Chapter 11
Linear Regression and Correlation

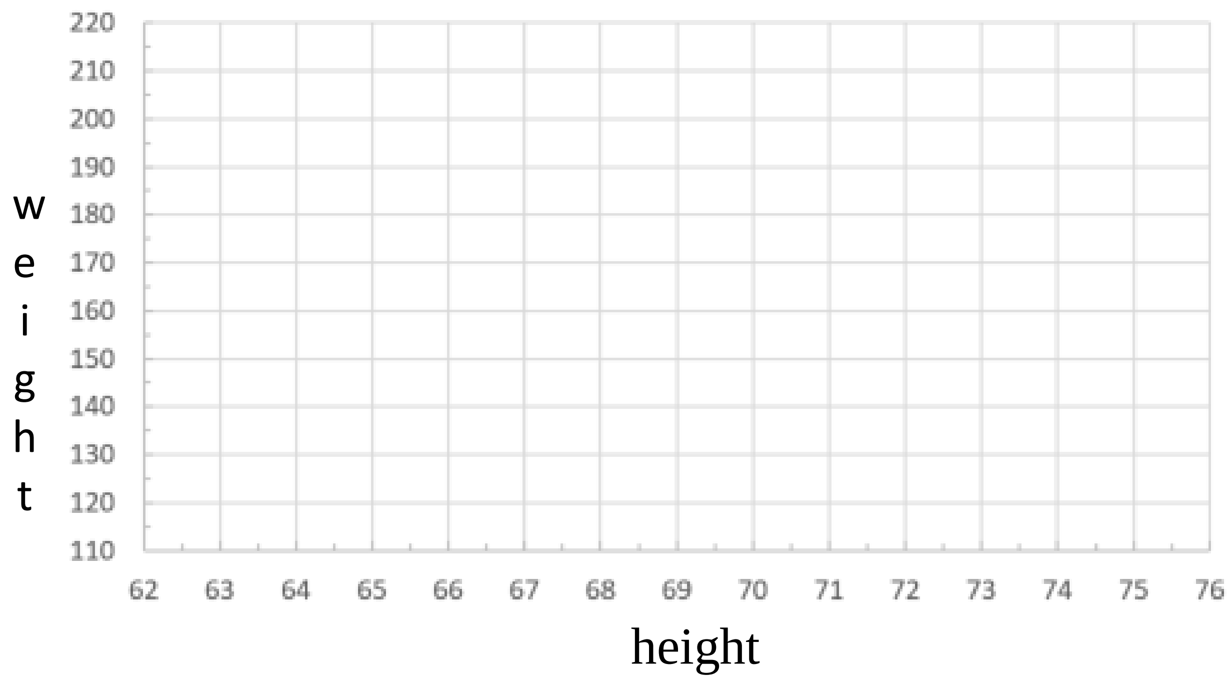
STAT 441/541 Statistical Methods II

Sections Covered in Chapter 11

- Section 11.1 Introduction and Abstract of Research Study
- Section 11.2 Estimating Model Parameters
- Section 11.3 Inferences About Regression Parameters
- Section 11.4 Predicting New y-values Using Regression
- Section 11.6 Correlation

Example: Predict weight using height

	A	B
1	height	weight
2	63	127
3	64	121
4	66	142
5	69	157
6	69	162
7	71	156
8	71	169
9	72	165
10	73	181
11	75	208



For each observation, plot height using the horizontal axis and the corresponding weight using the vertical axis. This is called a scatterplot.

What is simple linear regression?

There is a single independent variable x and the equation for predicting a dependent variable y is a linear function of x.

What is the simple linear regression model?

The model is

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

y is the dependent variable

x is the independent variable

β_0 is the true y-intercept (the value of the line when $x = 0$)

β_1 is the true slope of the line (the predicted change in y corresponding to a one-unit increase in x)

ε is random error

What is the prediction equation for simple linear regression?

The equation is

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

$\hat{y}$ are predicted values of the dependent variable

x are values of the independent variable

$\hat{\beta}_0$ is the estimated y-intercept

$\hat{\beta}_1$ is the estimated slope of the line

What are the four formal assumptions for simple linear regression analysis?

1. The model has been properly specified
2. The errors have the same variance, that is, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$ for all i
3. The errors are independent of each other
4. The errors are all normally distributed, that is, ε_i is normally distributed for all i

In statistical notation: $\varepsilon_i \sim Normal\left(0, \sigma_\varepsilon^2\right)$

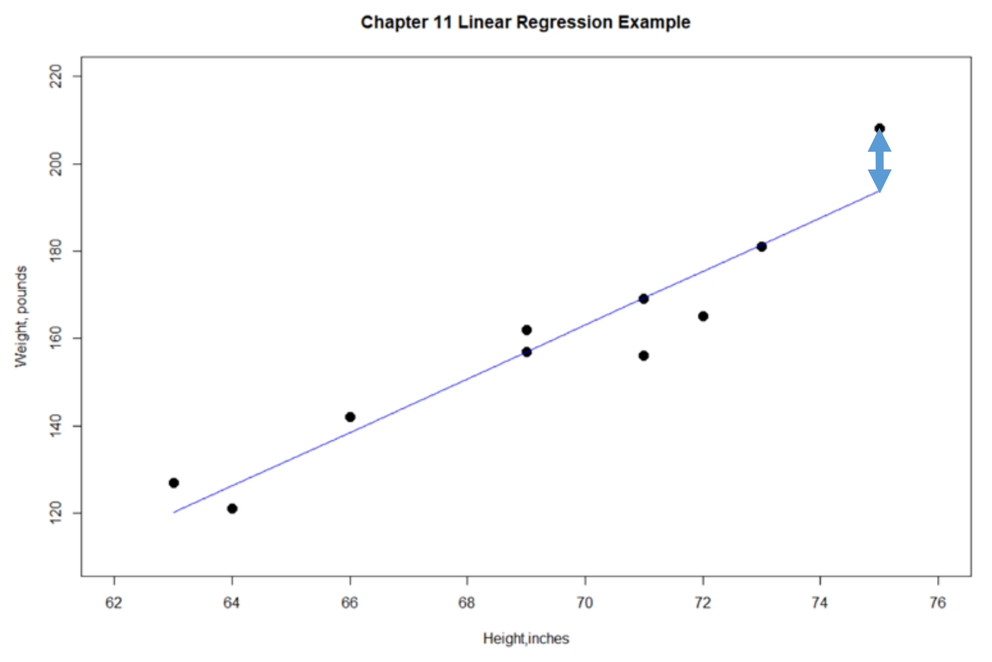
Transformations

- For simple linear regression, if the relationship between x and y is not linear, then it can often be "straightened out" by transforming the independent variable, dependent variable, or both
- The text provides several graphs and "Steps for choosing a transformation"
- The regression analysis is then performed on the transformed variables(s) as long as the assumptions are met

What is a residual?

- A residual is defined as an observed value of y minus its predicted value, that is, $y - \hat{y}$
- Residuals measure how far each observed value is from the regression line (parallel to y-axis)
- Residuals are used to estimate the common variance σ_{ε}^{2}

The residual for Height=75 is 208 – 193.8 = 14.2



How do we estimate the true error variance σ_ε^2?

The estimated variance around the line is

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

(Note the similarity with a sample standard deviation)

Observations with high leverage and influence

- An observation that affects the estimate of the regression slope is classified as high leverage or high influence
- What is a high leverage point?
    - An observations that has a very high or very low value of the independent variable (outliers in the x direction)
- What is a high influence point?
    - An observation that is a high leverage point and also has a very high or very low value of the dependent variable (outliers in the y direction)

What three terms in the simple regression model are estimated based on limited data?

Recall the simple linear regression model:

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

- The slope, B_1
- The intercept, B_0
- The variance of the random errors, σ_ε^2

What two concepts apply to regression summary figures?

1. Hypothesis tests
2. Confidence intervals

Both use the t distribution

Summary of a statistical test for β_1

- Hypotheses
  - Case 1: H_0: $\beta_1 \le 0$ versus H_a: $\beta_1 > 0$
  - Case 2: H_0: $\beta_1 \ge 0$ versus H_a: $\beta_1 < 0$
  - Case 3: H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$
- Test Statistic: t value from software output
- Compare *p*-value from output to significance level α
  - Reject the null hypothesis H_0 if *p*-value $\le \alpha$
    (If *p*-value is low, H_0 must go)
  - Fail to reject the null hypothesis H_0 if *p*-value > α
    (If *p*-value is high, with H_0 we must comply)
- Check assumptions and draw conclusions

Most common test of the true slope parameter (a t test)

The most common test is

H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$

This tests whether the independent variable x should be in the model.

Example: height and weight data

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -266.5344    51.0320  -5.223    8e-04 ***
height         6.1376     0.7353   8.347 3.21e-05 ***
```

Since p-value = 0.0000321 < 0.05, we reject the null hypothesis and conclude $\beta_1 \neq 0$. There is a significant relationship between height and weight. For a one inch increase in height, the average weight increases by 6.1 pounds.

F Test for Predictive Value of a Regression Model

This tests the null hypothesis that all independent variables have no value in predicting y (more useful for multiple regression; for simple linear regression this is same as the t test for β_1)

Example: height and weight data

```
Residual standard error: 8.641 on 8 degrees of freedom
Multiple R-squared:  0.897,     Adjusted R-squared:  0.8841
F-statistic: 69.67 on 1 and 8 DF,  p-value: 3.214e-05
```

The F test for using height to predict weight. The p-value is 0.00003214 so the null hypothesis that the model has no predictive value is rejected at significance level $\alpha = 0.05$. We conclude that height has value in predicting weight.

Is it useful to interpret the intercept β_0?

The intercept term in the model, β_0, is the value of y when $x = 0$.

The intercept is interpretable only when $x = 0$ is meaningful.

Example: For the height and weight data,

$$\hat{\beta}_0 = -266.5344$$

This is not interpretable since a height of 0 inches is not meaningful.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -266.5344    51.0320  -5.223    8e-04 ***
height         6.1376     0.7353   8.347 3.21e-05 ***
```

What are two interpretations of a y prediction for a given x value?

- The average response value [$E(y)$] of the population of all possible values for a specific x value
- The response value y_{n+1} for a specific x value

In the road-resurfacing example in our text, the county highway director wants to predict the cost of a new contract for $x = 6$ miles that is up for bids.

The average cost $E(y)$ of *all resurfacing* contracts for 6 miles of road will be $20,000.

The cost y of *this specific* resurfacing contract for 6 miles of road will be $20,000.

Confidence Interval for mean response $E(y_{n+1})$

We will use software to compute a confidence interval on the mean response for a specified value of the independent variable

Example: Develop a 95% confidence interval of the mean weight for height = 68 inches

First, create a new data frame that sets the height value,

Second, use the predict function and set the interval type as "confidence" using the default 0.95 confidence level

Third, interpret the interval: We are 95% confident that the interval from 144.1452 to 157.4971 captures the mean weight in pounds for a height of 68 inches

```
> # confidence interval for mean weight given height=68 inches
> newdatamu <- data.frame(height=68)
> predict(model,newdatamu,interval="confidence")
       fit      lwr      upr
1 150.8211 144.1452 157.4971
```

Prediction Interval for y_{n+1}

- We will use software to compute a prediciton interval on the response value for a specified value of the independent variable

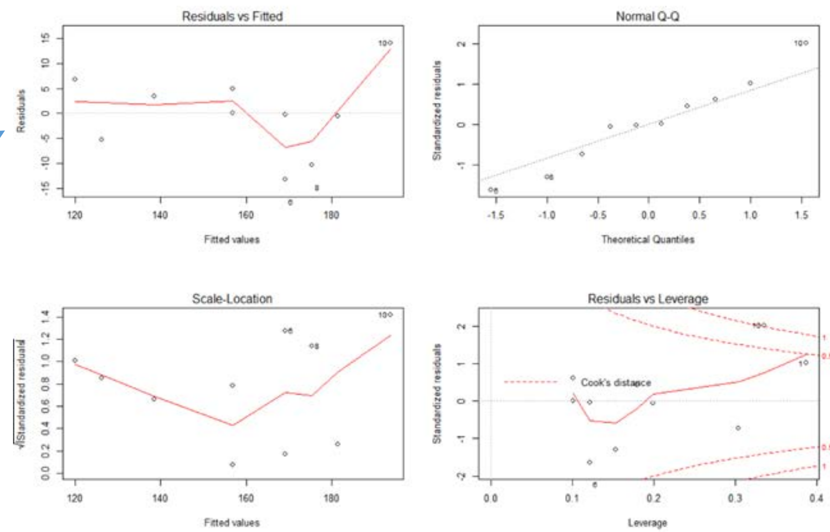
Example: Develop a 95% prediction interval of the weight in pounds for height = 68 inches
First, create a new data frame that sets the height value,
Second, use the predict function and set the interval type as “predict” using the default 0.95 confidence level
Third, interpret the interval: We are 95% confident that the interval from 129.8056 to 171.8367 captures the individual weight in pounds for a height of 68 inches

```
> # prediction interval of weight for height=68 inches
> newdatay <- data.frame(height=68)
> predict(model,newdatay,interval="predict")
       fit      lwr      upr
1 150.8211 129.8056 171.8367
```

What are residual plots?

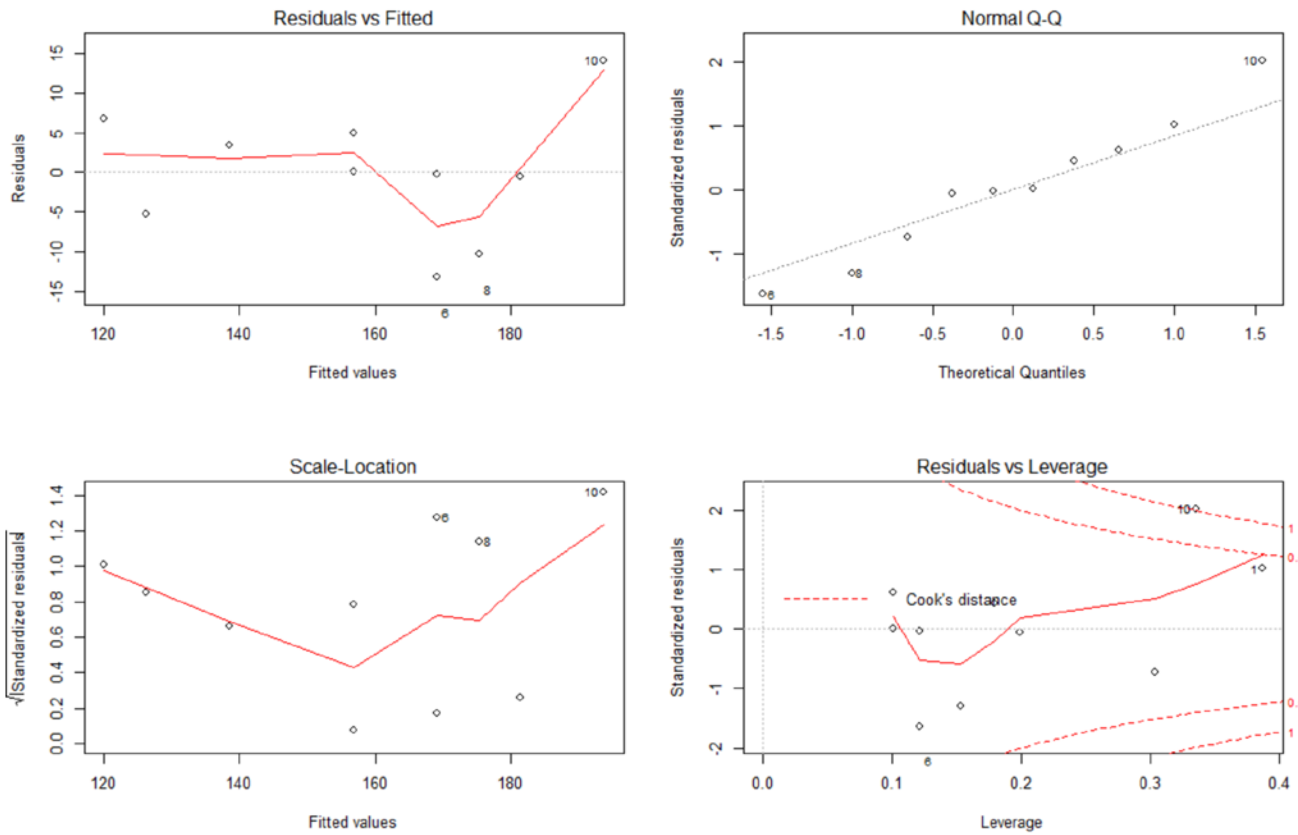
- A residual plot has residuals on the vertical axis and predicted values on the horizontal axis. Look for:
- Outliers or erroneous observations. Look for data points with unusually high (in absolute value) residuals.
- Violation of assumptions. Look for non-random patterns in the residuals.

R output labels this plot as “Residuals vs Fitted”



What are two ways to check the constant variance assumption?

(1) The residual plot and (2) the Scale-Location plot.



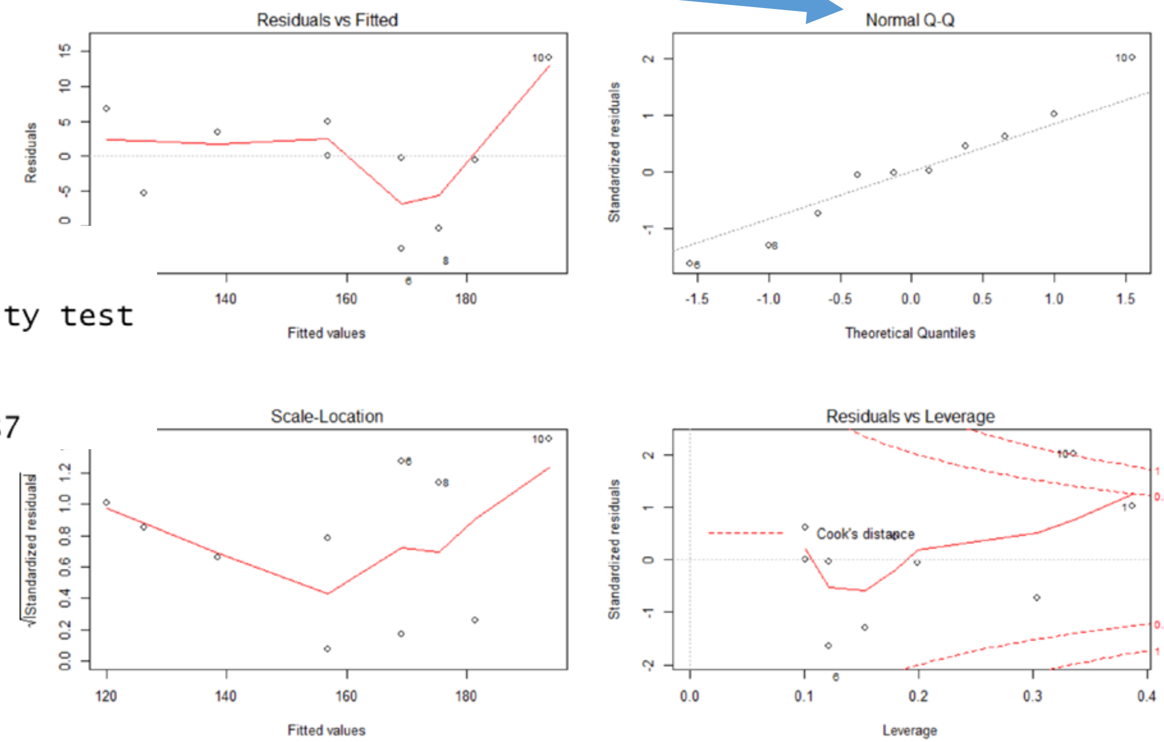
How can we check for normality of errors?

(1) The Normal Q-Q plot and (2) the Shapiro-Wilk test for normality

```
> shapiro.test(resid(model))

        Shapiro-Wilk normality test

data:  resid(model)
W = 0.97445, p-value = 0.9287
```



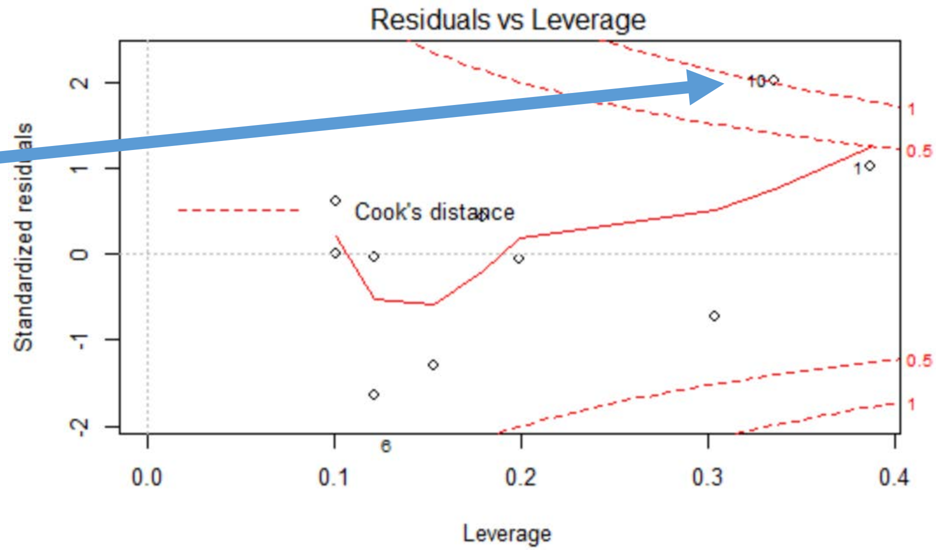
Cook's Distance for Potential Outliers

- Cook's distance is a measure of influence that considers the effect of a single observation on the model as a whole is. It is a measure of the overall influence of an observation on the model and values greater than one (1) may be cause for concern.
- Cook's distance, sometimes denoted by D_i, depends on both the residual and the leverage. That is, both the *x* value and the *y* value of the data point play a role in the calculation of Cook's distance.
- In short:
  - D_i directly summarizes how much *all* of the fitted values change when the i^{th} observation is deleted.
  - A data point having a large D_i indicates that the data point strongly influences the fitted values.

Using Cook's Distance

- We must rely on guidelines for deciding when a Cook's distance measure is large enough to warrant treating a data point as influential.
- The guidelines commonly used are:
- If D_i is greater than 0.5, then the i^{th} data point is worthy of further investigation as it **may be influential**.
- If D_i is greater than 1, then the i^{th} data point is **quite likely to be influential**.
- Or, if D_i sticks out like a sore thumb from the other D_i values, it is **almost certainly influential**.

We see that Observation 10 is a high influence point



Lack of Fit (LOF) for the Simple Linear Regression model

- What is LOF? To test if $y = \beta_0 + \beta_1 x + \varepsilon$

is an appropriate model.

- When can we test for LOF? When there is more than one observation per level of the independent variable.
- What are the two parts of SS(Error)?
  - SSP_{exp} is the sum of squares Pure Experimental Error pooled over each level of the independent variable
  - $SS(Error) = SSP_{exp} + SS_{Lack}$
  - $SS_{Lack} = SS(Error) - SSP_{exp}$

A Test for Lack of Fit in Linear Regression

- H_0: A linear regression model is appropriate
- H_a: A linear regression model is not appropriate
- Test Statistic T.S.: $F = \frac{MS_{Lack}}{MSP_{exp}}$
- Conclusion: If the F test is significant at a specified alpha (e.g. $\alpha = 0.01$), then the linear regression model is inadequate. A nonsignificant result indicates that there is insufficient evidence to suggest that the linear regression model is inappropriate.
- Results may be summarized in an Analysis of Variance Table for Regression Analysis

ANOVA for Simple Linear Regression

Source	df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	1	SS(Regression)	$\mathrm{MS(Regression)} = \frac{\mathrm{SS(Regression)}}{1}$	$F_{model} = \frac{\mathrm{MS(Regression)}}{\mathrm{MS(Residual)}}$	
Residual	n-2	SS(Residual)	$MS(Residual) = \frac{\mathrm{SS(Residual)}}{n-2}$		
Lack of Fit	n-2-$\sum_i (n_i - 1)$	$SS_{Lack} = SS(Residual) - SSP_{exp}$	$MS_{Lack} = \frac{SS_{Lack}}{n-2-\sum_i(n_i-1)}$	$F_{LOF} = \frac{MS_{Lack}}{MSP_{exp}}$	
Pure Experimental Error	$\sum_i (n_i - 1)$	SSP_{exp}	$MSP_{exp} = \frac{SSP_{exp}}{\sum_i(n_i-1)}$		
Total (Corrected)	n-1	SS(Total)			

See Section 11.5 Examining Lack of Fit in Linear Regression and Handout Example 11-1 Lack of Fit Using RStudio

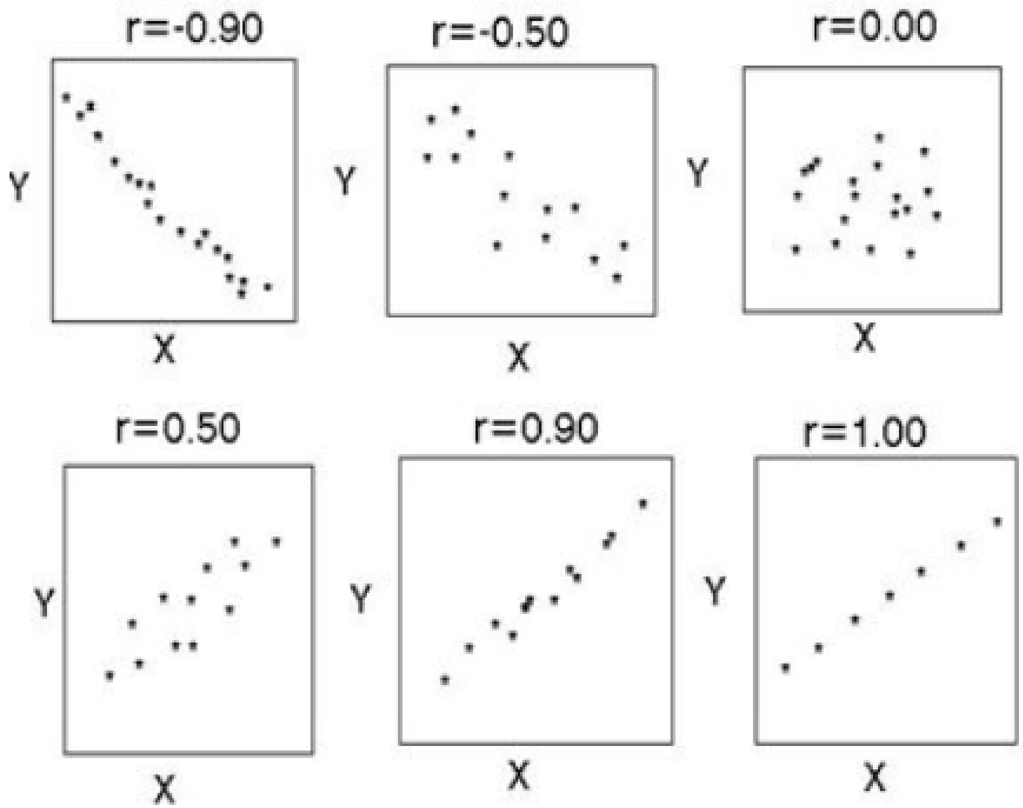
Scatterplot

- A plot of each observation or (x,y) point
- The independent variable, x, is along the horizontal axis
- The dependent variable, y, is along the vertical axis
- Examine the pattern of points to determine if they basically follow a straight line or if there is a definite curve
- Also look to see whether there are any potential outliers falling far from the general pattern of the data

What is correlation?

- Definition: Correlation measures the strength of the linear relation between x and y
- The formula for the sample correlation coefficient, r, can be found in our text. We will use software to compute sample correlation coefficients.
- Correlation coefficients range from -1 to +1
- Values over zero indicate positive correlation and values below zero indicate a negative correlation

Examples of various sample correlation coefficients

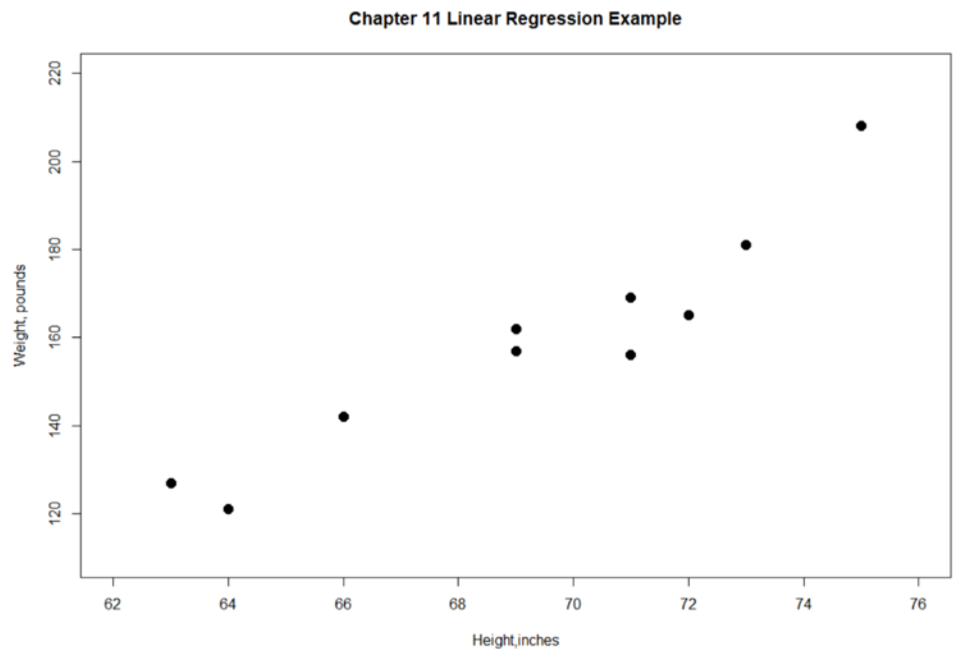


https://www.simplypsychology.org/correlation.html

Note: Figure 11.21 in our text displays 15 scatterplots for various values of the sample correlation coefficient

Example: height and weight

First, examine a scatterplot to verify there is a linear relationship



Second, use software to compute the sample correlation coefficient

```
> cor(dataobj$height, dataobj$weight, method="pearson")
[1] 0.9470984
```

$$r_{yx} = 0.947$$

Interpretation: There is a very strong linear relationship between weight and height.

A statistical test for population correlation coefficient ρ_{yx}

- Hypotheses
  - Case 1: $H_0: \rho_{yx} \leq 0$ versus $H_a: \rho_{yx} > 0$
  - Case 2: $H_0: \rho_{yx} \geq 0$ versus $H_a: \rho_{yx} < 0$
  - Case 3: $H_0: \rho_{yx} = 0$ versus $H_a: \rho_{yx} \neq 0$
- Test Statistic: t value from software output
- Compare *p*-value from output to significance level α
  - Reject the null hypothesis H_0 if *p*-value $\leq \alpha$
    (If *p*-value is low, H_0 must go)
  - Fail to reject the null hypothesis H_0 if *p*-value > α
    (If *p*-value is high, with H_0 we must comply)
- Check assumptions and draw conclusions

Coefficient of Determination

Definition: The proportionate reduction in error for regression is called the coefficient of determination.

For simple linear regression, the coefficient of determination is the square of the correlation coefficient.

Example: height and weight

```
Residual standard error: 8.641 on 8 degrees of freedom
Multiple R-squared:  0.897,     Adjusted R-squared:  0.8841
F-statistic: 69.67 on 1 and 8 DF,  p-value: 3.214e-05
```

Interpretation: Using height in the regression model explains 89.7% of the variation in weight.