n_2$
- Recall test statistic

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{N - 2}$$

Linear Regression and 2 Sample t-test

- Let

$$N = n_1 + n_2$$

$$(Y_1, \ldots, Y_{n_1}) = (Y_{11}, \ldots, Y_{1n_1})$$

$$(Y_{n_1+1}, \ldots, Y_N) = (Y_{21}, \ldots, Y_{2n_2})$$

$$X_i = \begin{cases} 1 & \text{if group 1} \\ 0 & \text{if group 2} \end{cases}$$

Linear Regression and 2 Sample t-test

- Consider the regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i; i = 1, 2, 3, \ldots, N$$

- Note

$$\sum_i (X_i - \bar{X})^2 = \sum X_i^2 - N\bar{X}^2$$

$$= n_1 - N\left(\frac{n_1}{N}\right)^2 = n_1\left(1 - \frac{n_1}{N}\right) = \frac{n_1 n_2}{N}$$

Linear Regression and 2 Sample t-test

- Can show that
$$\hat{\beta} = \bar{Y}_1 - \bar{Y}_2$$
- and
$$\text{MSE} = s_p^2$$

Linear Regression and 2 Sample t-test

- Therefore:

$$t = \frac{\hat{\beta}}{\sqrt{\text{MSE}/\sum_i (X_i - \bar{X})^2}} = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{N/(n_1 n_2)}}$$

Example: A study is conducted to compare the effect of a new drug on shrinking tumor size. 20 patients are enrolled in this study, in which 10 are assigned to placebo and 10 are assigned to the new drug, and the tumor size of each patient is measured after two weeks. Test if there is any difference in the mean tumor size at $\alpha = 0.05$.

```
> placebo
 [1] 49.43952 49.76982 51.55871 50.07051 50.12929 51.71506
 [8] 48.73494 49.31315 49.55434
> drug
 [1] 49.22408 48.35981 48.40077 48.11068 47.44416 49.78691
 [8] 46.03338 48.70136 47.52721
```

R Code and Output

```
> t.test(placebo,drug,var.equal=TRUE)

Two Sample t-test

data:  placebo and drug
t = 4.1858, df = 18, p-value = 0.0005554
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
 0.9294306 2.8025768
sample estimates:
mean of x mean of y
 50.07463  48.20862
```

R Code and Output

```
> grp <- rep(c(0,1),each=10)
> fit <- lm(c(placebo,drug)~grp)
> summary(fit)

Call:
lm(formula = c(placebo, drug) ~ grp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17524 -0.64668  0.02527  0.41290  1.64044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.0746     0.3152 158.855  < 2e-16 ***
grp          -1.8660     0.4458  -4.186 0.000555 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.9968 on 18 degrees of freedom
```

Diagnostics

- Assumptions for linear regression
  1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
  2. X's are fixed constants
  3. ϵ_i iid $\sim N(0, \sigma^2)$

Diagnostics

- Assumptions: Linear model and homogeneity of variance
- *Residual plot*: Scatterplot of

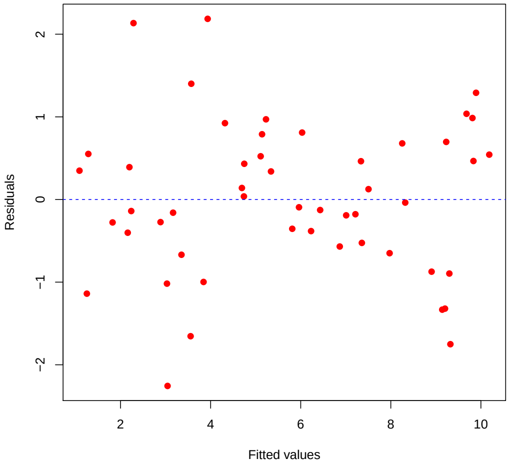
$$(\hat{Y}_i, r_i) = (\hat{Y}_i, Y_i - \hat{Y}_i)$$

- If we see lack of homogeneity of variance or linearity, consider transformations

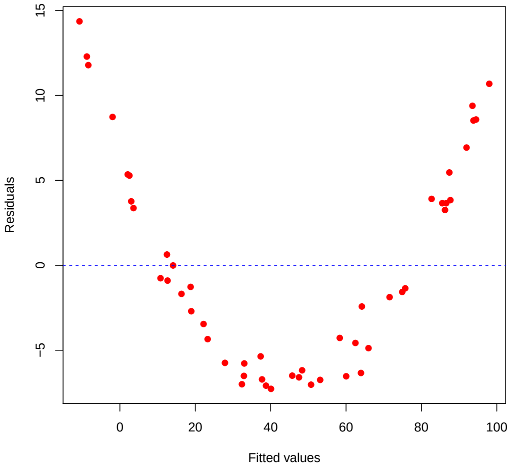
Diagnostics

- The following three slides are prototypical residual plots indicating
  1. linear regression model is appropriate
  2. assumption of linearity questionable
  3. assumption of constant variance questionable

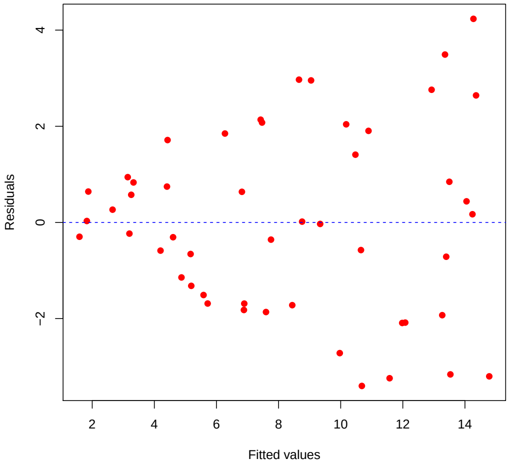
Residual plots



Residual plots



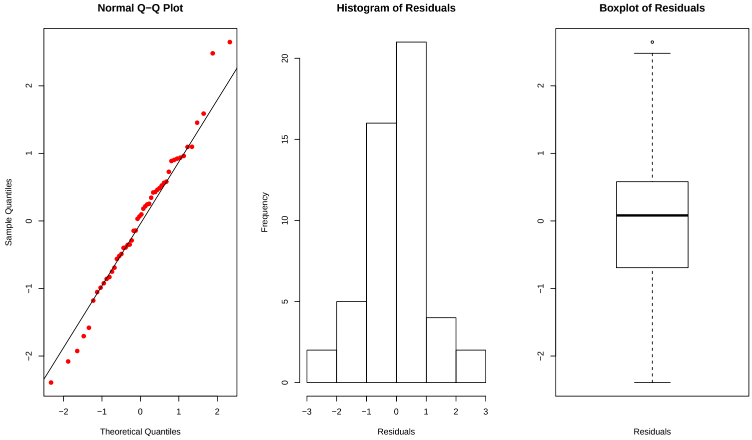
Residual plots



Normality Diagnostics

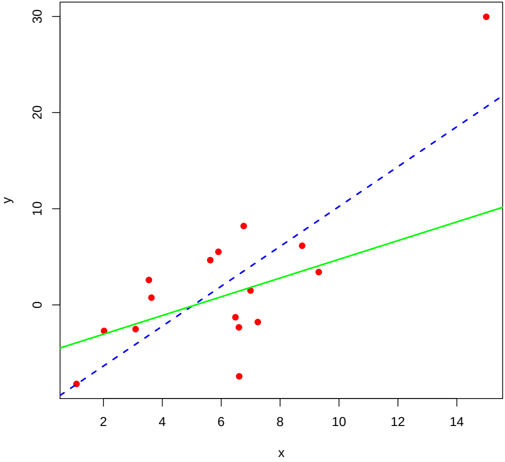
- Assumption: ϵ_i's are normally distributed
- This assumption is not as important if N is large (CLT)
- Inference robust to small departures from normality
- Violations of other assumptions can suggest non-normality
- qq-plot, histogram, boxplot of residuals

Residual plots



Regression: Diagnostics

- Beware influential observations; always check scatterplot



Remedial Measures

- Transformations, e.g., Box Cox transformation

$$\frac{y^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0$$

$$\log(y) \text{ if } \lambda = 0$$

- Multiple regression, e.g., $Y = \alpha + \beta_1 X + \beta_2 X^2$
- Nonparametric procedures, e.g., Kendall's tau
- More sophisticated models allowing for
  - dependencies/clusters (e.g., GEE)
  - heterogeneity of variance (e.g., weight least squares)

Box Cox transformation

```
> library(MASS)
> y <- c(1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 6, 7, 8)
> x <- c(7, 7, 8, 3, 2, 4, 4, 6, 6, 7, 5, 3, 3, 5, 8)
> fit <- lm(y~x)
> plot(fit)
> bc1 <- boxcox(y ~ x)
> lambda1 <- bc1$x[which.max(bc1$y)]
> fit.new <- lm(((y^lambda1-1)/lambda1) ~ x)
> plot(fit.new)
```

Reference
https://www.statology.org/box-cox-transformation-in-r/