STA302/1001 - Methods of Data Analysis I

(Week 02 - Lecture B)

Wei (Becky) Lin

May 22-26, 2017

Review of Week 2 -Lecture A

- Review of distribution theory.
- Inference for SLR.
- Interval estimation of mean response.
- Prediction interval.
- Difference between prediction interval and confidence interval.
- ANOVA approach.

Week 02 - Lecture B: Learning objectives & Outcomes

- Normal correlation model
- Correlation coefficient, r
- Inference for Correlation Coefficient
- Coefficient of determination, R^2
- Interpretation of R^2
- Spearman correlation coefficient, r_S
- r versus r_S
- Regression with dummy variable
- Ch3: diagnostic of predictor variable
- Influential point and definition of leverage of a data point
- properties associated with definition of leverage

Correlation Analysis

- For linear regression model, we assume that the X values are known constants
  - E.g. X = amount of fertilizer, Y = crop yield
  - This assumes level of X can be controlled
- In many cases, the X is also random
  - E.g. X = height of a person, Y = weight of a person
  - The X variable cannot be controlled, both are random
- An alternative to regression in such cases is *Correlation Analysis*

Normal correlation model

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \rho\sigma_Y\sigma_X \\ \rho\sigma_Y\sigma_X & \sigma_X^2 \end{pmatrix} \right)$$

- Data (Y_i, X_i) is from bivariate normal distribution with density

$$f(y_i, x_i) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_i-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{y_i-\mu_y}{\sigma_y}\right)\left(\frac{x_i-\mu_x}{\sigma_x}\right) + \left(\frac{x_i-\mu_x}{\sigma_x}\right)^2 \right] \right\}$$

  - Marginal distribution for X: $X \sim N(\mu_x, \sigma_x^2)$ $\leftarrow f_X(x) = \int_Y f(x,y)\,dy$
  - Marginal distribution for Y: $Y \sim N(\mu_y, \sigma_y^2)$ $\leftarrow f_Y(y) = \int_X f(x,y)\,dx$
  - Covariance between X and Y: $\mathrm{Cov}(X, Y) = \rho\sigma_x\sigma_y$ where

$$\rho = \frac{Cov(X,Y)}{\sigma_x\sigma_y} = cor(X,Y)$$

  - Relationship between X, Y is reflected by the coefficient of correlation parameter ρ

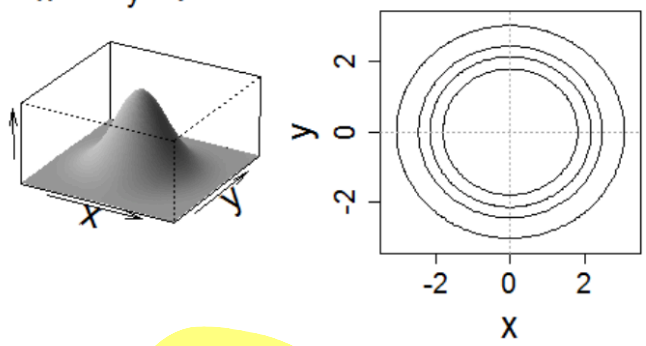
Normal correlation model

- MLE estimation for correlation model
  - $\hat{\mu}_x = \bar{X}$
  - $\hat{\mu}_y = \bar{Y}$
  - $\hat{\sigma}_x^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 = S_{XX}/n$: biased estimator
  - $\hat{\sigma}_y^2 = \frac{1}{n}\sum(Y_i - \bar{Y})^2 = S_{YY}/n$: biased estimator
  - $$\hat{\rho} = r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum(X_i - \bar{X})^2][\sum(Y_i - \bar{Y})^2]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$ *
  - r is the Pearson Correlation Coefficient.

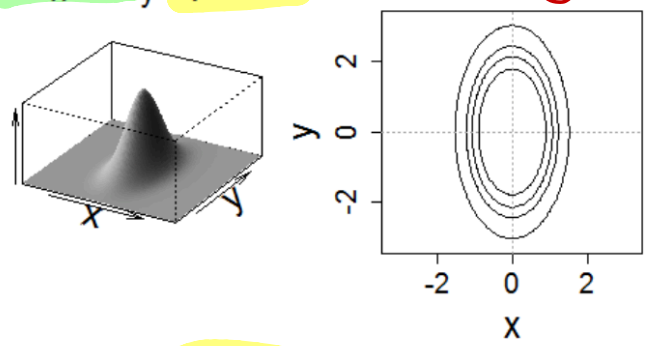
Normal correlation model- Bivariate Normal

project $f(x,y)$ down to x-y plane.

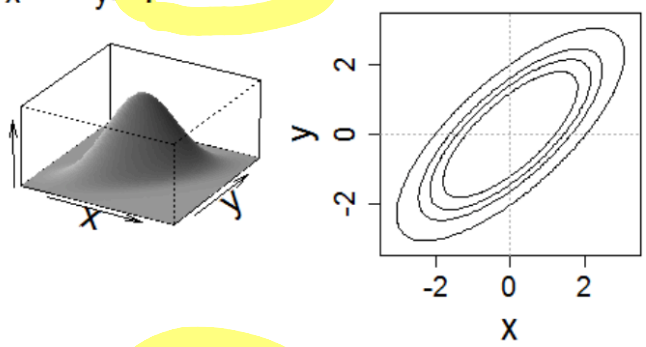
$\sigma_x = \sigma_y$, $\rho = 0$



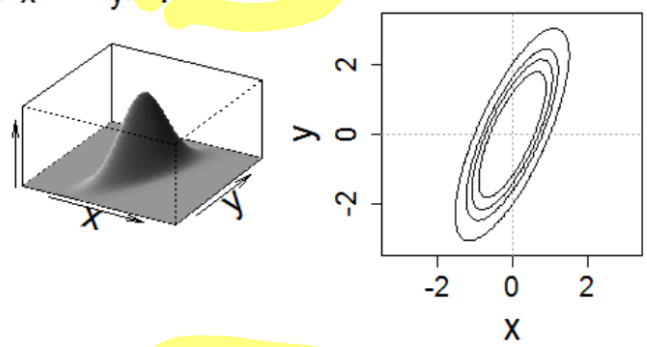
$2\sigma_x = \sigma_y$, $\rho = 0$



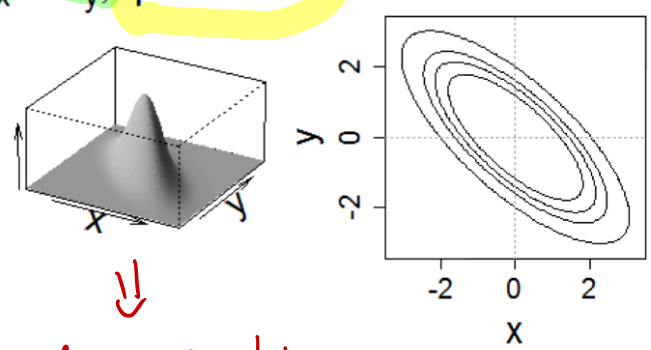
$\sigma_x = \sigma_y$, $\rho = 0.75$



$2\sigma_x = \sigma_y$, $\rho = 0.75$



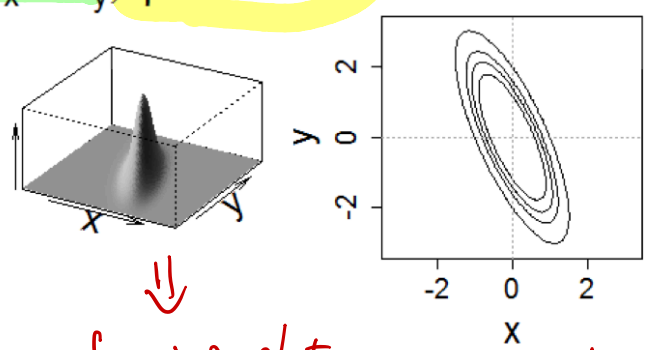
$\sigma_x = \sigma_y$, $\rho = -0.75$



$f(x,y)$ plot

contour

$2\sigma_x = \sigma_y$, $\rho = -0.75$



$f(x,y)$ plot

contour

Normal correlation model - Conditional Inference

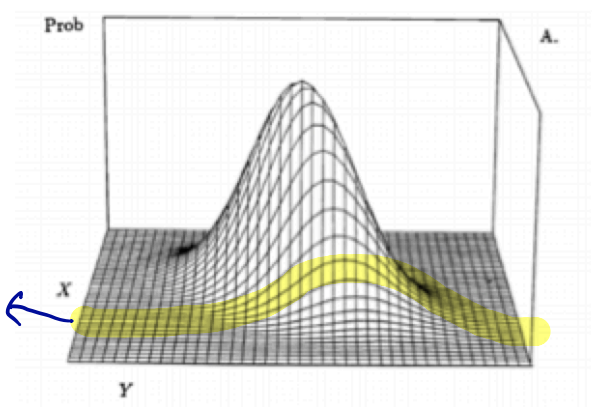
- The most common use of normal correlation model is to make conditional inference
  - Given an observed value of X, what can we say about Y?

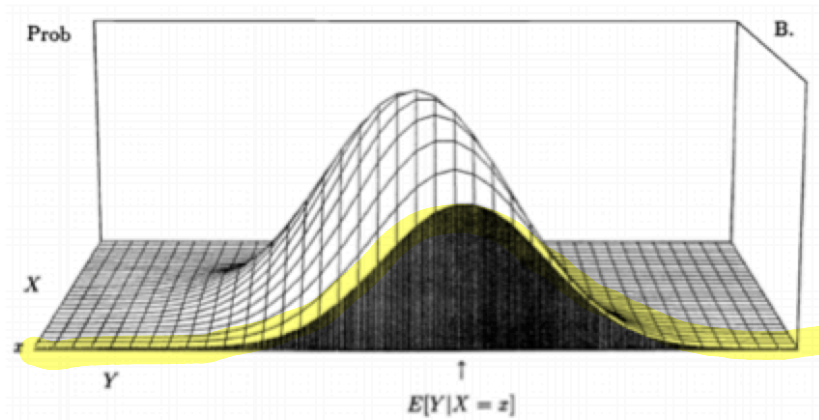
$$\frac{f(x,y)}{f(x)} = f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2_{y|x}}} \exp\left\{ -\frac{1}{2}\left(\frac{y-\alpha_{y|x}-\beta_{y|x}x}{\sigma_{y|x}}\right)^2 \right\}$$

f(Y|x) is the condtl density function.

  - $\alpha_{y|x} = \mu_y - \mu_x\beta_{y|x}, \quad \beta_{y|x} = \rho\frac{\sigma_y}{\sigma_x}, \quad \sigma^2_{y|x} = \sigma^2_y(1-\rho^2)$
  - For given X, Y follows normal with constant variance, and mean is a linear function of X.
  - Normal correlation model is closely related to simple linear regression with normal errors

X=x





density for Y|X=x

Normal correlation model - Conditional Inference

Three important characteristics of the conditional distribution of $Y|X$

- Normality:

$$(Y|X) \sim N(\alpha_{y|x} + \beta_{y|x}X, \sigma^2_{y|x})$$

SLR:

$Y \sim N(\beta_0+\beta_1 X, \sigma^2)$

- Linear regression: the mean of the conditional probability distribution of Y fall on a straight line and is a linear function of X:

$$E(Y|X) = \alpha_{y|x} + \beta_{y|x}X$$

SLR:

$E(Y) = \beta_0+\beta_1 X$

- Constant variance: all the conditional probability distributions of Y given different level of X have the same variance.

$$Var(Y|X) = \sigma^2_{y|x} = \sigma^2_y(1-\rho^2)$$

SLR

$V(Y) = V(\varepsilon_i) = \sigma^2$

- free of X;
- view it as a constant variance

.

Normal correlation model

Note that $SSTO = \sum(Y_i - \bar{Y})^2 = S_{YY}$

Show that

$$\textcircled{1}\ \hat{\beta}_{y|x} = r\sqrt{S_{YY}/S_{XX}} = b_1, \quad \textcircled{2}\ \hat{\alpha}_{y|x} = \bar{Y} - b_1\bar{X} = b_0$$

$$\textcircled{3}\ \hat{\sigma}^2_{y|x} = \hat{\sigma}^2_y(1-\hat{\rho}^2) = \frac{SSE}{n} \neq MSE$$

Proof:

- $r = \dfrac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\left[\sum_{i=1}^n (X_i-\bar{X})^2 \sum_{i=1}^n (Y_i-\bar{Y})^2\right]^{1/2}} = \dfrac{S_{XY}}{[S_{XX}S_{YY}]^{\frac{1}{2}}} = \boxed{\dfrac{S_{XY}}{S_{XX}}}\left(\dfrac{S_{XX}}{S_{YY}}\right)^{\frac{1}{2}}$ (b_1)

$$\Rightarrow r = b_1\left(\frac{S_{XX}}{S_{YY}}\right)^{\frac{1}{2}} \Leftrightarrow \boxed{b_1 = r\left(\frac{S_{YY}}{S_{XX}}\right)^{\frac{1}{2}}}$$

$$\textcircled{1}\ \hat{\beta}_{Y|X} = r\left(\frac{S_{YY}}{S_{XX}}\right)^{\frac{1}{2}} = b_1\left(\frac{S_{XX}}{S_{YY}}\right)^{\frac{1}{2}}\left(\frac{S_{YY}}{S_{XX}}\right)^{\frac{1}{2}} = b_1$$

$$\textcircled{2}\ \hat{\alpha}_{Y|X} = \hat{\mu}_Y - \hat{\mu}_X\hat{\beta}_{Y|X} = \bar{Y} - \hat{\beta}_{Y|X}\bar{X} = \bar{Y} - b_1\bar{X} = b_0$$

$$\textcircled{3}\ \sigma^2_{Y|X} = \frac{S_{YY}}{n}(1-r^2) = \frac{S_{YY}}{n}\left(1 - \frac{b_1^2 S_{XX}}{S_{YY}}\right) = \frac{1}{n}(S_{YY} - b_1^2 S_{XX})$$

$$= \frac{1}{n}(SSTO - SSR)$$

$$= \frac{SSE}{n}$$

Q.E.D.

Week 3 ↓ SSR

Normal Correlation model

- Conditional inference for normal correlation model is equivalent to (MLE) inference for linear regression. So, what is the difference?
  - Interpretation of sampling distributions (in Correlation Analysis X is not fixed, data (X,Y) is from the bivariate Normal distribution.)
  - Equivalence holds only for Normality
  - Correlation analysis does not distinguish between response and explanatory variables. Moreover, it does not afford a causal relationship between X and Y (we cannot control X)
  - If experimental data is presented, the linear model does afford a causal relationship.
- Can we still use normal error model if (Y,X) are not bivariate normal? It can be shown that all results on estimation, testing and prediction obtained from SLR apply if both Y and X are random and the following two conditions hold:
  - $Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ ✓
  - The distribution for RV X , $f(x_i)$, does not involve the parameters $\beta_0, \beta_1, \sigma^2$ ✓

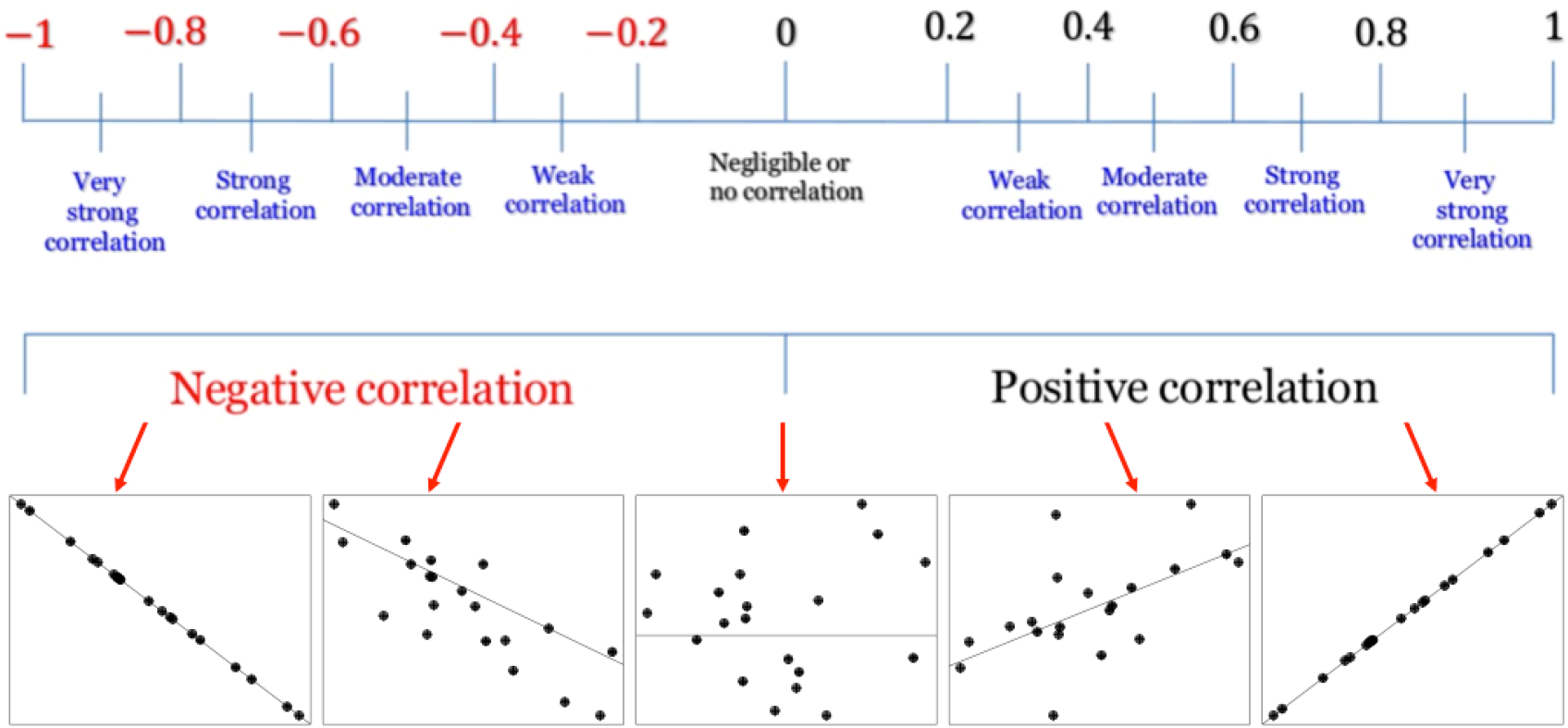
(Pearson) Coefficient of Correlation

- r is the basic quantity in correlation analysis

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum(X_i - \bar{X})^2][\sum(Y_i - \bar{Y})^2]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

- r takes values in [-1,1], the $\pm$ signs are used for positive and negative linear correlations, respectively.
- r measures strength & direction of linear relationship
  - Value of -1 or $+1$ indicates a perfect positive or negative correlation, all data points all lie exactly on a straight line
  - Value of 0.0 indicates no linear correlation
  - Positive values indicate a direct relationship, X and Y move in the same directions.
  - Negative values indicate an inverse relationship, X and Y move in opposite directions.
- Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

Interprestation of Correlation coefficient



- $0.2 \leq |r| \leq 0.4$: weak correlation.
- $0.4 < |r| \leq 0.6$: moderate correlation.
- $0.6 < |r| \leq 0.8$: strong correlation.
- $|r| > 0.8$: very strong correlation.

Inference on Correlation Coefficient

Testing: when we assume the population is bivariate normal,

↑ need this assumption since $\rho = cor(X, Y)$ ↘ RVs.

$$H_0 : \rho = 0, \quad H_a : \rho \neq 0$$

- If (Y, X) is jointly Normal distributed, then $\rho = 0$ implies $X \perp Y$.
- Test statistics

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\Big|_{H_0} \sim t_{n-2}$$

- Make decision: reject H_0 if $|t^*| > t_{1-\alpha/2;n-2}$
- Test t^* is equivalent to t-test for $H_0 : \beta_1 = 0$

Inference on Correlation Coefficient

Show under H_0,

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1 - 0}{s(b_1)}$$

→ test statistic for: $H_0: \beta_1 = 0$ VS $H_a: \beta_1 \neq 0$

Proof:

From slide 10, we have

$$r = b_1 \left(\frac{S_{xx}}{S_{YY}}\right)^{\frac{1}{2}}, \quad 1 - r^2 = 1 - \frac{b_1^2 S_{xx}}{S_{YY}} = 1 - \frac{SSR}{SSTO} = \frac{SSE}{SSTO} \qquad S_{YY} = SSTO$$

$$\Rightarrow t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1 \left(\frac{S_{xx}}{S_{YY}}\right)^{\frac{1}{2}} \sqrt{n-2}}{(SSE/SSTO)^{\frac{1}{2}}} = \frac{b_1 (S_{xx})^{\frac{1}{2}} \sqrt{n-2} / (SSTO)^{\frac{1}{2}}}{(SSE/SSTO)^{\frac{1}{2}}}$$

$$= \frac{b_1 (S_{xx})^{\frac{1}{2}} \sqrt{n-2}}{(SSE)^{\frac{1}{2}}} = \frac{b_1 \sqrt{S_{xx}}}{\sqrt{SSE/(n-2)}} = \frac{b_1 \sqrt{S_{xx}}}{\sqrt{MSE}}$$

$$= \frac{b_1}{\sqrt{MSE/S_{xx}}} = \frac{b_1}{s(b_1)}$$

Since $s^2(b_1) = \frac{MSE}{S_{xx}}$ and $s(b_1) = \sqrt{\frac{MSE}{S_{xx}}}$

Inference on Correlation Coefficient

Interval estimation using Fisher Z transformation

Fisher Z transformation

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

When $n > 25$, the distribution of Z is approximately normal with mean and variance

$$\mu_Z = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right), \quad \sigma_Z^2 = \frac{1}{n-3}$$

That is

$$\frac{Z-\mu_Z}{\sigma_Z} \sim N(0,1) \quad \text{given n>25}$$

Therefore, 1-α confidence limits for μ_Z are

$$Z \pm Z_{1-\alpha/2}\sigma_Z$$

A further transformation back to ρ gives the CI for ρ.

Coefficient of Determination

Define R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- The measure R^2 is called the coefficient of determination.
- R^2 is a proportion. Since $0 \leq SSE \leq SSTO$, it follows that $0 \leq R^2 \leq 1$
- R^2 measures strength of linear relationship. $R^2 = 1$, all data points fall perfectly on the regression line and X accounts for all of the variation in Y. $R^2 = 0$, the estimated regression line is pefectly horizontal. X accounts none of the variation in Y.
- Applies to both correlation and regression analysis. Show $R^2 = r^2$
- Interpretation of R^2: *
  - "$R^2 \times$ 100 percent of the variation in Y is reduced by taking into account predictor X"
  - "$R^2 \times$ 100 percent of the variation in Y is 'explained by' the variation in predictor X"

Show $R^2 = r^2$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = r^2$$

Proof:

$SSTO = \Sigma(Y_i - \bar{Y})^2 = S_{YY}$

$r = b_1 \left(\frac{S_{XX}}{S_{YY}}\right)^{1/2} \Rightarrow r^2 = \frac{b_1^2 S_{XX}}{S_{YY}}$

$= \frac{SSR}{SSTO}$

$= R^2$

From R^2 to get r

$$r = \pm\sqrt{R^2} = \text{sign}(b_1)\sqrt{R^2}$$

- If b_1 is negative, then r takes a negative sign.
- If b_1 is positive, then r takes a positive sign.

$\text{sign}(r) = \text{sign}(b_1)$

$$r = b_1\sqrt{\frac{S_{XX}}{S_{YY}}}$$

- The estimated slope b_1 of the regression line and the correlation coefficient r always share the same sign.
- If the estimated slope b_1 of the regression line is 0, then the correlation coefficient r must also be 0.

Limitations of R^2

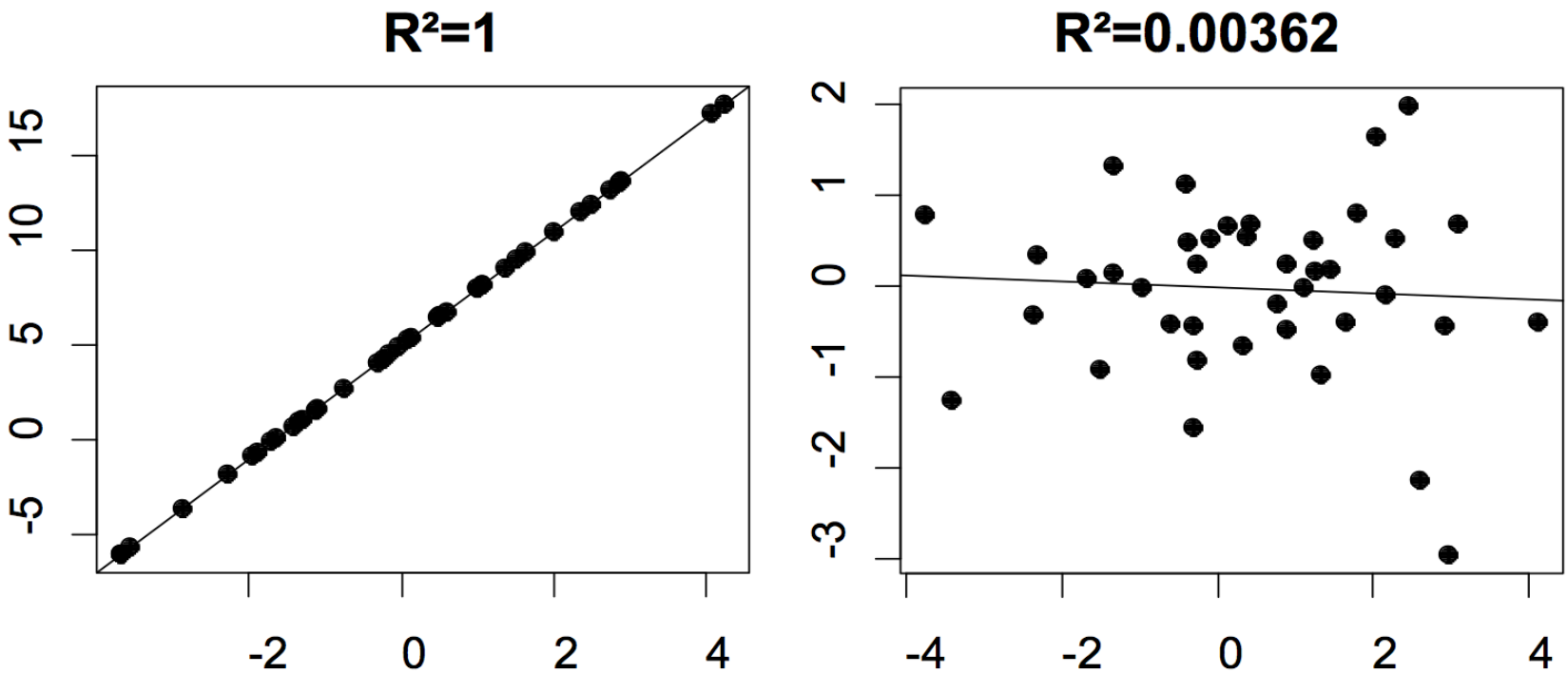
- R^2 is frequently used for assessing and comparing model fits, but
  - A high R^2 does not necessarily imply that you can make useful predictions.
  - A high R^2 does not necessarily imply that the estimated line is a good fit.
  - A low R^2 does not necessarily imply that X and Y are not related or independent.
- R^2 measures degree of *linear* association but does not measure the evidence of a linear relationship between X and Y.
- R^2 usually can be made larger by including a larger number of predictors.(More predictors, MSE goes down,SSTO remains unchanged, so R^2 goes up)
- Adjusted R^2, R_a^2(p=number of predictors in the model+1, +1 for β_0)

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO} = 1 - (n-1)\frac{MSE}{SSTO}$$

$$R_a^2 = 1 - \frac{(1-R^2)(n-1)}{n-p}$$

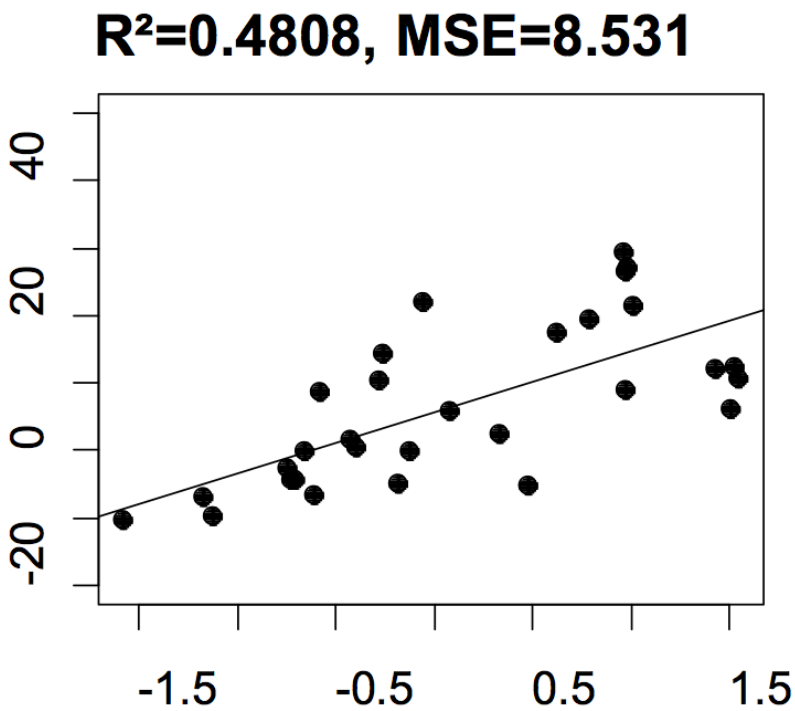
Interpretation of R^2

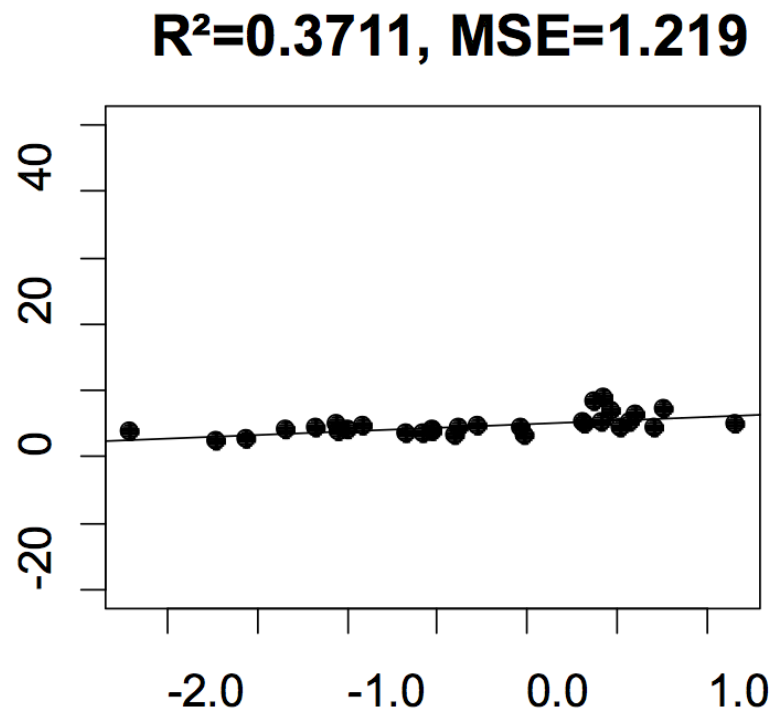
- R^2 measures degree of *linear* association between X and Y.



Interpretation of R^2

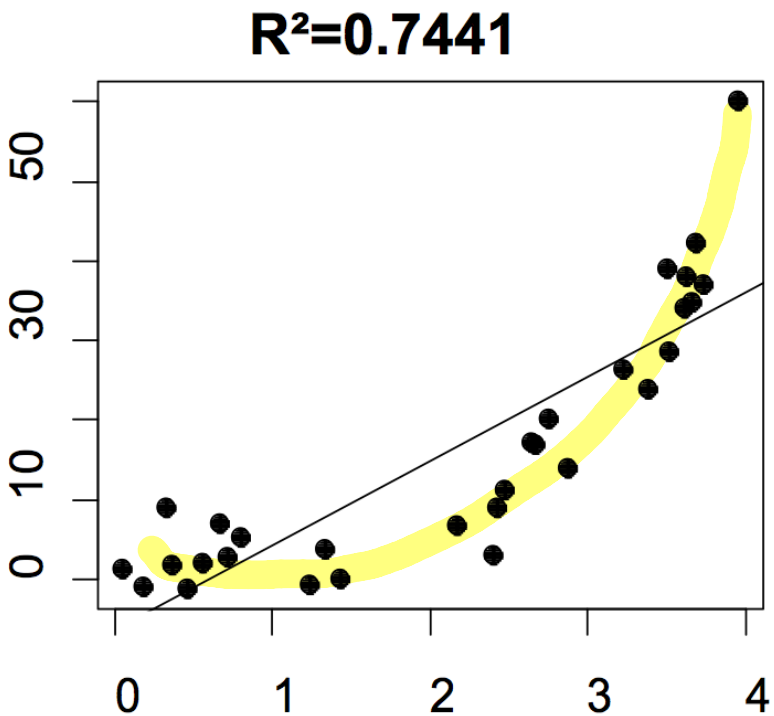
- R^2 measures only a relative reduction from SSTO.
- R^2 might be large but MSE may still be too large for inference to be useful in prediction.
- R^2 might be small but MSE may still be small which is useful in prediction

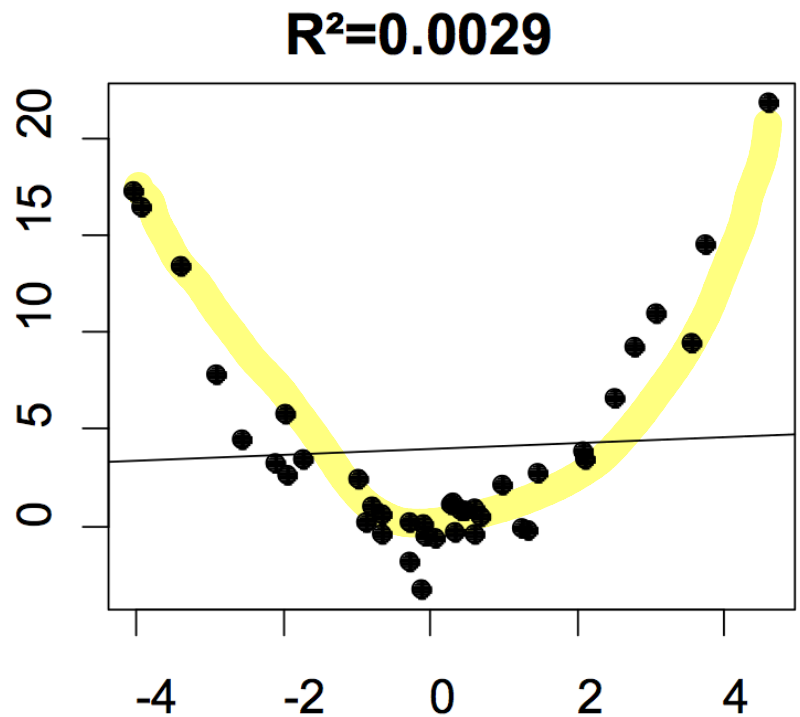




Interpretation of R^2

- R^2 measures degree of *linear* association.
- R^2 might be large or small if the true regression association between X and Y is curvilinear.





Spearman Rank Correlation

- Pearson's correlation measures the linear relationship between X and Y.
- If the relationship is not linear, it is not a good choice for a summary statistics.
- Provided the relationship is *monotonic*, we can use Spearman's correlation
  - Monotonic: dependent variable Y either never increases or never decreases as its independent variable X increases.
- Spearman Correlation Coefficient:

$$r_s = \frac{\sum(R_i^y - \bar{R}^y)(R_i^x - \bar{R}^x)}{[\sum(R_i^y - \bar{R}^y)^2 \sum(R_i^x - \bar{R}^x)^2]^{1/2}} \in [-1, 1]$$

  - R_i^y is the rank value of Y_i, so is R_i^x for X_i. E.g. $Y = (2, 5, 3), R^y = (1, 3, 2)$.
  - $\bar{R}^y, \bar{R}^x$ are the means of R_i^y, R_i^x respectively. $\bar{R}^y = \bar{R}^x = (n+1)/2$

r versus r_p

- Pearson correlation coefficient

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2]^{1/2}}$$

  - Assume normality: both X and Y are normally distributed.
  - Assume linearity: a straight line relationship between each of the variables in the analysis.
  - Assume homoscedasticity: data is normally distributed about the regression line.

- Spearman correlation coefficient

$$r_s = \frac{\sum(R_i^y - \bar{R}^y)(R_i^x - \bar{R}^x)}{[\sum(R_i^y - \bar{R}^y)^2 \sum(R_i^x - \bar{R}^x)^2]^{1/2}}$$

  - A non-parametric method to measure the degree of association between X and Y.
  - No assumptions about the distribution of the data.
  - Assume only a monotonic relationship between X and Y.

Example 1: r versus r_p

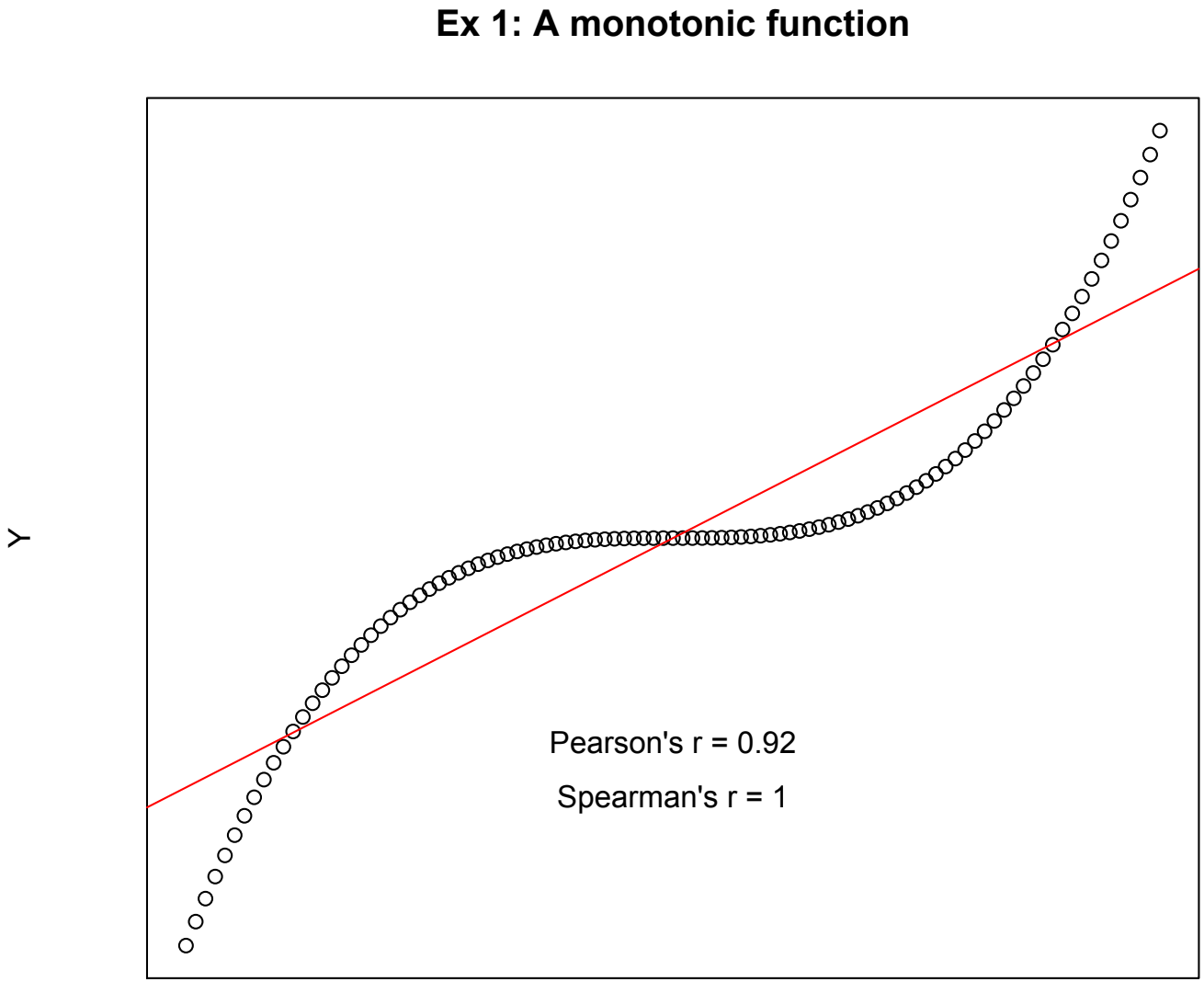
```
X <- -50:50
Y <- X^3
fit <- lm(Y ~ X)

rP <- sqrt(summary(fit)$r.squared)
rP2 <- cor(Y, X, method="pearson")
rS <- cor(Y, X, method="spearman")
rS2 <- cor(order(Y), X, method="pearson")
cbind(rP,rP2,rS,rS2)
```

$R^2 = r^2$ ⇐ } are the same
using cor(X,Y) ⇐ }
specifies method ⟶
apply r formula to rank values ⟶

```
##              rP       rP2 rS rS2
## [1,] 0.916575 0.916575   1   1
```

```
# plot(X, Y, xaxt="n", yaxt="n", main="Ex 1: A monotonic function")
# abline(coef(fit), col="red")
# text(median(X), min(Y) + 0.25*(max(Y)-min(Y)),
#      paste("Pearson's r =", round(rP, 2)))
# text(median(X), min(Y) + 0.18*(max(Y)-min(Y)),
#      paste("Spearman's r =", round(rS, 2)))
```

Example 2: r versus r_p



Example 1: r versus r_p

```r
set.seed(2016)
X <- seq(0, 10, 0.1)
Y <- 2*X^4 + rnorm(length(X), 0, 625)
fit <- lm(Y ~ X)

rP <- sqrt(summary(fit)$r.squared)
rP2 <- cor(Y, X, method="pearson")
rS <- cor(Y, X, method="spearman")
rS2 <- cor(order(Y), X, method="pearson")
cbind(rP,rP2,rS,rS2)
```

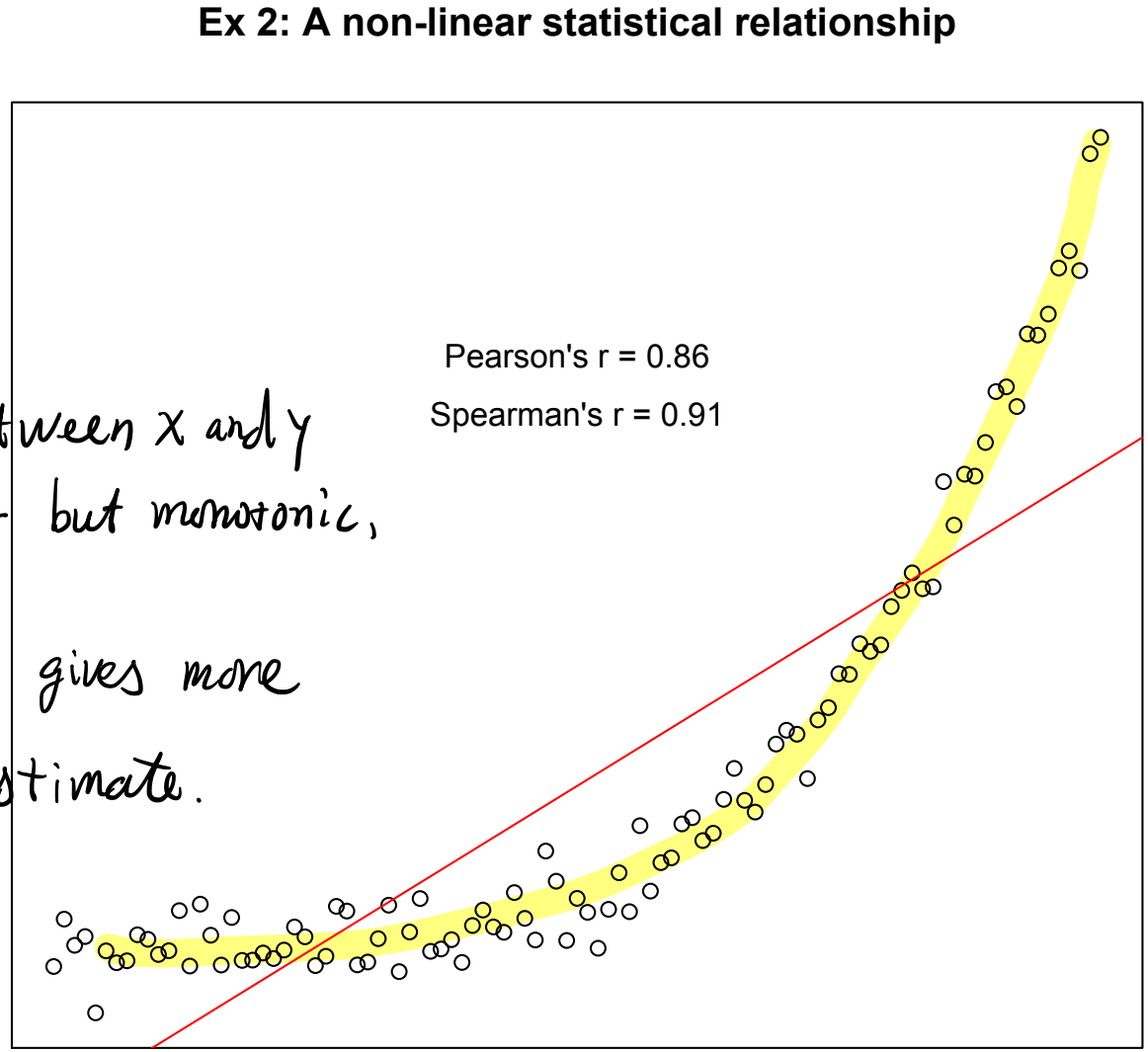
same result

same result

```
##              rP       rP2       rS      rS2
## [1,] 0.8613084 0.8613084 0.909074 0.909074
```

```r
# plot(X, Y, xaxt="n", yaxt="n",
#      main="Ex 2: A non-linear statistical relationship")
# abline(coef(fit), col="red")
# text(median(X), min(Y) + 0.75*(max(Y)-min(Y)),
#      paste("Pearson's r =", round(rP, 2)))
# text(median(X), min(Y) + 0.68*(max(Y)-min(Y)),
#      paste("Spearman's r =", round(rS, 2)))
```

Example 2: r versus r_p



The true association between X and Y is curvilinear but monotonic,

> Spearman's r gives more accurate estimate.

Regression with Dummy Variable

Regression with Dummy Variable

- So far we have looked at data sets involving two numerical variables, X and Y.
- What if one (or both) of the variables are categorical instead?
  - for example: binary variable, takes value 0 or 1.
- If both variables are binary, can use a 2×2 contingency table.
- If Y, the dependent variable is binary, can use Logistic Regression.
  - Both techniques covered in STA303.

Regression with Dummy Variable

- We will focus on the case where the predictor variable X is binary.
- Example: Height and gender

Sex	Height (cm)
male	183
male	178
female	172
female	154
female	159
female	165

iFemale	Height (cm)
0	183
0	178
1	172
1	154
1	159
1	165

has two levels

need only one dummy variable.

Regression with Dummy Variable

- For a categorical variable with two levels, we need only one indicator variable
  - Continue the example: X takes value 1 for female and "0"for male.
- For a categorical variable with K levels, we need $K-1$ indicator variables.
  - For a variable takes value 1, 2 or 3.

$$X_1 = \begin{cases} 1, & X = 2 \\ 0, & X = 1 \text{ or } 3 \end{cases}, \quad X_2 = \begin{cases} 1, & X = 3 \\ 0, & X = 1 \text{ or } 2 \end{cases}$$

  - Reference level is 1,this occurs when $X_1 = 0$ and $X_2 = 0$.

Regression with Dummy Variable

The Example of height vs gender, model

$$Y_i = \beta_0 + \beta_1 X + \epsilon_i$$

β_0 = mean of Y for male (refrence) group

β_1 = mean of Y for female − mean of Y for male

- All of the same assumptions hold for this model
- X here is the female indicator

$$X = \begin{cases} 1, & Female \\ 0, & Male \end{cases}$$

b_0 = avg male height

b_1 = avg female height − avg male height

- The slope β_1 gives the true difference in means between males and females, which can be estimated by b_1 from data.
- Proof:
  - $E(Y|female) = E(Y|X == 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$
  - $E(Y|male) = E(Y|X == 0) = \beta_0 + \beta_1 \times 0 = \beta_0$
  - $\beta_1 = E(Y|female) - E(Y|male) = E(Y|X == 1) - E(Y|X == 0)$

Example: height vs gender

```r
X = c( rep(0,2),rep(1,4))          # X==1 for female, X==0 for male
Y= c( 183,178,172,154,159,165)     # height (cm) of a subject
summary(lm(Y~X))                   # summary of a SLR model
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##    1    2    3    4    5    6
##  2.5 -2.5  9.5 -8.5 -3.5  2.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  180.500      4.918  36.701 3.29e-06 ***
## X            -18.000      6.023  -2.988   0.0404 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.955 on 4 degrees of freedom
## Multiple R-squared:  0.6906, Adjusted R-squared:  0.6133
## F-statistic:  8.93 on 1 and 4 DF,  p-value: 0.0404
```

= 162.5 − 180.5

```r
c(mean(Y[X==0]), mean(Y[X==1]))
```

From data, find avg for male, female

```
## [1] 180.5 162.5
```

Regression with Dummy variables: Iris data set in R

```
data(iris) # load in data set: iris
head(iris) # have a look of the first 6 data lines
```

Categorical variable ↓ has 3 categories

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
is.factor(iris$Species) # is Species a factor variable in this data set
```

```
## [1] TRUE
```

```
levels(iris$Species)    # see detail of level information of Species
```

```
## [1] "setosa"     "versicolor" "virginica"
```

Level 1 Level 2 Level 3

Regression with Dummy variables: Iris data set in R

```
tapply(iris$Petal.Width,iris$Species,mean) # mean for each species
```

```
##     setosa versicolor  virginica
##      0.246      1.326      2.026
```

Model:

- Patel.Width $=\beta_0 + \beta_1 Sp_1 + \beta_2 Sp_2 + \epsilon_i$. Let PW denotes Patel.Width.
- where species $\in$ (setosa, versicolor, virginica)

$$Sp_1 = \begin{cases} 1, & \text{Species is versicolor} \\ 0, & \text{otherwise} \end{cases}, Sp_2 = \begin{cases} 1, & \text{Species is virginica} \\ 0, & \text{otherwise} \end{cases}$$

  - when $Sp_1 = Sp_2 = 0$, this refers to the base level species: setosa

- $E(PW|setosa) = E(PW|Sp_1 = Sp_2 = 0) = \beta_0, \hat{\beta}_0 = b_0 = 0.246$
- $E(PW|versicolor) = E(PW|Sp_1 = 1, Sp_2 = 0) = \beta_0 + \beta_1$
- $E(PW|virginica) = E(PW|Sp_1 = 0, Sp_2 = 1) = \beta_0 + \beta_2$
- $\beta_1 = E(PW|versicolor) - E(PW|setosa), \hat{\beta}_1 = b_1 = 1.326 - 0.246 = 1.08$
- $\beta_2 = E(PW|virginica) - E(PW|setosa), \hat{\beta}_2 = b_2 = 2.026 - 0.246 = 1.78$
- $\widehat{PW} = 0.246 + 1.08 Sp_1 + 1.78 Sp_2$

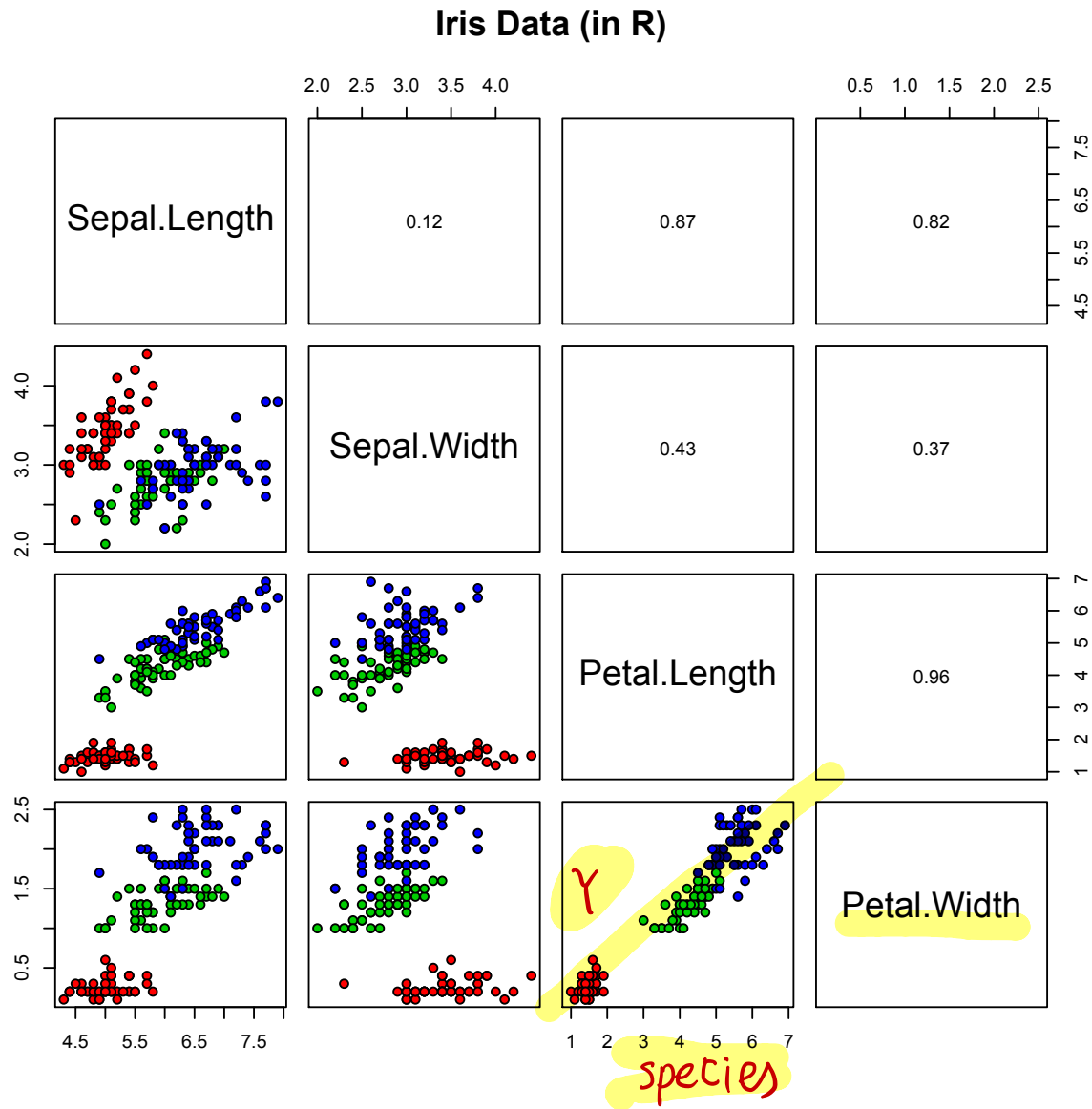
Regression with Dummy variables: Iris data set in R

R code to generate Pairs Scatter Plots

```
pairs(iris[1:4], main = "Iris Data (in R)", pch = 21,
      bg = c("red","green3","blue")[unclass(iris$Species)] )

# change the upper pannel with with Peareson correlation coefficient
panel.pearson <- function(x, y, ...) {
  horizontal <- (par("usr")[1] + par("usr")[2]) / 2;
  vertical <- (par("usr")[3] + par("usr")[4]) / 2;
  text(horizontal, vertical, format(abs(cor(x,y)), digits=2))
  }

pairs(iris[1:4], main = "Iris Data (in R)", pch = 21,
      bg = c("red","green3","blue")[unclass(iris$Species)],
      upper.panel=panel.pearson)
```

Regression with Dummy variables: Iris data set in R

- Pairs scatter plot



Regression with Dummy variables: Iris data set in R

```
fit=with(iris,lm(Petal.Width~factor(Species)))
summary( fit )
```

```
##
## Call:
## lm(formula = Petal.Width ~ factor(Species))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.626 -0.126 -0.026   0.154   0.474
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.24600    0.02894    8.50 1.96e-14 ***
## factor(Species)versicolor   1.08000    0.04093   26.39  < 2e-16 ***
## factor(Species)virginica    1.78000    0.04093   43.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2047 on 147 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.9279
## F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

compare with slide 37

Regression with Dummy variables: Iris data set in R

```
PW=iris$Petal.Width
Sp1=ifelse(iris$Species=="versicolor", 1, 0)
Sp2=ifelse(iris$Species=="virginica", 1, 0)
summary(lm(PW~Sp1+Sp2))
```

} create the two dummy variables by our own

```
##
## Call:
## lm(formula = PW ~ Sp1 + Sp2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.626 -0.126 -0.026  0.154  0.474
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.24600    0.02894    8.50 1.96e-14 ***
## Sp1          1.08000    0.04093   26.39  < 2e-16 ***
## Sp2          1.78000    0.04093   43.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2047 on 147 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.9279
## F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

same as slide 37, 40

R commands for Dummy variables

- Creating dummary variables by hand
  - in Iris data set, variable species has 3 levels, we created two dummy variables Sp_1, Sp_2.
- Letting R do things automatically
  - factor(iris$Species): this command automatically created two dummy variables.
- The use of *factor*:
  - factor() is not needed in this Iris example, because "is.factor(iris$Species)" is true.
  - It is essential to use factor() if the coding of the categories is numerical!
  - To be safe, you can always use factor.

Chapter 3
Diagnostics and Remedial Measures

Estimation and Inference

- We have seen how to estimate model parameters for an OLS regression model.
- We have also performed inference on those estimates, as well as for predictions.
  - we derived lots properties about both
- All of these results, derivations and properties hinge on the assumptions of our regression model.
- We should check these assumptions.

Regression Diagnostics

- Diagnostic procedures are used for:
  - Assessing a model's appropriateness (fit)
  - Checking if model's assumptions are reasonable.
  - Finding observations that are problematic
  - Identifying ways to improve model
- We focus on graphical procedures, mostly using model residuals (e_i)
  - Difference between e_i (observed residual) and ϵ_i (true error term)

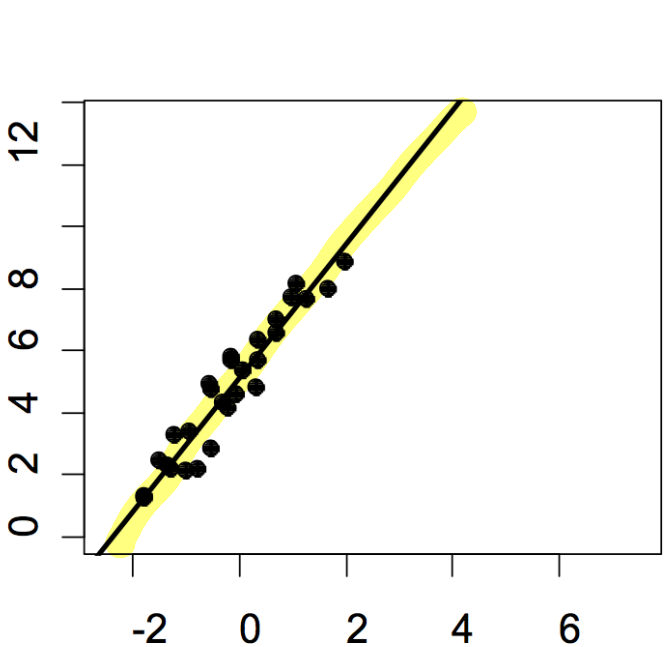
$$e_i = Y_i - \hat{Y}_i; \quad \epsilon_i = Y_i - E(Y_i)$$

Diagnostics for predictor variable X

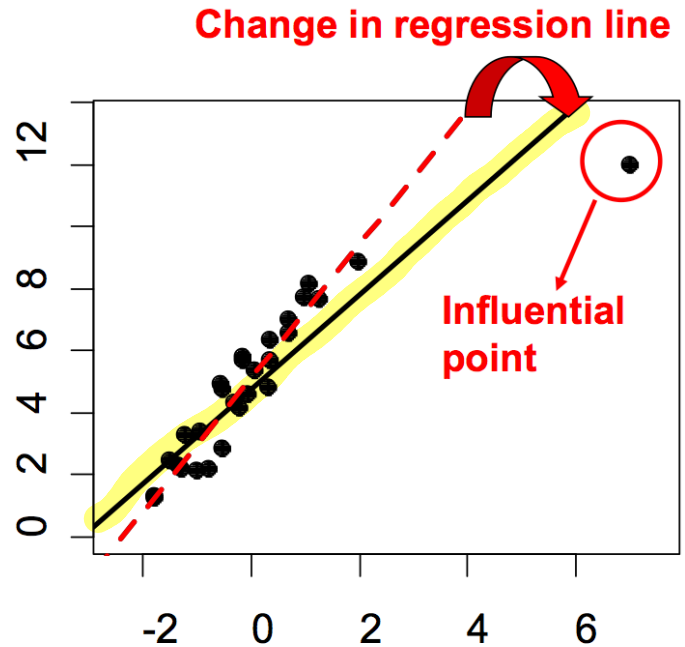
- Before fitting a linear regression model, we examine the predictor variable X.
  - Look for any outlying X values that are potential influence points
    - An observation is influencial if the estimates change substantially when the point is omitted.
    - R graphic functions: dotchart() for dot chart; stem() for stem-and-leaf plot; boxplot(), hist()
  - Loof for dependencies in X, when they are measured in sequence (e.g. time series)
    - sequence plot

Diagnostics for X

- Effect of influential point: with and without the influential point, the estimates (b_0, b_1) change a lot.



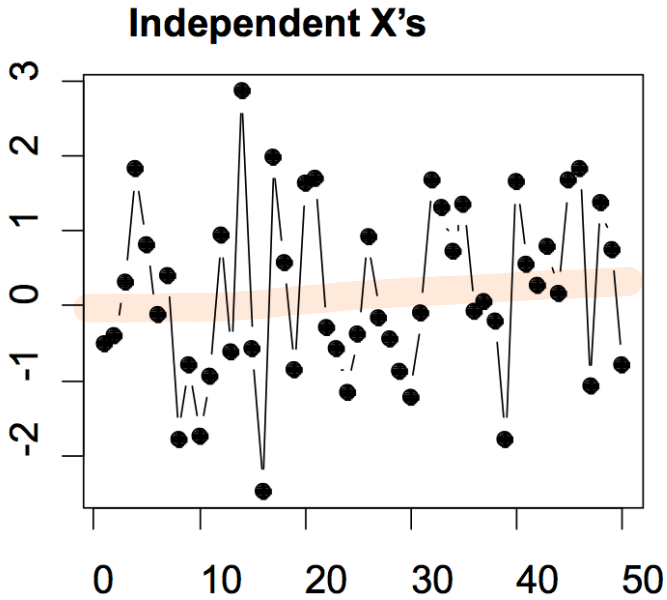
Fitting without influential pt

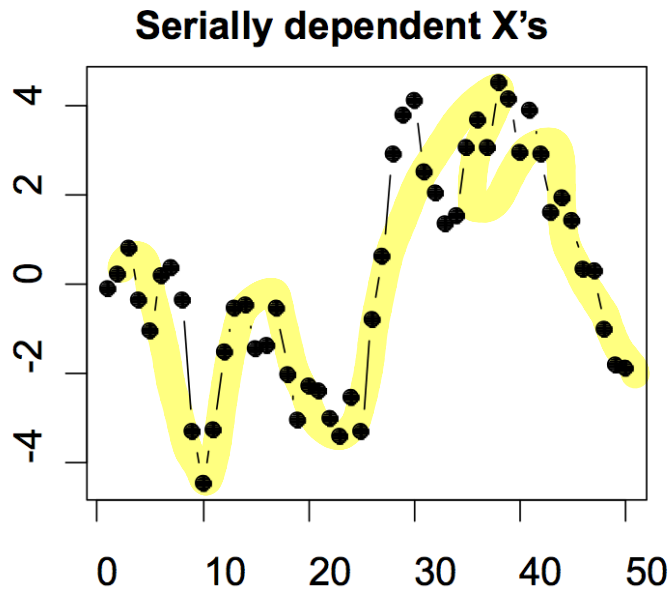


Fitting with influential pt

Diagnostics for X

- Sequence plots





data points bounce around $x = 0$.

clear pattern.

Assessing Leverage

- Leverage points have X values far from the mean
- We know that $\hat{Y}_i$ can be written as a linear combination of the Y_i's

$$\hat{Y}_i = b_0 + b_1 X_i = \bar{Y} + b_1(X_i - \bar{X}) = \sum_{j=1}^{n}[\frac{1}{n} + k_j(X_i - \bar{X})]Y_j$$

$$= \sum_{j=1}^{n}[\frac{1}{n} + \frac{(X_j - \bar{X})}{S_{XX}}(X_i - \bar{X})]Y_j$$

$$= \sum_{j=1}^{n}[\frac{1}{n} + \frac{(X_j - \bar{X})(X_i - \bar{X})}{S_{XX}}]Y_j = \sum_{j=1}^{n} h_{ij} Y_j$$

$\Rightarrow \hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \cdots + h_{in}Y_n$

$\hat{Y}_i$ is a weighted sum.

where $h_{ij} = \frac{1}{n} + \frac{(X_j - \bar{X})(X_i - \bar{X})}{S_{XX}}$

- Define the leverage of the i^{th} data point as

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \in (0, 1)$$

Assessing Leverage

$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$ implies

$$\begin{aligned}
\hat{Y}_1 &= h_{11}Y_1 + h_{12}Y_2 + \cdots + h_{1n}Y_n \\
\hat{Y}_2 &= h_{21}Y_1 + h_{22}Y_2 + \cdots + h_{2n}Y_n \\
\hat{Y}_3 &= h_{31}Y_1 + h_{32}Y_2 + \cdots + h_{3n}Y_n \\
&\vdots \\
\hat{Y}_n &= h_{n1}Y_1 + h_{n2}Y_2 + \cdots + h_{nn}Y_n
\end{aligned} \iff$$

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{22} & h_{22} & h_{23} & \cdots & h_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \cdots & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$\text{avg} = \frac{2}{n}$

$\sum_{j=1}^{n} h_{ij} = 1$

$\iff \hat{Y} = AY$

next slide we show:

① sum of each row of $A = 1$

② Average of the diagonal elements $= \frac{2}{n}$

Assessing Leverage

Show $\sum_{j=1}^{n} h_{ij} = 1$ and $\bar{h}_{ii} = 2/n$

Proof:

$$h_{ij} = \frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}}$$

$$\Rightarrow \text{①} \quad \sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n} \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}}\right)$$

$$= \frac{n}{n} + \frac{(x_i - \bar{x})}{S_{xx}} \sum_{j=1}^{n} (x_j - \bar{x}) = 1 + 0 = 1$$

$$\text{②} \quad \bar{h}_{ii} = \frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

$$= \frac{1}{n}\left(\frac{n}{n} + \frac{\sum (x_i - \bar{x})^2}{S_{xx}}\right)$$

$$= \frac{1}{n}(1 + 1)$$

$$= 2/n$$

Assessing Leverage

Show $h_{ij} = h_{ji}$ and $\sum_j h_{ij}^2 = h_{ii}$

Proof:

① Trivial: $h_{ij} = \frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} = h_{ji}$

② $\sum_{j=1}^n h_{ij}^2 = \sum_{j=1}^n \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}} \right)^2$

$= \sum_{j=1}^n \left(\frac{1}{n^2} + \frac{2}{n} \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} + \frac{(x_j - \bar{x})^2 (x_i - \bar{x})^2}{(S_{xx})^2} \right)$

$= \frac{n}{n^2} + \frac{2}{n} \frac{(x_i - \bar{x})}{S_{xx}} \underbrace{\sum_{j=1}^n (x_j - \bar{x})}_{=0} + \frac{(x_i - \bar{x})^2}{(S_{xx})^2} \underbrace{\sum_{j=1}^n (x_j - \bar{x})^2}_{= S_{xx}}$

$= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$

$= h_{ii}$

Q.E.D ^^ LoL .....

Assessing leverage

① $h_{i1} + \cdots + h_{in} = 1$

② if $h_{ii} \simeq 1 \Rightarrow Y_i$ contributes a lot to $\hat{Y}_i$

- If $h_{ii} \approx 1$, Y_i has a large effect on $\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \cdots + h_{in}Y_n$
  - The line will be close to Y_i no matter what the rest of the data look like (since $\hat{Y}_i \simeq h_{ii}Y_i \simeq Y_i$ for $h_{ii} \simeq$)
  - h_{ii} only a function of the X's
- Classify X_i (rather arbitrarily) as a high leverage point if

$$h_{ii} > 2\bar{h}_{ii} = \frac{4}{n}$$

- Some leverage points are better than others
  - They fall on (or close to) the regression line
  - The are not outliers
  - The are not influential
  - They don't change slope estimate by much
    - They can make R^2 higher

Practice problems and upcoming topics

- Practice problems after today's lecture: Chapter 2: 2.21, 2.22, 2.24, *2.25*, 2.35, 2.36, 2.51, 2.52, 2.54, 2.56, Ch3: 3.1,3.2.
- Upcoming topics
  - Diagnostic for residual
  - Influence Metrics: DEFITS, DEBETAS,COOK's distance
  - Variable transformation
- Reading for next lecture: CH3.3-3.8