SLR

Lecture 1 (Jan 9 2024)

Most of this is a review of chapter 3 from STAT 2507.

Review: Line of Best Fit
When we perform a study, we care about a certain population. We can take a sample of size n and use this to find a line of best fit.

$$Y = a + bX$$

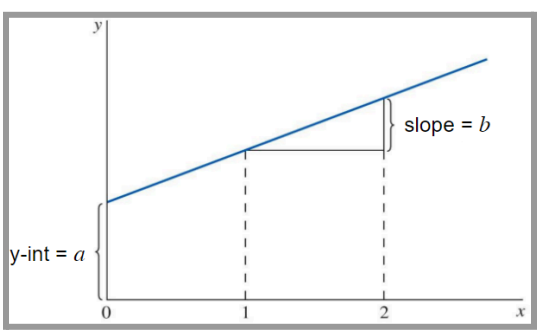
- X is the **independent** variable
- Y is the **dependant** variable (Y depends on X)

Q: Are we **only** interested in this fitted line from the sample?
A: No, we want to know what this sample tells us about the **population**
- Inference: making a conclusion about a population based on a sample

The Deterministic Model
In this model, every data point is **exactly on** the line graphed by the equation below.
- Thus, it is a "perfect model".
- Note that a **perfect** model will never exist in statistics, but it is possible mathematically.

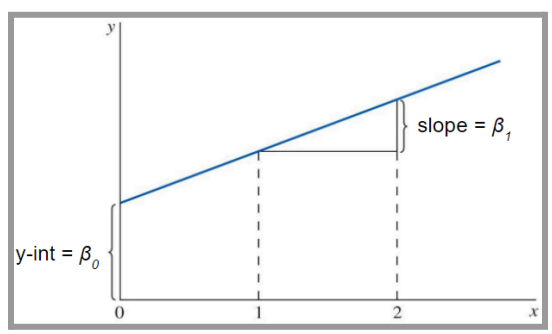
$$Y = \beta_0 + \beta_1 X$$

- We are introducing some new variables, β_0 and β_1, which replace a and b. These are represented by a lowercase beta letter, with a subscript of 0 or 1.
- X and Y still represent the independent and dependent variables.



LINE OF BEST FIT: $Y = a + bX$

VS.



DETERMINISTIC MODEL: $Y = \beta_0 + \beta_1 X$

Review: commenting on the nature of data

If we have two variables X and Y, the first step is to create a **scatter plot** of the data.

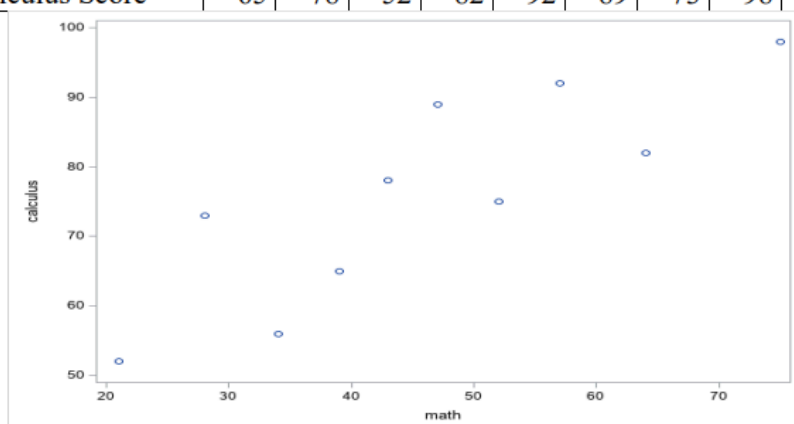
Then, we comment on:

- direction (positive or negative)
- nature of relationship (linear, quadratic, etc.)
- strength of relationship (weak, strong, etc.)
    - Note that strength is subjective, so direction and nature are more important

Exercise 1

Consider the mathematics achievement test scores, X, for a random sample of 10 Grade 12 students along with their final calculus grades, Y. Use SAS to produce a scatter plot for the bivariate data and comment on the nature of the relationship between Y and X.

Mathematics Achievement Test Score	39	43	21	64	57	47	28	75	34	52
Final Calculus Score	65	78	52	82	92	89	73	98	56	75



I would say that the relationship between Y and X is *positive* and *linear*. Everyone will have a different interpretation for the strength of the relationship; I think it is *moderate*.

The Probabilistic Model

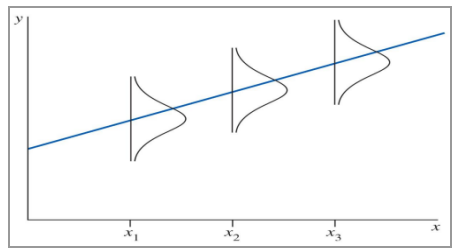
Realistically, not all points will lie exactly on any straight line we pick. This is why we use the probabilistic model, which adds an error component to account for deviations from the line.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- ε represents the **error term**, and is represented by a lowercase epsilon.
- The $\beta_0 + \beta_1 X$ part of the model is called the **line of means** since it describes the **average value** of Y at a given value of X.
- The error, ε, allows each individual response to deviate from this average by an amount.

The error terms are **assumed** to satisfy the following:

1. They are independent
2. They have a mean of 0
3. They have a constant variance of σ^2
4. They are normally distributed
    - This can be seen in the graph to the right



The Simple Linear Regression Model (SLR)

You may be asked to "state the simple linear regression model and its assumptions". This can be stated, with minimal writing, by writing both statements below:

① $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

② $i.\,i.\,d.$ ←$i.\,i.\,d.$ stands for **independent and identically distributed**

$\varepsilon_i \sim N(\mu = 0, \sigma^2 = \sigma^2)$

The above states that "the error terms are independently and identically distributed as normal with a mean of 0 and a constant variance of σ^2".

Least-Squares Method

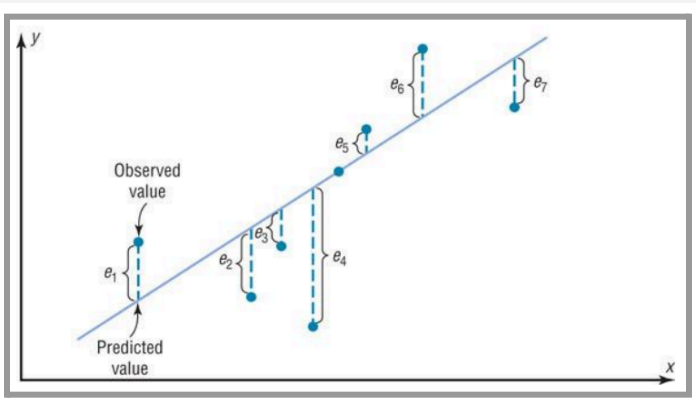
- We use this method to find the equation of the line of the SLR, aka the line of best fit.
- The line of best fit is the line that minimizes the **sum of squares of the deviations** of the observed values, Y, from those predicted.
- If we can't find the equation for the **population** ($y = \beta_0 + \beta_1 x + \varepsilon$), we find the equation for the **sample** ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) instead.

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

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates for β_0 and β_1.

The sum of squared deviations is called the sum of squares for errors (SSE).

$$SSE = \Sigma e_i^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$



Least-Squares Estimators of β_0 and β_1

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

where:

$$S_{xy} = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}$$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

<u>Exercise 2</u>

Method 1 (deriving line of best fit from SAS output)

Find the line of best fit for the application in Exercise 1.

Mathematics Achievement Test Score	39	43	21	64	57	47	28	75	34	52
Final Calculus Score	65	78	52	82	92	89	73	98	56	75

You should be able to look at the following output and derive the line of best fit from it.

this is an example output from SAS.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
		x-int			
Intercept	1	40.78416	8.50686	4.79	0.0014
mathscore	1	0.76556	0.17498	4.38	0.0024
		slope			

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

$$\hat{y} = 40.7842 + 0.7656$$

∴ The line of best fit for the application in Exercise 1 is $\hat{y} = 40.7842 + 0.7656x$.

Exercise 2 Cont…

Method 2 (calculating line of best fit by hand)

Find the line of best fit for the application in Exercise 1.

Mathematics Achievement Test Score	39	43	21	64	57	47	28	75	34	52
Final Calculus Score	65	78	52	82	92	89	73	98	56	75

- Recall that X represents the mathematics achievement test score, and Y represents the final calculus score.
- We can see that $n = 10$.

Step 1. Calculate $\hat{\beta}_1$.

- You need to calculate $\hat{\beta}_1$ first since you need $\hat{\beta}_1$ to calculate $\hat{\beta}_0$.
- It is recommended to calculate S_{xy} and S_{xx} on the side since they will be used in multiple calculations.
- $\hat{\beta}_1$ is the estimator, and b_1 is the estimate. We use $\hat{\beta}_1$ in the equation, but we switch to using b_1 once it takes on a value.

a. Calculate S_{xy}

$$S_{xy} = \sum_{i=1}^{10} x_i y_i - \frac{\sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{10} = (39)(65) + \ldots + (52)(75) - \frac{(39+\ldots+52)(65+\ldots+75)}{10}$$

$$= 36854 - \frac{(460)(760)}{10} = 36854 - 34960 = 1894$$

Note that $\sum_{i=1}^{10} x_i = 460$ and $\sum_{i=1}^{10} y_i = 760$ since we will use these values in multiple calculations.

b. Calculate S_{xx}

$$S_{xx} = \sum_{i=1}^{10} x_i^2 - \frac{(\sum_{i=1}^{10} x_i)^2}{10} = (39)^2 + \ldots + (52)^2 - \frac{(460)^2}{10}$$

$$= 23634 - \frac{211600}{10} = 23634 - 21160 = 2474$$

c. Sub values of S_{xy} and S_{xx} into equation to calculate $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \Rightarrow \quad b_1 = \frac{1894}{2474} = 0.7656$$

Step 2. Calculate $\hat{\beta}_0$.

- $\hat{\beta}_0$ is the estimator, and b_0 is the estimate. We use $\hat{\beta}_0$ in the equation, but we switch to using b_0 once it takes on a value.

a. **Calculate $\bar{y}$**

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{760}{10} = 76$$

b. **Calculate $\bar{x}$**

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{460}{10} = 46$$

c. **Sub values of $\bar{y}$, $\hat{\beta}_1$, and $\bar{x}$ into equation to calculate $\hat{\beta}_0$**

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

$$b_0 = 76 - (0.7656)(46) = 40.7842$$

Step 3. Find the line of best fit.

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

$$\hat{y} = 40.7842 + 0.7656x$$

∴ The line of best fit for the application in Exercise 1 is $\hat{y} = 40.7842 + 0.7656x$.

Q: What does this equation tell us?

A: $\hat{y} = 40.7842 + 0.7656x$

- **On average**, if a student's achievement test score (X) is 0, we expect the same student to get a final calculus score (Y) of 40.7842.
- An increase of one mark to a student's achievement test score (X) corresponds, **on average**, to an increase of 0.7656 marks on the same student's final calculus score (Y).

Analysis of Variance for Simple Linear Regression (ANOVA)

Recall that **regression** relates a dependent variable to one or more (explanatory) independent variables. We care about the dependant variable, Y, since it tells us if X is a factor that affects Y.

Total Variation / Total Sum of Squares (TSS)

All of the variation in variable Y. It can be computed using the following equation:

$$TSS = S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

It can be separated into **two different sources** of variation:

1. Sum of squares for regression (SSR): variation that **is** explained by X
2. Sum of squares for error (SSE): variation that **is not** explained by X

$$TSS = SSR + SSE$$

Sum of Squares for Regression (SSR)

Variation that is explained by the independent variable X.

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = \frac{(S_{xy})^2}{S_{xx}}$$

Sum of Squares for Error (SSE)

Variation that is not explained by X.

$$SSE = TSS - SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

Understanding the ANOVA Table

- SSR, SSE, and TSS have their own degrees of freedom, which are in the table below.
- The degrees of freedom are simply the terms in the denominators.
  - As an example/review: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$. t has $n - 1$ degrees of freedom because the denominator of s is $n - 1$. $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$.
- When SSR and SSE are divided by their own degree of freedom, the resulting estimates are called **mean squares**. $SSR \text{ -> } MSR$ and $SSE \text{ -> } MSE$
- MSE, also denoted as s^2, is an **unbiased estimator** for the constant variance, σ^2.
  - That is, $E(MSE) = E(s^2) = \sigma^2$ (the **expected value** of MSE is σ^2).
  - Do not mix up MSE / s^2 with the sample variance, which is also denoted as s^2.

ANOVA for simple regression is shown in the table below:

Source	degrees of freedom (df)	Sum of Squares (SS)	Mean Squares (MS)
Regression	1	$SSR = \frac{(S_{xy})^2}{S_{xx}}$	$MSR = \frac{SSR}{1}$
Error	$n - 2$	$SSE = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$	$MSE = \frac{SSE}{n-2}$
Total	$n - 1$	$TSS = S_{yy}$	

Lecture 2 (Jan 11 2024)

<u>Exercise 3</u>

Construct an ANOVA table for the application in Exercise 1.

As a reminder, we calculated the following values in Exercise 2:

$\sum_{i=1}^{10} y_i = 760$ $\sum_{i=1}^{10} x_i = 460$ $\sum_{i=1}^{10} y_i^2 = 59816$ $\sum_{i=1}^{10} x_i^2 = 23634$ $\sum_{i=1}^{10} x_i y_i = 36854$

$S_{xx} = 2474$ $S_{xy} = 1894$

We need to calculate TSS, SSR, and SSE. This can be done in any order.

$$TSS = S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 59816 - \frac{760^2}{10} = 2056$$

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474} = 1449.9741$$

$$SSE = TSS - SSR = 2056 - 1449.9741 = 606.0259$$

Now we can fill in the table:

Source	degrees of freedom (df)	Sum of Squares (SS)	Mean Squares (MS)
Regression	1	1449.9741	$\frac{1449.9741}{1} = 1449.9741$
Error	$n - 2 = 10 - 2 = 8$	606.0259	$\frac{606.0259}{8} = 75.7532$
Total	$n - 1 = 10 - 1 = 9$	2056	

Note: MSE estimates σ^2. Thus, $MSE = \sigma^2 = 75.7532$.

Testing how useful our model is

Is the model $Y = \beta_0 + \beta_1 X + \varepsilon$ useful in predicting Y?

- This can be answered by a **hypothesis test** or **confidence interval** for β_1.
- Recall that $\hat{\beta}_1$ is the least squares estimator for β_1.
- $\hat{\beta}_1$ is an unbiased estimator for β_1, so $E(\hat{\beta}_1) = \beta_1$.
- Standard error of $\hat{\beta}_1$ is $\sqrt{\frac{\sigma^2}{S_{xx}}}$.
- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$ due to our assumption that $\varepsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- Therefore, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2}$ since MSE has $n - 2$ degrees of freedom.

Conducting a hypothesis test for β_1 (t-test)

1. State your **hypotheses**. It will always be the same: $H_0: \beta_1 = 0$ and $H_a: \beta_1 \neq 0$.
2. Compute the **test statistic**. $t = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{MSE}{S_{xx}}}}$
   - When all assumptions of the SLR model are **valid**, the test statistic follows a t distribution with $n - 2$ degrees of freedom.
   - Note: positive t = positive relationship; negative t = negative relationship
3. Determine the **rejection region**.
   - If SAS gives us the p-value, always use the p-value.
     - Reject H_0 if p-value $< a$.
   - If we are calculating by hand, use the rejection value.
     - Reject H_0 if $t > t_{a/2, n-2}$, $t <- t_{a/2, n-2}$, or if $|t| > t_{a/2, n-2}$ (all equivalent).

Finding the Confidence Interval (CI) for β_1

The $(1 - a)100\%$ CI for β_1 is given by $\hat{\beta}_1 \pm t_{a/2, n-2} \sqrt{\frac{MSE}{S_{xx}}}$.

- We reject H_0 if the interval does not contain 0.
- We fail to reject H_0 if the interval contains 0.

Exercise 4

Consider Exercise 1. At $a = 0.05$, is there enough evidence to conclude that math achievement score is useful in predicting final calculus grade?

Step 1. State hypotheses.

$H_0: \beta_1 = 0, \quad H_a: \beta_1 \neq 0.$

Step 2. Compute test statistic.

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{0.7656 - 0}{\sqrt{\frac{75.7532}{2474}}} = 4.3752$$

Step 3. Determine rejection region.

$t_{0.05/2,\, n-2} = t_{0.025,\, 8} = 2.306$

Concluding Statement

Since $|4.3752| > 2.306$, at $a = 0.05$, we have enough evidence to reject H_0, accept H_a, and conclude that math achievement test scores are useful at predicting calculus final scores.

Exercise 5

Consider Exercise 1. Find the 95% confidence interval for β_1. Can this interval be used to test the hypotheses in Exercise 1? Explain.

- 95% CI $\Rightarrow a = 0.05$
- Yes, this interval can be used to test the hypotheses. We reject H_0 if the interval does not contain 0.

$$\hat{\beta}_1 \pm t_{0.05/2,\, 10-2}\sqrt{\frac{MSE}{S_{xx}}} = \hat{\beta}_1 \pm t_{0.025,\, 8}\sqrt{\frac{MSE}{S_{xx}}}$$

$$= 0.7656 \pm 2.306\sqrt{\frac{75.7532}{2474}} = 0.7656 \pm 0.4035 = (0.3621, 1.1691)$$

$\Uparrow$

Since the interval does not include 0, we reject H_0.

Concluding Statement

Therefore, we can say with 95% confidence that when math achievement test score increases by 1 mark, the mean calculus mark increases by anywhere from 0.3621 to 1.1691.

Difference between t-test and F-test

The hypotheses $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$ can also be tested using the F-test.

- The t-test and F-test test different hypotheses.
  - The t-test hypotheses are $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$.
  - The F-test hypotheses are $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ vs. $H_a: at\ least\ one\ \beta\ not\ 0$.
- In SLR, the t-test and F-test are equivalent. This is because these tests are equivalent when there is **one parameter**.
- The relationship between the two tests is that $F = t^2$.
- If you are told to "use ANOVA to test the model", do an F-test. ANOVA gives the F-value.

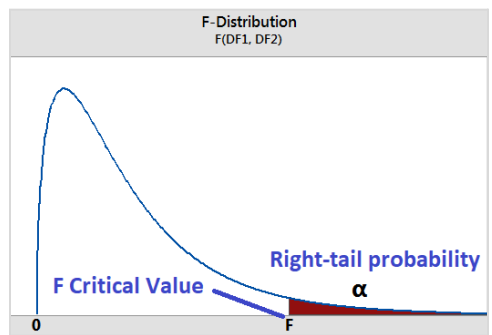
Conducting a hypothesis test for β_1 (F-test)

1. State your hypotheses. $H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$
2. Compute the test statistic. $F = \frac{MSR}{MSE}$
   - When all assumptions of the SLR model are valid, the test statistic follows an F distribution with 1 numerator degree of freedom (df_1) and $n - 2$ denominator degrees of freedom (df_2).
3. Determine the rejection region. Reject H_0 if $F > F_{a,\,1,\,n-2}$

The F distribution

The F-distribution is right-sided and always positive.

- Always positive since $F = \frac{MSR}{MSE}$ and both MSR and MSE are always positive since they're sources of variability
- $F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \Rightarrow F_{a,\,1,\,n-2}$
  - Always a one-sided test, so a not $a/2$.



Exercise 6

Test the hypotheses using the ANOVA table obtained in Exercise 4 using the F-test.

$H_0: \beta_1 = 0, \qquad H_a: \beta_1 \neq 0$

Method 1: using F value

- Reject H_0 if $F > F_{a,\,1,\,n-2}$

You can calculate the F value by using the formula…

Source	df	SS	MS	F
Regression	1	1449.9741	1449.9741	$F = \frac{MSR}{MSE} = \frac{1449.9741}{75.7532} = 19.1408$
Error	8	606.0259	75.7532	
Total	9	2056		

…or by squaring t….

$$t^2 = F$$

$(4.3752)^2 = 19.1424 \simeq 19.1408$ (calculated value)

…or by using SAS.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1449.97413	1449.97413	19.14	0.0024
Error	8	606.02587	75.75323		
Corrected Total	9	2056.00000			

Then you determine the rejection region.

$F_{a,\,1,\,n-2} = F_{0.05,\,1,\,8} = 5.32$

Since $19.1408 > 5.32$, at $a = 0.05$, we have enough evidence to reject H_0, accept H_a, and conclude that math achievement test scores are useful at predicting calculus final scores.

Exercise 6 Cont…

Method 2: using $Pr(> F)$ provided by SAS

- Reject H_0 if $(Pr > F) < a$

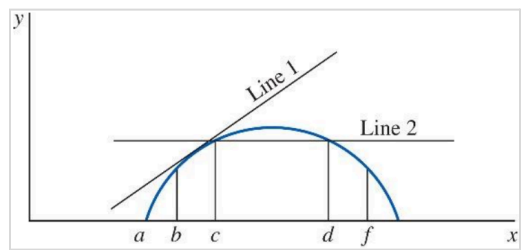
Table provided by SAS.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1449.97413	1449.97413	19.14	0.0024
Error	8	606.02587	75.75323		
Corrected Total	9	2056.00000			

Since $0.0024 < 0.05$, at $a = 0.05$, we have enough evidence to reject H_0, accept H_a, and conclude that math achievement test scores are useful at predicting calculus final scores.

Points to keep in mind



1. Your results are based on a sample. More investigation may be required.
2. It is possible that you fit the wrong model, leading to the wrong results.
   - When only looking at the data in the interval b to c, you may think that the relationship is linear. This is not the case.
   - When only looking at the data in the interval c to d, you may think that there is no relationship. This is not the case.
3. Do not extrapolate when the value of X is outside the range of X obtained from the sample.
   - If you assume a linear relationship from b to c and extrapolate values outside of that interval, your extrapolations will be inaccurate (shown by Line 1).
   - Line 1 is an valid estimation between b and c. Likewise, Line 2 is only valid between c and d.
4. Correlation does not imply causation. Finding a relationship between X and Y does not mean that X causes Y.
   - For example, there is no direct correlation between owning more TVs and living longer.

Lecture 3 (Jan 16 2024)

SLR Assumptions and how to test them

In order for inferences (i.e. hypothesis testing, CIs) to be valid, the following assumptions must hold. All assumptions are to be tested using graphical methods.

1. **The error terms are independent.**
   - This may be violated if the data is collected at regular time intervals. In such cases, we use time series methods. We will not do this in STAT 2509.
   - Ex: We are measuring the heights of students (Y). If we don't collect all of our data at the same time, the students sampled at later times will have had more time to grow, thus skewing the data.
2. **The error terms have a mean of 0.**
   - Plot of residuals vs. X
3. **The error terms have a constant variance σ^2.**
   - Plot of residuals vs. fitted value
4. **The error terms are normally distributed.**
   - Normal plot of residuals (QQ-Plot)
5. **The relationship between Y and X is linear (implicit assumption).**
   - Plot of residuals vs. X

What are residuals?

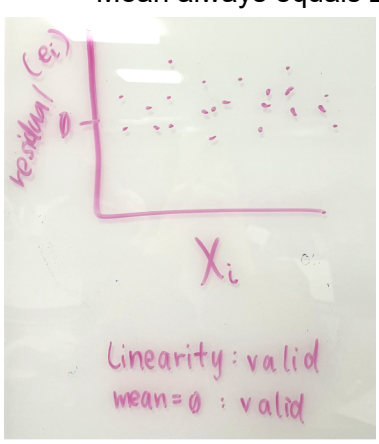
Residuals, denoted as e_i, are the estimators for error terms.

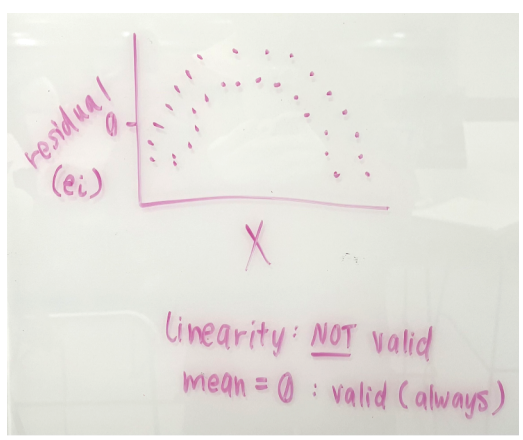
$$e_i = y_i - \hat{y}_i$$

Residual analysis is the process of analysing residuals to see if all of the assumptions are valid.

Plot of residuals vs. X (tests linearity and if mean equals zero)

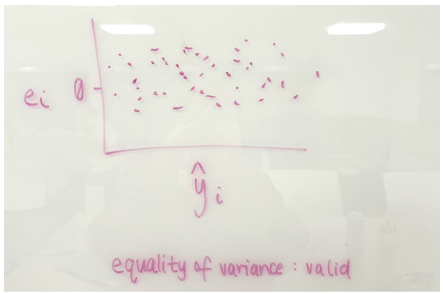
- Both assumptions are valid if there is a random scatter of points above and below 0
- Linearity assumption is violated if there is a visible pattern
- Mean always equals zero

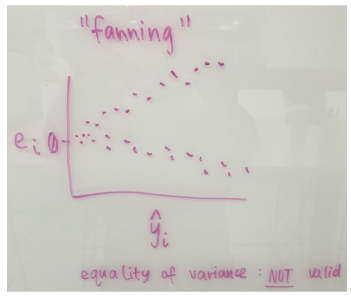


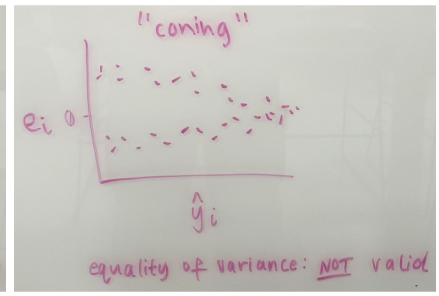


Plot of residuals vs. fitted value/predicted values (tests equality of variance)

- Equality of variance is valid if there is a random scatter of points above and below 0
- Equality of variance is violated if there is any coning or fanning

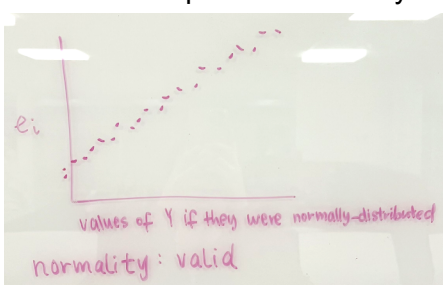


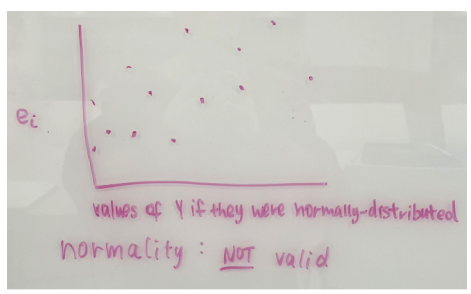




QQ-Plot/Normal plot of residuals (tests normality)

- Assumption of normality is valid if most points are close to the line

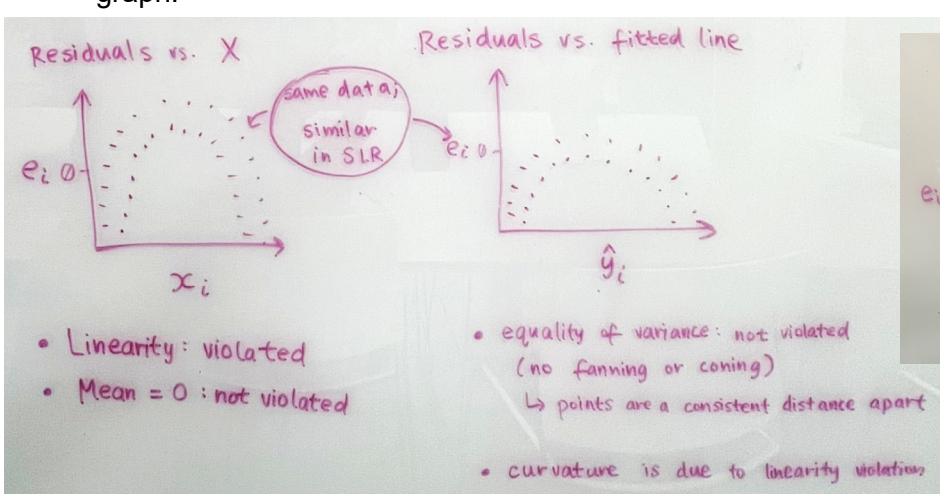


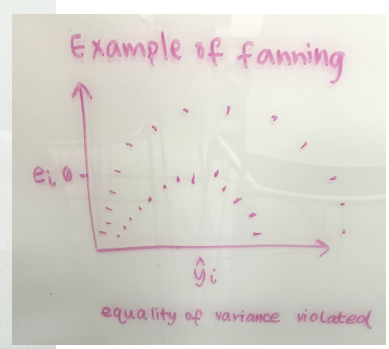


Some similarities between plots

$\hat{y}_i$ is just a weighted x_i with a vertical transformation by β_0: $\hat{y}_i = \beta_0 + \beta_1 x_i$.

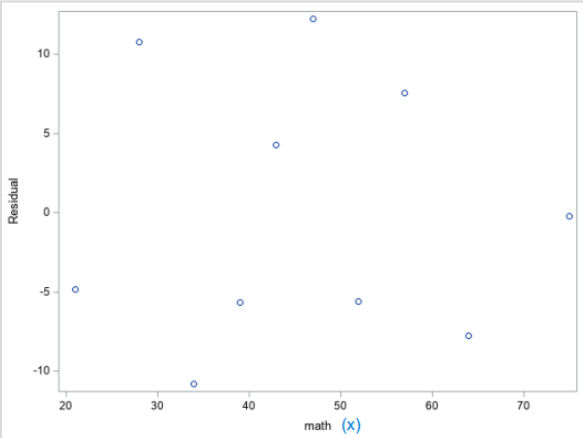
- This is why the Residuals vs. X graph looks similar to the Residuals vs. Fitted Value graph.



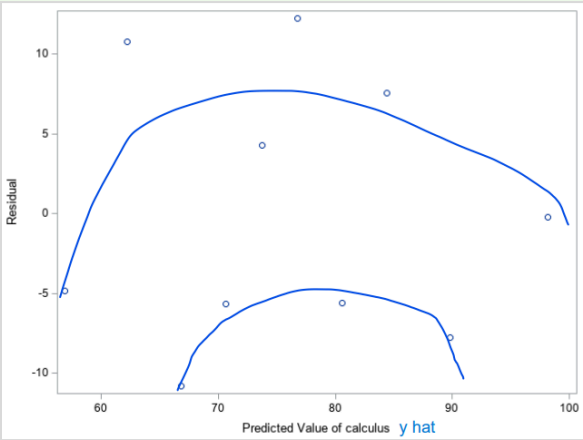


Exercise 7

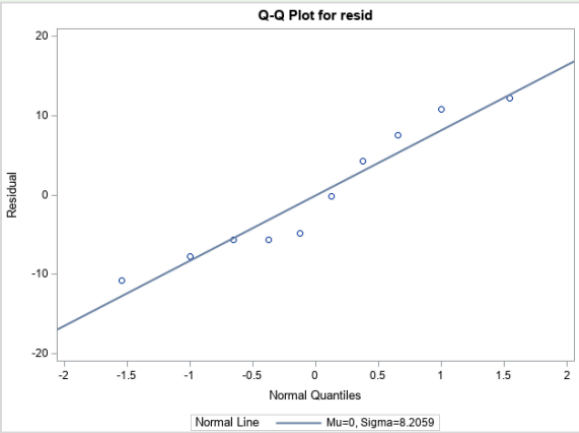
Test the assumptions for the math and calculus grades application.



Since there is a random scatter of points above and below zero, mean equals zero and linearity assumptions are valid.



Since there is a random scatter of points above and below zero (and no coning or fanning), equality of variance assumption is valid.



Since the points are close to the line, the normality assumption is valid/not violated.

Review: Correlation Coefficient

Recall that the correlation coefficient, r, measures the strength of the <u>linear</u> relationship between Y and X. $-1 \leq r \leq 1$, with -1 being a perfect negative relationship, 0 being no relationship, and 1 being a perfect positive relationship.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Coefficient of Determination

The coefficient of determination, r^2, measures the percent of variability in the dependent variable, Y, that is explained by the model (if multiple independent variables) or by X (SLR).

- Unlike the correlation coefficient r, the coefficient of determination measures the strength of relationship regardless of linearity.

$$R^2 = r^2 = (r)^2 = \frac{SSR}{TSS} = \frac{(S_{xy})^2}{S_{xx} S_{yy}}$$

- If you have your ANOVA table, use $R^2 = \frac{SSR}{TSS}$.
- Otherwise, calculate r and use $R^2 = \frac{(S_{xy})^2}{S_{xx} S_{yy}}$.

<u>Exercise 8</u>

Find the coefficient of determination for the math and calculus grades application.

$$R^2 = \frac{SSR}{TSS} = \frac{1449.9741}{2056} = 0.7052$$

Therefore, 70.52% of the variability in calculus marks is explained by math achievement test scores. (Note: don't say "explained by the model", since there is only one X in SLR.)

Bonus: what is the value of r? What can we conclude about the relationship between Y and X?

$$r = \pm\sqrt{R^2} = \pm\sqrt{0.7052} = \pm 0.8398$$

We need to use the regression equation to determine the direction of the relationship.

- If $\hat{b}_1 > 0$, then relationship is positive (positive slope)
- If $\hat{b}_1 < 0$, then relationship is negative (negative slope)

$\hat{y} = 40.7842 + 0.7656x$

$b_1 > 0$, so the relationship is positive. So, $r = 0.8398$.

Therefore, we can say that there is a strong, <u>positive</u>, <u>linear</u> relationship between calculus marks and math achievement test scores.

Prediction using the fitted line

- Once the fitted line is tested for all assumptions, it can be used for prediction.
- We want to create an interval containing the true <u>average</u> (CI) and the true <u>value</u> (PI).
- IMPORTANT!! Predictions can only be made for X values that are within the range of X covered by the sample's data points.
    - There must be least one X value greater and at least one X value less than x_0.

Confidence Interval (CI) for the average value of E(Y|X) when x=x$_0$

- Interval predicted to contain true **average**

$(1-a)100\%$ CI: $\hat{y} \pm t_{a/2,\, n-2}\sqrt{MSE(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})}$

Prediction Interval (PI) for the value of y when x=x$_0$

- Interval predicted to contain true **value**

$(1-a)100\%$ PI: $\hat{y} \pm t_{a/2,\, n-2}\sqrt{MSE(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})}$

<u>Exercise 9</u>

Using the fitted line for the math and calculus grades application, find the predicted value of the final calculus grade when the math achievement score is 55. Also, find the 95% confidence and prediction intervals.

$t_{a/2,\, n-2} = t_{0.05/2,\, 10-2} = t_{0.025,\, 8} = 2.306$

Check if $x_0 = 55$ is between the highest and lowest values of X in the sample

Mathematics Achievement Test Score	39	43	21	64	57	47	28	75	34	52
Final Calculus Score	65	78	52	82	92	89	73	98	56	75

There is at least one X value greater than x_0 (Ex: 57) and at least one X value less than x_0 (Ex: 39), so we can keep going.

Find predicted value by subbing x$_0$ into fitted line

Recall that the fitted line is $\hat{y} = 40.7842 + 0.7656x$.

$\hat{y}_{55} = 40.7842 + 0.7656(55) = 82.8922$ ← predicted value

Therefore, the predicted value of Y when $X = 55$ is 82.8922.

Calculate <u>shared term</u> between CI and PI to make finding both quicker

$$MSE\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right) = 75.7532\left(\frac{1}{10} + \frac{(55-\frac{460}{10})^2}{2474}\right) = 10.0555$$

Calculate CI for E(Y|X=55)

$$\hat{y}_{55} \pm t_{0.025,\,8}\sqrt{MSE\left(\frac{1}{10} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$= 82.8922 \pm 2.306\sqrt{10.0555}$

$= 82.8922 \pm 7.3124$ ← this is the "margin of error"

$= (75.5798,\ 90.2046)$

Therefore, we can say with 95% confidence that when the math achievement test score is 55, the mean calculus mark is anywhere from 75.58 to 90.20.

- You can round to 2 decimal places in the concluding statement

---

Calculate PI for Y when X = 55

$$\hat{y}_{55} \pm t_{0.025,\,8}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$= 82.8922 \pm 2.306\sqrt{MSE + 10.0555}$ ← note that we add another MSE!

$= 82.8922 \pm 2.306\sqrt{75.7532 + 10.0555}$

$= 82.8922 \pm 21.3612$ ← margin of error

$= (61.5292,\ 104.2516)$ ← don't need to cap the upper bound to 100 marks, ok to leave as is!

Therefore, we can say with 95% confidence that when the math achievement test score is 55, the calculus mark is anywhere from 61.53 to 104.25.

- Notice that this concluding statement is very similar to the one for the CI, just with different interval bounds and "calculus mark" instead of "mean calculus mark" (not an average!).

Lecture 4 (Jan 18 2024)

Recall that, in SLR:

total variability in Y → $TSS = SSR + SSE$ ← **remaining variability…**

↑ is it really an error, or did we choose the wrong model?

variability explained by X

So we introduce a new equation:

$SSE = SSPE + SSLF$ ← lack of fit (possibly due to wrong model)

↑

pure error (X can't explain variability)

Lack of Fit Test (for SLR)

- Tests how well the linear model fits. (is there a better model?)
- Only appropriate when at least one X value has multiple Y values.
- Note: if some/all SLR assumptions are violated, this could be an indication that the linear model does not fit.

$n = number\ of\ data\ points$
$m = number\ of\ unique\ x$s
Note 1: n always $\leq m$
Note 2: the number of data points per "row" may vary.

$SSLF = SSE - SSPE$ with $df_1 = m - 2$

$SSPE = \Sigma i\, \Sigma j (y_{ij} - \overline{y_i})^2$ with $df_2 = n - m$

X	Y	$Y - \overline{Y}$
5	10	$10 - 10 = 0$
6	11	$11 - 11 = 0$
7	12	$12 - 12 = 0$
8	13	$13 - 13 = 0$

A lack of fit test is <u>not appropriate</u> if each X has one Y. This is because the difference, $Y - \overline{Y}$, would always be 0, as shown on the left.

	x_i	y_{ij}	$\overline{y_i}$
$n_1 = 2$	—	1, 2	1.5
$n_2 = 3$	—	1, 3, 5	4.5
$n_3 = 1$	—	1	0
$n_4 = 1$	—	2	0
$n_5 = 1$	—	7	0

The test question may resemble something like the following. Meaning, not every X value will have that many Y values (this saves us a lot of time). Note: X values not written out.

Conducting a hypothesis test for a lack of fit test

1. State your hypotheses. This is the only time when we <u>do not want to reject</u> H_0.
    a. H_0: The model $Y = \beta_0 + \beta_1 X + \varepsilon$ is appropriate
    b. H_a: The model $Y = \beta_0 + \beta_1 X + \varepsilon$ is not appropriate
2. Compute the test statistic. $F = \frac{MSLF}{MSPE}$
3. Determine the rejection region. Reject H_0 if $F > F_{a, m-2, n-m}$
    - If using SAS, reject H_0 if $Pr(> F) > \alpha$.

Exercise 10

The manufacturer of a new brand of thermal panes examined the amount of heat loss by random assignment of three different panes to each outdoor temperature being considered. For each trial, the window temperature was controlled at 68°F and 50% relative humidity. The data are in the table below.

Outdoor Temperature (^{0}F)	Heat Loss
20	86, 80, 77
30	78, 84, 75
40	78, 69, 76
50	62, 53, 57
60	33, 38, 43

X: outdoor temperature; Y: heat loss

$m = 5$

$n = 15$

$n_1 = n_2 = n_3 = n_4 = n_5 = 3$ ← not always equal!

Another way to represent the table above is shown below. It could even be unsorted!

X	Y
20	86
20	80
20	77
30	78
…	…
60	43

A. Fit the simple linear regression model and test the assumptions.

x_i	y_{ij}	$\overline{y_i}$
20	86, 80, 77	$\overline{y_1} = \frac{86+80+77}{3} = 81$
30	78, 84, 75	$\overline{y_2} = \frac{78+84+75}{3} = 79$
40	78, 69, 76	$\overline{y_3} = \frac{78+69+76}{3} = 74.3333$
50	62, 53, 57	$\overline{y_4} = \frac{62+53+57}{3} = 57.3333$
60	33, 38, 43	$\overline{y_5} = \frac{33+38+43}{3} = 38$

Calculate SSPE

$SSPE = \Sigma i \, \Sigma j (y_{ij} - \overline{y}_i)^2$

m number of ys per X (replace number with n_i if not consistent)

$\downarrow \quad \downarrow$

$$= \sum_{i=1}^{5} \sum_{j=1}^{3} (y_{ij} - \overline{y}_i)^2$$

$$= (86 - 81)^2 + (80 - 81)^2 + (77 - 81)^2 + (78 - 79)^2 + \ldots + (33 - 38)^2$$

$$= 219.3333$$

Calculate SSLF

SSE (given by SAS) $= 709.3$

$$SSLF = SSE - SSPE$$
$$= 709.3 - 219.3333$$
$$= 489.9667$$

Source	df	SS	MS	F
Lack of fit	$m - 2$ $= 5 - 2 = 3$	489.9667	$\frac{489.9667}{3} = 163.3222$	$\frac{MSLF}{MSPE} = \frac{163.3222}{21.9333} = 7.4463$
Pure error	$n - m$ $= 15 - 5 = 10$	219.3333	$\frac{219.3333}{10} = 21.9333$	
Error	$n - 2$ or $df_1 + df_2$ $= 15 - 2 = 13$	709.3		

B. Perform a lack of fit test at $\alpha = 0.05$.

State hypothesis

H_0: The model $Y = \beta_0 + \beta_1 X + \varepsilon$ is appropriate

H_a: The model $Y = \beta_0 + \beta_1 X + \varepsilon$ is not appropriate

Determine rejection region

$F_{\alpha, m-2, n-m} = F_{0.05, 3, 10} = 3.71$

Reject H_0 if $F > F_{a, m-2, n-m}$.

Since $7.4463 > 3.71$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that the model $Y = \beta_0 + \beta_1 X + \varepsilon$ is not appropriate.

If **using SAS**, your conclusion will change to the following:

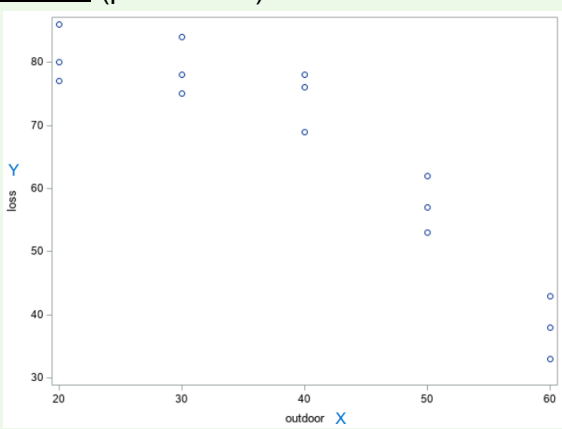
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3477.63333	3477.63333	63.74	<.0001
Error	13	709.30000	54.56154		
Lack of Fit	3	489.96667	163.32222	7.45	0.0066
Pure Error	10	219.33333	21.93333		
Corrected Total	14	4186.93333			

Since $0.0066 > 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that the model $Y = \beta_0 + \beta_1 X + \varepsilon$ is not appropriate.

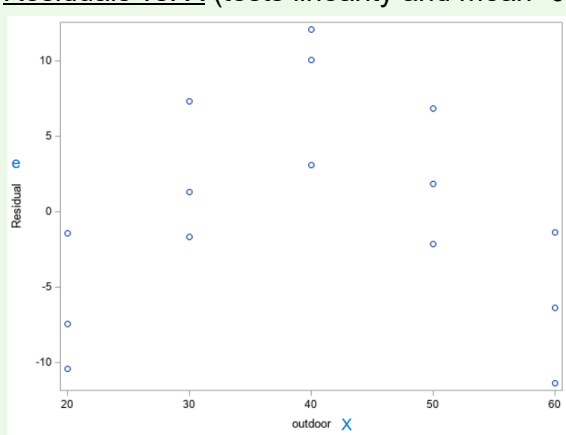
Now that we have concluded that a linear model is not appropriate, let's see which SLR assumptions were violated.

<u>Y vs. X</u> (plot of data)



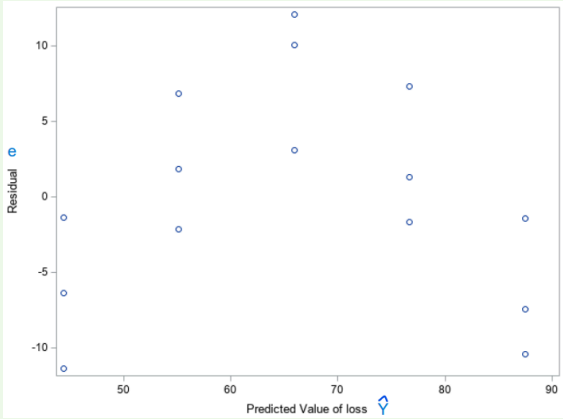
Looks sort of linear… but is it?

<u>Residuals vs. X</u> (tests linearity and mean=0)



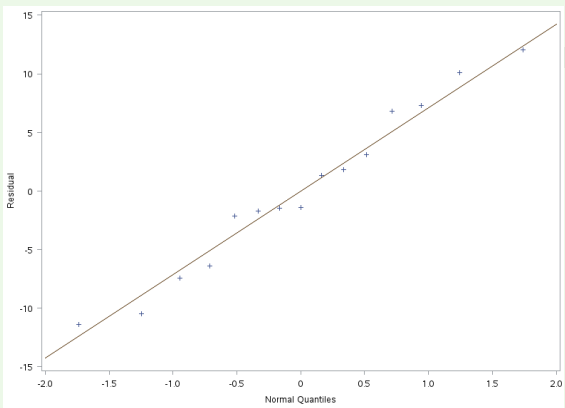
Linearity is violated (there is a curve)
Mean is always equal to 0

Residuals vs. Predicted Value of Y (tests equality of variance)



Equality of variance is valid (lines are equidistant)

Normal plot of residuals



Normality assumption valid

Bonus: calculate R^2 and interpret

$R^2 = \frac{SSR\ (from\ SAS)}{TSS\ (from\ SAS)} = \frac{3477.6333}{4186.9333} = 0.8306$ ← linear model is strong! but it's not the best.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
		SSR			
Model	1	3477.63333	3477.63333	63.74	<.0001
Error	13	709.30000	54.56154		
Lack of Fit	3	489.96667	163.32222	7.45	0.0066
Pure Error	10	219.33333	21.93333		
Corrected Total	14	4186.93333			
		TSS			

It was shown in class using SAS that $Y = X^2$ is an almost perfect model. We are not expected to be able to do this.

MLR - Part 1

Multiple Linear Regression Model

- Relationship between a dependent variable (Y) with <u>more than one</u> independent variable (X_i).
- In the form $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$
- The assumptions for this model are the same as SLR.
- If you are asked to "state the MLR model and its assumptions", write the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \text{ , } \varepsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Linear equations in statistics

In stats, models are linear if β_i are linear. This is because x_i are fixed, we know their value, and this simplifies the equation into a linear equation.

- $Y_i = \beta_0 + \beta_1 sin\, X_i + \beta_2 e^{X_2} + \varepsilon_i$ is linear.
- $Y_i = \beta_0 + cos\beta_1(X_1) + \varepsilon_i$ is not linear.

<u>Exercise 1</u>

Indicate whether each model is linear or nonlinear.

The following are other forms of multiple regression models

- *First Order Model*: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- *Interaction Model*: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- *Second Order Model*: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- *Complete Second Oder Model with Two Independent Variables:* $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \varepsilon$

They are all linear.

Lecture 5 (Jan 23 2024)

Review: Interpretation of β_1 in SLR

β_1 is how much the mean of Y increases with an increase of 1 in X.

Interpretation of B_j in MLR

β_j is how much Y increases with an increase of 1 in X_j, when all other independent variables in the model are fixed.

Testing assumptions in MLR Model

- Testing linearity
  - **Residuals vs. X** (one graph for each X)
  - The scatter plots for individual Xs may show a weak relationship
  - Same analysis as in SLR
- Testing equality of variance and mean equals zero
  - **Residuals vs. Predicted Values**
  - Equal variance and mean equal zero: random scatter of points above and below zero
- Testing normality
  - **QQ-Plot** / Histogram of Residuals
  - Same analysis as in SLR

<u>Exercise 2</u>

How do real estate agents decide on the asking price for a newly listed condo? A computer database in a small community contains the listed selling price, Y (in thousands of dollars), the amount of living area x_1 (in hundreds of square metres), and the numbers of floors x_2, bedrooms x_3, and bathrooms x_4, for 15 randomly selected condos currently on the market. The data is shown in the following table.

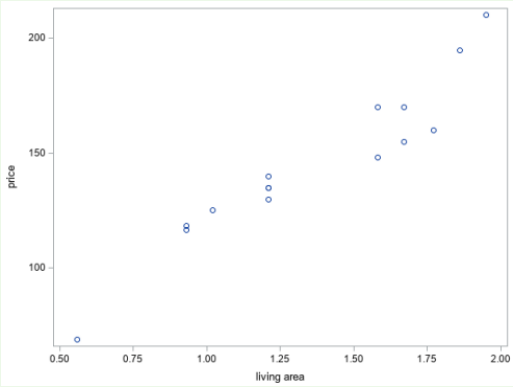
List Price, Y	Living Area, x_1	Floors, x_2	Bedrooms, x_3	Baths, $x4$
69	0.56	1	2	1
118.5	0.93	1	2	2
116.5	0.93	1	3	2
125	1.02	1	3	2
129.9	1.21	1	3	1.7
135	1.21	2	3	2.5
139.9	1.21	1	3	2
147.9	1.58	2	3	2.5
160	1.77	2	3	2
169.9	1.67	1	3	2
134.9	1.21	1	4	2
155	1.67	1	4	2
169.9	1.58	2	4	3
194.5	1.86	2	4	3
209.9	1.95	2	4	3

A. Give the MLR model and its assumptions for this application.

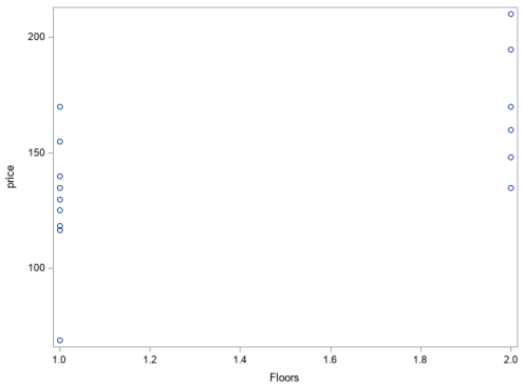
$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

B. Using SAS, test all assumptions.

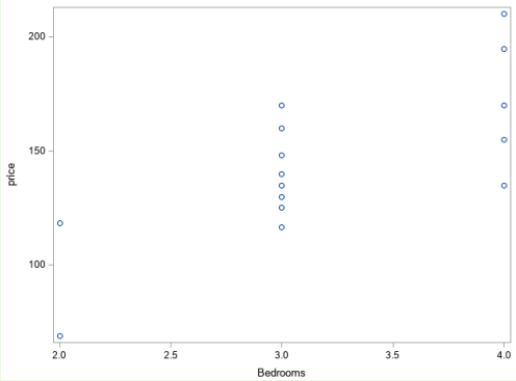
Price vs. living area
→ linear



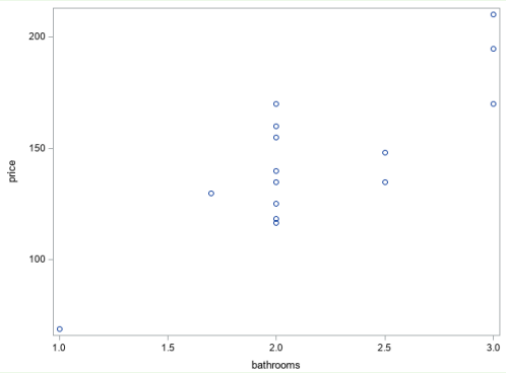
Price vs. floors
→ linear (discrete, 2 options)



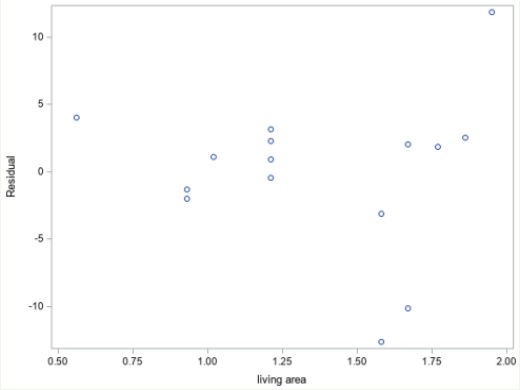
Price vs. bedrooms
→ linear (discrete, 3 options)



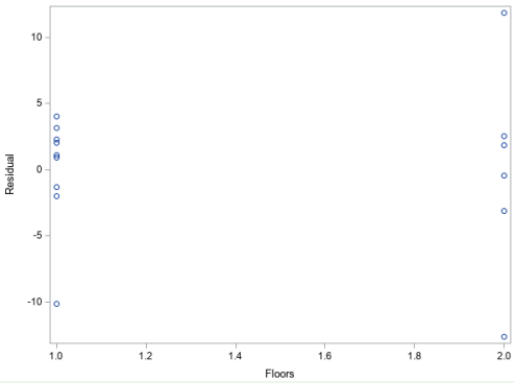
Price vs. bathrooms
→ linear



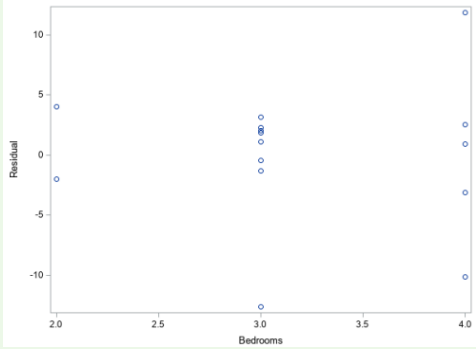
Residual vs. living area
→ linearity valid (points above and below 0, no pattern)



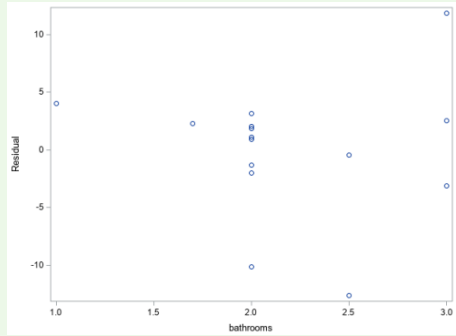
Residual vs. floors
→ linearity valid (points above and below 0, no pattern)



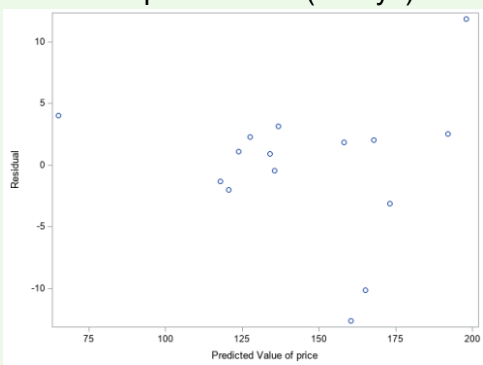
Residual vs. bedrooms
→ linearity valid (points above and below 0, no pattern)



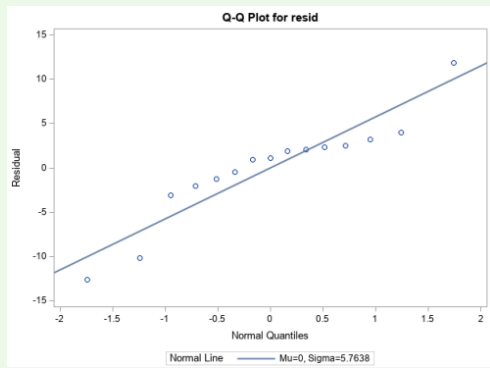
Residual vs. bathrooms
→ linearity valid (points above and below 0, no pattern)



Residual vs. predicted value of price
→ equality of variance valid (no fanning or coning)
→ mean equals 0 valid (always)



Q-Q plot for residual
→ normality assumption valid (most points near line)



All MLR assumptions are valid.

ANOVA for MLR

- TSS is same as in SLR: $TSS = S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$
  - This is the only one you can calculate by hand, since it only involves Ys.
- Formulas for SSR and SSE are different. We will get these values using SAS output.
- MSE is still an unbiased estimator for constant variance, σ^2.
- We need to be able to complete ANOVA table, given an incomplete table or SAS output.

Source	df	SS	MS	F
Regression	k: num. of Xs	SSR	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
Error	$n - k - 1$	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total	$n - 1$	TSS		

F-test for MLR (Whole Model)

- Tests the usefulness of the model (all Xs together). Not the same as testing usefulness of one X. In MLR, F-test and t-test are different.

1. State your hypothesis.
   a. H_0: All B_j's are equal to zero ($\beta_1 = \beta_2 = \ldots = \beta_k = 0$)
   b. H_α: At least $\beta_j \neq 0$ ($j = 1, 2, \ldots, k$)
2. Get test statistic, F, from SAS.
3. Determine rejection region.
   - Reject H_0 if $F > F_{\alpha, k, n-k-1}$ (conclude that the model is useful).
   - Reject H_0 if p-value $< \alpha$.

Q: Why can't B_j all equal zero?

A: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i$ would become $Y_i = \beta_0 + \varepsilon_i$. This model is useless.

t-test for MLR (Individual Variables)

- Tests the usefulness of one X_j individually. (Can X_j be removed from the model?)

1. State your hypothesis.
   a. H_0: $\beta_j = 0$
   b. H_α: $\beta_j \neq 0$ ($j = 1, 2, \ldots, k$)
2. Compute the test statistic. $t = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{B}_j)}$.
   - $\beta_j = 0$, as H_0 states.
   - $SE(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$.
3. Determine the rejection region. Reject H_0 if $|t| > t_{\alpha/2, n-k-1}$.

<u>Exercise 3</u>

Consider Exercise 2. Using the SAS output provided, answer the following questions.

Dependent Variable: price price

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	15917	3979.27075	85.56	<.0001
Error	10	465.09433	46.50943		
Corrected Total	14	16382			B)

Root MSE	6.81978	R-Square	0.9716
Dependent Mean	145.05333	Adj R-Sq	0.9603
Coeff Var	4.70157		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	18.59653	9.16469	2.03	0.0699
living_area	living area	1	67.67839	7.79014	8.69	<.0001
Floors	Floors	1	-16.50848	6.19755	-2.66	0.0237
Bedrooms	Bedrooms	1	-2.73037	4.47728	-0.61	0.5556
bathrooms	bathrooms	1	30.47877	6.81703	4.47	0.0012

A)

C) to F)

A. Give the fitted model.

$df = 4$, so $k = 4$ (there are 4 Xs)

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\beta}_4 X_{i4}$ ← Note: the subscript $i\ j$ means which variable (i) and which observation (j)

$\hat{Y}_i = 18.59653 + 67.67839X_{i1} - 16.50848X_{i2} - 2.73037X_{i3} + 30.47877X_{i4}$ is the fitted model.

B. At $\alpha = 0.1$, is the **<u>model</u> useful in predicting list price?**

↪ "model", so use F-test

<u>Hypothesis</u>

H_0: All B_j's are equal to zero ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$), H_α: At least one $\beta_j \neq 0$ ($j = 1, 2, 3, 4$)

<u>Using F-value</u>

F-value: 85.56

$F_{\alpha, k, n-k-1} = F_{0.1, 4, 15-4-1} = F_{0.1, 4, 10} = 2.61$

Reject H_0 if $F > F_{\alpha, k, n-k-1}$

Since $85.56 > 2.61$, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that the <u>model</u> is useful at predicting selling price.

↓

living area, floors, …, bathrooms

(don't need to state all Xs for MLR)

<u>Using p-value</u>

p-value: $<.0001$

$\alpha = 0.1$

Reject H_0 if p-value $< \alpha$

Since p-value < 0.1, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that the model is useful at predicting selling price.

C. At $\alpha = 0.1$, is <u>living area</u> useful in predicting list price?

↪ individual X, so use t-test

<u>Hypothesis</u>

$H_0: \beta_1 = 0$, $H_\alpha: \beta_1 \neq 0$

<u>Using rejection point (by hand)</u>	<u>Using p-value (SAS)</u>
t-value: 8.69 $t_{a/2,\, n-k-1} = t_{0.1/2,\, 15-4-1} = t_{0.05,\, 10} = 1.812$ ↑ same for parts C, D, E, F Reject H_0 if $\lvert t \rvert > t_{a/2,\, n-k-1}$ Since $\lvert 8.69 \rvert > 1.812$, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that <u>living area</u> is useful at predicting selling price. ↓ don't say "the model" for one X	p-value: $<.0001$ $\alpha = 0.1$ Reject H_0 if p-value $< \alpha$ Since p-value < 0.1, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that <u>living area</u> is useful at predicting selling price.

D. At $\alpha = 0.1$, is <u>number of floors</u> useful in predicting list price?

↪ individual X, so use t-test

<u>Hypothesis</u>

$H_0: \beta_2 = 0$, $H_\alpha: \beta_2 \neq 0$

<u>Using rejection point (by hand)</u>	<u>Using p-value (SAS)</u>
t-value: -2.66 $t_{a/2,\, n-k-1} = t_{0.1/2,\, 15-4-1} = t_{0.05,\, 10} = 1.812$ Reject H_0 if $\lvert t \rvert > t_{a/2,\, n-k-1}$ Since $\lvert -2.66 \rvert > 1.812$, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that <u>number of floors</u> is useful at predicting selling price.	p-value: 0.0237 $\alpha = 0.1$ Reject H_0 if p-value $< \alpha$ Since p-value < 0.1, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that <u>number of floors</u> is useful at predicting selling price.

E. At $\alpha = 0.1$, is <u>number of bedrooms</u> useful in predicting list price?

↪ individual X, so use t-test

<u>Hypothesis</u>

$H_0: \beta_3 = 0, H_\alpha: \beta_3 \neq 0$

<u>Using rejection point (by hand)</u>	<u>Using p-value (SAS)</u>
t-value: -0.61 $t_{a/2, n-k-1} = t_{0.1/2, 15-4-1} = t_{0.05, 10} = 1.812$ Reject H_0 if $\lvert t \rvert > t_{a/2, n-k-1}$ Since $\lvert -0.61 \rvert < 1.812$, at $\alpha = 0.1$, we have **do not** have enough evidence to reject H_0. Therefore, we **can't** conclude that <u>number of bedrooms</u> is useful at predicting selling price.	p-value: 0.5556 $\alpha = 0.1$ Reject H_0 if p-value $< \alpha$ Since $0.5556 > 0.1$, at $\alpha = 0.1$, we **do not** have enough evidence to reject H_0. Therefore, we **can't** conclude that <u>number of bedrooms</u> is useful at predicting selling price.

F. At $\alpha = 0.1$, is **<u>number of bathrooms</u> useful in predicting list price?**

↪ individual X, so use t-test

<u>Hypothesis</u>

$H_0: \beta_4 = 0, H_\alpha: \beta_4 \neq 0$

<u>Using rejection point (by hand)</u>	<u>Using p-value (SAS)</u>
t-value: 4.47 $t_{a/2, n-k-1} = t_{0.1/2, 15-4-1} = t_{0.05, 10} = 1.812$ Reject H_0 if $\lvert t \rvert > t_{a/2, n-k-1}$ Since $\lvert 4.47 \rvert > 1.812$, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that <u>number of bathrooms</u> is useful at predicting selling price.	p-value: 0.0012 $\alpha = 0.1$ Reject H_0 if p-value $< \alpha$ Since p-value < 0.1, at $\alpha = 0.1$, we have enough evidence to reject H_0, accept H_α, and conclude that <u>number of bathrooms</u> is useful at predicting selling price.

Intuition Test

If we were testing $a = 0.01$ instead, then we will make a different conclusion about the usefulness of individual Xs:

- living area is useful
- number of floors is **not** useful
- number of bedrooms is **not** useful
- number of bathrooms is useful

Q: Can we remove both number of floors and number of bedrooms simultaneously?

A: No, we need to remove them one at a time. We will learn more about this later.

Lecture 6 (Jan 25 2024)

Re: Assignment 1, question 3 (g)
This question asks you if the t-test is consistent with the confidence interval. We will use some fake data here to demonstrate what you should write.

H_0: $\beta_1 = 0$, H_α: $\beta_1 \neq 0$

<u>Confidence interval</u>
95% CI for β_1: $(7.6353,\ 9.3776)$ ← hypothesized value falls outside CI, so reject H_0

<u>t-test</u>
$t = 6.3156$, and $6.3156 > t_{a/2,\, n-2}$ ← reject H_0

Since the 95% CI for β_1 doesn't contain 0, H_0 is rejected which agrees with the t-test.

Confidence interval for β_j

- The $(1 - \alpha) \cdot 100$ CI for β_j is given by $\hat{\beta}_j \pm t_{a/2,\, n-k-1}\, SE(\hat{\beta}_j)$.
- This will not be a big focus of the class.

Coefficient of Determination, R^2, in MLR

- The percentage of variability in Y that's explained by the <u>model</u>.
- Computed using the same formulas used in SLR: $R^2 = \frac{SSR}{TSS}$.

The Adjusted Value of R^2

- The percentage of variability in Y that's explained by the model, adjusted by the number of Xs in the model.
- Computed using $R_{adj}^{\ 2} = 1 - \frac{SSE\ /\ (n-k-1)}{TSS\ /\ (n-1)}$.

Recall that $TSS = \sum_{i=1}^{n} y_i^{\ 2} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$. As you can see from the formula, TSS depends solely on the dependent variable (Y). However, as more independent variables are added to the model, SSR increases and R^2 also increases as a result. $R_{adj}^{\ 2}$ accounts for the number of independent variables in the model by factoring in the degrees of freedom.

As more Xs in model: $TSS = SSR \uparrow + SSE \downarrow$

Exercise 4

Find R^2 and R_{adj}^2 for the applications in Exercise 2.

$R^2 = \frac{SSR}{TSS} = \frac{15917}{16382} = 0.9716$

Therefore, 97% of the variability in selling price is explained by the model.

$R_{adj}^2 = 1 - \frac{SSE\,/\,(n-k-1)}{TSS\,/\,(n-1)} = 1 - \frac{465.09433\,/\,(15-4-1)}{16382\,/\,(15-1)} = 1 - \frac{465.09433\,/\,10}{16382\,/\,14} = 1 - 0.03975 = 0.9603$

Therefore, 96% of the variability in selling price is explained by the model, adjusted by the number of Xs in the model.

MLR - Part 2

Partial F-test for MLR (some Xs)

- Recall that a t-test for MLR tests whether x_i can be removed from the model.
- We use a partial F-test to test whether a set of independent variables can be removed from the model <u>simultaneously</u>.

1. State hypothesis. We are asking if the full model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$ can be reduced to the model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_g X_g + \varepsilon$.
   a. $H_0: \beta_{g+1} = \beta_{g+2} = \ldots = \beta_k = 0$.
   b. H_α: at least one of $\beta_{g+1}, \beta_{g+2}, \ldots, \beta_k$ is different from zero.
2. Compute F-statistic. We need either SSE or SSR for both the full and reduced models.
   a. Method 1: using SSE

$$F_{part} = \frac{MS_{drop}}{MSE_f} = \frac{\frac{SSR_f - SSR_r}{df_{SSR_f} - df_{SSR_r}}}{\frac{SSE_f}{df_{SSE_f}}} \sim F_{a,\,k-g,\,n-k-1}$$

   b. Method 2: using SSR

$$F_{drop} = \frac{MS_{drop}}{MSE_f} = \frac{\frac{SSE_r - SSE_f}{df_{SSE_r} - df_{SSE_f}}}{\frac{SSE_f}{df_{SSE_f}}} \sim F_{a,\,k-g,\,n-k-1}$$

3. Test rejection region.
   a. Using SSE: Reject H_0 if $F_{part} > F_{a,\,k-g,\,n-k-1}$.
   b. Using SSR: Reject H_0 if $F_{drop} > F_{a,\,k-g,\,n-k-1}$.

Exercise 1

At $\alpha = 0.01$, is there evidence to conclude that number of floors and number of bedrooms are not useful in the model? → can we remove X_2 and X_3?

List Price, Y	Living Area, x_1	Floors, x_2	Bedrooms, x_3	Baths, $x4$
69	0.56	1	2	1
118.5	0.93	1	2	2
116.5	0.93	1	3	2
125	1.02	1	3	2
129.9	1.21	1	3	1.7
135	1.21	2	3	2.5
139.9	1.21	1	3	2
147.9	1.58	2	3	2.5
160	1.77	2	3	2
169.9	1.67	1	3	2
134.9	1.21	1	4	2
155	1.67	1	4	2
169.9	1.58	2	4	3
194.5	1.86	2	4	3
209.9	1.95	2	4	3

Full Model:

The REG Procedure
Model: MODEL1
Dependent Variable: price price

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	15917	3979.27075	85.56	<.0001
Error	10	465.09433	46.50943		
Corrected Total	14	16382			

# of Xs --> full model

Reduced Model

The REG Procedure
Model: MODEL1
Dependent Variable: price price

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15573	7786.31965	115.42	<.0001
Error	12	809.53804	67.46150		
Corrected Total	14	16382			

2 =/= # of Xs --> reduced

Full model: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i$ → $k = 4$ (number of Xs in full model)

Reduced model: $Y_i = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon_i$ → $g = 2$ (number of Xs in reduced model)

State hypothesis

$H_0: \beta_2 = \beta_3 = 0$, H_α: At least one $\beta_j \neq 0$ $(j = 2, 3)$

Find F-statistic

$$F_{part} = \frac{(SSR_F - SSR_r) / (k-g)}{SSE_F / (n-k-1)} = \frac{(15917-15573) / (4-2)}{465.09433 / (15-4-1)} = \frac{344 / 2}{465.09433 / 10} = 3.6982$$

Find rejection point

$$F_{a,\, k-g,\, n-k-1} = F_{0.01,\, 4-2,\, 15-4-1} = F_{0.01,\, 2,\, 10} = 7.56$$

Concluding statement

Since $3.6982 < 7.56$, at $\alpha = 0.01$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that number of floors and number of bedrooms are **not** useful in the model. (We also **can't** conclude the number of floors and number of bedrooms **are not** useful in the model.)

Lecture 7 (Jan 30 2024)

Review: difference between F-test, t-test, and partial F-test

- F-test: testing whole model (all Xs)
- t-test: testing individual X (one X)
- Partial F-test: testing partial model (some but not all Xs)

Introducing qualitative independent variables

- The linear regression models that we have seen so far have had <u>quantitative</u> dependent (Y) and independent (X) variables.
  - These are numerical, measurable values.
  - We have been representing quantitative independent variables using X_j.
- However, independent variables can also be <u>qualitative</u>.
  - These are descriptive values that can't be expressed numerically.
  - We will be using D_j to represent qualitative variables.

Indicator ("Dummy") Variables

- We use indicator variables to represent qualitative values.
- The number of indicator variables that we create is $number\ of\ levels - 1$.
- Each indicator variable follows this coding: $x_j = \begin{cases} 1\ if\ in\ level/category \\ 0\ otherwise \end{cases}$
- Using indicator variables results in $[number\ of\ levels]$ **parallel** lines.

Example with indicator variables

If you have 3 categories (low, medium, high), you make $3 - 1 = 2$ dummy variables (D_1, D_2).

$$D_1 = \begin{cases} 1\ if\ low \\ 0\ otherwise \end{cases}, \quad D_2 = \begin{cases} 1\ if\ medium \\ 0\ otherwise \end{cases}$$

<u>How do 2 indicator variables represent 3 levels?</u>

Level	D_1	D_2
Low	1	0
Medium	0	1
High	0	0

<u>Why do we use 0 and 1?</u>

When you multiply a value by 1, the value is unchanged.

- Ex: $Ahmed \times 1 = Ahmed$

When you multiply a value by 0, it makes the value "disappear".

- Ex: $Ahmed \times 0 = 0$

We don't use another value since it would result in a changed value.

- Ex: $Ahmed \times 2 = 2Ahmed \rightarrow$ this is too many $Ahmed$s

Our model with indicator variables is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \varepsilon$

If low ($D_1 = 1,\ D_2 = 0$):

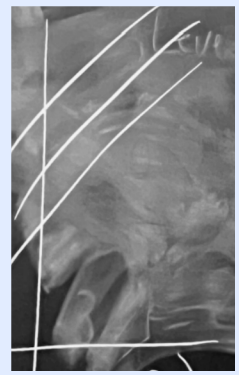
$Y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(0) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 + \varepsilon = (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon$

If medium ($D_1 = 0,\ D_2 = 1$):

$Y = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(1) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_3 + \varepsilon = (\beta_0 + \beta_3) + \beta_1 X_1 + \varepsilon$

If high ($D_1 = 0,\ D_2 = 0$):

$Y = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(0) + \varepsilon = \beta_0 + \beta_1 X_1 + \varepsilon = (\beta_0) + \beta_1 X_1 + \varepsilon$



- Notice that all of these lines have the same slope (β_1) and have different x-intercepts. These lines are parallel.
- Using indicator variables allows for the mean response to differ by levels.
- In a scatter plot, this example would show up as <u>3 parallel lines</u>.

Adding "Interaction Terms" to the Model

- Interaction terms in the model are created when **multiplying** two independent variables.
- In this course, we will focus on interaction between one <u>qualitative</u> independent variable and one <u>quantitative</u> independent variable.
- If your model includes interaction, the $[number\ of\ levels]$ lines **will not be parallel**.

Is a qualitative variable or interaction significant?

If we are dealing with two or more dummy variables, then the following applies:

1. If all p-values $> \alpha$, do a partial F-test.
    - This means that they're not significant individually. But, they may be significant together.
2. If at least one p-value $< \alpha$, conclude that the qualitative variable or interaction is significant.
3. If interaction is significant, the variables creating it are kept in the model.

Analyzing Scatter Plots for Interaction

If you are asked to "create a scatter plot and comment on if there is interaction", look out for <u>parallel lines with the same pattern</u> (this indicates that there is no interaction). You can say something like "since the $[number\ of\ lines]$ lines seem parallel, there is no interaction between $[quantitative\ variable]$ and $[qualitative\ variable]$".

Example with indicator variables and interaction

Assume that your model includes one quantitative variable (X_1) and one qualitative variable with 2 categories. We will create $2 - 1 = 1$ indicator variable (D_1) where $D_1 = \begin{cases} 1 \; if \; level \; 1 \\ 0 \; otherwise \end{cases}$.

<u>How does 1 indicator variable represent 2 categories?</u>

Category	D_1
Category 1	1
Category 2	0

Our model with interaction is: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 X_1 D_1$

↑ with expected value, no error term

↑ the interaction term, created by multiplying X_1 and D_1

<u>How does this affect the models for each category?</u>

Category 1 ($D_1 = 1$):

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3 X_1(1) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1$

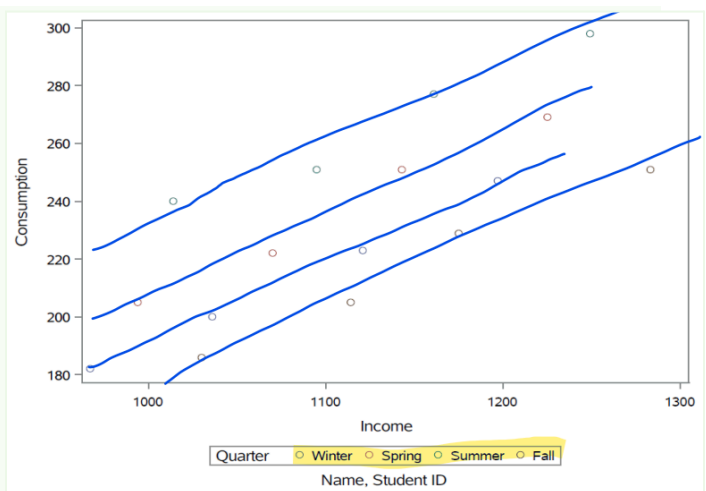
Category 2 ($D_1 = 0$):

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3 X_1(0) = \beta_0 + \beta_1 X_1$

- Notice that these lines have different slopes and different x-intercepts. These lines are <u>not</u> parallel.

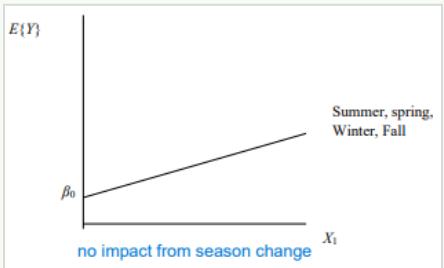
<u>Exercise 2</u>

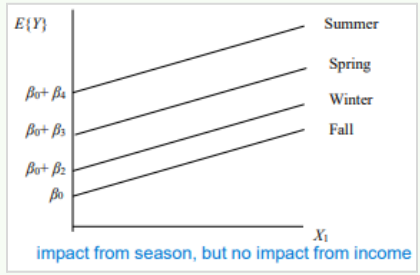
In order to establish a relationship between quarterly amount of beer purchased, Y, in millions of litres, and quarterly per capita disposable income, X_1, in dollars, data was collected over $n = 16$ consecutive quarters.

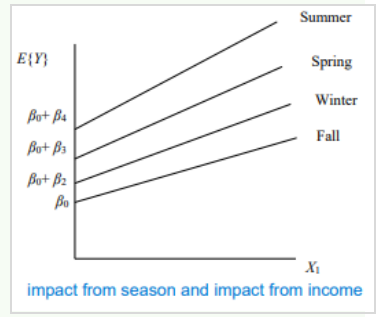


each season is represented by a different colour

Notice how adding indicator variables and interaction changes the model.







A. At $\alpha = 0.05$, is interaction between income and quarter useful in the model?

<u>State model (with interaction)</u>

4 seasons $\Rightarrow 4 - 1 = 3$ dummy variables (X_2, X_3, X_4)

X_1: $income$ $\quad X_2 = \begin{cases} 1 \; if \; winter \\ 0 \; otherwise \end{cases}$ $\quad X_3 = \begin{cases} 1 \; if \; spring \\ 0 \; otherwise \end{cases}$ $\quad X_4 = \begin{cases} 1 \; if \; summer \\ 0 \; otherwise \end{cases}$

Full model (FM): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \varepsilon$

($k = 7$) $\quad$ income ↩ $\quad$ different seasons ↩ $\quad$ ↪ interaction terms

Reduced mode (RM): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

($g = 4$)

Note: k and g are the number of terms in their respective models (not including β_0). $k - g$ is the number of terms we "kick out" of the full model to get the reduced model. In this exercise, $k - g = 7 - 4 = 3$ since we removed the 3 interaction terms to get the reduced model.

<u>State hypothesis</u>

$H_0: \beta_5 = \beta_6 = \beta_7 = 0$, H_α: At least one $\beta_j \neq 0$ ($j = 5, 6, 7$)

All p-values $> \alpha$, so we will perform a partial F-test.

FM:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-85.78322	25.77751	-3.33	0.0104
Income	Income	1	0.26383	0.02233	11.81	<.0001
x2		1	-5.19789	36.42715	-0.14	0.8901
x3		1	3.50290	37.12400	0.09	0.9271
x4		1	60.75501	37.35254	1.63	0.1425
income_x2		1	0.01757	0.03260	0.54	0.6046
income_x3		1	0.02411	0.03281	0.73	0.4835
income_x4		1	-0.00578	0.03268	-0.18	0.8640

Calculate F-statistic

FM:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	16238	2319.75571	136.75	<.0001
Error	8	135.71003	16.96375		
Corrected Total	15	16374			

RM:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	16220	4055.03321	289.90	<.0001
Error	11	153.86717	13.98792		
Corrected Total	15	16374			

$$F_{part} = \frac{(SSR_F - SSR_r) / (k-g)}{SSE_F / (n-k-1)} = \frac{(16238-16220) / (7-4)}{135.71003 / (16-7-1)} = 0.3537$$

Test rejection region

$$F_{a,\, k-g,\, n-k-1} = F_{0.05,\, 7-4,\, 16-7-1} = F_{0.05,\, 3,\, 8} = 4.07$$

Since $0.3537 < 4.07$, at $\alpha = 0.05$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that interaction between income and quarter (season) is significant.

Recall that the graph of Consumption vs. Income above contains four parallel lines. Our Partial F-test and our graph match, since the presence of parallel lines in the graph signifies that there is no interaction (which is what we concluded from the F-test).

Lecture 8 (Feb 1 2024)

B. At $\alpha = 0.05$, is quarter useful in the model?

We concluded in Part A that there is not enough evidence to conclude that interaction is significant. Therefore, we will take the interaction terms out of our full model and only test the terms for the different seasons.

(New) FM: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

$(k = 4)$ income ↩ different seasons ↩

(New) RM: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

$(g = 1)$ income ↩

State hypothesis

$H_0: \beta_2 = \beta_3 = \beta_4 = 0$, H_α: At least one $\beta_j \neq 0$ $(j = 2, 3, 4)$

Note: from the graph of Consumption vs. Income above, it is reasonable to hypothesize that quarter does matter (since there are 4 visible lines). If quarter didn't matter, then the data points would be scattered everywhere with no noticeable patterns.

Method 1: notice that at least one p-value $<$ α (fastest method)

FM:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-95.71423	12.39890	-7.72	<.0001
Income	Income	1	0.27246	0.01065	25.57	<.0001
x2		1	14.39025	2.74847	5.24	0.0003
x3		1	30.57951	2.68309	11.40	<.0001
x4		1	54.40353	2.65383	20.50	<.0001

Since all p-values < 0.05, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that quarter (season) is useful at predicting beer consumption.

Method 2: perform a partial F-test (not necessary if you do method 1)

Calculate F-statistic

FM:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	16220	4055.03321	289.90	<.0001
Error	11	153.86717	13.98792		
Corrected Total	15	16374			

RM:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9689.26544	9689.26544	20.29	0.0005
Error	14	6684.73456	477.48104		
Corrected Total	15	16374			

$$F_{part} = \frac{(SSR_F - SSR_r) / (k-g)}{SSE_F / (n-k-1)} = \frac{(16220 - 9689.26544) / (4-1)}{153.86717 / (16-4-1)} = 155.6279$$

Test rejection region

$$F_{a,\, k-g,\, n-k-1} = F_{0.05,\, 4-1,\, 16-4-1} = F_{0.05,\, 3,\, 11} = 3.59$$

Since $155.6279 > 3.59$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that quarter (season) is useful at predicting beer consumption.

C. Based on A and B, which model do you choose?

Model 1 is not good since it contains insignificant variables and interaction.

We can't conclude that model 2 is good just because Model 1 is bad. We still need to justify why Model 2 is a good model.

We have already done some of the necessary tests:

- We have tested X_2, X_3, X_4 (quarter) and concluded that it is significant (in Part B).

There are two things we still have to do:

- Test the whole model.
- Test if X_1 (per capita disposable income) is significant.

Testing the whole model (F-test)

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, H_α: At least one $\beta_j \neq 0$ $(j = 1, 2, 3, 4)$

FM:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	16220	4055.03321	289.90	<.0001
Error	11	153.86717	13.98792		
Corrected Total	15	16374			

Since p-value $< .001 < 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that Model 2 is useful at predicting beer consumption.

Testing if income is significant (t-test)

$H_0: \beta_1 = 0$, $H_\alpha: \beta_1 \neq 0$

FM:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-95.71423	12.39890	-7.72	<.0001
Income	Income	1	0.27246	0.01065	25.57	<.0001
x2		1	14.39025	2.74847	5.24	0.0003
x3		1	30.57951	2.68309	11.40	<.0001
x4		1	54.40353	2.65383	20.50	<.0001

Since p-value $< .001 < 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that per capita disposable income is useful at predicting beer consumption.

Since Model 2 is a good model (and Model 1 is not), I choose Model 2.

Test 1 notes:

- Dummy variables will not be part of Test 1.
- Partial F-test may or may not be included in Test 1.

MLR - Part 3

Multicollinearity

Sometimes individual variables are correlated. Let's see how this affects our regression model.

If $X_2 = 2X_1$ then:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 (2X_1) + \varepsilon$

$Y = \beta_0 + (\beta_1 + 2\beta_2) X_1 + \varepsilon$

- Multicollinearity increases variability, which widens confidence and prediction intervals. This is NOT GOOD.
- If two independent variables are highly correlated, there is no reason to collect both variables (one is enough in the model).
- We investigate multicollinearity using the variance inflation factor (VIF).

Variance Inflation Factor (VIF)

If the largest VIF is higher than 10, multicollinearity is an issue.

$VIF_j = \frac{1}{1 - R_j^2}$

- R_j^2 is the coefficient of determination for a regression model that relates X_j to the rest of the independent variables.

Example: testing if X_1, X_2, X_3 are correlated

Our coefficients of determination are R_1^2, R_2^2, R_3^2.

<u>R_1^2</u>

For X_1: $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$ $\rightarrow$ $VIF_1 = \frac{1}{1 - R_1^2}$

<u>R_2^2</u>

For X_2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$ $\rightarrow$ $VIF_2 = \frac{1}{1 - R_2^2}$

<u>R_3^2</u>

For X_3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ $\rightarrow$ $VIF_3 = \frac{1}{1 - R_3^2}$

Exercise 1

Data set is on Brightspace.

Use SAS to obtain the variance inflation factor and comment on what you obtain.

Parameter Estimates

	Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
	Intercept	Intercept	1	-1507.81373	778.63494	-1.94	0.0707	0
X1	Time	Time	1	2.00957	1.93065	1.04	0.3134	3.34262
	MktPoten	MktPoten	1	0.03720	0.00820	4.54	0.0003	1.97762
	Adver	Adver	1	0.15099	0.04711	3.21	0.0055	1.91021
	MktShare	MktShare	1	199.02354	67.02792	2.97	0.0090	3.23576
	Change	Change	1	290.85513	186.78200	1.56	0.1390	1.60173
	Accts	Accts	1	5.55096	4.77555	1.16	0.2621	5.63932
	WkLoad	WkLoad	1	19.79389	33.67669	0.59	0.5649	1.81835
X8	Rating	Rating	1	8.18928	128.50561	0.06	0.9500	1.80856

Since none of the VIFs are higher than 10, multicollinearity is not an issue.

However, notice that many of the p-values are extremely high. We can't accept the model as it is. If we try to kick out all of the independent variables with high p-values and we can't, we can use one of the model building methods below.

Model Building Methods

1. Forward selection
2. Backward selection
3. Stepwise regression

Lecture 9 (Feb 6 2024)

Forward Selection (add Xs)

1. Fit all models with **one** X. Test the strongest model (lowest p-value or highest F value).
   - If p-value $< \alpha_{enter}$, this model becomes our starting model.
   - If p-value $> \alpha_{enter}$, the process terminates (no useful model).
2. Fit all models with **two** Xs which include chosen X from step 1. Test the strongest model.
   - If p-value $< \alpha_{enter}$, this model replaces the model from step 1.
   - If p-value $> \alpha_{enter}$, the model from step 1 is our final model.
3. Fit all models with **three** Xs which include chosen Xs from steps 1 and 2.
   - If p-value $< \alpha_{enter}$, this model replaces the model from step 2.
   - If p-value $> \alpha_{enter}$, the model from step 2 is our final model.

…

k. Fit all models with k Xs (this would be the full model).
   - If p-value $< \alpha_{enter}$, this model is our final model.
   - If p-value $> \alpha_{enter}$, the model from step $k - 1$ is our final model.

Cons of Forward Selection

- Once an X enters the model, you cannot take it out. Just because an X is strong individually does not mean that it is a part of the strongest model!
- If you were to add a qualitative variable with at least 2 indicator variables, this method would enter them into the model separately. This means that the final model may not include all of the indicator variables belonging to one qualitative variable (not good).

Backward Elimination (remove Xs)

1. Fit the full model (all Xs). Test the weakest X (highest p-value or lowest F value).
    - If the highest p-value $> \alpha_{exit}$, remove the X associated with this p-value.
    - If all Xs have p-values $< \alpha_{exit}$, the full model (starting model) is the final model.
2. Fit the model with $k - 1$ Xs which does not include the X that was removed in step 1. Test the weakest X (highest p-value or lowest F value).
    - If the highest p-value $> \alpha_{exit}$, remove the X associated with this p-value.
    - If all Xs have p-values $< \alpha_{exit}$, this model is the final model.
3. Repeat until all Xs in the model have p-values $< \alpha_{exit}$. This is the final model.

Cons of Backward Selection

- Once an X exits the model, you cannot put it back in. Just because an X is weak individually does not mean that it is not a part of the strongest model!

Stepwise Regression (continuously add and remove Xs)

1. Fit all models with one X. Test the strongest model (lowest p-value or highest F value).
    - If p-value $< \alpha_{enter}$, this model becomes our starting model.
    - If p-value $> \alpha_{enter}$, the process terminates (no useful model).
2. Fit all models with two Xs which include X chosen in step 1. Keep the model in which:
    - New X has lowest p-value AND
    - this p-value $< \alpha_{enter}$ AND
    - X already in model have p-value $< \alpha_{exit}$.
    - If no new added X is significant, the model chosen in step 1 is the final model.
3. Repeat until new X has p-value $< \alpha_{enter}$ and X already in the model has p-value $< \alpha_{exit}$.
    - When this occurs, you have your final model.

Pros of Stepwise Regression

- Stepwise Regression solves the problems with entering and exiting that Forward Selection and Backward Elimination cause.

Exercise 2
Consider Exercise 1. Use all 3 modelling building methods to determine the best model. Use SAS.

Full SAS output included in lecture slides.

Full model includes: Time, MktPoten, Adver, MktShare, Change, Accts, WkLoad, Rating

Forward selection model includes: Accts, Adver, Mktpoten, MktShare
Backward elimination model includes: Time, MktPoten, Adver, MktShare
Stepwise regression model includes: Accts, Adver, MktPoten, MktShare

Model Selection
Say, for example, that we have a model with 50 variables that is only slightly better than a model with 7 variables. We would save so much time using the model with 7 variables. How do we decide which model to use? → Model selection

When more than one model is useful, the model that satisfies the following is the best model:

1. Largest R^2 or R_{adj}^2
2. Smallest MSE
3. Mallow's Cp closest to p
    - p is the number of βs in the model, which is equivalent to $k + 1$
    - $Cp = \frac{SSE_p}{MSE_f} - (n - 2p)$
        - SSE_p is the SSE for the model Cp is computed for.
        - MSE_f is the MSE for the full model.

Exercise 3
The sales manager of a company wishes to evaluate the performance of the company's sales representatives. Each sales representative is responsible for one sales territory, and the sales manager believes that the performance, y, of a sales representative can be evaluated by using the yearly sales of the company's product in the representative's sales territory. The manager feels that the sales performance, y, depends on the following variables:

X_1 = sales of the company's product and competing products in the sales territory
X_2 = dollar advertising expenditure in the territory
X_3 = weighted average of the company's market share in the territory for the previous four years
X_4 = number of accounts handled by the representative

Independent Variables in the model	*SSR*	*SSE*	*SSE df*	*MSE*	c_p
X_1	14788185	26591364	23	1156146	108.105
X_2	14707456	26672093	23	1159656	108.4969
X_3	9673814	31705735	23	1378510	132.936
X_4	23524074	17855475	23	776325	65.69095
X_1, X_2	25121971	16257578	22	738981	59.93293
X_1, X_3	30871719	10507830	22	477629	32.01706
X_1, X_4	26537074	14842475	22	674658	53.06239
X_2, X_3	19431324	21948225	22	997647	87.56185
X_2, X_4	32073320	9306229	22	423010	26.18311
X_3, X_4	25118452	16261097	22	739141	59.95001
X_1, X_2, X_3	35130240	6249309	21	297586	13.34131
X_1, X_2, X_4	34250796	7128753	21	339464	17.61114
X_1, X_3, X_4	32629834	8749715	21	416653	25.48115
X_2, X_3, X_4	32532485	8847064	21	421289	25.95379
X_1, X_2, X_3, X_4	37260202	4119346	20	205967	5.000029

B. Find the best model based on the 3 criteria for model selection. Assume all models are significant (model overall and individual Xs).

Reminder: $SSE\ df = n - k - 1$. We can use this to find n for the equation above.

We know that $SSE\ df(X_2, X_3) = 22$. We also know that $k = 2$. This means that:

$SSE\ df = n - k - 1$

$22 = n - 2 - 1$

$n = 25$

Mallow's Cp for model with X_2, X_3

$Cp = \frac{SSE_p}{MSE_f} - (n - 2p) = \frac{21948225}{205967} - (25 - 2(3)) = 87.5619$

Mallow's Cp for model with X_1, X_2, X_3

We want the model with 3 Xs (including X_2 and X_3) which has the <u>lowest MSE</u>, not counting the full model since we're trying to reduce the full model.

$MSE(X_1, X_2, X_3) = 297586$ ← lowest MSE

$MSE(X_2, X_3, X_4) = 421289$

$|Cp - p| = |13.34131 - 4| = |9.34131| = 9.34131$

Therefore, the best model (not including the full model) is the one with X_1, X_2, X_3.

Lecture 10 (Feb 8 2024)

A. Determine the subset of variables that is selected as best by the Forward Selection Procedure using $F_0^* = 1.9$ (to-add variable). Show your steps.

$F_0^* = 1.9$

We want the highest SSR since high SSR = high MSR = high F value = low p-value = strongest model.
$F = \frac{MSR}{MSE}$

Step 1 (one X)

$SSR(X_1) = 14788185$
$SSR(X_2) = 14707456$
$SSR(X_3) = 9673814$
$SSR(X_4) = 23524074 \leftarrow$ highest SSR. Our model becomes $Y = \beta_0 + \beta_4 X_4 + \varepsilon$

<u>Test model with F-test</u>
$H_0: \beta_4 = 0,\ H_\alpha: \beta_4 \neq 0$
$F_1 = \frac{MSR}{MSE} = \frac{SSR\ /\ (k-g)}{MSE} = \frac{23524074\ /\ (2-1)}{776325} = 30.3018$ ← Note: $k - g$ is always 1
Since $30.3018 > 1.9$, X_4 is kept.

Step 2 (two Xs)

$SSR(X_1, X_4) = 26537074$
$SSR(X_2, X_4) = 32073320 \leftarrow$ highest SSR. Our full model becomes $Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$
$SSR(X_3, X_4) = 25118452$

<u>Test model with Partial F-test</u>
$H_0: \beta_2 = 0,\ H_\alpha: \beta_2 \neq 0$
$F_2 = \frac{(SSR_F - SSR_r)\ /\ (k-g)}{MSE_F} = \frac{[SSR(X_2, X_4) - SSR(X_4)]\ /\ 1}{MSE(X_2, X_4)} = \frac{32073320 - 23524074}{423010} = 20.2105$
Since $20.2105 > 1.9$, X_2 is kept.

Step 3 (three Xs)

$SSR(X_1, X_2, X_4) = 34250796 \leftarrow$ highest SSR. Our full model becomes $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$
$SSR(X_2, X_3, X_4) = 32532485$

<u>Test model with Partial F-test</u>
$H_0: \beta_1 = 0,\ H_\alpha: \beta_1 \neq 0$
$F_3 = \frac{(SSR_F - SSR_r)\ /\ (k-g)}{MSE_F} = \frac{[SSR(X_1, X_2, X_4) - SSR(X_2, X_4)]\ /\ 1}{MSE(X_1, X_2, X_4)} = \frac{34250796 - 32073320}{339464} = 6.4145$

Since $6.4145 > 1.9$, X_1 is kept.

Step 4 (four Xs)

When we add one more X we have the full model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$.

<u>Test model with Partial F-test</u>

$H_0: \beta_3 = 0,\ H_\alpha: \beta_3 \neq 0$

$$F_3 = \frac{(SSR_F - SSR_r) / (k-g)}{MSE_F} = \frac{[SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_4)] / 1}{MSE(X_1, X_2, X_3, X_4)} = \frac{37260202 - 34250796}{205967} = 14.6111$$

Since $14.6111 > 1.9$, X_3 is kept.

Therefore, the best model is the full model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$. It has the lowest MSE, and its C_p is the closest to 1.9.

Analysis of Variance - Part 1

Types of studies

- Observational study: we are observing something that naturally occurs, without interfering.
  - Ex: eye colour
- Experimental study: we intentionally interfere with some aspect of the study.
  - Ex: drug trial

A study can be both observational and experimental.

Other definitions

Experimental unit: the object on which measurement(s) are taken.
Response variable (dependent variable): the variable of interest that is measured on each object in the study.
Factors (independent variables): A variable whose relationship to the dependent variable is being studied.
Factor levels: the different categories that a factor can take on.
Treatment: a combination of factor levels

<u>Example.</u> Athletes want to gain muscle mass.
Factor 1: diet 1, diet 2 ← factor 1 has 2 levels
Factor 2: exercise 1, exercise 2 ← factor 2 has 2 levels

	Exercise 1	Exercise 2
Diet 1	D_1E_1	D_1E_2
Diet 2	D_2E_1	D_2E_2

The treatments are: D_1E_1, D_1E_2, D_2E_1, D_2E_2.

<u>Exercise 1</u>
State the kind of study, experimental unit, factor(s), factor levels, treatments, and response variable for the following studies.

A. A group of people was randomly split into an experimental group and a control group. The control group was given an aptitude test after having eaten breakfast. The experimental group was given the same test without having breakfast.

Kind of study:	experimental (controlled breakfast)
Experimental unit:	a person
Factor(s) and their levels:	breakfast status (yes, no)
Treatments:	breakfast, no breakfast (or you can write "same as factors")
Response variable:	aptitude test score

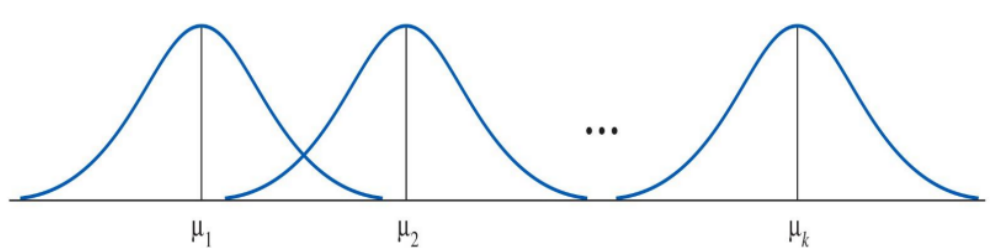
Note: if there is one factor, the factor levels are the same as the treatment options.

B. In a study of the effect of <u>applicant's eye contact (yes, no)</u> and personnel officer's <u>gender (male, female, other)</u> on the personnel officer's assessment of likely success of an applicant, personnel officers were shown a front view of photograph of an applicant's face and were asked to give the person in the photograph a success rating score. Half of the officers in each gender group were chosen at random to receive a version of the photograph in which the applicant made eye contact with the counselors. The other half received a version in which there was no contact. Data were collected on success rate.

Kind of study:	observational (gender), experimental (eye contact)
Experimental unit:	personnel officer
Factor(s) and their levels:	gender (female, male, other) eye contact (yes, no)
Treatments:	FY, FN, MY, MN, OY, ON
Response variable:	success rate

Completely Randomized Design (One-Way CRD)

- k independent samples are drawn from k populations.
- Tests one factor which has k levels. The factor is the population where each sample came from.
- The goal is to investigate differences between the k-means.
    - If $k = 2$, use t-test.
    - If $k \geq 3$, we need a new method.



Assumptions for One-Way CRD

The assumptions on the k populations are:

1. Each probability distribution is normal.
2. Each probability distribution has the same variance.
3. The responses for each factor level are random selections from the corresponding probability distribution and are independent of the responses for any other factor level.

Total Variation in an Experiment (TSS)

Let y_{ij} be the jth measurement ($j = 1, 2, \ldots, n_i$) in the ith factor level ($i = 1, 2, \ldots, k$).

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^{\ 2} - \frac{(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij})^2}{n} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^{\ 2} - \frac{(GT)^2}{n}$$

where: $n = n_1 + \ldots + n_k$ (overall sample size)

$GT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$ (grand total)

Sources of Variance

Total variance (TSS) comes from two sources:

- Differences <u>between</u> the different samples (SST)
- Differences <u>within</u> each sample (SSE)

$TSS = SST + SSE$

μ_1	μ_2	μ_3
$y_{1\,1}$	$y_{2\,1}$	$y_{3\,1}$
$y_{1\,2}$	$y_{2\,2}$	$y_{3\,2}$
…	…	…
$y_{1\,n_1}$	$y_{2\,n_2}$	$y_{3\,n_3}$

Variability within treatments (compared to μ_i)

"how each observation differs from its own mean"

Sum of Squares Error

$SSE = (n_1 - 1)s_1^{\ 2} + \ldots + (n_k - 1)s_k^{\ 2}$

where $s_k^{\ 2}$ is the sample variance

Difference between treatments (compared to μ)

"how category mean differs from population mean"

Sum of Squares Treatment

$$SST = \sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2 = \sum_{i=1}^{k} \frac{r_1^{\ 2}}{n_i} - \frac{(GT)^2}{n}$$

where T_i is the total of the observations for treatment i

ANOVA for One-Way CRD

Source	df	SS	MS	F
Treatment	$k - 1$	SST	MST	$\frac{MST}{MSE}$
Error	$n - k$	SSE	MSE	
Total	$n - 1$	TSS		

Hypothesis Test for One-Way CRD

1. State hypothesis.
    a. H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
    b. H_α: at least one μ_1 is different ($i = 1, 2, \ldots, k$)
2. Test statistic. $F = \frac{MST}{MSE}$
3. Rejection region.
    - Reject H_0 if $F > F_{a,\, k-1,\, n-k}$.
    - Fail to reject H_0 otherwise.

Lecture 11 (Feb 13 2024)

Exercise 2

In an experiment to determine the effect of nutrition on the attention spans of elementary school students, a group of 15 students were randomly assigned to each of three meal plans: no breakfast, light breakfast, and full breakfast. Their attention span (in minutes) were recorded during a morning reading period and are summarized in the following table.

No Breakfast	Light Breakfast	Full Breakfast
8	14	10
7	16	12
9	12	16
13	17	15
10	11	12

At $\alpha = 0.05$, is there evidence to conclude that the average attention span differs by breakfast status?

This is an experimental study with 1 factor (breakfast status) which has 3 levels (no, light, full).

- $n_1 = n_2 = n_3 = 5$
- $n = n_1 + n_2 + n_3 = 15$
- $k = 3$

Find GT

$$T_1 = \sum_{j=1}^{5} y_{1j} = 8 + 7 + 9 + 13 + 10 = 47$$

$$T_2 = \sum_{j=1}^{5} y_{2j} = 14 + 16 + 12 + 17 + 11 = 70$$

$$T_3 = \sum_{j=1}^{5} y_{3j} = 10 + 12 + 16 + 15 + 12 = 65$$

$$GT = \sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij} = T_1 + T_2 + T_3 = 47 + 70 + 65 = 182$$

Find TSS

$$TSS = \sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}^{\ 2} - \frac{(GT)^2}{n} = 8^2 + \ldots + 12^2 - \frac{182^2}{15} = 2338 - 2208.2667 = 129.7333$$

Note that $\frac{(GT)^2}{n} = \frac{182^2}{15} = 2208.2667$, since we will be using this value again (for SST).

Find SST

$$SST = [\sum_{i=1}^{3} \frac{T_i^{\ 2}}{n_i}] - \frac{(GT)^2}{n} = \frac{47^2}{5} + \frac{70^2}{5} + \frac{65^2}{5} - 2208.2667 = 58.5333$$

Find SSE

$SSE = TSS - SST = 129.7333 - 58.5333 = 71.2$

Create ANOVA Table

Source	df	SS	MS	F
Treatment	$3 - 1 = 2$	58.5333	$\frac{58.5333}{2} = 29.2667$	$\frac{29.2667}{5.9333} = 4.9326$
Error	$15 - 3 = 12$ or $14 - 2 = 12$	71.2	$\frac{71.2}{12} = 5.9333$	
Total	$15 - 1 = 14$	129.7333		

Hypothesis Test

H_0: $\mu_1 = \mu_2 = \mu_3$, H_α: At least one μ_i differs ($i = 1, 2, 3$)

$F_{\alpha, k-1, n-k} = F_{0.05, 2, 12} = 3.89$

Since $4.9326 > 3.89$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that mean attention span differs by breakfast status.

Lecture 12 (Feb 15 2024)

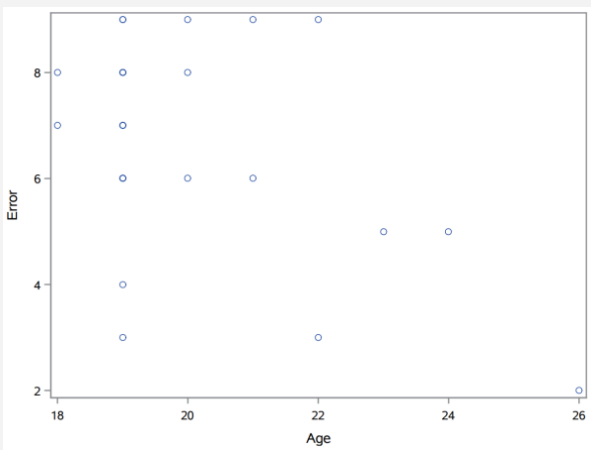
Note: If you are doing an F-test and your desired degree of freedom is between two values on the table, choose the value that makes rejecting the null hypothesis harder (the larger value).

Practice Problem 1

A Canadian youth training program tested 27 undergraduate students to examine the relationship between age and errors made on a flight simulator test. The ages of the participants ranged from 18 to 26 years. The number of errors made was found for each individual. The data collected on age, x, and number of errors, y, were used to compute the sums below.

$\Sigma x_i = 540 \qquad \Sigma y_i = 179 \qquad \Sigma x_i^2 = 10894 \qquad \Sigma y_i^2 = 1289 \qquad \Sigma x_i y_i = 3534$

A. Comment on the relationship using the scatter plot of age versus number of errors.



Based on the graph, there is a weak negative linear relationship between age and number of errors.

Note: Since the relationship is weak, Ahmed would accept any response (negative/positive/no relationship) for direction of relationship. Also make sure to comment on if you think the relationship is linear or not.

Tip: later on in the problem, once you have the regression model, the coefficient of β_1 (slope) will confirm what the direction of the relationship is.

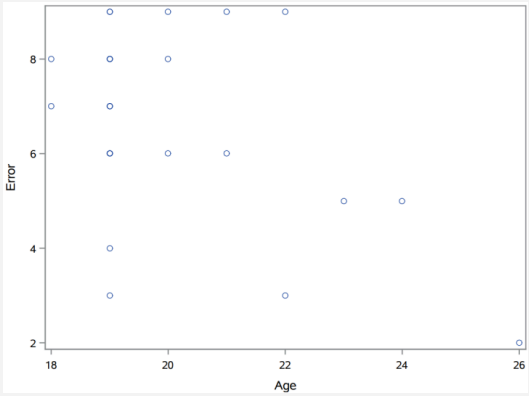
B. State the simple linear regression model along with the assumptions.

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i, \quad \overset{\text{i.i.d}}{\varepsilon_i \sim N(0, \sigma^2)}$$

Note: including subscript i is optional

C. Using the following 3 plots, test the assumptions.

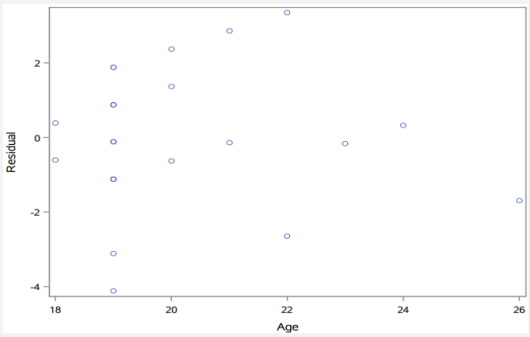
Age vs. residuals



Since there is a random scatter of points above and below zero, linearity assumption and mean equal zero assumption are valid.

Note: you can say “valid” or “not violated”, either is fine.

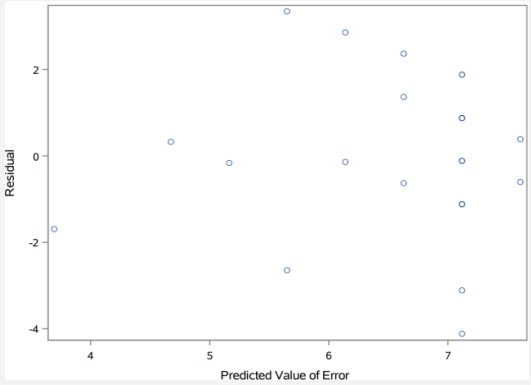
Pred. value vs. residuals



Since there is a random scatter of points above and below zero, equality of variance assumption is valid.

Notes: if we see a pattern, linearity is not valid. if we see coning/fanning, equality of variance is not valid.

Q-Q Plot



Since the points are close to the line, normality assumption is valid.

D. Find the least square estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ and give the least square fitted regression line.

$$S_{xy} = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n} = 3534 - \frac{(540)(179)}{27} = -46$$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 10894 - \frac{(540)^2}{27} = 94$$

$$\beta_1 = \frac{S_{xy}}{S_{xx}} \quad \Rightarrow \quad b_1 = \frac{-46}{94} = -0.4894$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad \Rightarrow \quad b_1 = \frac{179}{27} - (-0.4894)(\frac{540}{27}) = 16.4176$$

Regression line: $\hat{y} = 16.4176 - 0.4894x$

E. Construct the ANOVA table and <u>use it to test</u> if there is a significant linear relationship between number of errors and age. Use $\alpha = 0.05$.

$$TSS = S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 1289 - \frac{(179)^2}{27} = 102.2963$$

$$SSR = \frac{(S_{xy})^2}{S_{xx}} = \frac{(-46)^2}{94} = 22.5106$$

$$SSE = TSS - SSR = 102.2963 - 22.5106 = 79.7857$$

Source	df	SS	MS	F
Regression	$num\ of\ X$s $= 1$	22.5106	$\frac{22.5106}{1} = 22.5106$	$\frac{22.5106}{3.1914} = 7.0535$
Error	$n - 2 = 27 - 2 = 25$	79.7857	$\frac{79.7857}{25} = 3.1914$	
Total	$n - 1 = 27 - 1 = 26$	102.2963		

Note: Since we have been asked to use the ANOVA table to test the relationship, we need to do an F-test, not a t-test. We do an F-test since the ANOVA table gives us the F-statistic.

$H_0: \beta_1 = 0, \ H_\alpha: \beta_1 \neq 0$

$F_{0.05,\,1,\,25} = 4.24$

Since $7.0535 > 4.24$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that there is a significant linear relationship between number of errors and age.

F. Find a 95% confidence interval for the mean value of the number of errors for a person aged 25.

$\hat{y} = 16.4176 - 0.4894(25) = 4.1826$

$t_{\alpha/2, n-2} = t_{0.05/2, 27-2} = t_{0.025, 25} = 2.060$

$SE = \sqrt{MSE(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})} = \sqrt{3.1914(\frac{1}{27} + \frac{(25 - \frac{540}{27})^2}{94})} = 0.9833$

<u>95% CI for E(Y|X=25)</u>

$\hat{y} \pm t_{\alpha/2, n-2}\sqrt{MSE(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}$

$= 4.1826 \pm (2.060)(0.9833)$

$= 4.1826 \pm 2.0256$

$= (2.157,\ 6.2082)$

Therefore, we can say with 95% confidence that when age is 25, the average number of errors is anywhere from 2.157 to 6.2082.

G. Find the predicted number of errors for a person aged 35.

Since 35 is not in the range 18 to 26, it is not appropriate to use the fitted line to predict.

Practice Problem 2

Identifying individuals with a high risk of Alzheimer's disease usually involves a long series of cognitive tests. However, researchers have developed a 7-Minute screen, which is a quick and easy way to accomplish the same goal. To address whether the 7-Minute Screen is as effective, both tests are administered to a group of patients (Ijuin et al., 2008). The following are data similar to those obtained from the study.

Patient		1	2	3	4	5	6	7	8	9
Risk Score	7-Minute Screen (x)	3	8	10	8	4	7	4	5	14
	Cognitive Series (y)	11	20	22	20	14	14	9	20	25

Given $SSE = 66.3719$, conduct a lack of fit test at $\alpha = 0.05$.

Step 1: solve for SSPE and SSLF

n_i	x	y_{ij}	$\overline{y}_i$	$y_{ij} - \overline{y}_i$
$n_1 = 1$	3	11	11	$11 - 11 = 0$
$n_2 = 2$	4	9, 14	$\frac{9+14}{2} = 11.5$	$9 - 11.5 =- 2.5,\ 14 - 11.5 = 2.5$
$n_3 = 1$	5	20	20	$20 - 20 = 0$
$n_4 = 1$	7	14	14	$14 - 14 = 0$
$n_5 = 2$	8	20, 20	$\frac{20+20}{2} = 20$	$20 - 20 = 0,\ 20 - 20 = 0$
$n_6 = 1$	10	22	22	$22 - 22 = 0$
$n_7 = 1$	14	25	25	$25 - 25 = 0$

$m = 7$ and $n = 9$

$$SSPE = \sum_{i=1}^{7}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$
$$= 0^2 + (-2.5)^2 + (2.5)^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2$$
$$= 6.25 + 6.25$$
$$= 12.5$$

$$SSLF = SSE - SSPE = 66.3719 - 12.5 = 53.8719$$

Step 2: Create ANOVA table and solve for F-statistic

Source	df	SS	MS	F
Lack of fit	$m - 2 = 7 - 2 = 5$	53.8719	$\frac{53.8719}{5} = 10.7744$	$\frac{10.7744}{6.25} = 1.7239$
Pure error	$n - m = 9 - 7 = 2$	12.5	$\frac{12.5}{2} = 6.25$	
Error	$df_1 + df_2 = 5 + 2 = 7$	66.3719		

Step 3: Perform F-test

H_0: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ is appropriate , H_α: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ is not appropriate

$F_{\alpha, m-2, n-m} = F_{0.05, 5, 2} = 19.30$

Since $1.7239 < 19.30$, at $\alpha = 0.05$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that the model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ is **not** appropriate.

Note: we can't conclude that the model is appropriate.

Confidence intervals for pairwise differences between treatment means

- Recall that failing to reject H_0 tells us that at least one μ_1 differs.
- Accepting H_α doesn't tell us that all treatment means differ, nor does it explain the nature of the difference.
- We can learn more about the differences in μ_i by using pairwise differences.

I. Fisher's Least Significant Difference (LSD) Method

The $(1 - \alpha)100$ Fisher's LSD CI for $\mu_i - \mu_j$ is given by $(\overline{y}_i - \overline{y}_j) \pm t_{\alpha/2, n-k}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$.

Note: This method is not ideal so we will not be using it after we learn it.

Review from 2507: interpreting the result of $\mu_i - \mu_j$

$(-, -) \quad \Rightarrow \quad \mu_i < \mu_j$

$(-, +) \quad \Rightarrow \quad \mu_i = \mu_j$

$(+, +) \quad \Rightarrow \quad \mu_i > \mu_j$

Exercise 3

In an experiment to determine the effect of nutrition on the attention spans of elementary school students, a group of 15 students were randomly assigned to each of three meal plans: no breakfast, light breakfast, and full breakfast. Their attention span (in minutes) were recorded during a morning reading period and are summarized in the following table.

No Breakfast	Light Breakfast	Full Breakfast
8	14	10
7	16	12
9	12	16
13	17	15
10	11	12

Find one pairwise 95% comparison for the treatment means in the aptitude test scores problem using Fisher's LSD method.

Note: there are $C_2^3 = 3$ pairwise comparisons that can be made.

95% CI for $\mu_2 - \mu_3$ (light breakfast vs. full breakfast)

We calculated T_1, T_2, T_3, MSE in Lecture 11, Exercise 2.

$T_1 = 47 \quad \Rightarrow \quad \overline{y_1} = \frac{T_2}{n_2} = \frac{47}{5} = 9.4$ ← not used in this problem but used in Ex 4

$T_2 = 70 \quad \Rightarrow \quad \overline{y_2} = \frac{T_2}{n_2} = \frac{70}{5} = 14$

$T_3 = 65 \quad \Rightarrow \quad \overline{y_3} = \frac{T_3}{n_3} = \frac{65}{5} = 13$

$$(\overline{y_2} - \overline{y_3}) \pm t_{0.05/2,\,15-3}\sqrt{MSE(\frac{1}{n_2} + \frac{1}{n_3})}$$

$$= (14 - 13) \pm 2.179\sqrt{5.9333(\frac{1}{5} + \frac{1}{5})}$$

$$= 1 \pm 3.3569$$

$$= (-2.3569,\ 4.3569)$$

Therefore, we can say with 95% confidence that there is no difference in the average attention span between light breakfast and full breakfast.

Problems with Fisher's LSD Method

Fisher's LSD method only works for one pairwise difference. If you attempt to use it for simultaneous pairwise comparisons, the confidence level will get lower with more comparisons. To keep the confidence level the same for simultaneous comparisons, we use Tukey's Method.

II. Tukey's Simultaneous Pairwise Comparisons

The $(1 - \alpha)100$ Tukey's CI for $\mu_i - \mu_j$ is given by $(\overline{y_i} - \overline{y_j}) \pm q_{\alpha(k,\,n-k)}\sqrt{\frac{MSE}{2}(\frac{1}{n_i} + \frac{1}{n_j})}$.

Note: $q_{\alpha(k,\,n-k)}$ is obtained from the Tukey Table (available on Brightspace).

<u>Exercise 4 (by hand)</u>

We are still working with the breakfast data. Find all pairwise 95% comparisons for the treatment means in the aptitude test scores problem, if appropriate. Use Tukey's method.

Recall, from Exercise 3 that: $\overline{y_1} = 9.4,\ \overline{y_2} = 14,\ \overline{y_3} = 13.$

<u>Find value of error term (value after $\pm$ sign)</u>

$$(\overline{y_i} - \overline{y_j}) \pm q_{\alpha(k,\, n-k)} \sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$= (\overline{y_i} - \overline{y_j}) \pm q_{0.05(3,\, 15-3)} \sqrt{\frac{5.9333}{2}\left(\frac{1}{5} + \frac{1}{5}\right)}$$

$$= (\overline{y_i} - \overline{y_j}) \pm 3.77(1.0893)$$

$$= (\overline{y_i} - \overline{y_j}) \pm 4.1067$$

<u>95% CI for $\mu_1 - \mu_2$ (no breakfast vs. light breakfast)</u>

$$= (\overline{y_1} - \overline{y_2}) \pm 4.1067$$

$$= (9.4 - 14) \pm 4.1067$$

$$= -4.6 \pm 4.1067$$

$$= (-8.7067, -0.4933) \qquad \leftarrow \mu_1 < \mu_2$$

<u>95% CI for $\mu_1 - \mu_3$ (no breakfast vs. full breakfast)</u>

$$= (\overline{y_1} - \overline{y_3}) \pm 4.1067$$

$$= (9.4 - 13) \pm 4.1067$$

$$= -3.6 \pm 4.1067$$

$$= (-7.7067, 0.5067) \qquad \leftarrow \mu_1 = \mu_3$$

<u>95% CI for $\mu_2 - \mu_3$ (light breakfast vs. full breakfast)</u>

$$= (\overline{y_2} - \overline{y_3}) \pm 4.1067$$

$$= (14 - 13) \pm 4.1067$$

$$= 1 \pm 4.1067$$

$$= (-3.1067, 5.1067) \qquad \leftarrow \mu_2 = \mu_3$$

Therefore, we can say with 95% confidence that ① the mean attention span with no breakfast is lower than that with light breakfast and ② there is no difference from the remaining pairwise comparisons.

Note: you want to ① mention all differences and ② group all pairs with no differences together.

Lecture 14 (Feb 29 2024)

Test 1 Review

The MLR model is $Y = \beta_0 + \beta_1 X_x + \ldots + \beta_k X_k + \varepsilon$.

<u>F-test</u> (model / all Xs)

$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$, H_α: At least one $\beta_j \neq 0$ $(j = 1, \ldots, k)$

<u>t-test</u> (one X at a time)

$H_0: \beta_j = 0$, $H_\alpha: \beta_j \neq 0$

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

<u>Partial F-test</u> (more than one X but not all Xs; can be used for one X)

$H_0: \beta_1 = \beta_2 = \ldots = \beta_g = 0$, H_α: At least one $\beta_j \neq 0$ $(j = 1, \ldots, g)$ ← js won't necessarily be ordered

<u>Practice Problem 3</u>

The following SAS output is the result of analyzing scores (0 to 100) given to 12 brands of vegetarian burgers based on their price, calories, fat content, and sodium content. You may assume all assumptions are valid for all 3 models. Use $\alpha = 0.05$ for all tests.

A. Is Model 1 useful for predicting score? → F-test

Model 1 : price, calories, fat content, sodium content

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1128.43227	282.10807	1.74	0.2443
Error	7	1132.56773	161.79539		
Corrected Total	11	2261.00000			

Root MSE	12.71988	R-Square	0.4991
Dependent Mean	43.50000	Adj R-Sq	0.2128
Coeff Var	29.24111		

<u>State hypothesis</u>

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, H_α: At least one β_j differs $(j = 1, 2, 3, 4)$

<u>Test rejection region</u>

Since $0.2443 > 0.05$, at $\alpha = 0.05$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that Model 1 is useful at predicting score.

B. Can price (X_1) and sodium content (X_4) be <u>removed from Model 1</u>? → Partial F-test

Note: The problem is written in terms of H_0; we can't write our conclusion in terms of H_0.

Model 1 : price, calories, fat content, sodium content

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1128.43227	282.10807	1.74	0.2443
Error	7	1132.56773	161.79539		
Corrected Total	11	2261.00000			

Root MSE	12.71988	R-Square	0.4991
Dependent Mean	43.50000	Adj R-Sq	0.2128
Coeff Var	29.24111		

Model 2 : calories, fat content

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	842.17198	421.08599	2.67	0.1228
Error	9	1418.82802	157.64756		
Corrected Total	11	2261.00000			

Root MSE	12.55578	R-Square	0.3725
Dependent Mean	43.50000	Adj R-Sq	0.2330
Coeff Var	28.86386		

<u>State hypothesis</u>

$H_0: \beta_1 = \beta_4 = 0$, H_α: At least one $\beta_j \neq 0$ $(j = 1, 4)$

<u>State full and reduced models</u> (not required for full marks)

Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ $\leftarrow k = 4 = df\ SSR_F$

Reduced model: $Y + \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ $\leftarrow g = 2 = df\ SSR_r$

<u>Find test statistic</u>

$$F_{part} = \frac{(SSR_F - SSR_r) / (df\ SSR_F - df\ SSR_r)}{SSE_F / df\ SSE_F} = \frac{(1128.43227 - 842.17198) / (4-2)}{1132.56773 / 7} = 0.8846$$

Note: you can use either F_{part} or F_{drop} since they are equivalent tests.

$F_{0.05,\,2,\,7} = 4.74$

<u>Test rejection region</u>

Since $0.8846 < 4.74$, at $\alpha = 0.05$, we don't have enough evidence to reject H_0. Therefore, we can't conclude that price and sodium content <u>can't be removed from Model 1</u> / <u>are useful in Model 1</u>.

Note: Our conclusion is written in terms of H_α (rejecting H_α rather than accepting H_0).

C. Can calories (X_2) be removed from Model 2? → t-test

Model 2 : calories, fat content

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	842.17198	421.08599	2.67	0.1228
Error	9	1418.82802	157.64756		
Corrected Total	11	2261.00000			

Root MSE	12.55578	R-Square	0.3725
Dependent Mean	43.50000	Adj R-Sq	0.2330
Coeff Var	28.86386		

Model 3 : fat content

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	384.86441	384.86441	2.05	0.1826
Error	10	1876.13559	187.61356		
Corrected Total	11	2261.00000			

Root MSE	13.69721	R-Square	0.1702
Dependent Mean	43.50000	Adj R-Sq	0.0872
Coeff Var	31.48784		

<u>State hypothesis</u>

$H_0: \beta_2 = 0, \ H_\alpha: \beta_2 \neq 0$

Note: we would normally do a t-test, which requires the parameter estimates table since $t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$. Since we were given the ANOVA table, we must perform a Partial F-test instead.

<u>State full and reduced models</u>

Full model: $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ $\quad \leftarrow k = 2$

Reduced model: $Y = \beta_0 + \beta_3 X_3 + \varepsilon$ $\quad \leftarrow g = 1$

<u>Find test statistic</u>

$F_{part} = \frac{(842.17198 - 384.86441) \ / \ (2-1)}{1418.82802 \ / \ 9} = 2.9008$

$F_{0.05,\,1,\,9} = 5.12$

<u>Test rejection region</u>

Since $2.9008 < 5.12$, at $\alpha = 0.05$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that <u>calories can't be removed from the model</u> / <u>calories are useful in the model</u>.

D. Based on parts A, B, and C, have you found a model? Explain.

No, we have not found a useful model.

A: Model is not good overall.

B: We can kick 2 Xs out, so price and sodium content are not significant.

C: We can kick one more X out, so calories is not significant and Model 2 is not good.

D: SLR model (Model 3) is not good, so fat content is not significant.

Note: If asked about usefulness, don't comment on R_{adj}^2, which comments on strength.

Lecture 15 (Mar 5 2024)

Exercise 4 (using SAS)

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for span

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	5.933333
Critical Value of Studentized Range	3.77289
Minimum Significant Difference	4.11

Comparisons significant at the 0.05 level are indicated by *.**

Breakfast Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
2 - 3	1.000	-3.110	5.110	
2 - 1	4.600	0.490	8.710	***
3 - 2	-1.000	-5.110	3.110	
3 - 1	3.600	-0.510	7.710	
1 - 2	-4.600	-8.710	-0.490	***
1 - 3	-3.600	-7.710	0.510	

95% CI for $\mu_1 - \mu_2$ (no vs. light)

$= (-8.710, \ -0.490)$ $\leftarrow \mu_1 < \mu_2$

95% CI for $\mu_1 - \mu_3$ (no vs. full)

$= (-7.710, \ 0.510)$ $\leftarrow \mu_1 = \mu_3$

95% CI for $\mu_2 - \mu_3$ (light vs. full)

$= (-3.110, \ 5.110)$ $\leftarrow \mu_2 = \mu_3$

Therefore, we can say with 95% confidence that ① the mean attention span with no breakfast is lower than that with light breakfast and ② there is no difference from the remaining pairwise comparisons.

Exercise 5

This exercise has been left as homework. We may review this question before the test.

Analysis of Variance - Part 2

Cell Mean Model (ANOVA Model for One-Way CRD)

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \varepsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

Assumptions for Cell Mean

The assumptions are the same as the regression assumptions.

1. Error terms are normally distributed
2. Error terms are independent
3. Error terms have a mean of zero
4. Error terms have an equal variance, σ^2

Testing assumptions graphically

- Residuals vs. fitted values: equality of variance
- Q-Q plot: normality assumption

Note 1: Independence and mean equals zero assumptions are always valid.

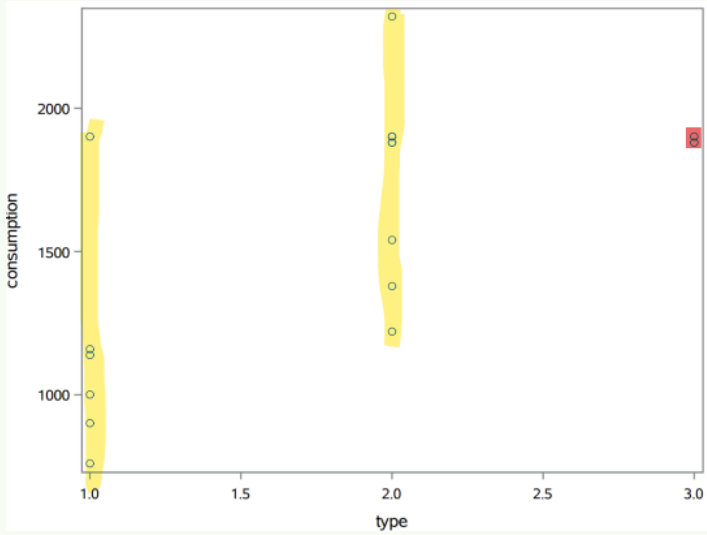
Note 2: Ahmed will not be testing us on these graphs.

<u>Exercise 1</u>

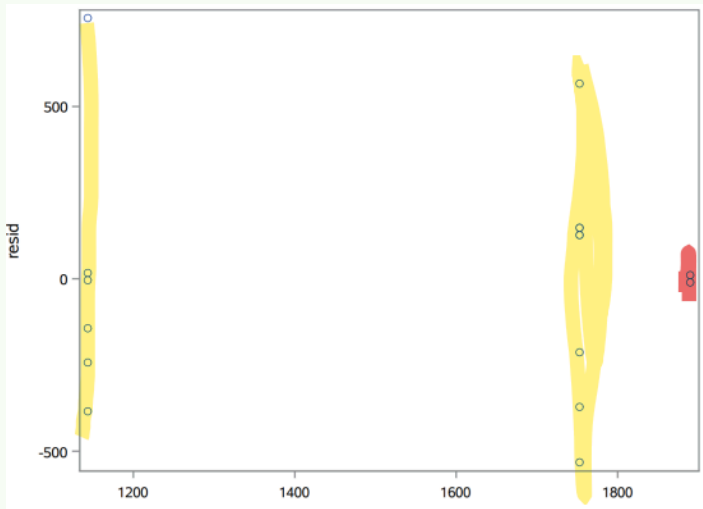
A study was done to compare the fuel consumption, in litres, of different types of hybrid vehicles that were driven 20,000 kilometres. Based on the information given below for the different models of hybrid cars, trucks, and SUVs, is there sufficient evidence to conclude that there is a difference in fuel consumption? Use $\alpha = 0.1$.

Hybrid Car	Hybrid Truck	Hybrid SUV
1000	2320	1880
900	1220	1900
1900	1380	1880
1160	1880	1900
1140	1900	
760	1880	
	1900	
	1540	

Note: I don't recall solving this question fully in class. I will ask around in case I missed it! If we did go over it, I will upload an updated PDF.

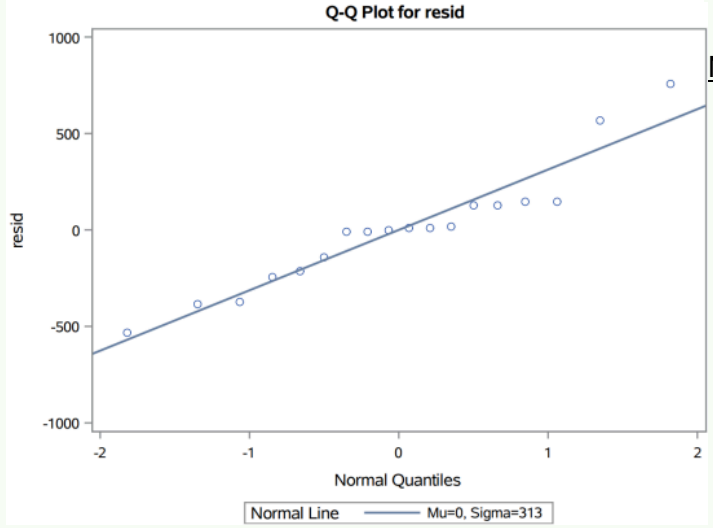


This is just the plot of Y vs. X. We can see if average consumption differs by type. Note that the sample size is very small for type 3 ($n_3 = 4$), so we should focus more on types 1 and 2 to make our conclusion. From the graph, it looks like average consumption does differ by type.



Is variance equal? Note that the sample size is very small for type 3 ($n_3 = 4$), so we should focus more on types 1 and 2 to make our conclusion. From the graph, it looks like <u>equality of variance assumption is valid</u>.

<u>Mean equals zero assumption is valid</u> (as always).



<u>Normality assumption is valid</u>.

Review from 2507: Sample variance

$$s_i^2 = \frac{\sum_{i=1}^{n_i} y_{ij}^2 - \frac{(\sum_{i=1}^{n_i} y_{ij})^2}{n_i}}{n_i - 1}$$

Each population will have its own sample variance.

Hartley's Test (tests equality of variance)

- State hypothesis.
  - $H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$ ← errors have equal variance
  - H_α: Not all σ_i^2 are equal ($i = 1, 2, \ldots, k$)

Note: we want to fail to reject H_0 (not enough evidence to say equality of variance is violated).

- Calculate test statistic, F_{max}.
  - $F_{max} = \frac{s_{max}^2}{s_{min}^2} \sim F_{max(k, [\overline{n}]-1)}$
  - $\overline{n} = \frac{\sum_{i=1}^{k} n_i}{k}$
  - $[\overline{n}]$ returns the integer portion of $\overline{n}$ (it's equivalent to the floor operation)
    - Example: $[\overline{3.9}] = 3$
- Test rejection region.
  - Reject H_0 if $F_{max} > F_{max(a,\, k,\, [\overline{n}]-1)}$ ← value obtained from Hartley's table

Note: Hartley's Table works with only two alpha levels: $\alpha = 0.05$ and $\alpha = 0.01$.

<u>Exercise 2</u>

Consider Exercise 1. Use Hartley's test to test the equality of the variance. Use $\alpha = 0.05$.

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$, H_α: Not all σ_i^2 are equal ($i = 1, 2, 3$)

S_i^2
159906.67
123650
133.33

$F_{max} = \frac{159906.67}{133.33} = 1199.33$

$n_1 = 6,\ n_2 = 8,\ n_3 = 4$

$\overline{n} = \frac{6+8+4}{3} = 6$

$[\overline{n}] = [6] = 6$

$F_{max(\alpha,\, k,\, [\overline{n}]-1)} = F_{0.05,\, 3,\, 6-1} = F_{0.05,\, 3,\, 5} = 10.8$

Since $1199.33 > 10.8$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that the equality of variance assumption is violated.

Note: This is bad. We should not use this data until equality of variance is no longer violated, through transforming the data. We don't cover how to do this in STAT 2509.

Kruskal-Wallis Test (for CRD; nonparametric alternative to F-test)
A "nonparametric" test is not sensitive to distributions, meaning it can be used when:

- normality is violated **and/or**
- equality of variance is violated

In order to perform a Kruskal-Wallis test, the data must be ranked accordingly:

1. Rank **all** Y_{ij} (data from all populations) from smallest to largest.
2. If there is a tie, the repeated values take on the average rank.

Example:	Values:	0.5	2	4.5	4.5	6.2	
	Rank:	1	2	3.5	3.5	5	← Since $\frac{3+4}{2} = 3.5$

Once the data is ranked, we can perform the hypothesis test.

- State hypothesis.
  - H_0: The medians are equal
  - H_α: At least one median is different
  - Note: mu with a tilde on top ($\tilde{\mu}_i$) is the symbol for median
- Calculate H statistic.
  - $H = \frac{12}{n(n+1)}\left[\sum_{i=1}^{k} \frac{T_{R_i}^{\ 2}}{n_i}\right] - 3(n+1) \sim \chi^2_{a,\,k-1}$ where $n = n_1 + n_2 + \ldots + n_k$ and T_{R_i} is the total rank for factor level i
- Test rejection region
  - Reject H_0 if $H > \chi^2_{a,\,k-1}$

Note: If $\sum_{i=1}^{k} T_{R_i} = \frac{n(n+1)}{2}$, your ranks are correct. Use this to **double check** your work!

Exercise 3

The data below are collected using a completely randomized design. The data are achievement test scores for four different groups of students, each group taught by a different teaching technique. Use the **Kruskal-Wallis test** to test if there is a difference in the distribution of achievement scores by technique. Use $\alpha = 0.05$.

Note: in the table below, I have written the entries in black and their ranks in grey.

Method 1	Method 2	Method 3	Method 4
65 3	75 9	59 1	94 20
87 18	69 5.5	78 11	96 22
73 8	83 16.5	67 4	99 23
79 12.5	81 14.5	62 2	95 21
81 14.5	72 7	83 16.5	
69 5.5	79 12.5	76 10	
	90 19		

$\frac{5+6}{2} = 5.5$

$\frac{12+13}{2} = 12.5$

$\frac{14+15}{2} = 14.5$

$\frac{16+17}{2} = 16.5$

$T_{R_1} = 3 + 18 + 8 + 12.5 + 14.5 + 5.5 = 61.5$

$T_{R_2} = 9 + 5.5 + 16.5 + 14.5 + 7 + 12.5 + 19 = 84$

$T_{R_3} = 1 + 11 + 4 + 2 + 16.5 + 10 = 44.5$

$T_{R_4} = 20 + 22 + 23 + 21 = 86$

$\sum_{i=1}^{4} T_{R_i} = 61.5 + 84 + 44.5 + 86 = 276$

Double check! $\frac{n(n+1)}{2} = \frac{23(23+1)}{2} = \frac{23(24)}{2} = 276$. $\sum_{i=1}^{k} T_{R_i} = \frac{n(n+1)}{2}$, so we did it right.

State hypothesis

H_0: $\tilde{\mu_1} = \tilde{\mu_2} = \tilde{\mu_3} = \tilde{\mu_4}$, H_α: At least one $\tilde{\mu_i}$ differs $(i = 1, 2, 3, 4)$

Calculate test statistic

$H = \frac{12}{(23)(24)}\left[\frac{61.5^2}{6} + \frac{84^2}{7} + \frac{44.5^2}{6} + \frac{86^2}{4}\right] - 3(23 + 1) = 10.9873$

Test rejection region

$\chi^2_{0.05,\,4-1} = \chi^2_{0.05,\,3} = 7.81473$ ← Chi table gives 5 decimal places, so keep all of them

Since $10.9873 > 7.81473$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that median scores differ by teaching method.

Bonferroni's Procedure (simultaneous pairwise comparisons for nonparametric distributions)
Given g possible pairwise comparisons, the CIs with a <u>family level</u> of at least $100(1-\alpha)\%$ can be calculated using $(\overline{R_i}-\overline{R_j}) \pm z_{1-a/2g}\sqrt{\frac{(n)(n+1)}{12}(\frac{1}{n_i})(\frac{1}{n_j})}$.

<u>Exercise 4</u>
Use Bonferroni's procedure to find all 95% pairwise comparisons.

- There are $C_2^4 = 6$ possible pairwise comparisons, so $g = 6$.
- Since we are finding 95% pairwise comparisons, $\alpha = 0.05$.

$z_{1-[\alpha/2g]} = z_{1-[0.05/(2\times6)]} = z_{0.9958} \Rightarrow z_{0.9959} = -2.64$

Note: If you are ever between two values on the z-table, choose the higher value since it makes it harder to reject H_0.

$\overline{R_1} = \frac{T_{R_1}}{n_1} = \frac{61.5}{6} = 10.25$

$\overline{R_2} = \frac{T_{R_2}}{n_2} = \frac{84}{7} = 12$

$\overline{R_3} = \frac{T_{R_3}}{n_3} = \frac{44.5}{6} = 7.4167$

$\overline{R_4} = \frac{T_{R_4}}{n_4} = \frac{86}{4} = 21.5$

<u>95% CI for $\tilde{\mu_1} - \tilde{\mu_2}$</u>

$= (10.25 - 12) \pm (-2.64)\sqrt{\frac{(23)(23+1)}{12}(\frac{1}{6})(\frac{1}{7})}$

$= -1.75 \pm 9.9616$

$= (-11.7116, 8.2116)$ ← no difference ($\tilde{\mu_1} = \tilde{\mu_2}$)

<u>95% CI for $\tilde{\mu_1} - \tilde{\mu_3}$</u>

$= (10.25 - 7.4167) \pm (-2.64)\sqrt{\frac{(23)(23+1)}{12}(\frac{1}{6})(\frac{1}{6})}$

$= 2.8333 \pm 10.3377$

$= (-7.5044, 13.171)$ ← no difference ($\tilde{\mu_1} = \tilde{\mu_3}$)

<u>95% CI for $\tilde{\mu_1} - \tilde{\mu_4}$</u>

$= (10.25 - 21.5) \pm (-2.64)\sqrt{\frac{(23)(23+1)}{12}(\frac{1}{6})(\frac{1}{4})}$

$= -11.25 \pm 11.5579$

$= (-22.8079, 0.3079)$ ← no difference ($\tilde{\mu_1} = \tilde{\mu_4}$)

Lecture 16 (Mar 7 2024)

$\underline{95\% \text{ CI for } \tilde{\mu_2} - \tilde{\mu_3}}$

$= (12 - 7.4167) \pm (-2.64)\sqrt{\frac{(23)(23+1)}{12}(\frac{1}{7})(\frac{1}{6})}$ ← we've calculated this value already!

$= 4.5833 \pm 9.9616$

$= (-5.3783, 17.4946)$ ← no difference ($\tilde{\mu_2} = \tilde{\mu_3}$)

$\underline{95\% \text{ CI for } \tilde{\mu_2} - \tilde{\mu_4}}$

$= (12 - 21.5) \pm (-2.64)\sqrt{\frac{(23)(23+1)}{12}(\frac{1}{7})(\frac{1}{4})}$

$= -9.5 \pm 11.2228$

$= (-20.7228, 1.7228)$ ← no difference ($\tilde{\mu_2} = \tilde{\mu_4}$)

$\underline{95\% \text{ CI for } \tilde{\mu_3} - \tilde{\mu_4}}$

$= (7.4167 - 21.5) \pm (-2.64)\sqrt{\frac{(23)(23+1)}{12}(\frac{1}{6})(\frac{1}{4})}$ ← we've calculated this value already!

$= 14.0833 \pm 11.5579$

$= (-25.6412, -2.5254)$ ← difference exists ($\tilde{\mu_3} < \tilde{\mu_4}$)

Therefore, we can say with 95% confidence that ① the median score under Method 3 is lower than that under Method 4 and ② there are no differences between the medians from the remaining pairwise comparisons.

RBD

Randomized Block Design (Two-Way)

- In CRD, it is assumed that any variation is due to random variation or experimental error.
- If another source of variability is identified, this source will be used as a blocking factor. This type of design is known as RBD.
- Note: All assumptions required for CRD are also required for RBD.
- In RBD:
  - The blocking factor will have b blocks.
  - The number of treatments is k.
  - The total number of experiments is $n = bk$.
- The total variation is partitioned into 3 parts: $TSS = SSB + SST + SSE$.
  - SSB (sum of squares for blocks) measures variation between the block means.
  - SST (sum of squares for treatments) measures variation between the treatment means.
  - SSE (sum of squares for error) measures variation of the differences among the treatment observations within blocks.

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{b} {y_{ij}}^2 - \frac{(G.T.)^2}{bk}$$

$$SST = \sum_{i=1}^{k} \frac{{T_i}^2}{b} - \frac{(G.T.)^2}{bk} \text{ where } T_i \text{ is the total of the observations for treatment } i$$

$$SSB = \sum_{j=1}^{b} \frac{{B_j}^2}{k} - \frac{(G.T.)^2}{bk} \text{ where } B_j \text{ is the total of the observations for block } j$$

$$SSE = TSS - SST - SSB$$

ANOVA Table for RBD with k treatments and b blocks

Source	df	SS	MS	F
Treatments	$k-1$	SST	$MST = \frac{SST}{k-1}$	$F_T = \frac{MST}{MSE}$
Blocks	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Error	$(k-1)(b-1)$	SSE	$MSE = \frac{SSE}{(k-1)(b-1)}$	
Total	$n-1$ or $bk-1$	TSS		

Test treatment means for RBD

1. State hypothesis.
   - $H_0 : \mu_1 = \mu_2 = ... = \mu_k$
   - $H_\alpha =$ At least one μ_i is different $(i = 1, 2, ..., k)$
2. Compute F-statistic, F_T.
   - By creating ANOVA table or using SAS
3. Test rejection region.
   - Reject H_0 if $F_T > F_{a,k-1,(k-1)(b-1)}$

Test effectiveness/usefulness of blocking factor for RBD

1. State hypothesis.
   - $H_0 : \mu_1 = \mu_2 = ... = \mu_b$
   - $H_\alpha =$ At least one μ_j is different $(j = 1, 2, ..., b)$
2. Compute F-statistic, F_B.
   - By creating ANOVA table or using SAS
3. Test rejection region.
   - Reject H_0 if $F_B > F_{a,b-1,(k-1)(b-1)}$

Exercise 1 (by hand)

A researcher conducted an experiment to compare the effects of three different insecticides on a variety of string beans. To obtain a sufficient amount of data, it was necessary to use four different plots of land. Since the plots had somewhat different soil fertility, drainage characteristics, and sheltering from winds, the researcher decided to conduct a randomized block design with plots serving as blocks. Each plot was subdivided into three rows. A suitable distance was maintained between rows within a plot so that the insecticides could be confined to a particular row. Each row was planted with 100 seeds and then maintained under the insecticide assigned to the row. The insecticides were randomly assigned to the rows within a plot so that each insecticide appeared in one row within all four plots. The number of seedlings that emerged per row were then recorded. The data are summarized below:

	Plot			
Insecticide	**1**	**2**	**3**	**4**
1	56	48	66	62
2	83	78	94	93
3	80	72	83	85

Treatment means: insecticide $\leftarrow k = 3$

Blocking variable: plot $\leftarrow b = 4$

$$T_1 = 56 + 48 + 66 + 62 = 232$$
$$T_2 = 83 + 78 + 94 + 93 = 348$$
$$T_3 = 80 + 72 + 83 + 85 = 320$$

$$B_1 = 56 + 83 + 80 = 219$$
$$B_2 = 48 + 78 + 72 = 198$$
$$B_3 = 66 + 94 + 83 = 243$$
$$B_4 = 62 + 93 + 85 = 240$$

Note: rows and columns may be flipped.
The rows are not always the treatment means,
and the columns are not always the blocking variable.
You need to be able to differentiate/identify them!

$$GT = \sum_{i=1}^{3}\sum_{j=1}^{4} y_{ij} = T_1 + T_2 + T_3 = 232 + 348 + 320 = 900$$

Note: you could also add up $B_1 + B_2 + B_3 + B_4$ to get GT. We added $T_1 + T_2 + T_3$ because it's quicker to do (adding 3 numbers vs. 4 numbers).

$$\begin{aligned} TSS &= \sum_{i=1}^{3}\sum_{j=1}^{4} {y_{ij}}^2 - \frac{(GT)^2}{bk} \\ &= 56^2 + 48^2 + \ldots + 85^2 - \frac{900^2}{(4)(3)} \\ &= 69796 - 67500 \\ &= 2296 \end{aligned}$$

$$SST = \sum_{i=1}^{3} \frac{{T_i}^2}{b} - \frac{(GT)^2}{bk} = \frac{232^2}{4} + \frac{348^2}{4} + \frac{320^2}{4} - 67500 = 1832$$

$$SSB = \sum_{j=1}^{4} \frac{{B_j}^2}{k} - \frac{(GT)^2}{bk} = \frac{219^2}{3} + \frac{198^2}{3} + \frac{243^2}{3} + \frac{240^2}{3} - 67500 = 438$$

$$SSE = TSS - SST - SSB = 2296 - 1832 - 438 = 26$$

ANOVA Table

Source	df	SS	MS	F
Treatment	$k-1$ $= 3-1 = 2$	1832	$\frac{1832}{2} = 916$	$F_T = \frac{916}{4.3333} = 211.3862$
Block	$b-1$ $= 4-1 = 3$	438	$\frac{438}{3} = 146$	$F_B = \frac{146}{4.3333} = 33.6926$
Error	$(k-1)(b-1)$ $= (2)(3) = 6$	26	$\frac{26}{6} = 4.3333$	
Total	$bk-1$ $= (4)(3) - 1 = 11$	2296		

A. At $\alpha = 0.05$, is there enough evidence that the effectiveness of the three insecticides differs?

$H_0 : \mu_1 = \mu_2 = \mu_3,\ H_\alpha : \text{At least one } \mu_i \text{ differs } (i = 1, 2, 3)$

$F_{0.05,2,6} = 5.14$

Since $211.3862 > 5.14$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that average yield differs by insecticide.

B. Was using plot as a blocking variable the right decision? Use $\alpha = 0.05$.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4,\ H_\alpha : \text{At least one } \mu_j \text{ differs } (j = 1, 2, 3, 4)$

$F_{0.05,3,6} = 4.76$

Since $33.6926 > 4.76$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that using plot as a blocking variable was useful.

Note: If we conclude that the blocking variable is not useful, the result from A is still valid. It just means that the next time we test the effectiveness of the insecticides, we should ① choose a different blocking variable or ② do one-way CRD instead. (This will not be tested)

Exercise 1 (using SAS output)

Dependent Variable: seed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2270.000000	454.000000	104.77	<.0001
Error	6	26.000000	4.333333		
Corrected Total	11	2296.000000			

R-Square	Coeff Var	Root MSE	seed Mean
0.988676	2.775555	2.081666	75.00000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
plot	3	438.000000	146.000000	33.69	0.0004
insecticide	2	1832.000000	916.000000	211.38	<.0001

A. At $\alpha = 0.05$, is there enough evidence that the effectiveness of the three insecticides differs?

$H_0 : \mu_1 = \mu_2 = \mu_3,\ \ H_\alpha : \text{At least one } \mu_i \text{ differs } (i = 1, 2, 3)$

$\text{p-value} < .0001$

Since $\text{p-value} < 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that average yield differs by insecticide.

B. Was using plot as a blocking variable the right decision? Use $\alpha = 0.05$.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4,\ \ H_\alpha : \text{At least one } \mu_j \text{ differs } (j = 1, 2, 3, 4)$

$\text{p-value} = 0.0004$

Since $0.0004 < 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that using plot as a blocking variable was useful.

Relative Efficiency (R.E.)

- A measure of how much more efficient an RBD is compared to a CRD.
- $$R.E. = \frac{(b-1)MSB + b(k-1)MSE}{(kb-1)MSE}$$
- If $R.E. > 1$, then the RBD is more efficient than the CRD.

Exercise 2

Compute the *R.E.* for the application in Exercise 1 and comment on the result you obtain.

$$R.E. = \frac{(4-1)(146) + (4)(3-1)(4.3333)}{[(3)(4)-1](4.3333)} = 9.9162$$

Since $R.E. > 1$, RBD is more efficient than CRD.

Tukey's Simultaneous Pairwise Comparisons of Treatment Means

The $(1-\alpha)100$ Tukey's CI for $\mu_i - \mu_j$ is given by $\overline{y_i} - \overline{y_j} \pm q_{a,k,(b-1)(k-1)}\sqrt{\frac{MSE}{b}}$.
Notice that the margin of error (in blue) is constant. Doing multiple pairwise comparisons is quick since we only need to calculate this value once.

<u>Exercise 3 (by hand)</u>
Find all Tukey pairwise comparisons for the treatment means in Exercise 1.
There are 3 treatment means, so we are finding ${}_3C_2 = 3$ confidence intervals.

$q_{0.05,3,(4-1)(3-1)} = q_{0.05,3,6} = 4.34$ ← From Tukey table

$ME = 4.34\sqrt{\frac{4.3333}{4}} = 4.5172$

Recall from Exercise 1 that $T_1 = 232$, $T_2 = 348$, $T_3 = 320$.

$\overline{y_1} = \frac{T_1}{b} = \frac{232}{4} = 58$ $\quad$ $\overline{y_2} = \frac{T_2}{b} = \frac{348}{4} = 87$ $\quad$ $\overline{y_3} = \frac{T_3}{b} = \frac{320}{4} = 80$

$\underline{95\% \text{ CI for } \mu_1 - \mu_2}$
$= (58 - 87) \pm 4.5172$
$= -29 \pm 4.5172$
$= (-33.5172, -24.4828)$ $\quad \leftarrow \mu_1 < \mu_2$

$\underline{95\% \text{ CI for } \mu_1 - \mu_3}$
$= (58 - 80) \pm 4.5172$
$= -22 \pm 4.5172$
$= (-26.5172, -17.4828)$ $\quad \leftarrow \mu_1 < \mu_3$

$\underline{95\% \text{ CI for } \mu_2 - \mu_3}$
$= (87 - 80) \pm 4.5172$
$= 7 \pm 4.5172$
$= (2.4828, 11.5172)$ $\quad \leftarrow \mu_2 > \mu_3$

Conclusion
Therefore, we can say with 95% confidence that ① the mean yield under insecticide 1 is lower than that under insecticide 2, ② the mean yield under insecticide 1 is lower than that under insecticide 3, ③ the mean yield under insecticide 2 is higher than that under insecticide 3.
Note 1: you need to discuss all of the confidence intervals in one paragraph (not separate statements) because you are confident in them <u>simultaneously</u>.

Exercise 3 (using SAS output)

```
Tukey's Studentized Range (HSD) Test for seed

 NOTE: This test controls the Type I experimentwise error rate.


          Alpha                                      0.05
          Error Degrees of Freedom                      6
          Error Mean Square                      4.333333
          Critical Value of Studentized Range     4.33917
          Minimum Significant Difference           4.5164
```

```
Comparisons significant at the 0.05 level are indicated by ***.


                     Difference
      insecticide       Between     Simultaneous 95%
      Comparison          Means    Confidence Limits

         2 - 3            7.000       2.484   11.516  ***
         2 - 1           29.000      24.484   33.516  ***
         3 - 2           -7.000     -11.516   -2.484  ***
         3 - 1           22.000      17.484   26.516  ***
         1 - 2          -29.000     -33.516  -24.484  ***
         1 - 3          -22.000     -26.516  -17.484  ***
```

Notes about SAS output for Tukey test:

- SAS thinks that order matters, so it does too many comparisons. Half of these are duplicates.
- When analyzing this output, the convention is to use the comparisons that have the smaller number first. For example, choose 2-3 instead of 3-2.
- Notice that all of the comparisons have *** on the right side. This means that there is a difference between means, i.e. the confidence interval does not include zero in it.

$$\underline{95\% \text{ CI for } \mu_1 - \mu_2}$$
$$= (-33.516, -24.484) \qquad \leftarrow \mu_1 < \mu_2$$

$$\underline{95\% \text{ CI for } \mu_1 - \mu_3}$$
$$= (-26.516, -17.484) \qquad \leftarrow \mu_1 < \mu_3$$

$$\underline{95\% \text{ CI for } \mu_2 - \mu_3}$$
$$= (2.484, 11.516) \qquad \leftarrow \mu_2 > \mu_3$$

Therefore, we can say with 95% confidence that ① the mean yield under insecticide 1 is lower than that under insecticide 2, ② the mean yield under insecticide 1 is lower than that under insecticide 3, ③ the mean yield under insecticide 2 is higher than that under insecticide 3.

Friedman Test (for RBD; nonparametric test)

- This test is the RBD equivalent of the Kruskal-Wallis test (which is used for CRD).
- This test can be used when normality and/or equality of variance are violated.

In order to perform a Friedman test, the data must be ranked and summed as follows:

1. Rank the k treatments within each block from smallest to largest.
2. If there is a tie, the repeated values take on the average rank.
3. Sum the ranks for each treatment to get T_{R_i}.

Then we can perform the hypothesis test:

1. State hypothesis.
   - $H_0: \text{ The medians are equal}$
   - $H\alpha: \text{ At least one median is different}$
2. Compute the test statistic, F_R.

$$F_R = \frac{12}{bk(k+1)}\left[\sum_{i=1}^{k} {T_{R_i}}^2\right] - 3b(k+1)$$

3. Test rejection region.
   - Reject H_0 if $F_R > \chi^2_{a,k-1}$

Note: If $\sum_{i=1}^{k} T_{R_i} = \frac{bk(k+1)}{2}$, your ranks are correct. Use this to double check your work!

<u>Exercise 1 (by hand)</u>

Suppose you wish to <u>compare the reaction times of people exposed to six different stimuli</u>. A reaction time measurement is obtained by subjecting a person to a stimulus and measuring the time until the person presents some specified reaction. The objective of the experiment is to determine whether differences exist in the reaction times to the stimuli used in the experiment. To <u>eliminate the person-to-person variation in reaction time</u>, four persons participated in the experiment and each person's reaction time (in seconds) was measured for each of the six stimuli. The data are summarized in the table below.

Use Friedman's test to determine if the data present sufficient evidence to indicate differences in the distributions of reaction times for the six stimuli. Use α = 0.05.

Treatment: stimuli (hint: "compare the reaction times of people exposed to six different stimuli")

Blocking var: people (hint: "eliminate the person-to-person variation in reaction time")

$k = 6$ and $b = 4$

IMPORTANT: Rank this → way (in block)! Sum this ↓ way (in treatment)!

Note: in the table below, the data is in black and their ranks are in grey.

	Stimulus (T)						
Subject (B)	**A**	**B**	**C**	**D**	**E**	**F**	
1	0.6 2.5	0.9 6	0.8 5	0.7 4	0.5 1	0.6 2.5	$\frac{2+3}{2} = 2.5$
2	0.7 3.5	1.1 6	0.7 3.5	0.8 5	0.5 1.5	0.5 1.5	$\frac{3+4}{2} = 3.5$
3	0.9 3	1.3 6	1.0 4.5	1.0 4.5	0.7 2	0.8 1	$\frac{1+2}{2} = 1.5$
4	0.5 2	0.7 5	0.8 6	0.6 3.5	0.4 1	0.6 3.5	$\frac{4+5}{2} = 4.5$
							$\frac{4+5}{2} = 4.5$

$T_{R_1} = 2.5 + 3.5 + 3 + 2 = 11$
$T_{R_2} = 6 + 6 + 6 + 5 = 23$
$T_{R_3} = 5 + 3.5 + 4.5 + 6 = 19$
$T_{R_4} = 4 + 5 + 4.5 + 3.5 = 17$
$T_{R_5} = 1 + 1.5 + 1 + 1 = 4.5$
$T_{R_6} = 2.5 + 1.5 + 2 + 3.5 = 9.5$

<u>Double check ranks</u> (optional)

$$\sum_{i=1}^{k} T_{R_i} = \frac{bk(k+1)}{2}$$

$$11 + 23 + 19 + 17 + 4.5 + 9.5 = \frac{(4)(6)(6+1)}{2}$$

$$84 = 84 \checkmark$$

<u>State hypothesis</u>

$H_0 : \tilde{\mu_1} = \tilde{\mu_2} = ... = \tilde{\mu_6}, \quad H_\alpha : \text{ At least one } \tilde{\mu_j} \text{ differs } (j = 1, 2, ..., 6)$

<u>Calculate test statistic</u>

$$F_R = \frac{12}{(4)(6)(6+1)} \left[11^2 + ... + 9.5^2\right] - 3(4)(6+1) = 16.75$$

<u>Test rejection region</u>

$\chi^2{}_{0.05,6-1} = \chi^2{}_{0.05,5} = 11.0705$

Since $16.75 > 11.0705$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that median reaction time differs by stimulus.

Exercise 1 (using SAS)

```
            Summary Statistics for stimulus by time
                    Controlling for subject

   Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1      4.4709    0.0345
    2        Row Mean Scores Differ     5     17.3704    0.0038
```

Note: the value of F_R given by SAS ($F_R = 17.3704$) is not exactly the same as the one we calculated by hand. That's ok, and we will always just use the p-value instead.

$$\text{p-value} = 0.0038$$

Since $0.0038 < 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that median reaction time differs by stimulus.

Simultaneous Pairwise Comparisons for RBD

- If the Friedman's test says that factor level medians differ, we want to know how they differ.
- Given g possible pairwise comparisons, the CIs with a <u>family level</u> of at least $100(1-\alpha)\%$ can be calculated using $(\overline{R_i} - \overline{R_j}) \pm z_{\alpha/2g} \cdot \sqrt{\dfrac{k(k+1)}{6b}}$.

Exercise 2 (altered from slide question)

Find <u>three</u> (not all) 95% pairwise comparisons for the application in Exercise 1.

$g = 3$

Note: Exercise 2 in the slides asks for <u>all</u> pairwise comparisons. There are $C_6^3 = 15$ possible pairwise comparisons. If we are solving the original question, then $g = 15$.

<u>Find z-value, calculate margin of error</u>

$$z_{0.05/2(3)} = z_{0.05/6} = z_{0.008333} \Rightarrow z_{0.0082} = -2.4$$

$$\text{Margin of error} = -2.4\sqrt{\frac{6(6+1)}{(6)4}} = -3.1749$$

<u>Calculate required $\overline{R_i}$ values</u>

$$\overline{R_1} = \frac{T_{R_1}}{b} = \frac{11}{4} = 2.75 \qquad \overline{R_2} = \frac{T_{R_2}}{b} = \frac{23}{4} = 5.75$$

$$\overline{R_3} = \frac{T_{R_3}}{b} = \frac{19}{4} = 4.75 \qquad \overline{R_4} = \frac{T_{R_4}}{b} = \frac{17}{4} = 4.25$$

$\underline{95\% \text{ CI for } \widetilde{\mu}_1 - \widetilde{\mu}_4}$
$= (2.75 - 4.25) \pm (-3.1749)$
$= -1.5 \pm (-3.1749)$
$= (-4.6749, 1.6749) \qquad \leftarrow \widetilde{\mu}_1 = \widetilde{\mu}_2$

$\underline{95\% \text{ CI for } \widetilde{\mu}_2 - \widetilde{\mu}_3}$
$= (5.75 - 4.75) \pm (-3.1749)$
$= 1 \pm (-3.1749)$
$= (-2.1749, 4.1749) \qquad \leftarrow \widetilde{\mu}_2 = \widetilde{\mu}_3$

$\underline{95\% \text{ CI for } \widetilde{\mu}_2 - \widetilde{\mu}_4}$
$= (5.75 - 4.25) \pm (-3.1749)$
$= 1.5 \pm (-3.1749)$
$= (-1.6749, 4.6749) \qquad \leftarrow \widetilde{\mu}_2 = \widetilde{\mu}_4$

Therefore, we can say with 95% confidence that there are no differences from the three CIs calculated.

2WAY

Two-Way Classification ($\mathrm{a} \times \mathrm{b}$ Factorial Experiment)

- The response variable is influenced by two factors (Factor A, Factor B).
- We are interested in knowing if the means differ by both factors together.
- There will be ab treatments since the treatments are now a combination of the levels of both factors.
- Each treatment will have r experimental units.

In this example, there are $2 \times 3 = 6$ treatments.

$b = 3$

$a = 2$

1 1	1 2	1 3
2 1	2 2	2 3

$\leftarrow SSA$ differs by row.

$\uparrow SSB$ differs by column. $\nwarrow SSAB$ differs by treatment (each box).

Two-Way ANOVA

The two-way model is $y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, $\varepsilon_{jik} \overset{i.i.d}{\sim} N(0, \sigma^2)$.

Note: All assumptions required for CRD are also required for 2-Way Classification.

The total variation is split into 4 parts: $TSS = SSA + SSB + SSAB + SSE$

- SSA is the sum of squares for Factor A (variation among Factor A means)
- SSB is the sum of squares for Factor B (variation among Factor B means)
    - Note: SSB is <u>not</u> sum of squares for blocks; don't mix them up!
- $SSAB$ is the sum of squares for interaction (variation among treatment means)
- SSE is the sum of squares for error (variation of the differences among the observations within each treatment)

$$TSS = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} {y_{ijk}}^2 - \frac{(GT)^2}{abr}$$

$$SSA = \sum_{i=1}^{a} \frac{{A_i}^2}{br} - \frac{(GT)^2}{abr}$$

$$SSB = \sum_{j=1}^{b} \frac{{B_j}^2}{ar} - \frac{(GT)^2}{abr}$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{{AB_{ij}}^2}{r} - \frac{(GT)^2}{abr} - SSA - SSB$$

$$SSE = TSS - SSA - SSB - SSAB$$

- $A_i = \sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$ is the total of the observations at the ith level of Factor A
- $B_j = \sum_{i=1}^{a}\sum_{k=1}^{r} y_{ijk}$ is the total of the observations at the jth level of Factor B
- $AB_{ij} = \sum_{k=1}^{r} y_{ijk}$ is the total of the r observations at the ith level of Factor A and the jth level of Factor B

ANOVA Table

Source	df	SS	MS	F
Factor A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$
Factor B	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Interaction	$(a-1)(b-1)$	$SSAB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error	$ab(r-1)$	SSE	MSE	
Total	$n-1 = abr-1$	TSS		

Testing Interaction

1. State hypothesis.
    - H_0: Factor A and Factor B do not interact
    - H_α: Factor A and Factor B interact
2. Compute test statistic, F_{AB}.
    - From ANOVA table
3. Test rejection region.
    - Reject H_0 if $F_{AB} > F_{\alpha,(a-1)(b-1),ab(r-1)}$

If interaction is **not** significant (failed to reject H_0), it is necessary to test the main effects.

Testing Factor A

1. State hypothesis.
    - $H_0 : \mu_{1.} = \mu_{2.} = ... = \mu_{a.}$
    - $H_\alpha : \text{at least one } \mu_{i.} \text{ is different } (i = 1, 2, ..., a)$
2. Compute test statistic, F_A.
    - From ANOVA Table
3. Test rejection region.
    - Reject H_0 if $F_A > F_{\alpha,a-1,ab(r-1)}$

Testing Factor B

1. State hypothesis.
    - $H_0 : \mu_{.1} = \mu_{.2} = ... = \mu_{.b}$
    - $H_\alpha : \text{at least one } \mu_{.j} \text{ is different } (j = 1, 2, ..., b)$
2. Compute test statistic, F_B.
    - From ANOVA Table
3. Test rejection region.
    - Reject H_0 if $F_B > F_{\alpha,b-1,ab(r-1)}$

Lecture 18 (Mar 14 2024)

Exercise 1 (by hand)

Suppose that two supervisors are each observed on three randomly selected days for each of three different shifts. The production was recorded in terms of units produced. The data is in the table below.

A: supervisor (2 levels: Super1, Super2) $a = 2$
B: Shift (3 levels: Day, Swing, Night) $b = 3$
Treatments: 1D, 1S, 1N, 2D, 2S, 2N $r = 3$ (num observations per treatment)

	Shift		
Supervisor	**Day**	**Swing**	**Night**
1	571 610 625	480 474 540	470 430 450
2	480 516 465	625 600 680	630 680 661

$A_1 = AB_{11} + AB_{12} + AB_{13} = 1806 + 1494 + 1350 = 4650$

$A_2 = AB_{21} + AB_{22} + AB_{23} = 1461 + 1806 + 1971 = 5238$

$B_1 = AB_{11} + AB_{21} = 1806 + 1461 = 3267$
$B_2 = AB_{12} + AB_{22} = 1494 + 1806 = 3300$
$B_3 = AB_{13} + AB_{23} = 1350 + 1971 = 3321$

$AB_{11} = 571 + 610 + 625 = 1806$
$AB_{12} = 480 + 474 + 540 = 1494$
$AB_{13} = 470 + 430 + 450 = 1350$
$AB_{21} = 480 + 516 + 465 = 1461$
$AB_{22} = 625 + 600 + 581 = 1806$
$AB_{23} = 630 + 680 + 661 = 1971$

$GT = A_1 + A_2 = 4650 + 5238 = 9888$ ← less work

or

$GT = B_1 + B_2 + B_3 = 3267 + 3300 + 3321 = 9888$

$$TSS = 571^2 + ... + 661^2 - \frac{9888^2}{(2)(3)(3)} = 109222$$

$$SSA = \frac{4650^2}{9} + \frac{5238^2}{9} - 5431808 = 19208$$

$$SSB = \frac{3267^2}{6} + \frac{3300^2}{6} + \frac{3321^2}{6} - 5431808 = 247$$

$$SSAB = \frac{1806^2}{3} + + \frac{1971^2}{3} - 5431808 - 19208 - 247 = 81127$$

$$SSE = 109222 - 19208 - 247 - 81127 = 8640$$

Source	df	SS	MS	F
Factor A (Supervisor)	$2-1=1$	19208	$\frac{19208}{1}=19208$	$\frac{19208}{720}=26.6778$
Factor B (Shift)	$3-1=2$	247	$\frac{247}{2}=123.5$	$\frac{123.5}{720}=0.1715$
Interaction	$(1)(2)=2$	81127	$\frac{81127}{2}=40563.5$	$\frac{40563.5}{720}=56.3382$
Error	$17-1-2-2=12$	8640	$\frac{8640}{12}=720$	
Total	$(2)(3)(3)-1=17$	109222		

A. At $\alpha = 0.05$, is there enough evidence to conclude that there is interaction between supervisor and shift?

$H_0: \widetilde{\mu_{11}} = \widetilde{\mu_{12}} = ... = \widetilde{\mu_{23}}$ or "no interaction between supervisor and shift"

$H_\alpha:$ At least one $\widetilde{\mu_{ij}}$ differs $(i = 1, 2, j = 1, 2, 3)$

$F_{0.05,2,12} = 3.89$

Since $56.3382 > 3.89$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that there is interaction between supervisor and shift.

B. If appropriate, test the main factor effects. Use $\alpha = 0.05$.

<u>Testing supervisor (Factor A)</u>

$H_0: \mu_{1.} = \mu_{2.}, \quad H_\alpha: \mu_{1.} \neq \mu_{2.}$

Note: the dot (.) in the subscript means to sum over j. No penalty for forgetting the dot. This helps us differentiate between having only one factor (which would be written as μ_1).

$F_{0.05,1,12} = 4.75$

Since $26.6778 > 4.75$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that average production differs by supervisor.

<u>Testing shift (Factor B)</u>

$H_0: \mu_{.1} = \mu_{.2} = \mu_{.3}, \quad H_\alpha:$ At least one $\mu_{.j}$ differs $(j = 1, 2, 3)$

Note: the dot (.) in the subscript means to sum over i. No penalty for forgetting the dot.

$F_{0.05,2,12} = 3.89$

Since $0.1715 > 3.89$, at $\alpha = 0.05$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that average production differs by shift.

Exercise 1 (using SAS)

The ANOVA Procedure

Class Level Information

Class	Levels	Values
super	2	1 2
shift	3	day night swing

Number of Observations Read	18
Number of Observations Used	18

The ANOVA Procedure

Dependent Variable: product

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	100582.0000	20116.4000	27.94	<.0001
Error	12	8640.0000	720.0000		
Corrected Total	17	109222.0000			

R-Square	Coeff Var	Root MSE	product Mean
0.920895	4.884615	26.83282	549.3333

Source	DF	Anova SS	Mean Square	F Value	Pr > F
super	1	19208.00000	19208.00000	26.68	0.0002
shift	2	247.00000	123.50000	0.17	0.8444
super*shift	2	81127.00000	40563.50000	56.34	<.0001

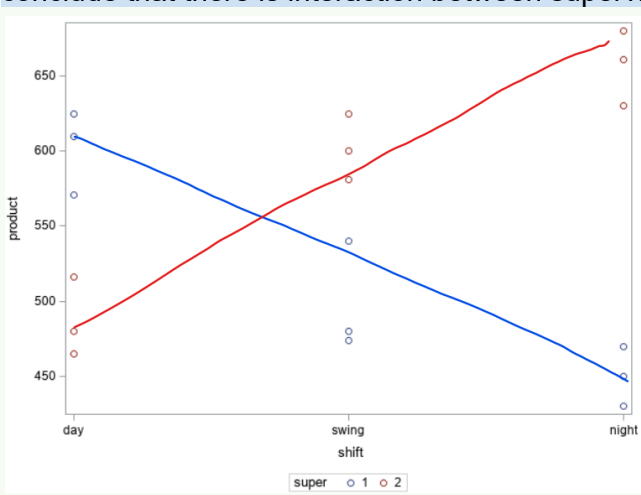
A. At $\alpha = 0.05$, is there enough evidence to conclude that there is interaction between supervisor and shift?

$H_0 : \widetilde{\mu_{11}} = \widetilde{\mu_{12}} = ... = \widetilde{\mu_{23}}$ or "no interaction between supervisor and shift"

$H_\alpha :$ At least one $\widetilde{\mu_{ij}}$ differs $(i = 1, 2, j = 1, 2, 3)$

$\text{p-value} < .0001$

Since $\text{p-value} < 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that there is interaction between supervisor and shift.



Graphically: "Since the two lines are not parallel, there is interaction between supervisor and shift".

B. If appropriate, test the main factor effects. Use $\alpha = 0.05$.

$\underline{\text{Testing supervisor (Factor A)}}$

$H_0 : \mu_{1.} = \mu_{2.}, \;\; H_\alpha : \mu_{1.} \neq \mu_{2.}$

$\text{p-value} = 0.0002$

Since $0.0002 < 0.05$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that average production differs by supervisor.

$\underline{\text{Testing shift (Factor B)}}$

$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3}, \;\; H_\alpha : \text{ At least one } \mu_{.j} \text{ differs } (j = 1, 2, 3)$

$\text{p-value} = 0.8444$

Since $0.8444 > 0.05$, at $\alpha = 0.05$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that average production differs by shift.

Tukey's Simultaneous Pairwise Comparisons (for Interaction)

- If interactions are significant: $(\overline{Y_{ij.}}) - (\overline{Y_{i'j'.}}) \pm q_{\alpha,ab,ab(r-1)}\sqrt{\dfrac{MSE}{r}}$
- If interactions are not significant,
  - compute for Factor A: $(\overline{Y_{i..}}) - (\overline{Y_{i'..}}) \pm q_{\alpha,a,ab(r-1)}\sqrt{\dfrac{MSE}{br}}$
  - compute for Factor B: $(\overline{Y_{.j.}}) - (\overline{Y_{.j'.}}) \pm q_{\alpha,b,ab(r-1)}\sqrt{\dfrac{MSE}{ar}}$

<u>Exercise 2</u>

Consider Exercise 1. Find all necessary 95% Tukey's simultaneous pairwise comparisons.

There are ${}_6C_2 = 15$ possible pairwise comparisons. We will only be calculating 3, to save time.

$q_{0.05,(2)(3),(2)(3)(3-1)} = q_{0.05,6,12} = 4.75$

$$ME = 4.75\sqrt{\frac{MSE}{r}} = 4.75\sqrt{\frac{720}{3}} = 73.5867$$

$$\overline{Y_{11.}} = \frac{AB_{11}}{r} = \frac{1806}{3} = 602 \qquad \overline{Y_{13.}} = \frac{AB_{13}}{r} = \frac{1350}{3} = 450$$

$$\overline{Y_{12.}} = \frac{AB_{12}}{r} = \frac{1494}{3} = 498 \qquad \overline{Y_{21.}} = \frac{AB_{21}}{r} = \frac{1461}{3} = 487$$

$\underline{95\% \text{ CI for } \mu_{11} - \mu_{12}}$

$= (602 - 498) \pm 73.5867$

$= 104 \pm 73.5867$

$= (30.4133, 177.5867) \qquad \leftarrow \mu_{11} > \mu_{12}$

$\underline{95\% \text{ CI for } \mu_{11} - \mu_{13}}$

$= (602 - 450) \pm 73.5867$

$= 152 \pm 73.5867$

$= (78.4133, 225.5867) \qquad \leftarrow \mu_{11} > \mu_{13}$

$\underline{95\% \text{ CI for } \mu_{11} - \mu_{21}}$

$= (602 - 487) \pm 73.5867$

$= 115 \pm 73.5867$

$= (41.4133, 188.5867) \qquad \leftarrow \mu_{11} > \mu_{21}$

… and so on…

The SAS output below includes all 15 CIs. The three that we calculated are highlighted.

Least Squares Means for Effect super*shift

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	152.000000	78.410152	225.589848
1	3	104.000000	30.410152	177.589848
1	4	115.000000	41.410152	188.589848
1	5	-55.000000	-128.589848	18.589848
1	6	0	-73.589848	73.589848
2	3	-48.000000	-121.589848	25.589848
2	4	-37.000000	-110.589848	36.589848
2	5	-207.000000	-280.589848	-133.410152
2	6	-152.000000	-225.589848	-78.410152
3	4	11.000000	-62.589848	84.589848
3	5	-159.000000	-232.589848	-85.410152
3	6	-104.000000	-177.589848	-30.410152
4	5	-170.000000	-243.589848	-96.410152
4	6	-115.000000	-188.589848	-41.410152
5	6	55.000000	-18.589848	128.589848

Lecture 19 (Mar 19 2024)

[notes have not been typed out]

Lecture 20 (Mar 21 2024)

Review: Qualitative variables (must have ≥ 3 levels)

1. If at least one p-value $< \alpha$, QV is significant.
2. If all p-values $> \alpha$, do partial F-test.
3. For interaction, 1 and 2 hold. Also, if an X interaction with QV is significant, X and QV should stay in the model.

A. Is interaction between campaign and $x1$ useful in Model 1?
$H_0 : \beta_6 = \beta_7 = 0$
(unfinished in class)

Model 1 and Model 2 are not useful models because we failed to reject H_0 in B and C.

Lecture 21 (Mar 26 2024)

Chapter 14 - Part 1

Multinomial experiment

If the variable measured is the number of outcomes in each category (counts), the experiment is a multinomial experiment.

- Ex: number of M&Ms of each colour from a bag of M&Ms
- Ex: number of people classified into each of five income brackets
- Ex: number of people who scored each letter grade from F to A+

Recall, from STAT 2507, that binomial experiments have two outcomes (success or failure). Binomial experiments are a special case of multinomial experiments, so we are actually already familiar with the characteristics of a multinomial experiment.

Characteristics of the Multinomial Experiment

- The experiment consists of n identical trials.
- The trials are independent.
- The probability that the outcome of a single trial falls into category i is p_i and remains constant from trial to trial. The probability must be between 0 and 1 for each of the k categories, and the sum of all k probabilities is $\sum_{i=1}^{k} p_i = 1$.
- The experiment counts the *observed* number of outcomes in each category, denoted as $O_1, O_2, ..., O_k$ with $O_1 + O_2 + ... + O_k = n$.

χ^2 Goodness of Fit Test

This test is used to check if the sample proportions match the expected proportions.

1. State hypothesis.
   - $H_0 : p_1 = p_{10},\ p_2 = p_{20}, ...,\ p_k = p_{k0}$
   - $H_\alpha : \text{At least one } p_i \text{ is different from } p_{i0}$
2. Calculate the expected count for each category, E_i.
   - $E_i = np_{i0}$
3. Compute test statistic, χ^2.
   - $$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
4. Test rejection region.
   - Reject H_0 if $\chi^2 > \chi^2_{\alpha,k-1}$

Exercise 1

A researcher designs an experiment in which a rat is attracted to the end of a ramp that leads to doors of three different colours. The researcher sends the rat down the ramp $n = 90$ times and observes the choices listed in the table below. Does the rat have (or acquire) a preference for one of the three doors? Use $\alpha = 0.05$.

	Door		
	Green	Red	Blue
Count	20	39	31

$\Rightarrow k = 3$

State hypothesis

$$H_0 : p_1 = p_2 = p_3 = \frac{1}{3}, \quad H_\alpha : \text{At least one } p_i \neq \frac{1}{3} \ (i = 1, 2, 3)$$

Calculate expected count for each category

$$E_1 = E_2 = E_3 = \frac{1}{3}(90) = 30$$

Compute test statistic

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = \frac{(20-30)^2}{30} + \frac{(39-30)^2}{30} + \frac{(31-30)^2}{30} = 6.0667$$

Test rejection region

$$\chi^2_{0.05,3-1} = \chi^2_{0.05,2} = 5.99147$$

$0.0482 < 0.05 \leftarrow$ using SAS

Since $6.0667 > 5.99147$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that the rat has or has acquired a preference for colour.

```
                    The FREQ Procedure

                 Cumulative    Cumulative
door    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
  1          20       22.22            20        22.22
  2          39       43.33            59        65.56
  3          31       34.44            90       100.00


                    Chi-Square Test
                 for Equal Proportions

                 Chi-Square     6.0667
                 DF                  2
                 Pr > ChiSq     0.0482

                 Sample Size = 90
```

Exercise 2

Statistics Canada reported on the proportions of Canadian university teachers/professors by religion. According to Statistics Canada, the proportion of Roman Catholic (R), Christian (C), no religion (N), Jewish (J), Muslim (M), and other (O) professors in Canadian universities are 0.288, 0.282, 0.297, 0.046, 0.024, and 0.064, respectively. To determine whether the actual population proportions fit this set of reported probabilities, a random sample of 200 Canadian professors was selected and their religions were recorded. The observed counts are given in the table below. Test the goodness of fit of these proportions. Use $\alpha = 0.10$.

	Religion						
	R	C	N	J	M	O	$\Rightarrow k = 6$
Count	59	60	60	8	5	8	$\Rightarrow n = 200$

State hypothesis

$H_0 : p_1 = 0.288, p_2 = 0.282, p_3 = 0.297, p_4 = 0.046, p_5 = 0.024, p_6 = 0.064$

$H_\alpha = \text{At least one } p_i \text{ is different from what is stated } (i = 1, 2, ..., 6)$

Calculate expected count for each category

$E_1 = 200(0.288) = 57.6 \qquad E_4 = 200(0.046) = 9.2$

$E_2 = 200(0.282) = 56.4 \qquad E_5 = 200(0.024) = 4.8$

$E_3 = 200(0.297) = 59.4 \qquad E_6 = 200(0.064) = 12.8$

Compute test statistic

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(59 - 57.6)^2}{57.6} + ... + \frac{(8 - 12.8)^2}{12.8} = 2.2347$$

Test rejection region

$\chi^2_{0.10,6-1} = \chi^2_{0.10,5} = 9.23623$

$0.8158 > 0.10 \leftarrow$ using SAS

Since $2.2347 < 9.23623$, at $\alpha = 0.10$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that the distribution by religion is different from the numbers provided by Statistics Canada.

The FREQ Procedure

religion	Test Frequency	Cumulative Percent	Cumulative Percent	Frequency	Percent
R	59	29.50	28.80	59	29.50
C	60	30.00	28.20	119	59.50
N	60	30.00	29.70	179	89.50
J	8	4.00	4.60	187	93.50
M	5	2.50	2.40	192	96.00
O	8	4.00	6.40	200	100.00

Chi-Square Test
for Specified Proportions

Chi-Square	2.2347
DF	5
Pr > ChiSq	0.8158

Sample Size = 200

Chapter 14 - Part 2

Contingency Tables: Two-Way Classification

Some experimental units are classified by two variables. This results in a contingency table with r rows and c columns.

	Shift			
Type of Defects	1	2	3	Total
A	15	26	33	74
B	21	31	17	69
C	45	34	49	128
D	13	5	20	38
Total	94	96	119	309

- Ex: a defective piece of furniture is classified according to the type of defect and the production shift that made it
- Ex: a professor is classified by professional rank and type of university (public or private) at which the professor works
- Ex: a patient is classified according to the type of preventative flu treatment received and whether the patient contracted the flu during the winter or not

The χ^2 Test of Independence

This checks if the two variables which classify an experiential unit are independent.

1. State hypothesis.
   a. Through words:
      - $H_0 : \text{the two classifications/variables are independent}$
      - $H_\alpha : \text{the two classifications/variables are dependent}$
   b. Through notation:
      - $H_0 : p_{ij} = p_i p_j \text{ (for all } i \text{ \& } j)$
      - $H_\alpha : p_{ij} \neq p_i p_j \text{ (for at least some } i \text{ \& } j)$
2. Calculate the expected count for each category, $\hat{E_{ij}}$.
   - $\hat{E_{ij}} = \frac{r_i c_j}{n}$ where:
     - r_i is the row total for row i in the contingency table
     - c_j is the column total for column j in the contingency table
   - This is the estimated expected number of elements that would be classified in row i and column j of the contingency table if the two classifications are statistically independent (under H_0)
3. Compute test statistic, χ^2.
   - $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - \hat{E_{ij}})^2}{\hat{E_{ij}}}$ where:
     - O_{ij} is the observed count corresponding to row i and column j of the contingency table
4. Test rejection region.
   - Reject H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$

Exercise 1

Consider the furniture defects problem given as one of the examples above. At $\alpha = 0.05$, does the data provide sufficient evidence to indicate that the type of furniture defect varies with the shift during which the piece of furniture is produced?

	Shift				
Type of Defect	**1**	**2**	**3**	**Total**	
A	15	26	33	$r_1 = 74$	$\leftarrow r = 4,\ c = 3$
B	21	31	17	$r_2 = 69$	
C	45	34	49	$r_3 = 128$	
D	13	5	20	$r_4 = 38$	
Total	$c_1 = 94$	$c_2 = 96$	$c_3 = 119$	$n = 309$	

State hypothesis

In notation

$H_0 : p_{ij} = p_i p_j$ for $i = 1, 2, 3, 4,\ j = 1, 2, 3$

$H_0 : p_{ij} \neq p_i p_j$ for at least some i & j $(i = 1, 2, 3, 4,\ j = 1, 2, 3)$

In words

H_0: type of defect and shift are independent

H_α: type of defect and shift are dependent

Calculate expected count for each category

Defect type A

$$\hat{E_{11}} = \frac{r_1 c_1}{n} = \frac{74(94)}{309} = 22.5113$$

$$\hat{E_{12}} = \frac{r_1 c_2}{n} = \frac{74(96)}{309} = 22.9903$$

$$\hat{E_{13}} = \frac{r_1 c_3}{n} = \frac{74(119)}{309} = 28.4984$$

Defect type B

$$\hat{E_{21}} = \frac{r_2 c_1}{n} = \frac{69(94)}{309} = 20.9903$$

$$\hat{E_{22}} = \frac{r_2 c_2}{n} = \frac{69(96)}{309} = 21.4369$$

$$\hat{E_{23}} = \frac{r_2 c_3}{n} = \frac{69(119)}{309} = 26.5728$$

Defect type C

$$\hat{E_{31}} = \frac{r_3 c_1}{n} = \frac{128(94)}{309} = 38.9385$$

$$\hat{E_{32}} = \frac{r_3 c_2}{n} = \frac{128(96)}{309} = 39.7670$$

$$\hat{E_{33}} = \frac{r_3 c_3}{n} = \frac{128(119)}{309} = 49.2945$$

Defect type D

$$\hat{E_{41}} = \frac{r_4 c_1}{n} = \frac{38(94)}{309} = 11.5599$$

$$\hat{E_{42}} = \frac{r_4 c_2}{n} = \frac{38(96)}{309} = 11.8058$$

$$\hat{E_{43}} = \frac{r_4 c_3}{n} = \frac{38(119)}{309} = 14.6343$$

Calculate test statistic

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(15-22.5113)^2}{22.5113} + \ldots + \frac{(20-14.6343)^2}{14.6343} = 19.1780$$

Test rejection region

$$\chi^2{}_{0.05,(4-1)(3-1)} = \chi^2{}_{0.05,6} = 12.5916$$

$0.0039 < 0.05 \leftarrow$ using SAS

Since $19.1780 > 12.5916$, at $\alpha = 0.05$, we have enough evidence to reject H_0, accept H_α, and conclude that type of defect depends on shift.

The FREQ Procedure

Table of Type by Shift

percent out of 309
percent in own row
percent in own column

Type	Shift 1	2	3	Total
A	15	26	33	74
	4.85	8.41	10.68	23.95
	20.27	35.14	44.59	
	15.96	27.08	27.73	
B	21	31	17	69
	6.80	10.03	5.50	22.33
	30.43	44.93	24.64	
	22.34	32.29	14.29	
C	45	34	49	128
	14.56	11.00	15.86	41.42
	35.16	26.56	38.28	
	47.87	35.42	41.18	
D	13	5	20	38
	4.21	1.62	6.47	12.30
	34.21	13.16	52.63	
	13.83	5.21	16.81	
Total	94	96	119	309
	30.42	31.07	38.51	100.00

Statistics for Table of Type by Shift

calculated by hand

Statistic	DF	Value	Prob
Chi-Square	6	19.1780	0.0039
Likelihood Ratio Chi-Square	6	20.3359	0.0024
Mantel-Haenszel Chi-Square	1	0.5396	0.4626
Phi Coefficient		0.2491	
Contingency Coefficient		0.2417	
Cramer's V		0.1762	

Sample Size = 309

Exercise 2

A randomly selected group of 30-year-olds is interviewed to determine whether the type of music most listened to by people in their age category is independent of the geographic location of their residence. Use the Chi-Square test of independence at $\alpha = 0.01$, and the following contingency table to determine whether music reference is independent of geographic location.

		Type of Music Preferred				
		Rock	**R&B**	**Country**	**Classical**	
Geographic Region	**Northwest**	140	32	5	18	$r_1 = 195$
	South	134	41	52	8	$r_2 = 235$
	West	154	27	8	13	$r_3 = 202$
		$c_1 = 428$	$c_2 = 100$	$c_3 = 65$	$c_4 = 39$	$n = 632$

Sum rows

$r_1 = 140 + 32 + 5 + 18 = 195$

$r_2 = 134 + 41 + 52 + 8 = 235$

$r_3 = 154 + 27 + 8 + 13 = 202$

$n = r_1 + r_2 + r_3 = 195 + 235 + 202 = 632$

Sum columns

$c_1 = 140 + 134 + 154 = 428$

$c_2 = 32 + 41 + 27 = 100$

$c_3 = 5 + 52 + 8 = 65$

$c_4 = 18 + 8 + 13 = 39$

Calculate all expected values

Northwest

$$\hat{E_{11}} = \frac{r_1 c_1}{n} = \frac{195(428)}{632} = 132.0570 \qquad \hat{E_{12}} = \frac{r_1 c_2}{n} = \frac{195(100)}{632} = 30.8544$$

$$\hat{E_{13}} = \frac{r_1 c_3}{n} = \frac{195(65)}{632} = 20.0554 \qquad \hat{E_{14}} = \frac{r_1 c_4}{n} = \frac{195(39)}{632} = 12.0332$$

South

$$\hat{E_{21}} = \frac{r_2 c_1}{n} = \frac{235(428)}{632} = 159.1456 \qquad \hat{E_{22}} = \frac{r_2 c_2}{n} = \frac{235(100)}{632} = 37.1835$$

$$\hat{E_{23}} = \frac{r_2 c_3}{n} = \frac{235(65)}{632} = 24.1693 \qquad \hat{E_{24}} = \frac{r_2 c_4}{n} = \frac{235(39)}{632} = 14.5016$$

West

$$\hat{E_{31}} = \frac{r_3 c_1}{n} = \frac{202(428)}{632} = 136.7975 \qquad \hat{E_{32}} = \frac{r_3 c_2}{n} = \frac{202(100)}{632} = 31.9620$$

$$\hat{E_{33}} = \frac{r_3 c_3}{n} = \frac{202(65)}{632} = 20.7753 \qquad \hat{E_{34}} = \frac{r_3 c_4}{n} = \frac{202(39)}{632} = 12.4652$$

Calculate χ^2 statistic

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{4}\frac{(O_{ij}-\hat{E_{ij}})^2}{\hat{E_{ij}}} = \frac{(140-132.0570)^2}{132.0570} + \ldots + \frac{(13-12.4652)^2}{12.4652} = 64.9198$$

Test rejection region

$$\chi^2{}_{0.01,(3-1)(4-1)} = \chi^2{}_{0.01,6} = 16.8119$$

$$\text{p-value} < .0001 < 0.01 \leftarrow \text{using SAS}$$

Since $64.9198 > 16.8119$, at $\alpha = 0.01$, we have enough evidence to reject H_0, accept H_α, and conclude that music preference depends on region.

The FREQ Procedure

Table of Region by Type

Region	Type				
Frequency Percent Row Pct Col Pct	Classica	Country	R&B	Rock	Total
Northwes	18 2.85 9.23 46.15	5 0.79 2.56 7.69	32 5.06 16.41 32.00	140 22.15 71.79 32.71	195 30.85
South	8 1.27 3.40 20.51	52 8.23 22.13 80.00	41 6.49 17.45 41.00	134 21.20 57.02 31.31	235 37.18
West	13 2.06 6.44 33.33	8 1.27 3.96 12.31	27 4.27 13.37 27.00	154 24.37 76.24 35.98	202 31.96
Total	39 6.17	65 10.28	100 15.82	428 67.72	632 100.00

Statistics for Table of Region by Type

Statistic	DF	Value	Prob
Chi-Square	6	64.9198	<.0001
Likelihood Ratio Chi-Square	6	64.5748	<.0001
Mantel-Haenszel Chi-Square	1	0.9725	0.3241
Phi Coefficient		0.3205	
Contingency Coefficient		0.3052	
Cramer's V		0.2266	

Sample Size = 632

Chapter 14 - Part 3

The χ^2 Test of Homogeneity

This tests if the c populations have the same distribution with respect to the categorical variable.

Note: this test is conducted in the same way as the χ^2 test of independence, just with a different hypothesis and conclusion.

1. State hypothesis.
   - H_0 : the column distributions are homogenous
   - H_α : the column distributions are not homogenous
2. Calculate the expected count for each category, $\hat{E_{ij}}$.
   - $\hat{E_{ij}} = \frac{r_i c_j}{n}$
3. Compute test statistic, χ^2.
   - $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - \hat{E_{ij}})^2}{\hat{E_{ij}}}$
4. Test rejection region.
   - Reject H_0 if $\chi^2 > \chi^2{}_{\alpha,(r-1)(c-1)}$

<u>Exercise 1</u>

A researcher selected 100 passengers from each of three airlines and asked them if the airline had lost their luggage on their last flight. The data are summarized in the table below. At $\alpha = 0.05$, test the claim that the proportion of passengers from each airline who lost luggage on the flight is the same for each airline.

	Airline 1	**Airline 2**	**Airline 3**	
Yes	10	7	4	$r_1 = 10 + 7 + 4 = 21$
No	90	93	96	$r_2 = 90 + 93 + 96 = 279$

$c_1 = c_2 = c_3 = 100$ (not always equal!)

<u>State hypothesis</u>

H_0 : the column distributions are homogenous

H_α : the column distributions are not homogenous

<u>Calculate expected values</u>

$\hat{E_{11}} = \hat{E_{12}} = \hat{E_{13}} = \frac{21(100)}{300} = 7$ $\qquad$ $\hat{E_{21}} = \hat{E_{22}} = \hat{E_{23}} = \frac{279(100)}{300} = 93$

Note: some of these values are equal because the cs are equal (not always the case).

Compute test statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

$$= \frac{(10-7)^2}{7} + \frac{(7-7)^2}{7} + \frac{(4-7)^2}{7} + \frac{(90-93)^2}{93} + \frac{(93-93)^2}{93} + \frac{(96-93)^2}{93}$$

$$= 2.7650$$

Test rejection region

$$\chi^2_{0.05,(2-1)(3-1)} = \chi^2_{0.05,2} = 5.99147$$

$0.2510 > 0.05 \leftarrow$ using SAS

Since $2.7650 < 5.99147$, at $\alpha = 0.05$, we **don't** have enough evidence to reject H_0. Therefore, we **can't** conclude that the distribution of lost luggage is heterogenous.

```
                    The FREQ Procedure

                Table of Lost by Airline

     Lost       Airline

     Frequency|
     Percent  |
     Row Pct  |
     Col Pct  |A       |B       |C       |  Total
     ---------+--------+--------+--------+
     N        |     90 |     93 |     96 |    279
              |  30.00 |  31.00 |  32.00 |  93.00
              |  32.26 |  33.33 |  34.41 |
              |  90.00 |  93.00 |  96.00 |
     ---------+--------+--------+--------+
     Y        |     10 |      7 |      4 |     21
              |   3.33 |   2.33 |   1.33 |   7.00
              |  47.62 |  33.33 |  19.05 |
              |  10.00 |   7.00 |   4.00 |
     ---------+--------+--------+--------+
     Total         100      100      100      300
                 33.33    33.33    33.33   100.00


       Statistics for Table of Lost by Airline

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     2      2.7650    0.2510
Likelihood Ratio Chi-Square    2      2.8502    0.2405
Mantel-Haenszel Chi-Square     1      2.7558    0.0969
Phi Coefficient                       0.0960
Contingency Coefficient               0.0956
Cramer's V                            0.0960

                  Sample Size = 300
```

Note: **this is the end of new material!** The rest of the lectures will be for review.

Lecture 23 (Apr 2 2024)

[solutions to review problem on brightspace, not typed out]

Lecture 24 (Apr 4 2024)

[went over practice problems]

Lecture 25 (Apr 9 2024)

[finished up 2-WAY practice problem and started MLR practice problem]