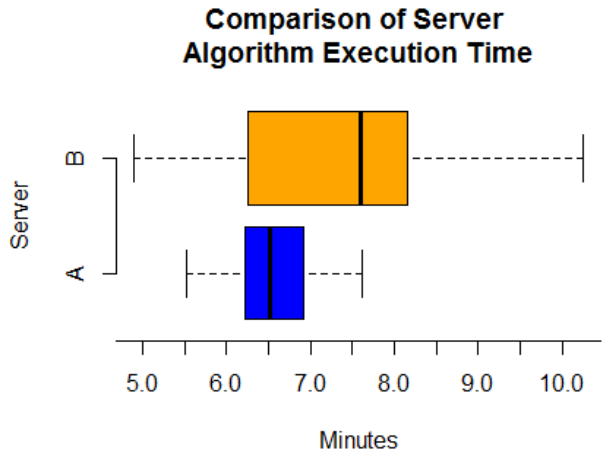
Week 7 Comparing multiple means

t Procedures for Two Independent Populations

Comparing Population Means

Suppose a business manager needs access to a server. Server A is supposedly faster than server B, but an account on A is more expensive. She would like to compare the speeds of the two servers to see if server A is worth the extra cost.



Two Sample *t* Confidence Interval

When to use:

want to estimate the difference in population means $\mu_1 - \mu_2$ from two independent population

Conditions for inference:

- representative samples from each of the population
- need sufficiently large sample sizes (same as the one sample t confidence interval)
- population need to be independent

The confidence interval for the difference in population means is:

Point estimate ± margin of error → Critical value x standard error estimate

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- μ_1 = pop mean from pop #1
- $\bar{x}_1$ = sample mean from sample #1
- s_1 = sample SD from sample #1
- n_1 = sample size from sample 1

Satterthwaite Approximate *t* Distribution

Satterthwaite degrees of freedom:

Used in an "unpooled" t procedure

Used when we do not want to assume the population standard deviations of the two populations of interest are equal (most of the time)

$$V = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n}\right)^2}{\left(\frac{s_1}{n_1}\right)^2/n_1 - 1 + \left(\frac{s_2^2}{n_2}\right)^2/n_2 - 1}$$

week 7 module has tool for calculating this

we are not using the conservative degrees

Conservative degrees of freedom:

Satterthwhaite can be tedious by hand. Sometimes "conservative" degrees of freedom is used.

Confidence Interval Example

Suppose a business manager needs access to a server. To compare the speed in minutes between two servers A and B, a computer algorithm is executed 30 times on server A and 30 times on server B.

week 7 tool

Random samples from Server A and B

	$\bar{x}$	s	n
Server A	6.5	0.5	30
Server B	7.3	1.3	30

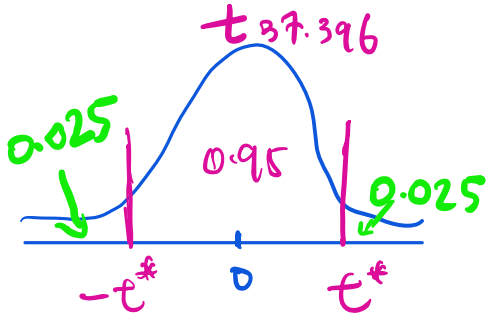
$\nu = 37.396$

Calculate a 95% confidence interval for $\mu_A - \mu_B$.

$$(\bar{X}_A - \bar{X}_B) \pm t^*_\nu \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$$

$$\Rightarrow (6.5 - 7.3) \pm \boxed{2.205} \sqrt{\frac{0.5^2}{30} + \frac{1.3^2}{30}}$$

$$\Rightarrow (-1.35, -0.285)$$



qt(0.975, 37.396)
= 2.025

Two Sample *t* Test

When to use:

want to test the difference in population means $\mu_1 - \mu_2$, from 2 independent population

Conditions required for inference:

- representative samples from each of the populations
- need sufficiently large sample size (same as one sample t confidence interval)
- populations need to be independent

equivalent correct statements

Null & Alternative hypotheses: $H_0: \mu_1 = \mu_2$, $H_0: \mu_1 - \mu_2 = 0$

$H_A: \mu_1 < \mu_2$ OR $H_A: \mu_1 - \mu_2 < 0$ | $H_A: \mu_1 > \mu_2$ OR $H_A: \mu_1 - \mu_2 > 0$ | OR $H_A: \mu_1 \neq \mu_2$ OR $H_A: \mu_1 - \mu_2 \neq 0$

Test statistic:

$$\frac{\text{Point estimate} - \text{hypothesized value}}{\text{standard error estimate of the point estimate}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Distribution of the Two Sample Test Statistic

The two sample t test statistic is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The distribution of this statistic follows:

t distribution with v df

satterthwaite

Two Sample *t* Test Example

Suppose a business manager needs access to a server. Server A is supposedly faster than server B, but an account on A is more expensive. She would like to compare the speeds of the two servers to see if server A is worth the extra cost.

Random samples from Server A and B

	$\bar{x}$	s	n
Server A	6.5	0.5	30
Server B	7.3	1.3	30

1. Random samples
2. Both samples (n=30)
3. server A and B independent

Determine the hypotheses needed for this test and check the conditions.

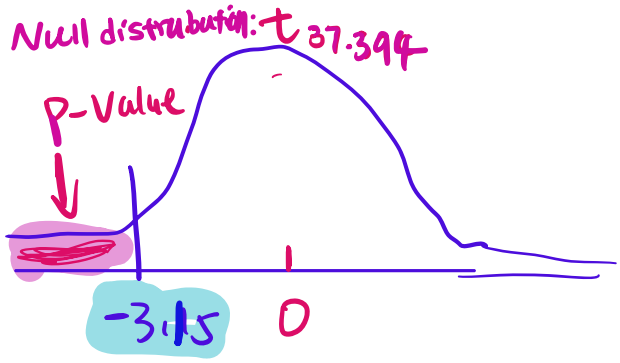
$H_0: \mu_A = \mu_B$ (or equivalently $\mu_A - \mu_B = 0$)

$H_A: \mu_A < \mu_B$ (" $\mu_A - \mu_B < 0$)

Using a significance level of 0.05, calculate the test statistic, the degrees of freedom, and p-value.

difference in sample means

$$t = \frac{6.5 - 7.3}{\sqrt{\frac{0.5^2}{30} + \frac{1.3^2}{30}}} = -3.15$$



P-Value R code:

pt(-3.15, 37.394)

= 0.0014

If there is no difference in the average run time of servers A and B, then the probability of observing a difference is sample means less than or equal to -0.8 minites is 0.14%.

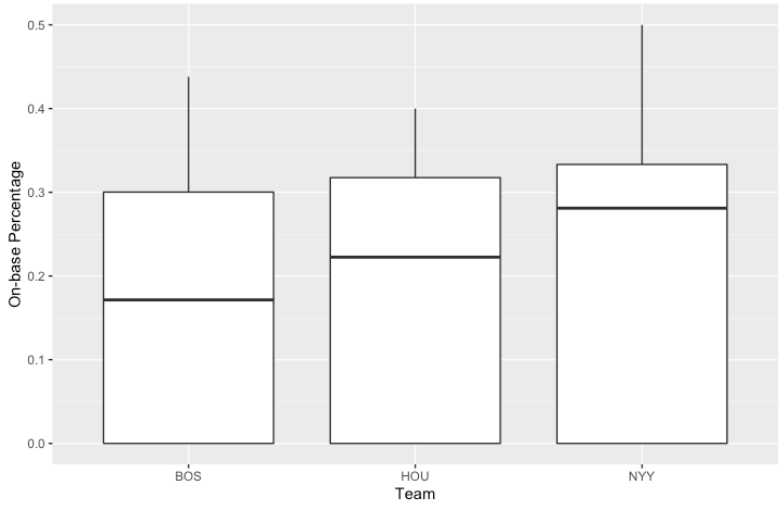
Analysis of Variance (ANOVA)

Comparing More than Two Means

In 2018, the Boston Red Sox, the Houston Astros, and the New York Yankees were ranked as the top three teams in the MLB.

The side-by-side boxplot displays the On-base Percentages for 30 randomly selected players from each of the 2018 top three teams.

According to the three graph, is there evidence that average On-base Percentages differ between the three teams?



Comparing Means with Mean Squares

- To compare more than two group means we compare the average between group variability to the average within group variability.
- Mean Squares represents the average variation between groups and average variation within groups.

Notation

Baseball Ex
K=3

- $k =$ # of groups (or Population of interest)
- $n =$ overall sample size
- $\bar{x} =$ overall sample mean (mean of all observations ignoring groups)

- n_i = sample size of ith group
- $\bar{x}_i$ = sample mean of ith group
- s_i = sample standard deviation of ith group

Mean Square Between Groups (MSG)	Mean Squared Error (MSE)
Average variability between group means	Average variability within groups
$MSG = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$MSE = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2$
$df_G = k-1$	$df_E = n-k$
Numerator degree of freedom	Denominator degrees of freedom

Single Factor ANOVA F Test → F distribution

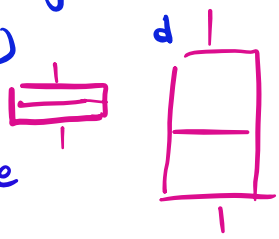
single variable

When to use:

Want to test for differences in averages between more than two populations

Conditions required for inference:

- Representatives samples from each of the populations of Intrest
- samples need to be sufficiently large ($n_i \geq 30$ or ($n_i < 30$) check if sample distributions suggest symmetric population)
- Constant-variance across groups



Null & Alternative hypotheses:

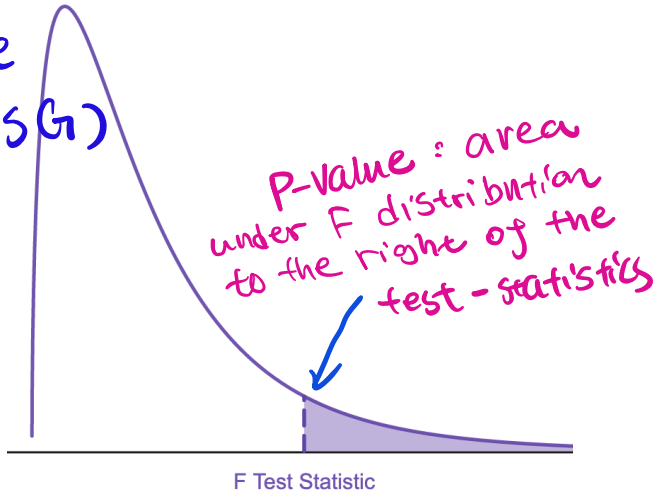
H_0: $\mu_1 = \mu_2 = \cdots \mu_k$ All population means are equal

H_A: At least one population means differs from at least one other.

or equivalently, At least two population means differ from each other.

Test Statistic: Ratio of the average between group variability (MSG) to the average within group variability.

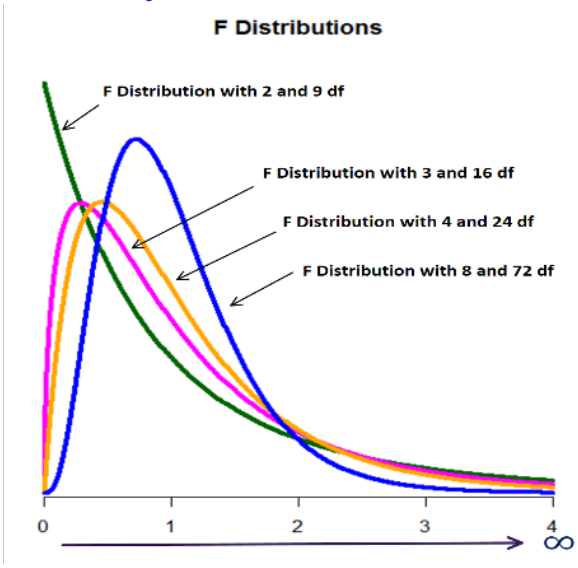
$$F = \frac{MSG}{MSE}$$



Null distribution of this test statistics is an F distribution with k-1 and n-k degrees of freedom

F Distributions

- Positively skewed distribution ranging from 0 to ∞
- The shape is defined by numerator df (v_1) and denominator df (v_2)
- Denoted: F_{v_1, v_2}
- For a single factor ANOVA F test, the p-value is always the area under the distribution curve to the Right of the F statistic in the distribution:



Finding *F* p-values using pf()

For an ANOVA F test, the p-value is the area under the curve of an F distribution with v1 and v2 degrees of freedom, to the right of the F statistic.

Function	Function Values	What does it do?
`pf(f,v1,v2)`	`f` = F statistic `v1` = numerator degrees of freedom `v2` = denominator degrees of freedom	This is the cumulative distribution function for an F distribution. $P(X \leq x)$

To find the p-value, you need to subtract the above function from 1!

$1 - Pf(F, v_1, v_2)$

ANOVA *F* Test Example

Set up the hypotheses to test whether the average On-base Percentage differed significantly between the Red Sox, Astros, and Yankees in 2018.

$H_0: \mu_R = \mu_A = \mu_Y$

H_A: At least one team's average on-base percentage differs from at least one other team

Performing an ANOVA F test by hand is tedious and unnecessary. R can do all the calculations for us in a few easy steps. Below is the R output for this hypothesis test.

degree of freedom

F statistics

	DF	Sum Sq	Mean Sq	F	Pr(>F)
Team	2	0.0223	0.01113	0.405	0.668
Residuals	87	2.3919	0.02749		

Row 1: between group variability

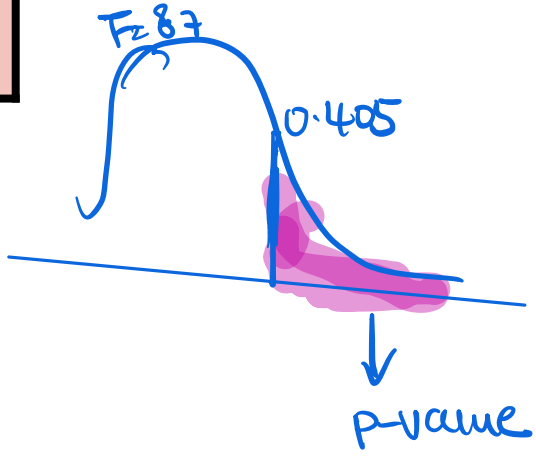
Row 2: within group variability

numerator

denominator

MSG

MSE



Using the R output, determine the df_G, df_E, MSG, MSE, the F statistic, and the p-value.

$df_G = 2$ $MSG = 0.0113$ $F = 0.405$

$df_E = 87$ $MSE = 0.027\overline{49}$ p-value $= 0.668$

Using the output, write a conclusion for the hypothesis test.

Because we have such a large p-value, we fail to reject the null hypothesis.

There is no evidence to suggest that the average on-base percentages differ between the three teams.

Using R for ANOVA

The following code templates can be used to perform an ANOVA in R.

Function	Function Values	What does it do?
`mod <- aov(response ~ treatment)`	`mod` saves the ANOVA model in an object called mod `response` = vector containing the response variable values `treatment` = vector that defines the groups for each recorded response	Calculates the values necessary to perform an ANOVA F test
`summary(mod)`	`mod` call the saved model stored above	Prints the ANOVA table

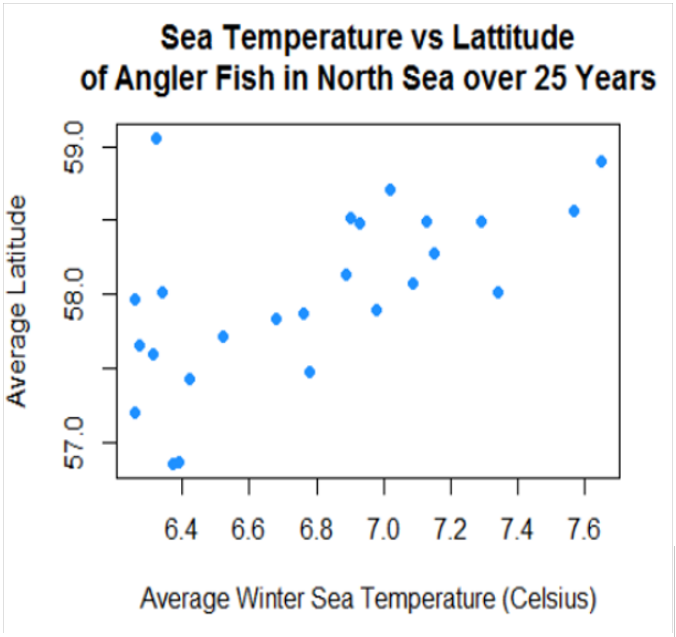
ST 314

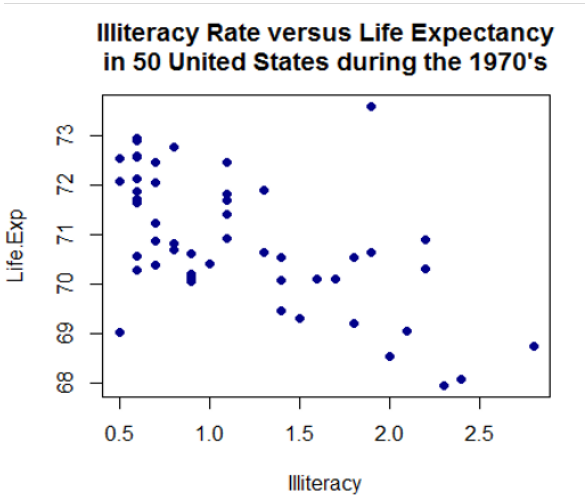
Week 8

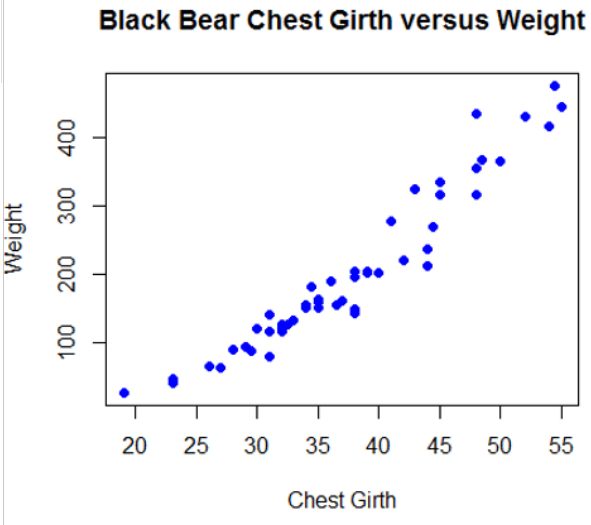
Simple Linear Regression

Bivariate Relationships

Bivariate data looks at the relationship between quantitative variables. Each unit has a measurement for both variables so the data is paired.







Explanatory and Response Variables

Variable Types for Bivariate Relationships

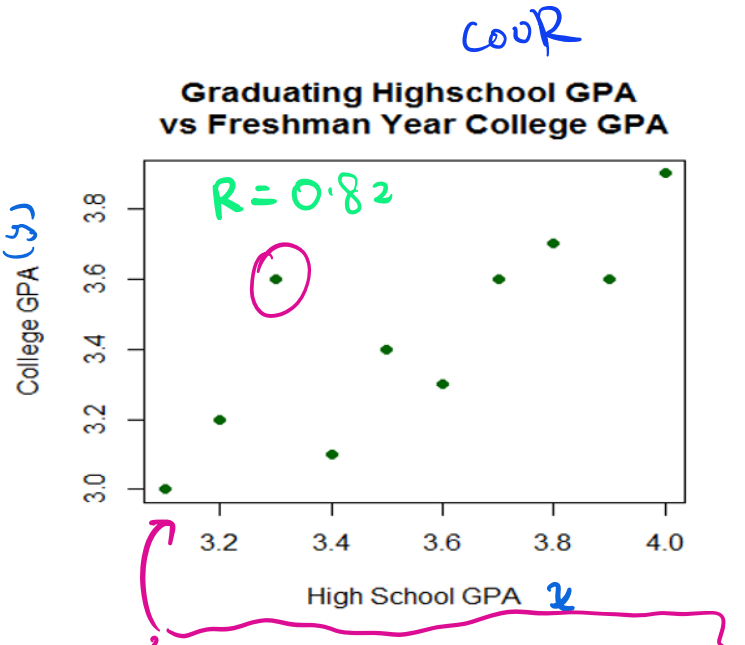
Explanatory

- May help explain or predict changes in response variable

- Quantitative
- Sometimes referred to as: x,
  - independent Variable
  - predictor variable

Response

- Variable to be estimated or predicte
- Quantitative
- Sometimes referred to as: y
  - dependent Variable



COOR

Strength: moderly strong
direction: positive
Form: linear
Outliers: 3.6

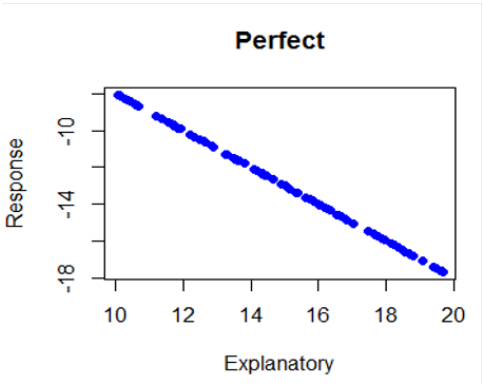
Interpreting Scatterplots

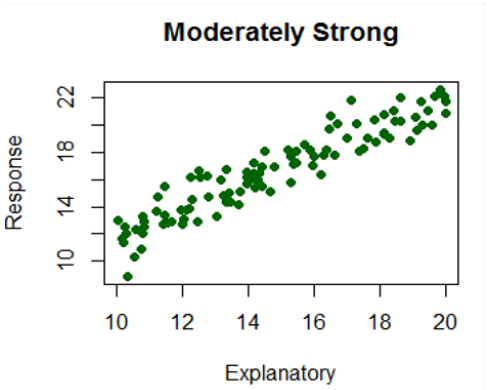
- Strength weak to strong
- Direction positive or negative
- Form linear or non-linear
- Outliers observations that fall outside the general pattern of the data

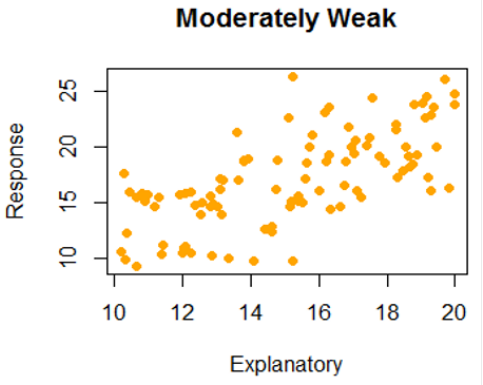
There is a moderate, positive, linear relationship between high school GPA and freshman year college GPA.

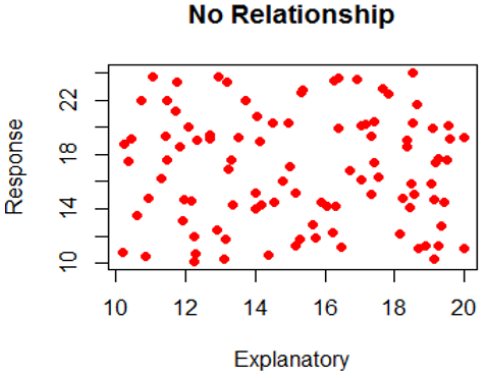
It appears there is an outlier at 3.3 and 3.6, where a student performed better in college than what we might expect.

1. Strength

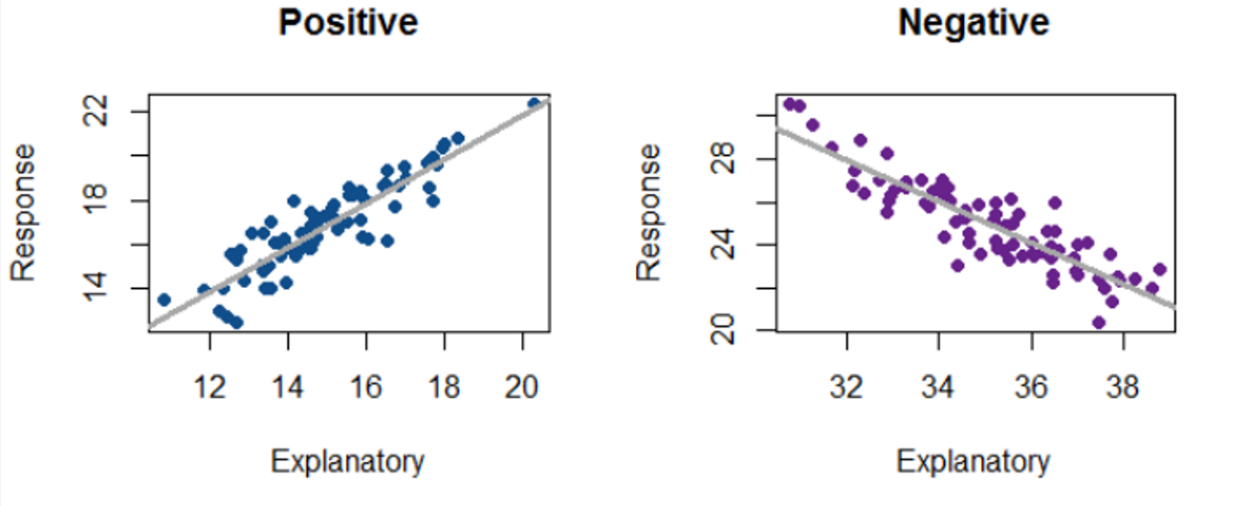




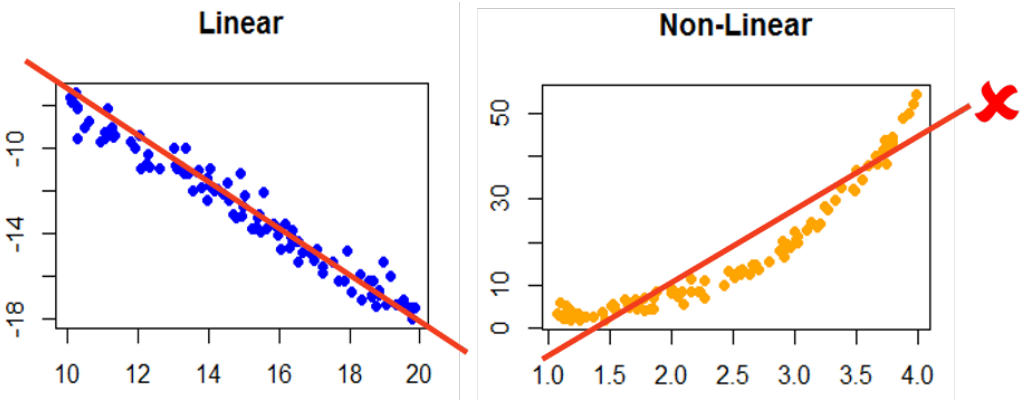




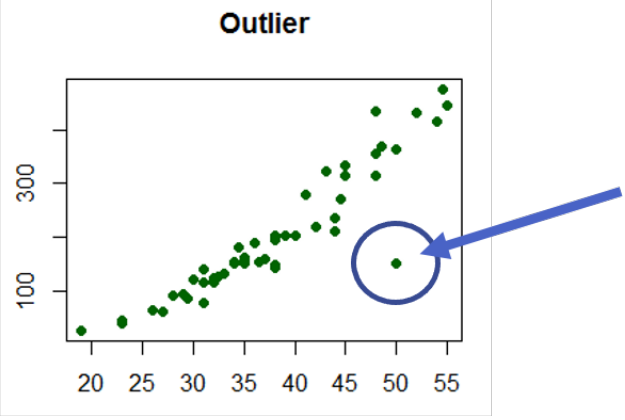
2. Direction



3. Form



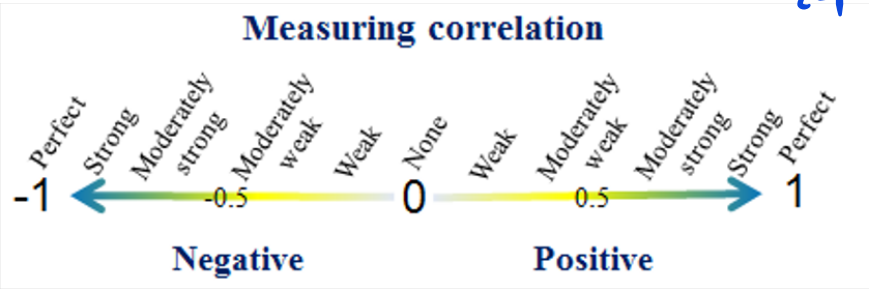
4. Outliers



Measuring Strength with Correlation

The correlation coefficient, *R*, measures the strength of a linear association between two quantitative variables.

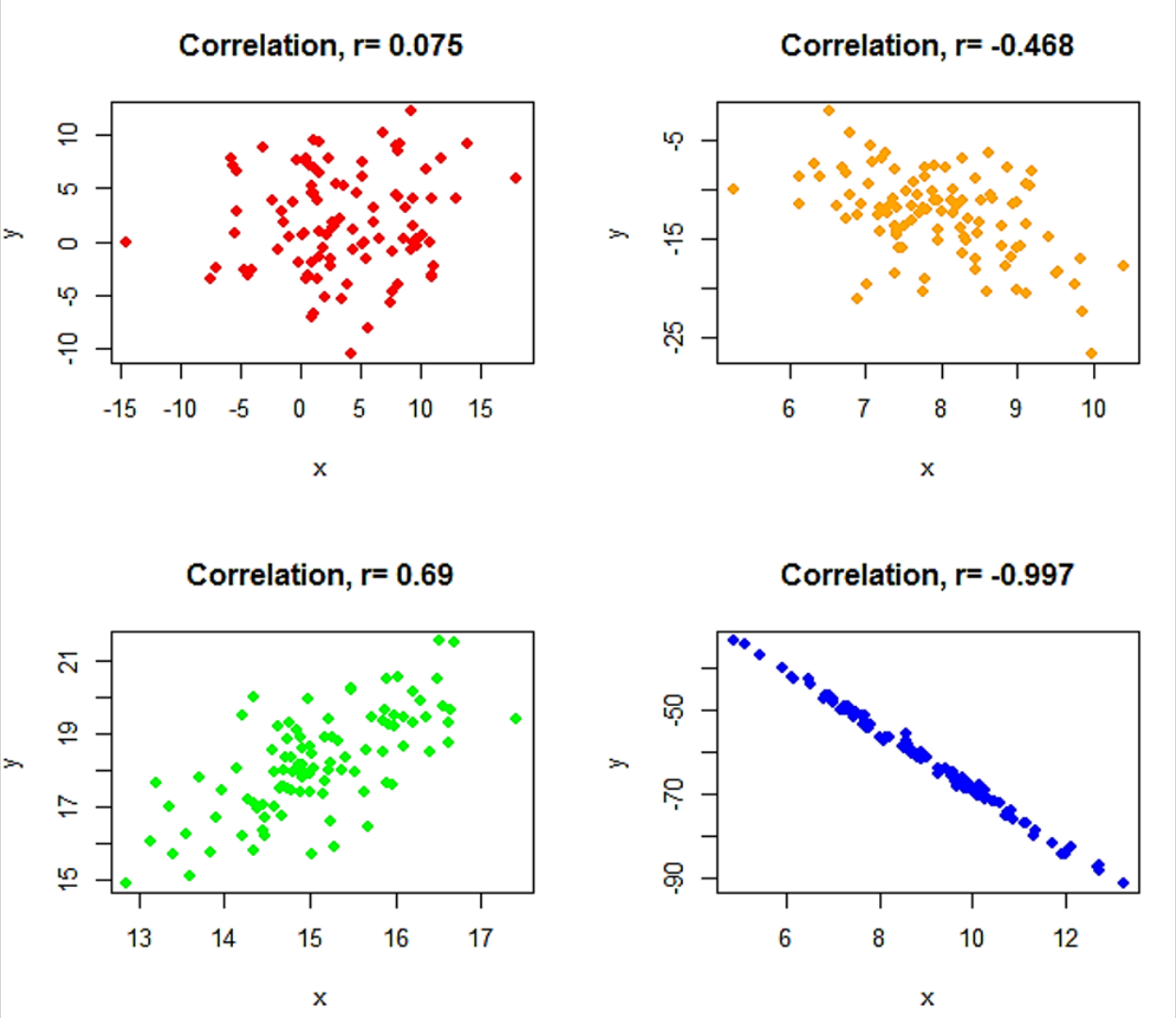
$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)$$



x_i = ith obs of explantory

$\bar{x}$ = mean of explantory

s_x = sd of explantory



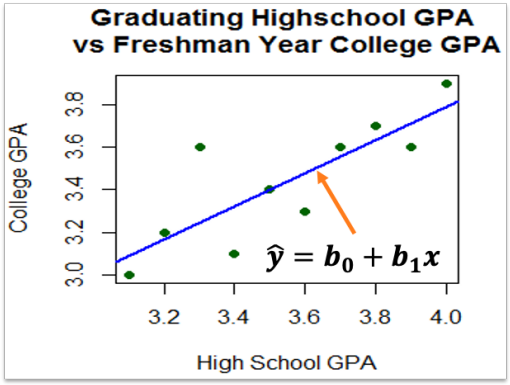
Least Squares Regression Line

A common goal in statistics is to model the relationship between quantitative variables.

The model that best fits the linear relationship is the least squares regression line:

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

- $\hat{y}$: predicted response variable
- b_0: estimate for y-intercept
- x: explanatory variable



What does “least squares” mean?

The LSRL is calculated by finding the smallest value for the sum of squared error, SSE, the sum of squared differences between the observed and predicted value of *y*.

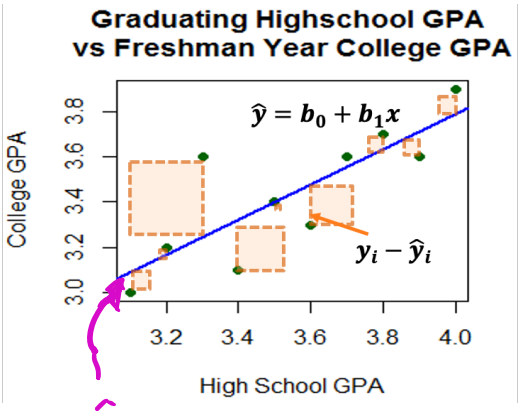
mean of response

The following are the outcomes for minimizing SSE:

$b_0 = \bar{y} - b_1\bar{x}$ → mean of explanatory

$b_1 = R\left(\frac{s_y}{s_x}\right)$ → standard dev. response / standard dev. explanatory

correlation coefficient



$\hat{y} = 0.69 + 0.78x$

Interpreting the LSRL

Interpreting the intercept:

expected value of the response when the explanatory variable is 0 ($x=0$)

Does not always have a practical interpretation

Interpreting the slope:

For a one unit increase in the explanatory variable, we expect the response to change by b_1

Example: The LSRL for the HS GPA vs. College GPA problem is $\hat{y} = 0.69 + 0.78x$. Interpret the slope and intercept estimates.

For every one point increase in HS GPA, we expect freshmen college GPA to increase by 0.78 points

Calculating the LSRL in R

Function	Function Values	What does it do?
`mod <- lm(y ~ x)`	`mod` saves the ANOVA model in an object called mod `x` = vector containing the explanatory variable values `y` = vector containing the response variable values	Computes the linear model for the explanatory and response variable relationship
`summary(mod)`	`mod` call the saved model stored above	Gives the LSRL estimates, standard errors and performs a t test on the slope. Provides R^2 and adjusted R^2 value and results from a model utility F test.

```
HSGPA <- c(3.4, 3.1, 3.7,3.3, 3.5, 4, 3.2, 3.8, 3.9, 3.6)
CollegeGPA <- c(3.1,3,3.6,3.6,3.4,3.9,3.2,3.7,3.6, 3.3)

# Use the lm() command to get the linear model.
# Take the summary() of lm() to get more information.
mod <- lm(CollegeGPA~HSGPA)
summary(mod)
```

Reading Statistical Software Output

R Output for Linear Regression

```
lm(formula = CollegeGPA ~ HSGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22364 -0.10636  0.01515  0.06045  0.35394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6861     0.6897   0.995  0.34902
HSGPA         0.7758     0.1937   4.006  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```

$\hat{y} = 0.6841 + 0.7758x$

Rounded to 2 decimals

$\hat{y} = 0.69 + 0.78x$

Basic Predictions from the LSRL

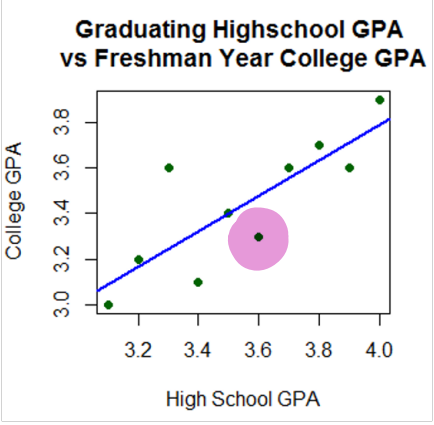
Let's use the LSRL to predict the freshman year college GPA of a student who earned a 3.6 high school GPA.

$\hat{y} = 0.69 + 0.78x$

$\hat{y} = 0.69 + 0.78(3.6) \approx 3.498$

Residuals

The difference between an observed value of the response and the value predicted by the regression equation for the same value of x is called the residual.



Residual = observed response − predicted response

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

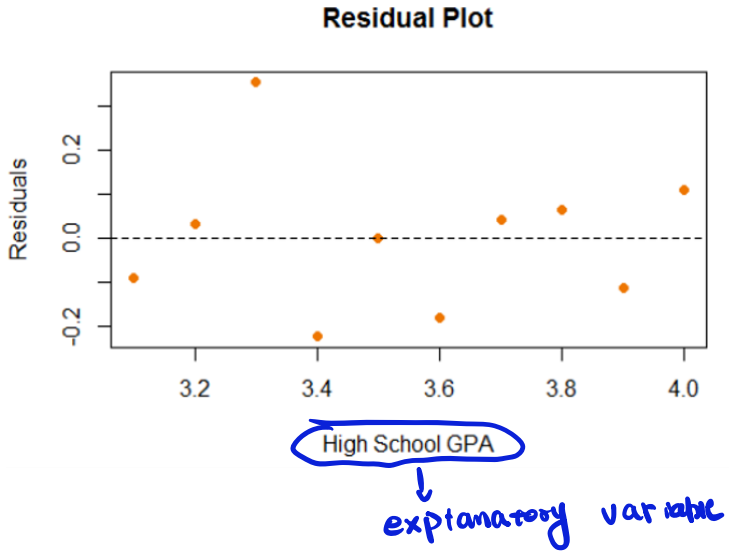
Example: Calculate the residual for the student with the HS GPA and College GPA of 3.6 and 3.3, respectively, given the regression line $\hat{y} = 0.69 + 0.78x$.

$e = 3.3 - 3.498 = -0.198$

Residual Plot

x_i y_i $\hat{y}_i$

HS GPA	College GPA	Fitted Value	Residual
3.4	3.1	3.342	-0.242
3.1	3	3.108	-0.108
3.7	3.6	3.576	0.024
3.3	3.6	3.264	0.336
3.5	3.4	3.42	-0.02
4.0	3.9	3.81	0.09
3.2	3.2	3.186	0.014
3.8	3.7	3.654	0.046
3.9	3.6	3.732	-0.132
3.6	3.3	3.498	-0.198



A residual plot is a scatterplot of all n residuals.

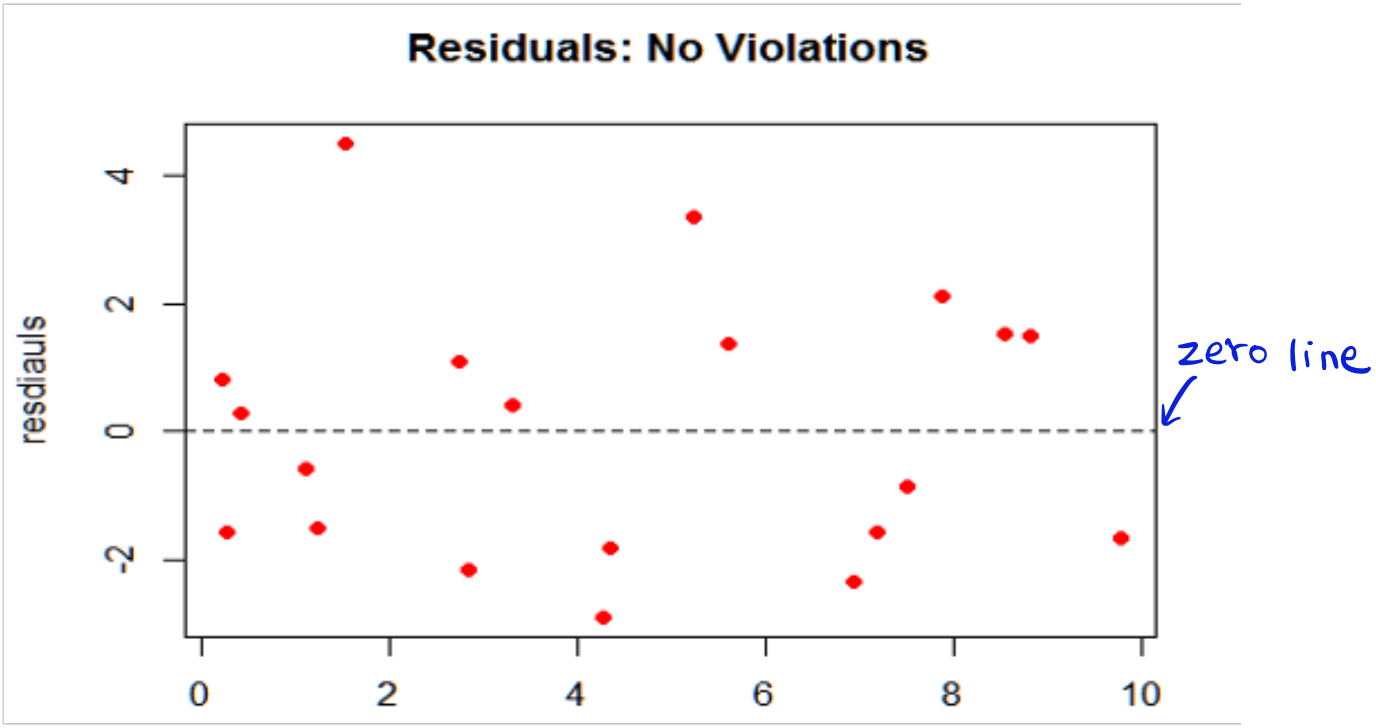
Used to assess:

- fit of regression line
- check necessary conditions for inference

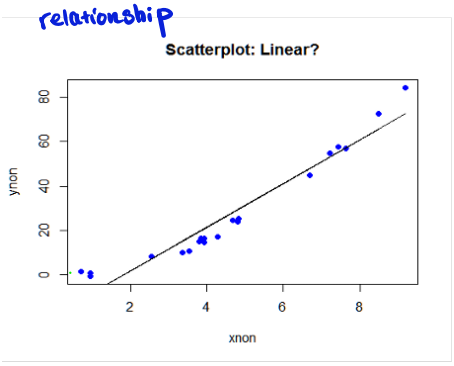
It is desirable to see no patterns in the residual plot.

Interpreting Residual Plots

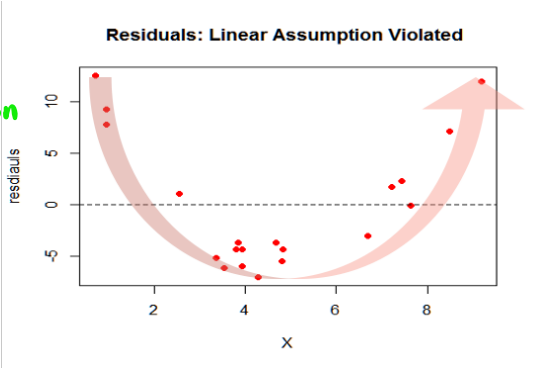
Condition	How to check
The relationship is linear in the population.	No "u" shape or curvature in the residuals.
The response varies normally about the population regression line.	Random scater of residuals above and below zero line.
Observations are independent.	No pattern in the residual plot Also, check the sampling asign (were data randomly collected)
The standard deviation of the responses is the same for all values of x. Constant variance around the LSRL	No Funnel shape in the residuals

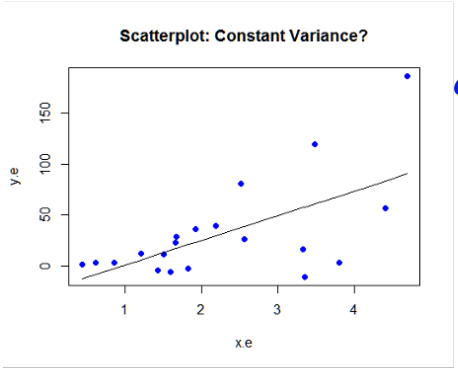


Scatterplot of Bivarite relationship

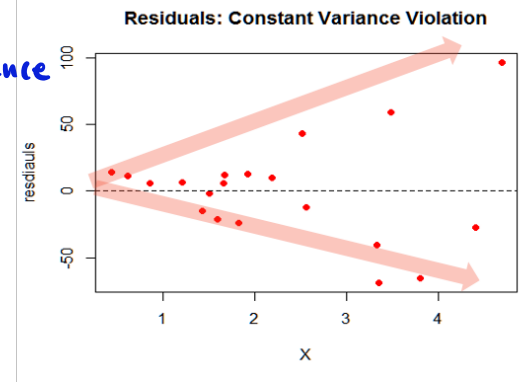


Linearity Condition Violated





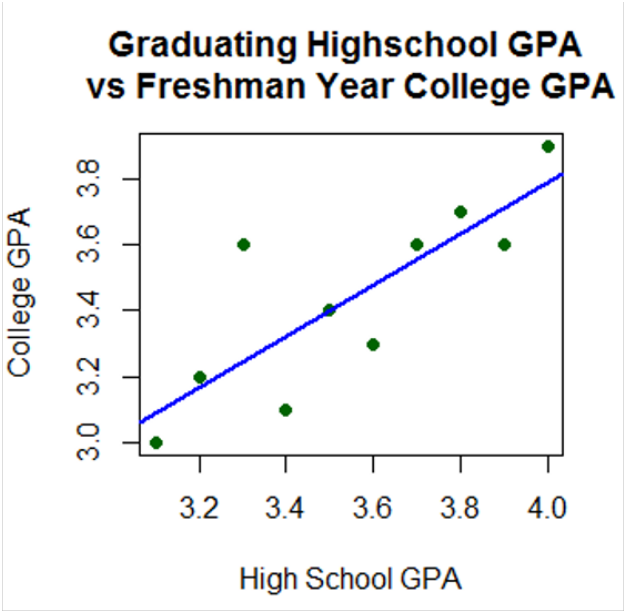
Constant Variance Condition Violated

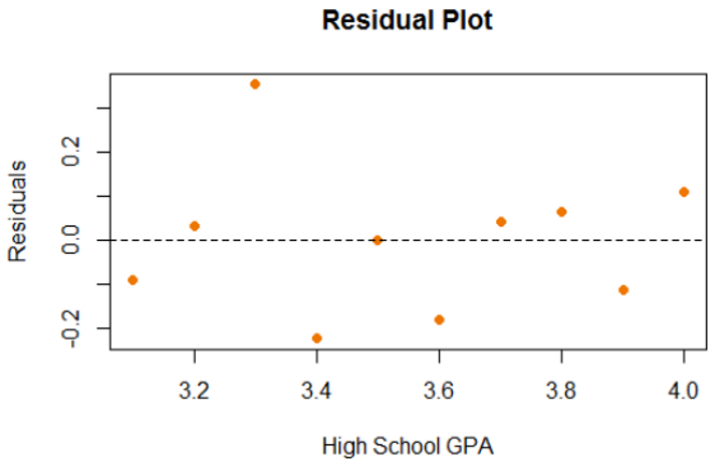


When we see these patterns in our residuals the least squares regression line is no longer reliable to make predictions about the data!

Checking Conditions for Inference Example

Given the sample of 10 students was obtained randomly. Based on the residuals are the conditions for inference satisfied, to make inference from our sampled data to the population?





Conditions are meet

ST 314

Week 9

Inference for Linear Regression

Estimating Parameters

The least squares regression line (LSRL) is based on sampled data so $\hat{y} = b_0 + b_1 x$ is the estimate for the true population regression line $\mu_y = \beta_0 + \beta_1 x_1$.

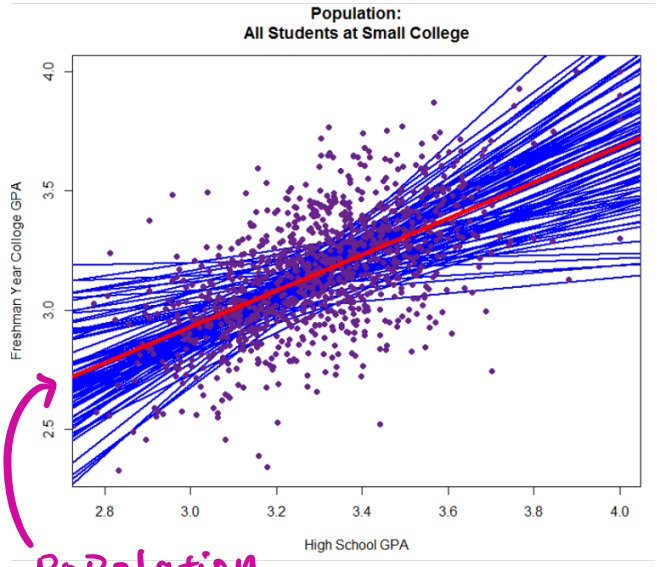
β_1 → true slope for population

β_0 → true y-int for pop

μ_y → average response for a given value of x

b_1 is the point estimate for β_1.

Random sampling implies that obtaining a different sample will obtain a different estimate for the LSRL.



Population regression equation

b_0 and b_1 are

random variables

b_0 and b_1 are unbiased point estimates for β_0 and β_1, respectively, and come with a certain amount of variability.

t Confidence Interval for β_1 → true slope pop

When to use:

want to estimate the slope parameter, β_1, of the population regression equation

Conditions required for inference:

The 4 conditions that can be assessed using the residual plot: linearity, normality, independence, constance variance (see the end of simple linear regression note)

n = # of observations

k = # of explanatory variable used in the model

Confidence interval for the slope of the population, β_1:

$b_1 \pm t^*_{n-(k+1)} (SE_{b_1})$

$t^*_{n-(k+1)}$: critical value

SE_{b_1}: standard error estimate for b_1

t Confidence Interval for β_1 Example

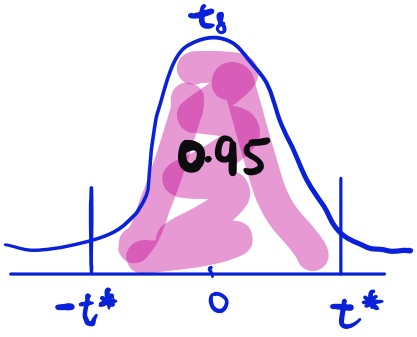
From the random sample of 10 students, calculate and interpret the 95% confidence interval for the true change in College GPA given a change in High School GPA. Use the software output where $b_1 = 0.7758$ and $SE_{b_1} = 0.1937$. Assume all of the conditions for inference are met.

```
Call:
lm(formula = CollegeGPA ~ HSGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22364 -0.10636  0.01515  0.06045  0.35394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6861     0.6897   0.995  0.34902
HSGPA         0.7758     0.1937   4.006  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```



$qt(0.975, 8) = 2.304$

$b_1 \pm t^*_{n-(k+1)} SE_{b_1}$

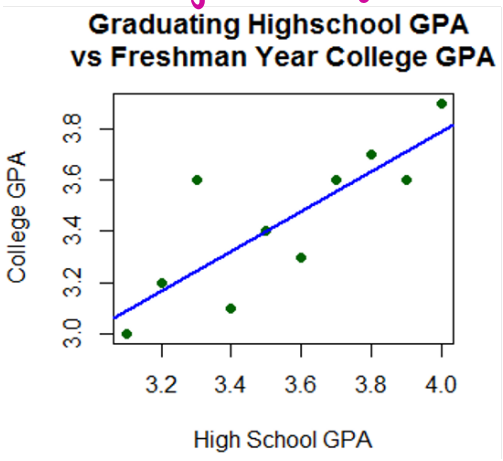
$0.7758 \pm 2.304(0.1937) = (0.329, 1.222)$

We are 95% confident that for every one point increase in Hs GPA, we expect freshmen college GPA to increase by 0.329 to 1.222 points with a point estimate of 0.7758 points

Testing for a Relationship

It is common to test whether the relationship between two quantitative variables is statistically significant.

Suppose a small college would like to know if their students high school GPA is a significant predictor of their freshman year college GPA. They take a random sample of 10 students. How can they answer this question?



t Test for the Slope β_1

When to use:

want to test whether the explanatory variable explains changes in the response variable

Conditions required for inference:

The 4 conditions that can be assessed using the residual plot: linearity, normality, independence, constance variance (see the end of simple linear regression note)

Null and alternative hypotheses:

$H_0: \beta_1 = 0$ (no relationship between x and y)

$H_A: \beta_1 \neq 0$

Test statistics $= \frac{b_1}{SE_{b_1}} \sim t_{n-(k+1)}$

null distribution

Testing for a Linear Relationship Example

Use the software output to test whether high school GPA is a significant predictor of freshman year college GPA at the small college. The data is from a random sample of 10 students. Assume all of the conditions for inference are met. Use $\alpha = 0.05$.

```
Call:
lm(formula = CollegeGPA ~ HSGPA)

Residuals:
      Min        1Q    Median        3Q       Max
-0.22364 -0.10636  0.01515  0.06045  0.35394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6861     0.6897   0.995  0.34902
HSGPA         0.7758     0.1937   4.006  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```

Hypotheses:

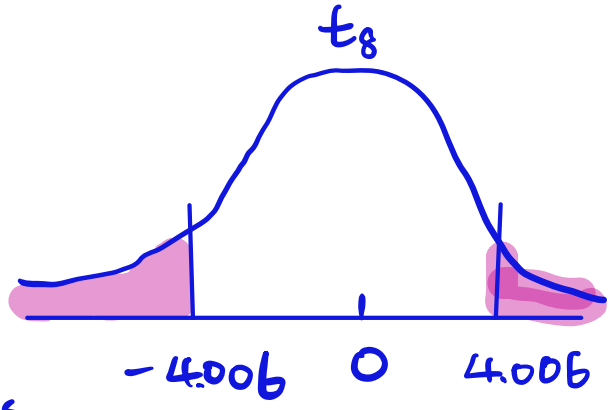
$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

Test statistic and p-value:

Test statistic: $t = 4.006$

P-value $= 0.003918$



Conclusion:

reject the null hypothesis

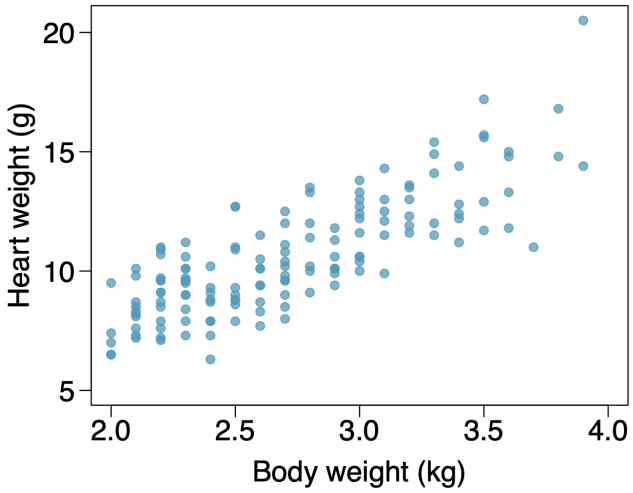
There is convincing evidence to suggest that HS GPA is a significant predictor of freshmen college GPA

Reading Statistical Software Output

	Estimate	Standard Error	*t* Statistic	p-value
Intercept				
Explanatory Variable				

Regression Inference Example

The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

Write the equation of the regression line.

Interpret the slope estimate.

Do the data provide strong evidence that an increase in body weight is associated with an increase in heart weight?
Construct a 95% confidence interval for the slope parameter β_1.

State the null and alternative hypotheses, report the p-value, and write a conclusion β_1.

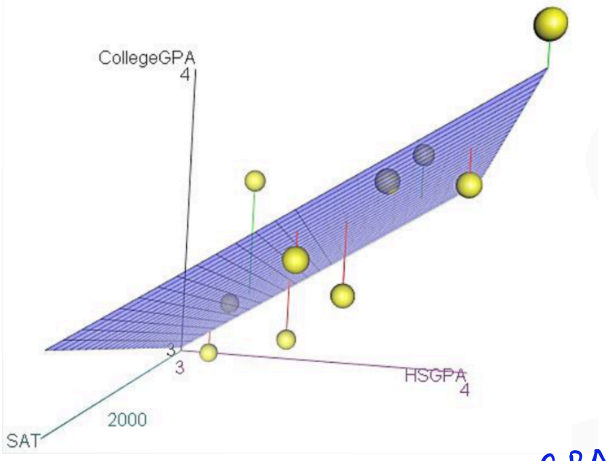
ST 314

Intro to Multiple Linear Regression

Multivariate Data

Along with HS GPA, suppose SAT scores may also be a good predictor of freshman year college GPA. Can we have more than one explanatory variable?

Yes

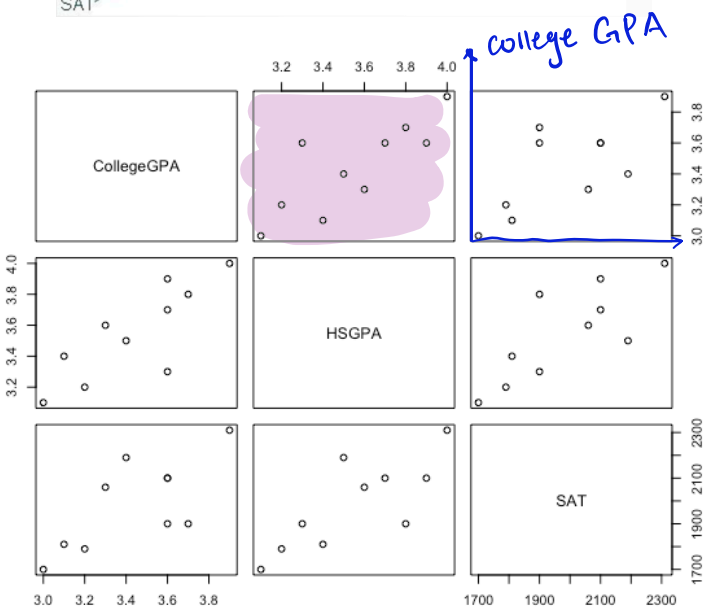


Visualizing Multivariate Data

As the number of explanatory variables increase, visualization can become tricky.

We can visualize bivariate relationships between all involved variables using a

Scatterplot Matrix.



Multiple Linear Regression Equation

A multiple linear regression (MLR) models the relationship between one

quantative response variable and multiple explanatory variables

Can be either quantative or categorical

ST 314

k = # of explanatory variable

General form of the estimated multiple linear regression model:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

$\hat{y}$: predicted response

b_0: y-intercept

b_1: estimate for slope coefficient on x_1

The coefficient estimates are determined by finding the model that yields the smallest sum of squared error (just like in simple linear regression).

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

MLR Example

Consider a random sample of 141 eBay auction sales for the Nintendo Wii game Mario Kart. We are interested in modeling total sale price from starting price of the bid, number of wheels, and condition.

response → y; x_1 → exp; x_2; x_3

total.price	start.price	wheels	used
51.55	0.99	1	0
37.04	0.99	1	1 → true
45.5	0.99	1	0
...			
54.51	1	2	0

quantitative; Categorical

x_3 { 1 if used; 0 otherwise (new)

Model Output:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	41.79739	0.99545	41.988	< 2e-16	***
start.price	0.12376	0.02566	4.824	3.70e-06	***
wheels	6.83250	0.50964	13.406	< 2e-16	***
used	-5.69207	0.85817	-6.633	7.02e-10	***

Write the estimated MLR equation using the model output.

$\hat{y} = 41.80 + 0.12\,\text{StartPrice} + 6.83\,\text{wheels} - 5.69\,\text{used}$

$\hat{y}$: predicted total sale price

used: $\hat{y} = (41.80 - 5.69) + 0.12\,\text{StartPrice} + 6.83\,\text{wheels}$

New: $\hat{y} = 41.80 + 0.12\,\text{StartPrice} + 6.83\,\text{wheels}$

Coefficient of Determination, R^2

How do we know if the model is a good fit?

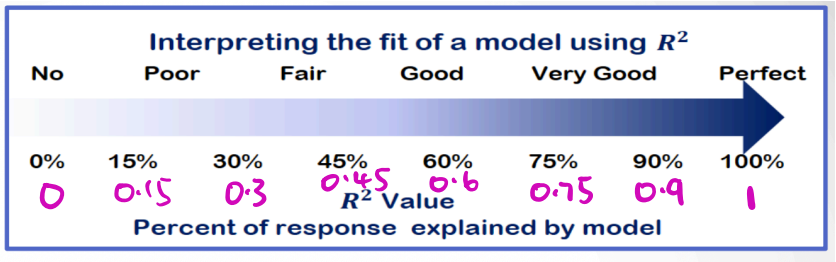
- The coefficient of Determination, R^2, defines the proportion of variability in the response variable that can be explained by the explanatory variables used in the model
- $R^2 = \frac{SSR}{SST}$

Total Sum of Sqaures (SST)	Model Sum of Squares (SSR)	Error Sum of Squares (SSE)
Total variability in response variable $\sum_{i=1}^{n}(y_i - \bar{y})^2$ y_i: Observed response value $\bar{y}$: average of observed response value	Variability in response explained by the model $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ $\hat{y}_i$: predicted response value	unexplained or left over variability in the response $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $\hat{y}_i$: predicted response value

$SST = SSR + SSE$

Interpreting R^2

The higher the R^2 value the better the explanatory variables are at explaining the variability in the response!



ST 314

Model Output:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.79739    0.99545  41.988  < 2e-16 ***
start.price  0.12376    0.02566   4.824 3.70e-06 ***
wheels       6.83250    0.50964  13.406  < 2e-16 ***
used        -5.69207    0.85817  -6.633 7.02e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.535 on 137 degrees of freedom
Multiple R-squared:  0.7577,    Adjusted R-squared:  0.7524
F-statistic: 142.8 on 3 and 137 DF,  p-value: < 2.2e-16
```

R^2

The explanatory variables in the model explain 75.77 % of the variation in total selling price for Mario Kart in the eBay Auctions.

R^2 = 0.7577

The model is a very good fit.

R^2 vs. Adjusted R^2

Cautions about R^2:

adding variables to the model will alway increase R^2 even if the variable are not good predictors of the response.

Adjusted R^2 accounts for

The number of expanatory variables used in the model

Adjusted R^2 = $1-\left(\frac{n-1}{n(k+1)}\right)\left(\frac{SSE}{SST}\right)$

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

R^2 vs. Adjusted R^2 Example

Compare the fits of the two models for predicting college GPA. What effect does adding SAT to the model have on the R^2 value? What about adjusted R^2? Which model is a "better" fit?

Model with HS GPA and SAT score as the explanatory variables:

```
Residual standard error: 0.1838 on 7 degrees of freedom
Multiple R-squared:  0.6821,    Adjusted R-squared:  0.5912
F-statistic: 7.508 on 2 and 7 DF,  p-value: 0.01812
```

Model with only HS GPA as the explanatory variable:

```
Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```

Adjusted R^2 Example

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.79739    0.99545  41.988  < 2e-16 ***
start.price  0.12376    0.02566   4.824 3.70e-06 ***
wheels       6.83250    0.50964  13.406  < 2e-16 ***
used        -5.69207    0.85817  -6.633 7.02e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.535 on 137 degrees of freedom
Multiple R-squared:  0.7577,    Adjusted R-squared:  0.7524
F-statistic: 142.8 on 3 and 137 DF,  p-value: < 2.2e-16
```

While accounting for sample size and the number of explanatory variables in the model, the explanatory variables in the model explain 75.24 % of the variation in total selling price for Mario Kart in the eBay Auctions.

The model is a very good fit.

ST 314 Practice Final Exam

The following questions are meant to give you sense of what the midterm exam might look like. This should not be your only source of study material! See Canvas for suggestions of other resources you can use to prepare for the exam.

Use for the following 4 questions. For each scenario indicate the matching test by filling in the corresponding letter in the _____. Each test can only be used once.

A. Two-sample t test	B. Single-factor ANOVA	C. One Sample t test	D. Simple linear Regression.

1) Matching test letter: _____ Jenny would like to know whether she has budgeted enough money per month for groceries while she is in college. She takes a representative random sample of 50 college students and asks them how much they spend a month on groceries per month. Jenny has budgeted $200. Which test can she use to test whether she has budgeted enough for groceries per month?
2) Matching test letter: _____ Research has found that there is a relationship between neck circumference and body fat percentage. A nationwide fitness franchise would like to be able to predict body fat percentage from neck circumference. Which type of analysis is appropriate for this scenario?
3) Matching test letter: _____ The lumen output is measure on 6 light bulbs each for 4 different brands of comparable bulbs. A contractor is interested in finding whether there is significant difference in the average lumen output for the different bulb brands. What type of procedure would you recommend?
4) Matching test letter: _____ A student would like test whether the average rent per room is higher for dwellings that are closer to campus versus further away. They take a random sample of 40 dwellings within 1 mile of their University campus, and another random sample of 40 dwellings more than 1 mile from campus. For each sample they record the cost of rent per bedroom. Which type of analysis is appropriate for this scenario?

5) If the p-value for a hypothesis test is very small, there is convincing evidence to support the alternative hypothesis.
    a. True
    b. False

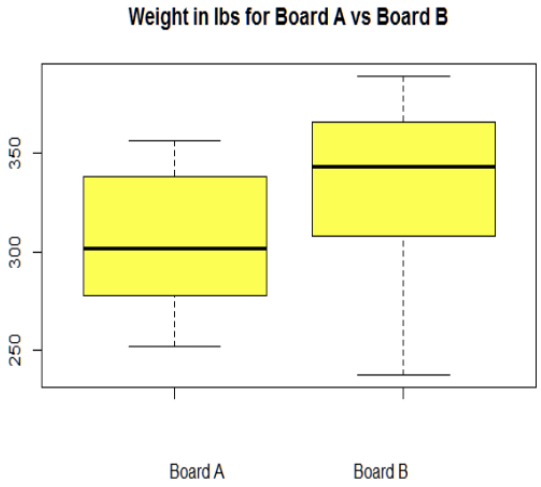
6) Suppose you have two single factor ANOVA experiments with the same degrees of freedom. The resulting F statistics are:
Experiment 1 $F = 5.68$
Experiment 2 $F = 20.15$
Which statement is true in regards to comparing Experiment 1 and Experiment 2?
   a. Procedure 1 has a smaller test statistic and therefore will result in stronger evidence in favor of the alternative hypothesis.
   b. Procedure 2 has a larger test statistic and therefore will result in strong evidence in favor of the alternative hypothesis.
   c. We should reject the null for both tests.
   d. Impossible to know with this information.

Use the following for the next two questions. In an experiment to study automobile engine operating efficiency for five different brands of gasoline, mpg was measured over a controlled distance and speed for eight cars in each group. Let μ_i represent the average mpg for the ith brand.

7) Identify the null and alternative hypotheses for the single factor ANVOA F test used to test if there is a difference in average mpg between the five brands.
   a. $H_0: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$ vs. $H_A:$ the average for a least one brand differs from the other four brands.
   b. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ vs. $H_A:$ the average for a least one brand differs from the other four brands.
   c. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ vs. $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$
   d. $H_0: \mu_1 = \mu_2 = \mu_3$ vs. $H_A: \mu_4 = \mu_5$

8) Assume mpg is normally distributed for all five gasoline brands. Identify the null distribution for the F test statistic in this hypothesis testing scenario.
   a. t distribution with 39 degrees of freedom
   b. Standard normal distribution
   c. F distribution with 4 and 35 degrees of freedom
   d. Left skewed distribution

Use the following to answer the next 3 questions. A lumber manufacturing company would like to compare the average weight capacity of two types of boards they currently manufacture. The sampled boards were obtained using a random mechanism.



Population	Sample mean	Sample Standard deviation	Sample Size
Board A	305.4 *lbs*	36.5	16
Board B	331.7 *lbs*	44.3	26

data: boardA and boardB
t = -2.0875, df = 36, p-value = 0.0439
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-51.8601685 -0.7621273
sample estimates:
mean of x mean of y
305.3603 331.6715

9) Describe the side-by-side box plot. Include a comparison of the two group, the center, shape, and spread of each.

10) State the null and alternative hypotheses to test whether there is a difference between the two boards.

11) Using the R output provided, using a significance level of 0.05, write a 4-part conclusion based on the software output for the hypothesis test.

Use the following to answer the next 3 questions. According to the National Institute on Alcohol Abuse and Alcoholism (NIAAA), and the National Institutes of Health (NIH), 41% of college students nationwide engage in "binge drinking" behavior, having 5 or more drinks in one occasion during the past two weeks. A college president wonders if the proportion of students enrolled at her college that binge drink is lower than the national proportion. In a commissioned study, 462 students are selected randomly from a list of all students enrolled at the college. Of these, 162 admitted to having engaged in binge drinking.

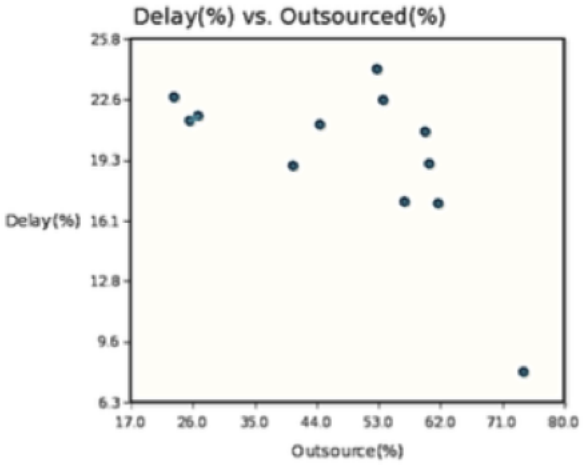
12) The college president is more interested in testing her suspicion that the proportion of students at her college that binge drink is lower than the national proportion of 0.41. Write

the null and alternative hypotheses for this test.

13) Write an expression for the test statistic needed to perform this hypothesis test. You do not need to evaluate the expression.

14) Which of the following conclusions is reasonable, based on a p-value of 0.0047?
    a. There is little evidence to support a conclusion that the proportion of students at this particular college that binge drink is lower than the national proportion of 0.41.
    b. There is no evidence that the proportion of binge drinking students at this college is lower than the national proportion of 0.41.
    c. There is strong evidence that the proportion of students at this college that binge drink is lower than the national proportion of 0.41.
    d. We can't reach any reasonable conclusion because the assumptions necessary for a significance test for a proportion are not met in this case.

Use the following information for the next 2 questions. Airlines have increasingly outsourced the maintenance of their planes to other companies. Flight delays are often due to maintenance problems. Critics are concerned that the maintenance may be done less carefully, such that outsourcing creates safety hazards and delays. The following is a simple linear regression analysis on data from 2005 and 2006 on the percent of outsourcing for 12 airlines and their respective percent of delayed flights. Does the data support the concerns of the critics?



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.48	3.296	8.338	<0.0001
Outsource	-0.1636	0.06488	-2.522	0.03030

15) Assuming the conditions are met, what is the least squares regression equation for estimated percent of flights delayed given the percent of maintenance outsourced by the airline.

a. $y = 27.48 - 0.1636x$
b. $x = 0.1636 - 27.48y$
c. $\hat{y} = 0.1636 - 27.48x$
d. $\hat{y} = 27.48 - 0.1636x$

16) From the regression analysis, which statement is true about the relationship between the percent of outsourced maintenance and percent of flight delays?

a. The critics were right. Outsourcing causes delays.
b. The LSRL estimates that, on average, for every 1% increase in outsources maintenance, flight delays increase by 27.48%.
c. The LSRL estimates that for every 1% increase in outsources maintenance the average percent of flight delays is expected to decrease by 0.1636.

17) Which of the following statements is FALSE about the correlation coefficient, *r*?

a. The correlation coefficient is a unitless number and must always lie between 0 and 1, inclusive.
b. The correlation coefficient can only describe the relationship between two quantitative variables.
c. If r =1, then there is a perfect positive association between x and y.
d. The correlation coefficient is a unitless number and must always lie between –1 and 1, inclusive.

Multiple Linear Regression Examples

Loan 50 Data Set

This data set represents 50 loans made through the Lending Club platform, which is a platform that allows individuals to lend to other individuals. Of course, not all loans are created equal. Someone who is essentially a sure bet to pay back a loan will have an easier time getting a loan with a low interest rate than someone who appears to be riskier. And for people who are very risky? They may not even get a loan offer, or they may not have accepted the loan offer due to a high interest rate. It is important to keep that last part in mind, since this data set only represents loans actually made, i.e. do not mistake this data for loan applications!

Suppose we are interested in predicting the interest rate on a loan given a set of additional information. We'll the following variables in this example:

response variable

explanatory variables

`interest_rate`	interest rate of the loan the applicant received	Quantitative
`verified_income`	type of verification of the applicant's income (source verified, verified, or not verified)	Categorical
`debt_to_income`	debt to income ratio	Quantitative
`total_credit_utilizied`	total credit balance, excluding a mortgage	
`public_record_bankrupt`	number of bankruptcies listed in the public record for this applicant	
`term`	the number of months of the loan the applicant received	

1. What would a multiple linear regression model with all six variables above look like?

$$\hat{y} = b_0 + b_{SV}x_{SV} + b_V x_V + b_D x_D + b_{TC} x_{TC} + b_P x_P + b_t x_t$$

$\hat{y}$: predicted interest rate

x_D: Debeto income

x_{TC}: total credit utilized

x_P: Pubic record bankrupt

x_t: term

$$x_{SV} = \begin{cases} 1 & \text{if source verified} \\ 0 & \text{if otherwise} \end{cases}$$

$$x_V = \begin{cases} 1 & \text{if verified} \\ 0 & \text{otherwise} \end{cases}$$

2. Fit the **full model** in R (see the R code available in the Week 10 module). Write out the full fitted model.

$$\hat{y} = 4.915 + 2.575x_{SV} + 7.486x_V + 4.815x_O - 0.00005x_{TC} - 3.271x_P + 0.11x_t$$

Verified:

$$\hat{y} = 4.915 + 7.486 + 4.185x_D - 0.00005x_{TC} - 3271x_P + 0.11x_t$$

3. What is the adjusted R^2 value for the full model above? Interpret this value.

0.2579

About 26% of variability in the response is explained by the model when accounting for the number of explanatory variables used

The full model above uses many explanatory variables. Is it really necessary to include all five explanatory variables? One method of determining which variables should be included is through a process of **backward elimination.** Backward elimination starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time until only variables that have significant p-values for their individual t-tests are included.

4. Using an $\alpha = 0.1$ significance level, which explanatory variables should be kept in the model using backward elimination.

Backward elimination is not the only mechanism that can be used to determine the "best" model. We could also compare adjusted R^2 values for different possible models and choose the model with the highest adjusted R^2 value.

5. Using the R code, fit all possible models using the five available explanatory variables. Which model yields the highest adjusted R^2?

Full model (adjust $R^2 = 0.2579$)

6. Plot the residuals and comment on whether the conditions required for inference are met.

7. Consider some options for improving the model fit. What are some ways in which we could transform the data to possibly improve the model?