Summary Statistics for Management and Economics lecture 1-13, tutorial work 1-13

Quantitative Research Methods (Aarhus Universitet)

Quantitative Research Methods

SPRING SEMESTER 2016

P-value< 0.05
Fobs> Fcrit

We REJECT H0

Table of Contents

When p-value is red, reject H0
F-test and X^2: use **α** to calculate crit (only t-test α/2)
CTRL+E to exclude an observation

P-value< 0.05
Fobs> Fcrit

We REJECT H0

P-value< 0.05
Fobs> Fcrit
We REJECT H0

When p-value is red, reject H0
F-test and X^2: use α to calculate crit (only t-test α/2)
CTRL+E to exclude an observation

Comparing two population means

↳ We use this when we want to find **the probability that the mean** of one sample is **greater** than the **mean** of another sample.

Assumptions:

- **Independent** random samples, drawn from 2 **normal** populations. If so, the **difference** between the 2 sample means will be *normally distributed.*

If the 2 populations are NOT both normally distributed, but the sample size are **>30, the difference between the sample means is* approximately *normal.*

To compute we need: 2 sample sizes, 2 means and 2 standard deviations.

- **Matched Pairs Experiment**

↳ **Comparing** 2 population means when an observation from one sample is **matched** with an observation from the second sample. (**Test if the means are equal)*

Objective: to compare 2 populations of **interval data**. (*Better than ANOVA because determines which* μ–mean *is greater)*

DECIDE IF THE DATA IS INDEPENDENT OR MATCHED.

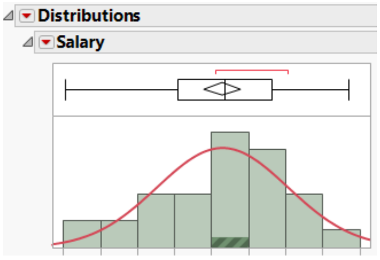
Is there natural *relationship* between *each pair* of observations that provides a logical reason to compare the first observation of sample 1 with the first observation of sample 2, and so on?

NOT Matched	Matched Pairs
Independent Samples (E.g. *comparing* if finance graduates have *higher* salaries than marketing graduates- we only look at the differences in their salaries)	**Matched Samples** (E.g. *comparing* if finance graduates have *higher* salaries than marketing graduates- we are comparing salaries of graduates with similar grades)
Hypotheses: $H_0: (\mu_1 - \mu_2) = 0$ $H_1: (\mu_1 - \mu_2) > 0$	***Hypotheses***: $H_0: (\mu_D) = 0$ $H_1: (\mu_D) > 0$, where $\mu_D = \mu_1 - \mu_2$
Equal-variances test statistic: $\boldsymbol{t} = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	**Test statistic for μ_D:** $\boldsymbol{t} = \dfrac{\overline{x_D} - \mu_D}{s_D/\sqrt{n_D}}$ with (n_D-1) degrees of freedom
Assumptions: **The differences are normally distributed**	*Assumptions*: **The differences are normally distributed**
In JMP: - Fit Fit Y by X (Y is the what we want to compare) **–** Δ Means/Anova/Pooled-t - Look at the **p-value** and conclude on the hypotheses (include the assumption).	*In JMP:* -Analyse-> Matched Pairs (Y is the 2 variables we are testing) - Look at the **p-value** and conclude on the hypothesis (include the assumption).

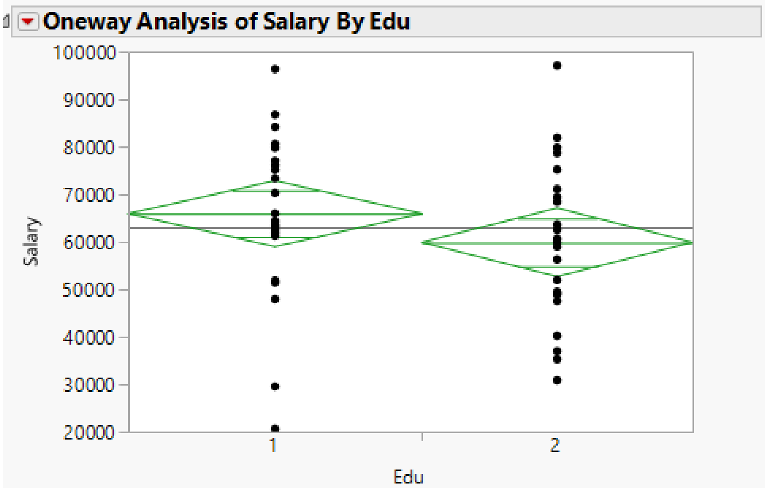
*For $\mathbf{t}_{crit}$ use $\frac{\sigma}{2}$ (double sided test).

NOT Matched:

1. **Assumption-** normally distributed Y (JMP- Distribution)



2. **Test in JMP**:



Oneway Anova

Summary of Fit

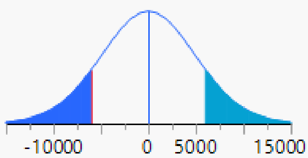
Rsquare	0.02977
Adj Rsquare	0.009557
Root Mean Square Error	17575.48
Mean of Response	63023.3
Observations (or Sum Wgts)	50

t Test

2-1

Assuming equal variances

Difference	-6038	t Ratio	-1.21358
Std Err Dif	4975	DF	48
Upper CL Dif	3965	Prob > |t|	0.2308
Lower CL Dif	-16041	Prob > t	0.8846
Confidence	0.95	Prob < t	0.1154



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Edu	1	454940115	454940115	1.4728	0.2308
Error	48	1.4827e+10	308897383		
C. Total	49	1.5282e+10			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	26	65921.4	3446.8	58991	72852
2	24	59883.7	3587.6	52670	67097

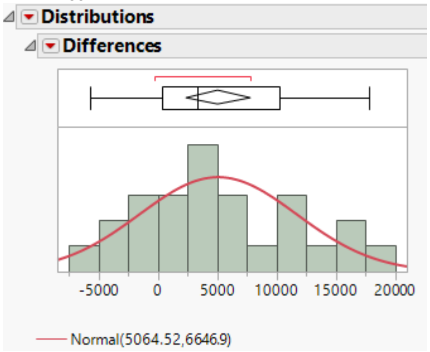
Std Error uses a pooled estimate of error variance

Conclusion:

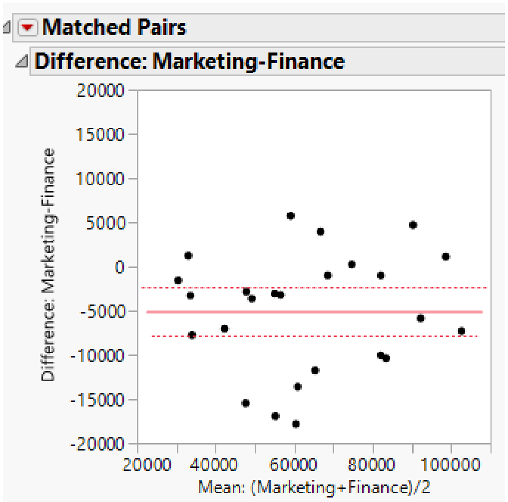
The **p-value** is 0.23 which is higher than our α level of 0.05, so we do not reject the null hypothesis. This means that at a **95% confidence level**, we do not have enough evidence to conclude that there is a difference between the salaries of finance and marketing graduates.

Matched Pairs:

1. **Assumption-** normally distributed DIFFERENCES *(*Make new column in JMP with the differences between the 2- edit Formula)*



2. **Test in JMP:**



Marketing	60373.7	t-Ratio	-3.80969
Finance	65438.2	DF	24
Mean Difference	-5064.5	Prob > |t|	0.0009*
Std Error	1329.38	Prob > t	0.9996
Upper 95%	-2320.8	Prob < t	0.0004*
Lower 95%	-7808.2		
N	25		
Correlation	0.95202		

Conclusion:

The **p-value** is 0.0009, which is very low, meaning we reject the null. So there is evidence that the finance graduates have higher salaries than marketing graduates. But taking into account the **normally distribution of differences assumption**, we cannot use this test because the data is very non-normal. So, our results are not reliable.

Comparing two variances (interval data)

↳ Comparing the variability. *E.g. judging the consistency of a production process, testing for quality.*

When comparing 2 populations, we look at **the ratio of variances**: $\frac{\sigma_1^2}{\sigma_2^2}$

- It can be used:
1. To test equality of variances (eg. Test if 2 portfolios have the same risk)
2. **First step** in deciding which *t-test for equality of means* to use.

F-distribution (with n-1 degrees of freedom): independent sampled data from 2 normal populations.

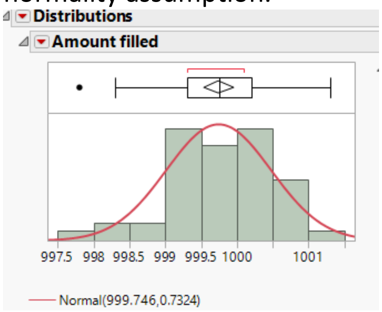
Comparing two variances
Assumptions: Independent sampled data from 2 normal populations. (*random data and normally distribution)*
Hypotheses: $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} < or > or \neq 1$
Test-statistic: $F = \frac{s_1^2}{s_2^2}$
Rejection region: $F > F_{\alpha,v_1,v_2}$, where α=0.05, $v_1 = n_1 - 1$ *(Calculate in F excel template)*
In JMP: **-** Create 2 columns for each variable (*e.g. machine)* - Tabulate: Vertical- the 2 variables (*machines*), Horizontal- Variance - Divide the 2 variances to find F_{obs} - Compare with F-distribution (*excel F),* look at p-value and conclude.
Interval estimator- Confidence interval: $\text{LCL} = \left(\frac{s_1^2}{s_2^2}\right) \cdot \frac{1}{F_{\alpha/2,v_1,v_2}}$; $\text{UCL} = \left(\frac{s_1^2}{s_2^2}\right) \cdot F_{\alpha/2,v_2,v_1}$ • *Required that the populations are normal.*

***If there is an empty cell, the observation must be removed.**

- **Assumptions:**

a) It is assumed that the data is randomly collected.

b) Normally distributed sample- The histograms appear to be sufficiently bell shaped to satisfy the normality assumption.



- In **JMP**:

	Variance
Machine 1	0.6333333333334
Machine 2	0.4527666667

$F = \frac{s_1^2}{s_2^2} = \frac{0.633}{0.452} = 1.40$

F_{crit}=$F_{\alpha,v_1,v_2} = F_{0.05,n_1-1,n_2-1}$ =1.98 (*From excel temp.*)

F_{obs}<F_{crit} => We do not have enough evidence to reject the null.

P-value=0.05 (*From excel temp.*)

- **Conclusions:**

Based on the **Test statistic F** conducted and **p-value**, the null is not rejected. So there is not enough evidence to conclude that the variance of machine 2 is less than the variance of machine 1. In other words, at a 5% significance level, there is no evidence that machine 2 is superior is its consistency.

- **Estimated interval:**

$$\mathrm{LCL} = \left(\frac{s_1^2}{s_2^2}\right) \cdot \frac{1}{F_{\alpha/2,v_1,v_2}} = 1.40 \cdot \frac{1}{2.27} = 0.61$$

$$\mathrm{UCL} = \left(\frac{s_1^2}{s_2^2}\right) \cdot F_{\alpha/2,v_2,v_1} = 1.40 \cdot 2.27 = 3.17$$

The 95% confidence interval estimate of the ratio of the two population variances is: (0.61;3.17).

**1 is in the interval.*

Chapter 14- ANOVA

ANOVA

↳compares **two or more** populations of interval data. It determines if *differences* exist between the *population means,* by analysing the **sample variance.**

One-way ANOVA

↳ **independently drawn samples with one factor** *(e.g. age)*.

One variable is nominal/ordinal (the factor/explanatory), the other is **continuous.*

F-distribution (with k-1 and n-k degrees of freedom).

One-way ANOVA
Assumptions: The random variable needs to be *normally distributed* with *equal variances*. *Independent* drawn samples. The errors are normally distributed. (*normally distribution - as many graphs as terms in H_0-, equal variances, independence- random sampling, normal distribution residuals,* trustworthiness & validity) *no. of terms in H_o=no. categories in factor level
Model: $y = \mu + \alpha_i + \varepsilon_i$ E.g.: $Cost = \mu + Bumper + \varepsilon$
Hypotheses: $H_0: \mu_1 = \mu_2 (= \mu_3 = \mu_4)$ $H_1: At\ least\ two\ means\ differ$
Significance level: α= 0.05

Test-statistic:

$$F = \frac{MST}{MSE} \sim F_{\alpha,k-1,n-k}$$

**Usually a small SST (and F) supports H₀.*

Rejection region:

$F > F_{\alpha,k-1,n-k}$, where α=0.05, $\nu_1 = k-1, \nu_2 = n-k$, where *k* is the number of factor levels and *n* is the number of observations *(Calculate in F excel template)*

In JMP:

- Check for normal distribution (*we get as many graphs as terms in H_0*)- **Distribution**: Y- dependent variable, *By*- independent variable

Y- CONTINIOUS, X- NOMINAL

- Check for equal variances: **Tabulate**: Vertical- the factor levels, Horizontal- Variance.
- **Hartley's test**: Take the biggest variance and divide it with the smallest variance to get F_{obs}. Compare it to F_{crit}, calculated in *excel template,* with $F(\frac{\alpha}{k(k-1)}; n_{max} - 1; n_{min} - 1)$

↳**Hypotheses:**

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

$H_1: At\ least\ two\ variances\ differ$

Test statistic:

$$F_{obs} = \frac{s_{max}^2}{s_{min}^2} \sim F_{(\frac{\alpha}{k(k-1)}; n_{max}-1; n_{min}-1)}$$

- **Fit Y by X**: Y- what we are analysing, X- factor levels, **△Means/ANOVA**
- **△Unequal Variances**, look at Brown-Forsythe and Levene tests. If the null is **not** rejected, we have equal variances across groups (*good)*.
- **△Save-> Save Residuals,** *normal distribution for the residuals:* **Distribution**: Y-residuals.
- Compare with F-distribution (*excel F),* look at p-value and conclude.
- Comment on R^2: how good is the model (*% the Xs explain the Y).*

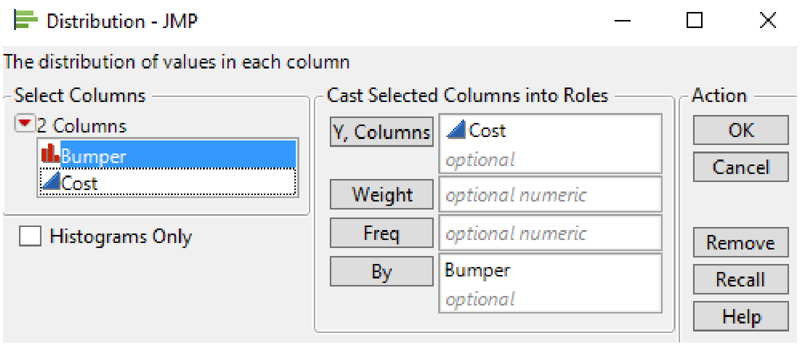
**It's NOT necessary that the sample sizes are equal ($n_1=n_2=...=n_k$)*

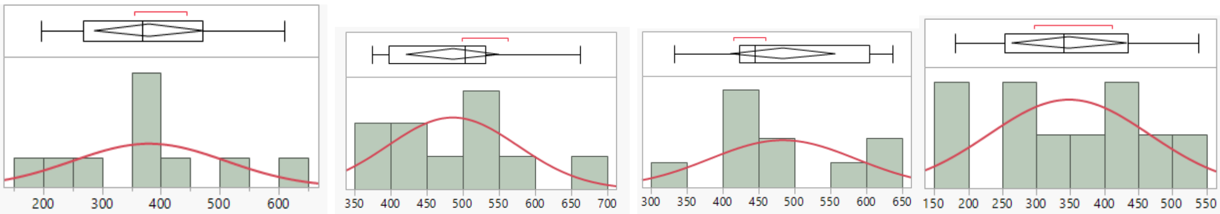
Y needs to be continuous.

Source of Variation	degrees of freedom	Sum of Squares	Mean Square
Treatments	*k*–1	SST	MST=SST/(*k*–1)
Error	*n*–*k*	SSE	MSE=SSE/(*n*–*k*)
Total	*n*–1	SS(Total)	

- **Assumptions:**

a) Gaussianity in each group (normally distributed samples).





The normallity assumption is not truly met, but it may still be reasonable to work with, so the analysis is continued.

b) Equal variances. (*Look also the Unequal Variances test- below)*

	Cost	
Bumper	Variance	N
1	16924.222222222	10
2	8197.4333333333	10
3	10426.177777778	10
4	14048.622222222	10

It can be said that the variances are roughly equal, since there are no big differences between each sample.

Homogeneity- Hartley's test

Hypotheses:

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

$H_1: At\ least\ two\ variances\ differ$

Test statistic:

$$F_{obs} = \frac{s_{max}^2}{s_{min}^2} = \frac{16924}{8197} = 2.06$$

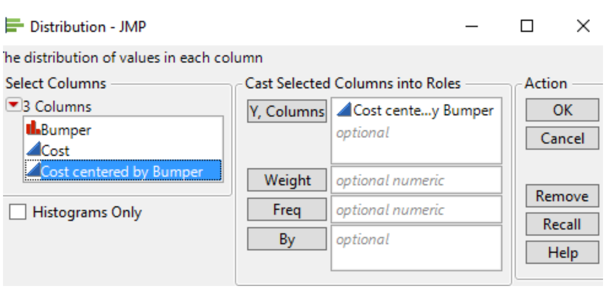
$$F\left(\frac{\alpha}{k(k-1)}; n_{max} - 1; n_{min} - 1\right) = F_{\frac{0.05}{4(4-1)};10-1;10-1} = 6.88$$

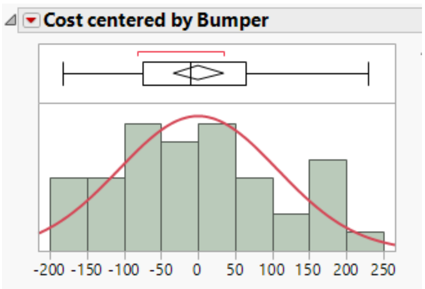
If $F_{obs}<F_{crit}$, the null is not rejected, so we have homogeneity across variances.

c) Independent drawn samples.

In this case the sample is random, so the independency assumption is fulfilled.

d) The errors are normally distributed.





The errors follow the bell shape, so the errors are normally distributed.

- In **JMP:**

Oneway Anova

Summary of Fit

Rsquare	0.25263
Adj Rsquare	0.190349
Root Mean Square Error	111.3513
Mean of Response	424.475
Observations (or Sum Wgts)	40

Analysis of Variance

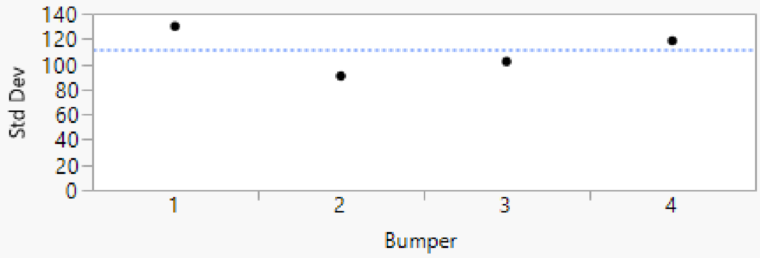
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Bumper	3	150883.88	50294.6	4.0563	0.0139*
Error	36	446368.10	12399.1		
C. Total	39	597251.98			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	10	380.000	35.212	308.59	451.41
2	10	485.900	35.212	414.49	557.31
3	10	483.800	35.212	412.39	555.21
4	10	348.200	35.212	276.79	419.61

Std Error uses a pooled estimate of error variance

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	10	130.0931	96.20000	96.00000
2	10	90.5397	70.52000	67.90000
3	10	102.1087	83.16000	78.40000
4	10	118.5269	95.20000	95.20000

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0.6202	3	36	0.6065
Brown-Forsythe	0.3815	3	36	0.7669
Levene	0.3626	3	36	0.7803
Bartlett	0.4332	3	.	0.7292

↳H_0 says that the variances are equal across groups. In this case we do not reject the null, so we assume equal variances, according to Brown-Forsythe and Levene tests.

- **Conclusions:**

F_{crit}=$F_{\alpha,k-1,n-k} = F_{0.05,3,36}$ =2.8662 (*From excel temp.*)
F_{obs}=4.05
F_{obs}>F_{crit} => We reject the null.
P-value=0.0139 (*ANOVA table in JMP)*

We have enough evidence to reject the null hypothesis, according to the **Test statistic F** conducted and **p-value**. This means that **at least 2 means differ in our factor levels**, so there is evidence that at least two bumpers differ.

• Fisher's least significant difference (LSD) method
• Bonferroni adjustment
• Tukey's multiple comparison method

↳ Determine **which** population means **differ**.

• Fisher's least significant difference (LSD) method (Type I error)

↳ compares the difference between means with LSD:

$(\overline{x_1} - \overline{x_2}) \sim LSD = t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, where $v = n - k$

All k sample sizes must be equal. If some sample sizes differ, LSD must be calculated for each combination.*

In **JMP:**

➔ One-way ANOVA: test assumption, test-statistic, conclude on F and p-value (see above).

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Bumper	3	150883.88	50294.6	4.0563	0.0139*
Error	36	446368.10	12399.1		
C. Total	39	597251.98			

↳ *At least two means differ.*

➔ **LSD method** to test which means differ:

– **∆Compare Means → Each Pair, Student's t**

– Look at <u>positive</u> numbers in the *Threshold Matrix* and <u>small p-values</u> (<0.05) in *Ordered Differences Report*. This shows the different means pairs.

Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

LSD Threshold Matrix

Abs(Dif)-LSD

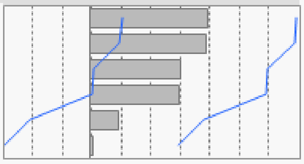
	2	3	1	4
2	-100.99	-98.89	4.91	36.71
3	-98.89	-100.99	2.81	34.61
1	4.91	2.81	-100.99	-69.19
4	36.71	34.61	-69.19	-100.99

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
2	4	137.7000	49.79782	36.7053	238.6947	0.0089*
3	4	135.6000	49.79782	34.6053	236.5947	0.0099*
2	1	105.9000	49.79782	4.9053	206.8947	0.0404*
3	1	103.8000	49.79782	2.8053	204.7947	0.0443*
1	4	31.8000	49.79782	-69.1947	132.7947	0.5271
2	3	2.1000	49.79782	-98.8947	103.0947	0.9666



Differ: μ_1 and μ_2, μ_1 and μ_3, μ_2 and μ_4, μ_3 and μ_4
Do not differ: μ_1 and μ_4, μ_2 and μ_3

• Bonferroni adjustment (Type II error)

↳ Better than Fisher's LSD, *reliable when we look at two or three pairs to compare*.

Same as LSD method, but we adjust the α-level:

$\alpha^* = \frac{\alpha}{k(k-1)}$, where k is the number of groups and α=0.05 (usually).

<u>In **JMP**:</u>

➔ One-way ANOVA: test assumption, test-statistic, conclude on F and p-value (see above).
↳ *At least two means differ.*

➔ **LSD method with Bonferroni adjustment** to test which means differ:

- **Calculate α^***
  $\alpha^* = \frac{\alpha \cdot 2}{k(k-1)} = \frac{0.05 \cdot 2}{4(4-1)} = 0.008$
- **∆Compare Means → Each Pair, Student's t**
- **∆Set alpha level** → put α^*
- Look at <u>positive</u> numbers in the *Threshold Matrix*. This shows the different means pairs.

Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

LSD Threshold Matrix

Abs(Dif)-LSD

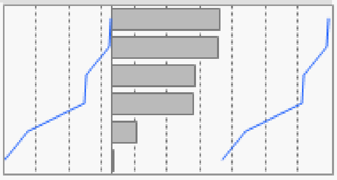
	2	3	1	4
2	-139.11	-137.01	-33.21	-1.41
3	-137.01	-139.11	-35.31	-3.51
1	-33.21	-35.31	-139.11	-107.31
4	-1.41	-3.51	-107.31	-139.11

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
2	4	137.7000	49.79782	-1.413	276.8130	0.0089*
3	4	135.6000	49.79782	-3.513	274.7130	0.0099*
2	1	105.9000	49.79782	-33.213	245.0130	0.0404*
3	1	103.8000	49.79782	-35.313	242.9130	0.0443*
1	4	31.8000	49.79782	-107.313	170.9130	0.5271
2	3	2.1000	49.79782	-137.013	141.2130	0.9666



No means differ.

<u>•Tukey's multiple comparison method</u>

↳ Better than Fisher's LSD and Bonferroni *if you look at all possible combinations*.

<u>*Test:*</u>

$\omega = q_\alpha(k,v)\sqrt{\frac{MSE}{n_g}}$, where $q = \frac{\overline{x_{max}} - \overline{x_{min}}}{s/\sqrt{n}}$

<u>In **JMP**:</u>

➔ One-way ANOVA: test assumption, test-statistic, conclude on F and p-value (see above).
↳ *At least two means differ.*

➔ <u>Tukey method</u>

- **∆Compare Means → All Pairs, Tukey's HSD**
- **∆Set alpha level** → put the initial α (=**0.05**)
- Look at <u>positive</u> numbers in the *Threshold Matrix* and <u>small p-values</u> (<0.05) in *Ordered Differences Report*. *(Also look at Lower and Upper CL)* This shows the different means pairs.

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

HSD Threshold Matrix

Abs(Dif)-HSD

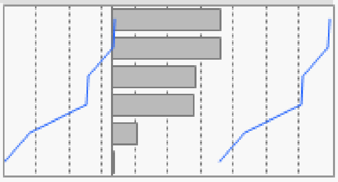
	2	3	1	4
2	-134.12	-132.02	-28.22	3.58
3	-132.02	-134.12	-30.32	1.48
1	-28.22	-30.32	-134.12	-102.32
4	3.58	1.48	-102.32	-134.12

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
2	4	137.7000	49.79782	3.583	271.8169	0.0423*
3	4	135.6000	49.79782	1.483	269.7169	0.0467*
2	1	105.9000	49.79782	-28.217	240.0169	0.1641
3	1	103.8000	49.79782	-30.317	237.9169	0.1775
1	4	31.8000	49.79782	-102.317	165.9169	0.9188
2	3	2.1000	49.79782	-132.017	136.2169	1.0000



Differ: μ_2 *and* μ_4, μ_3 *and* μ_4

Randomized Blocks

↳ Compares ***more*** *than two population means and we have a* ***matched group*** *of observations*. (*Matched pairs compares only 2).* Only **one factor.** *(e.g. age) Interval data.*

F-distribution (with k-1 and n-k-b+1 degrees of freedom).

Randomized Blocks
Assumptions: The random variable needs to be *normally distributed* with *equal variances*. *Independent* drawn samples. Normally distribute errors. (*normally distribution - as many graphs as terms in H_0-, equal variances, independence- random sampling, normal distribution residuals,* trustworthiness & validity*)* *no. of terms in H_o=no. categories in factor level
Model: $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ E.g. Reduction= $\mu + Drug + Group + \varepsilon$
Hypotheses: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ $H_1: At\ least\ two\ means\ differ$
Significance level: α= 0.05
Test-statistic: $F = \frac{MST}{MSE} \sim F_{\alpha,k-1,n-k-b+1}$
Rejection region: $F > F_{\alpha,k-1,n-k-b+1}$, where α=0.05, $\nu_1 = k-1, \nu_2 = n-k-b+1$, where *b* is the number of blocks, *k* is the number of factor levels and *n* is the number of observations *(Calculate in F excel template)*
In JMP: - Check for normal distribution (*as many graphs as terms in H_0)*- **Distribution**: Y- the dependent variable, *By*- independent variable.

- Check for equal variances: **Tabulate**: Vertical- the factor levels, Horizontal- Variance
- **Hartley's test**: Take the biggest variance and divide it with the smallest variance to get F_{obs}. Compare it to F_{crit}, calculated in *excel template,* with $F(\frac{\alpha}{k(k-1)}; n_i - 1; n_j - 1)$

↳**Hypotheses:**

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$$
$$H_1: At\ least\ two\ variances\ differ$$

Test statistic:

$$F_{obs} = \frac{s_{max}^2}{s_{min}^2}$$

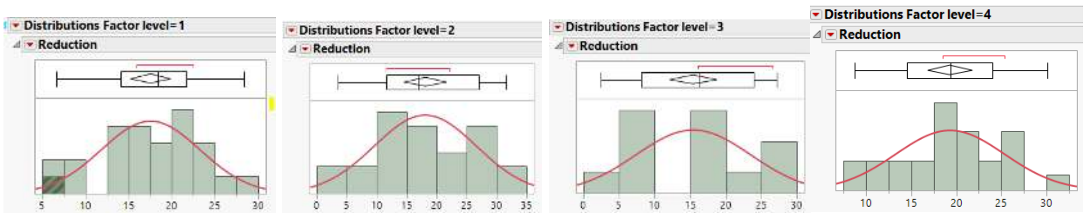
- **Fit Y by X**: Y- what we are analysing, X- factor levels, *Block*- matching groups, **∆Means/ANOVA**
- **∆Unequal Variances**, look at Brown-Forsythe and Levene tests. If the null is **not** rejected, we have equal variances across groups (*good)*.
- **∆Save-> Save Residuals,** *normal distribution for the residuals:* **Distribution**: Y-residuals.
- Compare with F-distribution (*excel F),* look at p-value and conclude.
- Comment on **Adjusted R^2** (or R^2 if only one X): how good is the model (*% the Xs explain the Y).*
- Look at the **Mean** to see which factor level is the best (*highest means*).

If there is one empty cell, the row must be removed.
*****Y needs to be continuous, X is nominal/ordinal.***

- **Assumptions:**

a) Gaussianity in each group (normally distributed samples).

Source of Variation	d.f.:	Sum of Squares	Mean Square	***F*** Statistic
Treatments	k–1	SST	MST=SST/(k–1)	F=MST/MSE
Blocks	b–1	SSB	MSB=SSB/(b-1)	F=MSB/MSE
Error	n–k–b+1	SSE	MSE=SSE/(n–k–b+1)	
Total	n–1	SS(Total)		



The normallity assumption is not truly met, so it means that we may consider collecting more data, because the results may not be entirely valid and reliable. But for now, it may still be reasonable to work with, so the analysis is continued.

b) Equal variances. (**Look also the Unequal Variances test- below)*

	Reduction	
Factor level	Variance	N
1	32.696766666667	25
2	73.244566666667	25
3	65.716766666667	25
4	36.309166666667	25

There may be a problem with this assumption, since the variances vary across the different drugs.

Homogeneity- Hartley's test

Hypotheses:

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

$H_1: At\ least\ two\ variances\ differ$

Test statistic:

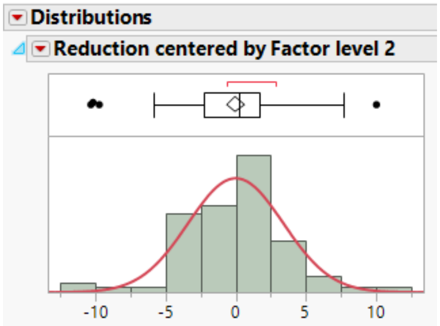
$$F_{obs} = \frac{s_{max}^2}{s_{min}^2} = \frac{73.24}{32.69} = 2.2$$

$$F\left(\frac{\alpha}{k(k-1)}; n_{max} - 1; n_{min} - 1\right) = F_{\frac{0.05}{4(4-1)};25-1;25-1} = 3.04$$

If $F_{obs} > F_{crit}$, so the null hypothesis is rejected. The homogeneity across variances assumption is not fulfilled.

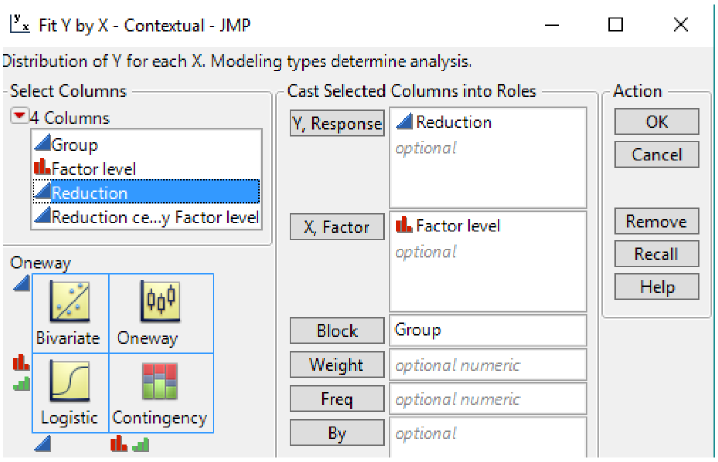
c) Independent drawn samples.
It is assumed that the observations are random and independent.

d) The errors are normally distributed.



The assumption is met.

- In **JMP:**



Oneway Anova

Summary of Fit

Rsquare	0.779734
Adj Rsquare	0.697134
Root Mean Square Error	3.983574
Mean of Response	17.603
Observations (or Sum Wgts)	100

Analysis of Variance

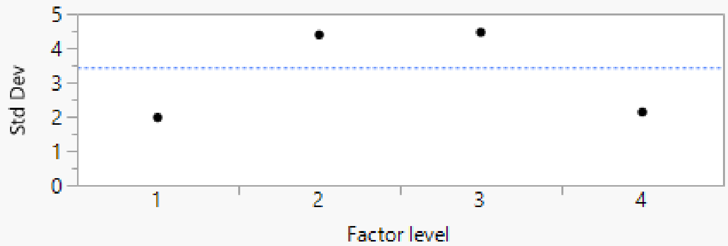
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Factor level	3	195.9547	65.318	4.1161	0.0094*
Group	24	3848.6566	160.361	10.1054	<.0001*
Error	72	1142.5578	15.869		
C. Total	99	5187.1691			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	25	17.5480	0.79671	15.960	19.136
2	25	18.0960	0.79671	16.508	19.684
3	25	15.4480	0.79671	13.860	17.036
4	25	19.3200	0.79671	17.732	20.908

Std Error uses a pooled estimate of error variance

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	25	1.974987	1.513600	1.496000
2	25	4.388826	3.508720	3.504000
3	25	4.459768	3.338000	3.286000
4	25	2.134168	1.640320	1.636000

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	4.2427	3	96	0.0073*
Brown-Forsythe	5.9302	3	96	0.0009*
Levene	6.3598	3	96	0.0006*
Bartlett	8.5013	3	.	<.0001*

↳H_0 says that the variances are equal across groups. In this case we reject the null, so we have unequal variances, according to Brown-Forsythe and Levene tests.

- **<u>Conclusions:</u>**

F_{crit}=$F_{\alpha,k-1,n-k-b+1}$= 2.73 (*From excel temp.)*
F_{obs}>F_{crit} => We reject the null.
P-value=0.0094 (*ANOVA table in JMP)*

We have enough evidence to reject the null hypothesis, according to the **Test statistic F** conducted and **p-value**. This means that **at least 2 means differ in our factor levels**, so there is evidence that at least two of the drugs differ. Looking at the means, drug 2 and drug 4 reveals the biggest reduction in cholesterol, but further testing is recommended to determine which is better.

Two-way ANOVA

↳ Compares *more than two population means and* **two factors** *(or more- e.g. age & gender). Interval data.*

F-distribution

Two-way ANOVA
Assumptions: The random variable needs to be *normally distributed* with *equal variances*. *Independent* drawn samples. Normally distributed errors. (*normally distribution - as many graphs as terms in H_0-, equal variances, independence- random sampling, Residual by Predicted Plot,* trustworthiness & validity*)* *no. of terms in H_o=no. categories in factor1*no. categories in factor 2
Model: $y_{ijh} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijh}$ E.g. No. jobs= $\mu + Education + Gender + Edu * Gender + \varepsilon$
Hypotheses: $H_0: \alpha = 0$ $H_1: \alpha \neq 0$ $H_0: \beta = 0$ $H_1: \beta \neq 0$ $H_0: \gamma = 0$ $H_1: \gamma \neq 0$
Significance level: α= 0.05
Test-statistic: $F = \frac{MS(\alpha)}{MSE} \sim F_{a-1;n-ab}$; $F = \frac{MS(\beta)}{MSE} \sim F_{b-1;n-ab}$; $F = \frac{MS(\delta)}{MSE} \sim F_{(a-1)(b-1);n-ab}$
In JMP: - Check for normal distribution - **Distribution**: Y- dependent variable, *By*- all independent variables - Check for equal variances: **Tabulate**: Vertical- the factor levels, Horizontal- Variance - **Hartley's test**: Take the biggest variance and divide it with the smallest variance to get F_{obs}. Compare it to F_{crit}, calculated in *excel template,* with $F(\frac{\alpha}{k(k-1)}; n_i - 1; n_j - 1)$ ↳**Hypotheses:** $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2 = \sigma_7^2 = \sigma_8^2$ $H_1: At\ least\ two\ variances\ differ$ **Test statistic:** $F_{obs} = \frac{s_{max}^2}{s_{min}^2}$ - **Fit Model**: Y- dependent variable, *Model effects (box down)*- independent variables and interaction. ∆**Save columns-> Save Residuals, Distribution** for the residuals. **-** Check for normal distribution among residuals –ALSO at the bottom of the Fit model window, copy **Residual by Predicted Plot** (**You can resize it by dragging to the sides)* - ∆**Estimates-> Show prediction expression:** copy the graph from the bottom and comment. **-** ∆**Factor Profiling-> Profiler:** copy the graphs from the bottom and comment on them. - Check if the independent variable are significant (p-value<0.05). If not, remove them *one at a time*: **Model Dialog** (*Start with the interaction!- DO NOT REMOVE A TERM FROM THE INTERACTION). - Again ∆**Estimates-> Show prediction expression,** if the model has changed.

- Compute $\boldsymbol{\alpha}^* = \frac{\alpha}{k(k-1)}$, for calculating Fcrit. (*k is the no. of factors in first X multiplied by the no. of factor in second X. Eg. Gender and Edu k=2*4=8)*
- Compare with F-distribution (*excel F),* look at p-value and conclude.
- Comment on **Adjusted R^2** (or R^2 if only one X): how good is the model (*% the Xs explain the Y).*
- Compare the **Mean** with the *Bonferroni adjustment* to see which groups differ from each other. Go to the right side of the model: **ΔLS Means Student's t** and *Shift* (REMEMBER TO PRESS *Shift* to change α to the calculated α*). There is a difference where is red. *Can also do Tukey.*

↳**ΔOrdered differences report:** look where there is no **0** in the Confidence Interval.

- **ΔLS Means Plot:** to look at the variation in the means.

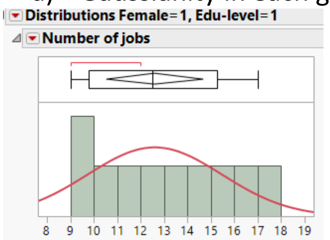
If there is one empty cell, the row must be removed.

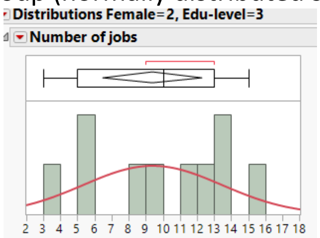
****Y needs to be continuous and Xs nominal/ordinal.***

Source of Variation	d.f.:	Sum of Squares	Mean Square	*F* Statistic
Factor A	a-1	SS(A)	MS(A)=SS(A)/(a-1)	F=MS(A)/MSE
Factor B	b–1	SS(B)	MS(B)=SS(B)/(b-1)	F=MS(B)/MSE
Interaction	(a-1)(b-1)	SS(AB)	MS(AB) = $\frac{SS(AB)}{[(a-1)(b-1)]}$	F=MS(AB)/MSE
Error	n–ab	SSE	MSE=SSE/(n–ab)	
Total	n–1	SS(Total)		

- **<u>Assumptions:</u>**

a) Gaussianity in each group (normally distributed samples).





The normallity assumption is not truly met, so it means that we may consider collecting more data, because the results may not be entirely valid and reliable. But for now, it may still be reasonable to work with, so the analysis is continued.

b) Homogeneity- Equal variances. (**Look also the Unequal Variances test- below)*

		Number of jobs	
Education	**Female**	**Variance**	**N**
Less than high school	1	8.2666666667	10
	2	8.2777777778	10
High school	1	8.6666666667	10
	2	9.7333333333	10
College	1	11.6	10
	2	16.488888889	10
University degree	1	5.3333333333	10
	2	12.322222222	10

There may be a problem with this assumption, since the variances vary, mostly across the females different levels of education.

Homogeneity- Hartley's test

Hypotheses:

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2 = \sigma_7^2 = \sigma_8^2$

$H_1: At\ least\ two\ variances\ differ$

Test statistic:

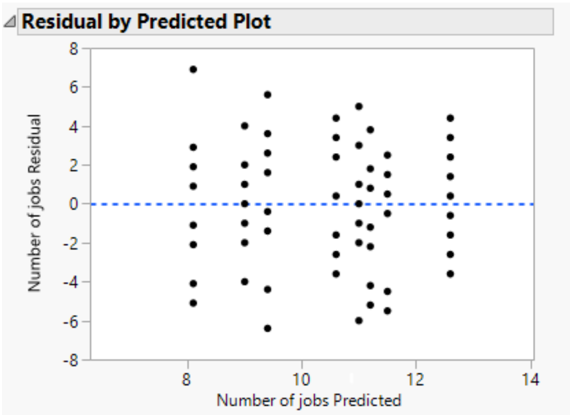
$$F_{obs} = \frac{s_{max}^2}{s_{min}^2} = \frac{73.24}{32.69} = 2.24$$

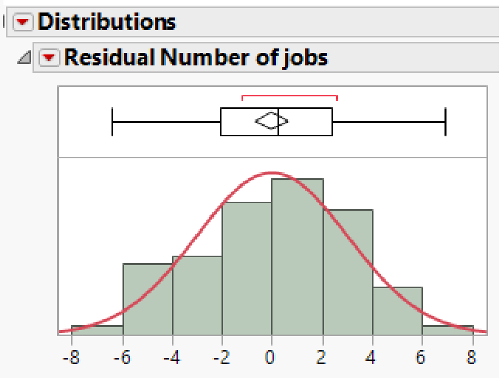
$$F\left(\frac{\alpha}{k(k-1)}; n_{max} - 1; n_{min} - 1\right) = F_{\frac{0.05}{8(8-1)};100-1;100-1} = 1.89$$

If $F_{obs} > F_{crit}$, so there the null hypothesis is rejected, so we have a problem with homogeneity across variances.

c) Independent drawn samples.
It is assumed that the observations are random and independent.

d) The errors are normally distributed.





The assumption seems to be met.

- In **JMP:**

Summary of Fit

RSquare	0.174351
RSquare Adj	0.094079
Root Mean Square Error	3.175864
Mean of Response	10.425
Observations (or Sum Wgts)	80

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	153.35000	21.9071	2.1720
Error	72	726.20000	10.0861	Prob > F
C. Total	79	879.55000		0.0467*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.425	0.355072	29.36	<.0001*
Female[1]	0.375	0.355072	1.06	0.2944
Edu-level[1]	1.625	0.615003	2.64	0.0101*
Edu-level[2]	0.675	0.615003	1.10	0.2761
Edu-level[3]	-0.425	0.615003	-0.69	0.4918
Female[1]*Edu-level[1]	0.175	0.615003	0.28	0.7768
Female[1]*Edu-level[2]	-0.475	0.615003	-0.77	0.4424
Female[1]*Edu-level[3]	0.225	0.615003	0.37	0.7155

Effect Tests

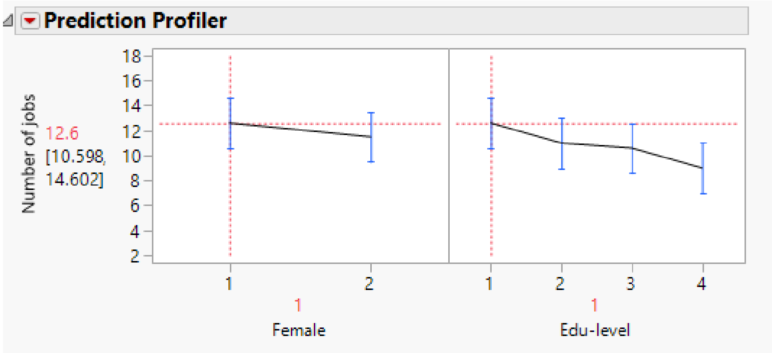
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Female	1	1	11.25000	1.1154	0.2944
Edu-level	3	3	135.85000	4.4897	0.0060*
Female*Edu-level	3	3	6.25000	0.2066	0.8915

The model is significant, since we have a p-value=0.0467 lower than 0.05.

Prediction Expression

$$
\begin{aligned}
&10.425\\
&+\text{Match}\left(\text{Female}\right)\begin{pmatrix} 1 \Rightarrow 0.375 \\ 2 \Rightarrow -0.375 \\ \text{else} \Rightarrow . \end{pmatrix}\\
&+\text{Match}\left(\text{Edu-level}\right)\begin{pmatrix} 1 \Rightarrow 1.625 \\ 2 \Rightarrow 0.675 \\ 3 \Rightarrow -0.425 \\ 4 \Rightarrow -1.875 \\ \text{else} \Rightarrow . \end{pmatrix}\\
&+\text{Match}\left(\text{Female}\right)\begin{pmatrix} 1 \Rightarrow \text{Match}\left(\text{Edu-level}\right)\begin{pmatrix} 1 \Rightarrow 0.175 \\ 2 \Rightarrow -0.475 \\ 3 \Rightarrow 0.225 \\ 4 \Rightarrow 0.075 \\ \text{else} \Rightarrow . \end{pmatrix} \\ 2 \Rightarrow \text{Match}\left(\text{Edu-level}\right)\begin{pmatrix} 1 \Rightarrow -0.175 \\ 2 \Rightarrow 0.475 \\ 3 \Rightarrow -0.225 \\ 4 \Rightarrow -0.075 \\ \text{else} \Rightarrow . \end{pmatrix} \\ \text{else} \Rightarrow . \end{pmatrix}
\end{aligned}
$$

The baseline is 10.425 number of jobs. If male, the number of jobs increases with 0.375 and for female, it decreases with -0.375. For the education level, the number of jobs increases with 1.625 for people that do not have high school, increases with 0.675 for people with high school and afterwards starts decreasing. And so on.



Looking at the Prediction Profiler, it can be concluded that the number of jobs man have is higher than for female. Also, as the education level increases, the number of jobs decreases.

➔ Looking at the Effect tests, we can see that the interaction Edu*Female is not significant: the p-value 0.89>0.05. So we remove it from the model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	147.10000	36.7750	3.7656
Error	75	732.45000	9.7660	Prob > F
C. Total	79	879.55000		0.0076*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	3	6.25000	2.0833	0.2066
Pure Error	72	726.20000	10.0861	Prob > F
Total Error	75	732.45000		0.8915
				Max RSq
				0.1744

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.425	0.349392	29.84	<.0001*
Female[1]	0.375	0.349392	1.07	0.2866
Edu-level[1]	1.625	0.605165	2.69	0.0089*
Edu-level[2]	0.675	0.605165	1.12	0.2682
Edu-level[3]	-0.425	0.605165	-0.70	0.4847

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Female	1	1	11.25000	1.1520	0.2866
Edu-level	3	3	135.85000	4.6368	0.0050*

➔ Now, the Female variable is still not significant, with a p-value of 0.28>0.05. So we remove it.

Summary of Fit

RSquare	0.154454
RSquare Adj	0.121077
Root Mean Square Error	3.128183
Mean of Response	10.425
Observations (or Sum Wgts)	80

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	135.85000	45.2833	4.6276
Error	76	743.70000	9.7855	Prob > F
C. Total	79	879.55000		0.0050*

Parameter Estimates

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Edu-level	3	3	135.85000	4.6276	0.0050*

➔ The new model is a One-way ANOVA and is significant, with a p-value of 0.005<0.05.

Prediction Expression

$$10.425 + \text{Match}\left(\text{Female}\right)\begin{bmatrix} 1 & \Rightarrow 0.375 \\ 2 & \Rightarrow -0.375 \\ \text{else} \Rightarrow . & \end{bmatrix} + \text{Match}\left(\text{Edu-level}\right)\begin{bmatrix} 1 & \Rightarrow 1.625 \\ 2 & \Rightarrow 0.675 \\ 3 & \Rightarrow -0.425 \\ 4 & \Rightarrow -1.875 \\ \text{else} \Rightarrow . & \end{bmatrix}$$

Same interpretation as before: The baseline is 10.425 number of jobs. If male, the number of jobs increases with 0.375 and for female, it decreases with -0.375. For the education level, the number of jobs increases with 1.625 for people that do not have high school, increases with 0.675 for people with high school and afterwards starts decreasing.

$$\alpha^* = \frac{\alpha}{k(k-1)} = \frac{0.05}{8(8-1)} = 0.001$$

- **Conclusions:**

$F_{crit}=F_{0.05;3;76} = 2.72$ *(From excel temp.)*
F_{obs}=4.62
$F_{obs}>F_{crit}$ => We reject the null.
P-value Education=0.005 (*ANOVA table in JMP)*

We have enough evidence to reject the null hypothesis, according to the **Test statistic F** conducted and **p-value**. This means that **at least 2 means differ in our factor levels**, so there is evidence that at least two of the drugs differ. Looking at the means, drug 2 and drug 4 reveals the biggest reduction in cholesterol, but further testing is recommended to determine which is better.

Bonferroni adjustment- test which means differ

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
1	12.050000	0.69948289	12.0500
2	11.100000	0.69948289	11.1000
3	10.000000	0.69948289	10.0000
4	8.550000	0.69948289	8.5500

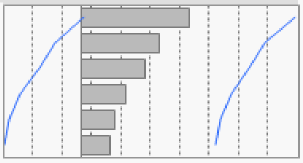
LSMeans Differences Student's t

α= 0.001 t= 3.4232

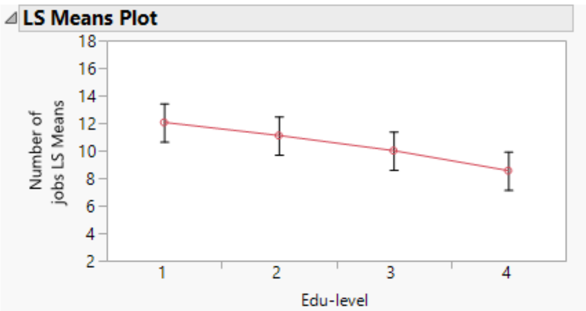
LSMean[i]	Mean[i]-Mean[j] / Std Err Dif / Lower CL Dif / Upper CL Dif	LSMean[j] 1	2	3	4
1	Mean[i]-Mean[j]	0	0.95	2.05	3.5
	Std Err Dif	0	0.98922	0.98922	0.98922
	Lower CL Dif	0	-2.4363	-1.3363	0.11371
	Upper CL Dif	0	4.33629	5.43629	6.88629
2	Mean[i]-Mean[j]	-0.95	0	1.1	2.55
	Std Err Dif	0.98922	0	0.98922	0.98922
	Lower CL Dif	-4.3363	0	-2.2863	-0.8363
	Upper CL Dif	2.43629	0	4.48629	5.93629
3	Mean[i]-Mean[j]	-2.05	-1.1	0	1.45
	Std Err Dif	0.98922	0.98922	0	0.98922
	Lower CL Dif	-5.4363	-4.4863	0	-1.9363
	Upper CL Dif	1.33629	2.28629	0	4.83629
4	Mean[i]-Mean[j]	-3.5	-2.55	-1.45	0
	Std Err Dif	0.98922	0.98922	0.98922	0
	Lower CL Dif	-6.8863	-5.9363	-4.8363	0
	Upper CL Dif	-0.1137	0.83629	1.93629	0

Level		Least Sq Mean
1	A	12.050000
2	A B	11.100000
3	A B	10.000000
4	B	8.550000

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
1	4	3.500000	0.9892182	0.11371	6.886289	0.0007*
2	4	2.550000	0.9892182	-0.83629	5.936289	0.0119*
1	3	2.050000	0.9892182	-1.33629	5.436289	0.0416*
3	4	1.450000	0.9892182	-1.93629	4.836289	0.1468
2	3	1.100000	0.9892182	-2.28629	4.486289	0.2696
1	2	0.950000	0.9892182	-2.43629	4.336289	0.3399

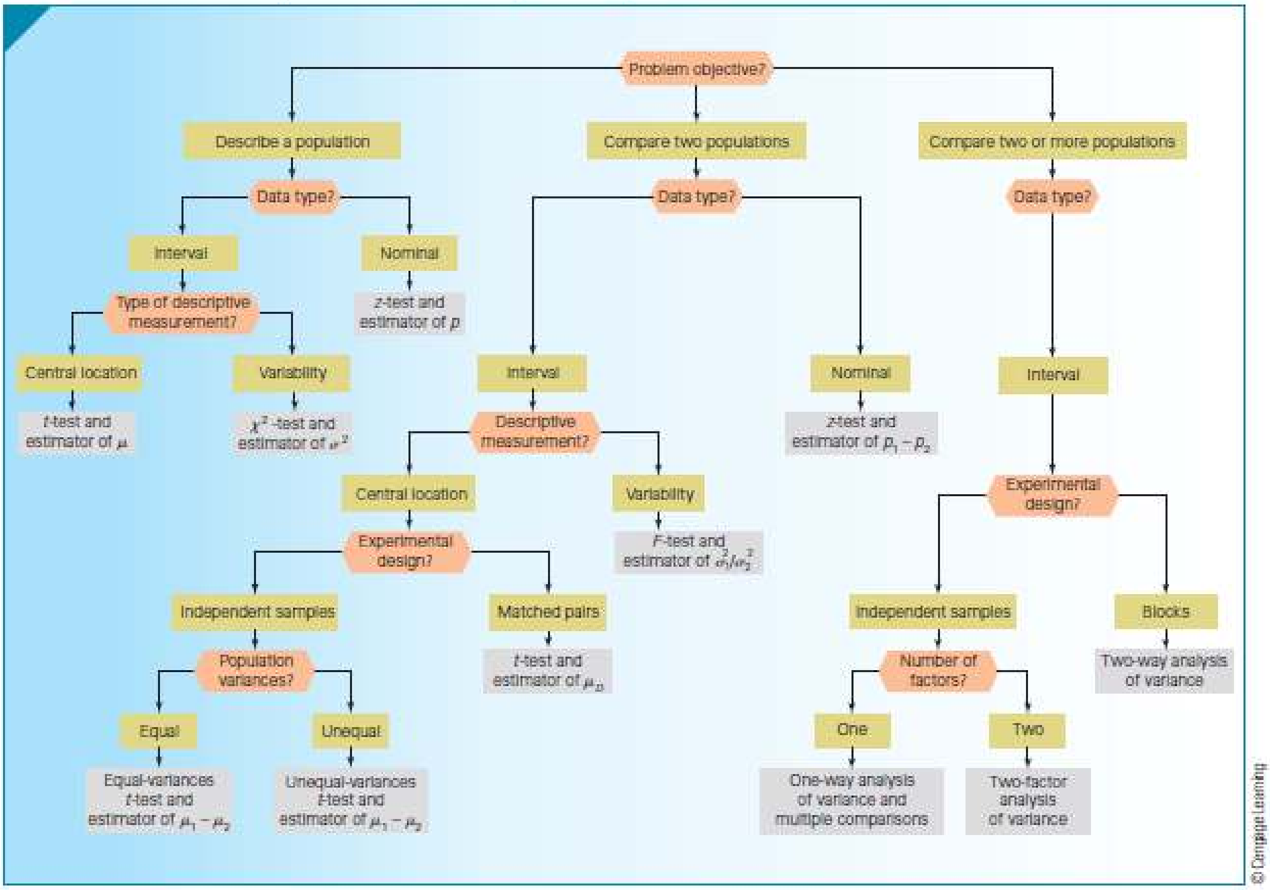


There is a difference in the no. of jobs between the people with no high school and people with university degree, which was expected from our previous analysis. This can also be confirmed by the Ordered differences report, since the only confidence interval that doesn't include 0 is corresponding to Education level 1 and 4.



Graphically, we can see that the mean for edu level 4 is different from mean of edu level 1.

FIGURE A14.1 Summary of Statistical Techniques in Chapters 12 to 14



Summary ANOVA

1. One way ANOVA- test whether the means of 2 or more population differ. The populations are characterized by one factor
2. Randomized blocks (2 way): the groups are matched such that the elements in each group have similar characteristics. Want to reduce variation caused by differences between experimental units.
3. 2 factor ANOVA: test whether the means of populations differ. The populations are characterized by 2 factors or more.

Goodness-of-Fit

↳ Tests if the *probabilities* of a multinomial distribution take a certain value. We deal with **nominal** random variables and describes ONE population of data.
Also tests: that the nominal variable follows a **uniform distribution**.

X^2-test (with k-1 degrees of freedom).

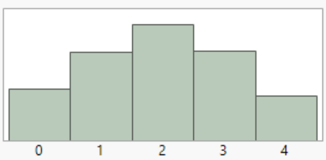
Goodness of Fit
Assumptions: *Rule of 5*: the **Expected** value in each cell >5 (otherwise *combine* cells to meet the assumption; recommend t0 gather more data)
Hypotheses: $H_0: \mathrm{p}_1 = \mathrm{a}_1, \mathrm{p}_2 = \mathrm{a}_2, \mathrm{p}_3 = \mathrm{a}_3 \ldots ..$ H_1: At least one p_i *is not equal to its specified value* a_i Where a are the values we want to test.
Significance level: α= 0.05
Test-statistic: $\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij}-e_{ij})^2}{e_{ij}} \sim X^2_{(k-1),\alpha}$, where *r* is the no. of rows and *c* the no. of columns
In JMP: - Check if the variable is **nominal.** - **Distribution: ∆Test probabilities and Confidence interval-** insert the probabilities (*if not given, use* 100/no. of variables- e.g. 5 variables, p=0.2) **-** Check for *Rule of 5 assumption:* Multiply the Total no. of observations with the calculated probability. If>5, the assumption is met. (*E.g. 700*0.2=140>5* ✓) - Look at *Pearson test:* for ChiSquare and p-value. Compare with Chi^2 crit (*excel X^2),* look at p-value and conclude.

All variables need to be nominal/ordinal.
*****Combine cells to get Expected>5.***

- **Assumptions:**

➔ *Rule of 5:* the expected should be higher than 5.
$\mathrm{e_{ij}} = 700 * 0.2 = 140 > 5$, so the assumption is fulfilled.

Distributions
Prev-jobs



Frequencies

Level	Count	Prob
0	92	0.13143
1	159	0.22714
2	209	0.29857
3	161	0.23000
4	79	0.11286
Total	700	1.00000

N Missing 0
5 Levels

Test Probabilities

Level	Estim Prob	Hypoth Prob
0	0.13143	0.20000
1	0.22714	0.20000
2	0.29857	0.20000
3	0.23000	0.20000
4	0.11286	0.20000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	85.3019	4	<.0001*
Pearson	82.7714	4	<.0001*

Method: Fix hypothesized values, rescale omitted

Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
0	92	0.13143	0.108398	0.158482	0.950
1	159	0.22714	0.197643	0.259621	0.950
2	209	0.29857	0.265844	0.333497	0.950
3	161	0.23000	0.200349	0.262598	0.950
4	79	0.11286	0.091499	0.138442	0.950
Total	700				

Note: Computed using score confidence intervals.

Looking at the Person test, the p-value is highly significant- smaller than 0.0001, which is implicitly smaller than 0.05, so H_0 is rejected. The X^2= 82.77.
$X^2_{(k-1),\alpha} = X^2_{(5-1),0.05} = 9.49$
$X^2_{crit} < X^2_{obs}$, so it confirms that the null hypothesis is rejected.

Conclusion:
Since the null is rejected, at least one probability is different. In other words, there is sufficient evidence to infer that the number of jobs of a person do not follow a uniform distribution. Also by looking at the distribution plot, the sample looks more normal, rather than uniform.

Contingency Tables

↳ Tests if there is a *dependence* between 2 or more populations (*are they independent?*). And if whether a *relation* exists between two or more populations of **nominal** random variables.

X^2-test (with (r-1)(c-1) degrees of freedom).

Contingency Table
Assumptions: *Rule of 5*: the **Expected** value in each cell >5 (otherwise *combine* cells to meet the assumption; recommend to gather more data)
Hypotheses: H_0: The two variables are independent H_1: The two variables are dependent
Significance level: α= 0.05
Test-statistic: $\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \sim X^2_{(r-1)(c-1),\alpha}$, where *r* is the no. of rows and *c* the no. of columns
In JMP: - Check if the variables are **nominal.** - **Fit Y by X**: *doesn't matter which is Y or X.* - Δ**Contingency Table:** *remove* Total%, Col%, Row% and *add* Expected, Deviation and Cell Chi^2. - Check for *Rule of 5 assumption* among the **Expected** count (>5). If it is not, combine some cells.**If cells are merged, the degrees of freedom will also change.* ↳ If Expected<5: make new column and merge 2 categories under one variable, using a Formula (*E.g. IF(Edu=1=>2, else=>Edu) – here education level 1 and 2 are merged)* - Look at the Contingency Table and Tests- *Pearson:* for ChiSquare and p-value. Compare with Chi^2 crit (*excel X²),* look at p-value and conclude.

All variables need to be nominal/ordinal.
****Combine cells to get Expected>5.***

Contingency Table

MBA Major	Count / Expected / Deviation / Cell Chi^2	Degree 1	2	3	4	Total
1	Count	31	8	12	10	61
	Expected	24.0789	12.4408	15.6513	8.82895	
	Deviation	6.92105	-4.4408	-3.6513	1.17105	
	Cell Chi^2	1.9893	1.5852	0.8518	0.1553	
2	Count	13	16	10	5	44
	Expected	17.3684	8.97368	11.2895	6.36842	
	Deviation	-4.3684	7.02632	-1.2895	-1.3684	
	Cell Chi^2	1.0987	5.5015	0.1473	0.2940	
3	Count	16	7	17	7	47
	Expected	18.5526	9.58553	12.0592	6.80263	
	Deviation	-2.5526	-2.5855	4.94079	0.19737	
	Cell Chi^2	0.3512	0.6974	2.0243	0.0057	
Total		60	31	39	22	152

Tests

N	DF	-LogLike	RSquare (U)
152	6	6.8905936	0.0343

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	13.781	0.0322*
Pearson	14.702	0.0227*

Assumptions:

➔ *The rule of 5:* the expected count in each cell needs to be higher than 5. Looking at the figure below, the assumption is met.

In JMP:

The p-value is highly significant- smaller than 0.05- so there is a dependency between the education and labour market experience.

X^2 is 14.7, according to the Pearson test.

$X^2_{(r-1)(c-1),\alpha} = X^2_{(3-1)(4-1),0.05} = 13$

$X^2_{obs}>X^2_{crit}$, so the null is rejected, reaching the same conclusion as the p-value: there is dependency between the bachelor degree and master. The test can be considered valid, since there were no problems with the assumptions.

Simple Linear Regression

↳ Predicts/ forecasts the value of one variable (Y) on the basis of other variables (Xs). All variables must be interval- **continuous.**

Deterministic model: determines the value of *Y* from the values of *Xs*. **No error term.**
Probabilistic model: method used to capture the *randomness* that is part of a real-life process. **Includes the error term** $(\varepsilon = actual - estimated)$**.**

Simple Linear Regression

Assumptions:

1. Random sample and reliable data. (*Can we generalise the results?)*
2. Variation in X variable. (*can be also seen under* **Parameter Estimates, STD Error** of the variable, if big x is in-variant to some extent)
3. No problem with multicollinearity, since there is only one independent variable in the model.
   Errors:
4. Normality of the error term ε. (*When the sample size is large, the assumption can be dropped based on Central Limit Theorem).*
5. The expected value ε is zero for the independent variable (Xs): $E(\, \varepsilon \mid X\,) = 0$
   ↳ **look for U-shape in the residual plot**=> not fulfilled**.**
6. Homoscedasticity (*constant variance)*- the variation around the regression line should be similar for all values of the independent variable (X): $Var(\, \varepsilon \mid X\,) = \sigma^2$ **(funnel shape)**
7. The error terms ε are independent of each other. (**patterns in the errors)**
8. Validity and trustworthiness. (*comment yourself)*

* If we wish to test for positive or negative linear relationships:
The null hypothesis remains: H0: β1 = 0.
H1: β1 < 0 (testing for a negative slope)
or
H1: β1 >0 (testing for a positive slope)

Model:

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

Slope coefficient formula: $b_1 = \frac{s_{xy}}{s_x^2}$

Intercept coefficient formula: $b_0 = \bar{y} - b_1\bar{x}$

Hypotheses:

Tests for linear relationship:	*Tests for positive relationship:*	*Tests for negative relationship:*
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_1 > 0$	$H_1: \beta_1 < 0$

Significance level:

α= 0.05

Test-statistic:

$t = \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{(n-2),\frac{\alpha}{2}}$, where $s_{b_1} = \frac{s}{\sqrt{(n-1)s_x^2}}$

For positive/negative relationships: t_{crit}=$t_{(n-2),\alpha}$ and p-value=p-value JMP/2

Prediction interval:

$$PI = \hat{y} \mp t_{\frac{\alpha}{2},n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

Confidence interval:

$$CI = \hat{y} \mp t_{\frac{\alpha}{2},n-2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

In JMP:

- Check if the variables are **continuous.**
- **Fit Y by X**: *Y- dependent, X- independent.*
- **Bivariate Fit** graph: check assumptions of variation in X and linear relationship between X&Y. Can also look at: **Graph Builder:** *Y- dependent, X- independent.*
- **ΔFit line:** to get the model

↳ Under ΔLinear Fit -> **Save Predicteds, Mean Confidence Limit Formula and Indiv Confidence limit formula.** To make a *forecast*, insert the given number in a new row under X: look at Predicted- *Expected value*, mean confidence interval for *Prediction interval* and for the individual *confidence interval*, both with 95% certainty.

↳ Under ΔLinear Fit ->**Save residuals (**if a residual>∓2***: outlier****- Sensitivity analysis: remove it and run the model again. Rsquare should increase. Compare before and after)*;

Plot Residuals: look at **Residual by Predicted Plot** to check for assumptions about errors (normality in errors, zero conditional mean, heteroscedasticity, independent errors).

- ➔ Asses how well the model fits the data, look at sum of squares for errors (SSE)-***Root Mean Square Error*** *in JMP.* The smaller it is, the better is the model. Compare it to ***Mean of Response*** in order to conclude if it's small=good.
- ➔ Interpret ***R square****:* how much of the variation in y is explained by variation in x. The bigger the better. (%)

Level-level, log-level, level- log and log-log Models:

1. Create new columns with formulas based on Y and X: log(Y) and log(X) *(log is under Transcendental in formula)*
2. Fit **Y by X** for all four combinations.
3. To decide which model is better, look at the **Bivariate Fit** graph which *follows the line* the most.

Level-level: *1 unit increase in x=>* $\boldsymbol{b_1}$ *increase in y. When* $b_1=0$*, y is* b_0*.*	Log-level: *1 unit increase in x=>* $\boldsymbol{b_1}$***%*** *increase in y. When* $b_1=0$*, y is* $b_0=e^{b0}$ *(exponential)*
Level-log: ***1%*** *increase in x=>* b_1 *increase in y. When* $b_1=0$*, y is* b_0*.*	Log-Log: ***1%*** *increase in x=>* $\boldsymbol{b_1}$***%*** *increase in y.*

**Log is the % change.*

****All variables need to be continuous.*** *But X can also be a* *<u>dummy variable.</u>*

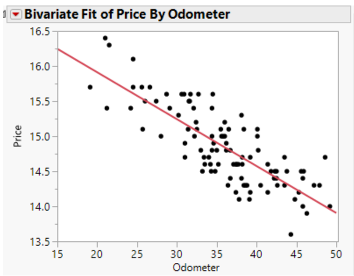
**We only test how X affects Y, not both!*

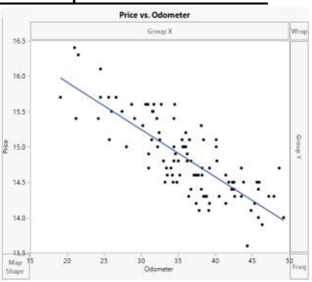
***Log can be used only with positive data.*

<u>Assumptions:</u>

1. <u>Random sample and reliable data.</u> *(Can we generalise the results?)*
   It is assumed that the data is collected randomly and that it is reliable.

2. <u>Variation in X and linear relationship between Y & X</u>

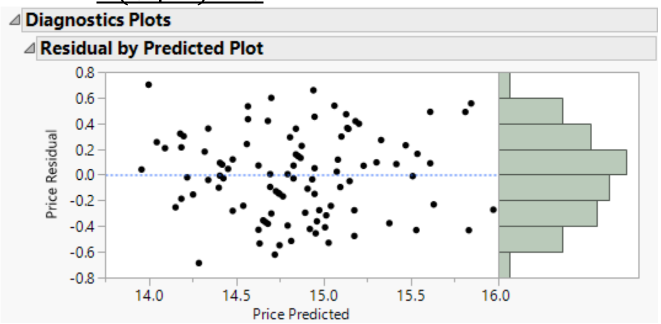




From the bivariate fir graph it can be seen that the independent variable- Odometer- is not constant and that the observations are following a line, indicating that there is a linear relationship between the two variables: Price and Odometer. So these assumptions are met.

3. <u>Multicollinearity</u>
   No problem with multicollinearity, since there is only one independent variable in the model.

4. Normality of the error term ε and the expected value ε is zero for the independent variable (Xs): $E(\varepsilon \mid X) = 0$



The residuals follow a bell shape with the mean close to zero, so the assumptions of normality and zero conditional mean are met.

5. Homoscedasticity (*constant variance)*- the variation around the regression line should be similar for all values of the independent variable (X): $Var(\varepsilon \mid X) = \sigma^2$
   There is not funnel shape present in the errors, so the assumption is met.

6. The error terms ε are independent of each other
   Again, there is no clear pattern, so it can be inferred that the errors are independent.

7. Validity and trustworthiness
   It is assumed that the data is valid and trustworthy.

In JMP:

Linear Fit

Price = 17.248727 - 0.0668609*Odometer

Summary of Fit

RSquare	0.648295
RSquare Adj	0.644707
Root Mean Square Error	0.326489
Mean of Response	14.841
Observations (or Sum Wgts)	100

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	19.255607	19.2556	180.6430
Error	98	10.446293	0.1066	Prob > F
C. Total	99	29.701900		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	17.248727	0.182093	94.73	<.0001*
Odometer	-0.066861	0.004975	-13.44	<.0001*

From Analysis of Variance table: The p-value smaller than 0.05, so the model is highly significant.
From Parameter Estimates table: The p-value smaller than 0.05, so the explanatory variable, Odometer, is also highly significant.

Prediction expression:
Price = 17.248727 - 0.0668609*Odometer

Conclusions:
RSquare is 0.65, meaning that 65% of the variation in price is explained by the odometer variables. This means that there are other variables influencing the price of the car, and further investigation would be necessary in order to determine them.

Root Mean Square Error is 0.33, while the **Mean of Response** is 14.84. So SSE is definitely smaller, meaning that the model fits the data quite well, having only little room for error.

t_{obs}= 13.44
t_{crit}=$t_{98,\frac{0.05}{2}}$= 1.98
T-test: t_{obs}> t_{crit} , so the null hypothesis is rejected, meaning that there is a linear relationship between Price and Odometer.

Looking at the model: If no miles are driven- Odometer=0- the price of a car is 17.248£. And if the slope parameter is increased by 1 unit- 1000 miles- the price is reduced by 67£, for each 1000 miles. In other words, the price per mile in terms of reduced value is 6.70 cents.

All in all, there is a highly significant relationship between the price of the car and odometer reading.

Expected Value, Prediction interval and Confidence interval (95%):
For x= 34.000 miles (*new column)*

Predicted Price	Lower 95% Mean Price	Upper 95% Mean Price	Lower 95% Indiv Price	Upper 95% Indiv Price
14.97545724	14.907693322	15.043221159	14.324017143	15.626897338

**In this exercise the numbers are in thousands, so be careful when interpreting the results.*

Multiple Regression

↳ Predicts/ forecasts the value of one variable (Y) on the basis of other multiple variables (Xs). All variables must be interval- **continuous.** *Possible also with X* **nominal- dummy variable,** but keep it **continuous.**

**It is expected that multiple regression model fits the data better than a simple regression model.*

Multiple Regression

First comment on each X included in the model- why is it relevant?

Assumptions:

1. Random sample and reliable data. (*Can we generalise the results?)*
2. Variation in X variable. (**Tabulate: Min** and **Max** of Xs)
3. Linearity between Y and X (**Graph Builder**: Y against all Xs- Lambda smoothing)
   Errors:
4. Normality of the error term ε. (Save Residuals- **Distribution** residuals. *When the sample size is large, the assumption can be dropped based on Central Limit Theorem).*
5. The expected value ε is zero for the independent variable (Xs): $E(\varepsilon \mid X) = 0$
   ↳ **Residual by predicted plot: look for U-shape in the residual plot**=> not fulfilled**.**
6. Homoscedasticity (*constant variance)*- the variation around the regression line should be similar for all values of the independent variable (X): $Var(\varepsilon \mid X) = \sigma^2$ **(funnel shape)**
7. The error terms ε are independent of each other. **(patterns in the errors)**
8. Validity and trustworthiness. (*comment yourself)*
9. Multicollinearity (**Multivariate Methods- Multivariate** all Xs; **CI of Correlation** check if there is 0 in the intervals) *normal correlation between x and x^2*

Model:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \varepsilon$

Quadratic e.g.:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$ or $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$

Hypotheses:

Tests for linear relationship:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \cdots = 0$

$H_1: At\ least\ one\ \beta_i \neq 0$

**If the null is true, no X is linearly related to y.*

Significance level:

α= 0.05

Test-statistic:

$F_{crit} = \frac{MSR}{MSE} \sim F_{\alpha,k,n-k-1}$

Dummy variables:

If Color == 1 ⇒ 1
else ⇒ 0

↳ Put in 1 the true value. (Male=1-Female)

A nominal variable can be transformed into dummy: *new column, new formula (if is under Conditional), but keep it **continuous** to perform regression analysis* e.g.:

*When transforming a variable into more dummies, keep one out to have it as a reference group, but include the rest, even if *are NOT significant. Mention that there is no evidence that the insignificant dummy affects Y, while the other Xs are constant.*

Interpret: in the final model after estimating the parameters, the X dummy variable can be only 1/0, e.g. : Price=b0+b1Color ⇔ Price= 43+ 2Color ⇔ Price= 45 when Color=1

In JMP (first model, afterwards assumptions):

- Check if the variables are **continuous.**
- Check normality for all variables: **Distribution,** Y- all variables.
- **Graph Builder** to check if quadratic: ∩ or ∪ shape; if yes, use $\mathbf{x^2}$ for the quadratic Xs (new column, new formula x*x) and create the interaction (also x_1*x_1) in the model (*polynomial).*
- **Fit Model:** *Y- dependent, Add- independent.* **∆Model Dialog:** Remove the insignificant variables, one by one, starting with the biggest p-value. (Check for possible ***outliers****- Sensitivity analysis: remove it and run the model again. Adjusted Rsquare should increase. Compare before and after)*;

You can group* **By *if it makes sense, before running the model.*

- **∆Estimates -> Show Prediction Expression:** the intercept may not have sense to interpret- how much is Y when all Xs are zero; for each variable, increasing one X with 1 units, holding the rest constant, y increases/decreases.

  *For *dummy variable* (e.g. gender): e.g. being a male increases/decreases y with ...

  Assumptions:

- **Tabulate: Min** and **Max** for the significant Xs- to check for *Variation in X.*
- **Graph Builder: Y** and **multiple Xs** to check for *Linearity between X and Y.* Remember to drag to max. **Lambda Variables** in the down left part, to make the line smooth.
- **∆Save columns** -> **Residuals.** Check for *Normality of the errors*- **Distribution- Residual.** Under **Fitter Normal** (*in right side)->* **∆Goodness of Fit:** if p-value> 0.05, the errors are normal.

H_0: the errors are normally distributed

H_1: the errors are not normally distributed

- Save the **Residual by Predicted Plot** from the bottom of the model analysis. Check for *Zero conditional mean (no U shape), Homoscedasticity (no funnel shape), the error terms are independent (no patterns).*
- **Analyze -> Multivariate Methods-> Multivariate: Y** all significant Xs. **∆CI of Correlation-** check for *Multicollinearity:* if correlations >∓0.2 and there is no 0 in the confidence interval => correlation between the 2 variables.
- **∆Save Columns** -> **Save Predicted Values, Mean Confidence Limit Formula and Indiv Confidence limit formula.** To make a *forecast*, insert the given number in a new row under X: look at Predicted- *Expected value*, mean confidence interval for *Prediction interval* and for the individual *confidence interval*, both with 95% certainty.

➔ Asses how well the model fits the data, look at sum of squares for errors (SSE)-***Root Mean Square Error*** *in JMP.* The smaller it is, the better is the model. Compare it to ***Mean of Response*** in order to conclude if it's small=good (*no heteroscedasticity*).

➔ Interpret ***Adjusted R square****:* how much of the variation in y is explained by variation in x. The bigger the better. (%)

➔ F-test (*the model has explanatory power?)* and p-value.

**large F indicates that most of the variation in Y is explained by the model, while a small F indicates that most of the variation in Y is unexplained.*

➔ To ***exclude*** some observations that meet a certain condition: **Ctrl+Shift+W** to select which rows (*put condition, add, ok)* and **Ctrl+E** to exclude them.

<u>Level-level, log-level, level- log and log-log Models:</u>

1. Create new columns with formulas based on Y and X: log(Y) and log(X) *(log is under Transcendental in formula)*
2. **Fit Model** for all four combinations.
3. To decide which model is better, look at **Adjusted R square** of each with the same Y and same number of Xs. If different, look at **F-ratio**. Also look at the assumptions.

Level-level: *1 unit increase in x=> b_1 increase in y. When b_1=0, y is b_0.*	Log-level: *1 unit increase in x=> b_1% increase in y. When b_1=0, y is $b_0=e^{b0}$ (exponential)*
Level-log: ***1%*** *increase in x=> b_1 increase in y. When b_1=0, y is b_0.*	Log-Log: ***1%*** *increase in x=> b_1% increase in y.*

**Log is the % change.*

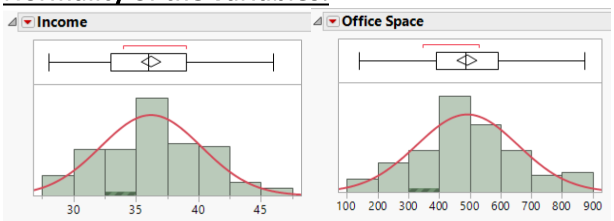
****All variables need to be continuous.*** *You can change a <u>dummy variable</u> to* ***continuous.*** (e.g. gender)

**We only test how X affects Y, not both!*

***Log can be used only with <u>positive</u> data.*

Source of Variation	degrees of freedom	Sums of Squares	Mean Squares	F-Statistic
Regression	k	SSR	MSR = SSR/k	F=MSR/MSE
Error	n–k–1	SSE	MSE = SSE/(n–k–1)	
Total	n–1	$\sum(y_i - \bar{y})^2$		

<u>Normality of the variables:</u>



In general, the variables are following a bell shape, meaning that they are normally distributed. The biggest issues are with variables Income and Office Space. But for now Gaussianity is considered met and the analysis is continued.

In JMP:

Summary of Fit

RSquare	0.525062
RSquare Adj	0.49442
Root Mean Square Error	5.512084
Mean of Response	45.739
Observations (or Sum Wgts)	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	3123.8320	520.639	17.1358
Error	93	2825.6259	30.383	Prob > F
C. Total	99	5949.4579		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	38.138575	6.992948	5.45	<.0001*
Number	-0.007618	0.001255	-6.07	<.0001*
Nearest	1.6462371	0.632837	2.60	0.0108*
Office Space	0.0197655	0.00341	5.80	<.0001*
Enrollment	0.2117829	0.133428	1.59	0.1159
Income	0.4131221	0.139552	2.96	0.0039*
Distance	-0.225258	0.178709	-1.26	0.2107

From *Analysis of Variance table*: The p-value smaller than 0.05, so the model is highly significant.

From *Parameter Estimates table*: There are 2 insignificant variables: Distance and Enrolment, with p-values higher than 0.05. We first remove the most insignificant one, in this case Distance.

After removing it, Enrolment is still insignificant, so it is also removed.

Summary of Fit

RSquare	0.505791
RSquare Adj	0.484982
Root Mean Square Error	5.563295
Mean of Response	45.739
Observations (or Sum Wgts)	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	3009.1836	752.296	24.3066
Error	95	2940.2743	30.950	Prob > F
C. Total	99	5949.4579		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	41.27315	6.414992	6.43	<.0001*
Number	-0.007863	0.00126	-6.24	<.0001*
Nearest	1.6536505	0.635316	2.60	0.0107*
Office Space	0.0196075	0.003439	5.70	<.0001*
Income	0.3993871	0.139886	2.86	0.0053*

The remaining variables – Number, Nearest, Office space and Income- are significant, with p-values smaller than 0.05.

Prediction expression:

41.2731501754566
+ -0.0078625222005 * Number
+ 1.65365049192422 * Nearest
+ 0.01960749156581 * Office Space
+ 0.39938712073332 * Income

Margin= 41.27- 0.008Number+ 1.65Nearest+ 0.02Office Space+ 0.4Income

Assumptions:

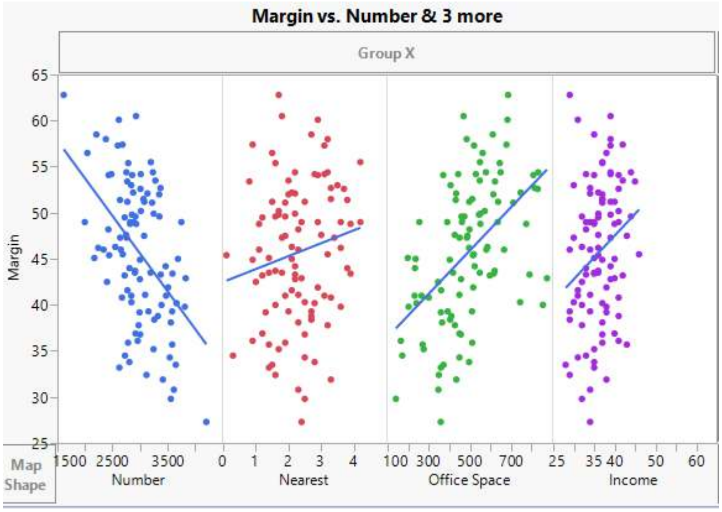
1. Random sample and reliable data. (*Can we generalise the results?)*
   It is assumed that the data is collected randomly and that it is reliable.

2. Variation in X

	Number	Nearest	Office Space	Income
Min	1613	0.1	140	28
Max	4214	4.2	875	60

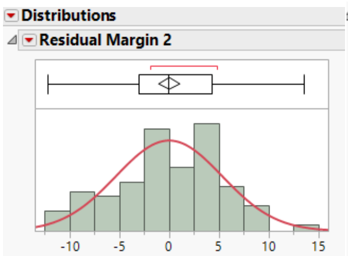
Looking at the minimum and maximum value under each variable, it is clear that there is variation among the observations: min≠ max.

3. Linear relationship between Y & X



Plotting the Y variable, Margin, against the independent variables, the linear relationships can be distinguished.
There seem to be problems with the assumptions across all the explanatory variables, since the observations are spread randomly and do not follow clearly a line.

4. Normality of the error terms ε



Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	2.096e-15	-1.060056	1.0600562
Dispersion	σ	5.342441	4.6906998	6.2061815

-2log(Likelihood) = 617.92423908687

Goodness-of-Fit Test

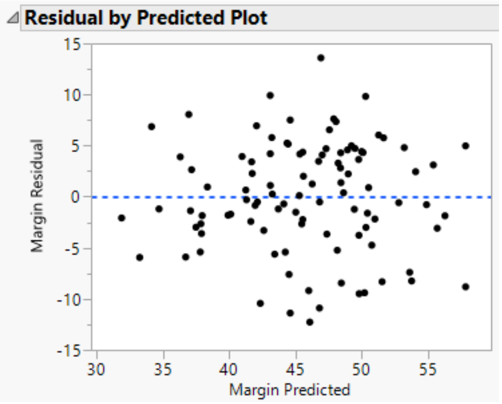
Shapiro-Wilk W Test

W	Prob<W
0.982945	0.2239

H_0: the errors are normally distributed
H_1: *the errors are not normally distributed*

The Goodness of Fit test gives a p-value of 0.22 which is bigger than the significance level of 0.5. So there is not enough evidence to reject the null hypothesis, meaning that the errors follow th normal distribution.

5. The expected value ε is zero for the independent variable (Xs): $E(\varepsilon \mid X) = 0$



There is no U shape in the errors, so the zero conditional mean is met.

6. Homoscedasticity (*constant variance*)- the variation around the regression line should be similar for all values of the independent variable (X): $Var(\varepsilon \mid X) = \sigma^2$
There is not funnel shape present in the errors, so the assumption is met.

7. The error terms ε are independent of each other
Again, there is no clear pattern, so it can be inferred that the errors are independent.

8. Validity and trustworthiness
It is assumed that the data is valid and trustworthy.

9. Multicollinearity

Multivariate

Correlations

	Number	Nearest	Office Space	Income
Number	1.0000	0.0507	-0.0937	0.0310
Nearest	0.0507	1.0000	0.0438	-0.0400
Office Space	-0.0937	0.0438	1.0000	0.1528
Income	0.0310	-0.0400	0.1528	1.0000

CI of Correlation

Variable	by Variable	Correlation	Lower 95%	Upper 95%
Nearest	Number	0.0507	-0.1462	0.2437
Office Space	Number	-0.0937	-0.2839	0.1036
Office Space	Nearest	0.0438	-0.1530	0.2372
Income	Number	0.0310	-0.1654	0.2251
Income	Nearest	-0.0400	-0.2336	0.1567
Income	Office Space	0.1528	-0.0439	0.3382

Rule of thumb: if the value is bigger than ∓0.2, it can be said with 95% confidence that there is a correlation between the 2 variables. This can be also checked by looking at the confidence intervals: if there is 0 in the confidence interval, we do NOT have correlation.
So the multicollinearity assumption is fulfilled.

Conclusions:
Adjusted RSquare is 0.48, meaning that 48% of the variation in operating margin is explained by the four explanatory variables. This means that there are other variables influencing the operating margin, and further research would be necessary in order to determine them.

Root Mean Square Error is 5.56, while the **Mean of Response** is 45.74. So SSE is much smaller, meaning that the model fits the data quite well, not having much room for error.

F-test:
F_{obs}= 24.3
$F_{crit}=F_{\alpha,k,n-k-1} = F_{0.05;4;95}= 2.47$
F-test: $F_{obs} > F_{crit}$, so the null hypothesis is rejected, meaning that there is a linear relationship between the margin and the four independent variables.

P-value: the model is significant, with a p-value of almost 0. So the null hypothesis is rejected. Thus it can be inferred that at least one independent variable has an impact on the operating margin. But after considering the p-values of each variable, all four are significant, with p-values smaller than 0.05.

Looking at the model: If there are no motels/ hotels and offices in the area, and the community would have no household income, than the operating margin would be 41.27%.
Number: for each additional motel in the area, the margin decreases by 0.008, assuming that the other explanatory variables in the model are held constant.
Nearest: for each additional mile to the closest competition, the margin increases by 1.65, assuming that the other explanatory variables in the model are held constant. And so on.

*For dummy variable (e.g. gender): e.g. being a male increases the margin with 13.

Expected Value, Prediction interval and Confidence interval (95% certainty):
For the give Xs (*new row, make sure the model includes only the given Xs)*

Pred Formula Margin	Lower 95% Indiv Margin	Upper 95% Indiv Margin	Lower 95% Mean Margin	Upper 95% Mean Margin
42.292067501	31.163640047	53.420494955	40.928218864	43.655916138

Logistic Regression

↳ Predicts/ forecasts the value of one variable (Y) on the basis of another (X). The explanatory variables Xs can be **binary nominal** or **continuous,** but classified as **continuous** and Y must be **binary nominal**.

*We ensure that the predicted probabilities of y are between **0 and 1.**

Logistic Regression

First comment on each X included in the model- why is it relevant?

Assumptions:

1. Probability of Success Y variable: **Tabulate: Y** and **% of Total** **-** *Probability of success should not be too extreme.*
2. Random sample and reliable data. (*Can we generalise the results?)*
3. Y must be binary nominal.
4. Variation in X variable. (**Distribution:** Xs- *make sure they are classified correctly*)
5. Validity and trustworthiness. (*comment yourself)*
6. Multicollinearity (**Multivariate Methods- Multivariate** all Xs; **CI of Correlation** check if there is 0 in the intervals) *normal correlation between x and x^2*

Model:

$\text{logit}(y) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \varepsilon$, where y is the odds ratio.

True model: $\ln(\hat{y}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \ldots$, where $\hat{y} = e^{\ln(\hat{y})}$

Odds ratio:

$$Odds\ ratio = \frac{Probability\ of\ event}{1 - Probability\ of\ event} = \frac{Probability\ of\ event}{Probability\ of\ failure}$$

Indicates how many times larger the probability of success (of the event happening) is than the probability of failure.

$$Probability\ of\ event = \frac{\hat{y}}{1+\hat{y}} = \frac{Odds\ ratio}{1 + Odds\ ratio}$$
$$= \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \cdots)}$$

You can calculate a specific probability by using given values of X in the estimated model.

Hypotheses:

Tests for linear relationship:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \cdots = 0$

$H_1: At\ least\ one\ \beta_i \neq 0$

**If the null is true, no X is linearly related to y.*

***Check how many Xs to have the correct no. of* βs.

Significance level:

α= 0.05

Test-statistic:

$F_{crit} = \frac{MSR}{MSE} \sim F_{\alpha,k,n-k-1}$

Dummy variables:

↳ Put in 1 the true value. (Male=1-Female)

A nominal variable can be transformed into dummy: *new column, new formula (if is under Conditional).*

If Color == 1 ⇒ 1
else ⇒ 0

*When transforming a variable into more dummies, keep one out to have it as a reference group, but include the rest, even if *are NOT significant. Mention that there is no evidence that the insignificant dummy affects Y, while the other Xs are constant.*

<u>Interpret:</u> in the final model after estimating the parameters, the X dummy variable can be only 1/0, e.g. : Price=b0+b1Color $\Leftrightarrow$ Price= 43+ 2Color $\Leftrightarrow$ Price= 45 when Color=1

<u>In JMP (first model, afterwards assumptions):</u>

- Check if Y is **binary nominal** and Xs **binary nominal or continuous,** but all Xs need to be classified as **continuous.**

- Reverse 1/Yes and 0/No in the **nominal variable Y: Column info-> Column Properties-> Value Ordering-** Make sure 1/Yes is first. *To test for the probability of success, because JMP automatically tests fort he probability of failure.*

- Check normality for Y: **Distribution**. *The distribution should be equal between the 2 groups.*

-

- **Fit Model:** *Y- dependent nominal, Add Xs- independent continuous.*

- **ΔOdds Ratios:** Look at *Unit odds Ratios-* if the X is increased by 1 unit, Y increases/decreases by **100(Odds ratio-1)=_%.** *Range Odds Ratios* compares the lowest and the highest value in X, and the Odds ratio tells the difference between them in %.

- **ΔConfusion Matrix:** There are 3+56=59 observations true, which the model predicts 1 to be positive and 129 to be negative.

Confusion Matrix

Training

Actual	Predicted	
Remedial	Yes	No
Yes	3	56
No	1	129

Hit rate=(3+129)/(3+56+1+129)= 70%

and

- **Misclassification Rate:** Hit ratio=100%- Misclassification rate%. *The Hit Rate needs to be higher than* **(%of Total if True/1/Yes * 25%)=**Total if true*(1+0.25)=**_%** in order to declare the model good.

- ΔSave Probability Formula: used to make a forecast. Insert the given numbers in a new row under Xs: look at **Lin[Yes]-** y prediction, **Prob[Yes]-** probability of success, **Prob[No]-** probability of failure, **Most Likely-** yes or no.

- **Graph Builder** to check if quadratic: ∩ or ∪ shape; if yes, use $\mathbf{x^2}$ for the quadratic Xs (new column, new formula x*x) and create the interaction (also x_1*x_1) in the model (*polynomial)*.

- **Analyze -> Multivariate Methods-> Multivariate: Y** all significant Xs. **ΔCI of Correlation-** check for *Multicollinearity:* if correlations >$\mp$0.2 and there is no 0 in the confidence interval => correlation between the 2 variables.

<u>Interpret:</u>

- ➔ *Whole model test:* Look at X^2 **p-value** to see if it is significant (<0.05)- So we can reject the null, of coefficients being equal to zero. **Entropy Rsquare%-** shows how much the model explains the variation in y, rather than a model without any Xs.
- ➔ *Parameter Estimates:* check Xs significant (<0.05), otherwise remove. Only comment on the sign of the coefficient: positive or negative impact on Y.
- ➔ *Unit Odds Ratios:* if X is increased by 1 unit => probability of Y changes by **100(Unit Odds Ratio-1)=_%**
- ➔ *Range Odds Ratios:* Compares the lowest value with the highest value, and the change between them in X.

- ➔ To ***exclude*** some observations that meet a certain condition: **Ctrl+Shift+W** to select which rows (*put condition, add, ok)* and **Ctrl+E** to exclude them.

All X variables need to be classified continuous, Y needs to be <u>binary nominal</u>.

**We only test how X affects Y, not both!*

****Log can be used only with <u>positive</u> data.*

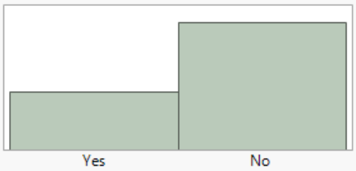
*****Do NOT remove related variables, such as age, before removing agesquare.**

Can use Contingency tables to determine if 2 variables are dependent, before running the logistic regression.

Normality of dependent variable Y:

Distributions

Remedial



Frequencies

Level	Count	Prob
Yes	59	0.31217
No	130	0.68783
Total	189	1.00000

N Missing 0
2 Levels

59 children have been assigned to remedial training, from the total of 189. The distribution is not equal between the two categories, which may be a problem, but the analysis is continued.

Assumptions:

1. Probability of Success Y variable:

Tabulate: Y and **% of Total** - *Probability of success should not be too extreme.*

Remedial	% of Total
Yes	31.22%
No	68.78%

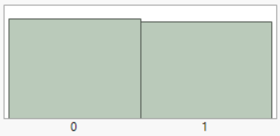
2. Random sample and reliable data. (*Can we generalise the results?)*
   It is assumed that the data is random and representative for the population.

3. Y must be binary nominal.
   Y is represented by Remedial and it is binary- the students have or have not received remedial training.

4. Variation in X variable. (**Distribution:** Xs- *make sure they are classified correctly*)

Distributions

Boy

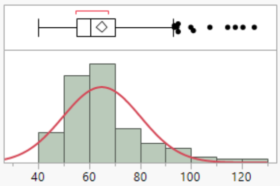


Frequencies

Level	Count	Prob
0	96	0.50794
1	93	0.49206
Total	189	1.00000

N Missing 0
2 Levels

Score



There is almost equal distribution between the boys and girls- 96 girls and 93 boys.
The distribution of score is not so well spread, being a focus around the 60 points mark and very few observations between 80 and 120. This may impact the final results.

5. Validity and trustworthiness. (*comment yourself)*
   It is assumed that the data provided is valid and trustworthy. In other words, there are no errors in the data.

6. Multicollinearity (**Multivariate Methods- Multivariate** all Xs; **CI of Correlation** check if there is 0 in the intervals)
   There is no correlation between Boy and Score, since the value is very low and there is 0 in the confidence interval. So there are no problems with multicollinearity.

Multivariate

Correlations

	Boy	Score
Boy	1.0000	-0.0745
Score	-0.0745	1.0000

CI of Correlation

Variable	by Variable	Correlation	Lower 95%	Upper 95%
Score	Boy	-0.0745	-0.2150	0.0689

In JMP:

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	5.00973	2	10.01945	0.0067*
Full	112.32627			
Reduced	117.33600			

RSquare (U)	0.0427
AICc	230.782
BIC	240.378
Observations (or Sum Wgts)	189

Measure	Training	Definition
Entropy RSquare	0.0427	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0726	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5943	∑ -Log(ρ[j])/n
RMSE	0.4513	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.4074	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3016	∑ (ρ[j]≠ρMax)/n
N	189	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	99	60.37850	120.757
Saturated	101	51.94778	Prob>ChiSq
Fitted	2	112.32627	0.0678

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.53399149	0.8108589	0.43	0.5102
Boy	0.6476915	0.3248274	3.98	0.0462*
Score	-0.0261372	0.0122332	4.56	0.0326*

For log odds of Yes/No

Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Boy	1	1	4.0381236	0.0445*
Score	1	1	5.20479255	0.0225*

Odds Ratios

For Remedial odds of Yes versus No

Tests and confidence intervals on odds ratios are likelihood ratio based.

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Boy	1.911124	1.016023	3.644726	0.5232523
Score	0.974201	0.949552	0.996503	1.0264818

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Boy	1.911124	1.016023	3.644726	0.5232523
Score	0.108429	0.012277	0.742503	9.2226605

Interpret:

Whole model test: Looking at X^2, it is significant: 0.0067<0.05. So we can reject the null, of coefficients being equal to zero. The *Entropy Rsquare* is 0.0427, so the model explains 4% more of the variation in y rather than a model without any explanatory variables.

Parameter Estimates: both boy and score are significant, with p-values< 0.05. Looking at the coefficients provided by JMP, boy has a positive influence with a value of 0.64 and score a negative influence, with a value of 0.03. Thus, boys are more likely to receive remedial training and children with high scores are less likely to receive it.

Unit Odds Ratios: Per unit changes are for boy 1.9 (the odds of being assigned remedial training are 1.9 higher than for a girl) and for score 0.97 (the one unit increase in the test score results is 0.97 times higher).

↳ So for a boy, the probability of remedial training *increases* by 100(1.91-1)= 91% and for one extra point in the score variable, the probability of remedial training *decreases* by 100(0.10-1)=-90%.

Range Odds Ratios: Compares the lowest value with the highest value, and the change between them, in the explanatory variable. So there is a 1.91% difference between a boy and a girl, and 0.10% difference between the lowest and highest scores.

Confusion Matrix

Training

Actual Remedial	Predicted Yes	Predicted No
Yes	3	56
No	1	129

Confusion Matrix: There are 3+56=59 observations that have taken the remedial training, which the model predicts that (3+1)=4 to receive remedial training. But, out of (1+129)=130 children not receiving the training, 129 are predicted correctly.

Hit rate=(3+129)/(3+56+1+129)= 70%

Hit rate= 1- Misclassification rate= 1- 0.30=70%

(**31.22%***25%)=0.3122*(1+0.25)=39%

39%< 70% so the model is very good.

Remedial	% of Total
Yes	31.22%
No	68.78%

Prediction:
For a boy with a score of 120, it can be said with 95% certainty that there is a 12% probability that he will receive remedial training and 88% that he will no. So most likely he will not receive the training.

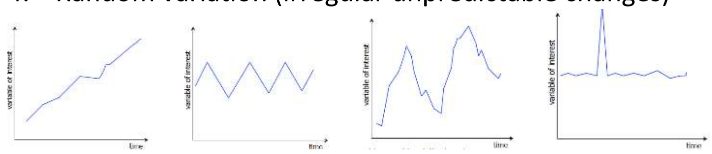
1	120	0	•	-1.954783204	0.1240327345	0.8759672655 No

Time Series and Forecasting

↳a variable measured over time, in sequential order, is analysed to detect *patterns* for **forecasting** future values.

Times Series components:

1. Long-term trend (steady variation)
2. Cyclical variation (wavelike pattern)
3. Seasonal variation (short term repetitive behaviour)
4. Random variation (irregular unpredictable changes)



Random variation can be reduced by smoothing:

➔ **Moving averages** *(forward looking)***-** the arithmetic mean of the values in that time period and those close to it.
E.g. 3 period: $\overline{x_t} = \frac{x_{t-1}+x_t+x_{t+1}}{3}$

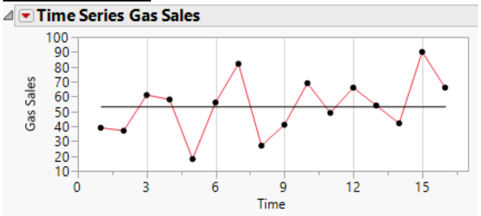
↳ *Bad because:* there are no moving averages for the first and last sets of time periods- we lose data in the ends; and it forgets most of the previous time-series values- only looks at those around it.

➔ **Exponential smoothing** *(backward looking)***:** $S_t = \omega y_t + (1-\omega)S_{t-1}$ for t≥2 and $S_1 = y_1$, where y_t is the time series at time t (*the original data)* and w is a smoothing constant (0≤ ω ≤1)
↳ *Solves the issues of moving averages.*
In JMP- !not reliable:
- **Analyse-> Modelling -> Time Series:** add Y Time series (*what we test)* and X Time (*time variable)* ***both continuous.*** *Doesn't matter what no. you choose.*
- **ΔSmoothing model-> Simple Exponential Smoothing.** ***Constraints:*** **Custom,** ***Level:*** **Fixed.** Insert given *weights.*
- **ΔSave Columns:** new columns appear. Look at ***Predicted*** too see the forecast for a specific period.

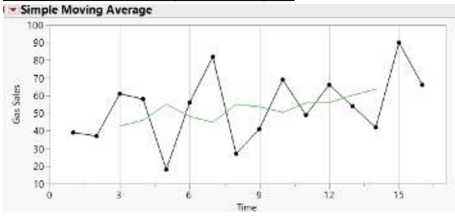
Time Series- Moving averages	
Model: *E.g. 3 period:* $\overline{x_t} = \frac{x_{t-1}+x_t+x_{t+1}}{3}$ *E.g. 5 period:* $\overline{x_t} = \frac{x_{t-2}+x_{t-1}+x_t+x_{t+1}+x_{t+2}}{5}$	
In JMP: - Check if the variables are **continuous.** - **Analyse-> Modelling -> Time Series:** add Y Time series (*what we test)* and X Time (*time variable)* ***both continuous.*** *Doesn't matter what no. you choose.*	
- **ΔSmoothing model-> Simple Moving Average.** Select ***Centred and double smoothed for even number of terms.*** *Can be done for more periods- 3,4,5 etc.* - **ΔSave to data table:** new column appears with the moving averages.	**-ΔARIMA** **- ΔSave Columns-** new columns appears, which can be used to forecast (*just look at the specific time, at the predicted value).*

All variables need to be continuous.

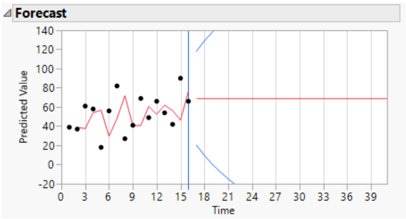
Time Series:



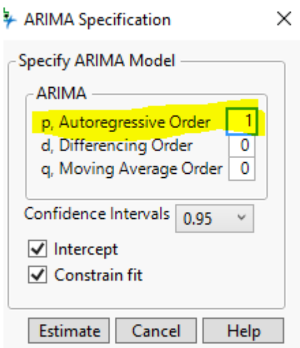
Simple moving average:



Exponential Smoothing:



ARIMA:



Model: AR(1)

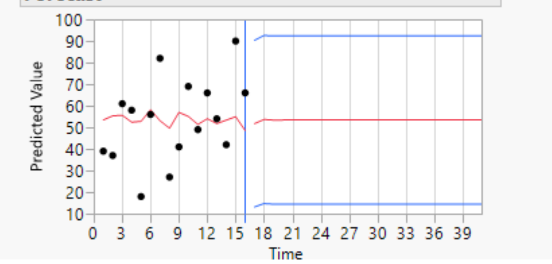
Model Summary

DF	14	Stable	Yes
Sum of Squared Errors	5429.89184	Invertible	Yes
Variance Estimate	387.849417		
Standard Deviation	19.6938929		
Akaike's 'A' Information Criterion	142.657325		
Schwarz's Bayesian Criterion	144.202502		
RSquare	0.01813114		
RSquare Adj	-0.0520024		
MAPE	37.5147229		
MAE	15.1585634		
-2LogLikelihood	138.657325		

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	-0.13327	0.247770	-0.54	0.5991	60.5747651
Intercept	0	53.45149	4.094173	13.06	<.0001*	

Forecast



Remember to select 1 in Autoregressive order!

To measure a *long-term trend* we use **regression analysis,** where X is ***Time.***
$y = \beta_0 + \beta_1 t + \varepsilon$

Seasonal variation occurs at specific time periods. We use **Seasonal indexes** to estimate the degree to which the seasons differ from one another.

Steps to compute Seasonal indexes:

1. Compute simple regression line $\hat{y}_t = b_0 + b_1 t$
2. For each time period, compute: $\frac{y_t}{\hat{y}_t}$
3. For each type of season, compute: *the average of ratios from step 2.*
4. Adjust the *averages from step 3*. so the average of all seasons is 1.

Seasonal indexes are used to remove the seasonal variation- ***Deseasonalizing:***

$$Seasonally\ Adjusted\ Time\ Series = \frac{Actual\ Time\ Series}{Seasonal\ Index}$$

Seasonal Indexes In JMP:

- Make sure the variables are **continuous**
- **Fit Model:** *Y*- what we test, *Add*- Period (*new column with consecutive values 1,2,3 etc.)*
- ∆**Save Predicted Formula:** new column appears.
- Create new column, **Formula:** $\frac{y_t}{\hat{y}_t}$ (original Y divided by predicted formula column). This is the **Seasonal index.**
- *Can use Graph Builder to plot the Seasonal Index over Period to see if there is a real trend (positive/negative).*

*To select the model with the ***greatest forecast accuracy,*** 2 methods can be used:

1. **Mean Absolute Deviation (MAD):** $MAD = \frac{\sum_{i=1}^{n} |y_t - F_t|}{n}$
2. **Sum of Squares for Forecast Error:** $SSE = \sum_{i=1}^{n} (y_t - F_t)^2$. *Use SSE if we want to avoid large errors.*

   Where n is the no. of time periods, y_t is the actual value of time series and F_t is the forecasted value.

Forecasting with Exponential Smoothing	Forecasting with Seasonal Indexes
When the time series displays *gradual or no trend* and there is *no seasonal variation*.	When the time series has *seasonal variation* and has a *long-term trend.*
Forecast for the period t+k (k=1,2,3..): $F_{t+k}=S_t$, where S_t is the exponentially smoothed value.	Forecast for the period t (*regression equation)*: $F_t=[b_0+b_1t]xSI_t$
The more into the future we get, the less accurate the predictions.	

Durbin Watson test

↳ **test for dependence in errors terms for Time Series when there is a natural ordering of the observations.**
↳ tests first-order autocorrelation- relationship that exists between consecutive residuals: e_{i-1} and e_i, where i is the time period.

Time Series- regression analysis
First comment on each X included in the model- why is it relevant? **Assumptions:** 1. Random sample and reliable data. (*Can we generalise the results?)* *Errors:* 2. Normality of the error term ε. (Save Residuals- **Distribution** residuals. *When the sample size is large, the assumption can be dropped based on Central Limit Theorem).* 3. The expected value ε is zero for the independent variable (Xs): $E(\varepsilon \mid X) = 0$ ↳ **Residual by predicted plot: look for U-shape in the residual plot**=> not fulfilled.

4. Homoscedasticity (*constant variance)*- the variation around the regression line should be similar for all values of the independent variable (X): $Var(\varepsilon \mid X) = \sigma^2$ **(funnel shape)**
5. The error terms ε are independent of each other. (**patterns in the errors)**
6. Validity and trustworthiness. (*comment yourself)*
7. Multicollinearity (**Multivariate Methods- Multivariate** all Xs; **CI of Correlation** check if there is 0 in the intervals) *normal correlation between x and* x^2

Model:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \varepsilon$

Hypotheses regression:

Tests for linear relationship:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \cdots = 0$

$H_1: At\ least\ one\ \beta_i \neq 0$

**If the null is true, no X is linearly related to y.*

Hypotheses Durbin Watson Test:

H_0: There is no first-order autocorrelation.

H_1: There is positive first-order autocorrelation. (small p-value, for JMP output)

* H_1: *There is either positive or negative first-order autocorrelation. (in general)*

Significance level:

α= 0.05

Test-statistic regression:

$$F_{crit} = \frac{MSR}{MSE} \sim F_{\alpha,k,n-k-1}$$

Test-statistic Durbin Watson Test:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

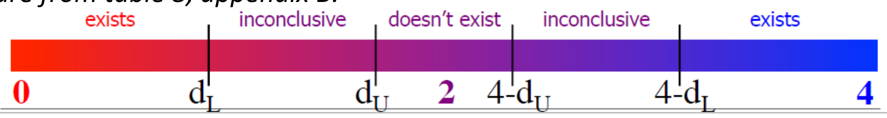
Where e_{i-1} and e_i are consecutive residuals, i is time period.

$0 \leq d \leq 4$

Interpret:

	Positive first order autocorrelation	**Negative first order autocorrelation**
There is enough evidence to show that first-order autocorrelation exists.	If $d<d_L$	If $d>4-d_L$
There is NOT enough evidence to show that first-order autocorrelation exists.	If $d>d_U$	$d<4-d_U$
Test is inconclusive.	If $d_L<d<d_U$	If $4-d_U<d<4-d_L$

**d_L and d_U are from table 8, appendix B.*



In JMP (first model, afterwards assumptions):

- Check if all variables are **continuous.**
- **Fit Model:** *Y- dependent, Add- independent (also Time).* **∆Model Dialog:** Remove the insignificant variables, one by one, starting with the biggest p-value.

You can group* **By *if it makes sense, before running the model.*

- **∆Row Diagnostics- Durbin Watson test-> ∆Significance p-value** *if very small, there is positive first-order autocorrelation. (if there is, use* ***autoregressive model*** *– see below)*

- △**Estimates -> Show Prediction Expression:** the intercept may not have sense to interpret- how much is Y when all Xs are zero; for each variable, increasing one X with 1 units, holding the rest constant, y increases/decreases.

*For *dummy variable* (e.g. gender): e.g. being a male increases/decreases y with ...

- △**Save columns-> Save Prediction Expression:** new column appears which can be used to forecast- *new row, insert given values.*

Assumptions:

- **Graph Builder: Y** and **X- time** to check for positive/negative trend in the residuals. Remember to drag to max. **Lambda Variables** in the down left part, to make the line smooth.
- △**Save columns** -> **Residuals.** Check for *Normality of the errors*- **Distribution- Residual.** Under **Fitter Normal** (*in right side)*-> △**Goodness of Fit:** if p-value> 0.05, the errors are normal.

H_0: the errors are normally distributed

H_1: the errors are not normally distributed

- **Distribution- residuals.**
- Save the **Residual by Predicted Plot** from the bottom of the model analysis. Check for *Zero conditional mean (no U shape), Homoscedasticity (no funnel shape), the error terms are independent (no patterns).*
- **Multivariate Methods- Multivariate** all Xs; **CI of Correlation** check if there is 0 in the intervals); *normal correlation between x and x^2.*

All variables need to be continuous.

In JMP:

Summary of Fit

RSquare	0.12003
RSquare Adj	0.016504
Root Mean Square Error	1711.676
Mean of Response	9315.3
Observations (or Sum Wgts)	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	6793798	3396899	1.1594
Error	17	49807214	2929836	Prob > F
C. Total	19	56601012		0.3373

Parameter Estimates

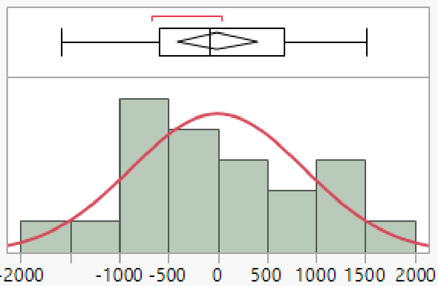
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8308.0114	903.7285	9.19	<.0001*
Snowfall	74.593249	51.57483	1.45	0.1663
Temperature	-8.753738	19.70436	-0.44	0.6625

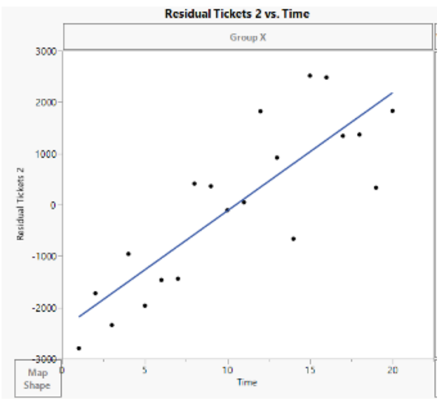
Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
0.5931403	20	0.5914	0.0002*

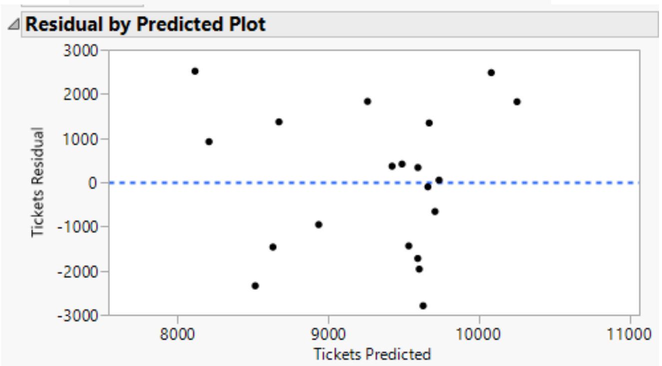
Distributions

Residual Tickets





Residual by Predicted Plot



Interpret:

Adjusted RSquare is 0.01, meaning that only 1% of the variation in the no. of tickets sold is explained by the explanatory variables, temperature and snowfall. This means that the model is very bad, and further investigation is needed to identify what affects the sales.

Root Mean Square Error is 1711, while the **Mean of Response** is 9315. So SSE is much smaller, meaning that the model fits the data quite well, not having much room for error.

P-value: the model is insignificant, with a p-value of 0.33. So there is not enough evidence to reject the null hypothesis.

Durbin Watson test: the p-value is very small. 0.0002, meaning that the null hypothesis is rejected. This means that there is positive first-order autocorrelation.

Distribution: the residuals of the dependent variable very roughly follow the bell shape, significant differences in the number of observations being recorded.

Residuals: from the *Residual by predicted plot* is can be seen a pattern in the errors- going up and down. This has also been deducted from the Durbin Watson test, that positive autocorrelation exists.

Time series: plotting the residuals over time, a positive trend is highlighted.

➔ **Time variable is added to improve the model:**

Summary of Fit

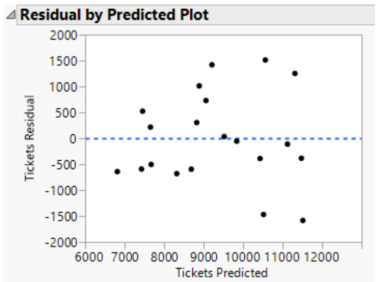
RSquare	0.74098
RSquare Adj	0.692414
Root Mean Square Error	957.2354
Mean of Response	9315.3
Observations (or Sum Wgts)	20

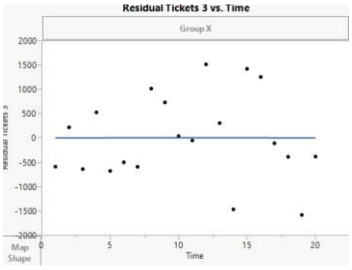
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	41940217	13980072	15.2571
Error	16	14660795	916299.68	Prob > F
C. Total	19	56601012		<.0001*

Parameter Estimates

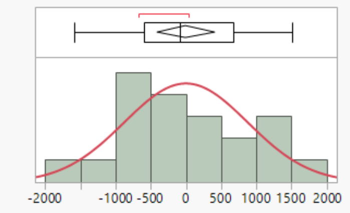
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5965.5876	631.2518	9.45	<.0001*
Snowfall	70.183059	28.85142	2.43	0.0271*
Temperature	-9.232802	11.01971	-0.84	0.4145
Time	229.96997	37.13209	6.19	<.0001*





Distributions

Residual Tickets 3



Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.8849875	20	0.0405	0.3512

<u>Interpret:</u>

P-value: the model is significant, with a p-value smaller than 0.0001. So there is enough evidence to reject the null hypothesis. Looking at the *Parameter estimates,* all variables are significant besides Temperature. So this will be removed.

Summary of Fit

RSquare	0.729616
RSquare Adj	0.697806
Root Mean Square Error	948.808
Mean of Response	9315.3
Observations (or Sum Wgts)	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	41296990	20648495	22.9367
Error	17	15304022	900236.59	Prob > F
C. Total	19	56601012		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5842.0569	608.3882	9.60	<.0001*
Snowfall	70.723427	28.59027	2.47	0.0242*
Time	229.75159	36.80427	6.24	<.0001*

The model is still significant, as well as all the remaining variables. So the model is valid.

Adjusted RSquare is now 0.70, meaning that 70% of the variation in the no. of tickets sold is explained by the explanatory variables, time and snowfall. The model has greatly improved.

Root Mean Square Error is 949, while the **Mean of Response** is 9315. So SSE is much smaller, meaning that the model fits the data quite well, not having much room for error.

F-test:
F_{obs}= 22.94
F_{crit}=$F_{\alpha,k,n-k-1} = F_{0.05;2;19}$= 3.52
F-test: F_{obs}> F_{crit}, so the null hypothesis is rejected, meaning that there is a linear relationship between the tickets sold and the two independent variables.

Durbin Watson test: the p-value is 0.35, so there is not enough evidence to reject the null hypothesis. This means that there is no positive first-order autocorrelation.

Residuals: the *Residual by predicted plot* does not show any specific pattern or shape in the errors.

Time series: plotting again the residuals over time, no positive nor negative trend is seen.

Multicollinearity:

Multivariate

Correlations

	Snowfall	Temperature	Time
Snowfall	1.0000	-0.0222	0.0245
Temperature	-0.0222	1.0000	0.0065
Time	0.0245	0.0065	1.0000

CI of Correlation

Variable	by Variable	Correlation	Lower 95%	Upper 95%
Temperature	Snowfall	-0.0222	-0.4602	0.4245
Time	Snowfall	0.0245	-0.4226	0.4620
Time	Temperature	0.0065	-0.4373	0.4477

Rule of thumb: if the value is bigger than ∓0.2, it can be said with 95% confidence that there is a correlation between the 2 variables. This can be also checked by looking at the confidence intervals: if there is 0 in the confidence interval, we do NOT have correlation.
So the multicollinearity assumption is fulfilled.

Prediction Expression

5842.05688875721 + 70.7234270312634 * Snowfall + 229.75159102608 * Time

Looking at the model: If there is no snow at time 1, the number of tickets sold would be (5842+230)= 6072. If the snowfall increases by 1 cm, there will be an increase in the no. of tickets sold of 70.
Snowfall: for each additional snow cm, the tickets sold increase by 70, assuming that the other explanatory variable in the model is held constant.
Time: for each additional time unit, the no. of tickets sold increase by 229, assuming that the other explanatory variable in the model is held constant.

Autoregressive Model

↳ when there is no obvious trend or seasonality, but we believe that there is a *correlation between consecutive residuals (from **Durbin-Watson test**)*.

Autoregressive forecasting model:
$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon$
Estimated Autoregressive forecasting model:
$\widehat{y_t} = b_0 + b_1 y_{t-1}$

In JMP:
-**New column, formula:** *Lag(Y,1)*
-**Fit Y by X,** where X is the Lag. **∆Fit line**
-**New row-** to forecast for the next period, put the estimated Lag value in the model:
Gas sales=61.48-0.13*66=52.9
-**Exclude** the last period: 16 and run the model again.
**Only under Fit Model we can run Durbin Watson test.*

Linear Fit

Linear Fit

Gas Sales = 61.483786 - 0.1346727*Gas_Lag

Summary of Fit

RSquare	0.018321
RSquare Adj	-0.05719
Root Mean Square Error	20.02745
Mean of Response	54.4
Observations (or Sum Wgts)	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	97.3145	97.315	0.2426
Error	13	5214.2855	401.099	Prob > F
C. Total	14	5311.6000		0.6305

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	61.483786	15.28286	4.02	0.0014*
Gas_Lag	-0.134673	0.273411	-0.49	0.6305

Linear Fit

Gas Sales = 68.39339 - 0.2915044*Gas_Lag

Summary of Fit

RSquare	0.067548
RSquare Adj	-0.00418
Root Mean Square Error	19.25292
Mean of Response	53.52667
Observations (or Sum Wgts)	15

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	349.0765	349.076	0.9417
Error	13	4818.7728	370.675	Prob > F
C. Total	14	5167.8493		0.3495

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	68.39339	16.10609	4.25	0.0010*
Gas_Lag	-0.291504	0.300387	-0.97	0.3495

Dictionary

Normal distribution= gaussian distribution
Mean =average; For population: μ, for sample: $\bar{x}$
Population= treatment
Standard deviation (σ)
Sample size (n)
x_{ij}- the i[th] observation is the j[th] sample.
$\bar{\bar{x}}$- grand mean (=the mean of all observations from all the populations)
Variance: for population: σ^2, for sample: s^2
Factor= population classification criteria (*e.g. age)*
Factor level= level under the classification criteria (*e.g. young, middle-aged, senior*)
SST= sum of squares for treatments/populations
SSE= sum of squares for error; measures the *amount of variation* in all groups. Measures how well the regression model fits the data.
MST= mean square for treatments
MSE= mean square for errors
SSB= sum of squares for blocks; measures the amount of variation between *blocks.*
Block= *matched* group of observations from each population.

Confidence interval estimator of $(\mu_1 - \mu_2)$: $(\overline{x_1} - \overline{x_2}) \mp t\alpha_{/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

↳If the interval excludes 0, the population means differ.

β_0, β_1- are population parameters