# MATH 3339
# Statistics for the Sciences

Sec 9.1-9.2

Wendy Wang, Ph.D.
wwang60@central.uh.edu

Lecture 5 - 3339

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Outline



UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

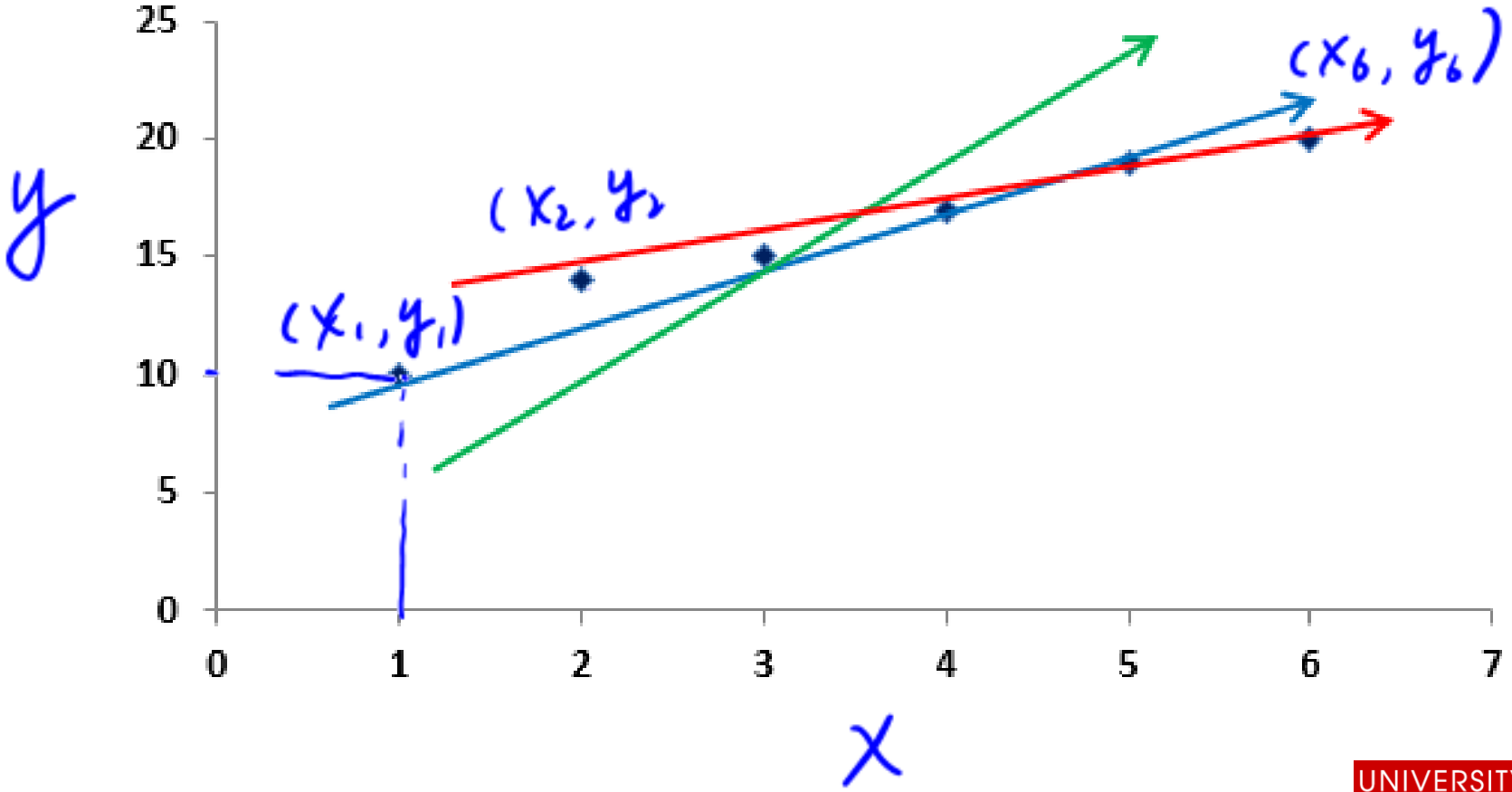
# Examining relationships

- Correlation measures the direction and strength of the straight-line relationship between two quantitative variables.
- If a scatterplot shows a linear relationship, we would like to summarize this overall pattern by drawing a line on the scatterplot.
- A **regression line** is a straight line that describes how a response variable $y$ changes as an explanatory variable $x$ changes.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Which line is the best?

There are many possible lines that we can fit to the data. Which line is the best to fit to the data in the scatter plot?

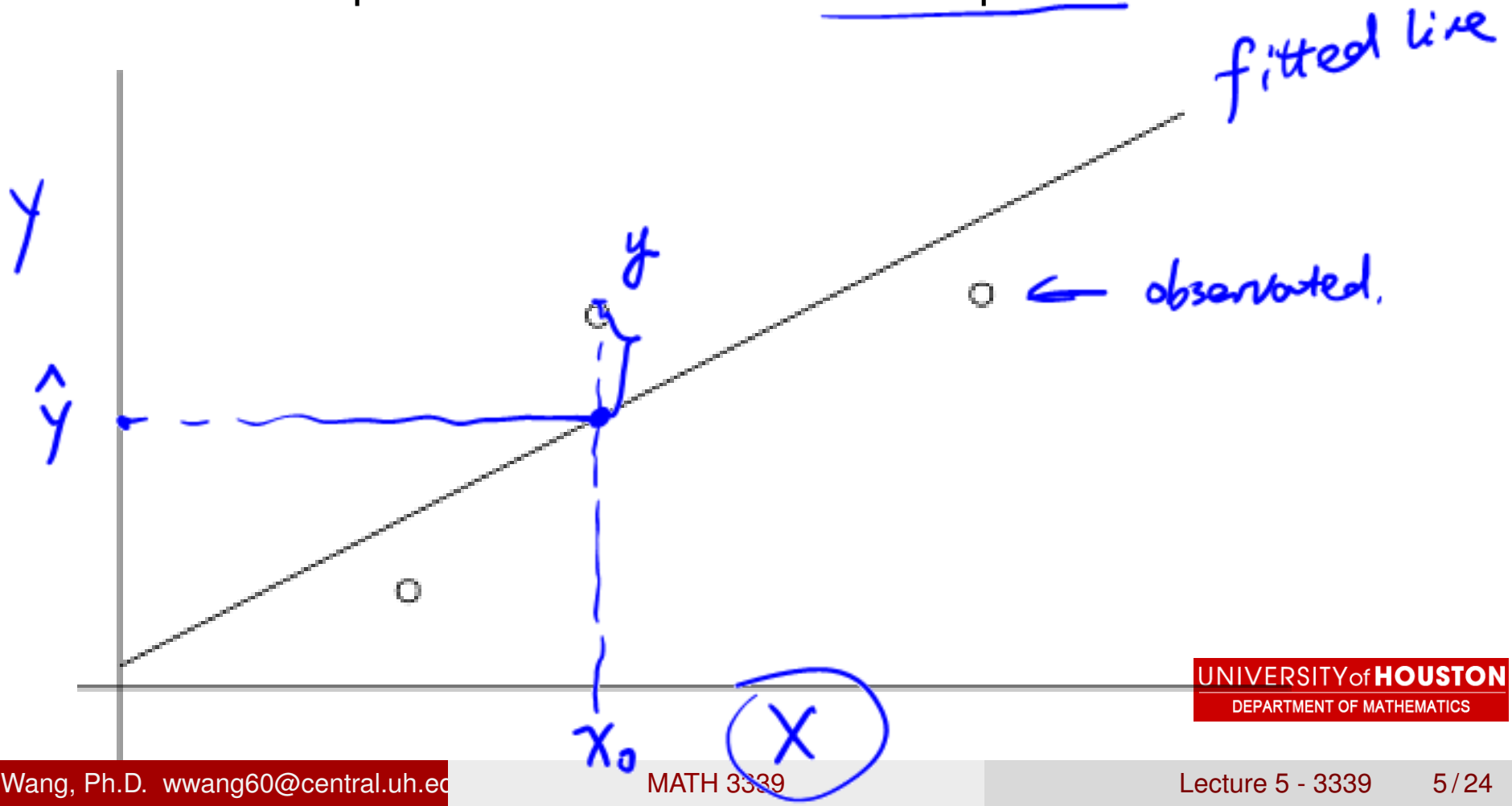



UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Which line is the best?

There are many possible lines that we can fit to the data. Which line is the best to fit to the data in the scatter plot?

We want a regression line that makes the vertical distances of the points in a scatter plot from the line as small as possible.



# Least-Squares regression

- The **least-squares regression line (LSRL)** of $Y$ on $X$ is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- The linear regression model is: $Y = \beta_0 + \beta_1 x + \varepsilon$
  - $Y$ is dependent variable (response).
  - $x$ is the independent variable (explanatory).
  - $\beta_0$ is the population intercept of the line.
  - $\beta_1$ is the population slope of the line.
  - $\varepsilon$ is the error term which is assumed to have mean value 0. This is a random variable that incorporates all variation in the dependent variable due to factors other than $x$.
  - The variability: $\sigma$ of the response $y$ about this line. More precisely, $\sigma$ is the standard deviation of the deviations of the errors, $\epsilon_i$ in the regression model.

"$\wedge$" : estimated value

- We will gather information from a sample so we will have the least squares estimates model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Calculating the Least Squares Estimates

- The least squares estimate of the slope coefficient $\beta_1$ of the true regression line is

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = cor(x, y) \cdot \frac{s_y}{s_x}$$

- The least squares estimate of the intercept $\beta_0$ of the true regression line is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Finding Equation for Coffee Sales

Given these values determine the least-square regression line (LSRL) equation for predicting number sold based on shelf space. (Dataset can be found at: `https://www.math.uh.edu/~cathy/Math3339/data/shelf.txt`)

- Explanatory variable: Shelf space ($X$) $\bar{x} = 12.5$ feet, $s_x = 5.83874$ feet.
- Response variable: Sales ($Y$) $\bar{y} = 237.5$ units sold, $s_y = 52.2451$ units sold.
- Correlation: $r = 0.827$

$$\hat{\beta}_1 = cor(X, Y) * \frac{s_y}{s_x} = 0.827 * \frac{52.2451}{5.83874} = 7.4$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 237.5 - 7.4 * 12.5 = 145$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# R code

lm( ) : linear model

```
> shelf.lm=lm(sold~space,data=shelf)
> summary(shelf.lm)
Call:
lm(formula = sold ~ space)

Residuals:
   Min     1Q Median     3Q    Max
-42.00 -26.75   5.50  21.75  41.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  145.000     21.783   6.657 5.66e-05 ***
space          7.400      1.591   4.652 0.000906 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.81 on 10 degrees of freedom
Multiple R-squared:  0.6839,Adjusted R-squared:  0.6523
F-statistic: 21.64 on 1 and 10 DF,  p-value: 0.000906
```

y~x

LSRL: $\hat{y} = 145 + 7.4x$

$\hat{sold} = 145 + 7.4 * space$

$\hat{\beta}_0$

$\hat{\beta}_1$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Example of Regression

Because elderly people may have difficulty standing to have their heights measured, a study looked at predicting overall height from height to the knee. Here are data (in centimeters, cm) for five elderly men:

| | | | | | |
|---|---|---|---|---|---|
| Knee Height (cm) | 57.7 | 47.4 | 43.5 | 44.8 | 55.2 |
| Overall Height(cm) | 192.1 | 153.3 | 146.4 | 162.7 | 169.1 |

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Results

The following is an output from R for the "lm" (linear models) function. Give the LSRL equation of the relationship between the knee height and overall height.

```
> height.lm=lm(overall~knee)
> summary(height.lm)
Call:
lm(formula = overall ~ knee)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.1315    38.3986   1.149   0.3338
knee          2.4254     0.7673   3.161   0.0508 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.766 on 3 degrees of freedom
Multiple R-squared:  0.7691,Adjusted R-squared:  0.6921
F-statistic: 9.992 on 1 and 3 DF,  p-value: 0.05083
```

What is the regression equation? $= 2.4254 * knee + 44.1315$

$\widehat{overall} = 44.1315 + 2.4254 * knee$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Is this good at predicting the response?

$r^2$ (**coefficient of determination**) is the percent (fraction) of variability in the response variable ($Y$) that is explained by the least-squares regression with the explanatory variable.

- This is a measure of how successful the regression equation was in predicting the response variable.
- The closer $r^2$ is to one (100%) the better our equation is at predicting the response variable.
- We will look later at how this is calculated.
- In the R output it is the **Multiple R-squared** value.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Sum of Squares

The Total Sum of Squares tells you how much variation there is in the dependent variable.

$$SS(tot) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The sum of squares residual (SS(resid)), is found by:

$$SS(resid) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

And,

$$r^2 = \frac{SS(Model)}{SS(tot)} = \frac{SS(tot) - SS(resid)}{SS(tot)} = 1 - \frac{SS(resid)}{SS(tot)}$$

Because of this, $r^2$ is interpreted as the proportion of observed $y$ variation that can be explained by the simple linear regression model.

# Is this good at predicting the response?

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line.

$$\text{residual} = \text{observed } y - \text{predicted } y$$

$$= y \text{ in data} - \hat{y}$$

- We can determine residuals for each observation.
- The closer the residuals are to zero, the better we are at predicting the response variable.
- We can plot the residuals for each observation, these are called the residual plots.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Example of residuals

① relation bet. x and y
② prediction

The regression equation to determine number of units sold ($y$) for coffee by shelf space ($x$) is: $\hat{y} = 145 + 7.4x$.
A store that has 10 feet of space sold 260 units of coffee. Determine the residual for this store.

1. Determine the predicted units sold for $x = 10$.

$\hat{y} = 145 + 7.4 * 10 = 219$

2. The observed units sold is the given value 260.

$y = 260$ for $x = 10$

3. The residual is the difference between the observed y and the predicted y.

residual = obs. $y$ − predicted $y$
$= 260 - 219 = 41$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Residuals of coffee sales

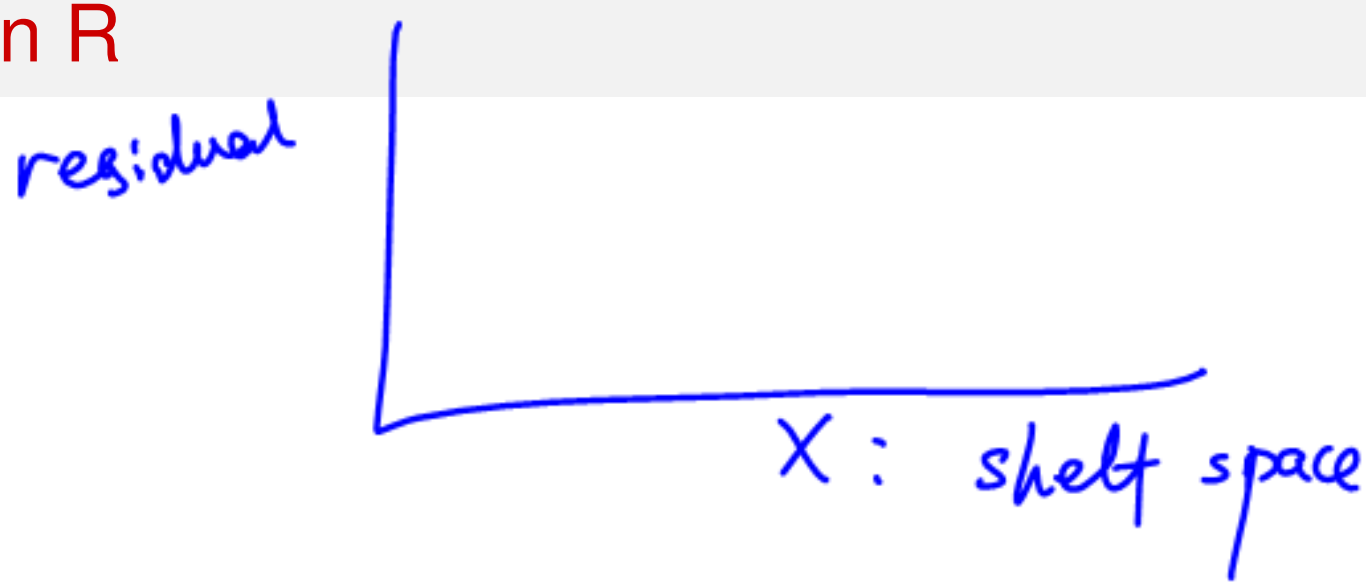
The regression equation is

Weekly Sales = 145 + 7.40 Shelf Space

| Shelf Space | Observed Weekly Sales | Predicted Weekly Sales | Residual |
|---|---|---|---|
| 5 | 160 | 182 | -22 |
| 5 | 220 | 182 | 38 |
| 5 | 140 | 182 | -42 |
| 10 | 190 | 219 | -29 |
| 10 | 240 | 219 | 21 |
| 10 | 260 | 219 | 41 |
| 15 | 230 | 256 | -26 |
| 15 | 270 | 256 | 14 |
| 15 | 280 | 256 | 24 |
| 20 | 260 | 293 | -33 |
| 20 | 290 | 293 | -3 |
| 20 | 310 | 293 | 17 |

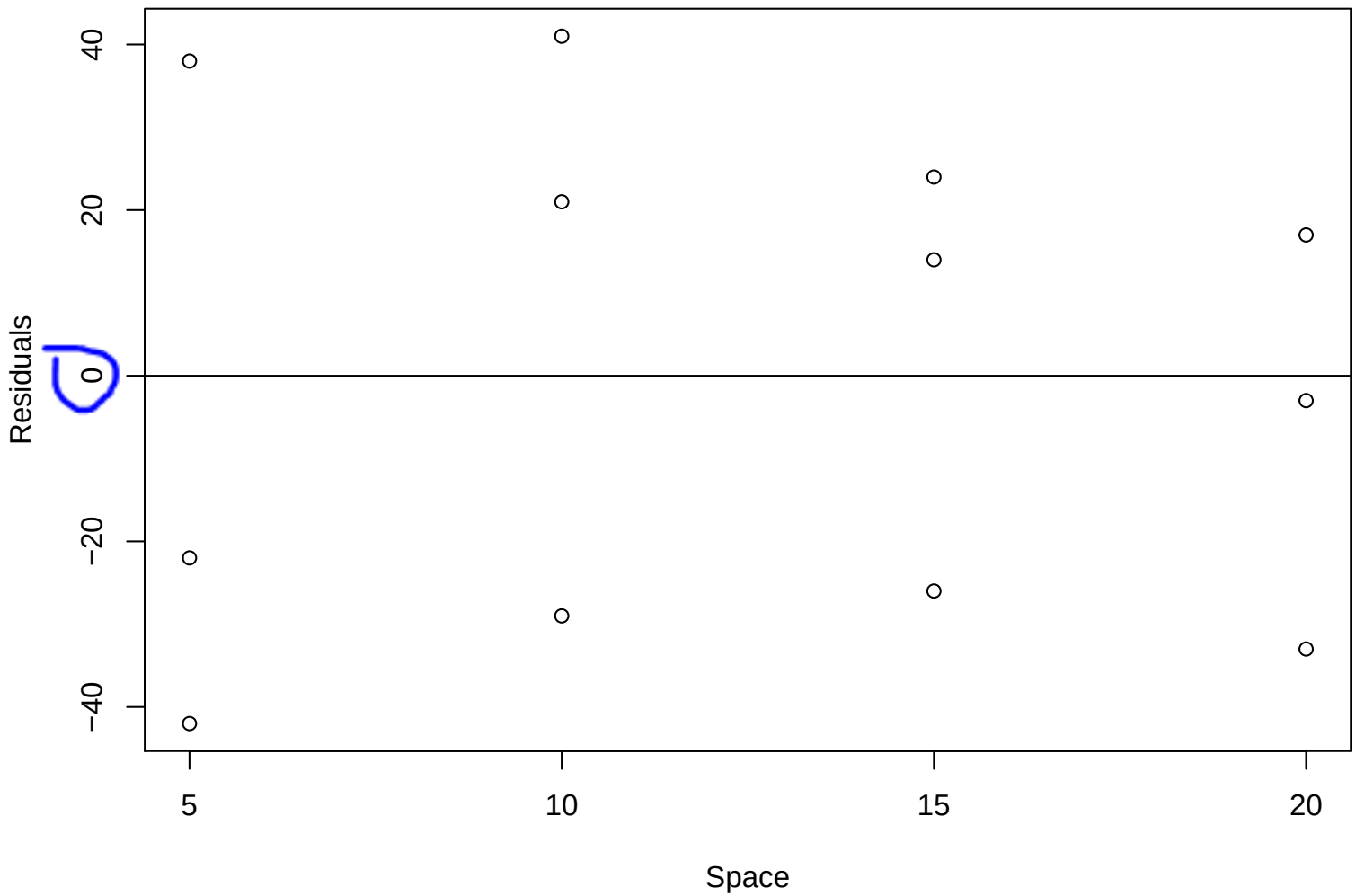
UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Residual plots in R




- Input the data.
- Create the linear models: name.lm=lm(y∼x)
- Create the plot: plot(x, resid(name.lm))

# Residual Plot



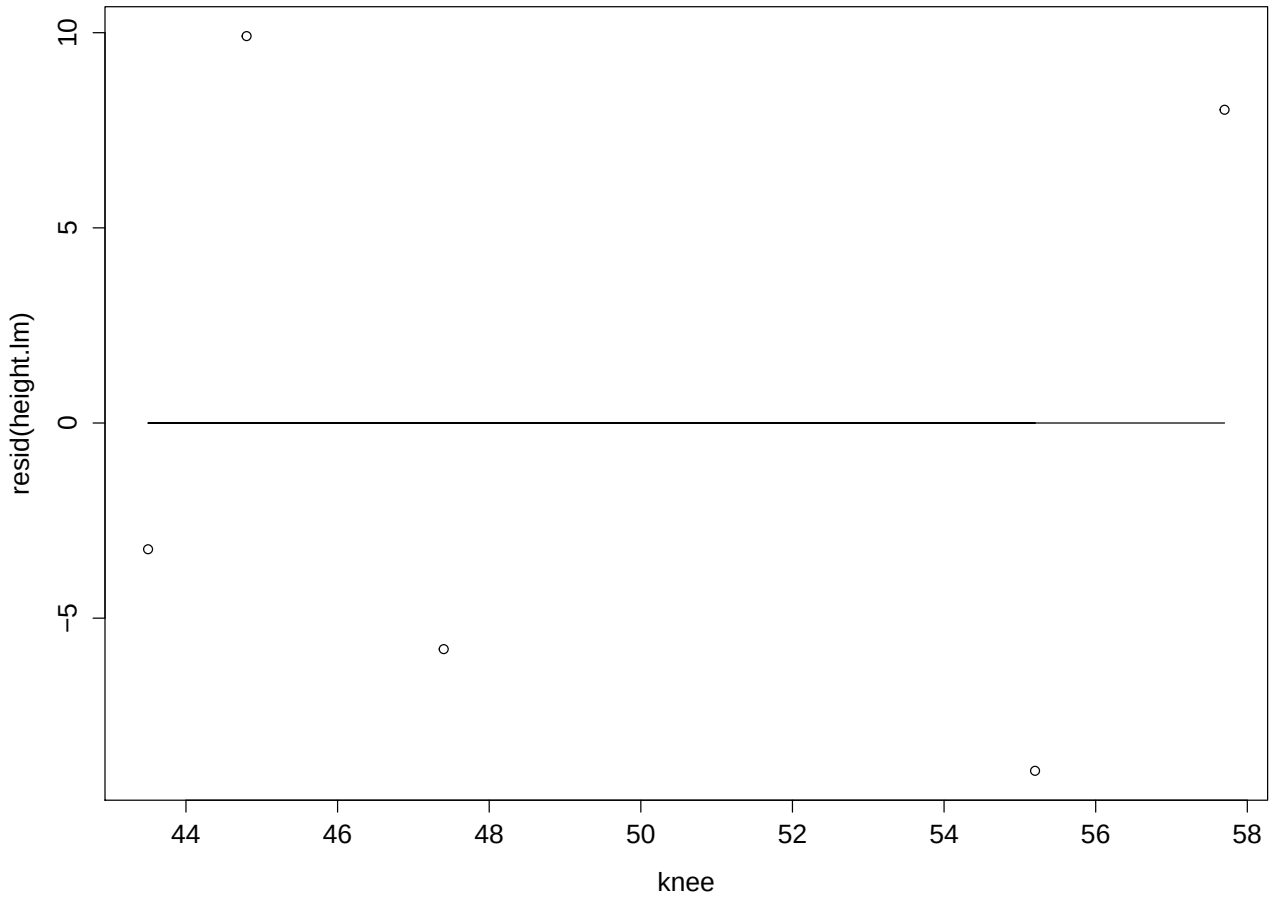
# Examining a residual plot

- A **curved pattern** shows that the relationship is not linear.
- **Increasing spread** about the zero line as $x$ increases indicates the prediction of $y$ will be less accurate for larger $x$. **Decreasing spread** about the zero line as $x$ increases indicates the prediction of $y$ to be more accurate for larger $x$.
- Individual points with larger residuals are considered outliers in the vertical ($y$) direction.
- Individual points that are extreme in the $x$ direction are considered outliers for the $x$-variable.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

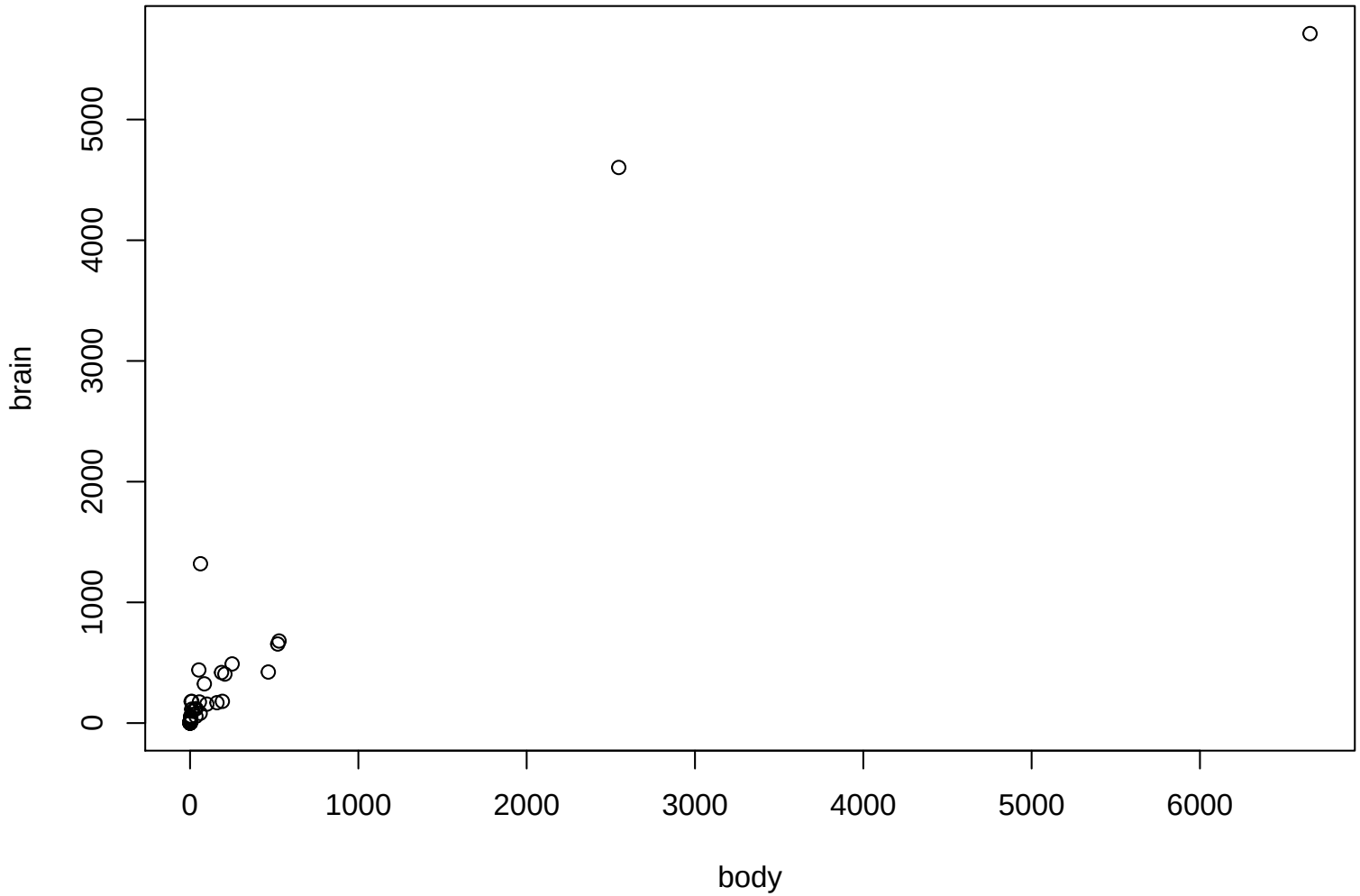
# Example of Residual Plot

The following is the residual plot of the model $x$ = knee height and $y$ = overall height. Would this linear model be best to use for this data? Explain.



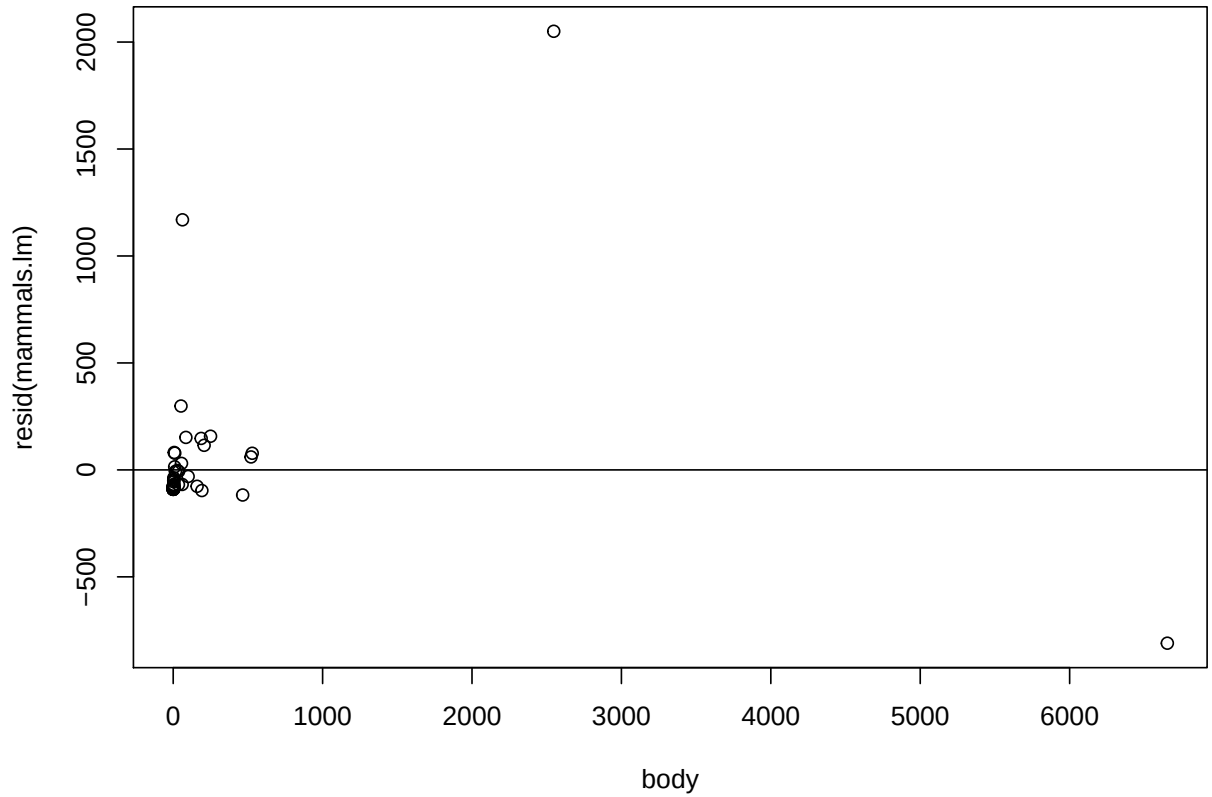

VERSITY of HOUSTON
EPARTMENT OF MATHEMATICS

# What if the scatterplot does not appear linear?



# Residual Plot

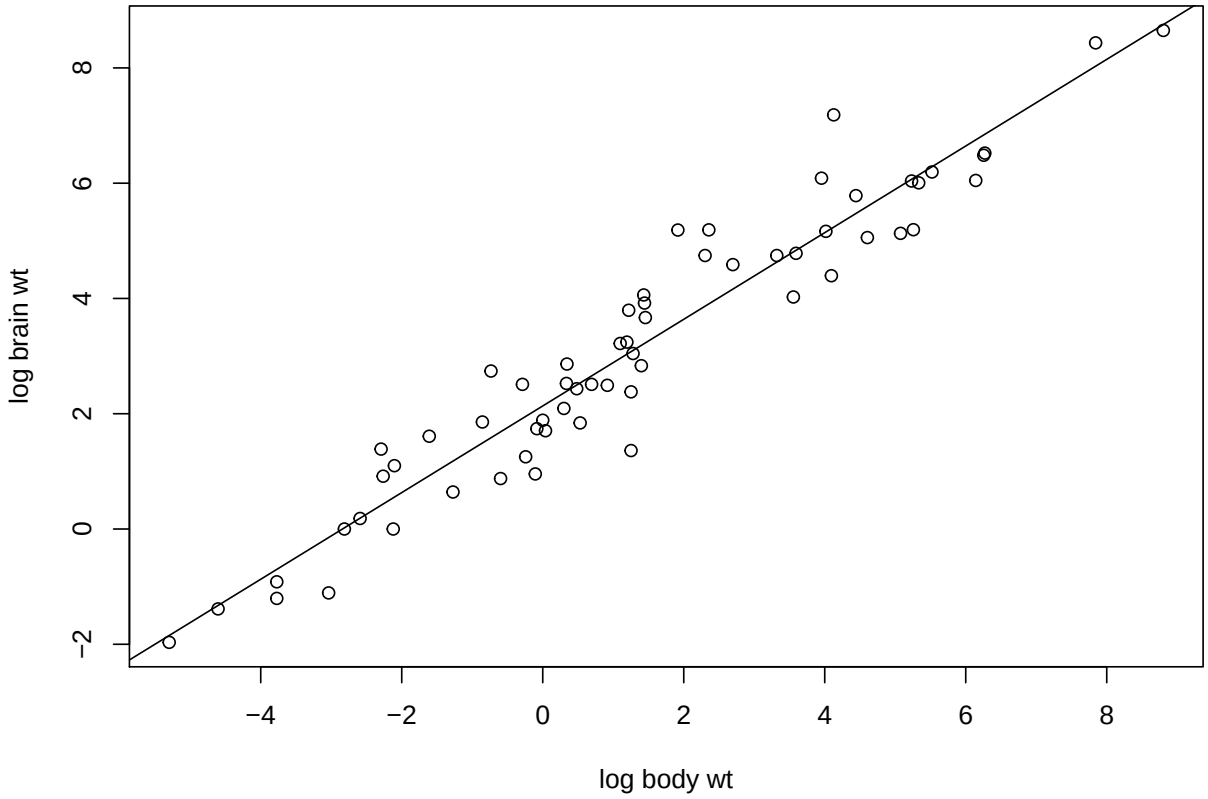
```
> mammals.lm=lm(brain~body)
> plot(body,resid(mammals.lm))
> abline(0,0)
```




UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

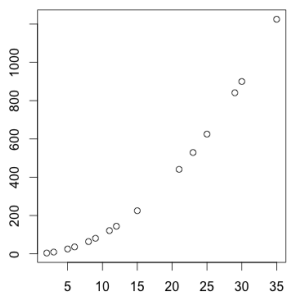
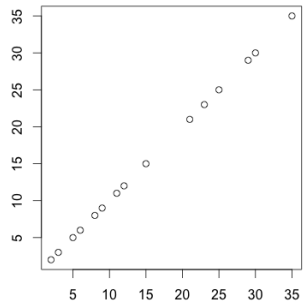
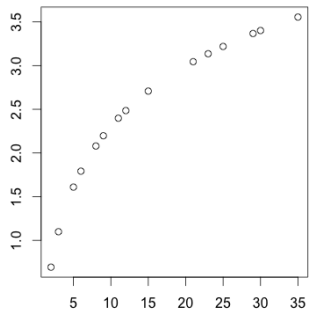
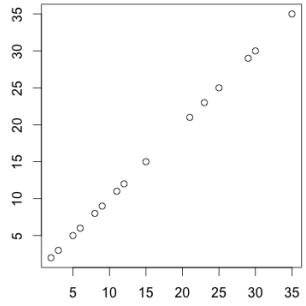
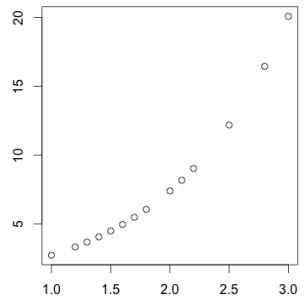
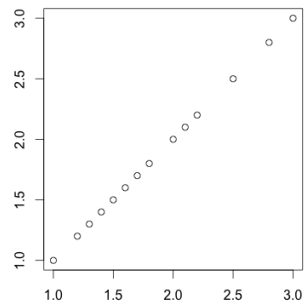
# Transform The Variables

```
plot(log(body),log(brain),xlab="log body wt",ylab="log brain wt")
abline(lm(log(brain)~log(body)))
```




UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

# Other Possible Transformations

| Original Scatter-plot | Type of plot and Transformation | Modified Scatter-plot |
|---|---|---|
|  | ← looks like $y = x^2$<br>transform by changing<br>$(x, y)$ into $(x, \sqrt{y})$<br>→ |  |
|  | ← looks like<br>$y = \log x$<br>transform by changing<br>$(x, y)$ into $(x, e^y)$<br>→ |  |
|  | ← looks like $y = e^x$<br>transform by changing<br>$(x, y)$ into $(x, \log y)$<br>→ |  |