Chapter 8: Correlation & Regression

- ANOVA and the two-sample t-test:
  ———apply to the situation where there is a quantitative response variable, and a predictor variable that indicates group membership, which we might think of a as categorical predictor variable.
- Categorical data analysis:
  ——- all variables are categorical, and we keep track of the counts of observation in each category or combination of categories.
- Correlation and Regression in this Chapter:
  ———we analyze cases where we have multiple quantitative variables.

Chapter 8: Correlation & Regression

In the simplest case, there are two quantitative variables. Examples include the following:

- heights of fathers and sons (this is a famous example from Galton, Darwin's cousin)
  —- Galton studied the relation between heights of parents and children and mentioned that "It appeared from these experiments that the offspring did not tend to resemble their parents in size, but always to be more mediocre than theirs - to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were small." He first mentioned "Regression to the mean"

- ages of husbands and wifes
- systolic versus diastolic pressure for a set of patients
- high school GPA and college GPA
- college GPA and GRE scores
- MCAT scores before and after a training course

In the past, we might have analyzed pre versus post data using a two-sample t-test to see whether there was a difference.

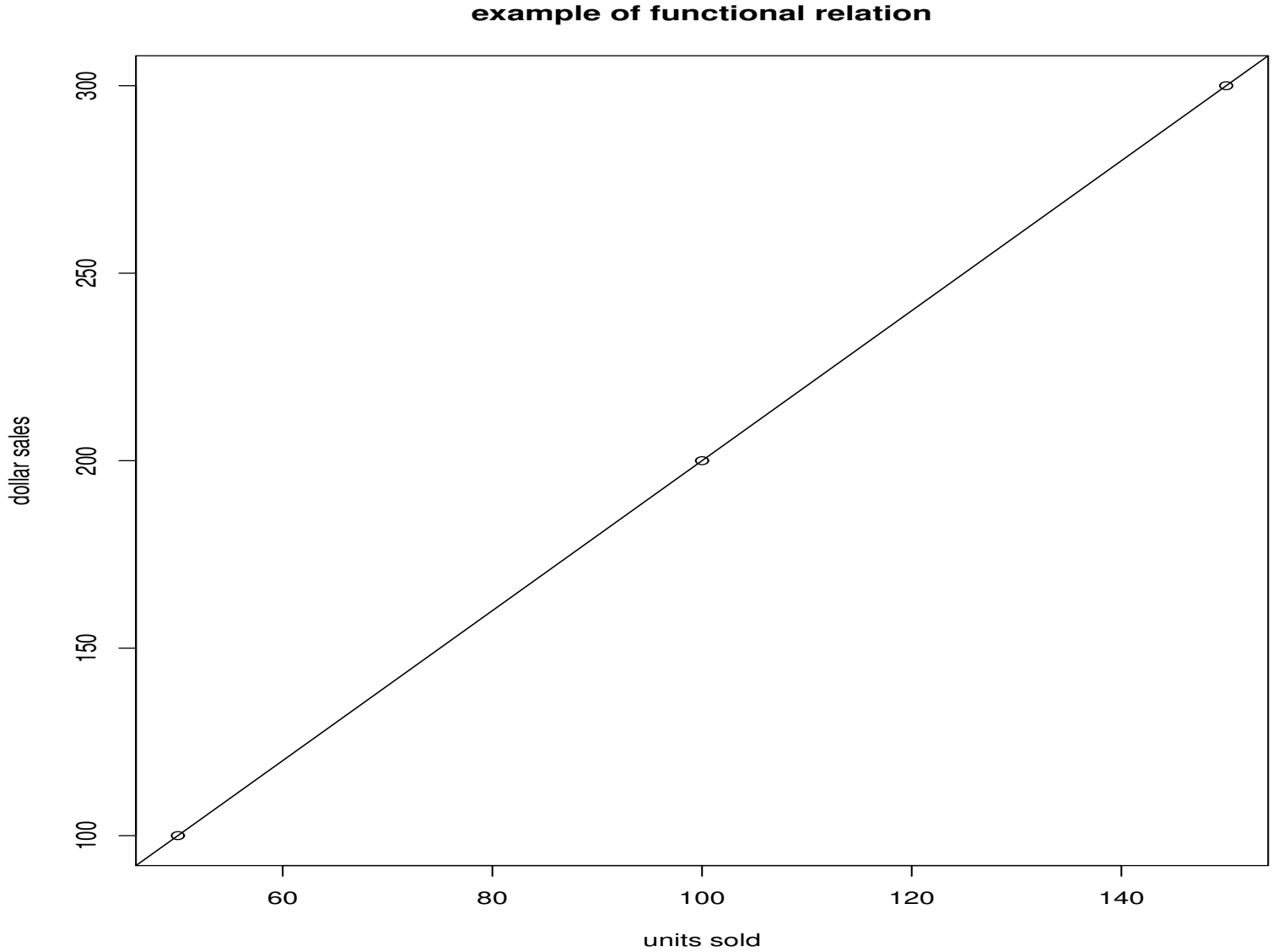
It is also possible to try to **quantify** the relationship—instead of just asking whether the two sets of scores are different, or getting an interval for the average difference, we can also try to **predict the new score based on the old score**, and the amount of improvement might depend on the old score.

Relations between variables

Functional relation between two variables: expressed by a mathematical formula

Example: consider a product's sale

- y: Dollar sales
- x: Number of units sold
- Selling price: \$2 per unit
- The relation between dollar sales and number of units sold is expressed by the equation $y = 2x$

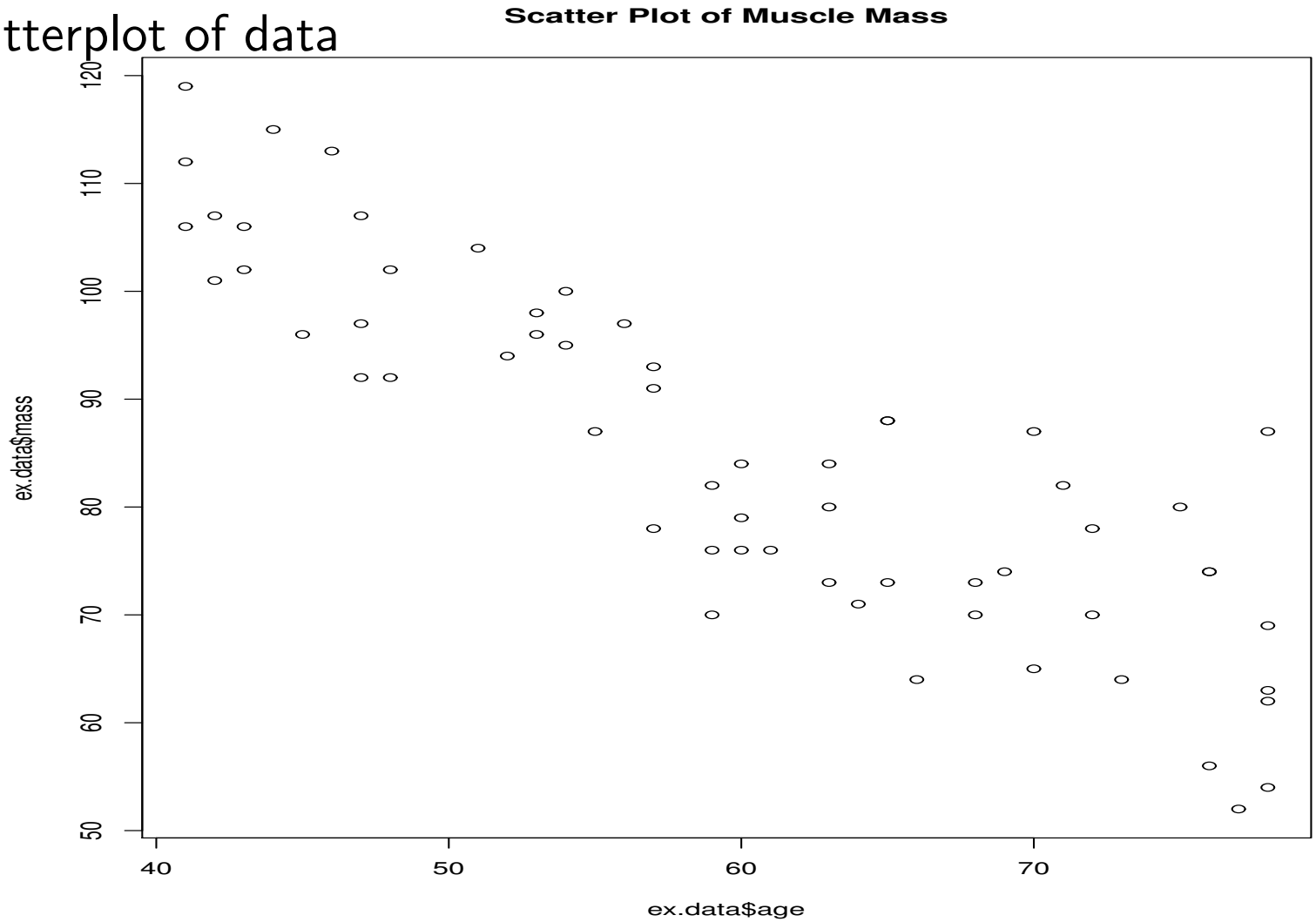

example of functional relation
dollar sales
100
150
200
250
300
60
80
100
120
140
units sold

Statistical relation between two variables

- Not a perfect relation
- In general, the observations for a statistical relation do not fall directly on the curve of relationship
- Statistical relation could be very useful, even though they do not have the exactitude of a functional relation

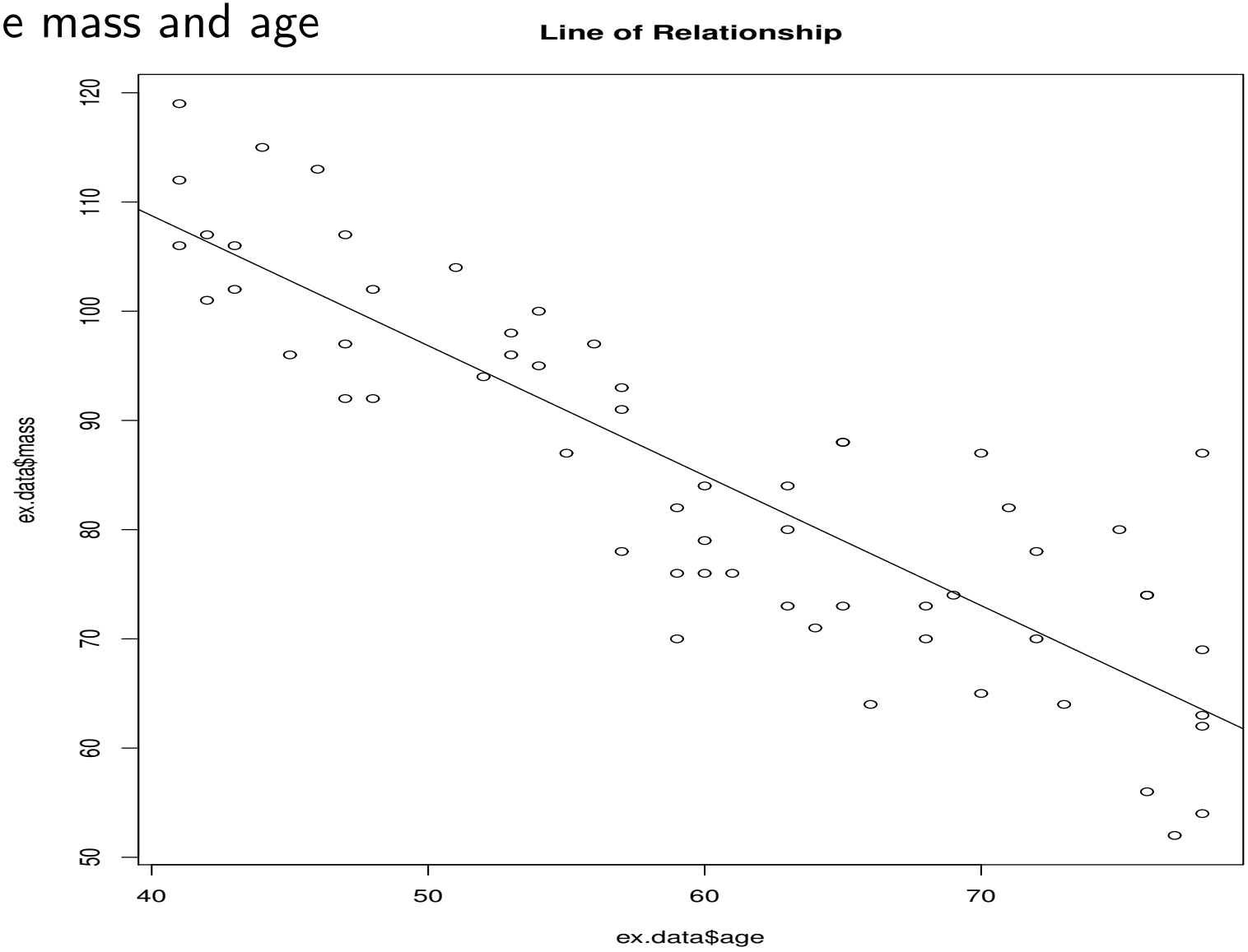
Example: A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79 with a total number of 60 women.

- Scatterplot of data



- In statistical terminology, each point in the scatter plot represents a trial or a case
- Plot suggests a negative relation between age and muscle mass in women
- Clearly, the relation is not a perfect one
- Variation in muscle mass is not accounted for by age

Plot a line of relationship that describes the statistical relation between muscle mass and age



- The line indicates the general tendency by which muscle mass vary with age
- Most of the points do not fall directly on the line of statistical relationship
- The scattering of points around the line represents variation in muscle mass that is not associated with age and that is usually considered to be of a random nature

Correlation

For n observations on two variables, the sample correlation is calculated by

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Here S_X and S_Y are the sample standard deviations

$$S_X = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)}, S_Y = \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)}$$

and

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

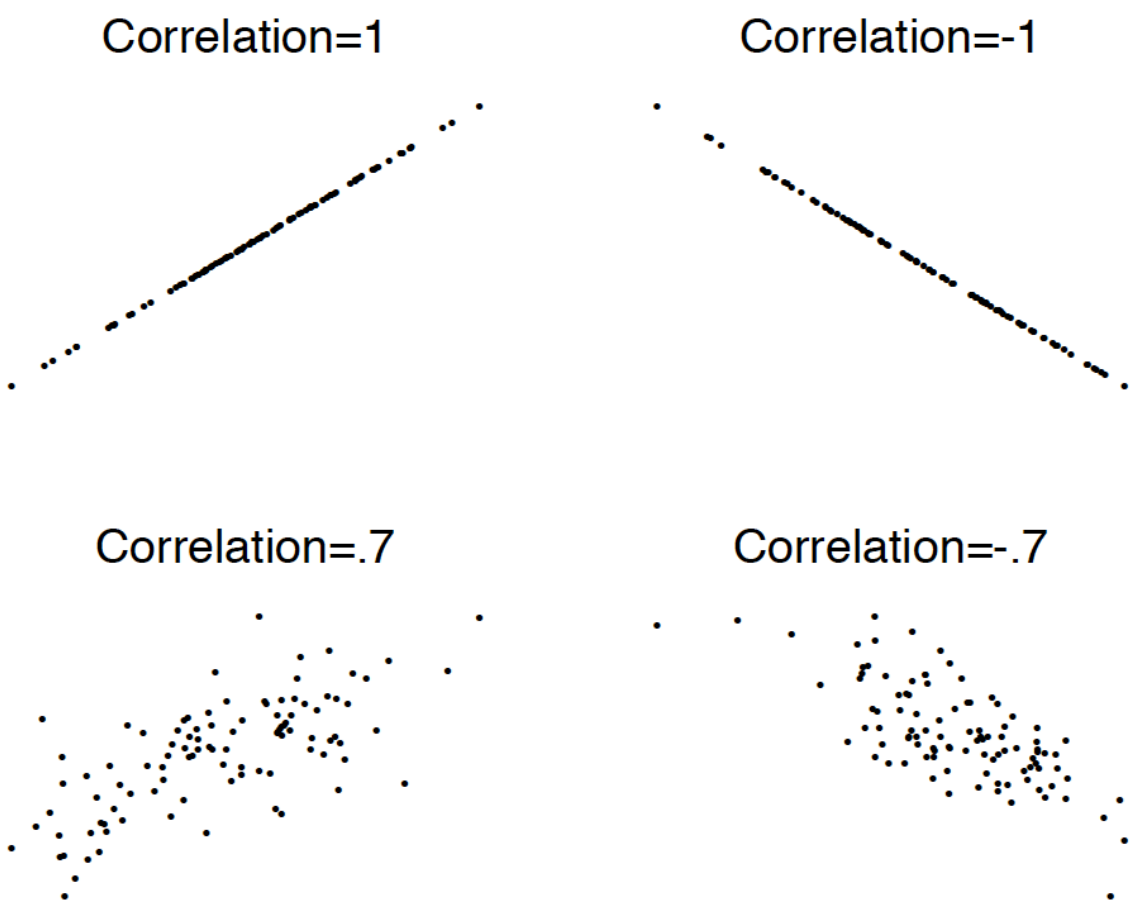
is the sample covariance. All the $(n-1)$ terms cancel out from the numerator and denominator when calculating r.

Correlation

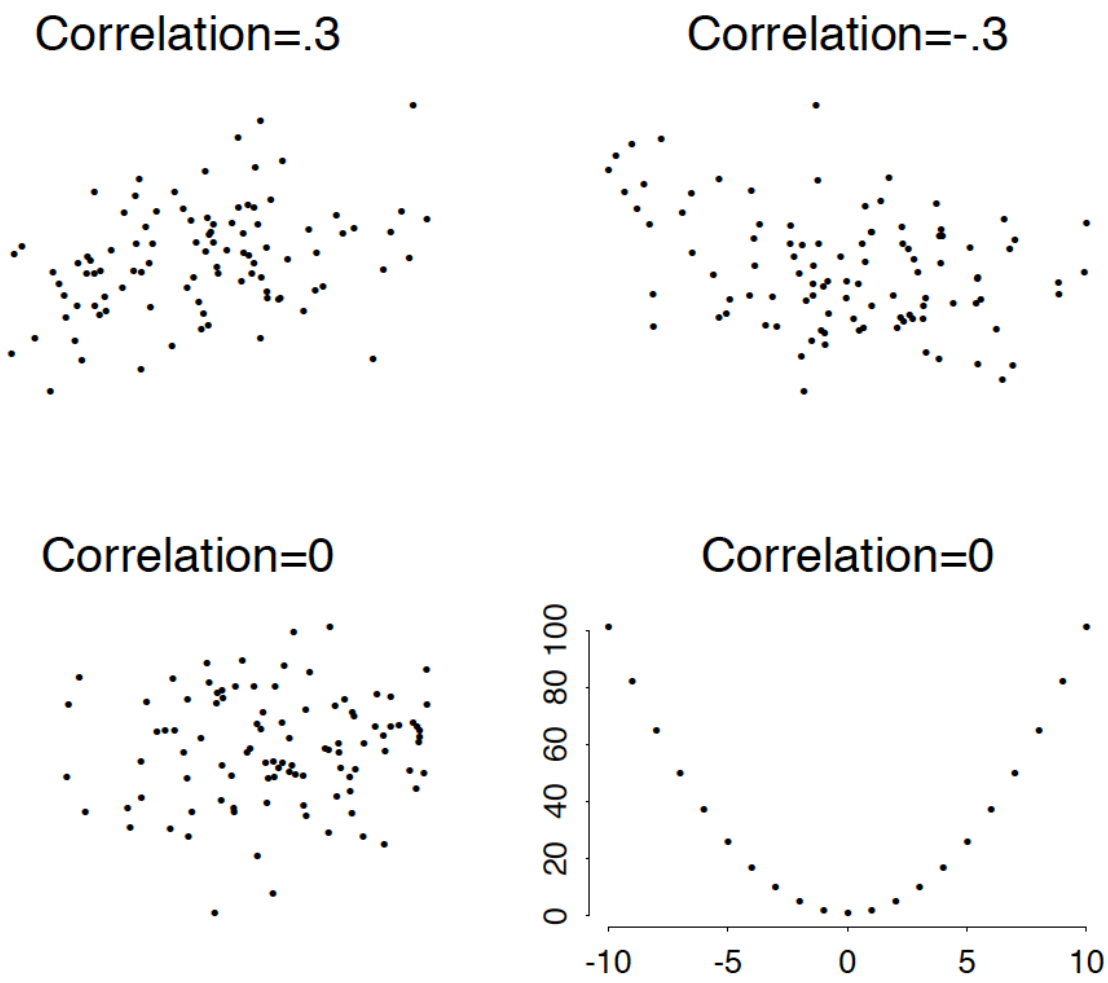
The correlation measures the linear relationship between variables X and Y. The sample correlation r between $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ has the following properties

- $-1 \leq r \leq 1$
- if Y_i tends to increase linearly with X_i, then $r > 0$
- if Y_i tends to decrease linearly with X_i, then $r < 0$
- if there is a perfect linear relationship between X and Y, then $r = 1$ (points fall on a line with positive slope)
- if there is a perfect negative relationship between X and Y, then $r = -1$ (points fall on a line with negative slope)
- the closer the points (X_i, Y_i) are to a straight line, the closer r is to 1 or -1
- r is not affected by linear transformations (i.e., converting from inches to centimeters, Fahrenheit to Celsius, etc.
- the correlation is symmetric: the correlation between X and Y is the same as the correlation between Y and X

Correlation



Correlation

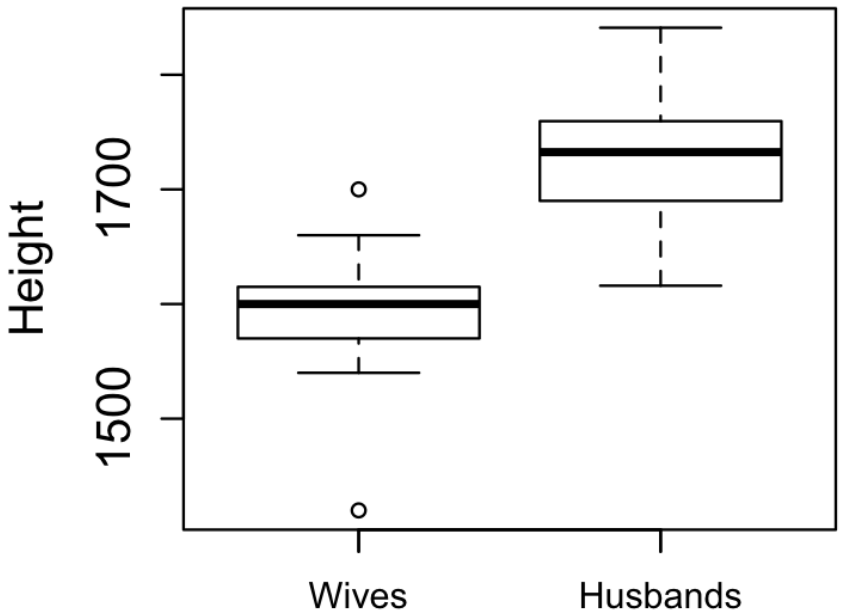


Correlations: husband and wife example

Here is some example data for husbands and wives. Heights are in cm.

Couple	HusbandAge	HusbandHeight	WifeAge	WifeHeight
1	49	180.9	43	159.0
2	25	184.1	28	156.0
3	40	165.9	30	162.0
4	52	177.9	57	154.0
5	58	161.6	52	142.0
6	32	169.5	27	166.0
7	43	173.0	52	161.0
8	47	174.0	43	158.0
9	31	168.5	23	161.0
10	26	173.5	25	159.0
11	40	171.3	39	161.0
12	35	173.6	32	170.0

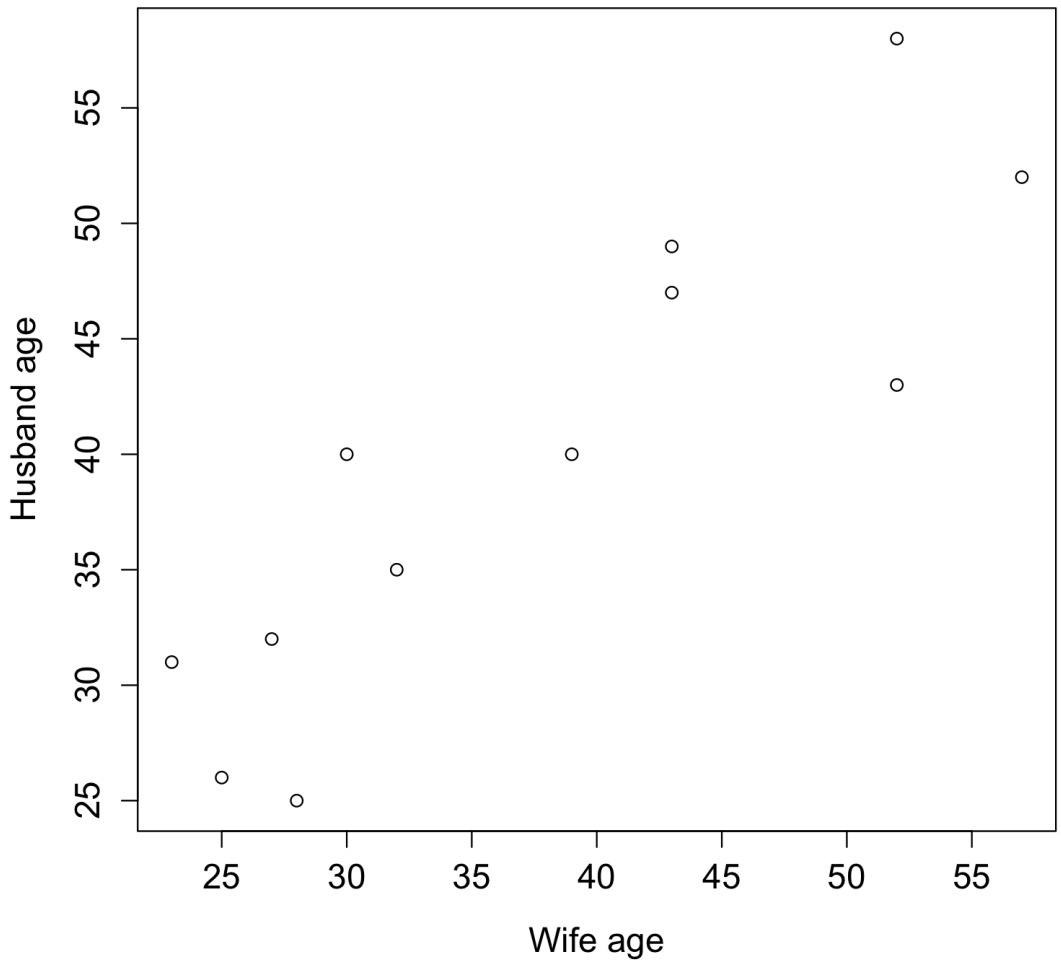
Correlation



Correlation: Husband and wife ages

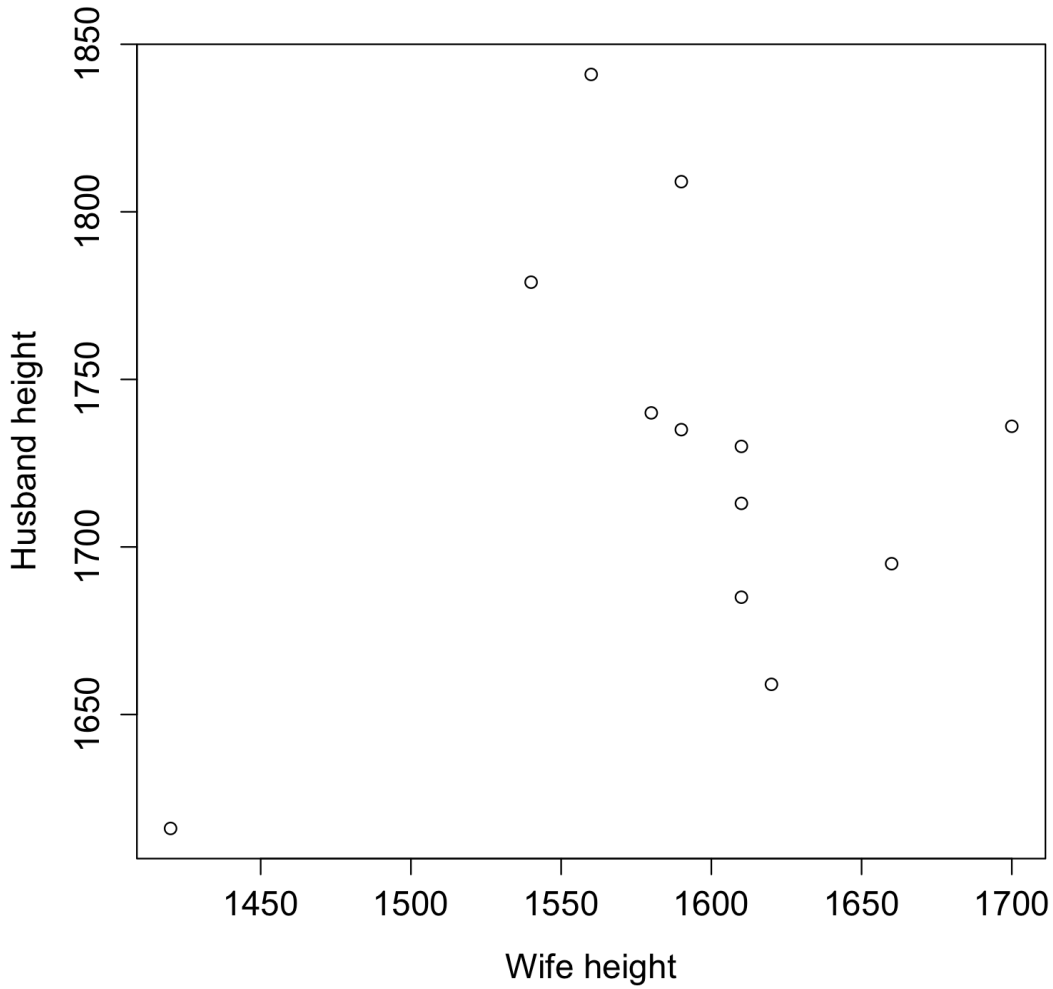
Correlation between husband and wife's age is 0.88.

```
cor(x$WifeAge,x$HusbandAge)
```



Correlation: Husband and wife heights

Correlation between husband and wife's height is 0.18 with outlier, and is -0.54 without outlier.



husband and wife example

The correlation is looking at something different than the t test.

- A t-test for this data might look at whether the husbands and wives had the same average age.
- The correlation looks at whether younger wives tend to have younger husbands and older husbands tend to have older wives. Similarly for height.
- Even if husbands tend to be taller than wives (higher mean height for husbands), that doesn't necessarily mean that there is a relationship between the heights for couples.

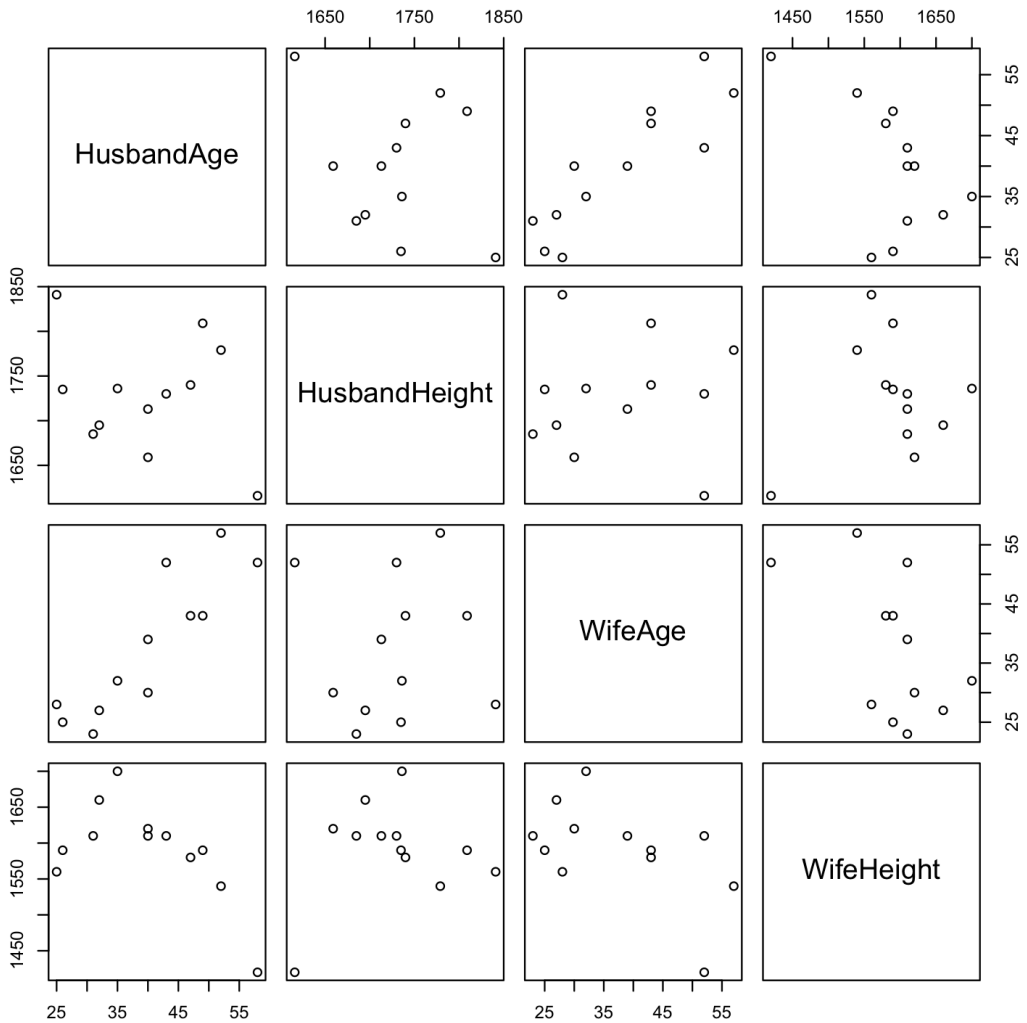
Pairwise Correlations

The pairwise correlations for an entire dataset can be done as follows.

```
options(digits=4) # done so that the output fits
#on the screen!
cor(x[,2:5])
              HusbandAge HusbandHeight  WifeAge WifeHeight
HusbandAge         1.0000      -0.24716  0.88003    -0.5741
HusbandHeight     -0.2472       1.00000  0.02124     0.1783
WifeAge            0.8800       0.02124  1.00000    -0.5370
WifeHeight        -0.5741       0.17834 -0.53699     1.0000
```

Correlation: scatterplot matrix

`pairs(x[,2:5])` allows you to look at all data simultaneously.



Correlation

CIs and hypothesis tests can be done for correlations using `cor.test()`. The test is usually based on testing whether the population correlation ρ is equal to 0, so

$$H_0 : \rho = 0$$

and you can have either a two-sided or one-sided alternative hypothesis. We think of r as a sample estimate of ρ. The test is based on a t-statistic,

$$t_{obs} = r\sqrt{\frac{n-2}{1-r^2}}$$

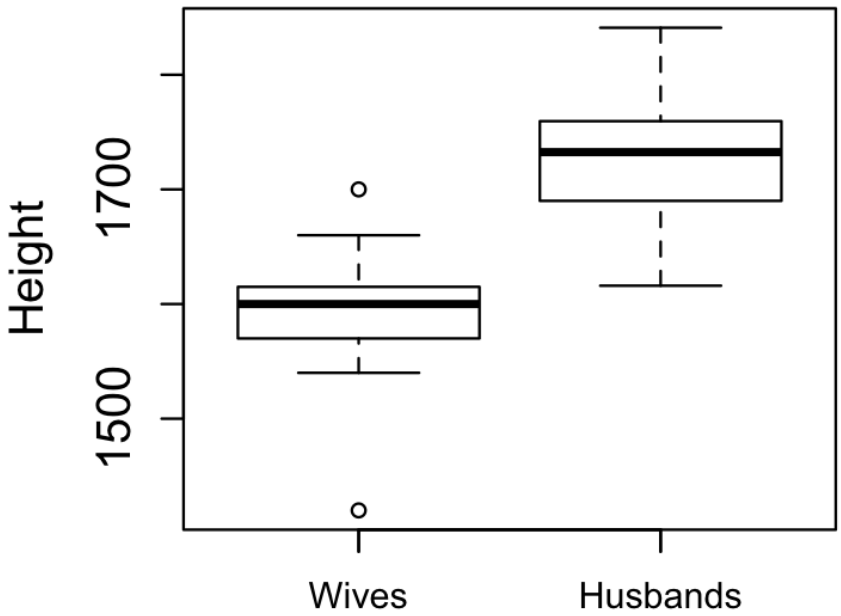
and this is compared to a t distribution with $n-2$ degrees of freedom. As usual, you can rely on R to do the test and get the CI.

Correlation

Assumptions:

- The t distribution derivation of the p-value and CI assume that the joint distribution of X and Y follow a bivariate normal distribution. ——-A sufficient condition for this is that X and Y each individually have normal distributions, but this is not a necessary condition
- We can do usual tests or diagnostics for normality.
- Similar to the t-test, the correlation is sensitive to outliers.

Correlation



Correlation

Shairpo-Wilk's tests for normality would all be not rejected, although the sample sizes are quite small for detecting deviations from normality:

```
> shapiro.test(x$HusbandAge)$p.value
[1] 0.8934
> shapiro.test(x$WifeAge)$p.value
[1] 0.2461
> shapiro.test(x$WifeHeight)$p.value
[1] 0.1304
> shapiro.test(x$HusbandHeight)$p.value
[1] 0.986
```

Correlation

Here we testing whether ages are significantly correlated.

```
> cor.test(x$WifeAge,x$HusbandAge)

Pearson's product-moment correlation

data:  x$WifeAge and x$HusbandAge
t = 5.9, df = 10, p-value = 2e-04
alternative hypothesis: true correlation is not
equal to 0
95 percent confidence interval:
 0.6185 0.9660
```

Correlation

Here we test whether heights are significantly correlated.

```
> cor.test(x$WifeHeight,x$HusbandHeight)

Pearson's product-moment correlation

data:  x$WifeHeight and x$HusbandHeight
t = 0.57, df = 10, p-value = 0.6
alternative hypothesis: true correlation is not
equal to 0
95 percent confidence interval:
 -0.4407  0.6824
sample estimates:
   cor
0.1783
```

Correlation

We might also test the heights with the bivariate outlier removed:

```
> cor.test(x$WifeHeight[x$WifeHeight>1450],
x$HusbandHeight[x$WifeHeight>1450])

Pearson's product-moment correlation
data:  x$WifeHeight[x$WifeHeight > 1450] and
x$HusbandHeight[x$WifeHeight > 1450]
t = -1.9, df = 9, p-value = 0.1
alternative hypothesis: true correlation is not
equal to 0
95 percent confidence interval:
 -0.8559  0.1078
sample estimates:
    cor
-0.5261
```

Correlation

- Removing the outlier changes the direction of the correlation (from positive to negative).
- The result is still not significant at the $\alpha = 0.05$ level, although the p-value is 0.1, suggesting slight evidence against the null hypothesis of no relationship between heights of husbands and wives.
- Note that the negative correlation here means that, with the one outlier couple removed, taller wives tended to be associated with shorter husbands and vice versa.

Correlation

A nonparametric approach for dealing with outliers or otherwise nonnormal distributions for the variables being correlated is to rank the data within each sample and then compute the usual correlation on the ranked data.
—- Note that in the Wilcoxon two-sample test, you pool the data first and then rank the data.
—-For the Spearman correlation, you rank each group separately.

The idea is that large observations will have large ranks in both groups, so that if the data is correlated, large ranks will tend to get paired with large ranks, and small ranks will tend to get paired with small ranks if the data is correlated.
If the data are uncorrelated, then the ranks will be random with respect to each other.

The Spearman correlation is implemented in `cor.test()` using the option `method=‘‘spearman’’`.

- the correlation is negative using the Spearman ranking even with the outlier ——recall that the correlation was positive using the usual (Pearson) correlation.
  ——the correlation was negative when the outlier was removed using the usual Person correlation.
- The results depended so much on the presence of a single observation, I would be more comfortable with the Spearman correlation for this example.

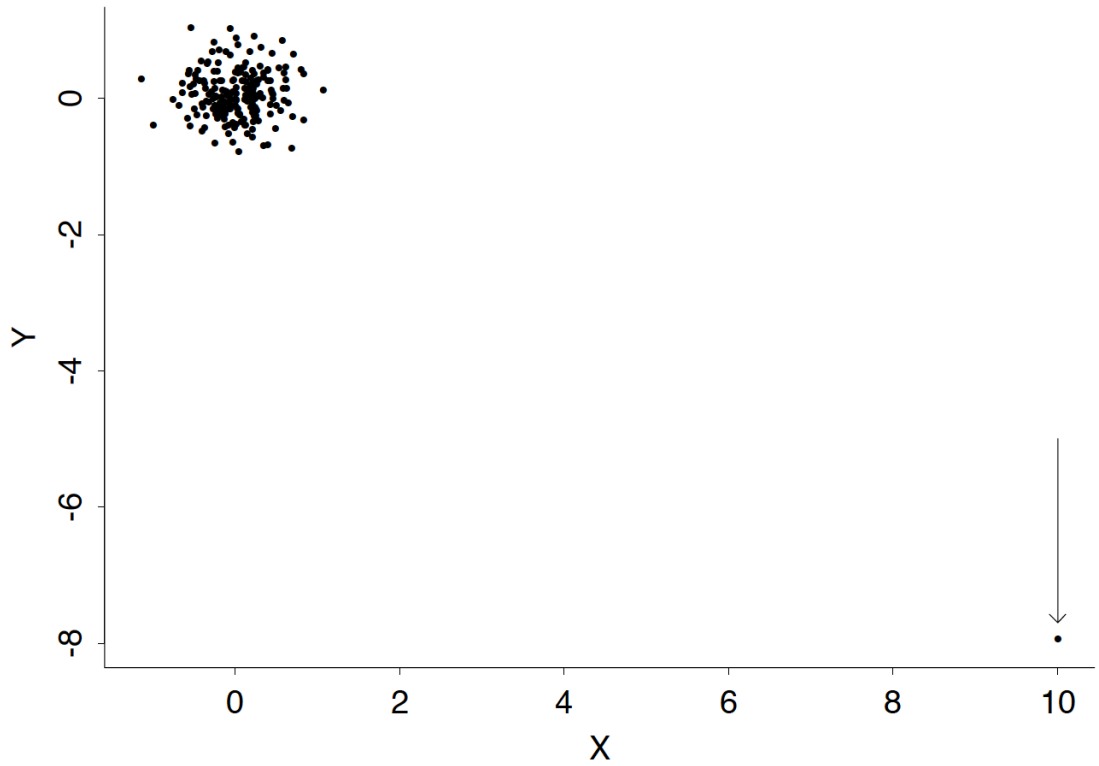
```
cor.test(x$WifeHeight,x$HusbandHeight,method="spearman")

Spearman's rank correlation rho

data:  x$WifeHeight and x$HusbandHeight
S = 370, p-value = 0.3
alternative hypothesis: true rho is not equal to 0
sample estimates:
    rho
-0.3034
```

Correlation

A more extreme example of an outlier. Here the correlation changes from 0 to negative.



Correlation

Comments:

- may have many variables (which often occurs)
- testing every pair of variables for a significant correlation leads to multiple comparison problems, for which you might want to use a Bonferroni correction
  ——-you may limit yourself to only testing a small number of pairs of variable that are interesting.

Regression

In the basic regression model, we assume that the average value of Y has a linear relationship to X, and we write

$$y = \beta_0 + \beta_1 x$$

Here β_0 is the coefficient and β_1 is the slope of the line. This is similar to equations of lines from courses like College Algebra where you write

$$y = a + bx$$

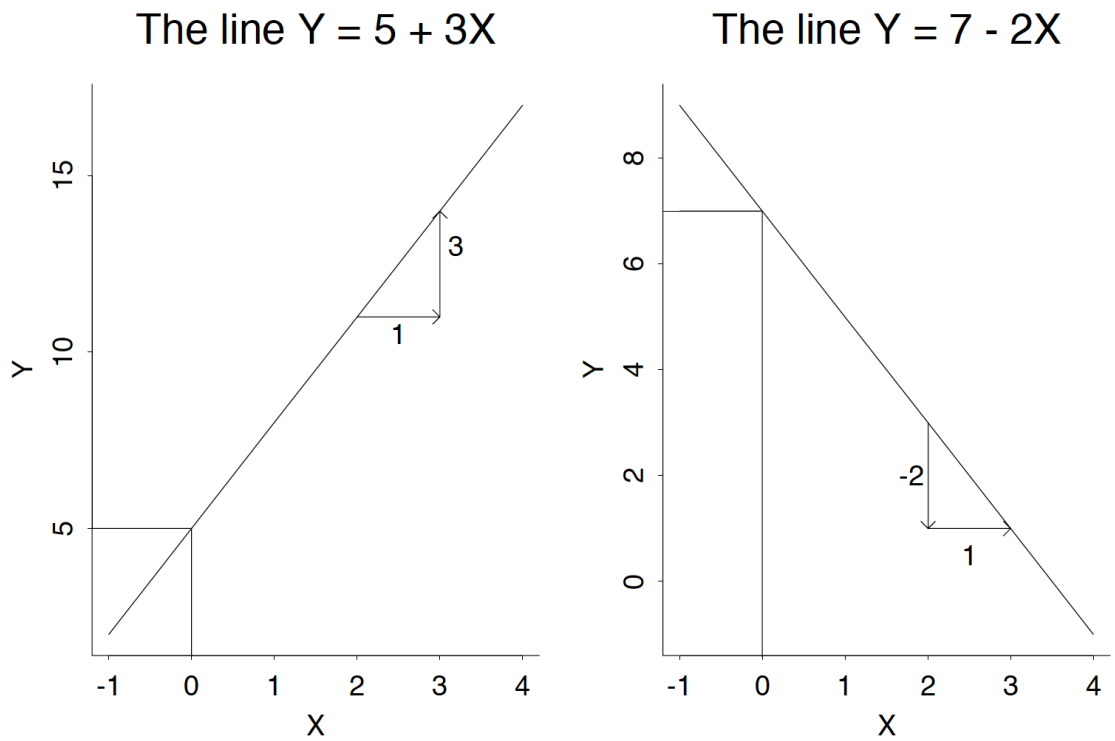
or

$$y = mx + b$$

But we think of β_0 and β_1 as unknown parameters, similar to μ for the mean of a normal distribution. One possible goal of a regression analysis is to make good estimates of β_0 and β_1.

Regression

Review of lines, slopes, and intercepts. The slope is the number of units that y changes for a change of 1 unit in x. The intercept (or y-intercept) is where the line intersects the y-axis.

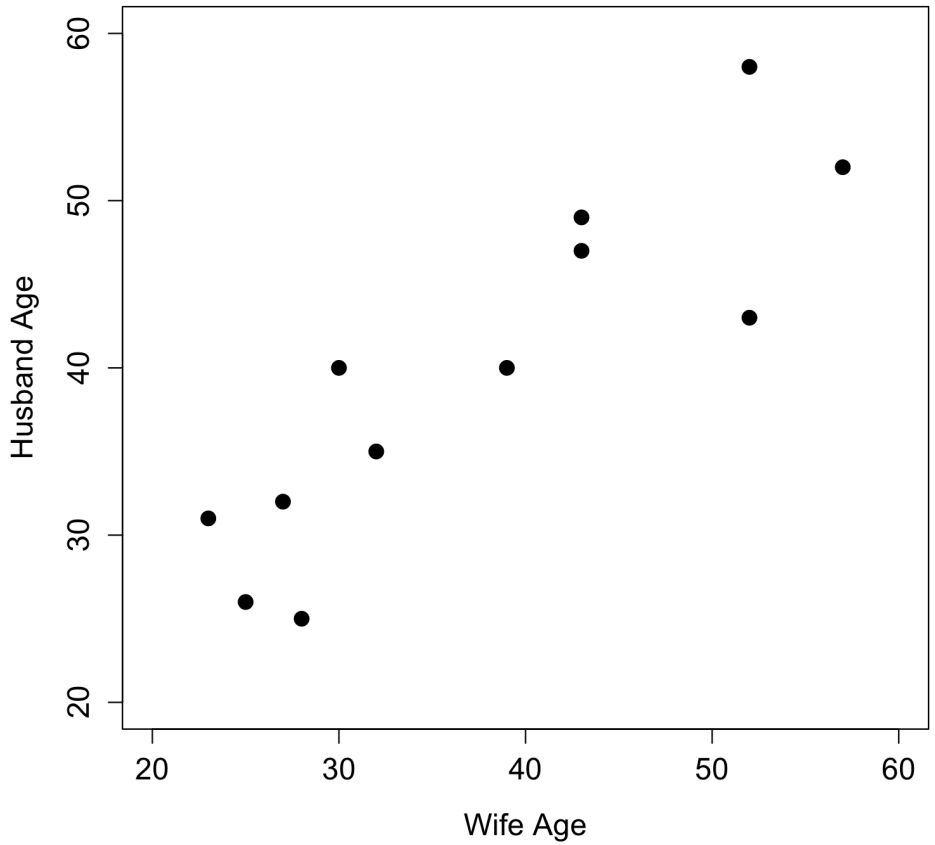


Regression

In real data, the points almost never fall exactly on a line, but there might be a line that describes the overall trend. (This is sometimes even called the trend line). Given a set of points, which line through the points is "best"?

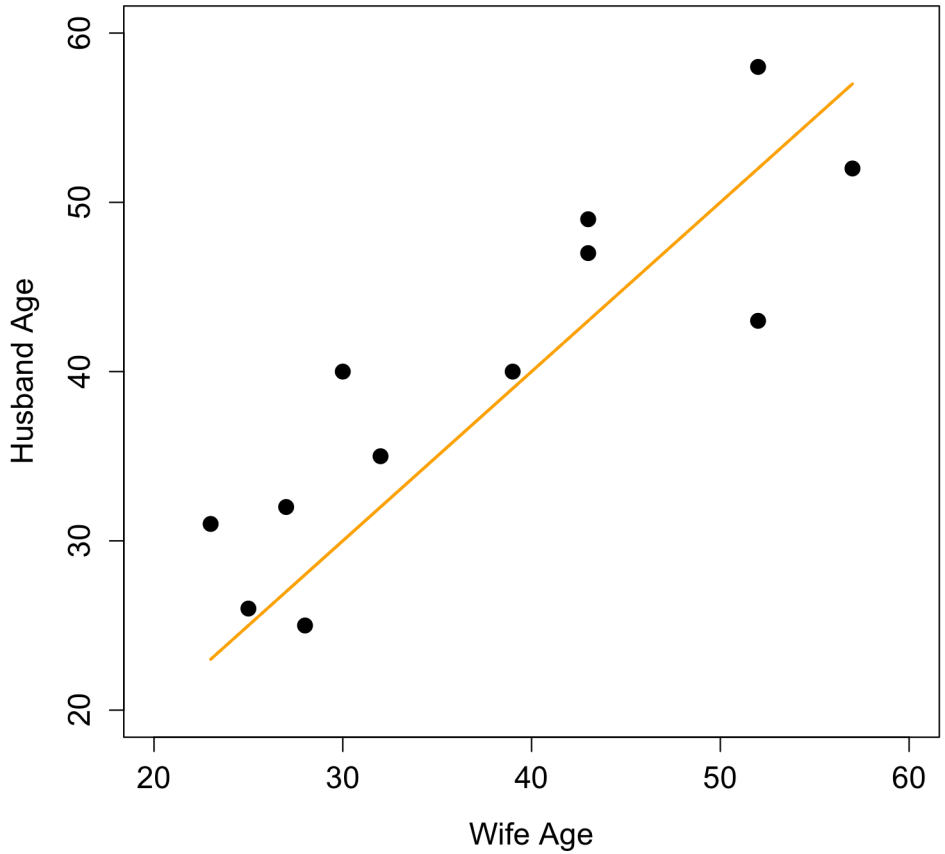
Regression

Husband and wife age example.

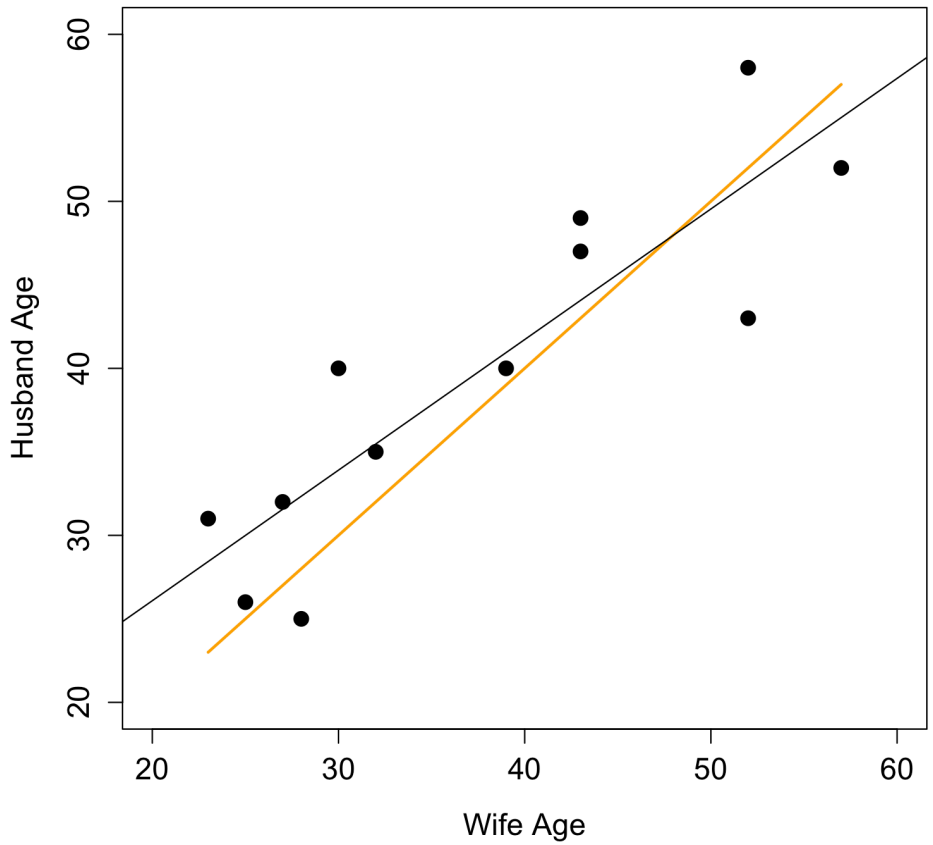


Regression

Husband and wife age example. Here we plot the line $y = x$. Note that 9 out of 12 points are above the line—for the majority of couples, the husband was older than the wife. The points seem a little shifted up compared to the line.

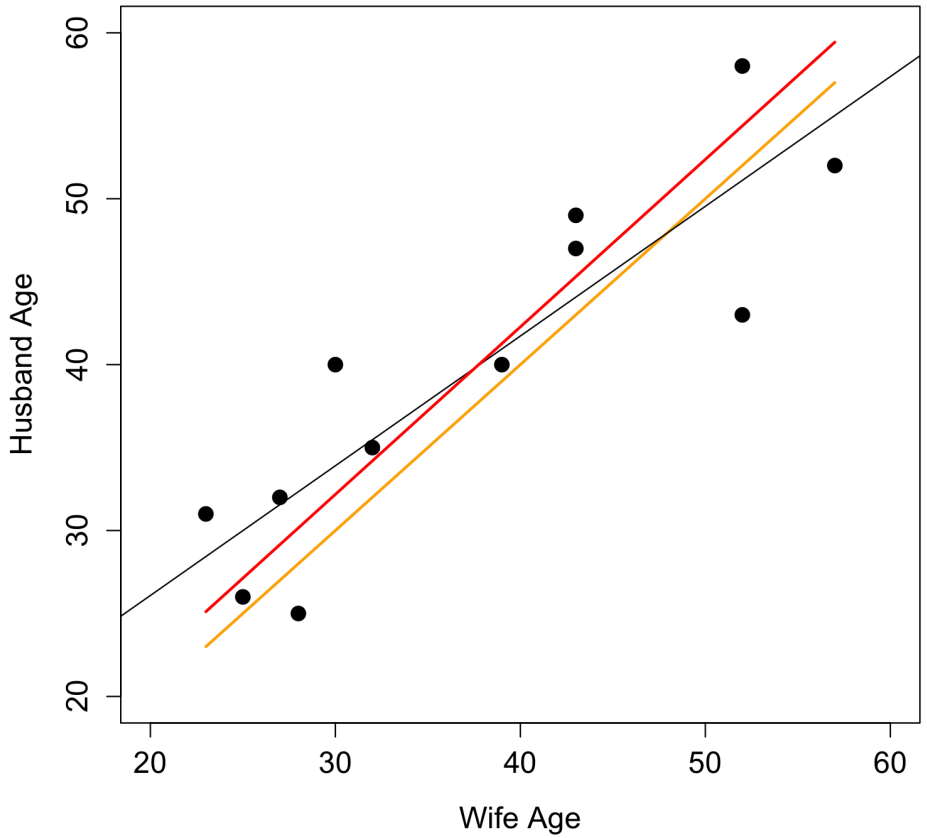


Now we've added the usual regression line in black. It has a smaller slope but a higher intercept. The lines seem to make similar predictions at higher ages, but the black line seems a better fit to the data for the lower ages. Although this doesn't always happen, exactly half of the points are above the black line.



Regression

It is a little difficult to tell visually which line is best. Here is a third line, which is based on regressing the wives' heights on the husbands heights.



Regression

It is difficult to tell which line is "best" or even what is meant by a best line through the data. What to do?
One possible solution to the problem is to consider all possible lines of the form

$$y = \beta_0 + \beta_1 x$$

or here

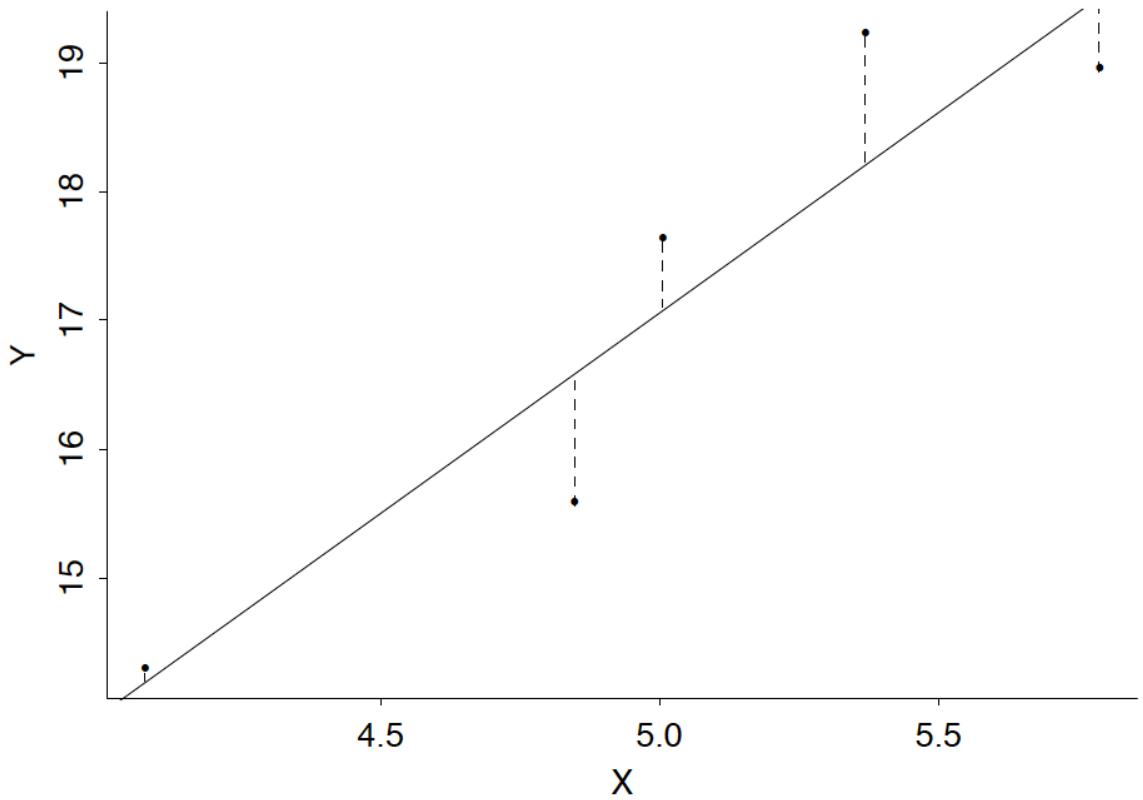
$$\text{Husband height} = \beta_0 + \beta_1 \times (\text{Wife height})$$

In other words, consider all possible choices of β_0 and β_1 and pick the one that minimizes some criterion. The most common criterion used is the **least squares** criterion—here you pick β_0 and β_1 that minimize

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

Regression

Graphically, this means minimizing the sum of squared deviations from each point to the line.



Regression

Rather than testing all possible choices of β_0 and β_1, formulas are known for the optimal choices to minimize the sum of squares. We think of these optimal values as estimates of the true, unknown population parameters β_0 and β_1. We use b_0 or $\widehat{\beta}_0$ to mean an estimate of β_0 and b_1 or $\widehat{\beta}_1$ to mean an estimate of β_1:

$$b_1 = \widehat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = r\frac{S_Y}{S_X}$$

$$b_0 = \widehat{\beta}_0 = \overline{y} - b_1\overline{x}$$

Here r is the Pearson (unranked) correlation coefficient, and S_X and S_Y are the sample standard deviations. Note that if r is positive if, and only if, b_1 is positive. Similarly, if one is negative the other is negative. In other words, r has the same sign as the slope of the regression line.

Regression

The equation for the regression line is

$$\widehat{y} = b_0 + b_1 x$$

where x is an value (not just values that were observed), and b_0 and b_1 were defined on the previous slide. The notation $\widehat{y}$ is used to mean the predicted or average value of y for the given x value. You can think of it as meaning the best guess for y if a new observation will have the given x value.

A special thing to note about the regression line is that it necessarily passes through the point $(\overline{x}, \overline{y})$.

Regression: scatterplot with least squares line

Options make the dots solid and a bit bigger.

```
plot(WifeAge,HusbandAge,xlim=c(20,60),ylim=c(20,60),xlab=
"Wife Age", ylab="Husband Age", pch=16,
cex=1.5,cex.axis=1.3,cex.lab=1.3)
abline(model1,lwd=3)
```

Regression: scatterplot with least squares line

You can always customize your plot by adding to it. For example you can add the point $(\overline{x}, \overline{y})$. You can also add reference lines, points, annotations using text at your own specified coordinates, etc.

```
points(mean(WifeAge),mean(HusbandAge),pch=17,col=''red'')
text(40,30,''r = 0.88'',cex=1.5)
text(25,55,''Hi Mom!'',cex=2)
lines(c(20,60),c(40,40),lty=2,lwd=2)
```

The points statement adds a red triangle at the mean of both ages, which is the point $(37.58, 39.83)$. If a single coordinate is specified by the `points()` function, it adds that point to the plot. To add a curve or line to a plot, you can use `points()` with x and y vectors (just like the original data). For `lines()`, you specify the beginning and ending x and y coordinates, and R fills in the line.

Regression

To fit a linear regression model in R, you can use the `lm()` command, which is similar to `aov()`.
The following assumes you have the file `couple.txt` in the same directory as your R session:

```
x <- read.table("couples.txt",header=T)
attach(x)
model1 <- lm(HusbandAge ~ WifeAge)
summary(model1)
```

Regression

```
Call:
lm(formula = HusbandAge ~ WifeAge)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1066 -3.2607 -0.0125  3.4311  6.8934

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4447     5.2350   1.995 0.073980 .
WifeAge       0.7820     0.1334   5.860 0.000159 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.197 on 10 degrees of freedom
Multiple R-squared:  0.7745, Adjusted R-squared:  0.7519
F-statistic: 34.34 on 1 and 10 DF,  p-value: 0.0001595
```

Regression

The `lm()` command generates a table similar to the ANOVA table generated by `aov()`.

To go through some elements in the table, it first fives the formula used to generate the output. This is useful when you have generated several models, say `model1, model2, model3, ...` and you can't remember how you generated the model. For example, you might have one model with an outlier removed, another model with one of the variables on a log-transformed scale, etc.

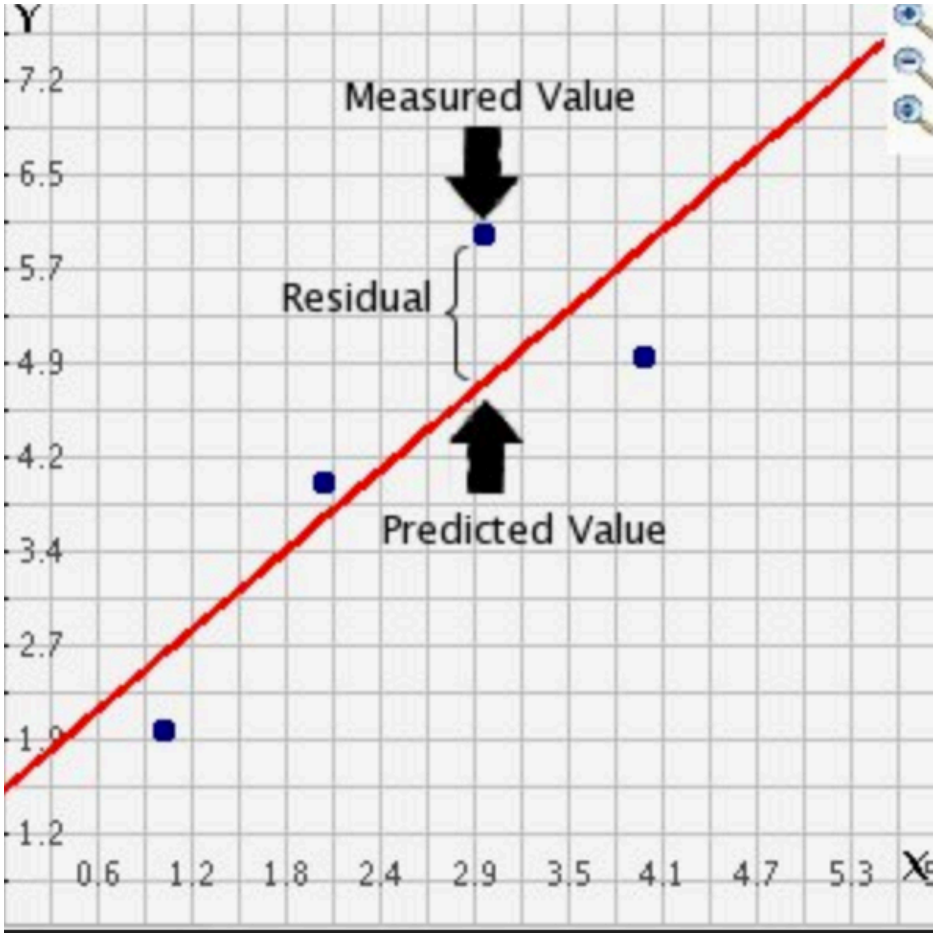
The next couple lines deal with **residuals**. Residuals are the difference between theobserved and fitted values, That is

$$y_i - \widehat{y}_i = y_i - (b_0 + b_1 x_i)$$

Regression

From the web:

www.shodor.org/media/M/T/I/mYzliZjY4ZDc0NjI3YWQ3YWVlM2MzZmUzN2MwOWY.jpg



Regression

The next part gives a table similar to ANOVA. Here we get the estimates for the coefficients, b_0 and b_1 in the first quantitative column. We also get standard errors for these, corresponding t-values and p-values. The p-values are based on testing the null hypotheses

$$H_0 : \beta_0 = 0$$

and

$$H_0 : \beta_1 = 0$$

The first null hypothesis says that the intercept is 0. For this problem, this is not very meaningful, as it would mean that the husband of 0-yr old woman would also be predicted to be a 0-yr old!

Often the intercept term is not very meaningful in the model. The second null hypothesis is that the slope is 0, which would mean that the wife's age increasing would not be associated with the husband's age increasing.

Regression

For this eample, we get a significant result for the wife's age. This means that the wife's age has some statistically signficant ability to predict the husband's age. The coefficients give the model

$$\text{Mean Husband's Age} = 10.4447 + 0.7820 \times (\text{Wife's Age})$$

The low p-value for the Wife's age, suggest that the coefficient 0.7820 is statistically significantly different from 0. This means that the data show there is evidence that the wife's age is associated with the husband's age. The coefficient of 0.7820 means that for each year of increase in the wife's age, the mean husband's age is predicted to increase by 0.782 years.

As an example, based on this model, a 20-yr old women who was married would be predicted to have a husband who was

$$10.4447 + (0.782)(30) = 33.9$$

or about 34 years old. A 50 yr-old women would be predicted to have husband who was

$$10.4447 + (0.782)(55) = 53.5$$

Regression

The fitted values are found by plugging in the observed x values (Wife ages) into the regression equation. This gives the expected husband ages for each wife. They are given automatically using

```
model1$fitted.values
       1       2       3        4        5        6        7        8
44.06894 32.33956 33.90348 55.01637 51.10658 31.55760 51.10658 44.06894
       9       10       11       12
28.42976 29.99368 40.94111 35.46740
x$WifeAge
 [1] 43 28 30 57 52 27 52 43 23 25 39 32
```

For example, if the wife is 43, the regression equation predicts $10.4447 + (0.782)(43) = 44.069$ for the husband age.

Regression

To see what is stored in `model1`, type

```
names(model1)
# [1] "coefficients"  "residuals"     "effects"       "rank"
# [5] "fitted.values" "assign"        "qr"            "df.residual
# [9] "xlevels"       "call"          "terms"         "model"
```

The residuals are also stored, which are the observed husband ages minus the fitted values.

$$e_i = y_i - \widehat{y}_i$$

Regression: ANOVA table

More details about the regression can be obtained using the `anova()` command on the `model1` object:

```
> anova(model1)
Analysis of Variance Table

Response: HusbandAge
           Df Sum Sq Mean Sq F value    Pr(>F)
WifeAge     1 927.53  927.53  34.336 0.0001595 ***
Residuals 10 270.13   27.01
```

Here the sum of squared residuals, `sum(model1$residuals`^{2}`)` is 270.13.

Regression: ANOVA table

Other components from the table are (SS means sums of squares):

$$\text{Residual SS} = \sum_{i=1}^{n} e_i^2$$

$$\text{Total SS} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$\text{Regression SS} = b_1 \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$\text{Regression SS} = \text{Total SS} - \text{Residual SS}$$

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = r^2$$

The Mean Square values in the table are the SS values divided by the degrees of freedom. The degrees of freedom is $n - 2$ for the residuals and 1 for the 1 predictor. The F statistic is MSR/MSE (Measn square for regression divided by mean square error), and the p-value can be based on the F statistic.

Regression

Note that $R^2 = 1$ occurs when the Regression SS is equal to the Total SS. This means that the Residual SS is 0, so all of the points fall on the line. In this case, $r = 1$ and $R^2 = 1$.

On the other hand, $R^2 = 0$ means that the Total SS is equal to the Residual SS, so the Regression SS is 0. We can think of the Regression SS and Residual SS as partitioning the Total SS:

$$\text{Total SS} = \text{Regression SS} + \text{Residual SS} \text{ or } \frac{\text{Regression SS}}{\text{Total SS}} + \frac{\text{Residual SS}}{\text{Total SS}} = 1$$

If a large proportion of the Total SS is from the Regression rather than from Residuals, then R^2 is high. It is common to say that R^2 is a measure of how much variation is *explained by* the predictor variable(s). This phrase should be used cautiously because it doesn't refer to a causal explanation.

Regression

For the husband and wife and example for ages, the R^2 value was 0.77. This means that 77% of the variation in husband ages was "explained by" variation in the wife ages. Since R^2 is just the correlation squared, regressing wife ages on husband ages would also result in $R^2 = 0.77$, and 77% of the variation in wife ages would be "explained by" variation in husband ages. Typically, you want the R^2 value to be high, since this means you can use one variable (or a set of variables) to predict another variable.

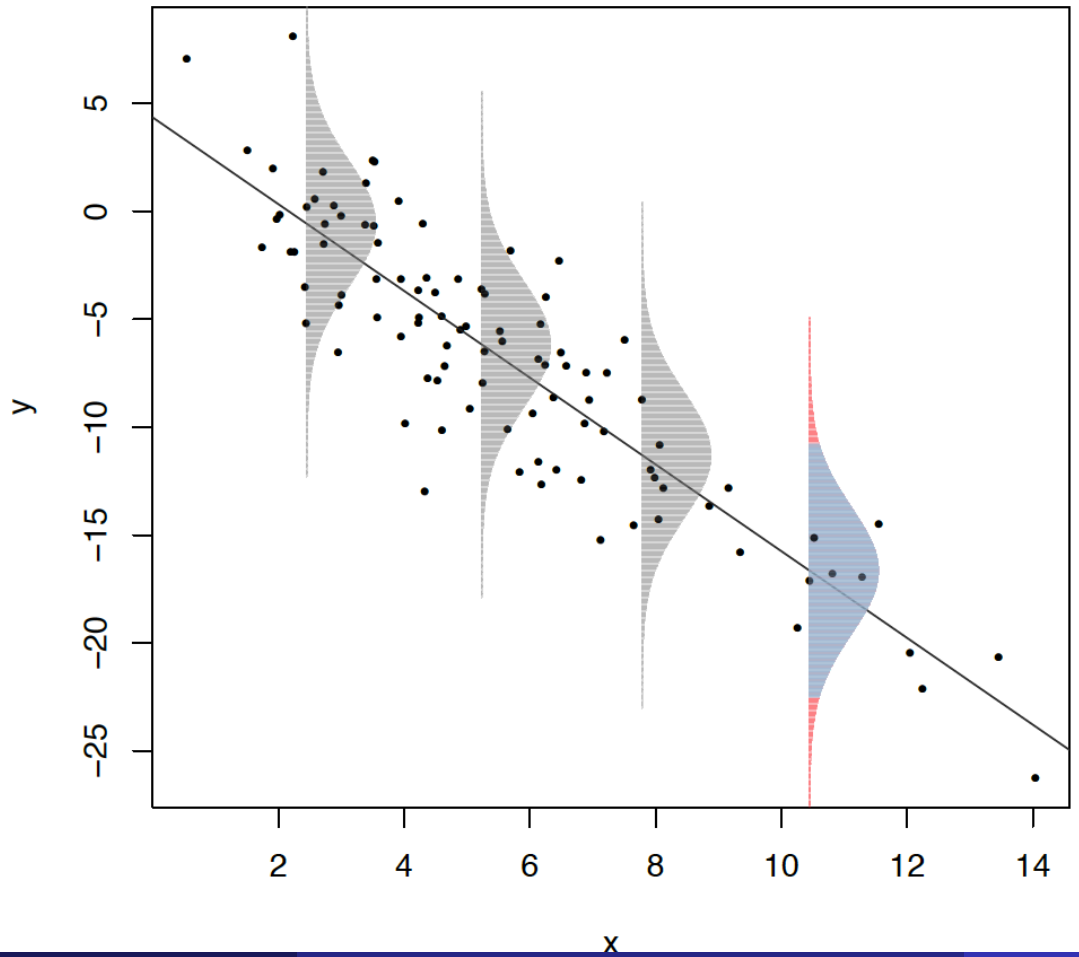
Regression

The least-squares line is mathematically well-defined and can be calculated without thinking about the data probabilitistically. However, p-values and tests of significance assume the data follow a probabilistic model with some assumptions. Assumptions for regression include the following:

- each pair (x_i, y_i) is independent
- The expected value of y is a linear function of x: $E(y) = \beta_0 + \beta_1 x$, sometimes denoted $\mu_{Y|X}$
- the variability of y is the same for each fixed value of x. This is sometimes denoted $\sigma^2_{y|x}$
- the distribuiton of y given x is normally distributed with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2_{y|x}$
- in the model, x is not treated as random

Regression

Note that the assumption that the variance is the same regardless of x is similar to the assumption of equal variance in ANOVA.



Regression

Less formally, the assumptions in their order of importance, are:

1. **Validity**. Most importantly, the data you are analyzing should map to the research question you are trying to answer. This sounds obvious but is often overlooked or ignored because it can be inconvenient.
2. **Additivity and Linearity**. The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors.
3. **Independence of errors** (i.e., residuals). This assumption depends on how the data were collected.
4. **Equal variance of errors**.
5. **Normality of errors**.

It is easy to focus on the last two (especially when teaching) because the first assumptions depend on the scientific context and are not possible to assess just looking at the data in a spreadsheet.

Regression

To get back to the regression model, the parameters of the model are β_0, β_1 and σ^2 (which we might call $\sigma^2_{Y|X}$, but it is the same for every x). Usually σ^2 is not directly of interest but is necessary to estimate in order to do hypothesis tests and confidence intervals for the other parameters.

$\sigma^2_{Y|X}$ is estimated by

$$s^2_{Y|X} = \text{Residual MS} = \frac{\text{Residual SS}}{\text{Residual df}} = \frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}$$

This formula is similar to the sample variance, but we subtract the predicted values for y instead of the mean for y, $\overline{y}$, and we divide by $n-2$ instead of dividing by $n-1$. We can think of this as two degrees of freedom being lost since β_0 and β_1 need to be estimated. Usually, the sample variance uses $n-1$ in the denominator due to one degree of freedom being lost for estimating μ_Y with $\overline{y}$.

Regression

Recall that there are observed residuals, which are observed minus fitted values, and unobserved residuals:

$$e_i = y_i - \widehat{y}_i = y_i - (b_0 + b_1)x_i$$
$$\varepsilon_i = y_i - E(y_i) = y_i - (\beta_0 + \beta_1)x_i$$

The difference in meaning here is whether the estimated versus unknown regression coefficients are used. We can think of e_i as an estimate of ε_i.

Regression

Two ways of writing the regression model are

$$E(y_i) = \beta_0 + \beta_1 x_i$$

and

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

Regression

To get a confidence interval for β_1, we can use

$$b_1 \pm t_{crit} SE(b_1)$$

where

$$SE(b_1) = \frac{s_{Y|X}}{\sqrt{\sum_i (x_i - \overline{x})^2}}$$

Here t_{crit} is based on the Residual df, which is $n - 2$.

To test the null hypothesis that $\beta_1 = \beta_{10}$ (i.e. a particular value for β_1, you can use the test statistic

$$t_{obs} = \frac{b_1 - \beta_{10}}{SE(b_1)}$$

and then compare to a critical value (or obtain a p-value) using $n - 2$ df (i.e., the Residual df).

Regression

The p-value based on the R output is for testing $H_0 : \beta_1 = 0$, which corresponds to a flat regression line. But the theory allows testing any particular slope. For the couples data, you might be interested in testing $H_0 : \beta_1 = 1$, which would mean that for every year older that the wife is, the husband's age is expected to increase by 1 year.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4447     5.2350   1.995 0.073980 .
WifeAge       0.7820     0.1334   5.860 0.000159 ***
```

To test $H_0 : \beta_1 = 1$, $t_{obs} = (0.782 - 1.0)/(.1334) = -1.634$. The critical value is `qt(.975,10) = 2.22`. So comparing $|-1.634|$ to 2.22 for a two-sided test, we see that the observed test statistic is not as extreme as the critical value, so we cannot conclude that the slope is significantly different from 1. For a p-value, we can use `pt(-1.634,10)*2 = 0.133`.

Regression

Instead of getting values such as the SE by hand from the R output, you can also save the output to a variable and extract the values. This reduces roundoff error and makes it easier to repeat the analysis in case the data changes. For example, from the previous example, we could use

```
model1.values <- summary(model1)
b1 <- model1.values$coefficients[2,1]
b1
#[1] 0.781959
se.b1 <- model1.values$coefficients[2,2]
t <- (b1-1)/se.b1
t
#[1] -1.633918
```

The object `model1.values$coefficients` here is a matrix object, so the values can be obtained from the rows and columns.

Regression

For the CI for this example, we have

```
df <- model1.values$fstatistic[3] # this is hard to find
t.crit <- qt(1-0.05/2, df)
CI.lower <- b1 - t.crit * se.b1
CI.upper <- b1 + t.crit * se.b1
print(c(CI.lower,CI.upper))
#[1] 0.4846212 1.0792968
```

Consistent with the hypothesis test, the CI includes 1.0, suggesting we can't be confident that the ages of husbands increase at a different rate from the ages of their wives.

Regression

As mentioned earlier, the R output tests $H_0 : \beta_1 = 0$, so you need to extract information to do a different test for the slope. We showed using a t-test for testing this null hypothesis, but it is also equivalent to an F test. Here the F statistic is t^2_{obs} when there is only 1 numerator degree of freedom (one predictor in the regression).

```
t <- (b1-0)/se.b1
t
#[1] 5.859709
t^2
#[1] 34.33619
```

which matches the F statistic from earlier output.

In addition, the p-value matches that for the correlation using `cor.test()`. Generally, the correlation will be significant if and only if the slope is signficantly different from 0.

Regression

Another common application of confidence intervals in regression is for the regression line itself. This means getting a confidence interval for the expected value of y for each value of x.
Here the CI for y given x is

$$b_0 + b_1 x \pm t_{crit} s_{Y|X} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$$

where the critical value is based on $n - 2$ degrees of freedom.

Regression

In addition to a confidence interval for the mean, you can made **prediction intervals** for a new observation. This gives a plausible interval for a new observation. Here there are two sources of uncertainty: uncertainty about the mean, and uncertainty about how much an individual observation deviates from the mean. As a result, the prediction interval is wider than the CI for the mean.

The prediction interval for y given x is

$$b_0 + b_1 x \pm t_{crit} s_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$$

Regression

For a particular wife age, such as 40, the CIs and PIs (prediction intervals) are done in R by

```
predict(model1,data.frame(WifeAge=40), interval="confidence",
level=.95)
#       fit      lwr      upr
#1 41.72307 38.30368 45.14245
predict(model1,data.frame(WifeAge=40), interval="prediction",
level=.95)
#       fit     lwr      upr
#1 41.72307 29.6482 53.79794
```

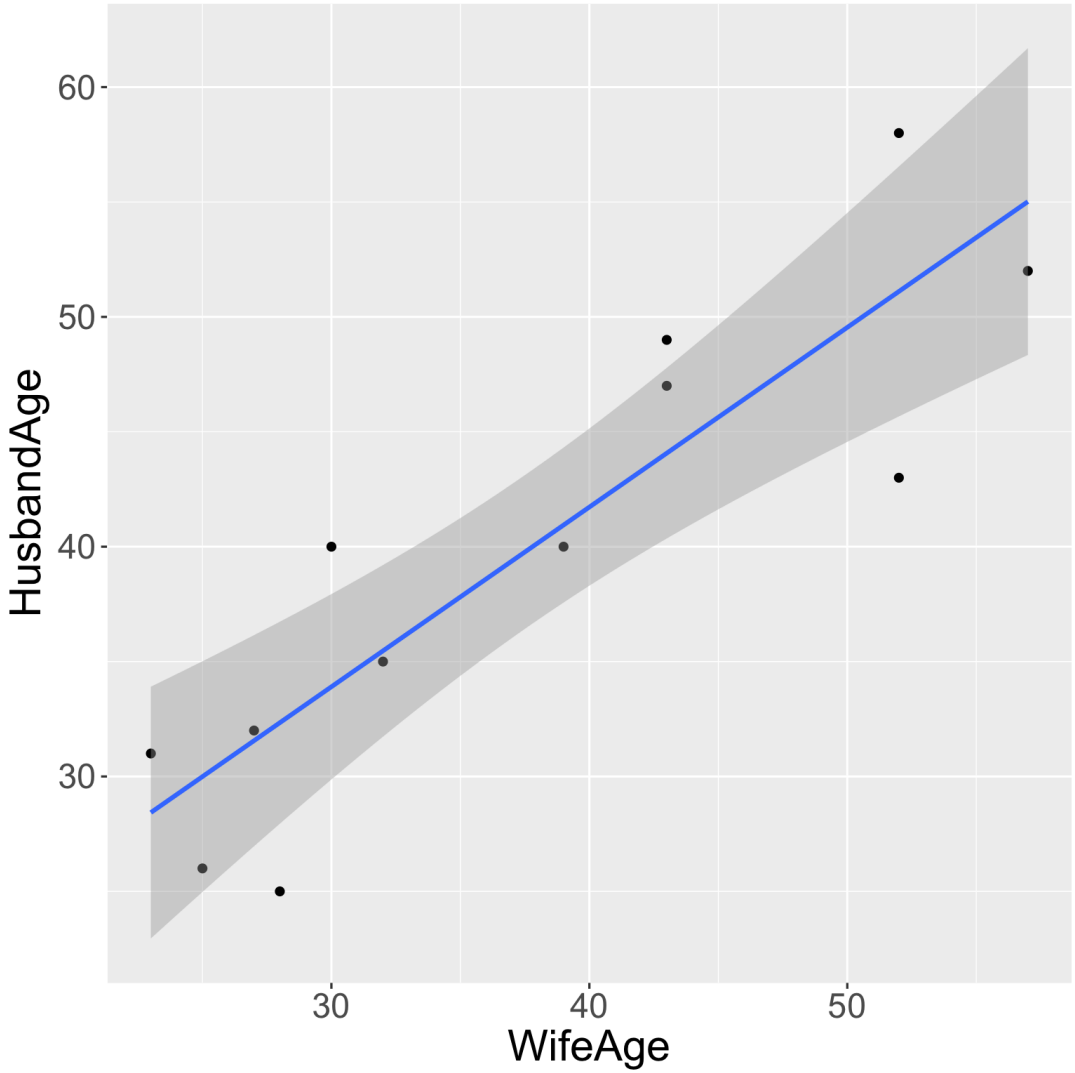
Here the predicted husband's age for a 40-yr old wife is 41.7 years. A CI for the mean age for the husband is (38.3,45.1), but a prediction interval is that 95% of husbands for a wife this age would be between 29.6 and 53.8 years old. There is quite a bit more uncertainty for an individual compared to the population average.

Regression

To plot the CIs at each point, here is some R code:

```
library(ggplot2)
p <- ggplot(x, aes(x = WifeAge, y = HusbandAge))
p <- p + geom_point()
p <- p + geom_smooth(method = lm, se = TRUE)
p <- p +  theme(text = element_text(size=20))
print(p)
```

Regression



Regression

Note that the confidence bands are narrow in the middle of the data. This is how these bands usually look. There is less uncertainty near the middle of the data than there is for more extreme values. You can also see this in the formula for the SE, which has $(x - \overline{x})^2$ inside the square root. This is the only place where x occurs by itself. The further it is from $\overline{x}$, the larger this value is, and therefore the larger the SE is.

Regression

There is a large literature on regression diagnostics. This involves checking the assumptions of the regression model. Some of the most important assumptions, such as that observations are independent observations from a single population (e.g., the population of married couples), cannot be checked just by looking at the data. Consequently, regression diagnostics often focus on what can be checked by looking at the data.

To review, the model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where

1. the observed data are a random sample
2. the average y value is linearly related to x
3. the variation in y given x is independent of x (the variability is the same for each level of x
4. the distribution of responses for each x is normally distributed with mean $\beta_0 + \beta_1 x$ (which means that ε is normal with mean 0)

Regression

The following plots show examples of what can happen in scatterplots:

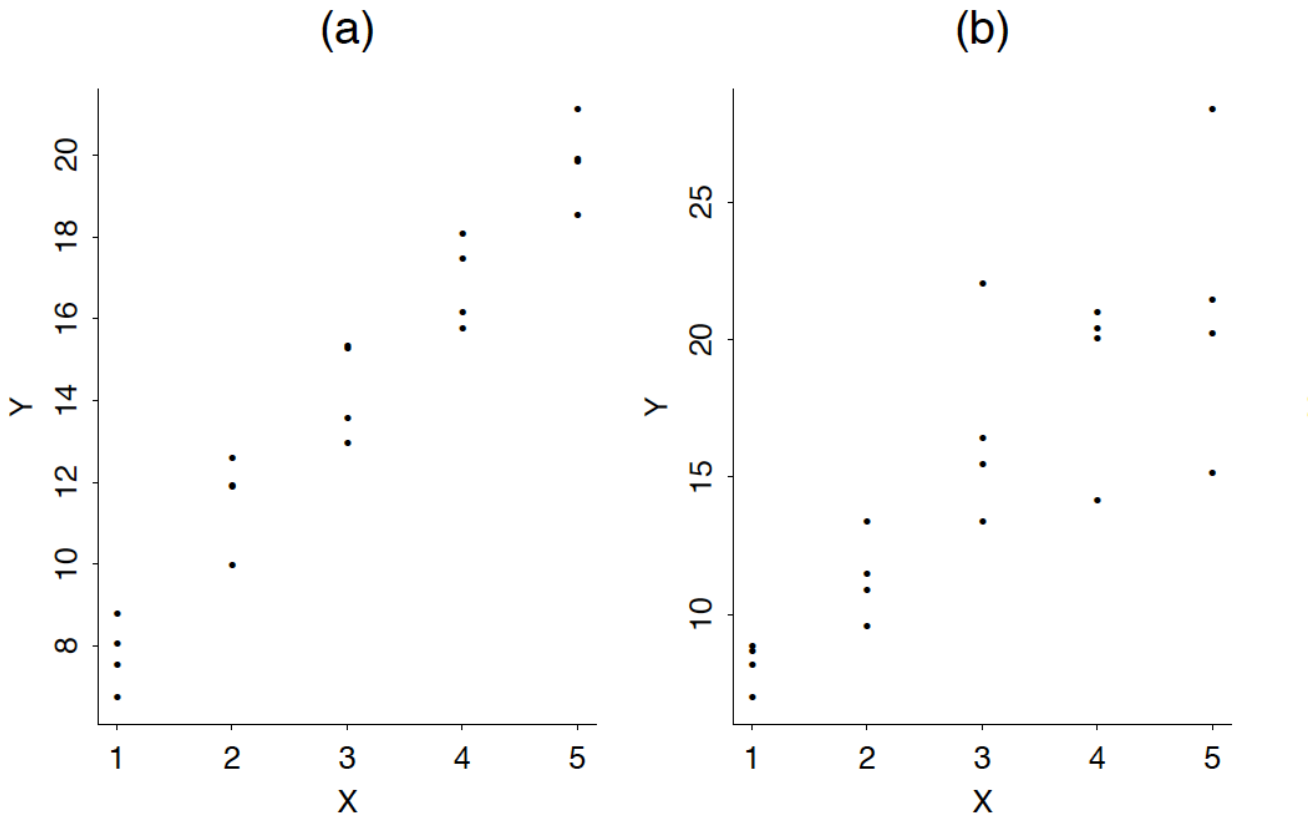
(a) Model assumptions appear to be satisfied

(b) The relationship appears linear, but the variance appears nonconstant

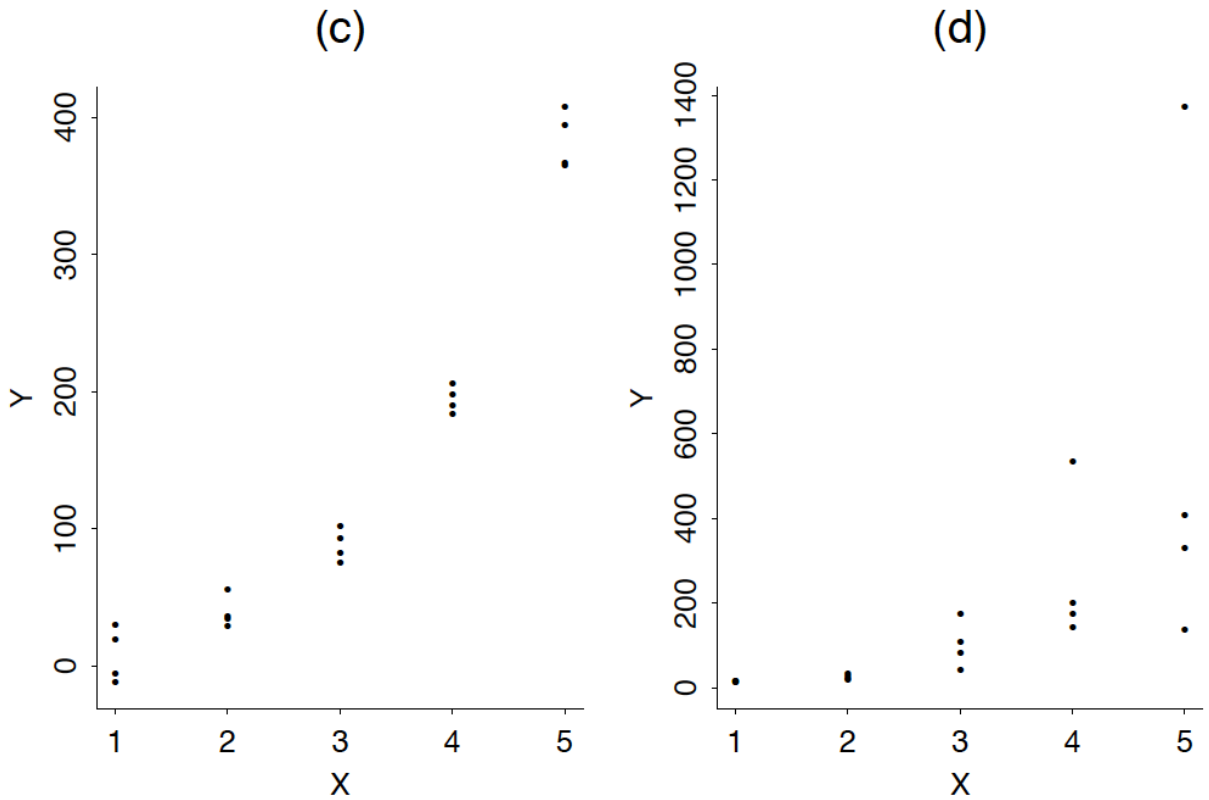
(c) The relationship appears nonlinear, althought the variance is appears to be constant

(d) The relationship is nonlinear and the variance is not constant

Regression



Regression


(c)
Y
0
100
200
300
400
1
2
3
4
5
X
(d)
Y
0
200
400
600
800
1000
1200
1400
1
2
3
4
5
X

Regression

Regression diagnostics are often based on examinging the residuals:

$$e_i = y_i - \widehat{y}_i = y_i - (b_0 + b_1 x_i)$$

Based on the model assumptions, the residuals should be normally distributed with mean 0 and some variance σ^2. To standardize the residuals, they are often divided by their standard error. Here r_i is called the **studentized residual** or **standardized residual**:

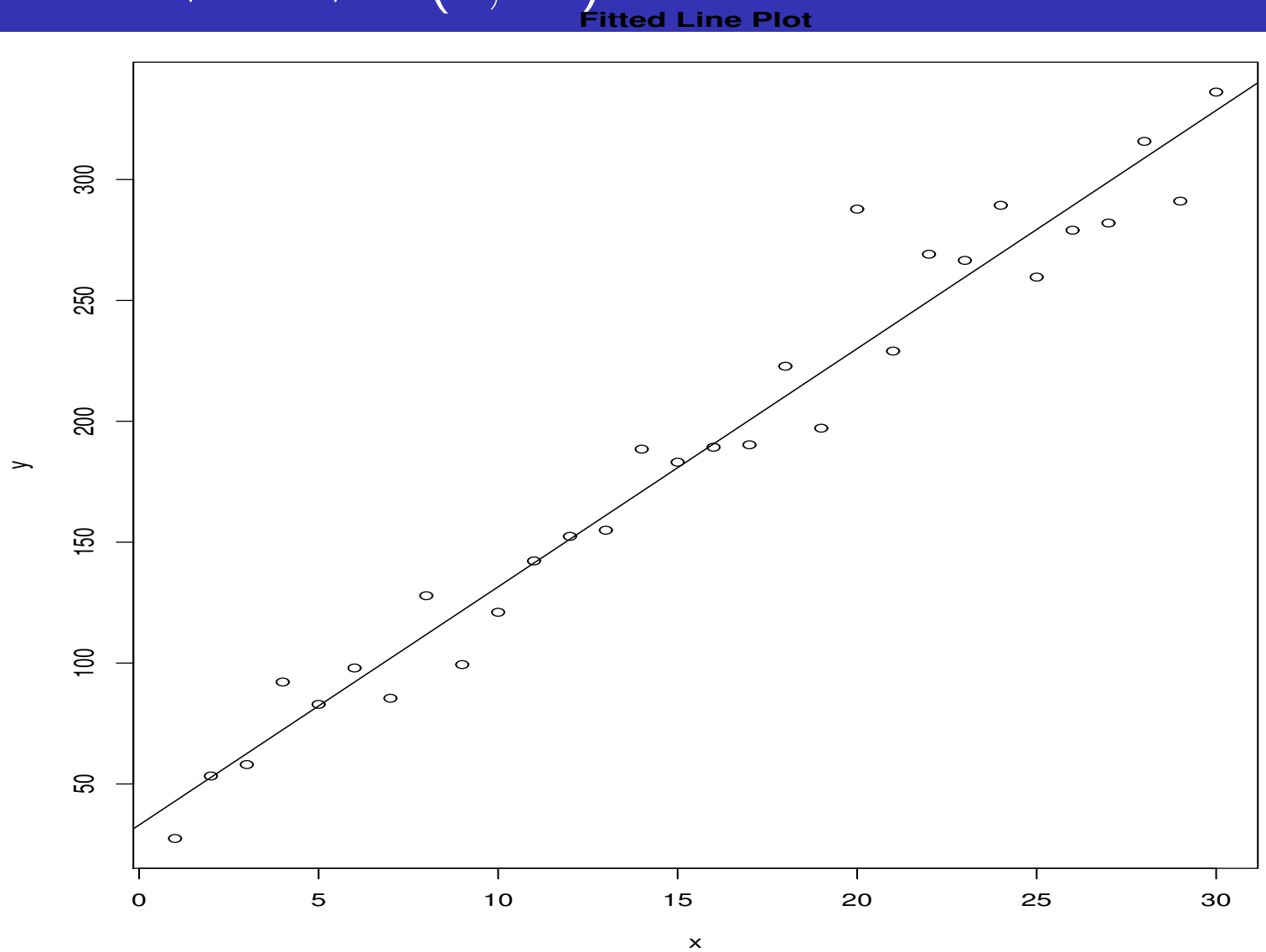
$$r_i = \frac{e_i}{SE(e_i)} = \frac{e_i}{s_{Y|X}\sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}}$$

Regression

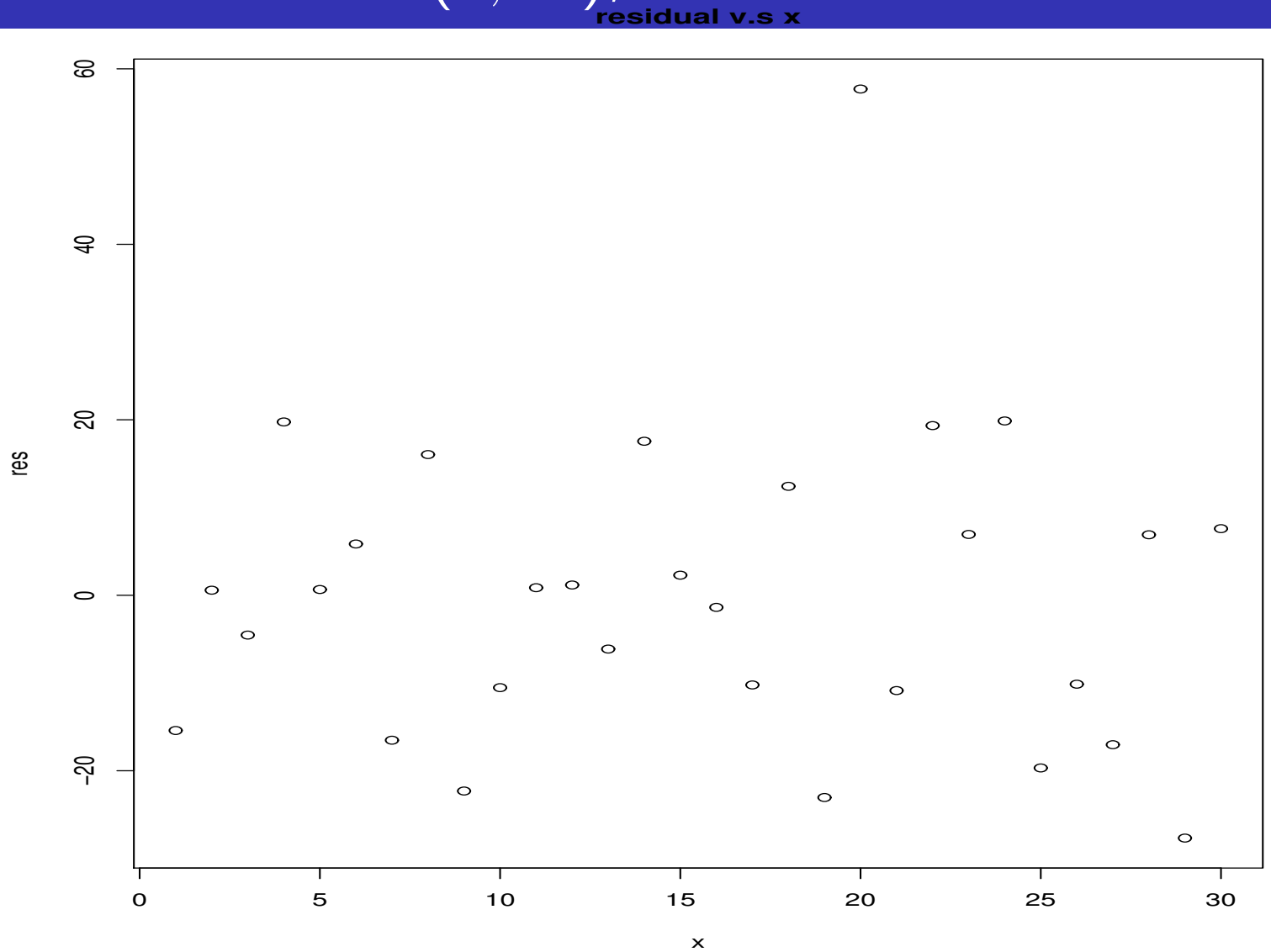
The studentized residuals are standard normal (if the model is correct), so most studentized residuals should be between -2 and +2, just like z-scores. **Often studentized residuals are plotted against the fitted values or versus x, $\widehat{y}_i$.**

If the linear regression function is **appropriate**, there should be **no obvious trend** in the residual plots.
If the linear regression function is **not appropriate**, then the residual plots **might show a trend**;

Plot of y v.s x, data were generated from $y = 10 * x + 30 + N(0, 25)$

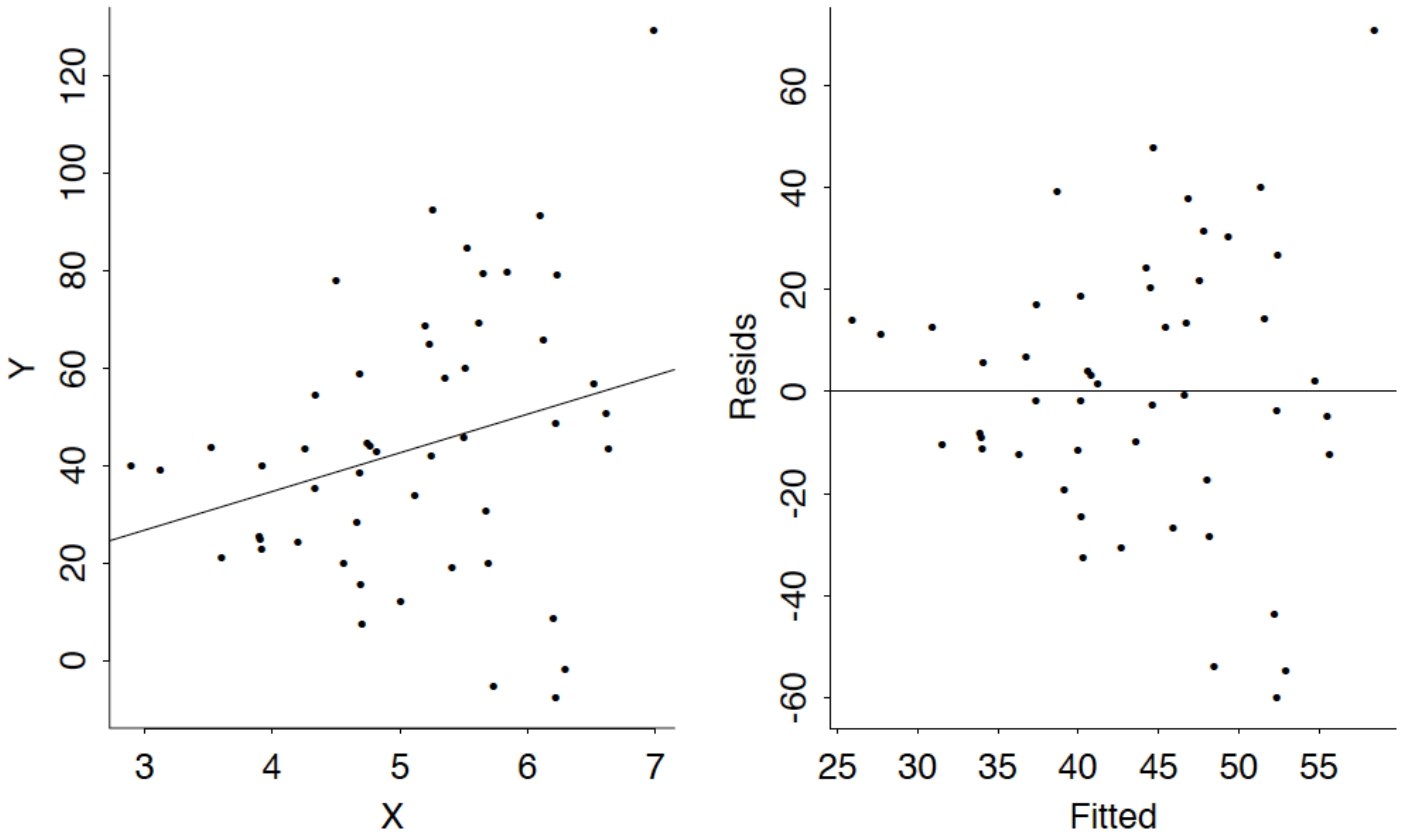


Plot of residual v.s x,data were generated from $y = 10 * x + 30 + N(0, 25)$, no obvious trend observed



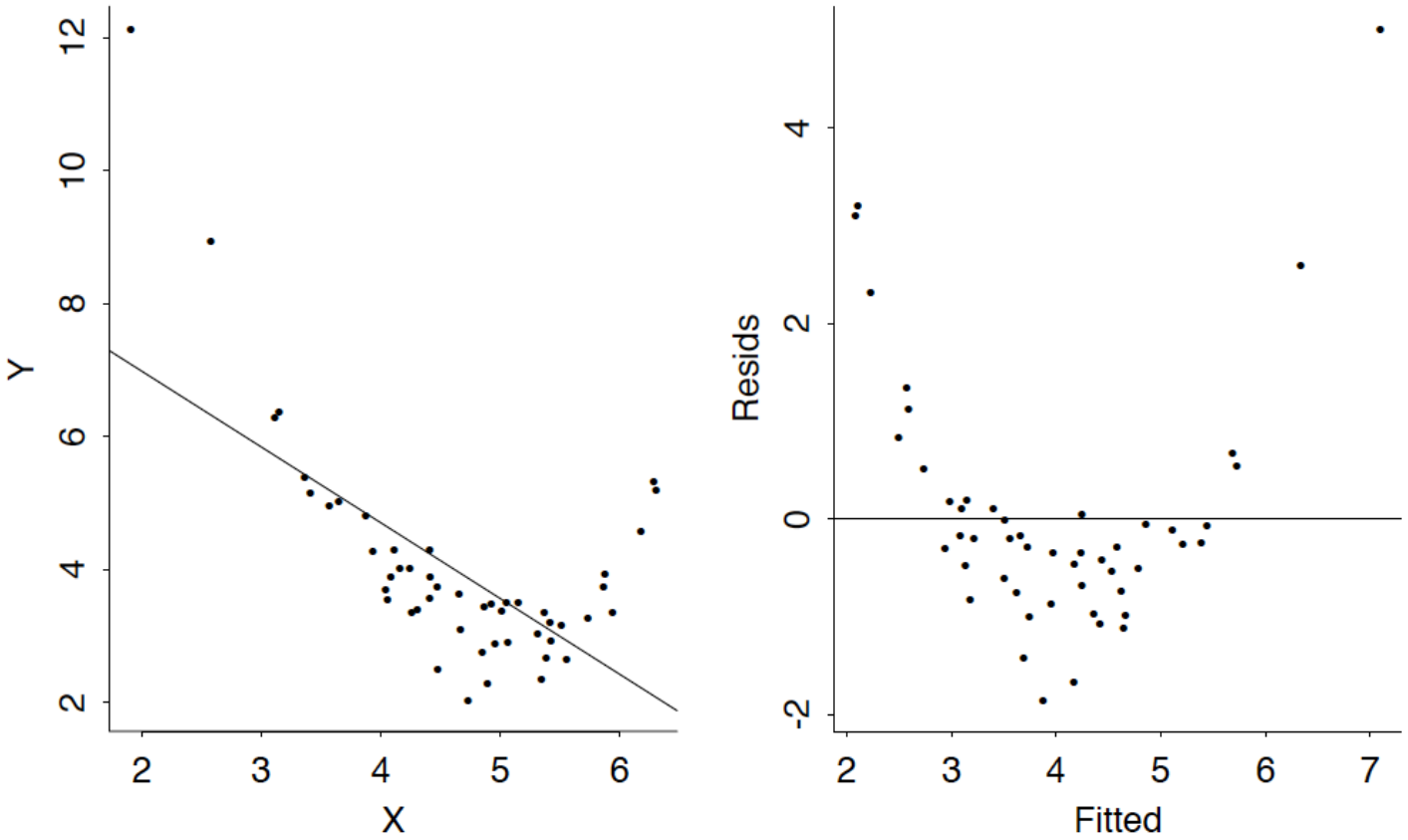
Regression

Bad residual plot (funnel shape).



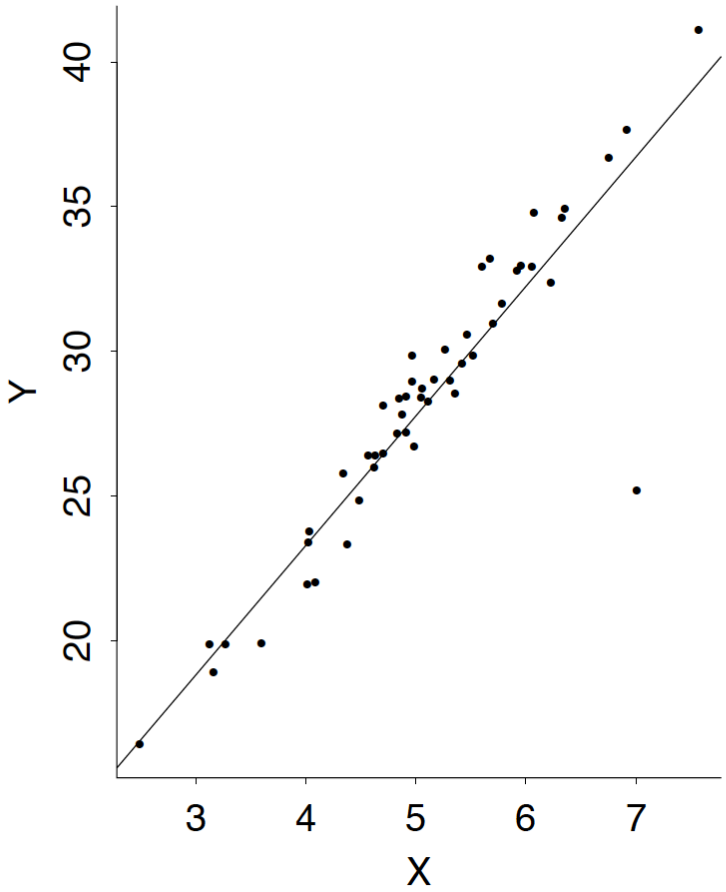
Regression

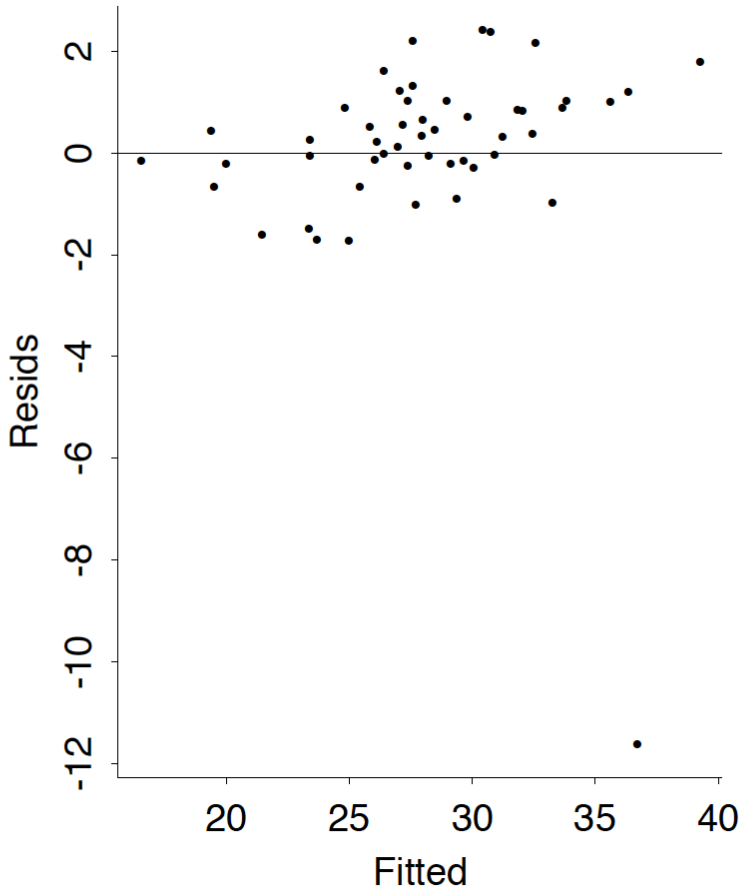
Bad residual plot (U-shape).



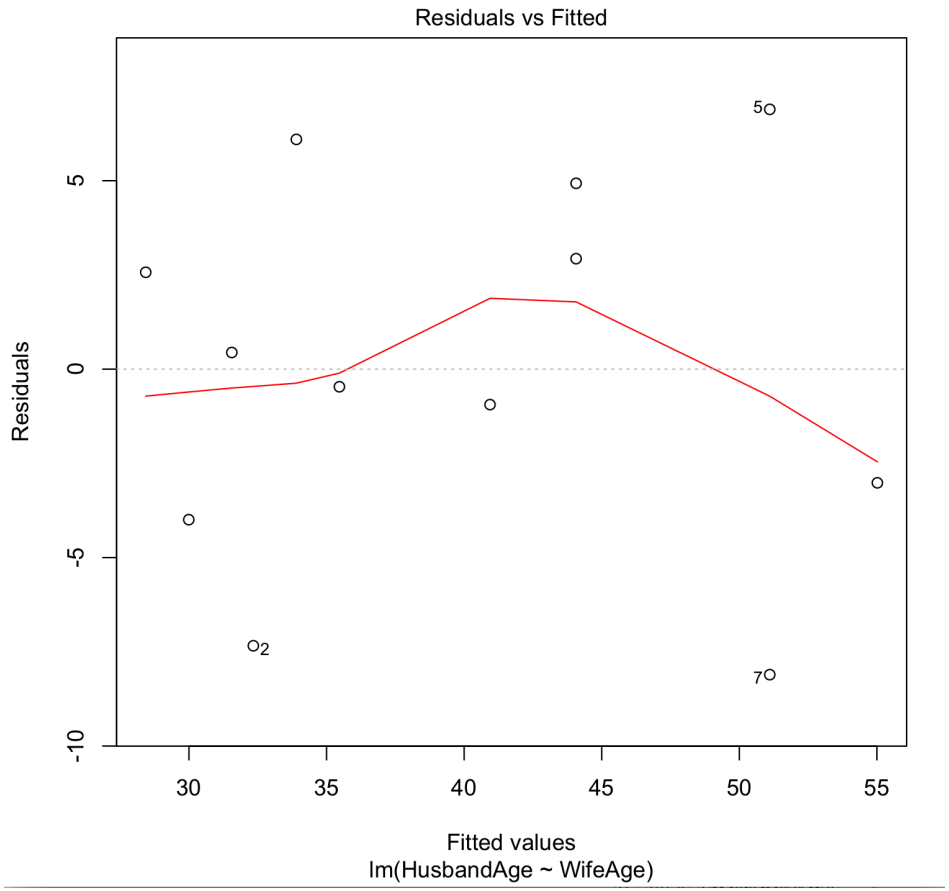
Regression

Residual plot with extreme outlier.





Typing `plot(model1)` (or whatever the name of your model from `lm()`) and then return several times, R will plot several diagnostic plots. The first is the residuals (not studentized) against the fitted values. This help you look for outliers, nonconstant variance and curvature in the residuals.



Regression

Another type of residual is the externally Studentized residual (also called Studentized deleted residual or deleted t-residual), which is based on rerunning the analysis without the ith observation, then determining what the difference would be between y_i and $b_0 + b_1 x_i$, where b_0 and b_1 are estimated with the pair (x_i, y_i) removed from the model. This could be done by tediously refitting the regression model n times for n pairs of data, but this can also be done automatically in the software, and there are computational ways to make it reasonably efficient.

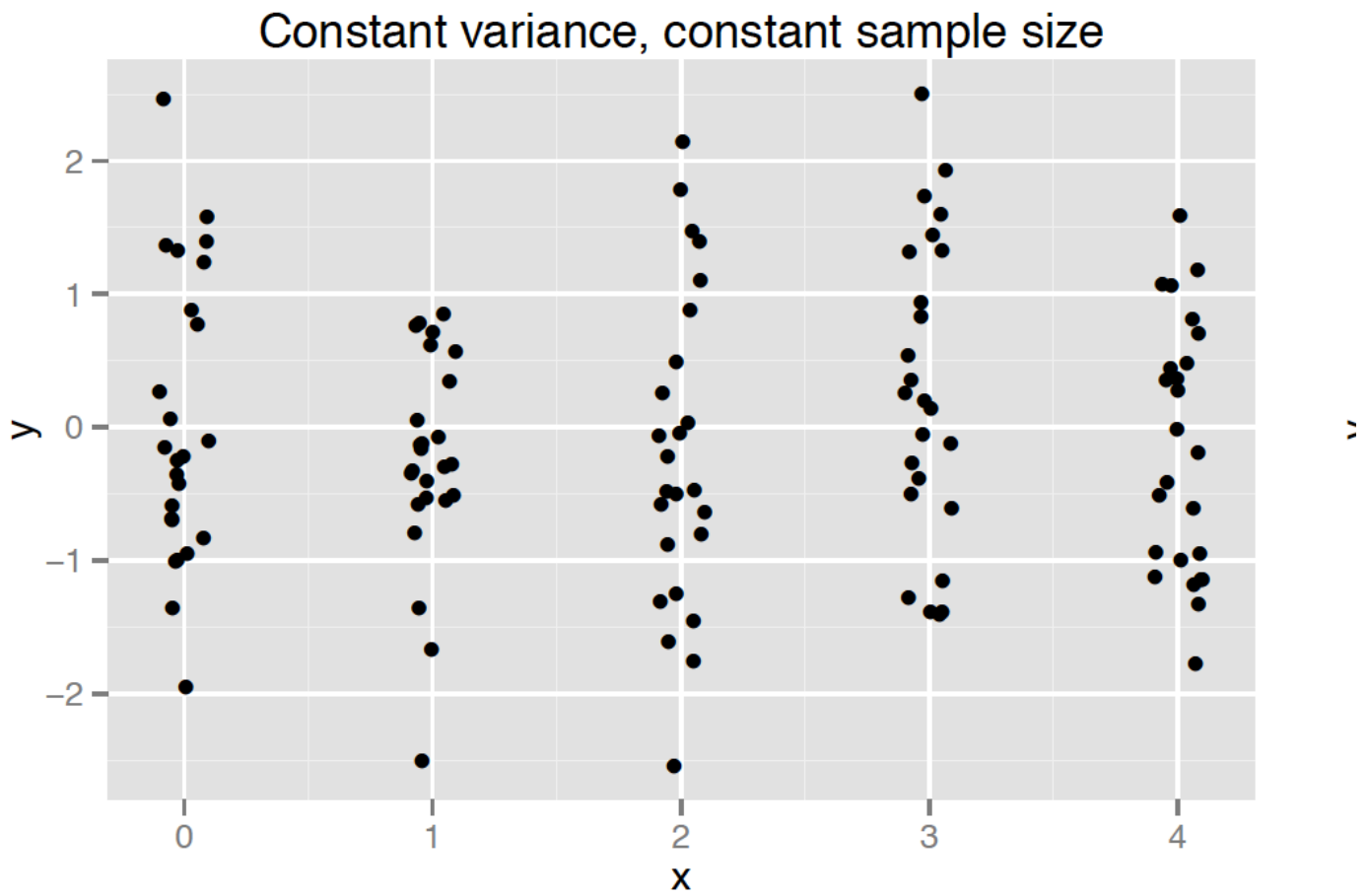
The point of doing this is that if an observation is outlier, it might have a large influence on the regression line, making its residual not as extreme as if the regression was fit without the line. The Studentized deleted residuals give a way of seeing which observations have the biggest impact on the regression line. If the model assumptions are correct (without extreme outliers), then the Studentized deleted residual has a t distribution with $n - 2$ degrees of freedom.

Regression

Something to be careful of is that if you have different numbers of observations for different values of x, then larger sample sizes will naturally have a larger range. Visually, this can be difficult to distinguish from nonconstant variance. The following examples are simulated.

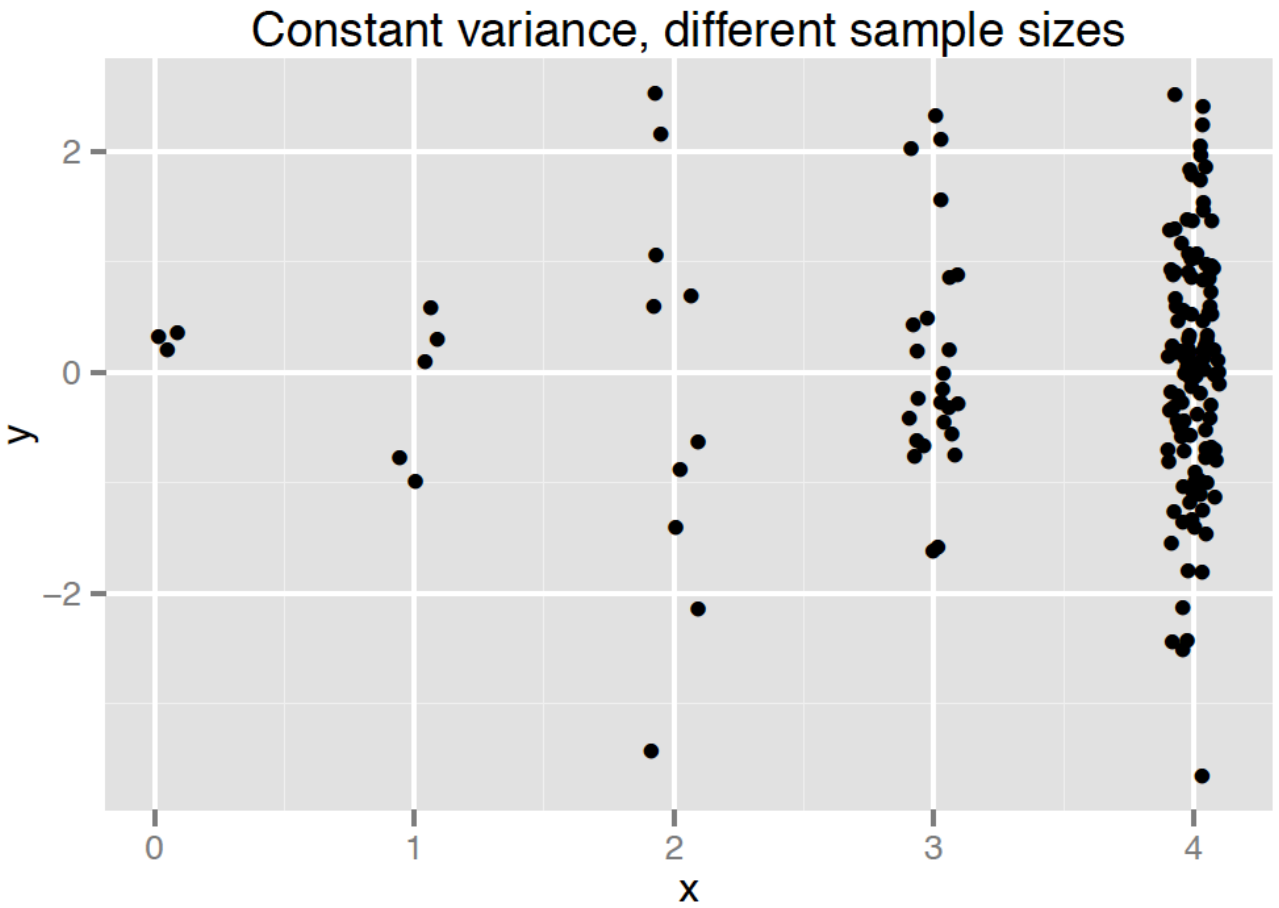
Regression

Residual plot with extreme outlier.



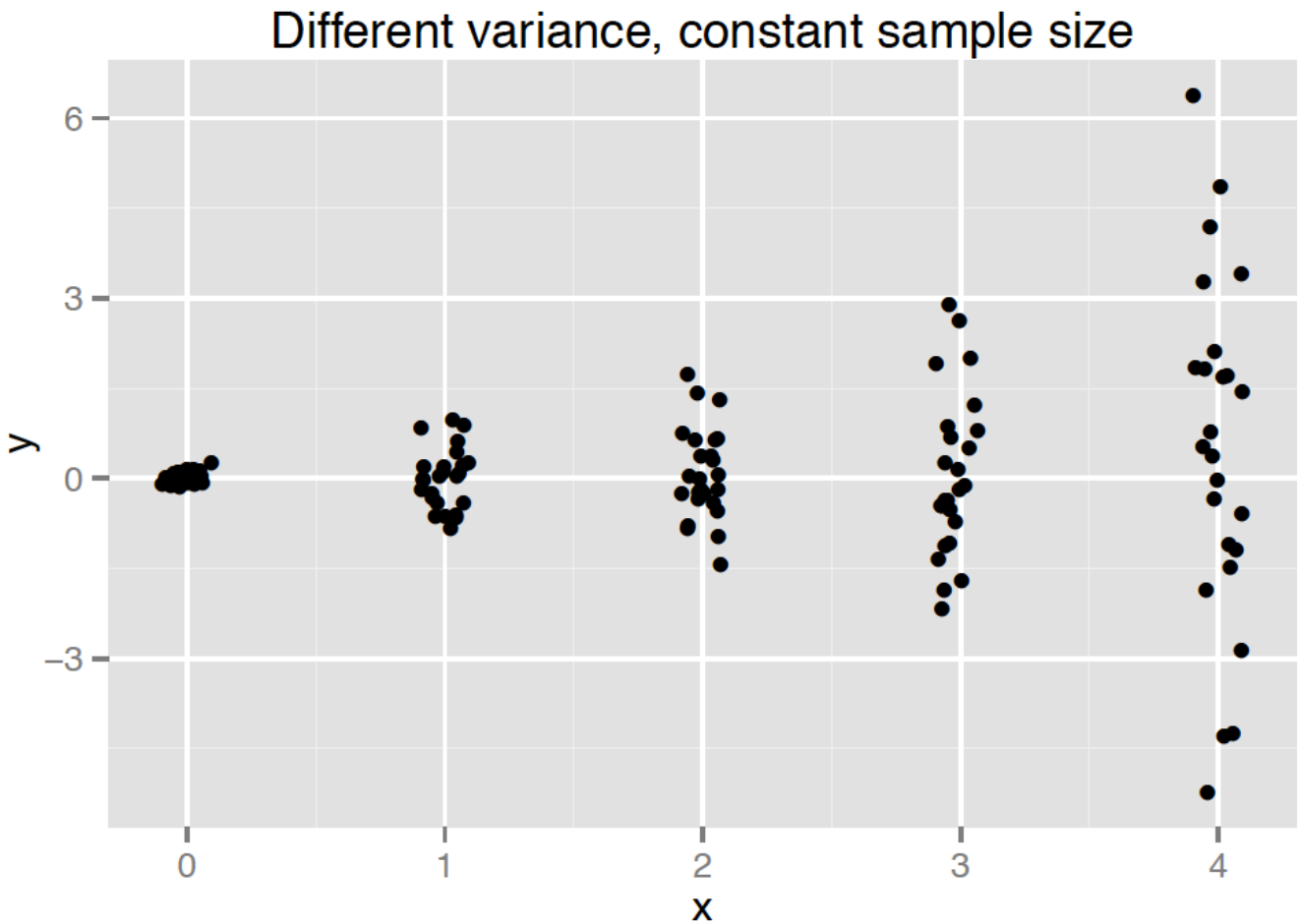
Regression

Residual plot with extreme outlier.



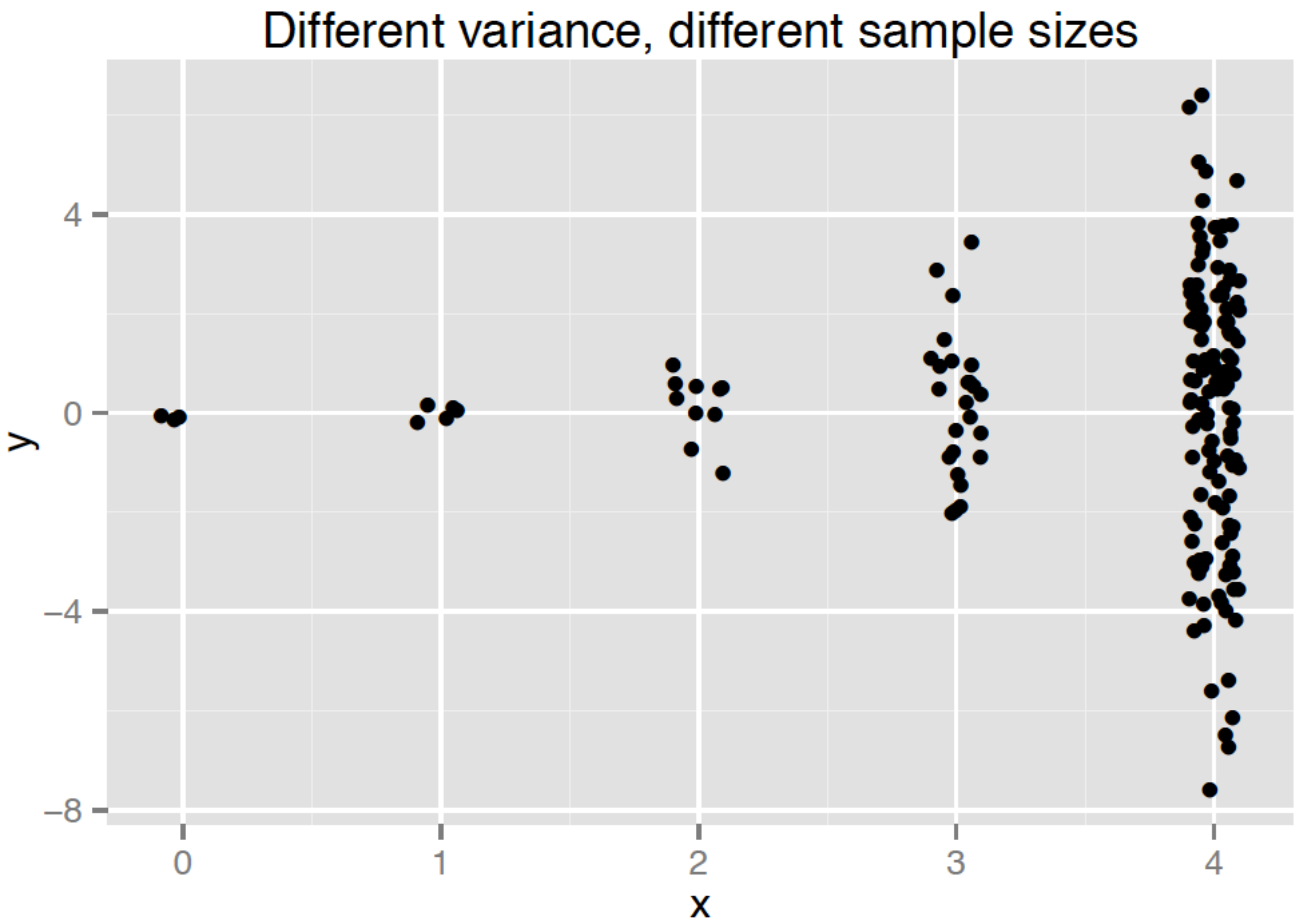
Regression

Residual plot with extreme outlier.



Regression

Residual plot with extreme outlier.



Regression

The normality assumption for regression is that the responses are normally distributed for each level of the predictor. Note that it is not assumed that the predictor follows any particular distribution. The predictor can be nonnormal, and can be chosen by the investigator in the case of experimental data. In medical data, the investigator might recruit individuals based on their predictors (for example, to get a certain age group), and then think of the response as random.

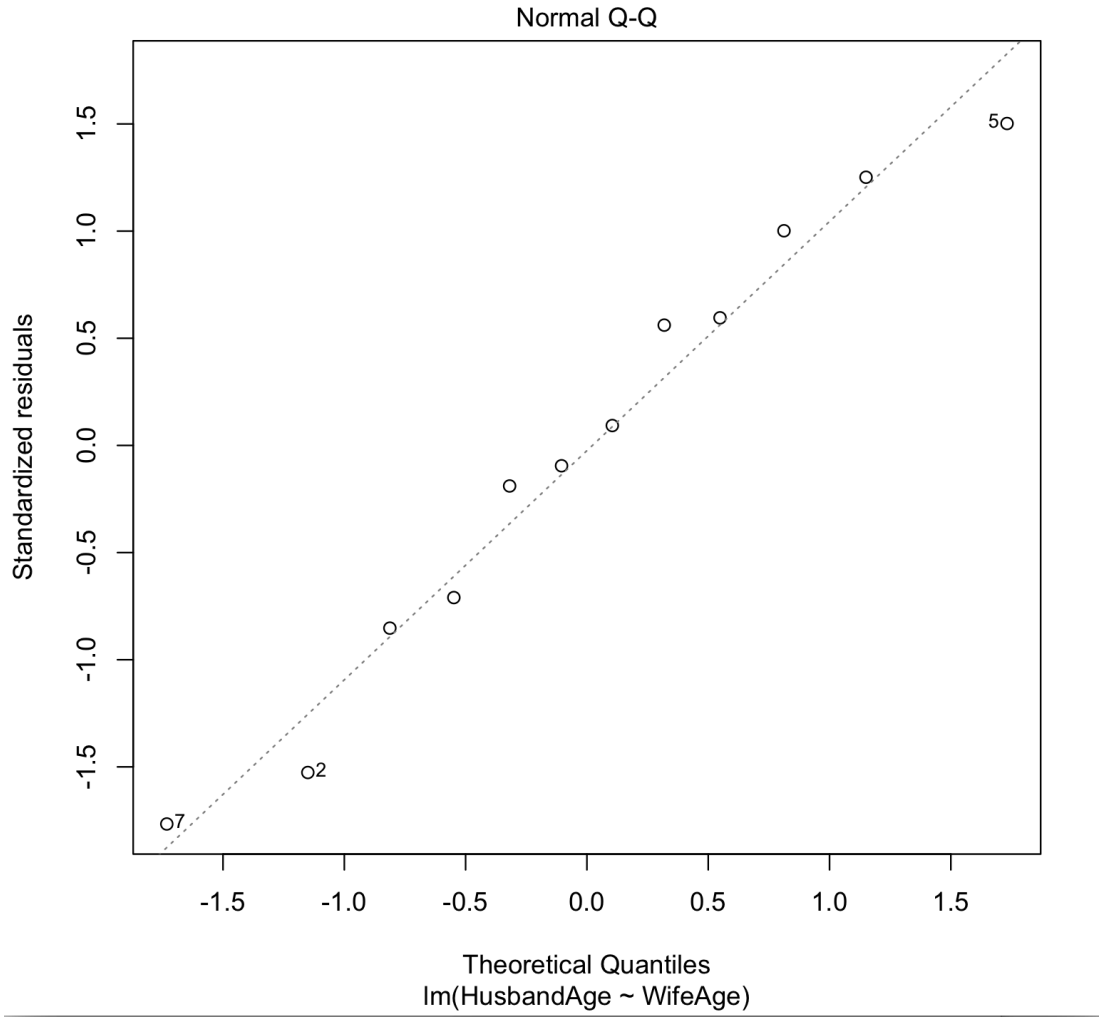
It is difficult to tell if the responses are really normal for each level of the predictor, especially if there is only one response for each x value (which happens frequently). However, the model also predicts that the residuals are normally distributed with mean 0 and constant variance. The indvidual values y_i come from different distributions (because they have different means), but the residuals all come from the same distribution according to the model. Consequently, you can check for normality of the residuals.

Regression

As an example, the QQ plot is also generated by typing `plot(model1)` and hitting return several times. The command `plot(model1)` will generate four plots. To see them all you might type `par(mfrow=c(2,2))` to put them in a 2×2 array first. You can also do a formal test on the residuals

```
par(mfrow=c(2,2))
plot(model1)
shapiro.test(model1$residual)
# Shapiro-Wilk normality test
#
#data:  model1$residual
#W = 0.95602, p-value = 0.7258
```

QQ plot


Normal Q-Q
Standardized residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
5
2
7
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(HusbandAge ~ WifeAge)

Regression

Studentized deleted residuals, can be obtained from R using `rstudent(model1)`. It helps to sort them to see which ones are most extreme. Note that studentized residuals can be obtained by `rstandard(model1)`.

```
rstudent(model1)
sort(rstudent(model1))
#         7           2          10           4          11          12
#-2.02038876 -1.65316857 -0.83992238 -0.69122350 -0.17987025 -0.09016353
#         6           9           8           1           3           5
# 0.08799631  0.54099532  0.57506585  1.00173208  1.29257824  1.61935539
```

Regression

The biggest outlier based on the resiudals has a z-score of -2.02, which is not very extreme for 12 observations from a normal distribution. This corresponds to observation 7, which is

```
> x[7,]
  Couple HusbandAge HusbandHeight WifeAge WifeHeight
7      7         43          1730      52       1610
```

The negative number is due to the husband being younger than expected given the wife's age.

Regression

If there are outliers in the data, you have to be careful about how to analyze them. If an outlier is due to an incorrect measurement or error in the data entry, then it makes sense to remove it. For a data entry error, you might be able to correct the entry by consulting with the investigators, for example if a decimal is put in the wrong place. This is preferable to simply removing the data altogether.

If an outlier corresponds to genuine data but is removed, I would typically analyze the data both with and without the outlier to see how much of a difference it makes. In some cases, keeping an outlier might not change conclusions in the model. Removing the outlier will tend to decrease variability in the data, which might make you underestimate variances and standard errors, and therefore incorrect conclude that you have signficance or greater evidence than you actually have.

Regression

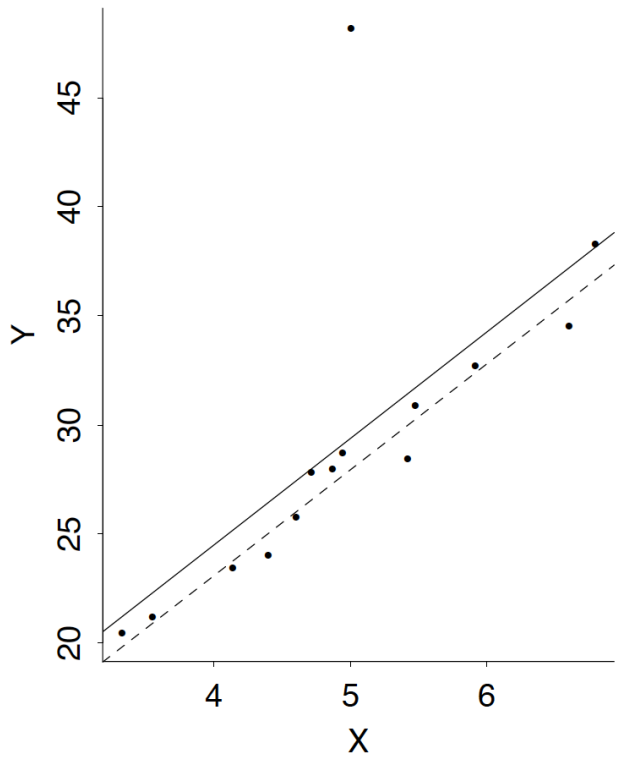
Removing an outlier that is a genuine observation also makes it less clear what population your sample represents. For the husband and wife example, if we removed a couple that appeared to be an outlier, we would be making inferences about the population of couples that do not have unusual ages or age combinations, rather than inferences about the more general population, which might include unusual ages.

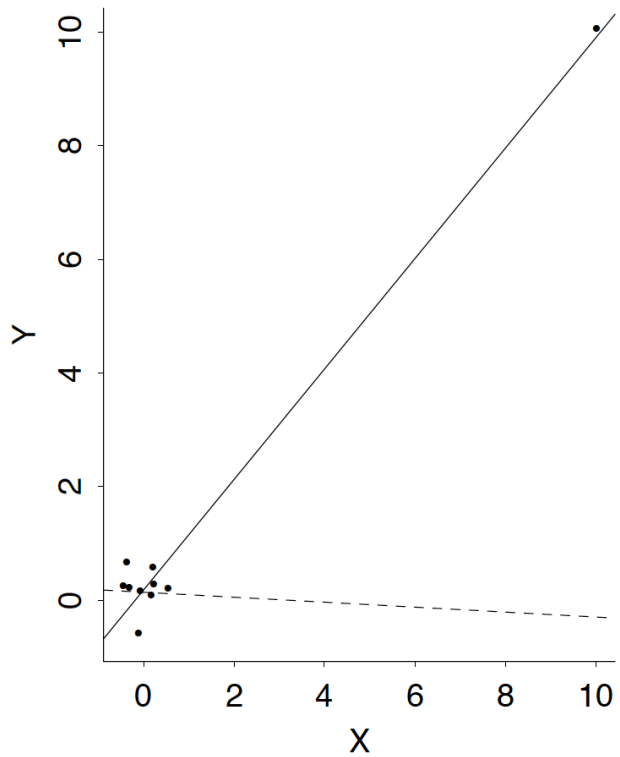
Regression

A concept from regression diagnostics is that of **influential observations**. These are observations that can have a large impact on the regression line if they are removed from the data. This is a slightly different concept from that of outliers. An influential observation might or might not be an outlier, and might or might not have a large residual.

In the next slide, the solid line is the regression line with the influential observation and the dotted line is the regression line with the influential observation removed.

Regression





Regression

Note that in the previous slide, the left plot has an observation with an unusal y value, but that the x value for this observation is not unusual. For the right plot, the outlier is unusual in both x and y values.

Typically, an unusual x value has more potential to greatly alter the regression line, so a measure called **influence** has been developed based on how unusual an observation is in the predictor variable(s), without taking into account the y variable.

Regression

To see measures of influence, you can type `influence(model1)`, where `model1` is whatever you saved your `lm()` call to. The leverage values themselves are obtained by `influence(model1)$hat`. Leverages are between 0 and 1, where values greater than about $2p/n$ or $3p/n$ (n is the number of observations in the data; p is the number of parameters including the intercept; in simple linear regression, $p = 2$), are considered large. If $3p/n$ is greater than 1, you can use 0.99 as a cutoff.

```
> influence(model1)$hat
> influence(model1)$hat
     1      2      3      4      5      6      7      8
0.1027 0.1439 0.1212 0.3319 0.2203 0.1572 0.2203 0.1027
     9     10     11     12
0.2235 0.1877 0.0847 0.1039
```

Regression

For the husband and wife age data, observation 4 has the highest leverage, and this corresponds to the couple with the highest age for the wife. Recall that for leverage, the y variable (husband's age) is not used. However, the value here is 0.33, which is not high for leverages. Recall that the observation with the greatest residual was observation 7.

```
> x[4,]
  Couple HusbandAge HusbandHeight WifeAge WifeHeight
4      4         52          1779      57       1540
```

Regression

The formula for the leverage is somewhat complicated (it is usually defined in terms of matrices), but to give some intuition, note that the z-scores for the wife's ages also give observation 4 as the most unusual, with a z-score of 1.65:

```
> (x$WifeAge-mean(x$WifeAge))/sd(x$WifeAge)
 [1]  0.461 -0.816 -0.646  1.653  1.228 -0.901  1.228
0.461 -1.242 -1.072 0.121 -0.475
```

Regression

Another measure of influence is **Cook's distance** or Cook's D. An expression for Cook's D is

$$D_j \propto \sum_i (\widehat{y}_i - \widehat{y}_{i[-j]})^2$$

Here $\widehat{y}_{i[-j]}$ is the predicted value of the ith observation when the regression line is computed with the jth observation removed from the data set. This statistic is based on the idea of recomputing the regression line n times, each time removing observation j, $j = 1, \ldots, n$ to get a statistic for how much removing the jth observation changes the regression line for the remaining.

The symbol $\propto$ means that the actual value is a multiple of the value that doesn't depend on j. There are different interpretations for what counts as a large value of D_j. One is values of $D_j > 1$. Another approach is to see if D_j is large for some j compared to other Cook's distances in the data.

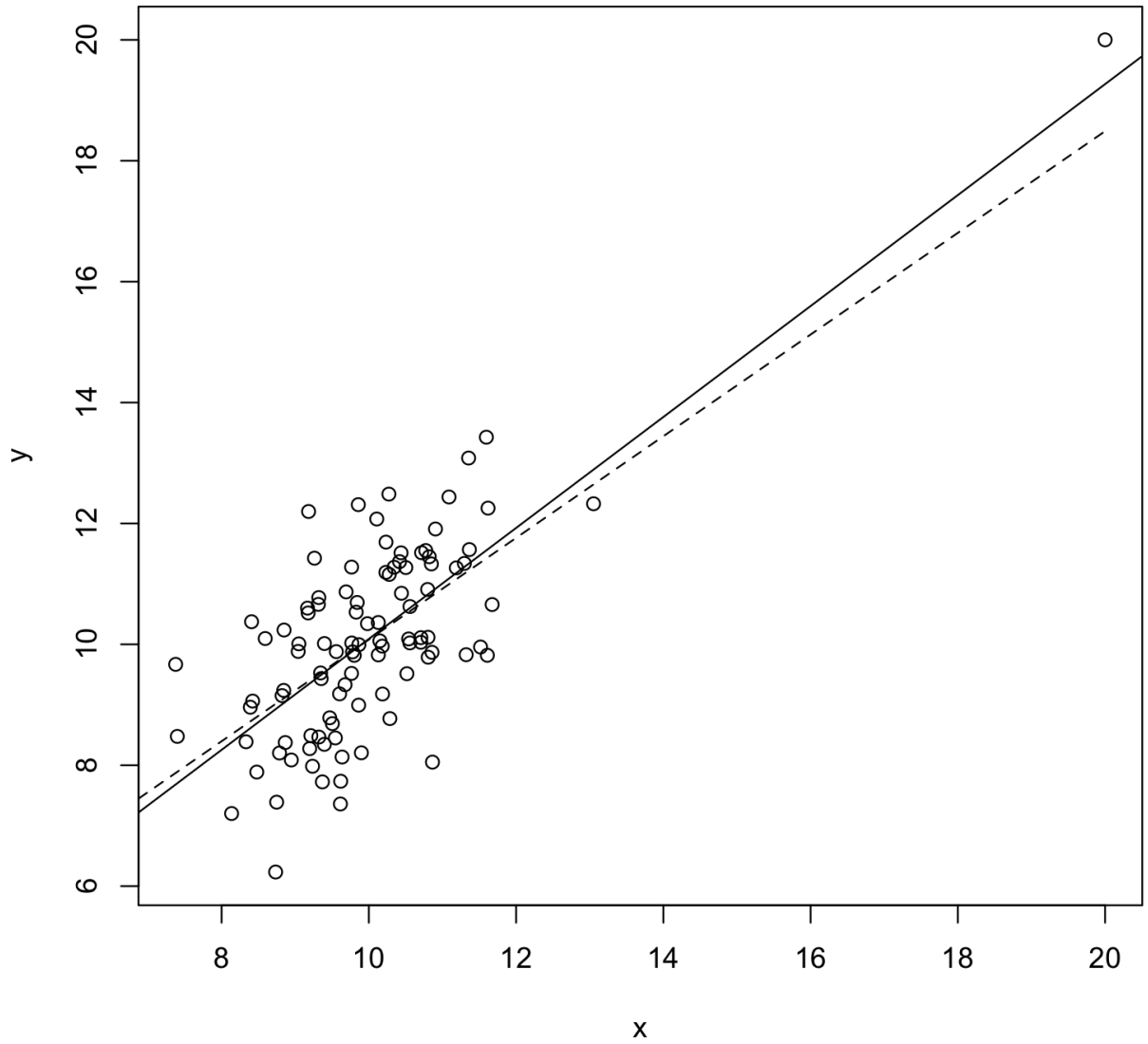
Regression

```
> cooks.distance(model1)
       1        2        3        4        5        6
0.057390 0.195728 0.108014 0.125200 0.318839 0.000802

       7        8        9       10       11       12
0.440936 0.020277 0.045336 0.083990 0.001656 0.000523
```

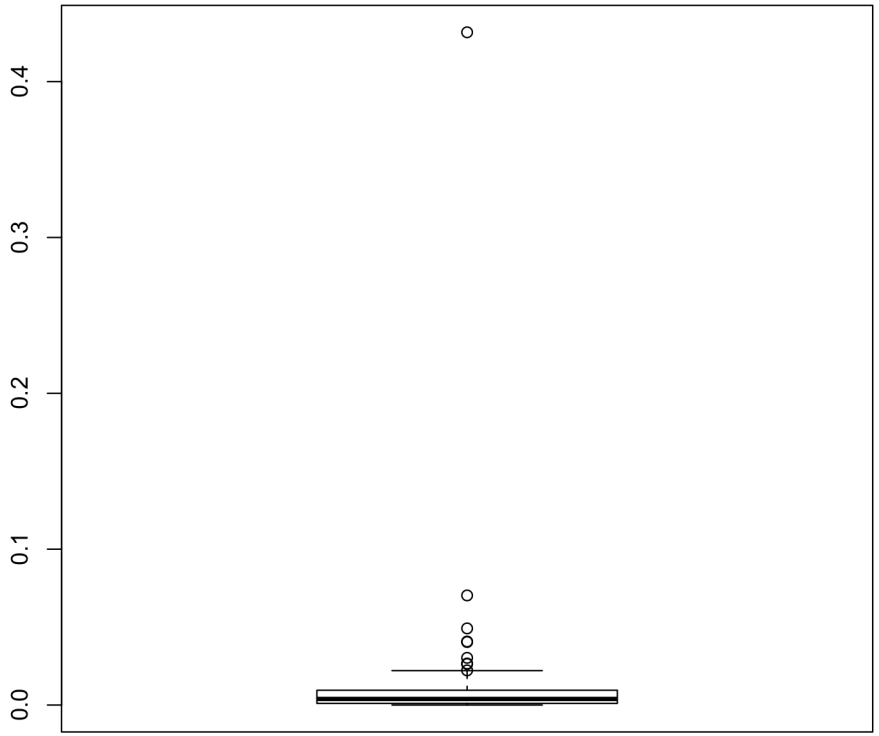
Regression

Cook's distance depend's on an observation being unusual for both x and y and having an influence on the regression line. In this case, observation 7 has the highest Cook's D, but it is not alarming. In the following example, an artificial data set with 101 observations has an outlier that is fairly consistent with the overall trend of the data with the outlier removed. However, Cook's D still picks out the outlier. As before, the solid line is with the outlier included.


y
x
6
8
10
12
14
16
18
20
8
10
12
14
16
18
20

Regression

```
a <- lm(y ~ x)
hist(cooks.distance(a))
```



Regression

Note that with one predictor we can pretty easily plot the data and visually see unusual observations. In multiple dimensions, with multiple predictor variables, this becomes more difficult, which makes these diagnostic techniques more valuable when there are multiple predictors. This will be explored more next semester.

The following is a summary for analyzing regression data.

Regression

1. **Plot the data**. With multiple predictors, a scatterplot matrix is the easiest way to do this.
2. Consider transformations of the data, such as logarithms. For count data, square roots are often used. Different transformations will be explored more next semester.
3. Fit the model, for example using `aov()` or `lm()`
4. Examine residual plots. Here you can look for
   - curvature in the residuals or other nonrandom patterns
   - nonconstant variance
   - outliers
   - normality of residuals

Regression

5. Check Cook's D values
6. If there are outliers or influential observations, only remove them if they are recording error etc. You may consider redoing the analysis with problematic points removed to compare the results though.