Overview

Stephanie Ranft S2459825

February 26, 2020

Statistics 2
PSBE2-07

Variance

Pooled variance

Consider a test between I independent samples. Whilst we cannot assume that they are all from the same population (and hence have the same variance), the Central Limit Theorem allows us to conclude that the pooled variance converges to the true variance. To see how this works, and in order to better understand when and where to use pooled variance:

$$s_p^2 = \frac{\sum_{i=1}^{I}(n_i-1)\,s_i^2}{\sum_{i=1}^{I}(n_i-1)}, \qquad \text{where } I \text{ is the total number of groups.} \tag{1}$$

$$= \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2 + \cdots + (n_I-1)\,s_I^2}{\underbrace{(n_1-1)+(n_2-1)+\cdots+(n_I-1)}_{n-I}} \tag{2}$$

If we were to assume that one of the samples was a lot bigger than the others, suppose sample k ($n_k >> n_i$ for every sample), then we can assume that the pooled variance will converge to s_k^2:

$$= \frac{\frac{n_1-1}{n_k-1}s_1^2 + \cdots + s_k^2 + \cdots + \frac{n_I-1}{n_k-1}s_I^2}{\frac{n_1-1}{n_k-1} + \cdots + \frac{n_k-1}{n_k-1} + \cdots + \frac{n_I-1}{n_k-1}} \tag{3}$$

$$\sim s_k^2. \tag{4}$$

The reason for this has something to do with the "power" of having a large n; that the sample is reliably similar to the population (lower standard error). This is an important point in sample collection, because if all samples except one have a really small size (<30) but one sample is substantially larger (>100), then the statistician can more readily believe the results of the largest sample (due to CLT - recall formula for SE). If we have that all of the samples are nearly the same size (choose $n_k \approx n_1 \approx \cdots \approx n_I$), then there is no dominating variance and we can assume the following:

$$s_p^2 \approx \frac{(n_k-1)\,s_1^2 + (n_k-1)\,s_2^2 + \cdots + (n_k-1)\,s_I^2}{(n_k-1)+(n_k-1)+\cdots+(n_k-1)} \tag{5}$$

$$= \frac{\cancel{(n_k-1)} \times \left(s_1^2 + s_2^2 + \cdots + s_I^2\right)}{I \times \cancel{(n_k-1)}} \tag{6}$$

$$= \frac{\sum_{i=1}^{I} s_i^2}{I} \tag{7}$$

$$= \bar{s}^2, \qquad \text{which is the } \textbf{mean of the variances}. \tag{8}$$

This indicates that the pooled variance is the weighted average of variances, where the highest weight is given to the largest sample size. This is to ensure that our pooled variance is the best estimator of the population variance.

Another way to look at this is to look at the fact that the sample variance for group i is $s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (n_i - 1)$, where j indexes the persons in a group.

$$\implies s_p^2 = \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2 + \cdots + \sum_{j=1}^{n_I} (y_{Ij} - \bar{y}_I)^2}{n - I} \tag{9}$$

$$= \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - I} = \frac{\text{SSE}}{\text{df}_E} = MSE. \tag{10}$$

It is important to know when exactly to use pooled v.s. unpooled variance, but if we can **assume that the variances of the populations are equal**, i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$, then we use pooled variance.
For $I = 2$ (comparing two means), we can use the t statistic to determine the results of our test:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\left(\underbrace{n_1 + n_2}_{n} - 2\right) \tag{11}$$

$$\implies t^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_p^2 \times \underbrace{\frac{n_1 + n_2}{n_1 n_2}}_{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{y}_1 - \bar{y}_2)^2 / (n_1 + n_2)}{s_p^2 / n_1 n_2} \tag{12}$$

If $n_1 \approx n_2$, then multiplying the t-statistic by itself gives us an F-statistic.

$$\implies t^2 \approx \frac{\frac{n}{2} (\bar{y}_1 - \bar{y}_2)^2}{s_p^2} \sim F(1, n_1 + n_2 - 2). \tag{13}$$

For $I > 2$ (comparing multiple means), we use the F statistic:

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\text{SSG}/\text{df}_G}{\text{SSE}/\text{df}_E} \tag{14}$$

$$= \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 / (I - 1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / \underbrace{(n - I)}_{\substack{(n-1)-(I-1) \\ = \text{df}_T - \text{df}_G}}} \tag{15}$$

$$= \frac{\sum_{i=1}^{I} n_i \times (\bar{y}_i - \bar{y})^2 / (I - 1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (n - I)} \sim F\left(\underbrace{I - 1}_{\text{df}_G}, \underbrace{n - I}_{\text{df}_T - \text{df}_G}\right) \tag{16}$$

More on this when we discuss ANOVA...

Unpooled variance

If we cannot see from the data that the independent samples are drawn from the same population, or that the I variances are similar, then we use unpooled variance. In general, we do not consider using this for $I > 2$ (for comparing more than two means). It is mathematically possible, but in practice it is seldom used and certainly not a part of the scope of this course. The main reason being that the calculations required to determine the degrees of freedom is quite complicated (see (18)), and most psychologists use a computer. So, the **only time you use unpooled variance is when you have observable differences between the two groups**. So, you could note that $s_1^2 >> s_2^2$ or perhaps the n_i of each group is low (<30) and unequal; it is something you need to determine for yourself. For example, if your test was about two machines from different manufacturers and whether they can complete the same task in the same amount of time, then you would use unpooled variance. If you were exploring behavioural data two culturally different countries, you would use unpooled variances.

Rule of thumb: if it's only two groups and you can assume/predict that the variances are unequal, use unpooled.

$$s_{up} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{17}$$

$$\implies t = \frac{\bar{y}_1 - \bar{y}_2}{s_{up}} \sim t(k), \qquad \text{where } k = \frac{(n_1 - 1) \times (n_2 - 1)}{(n_2 - 1) \times C^2 + (1 - C)^2 \times (n_1 - 1)} \quad \text{and } C = \frac{s_1^2/n_1}{s_{up}^2}. \tag{18}$$

Paired data

I don't think I need to refer much to this, however if you are testing whether there is any effect before and then after, consider this to be paired data. In that case, you don't use either pooled or unpooled variance, but look instead at the variance of differences. That is, transform your data from x and y to $d = x - y$, for instance.

ANOVA

One-way

In Table 1 is an incomplete tabulated output for a test comparing the means between groups. Can you fill in the missing values?

	SS	df	MS	F	sig.
G	91.467		45.733		0.021
E	276.400	27			
T	367.867				

Table 1: ANOVA one-way table

Some background on the test: a teacher wants to know if the starting level of her pupils affects the mean length of time to complete the exam. Formulate the null and alternative hypotheses.

$$H_0 : \tag{19}$$

$$H_a : \tag{20}$$

Summarise your findings of this test: In Table 1, can have that the sum of squares between the groups (SSG) is 91.467 and that the mean square between groups (MSG) is 45.733. You want to find out how these two are related, and notice that $45 \times 2 = 90$, which is indeed the degrees of freedom for groups ($\text{df}_G = 2$). Now, you can conclude that you have 3 groups in total ($I = 3$).
Moving on the row marked 'E': there is an evident relationship between the mean squared error within each group (MSE) of 276.400 and the degrees of freedom for the error (df_E) of 27. That is, $\text{df}_E \times 10 \approx \text{MSE}$; using the calculator, you find that MSE is equal to 10.237. Given that df_E ($= n - I$; $I = 3$) is 27, we know that each group has 30 participants ($n = 30$).
Now that we know n, and hence df_T is 29, we can calculate the variance of our data (MST) as 12.685.
In order to calculate our F statistic, we need to under exactly what it is: F is the ratio of variation between and within groups. There are three distinct cases which we will look at now.

$0 \leq \mathbf{F} < 1$

In this case (refer to Figure 1), we know that the variance in the data which can be explained by the differences between the groups, is less than the variance within the groups themselves. Either, there is not much difference between the groups, or there is a lot of variation within the groups. In both cases, you would need to perform post-hoc tests, such as contrasts, to confirm or deny your findings. Evidently if $F = 0$, then there is no variation between the groups, i.e. $\bar{y}_1 = \bar{y}_2 = \cdots = \bar{y}_I = \bar{y}$.

$$F = \frac{\text{MSG}}{\text{MSE}} < 1 \tag{21}$$

$$\implies \frac{\text{SSG}}{\text{df}_G} < \frac{\text{SSE}}{\text{df}_E} \tag{22}$$

$$\implies \text{variance between} < \text{variance within} \tag{23}$$

Figure 1

F ≈ 1

We can note the implications of an F near to one as 'good'; we conclude that the variance in the data is not solely due to variance between groups but equally within. We call this as 'good' because of the complications that arise when we find significant results (more on this in a moment). This F tells us that the variations between the groups is proportional to the variance within the groups themselves, so the effects of being in any particular group are not evident in the data. A very simple example of this is test scores between schools: you might want to test whether attending a more prestigious has any outcome on the results of the students themselves. So, you would have I equal to the number of schools (ranked in order of prestige) and J equal to the number of students at each school (in a particular graduating year). If you received an F value close to 1, then you would conclude that there is no significant advantage benefited to students who attend a prestigious school, in terms of grades. Additionally, note the following relationship:

$$\left.\begin{aligned} F &= \frac{\operatorname{Var} y - s_p^2}{s_p^2} = \frac{\operatorname{Var} y}{s_p^2} - 1 \\ F &\approx 1 \end{aligned}\right\} \implies \frac{\operatorname{Var} y}{s_p^2} \approx 2. \tag{24}$$

This says that, if the total variation in y is twice the size of the collective variation within each group, then any observed variation between the groups in our sample is acceptable (accept null hypothesis).

F > 1

In this case, we can conclude that we have a significant result: there is a great difference between the groups. With respect to the previous example, we would conclude that the prestige of a school has an effect on the grades of attendees; if F is **much** larger than 1, we could say that the effect is **profound** or immense. Within each school, there is not much variation in the data compared with the variation between the schools (e.g. $I = 3$): Figure 2 is a pictorial example of $F > 1$, which would lead us to assume that further tests (e.g. contrasts) need to be performed in order to determine which school benefits the greatest advantages to it's attendees.

So, back to our table! We can discern that we will have an $F > 1$, as $10 \times 4.5 = 45$, and indeed we have $F = 4.467$ (see the completed Table 2). Relating this to the significance of 0.021 found by the software, the probability of achieving an F more extreme than $F^*(2, 27 \,|\, \alpha = 0.05) = 3.354$ in any repetition is 0.021, which is less than our $\alpha = 0.05$.

The usual method of formulating the null and alternate hypotheses (as you may have figured, by now), is

$$H_0 : \mu_1 = \mu_2 = \mu_3, \text{ i.e. there is no difference between the three groups;} \tag{25}$$

$$H_a : \text{there exists a difference somewhere between the groups.} \tag{26}$$

Using the F statistic and p-value, we reject the null hypothesis at a 5% significance level in favour of the alternate hypothesis, and conclude that at least one of the groups is different from the others. In order to determine which group differs the most, we might formulate some contrasts based on our plot of the data, similar to the one in Figure 2. For example if one of the groups, say group three, was distinctly further away from groups one and two, we might decide to find an a, such that $0 \leq a \leq 1$ and test:

$$H_0 : \frac{a\,(\mu_1 + \mu_2)}{2} = (1 - a)\mu_3 \tag{27}$$

If we find that an a near to zero provides us with a insignificant results (i.e. not reject H_0), then we know that group three dominates. However, if an a near to one has this provision, then we know that a groups one and two equally dominate. Here, dominate means that they greatly contribute to the difference between groups (i.e. large comparable mean/s).

The background on the data set tells us that this is indeed about schooling, but only one particular test (limited factors $\implies$ ANOVA I, and not ANOVA II). We may conclude that the findings of the test imply that the starting level of the pupils does indeed affect the mean length of time to complete the exam. Without knowing any more information, such as individual group means $\bar{y}_i$, standard deviations sd_i and number per group n_i, we conclude our testing here. If this information was available to us, we could run some contrasts to find which particular starting level, beginner, intermediate or advanced, provided the biggest advantage on this particular exam.

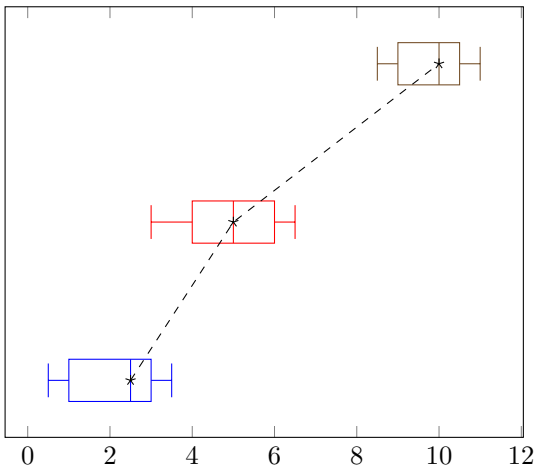


Figure 2: The box plots display the variance **within** groups, which is noticeably small, and the dashed line displays the variance **between** groups, which is noticeably large.

	SS	df	MS	F	sig.
G	91.467	2	45.733	4.467	0.021
E	276.400	27	10.237		
T	367.867	29	12.685		

Table 2: ANOVA one-way table (completed)

Two-way

Fun time!!!
So, not only do you have two or more groups (I), but we now consider that there might be multiple effects (A, B, $C, \dots$) and their interactions ($A \times B$, $A \times B \times C, \dots$). We will start with a simple example, and progress from there:
Perhaps a farmer wants to investigate the effects of manure (A) and nitrogen-based fertiliser (B), and also the combination of both ($A \times B$), on the yield of their corn. The output of an ANOVA II test is given below:

	SS	df	MS	F	sig.
A	15.842		15.842		0.029
B	17.298		17.298		0.035
A × B	3.872		3.872		0.273
E	48.000		3.000		
T	85.012	19	4.474		

Table 3: ANOVA two-way table

Can you fill in the table and draw some conclusions?

Summarise your findings: There is a pretty obvious relationship between the columns "SS" and "MS", so we can assume that the degrees of freedom for the factors and the interaction effect are 1. Meaning: that the factors A and B have two categories each, such as high and low levels (of manure and fertiliser). So, then we know that there are in total $2 \times 2 = 4$ groups: high levels of both, high levels of manure v.s. low levels of fertiliser, low levels of manure v.s. high levels of fertiliser, and low levels of both. Further, we can examine the calculation of df_E:

$$\mathrm{df}_E = \underbrace{(n-1) - (I-1) - (J-1) - \overbrace{(I-1)(J-1)}^{IJ-I-J+1}}_{\mathrm{df}_T - (\mathrm{df}_A + \mathrm{df}_B + \mathrm{df}_{A\times B})} = n - IJ. \tag{28}$$

We already have $I = 2 = J$, and $n = 19 + 1$, so $\mathrm{df}_E = 20 - 2 \times 2 = 16$. We remember how to calculate the F statistic for each factor and interaction:

$$F = \frac{\text{variance between}}{\text{variance within}} = \begin{cases} \dfrac{\mathrm{MSA}}{\mathrm{MSE}}, & \text{when testing if } A \text{ has a significant effect;} \\ \dfrac{\mathrm{MSB}}{\mathrm{MSE}}, & \text{when testing if } B \text{ has a significant effect;} \\ \dfrac{\mathrm{MSAB}}{\mathrm{MSE}}, & \text{when testing if an interaction of } A \text{ and } B \text{ has a significant effect.} \end{cases} \tag{29}$$

When we are talking about 'significant effect', we want to know if the variation between the groups is relatively equal to the variation within the groups. So, in our table we can input $F_A = 15.842/3 = 5.281$, $F_B = 17.298/3 = 5.77$ and $F_{A\times B} = 3.872/3 = 1.291$. So, given that our F statistic for factors A and B are well above 1, we can can conclude that the levels (each) of manure and fertiliser has a profound affect on corn yield. Further, this is evident by the p-values both being under our accepted 5% level. Now it becomes a little harder to discern the right answer concerning the existence on an interaction effect: we have that the F value is above 1.2 and perhaps would think of rejecting H_0, however we must consider the p-value! Repetitions of this study would yield more extreme F values for this interaction effect more than a quarter of time ($\mathbb{P}\left(F > f_{(1,16\,|\,0.05)}\right) = 0.273 > 0.25$), so we may safely conclude that there is no interaction effect. You can see what (graphically) denotes an effect in the slides from lecture 5 (slide 26-31).
If you were to advise the farmer, what advice would you give?

Previously we found that factors A and B both had a main effect, but there was no significant interaction effect. Now we are interested in conducting a basic two-way ANOVA without the interaction effect. In order to do this, we include the data from the interaction effect into the error terms:

$$\mathrm{SSE}_{\text{new}} = \mathrm{SSE} + \mathrm{SSAB} \qquad \mathrm{df}_{\mathrm{E,\ new}} = \mathrm{df}_{\mathrm{E}} + \mathrm{df}_{\mathrm{A\ \times\ B}} \tag{30}$$

$$\implies \mathrm{MSE}_{\text{new}} = \frac{\mathrm{SSE}_{\text{new}}}{\mathrm{df}_{\mathrm{E,\ new}}} \tag{31}$$

Something important to remember for the exam: the "spread of means" is not equal to the "mean spread of the data". The former refers to the variance between means (MSG) and the latter, the mean variance (MSE). The professor will use language like this in order to confuse you!!

Contrasts

Why do we perform contrasts? Simply put, it is to reduce the overall statistical error: if α% for each test and you have I groups then you will need to perform $I \times (I-1)/2$ tests, meaning that

$$\begin{aligned} \text{overall error rate} &= \mathbb{P}\left(\text{at least one false rejection}\right) && (32) \\ &= \mathbb{P}\left(\text{at least one Type I error} \,\middle|\, \frac{I\times(I-1)}{2} \text{ tests}\right) && (33) \\ &= 1 - \mathbb{P}\left(\text{no Type I errors} \,\middle|\, \frac{I\times(I-1)}{2} \text{ tests}\right) && (34) \\ &\approx 1 - (1-\alpha)^{I\times(I-1)/2} && (35) \end{aligned}$$

So, for a simple ANOVA I with 5 groups at 1% significance level, this means we need to perform $5\times4/2 = 10$ tests resulting in an error rate of $1 - 0.99^{10} = 0.0956$, i.e. 9.6% of a Type I error, which is higher than our accepted 5% level (chance capitalisation). It is possible to plan for this by introducing contrasts **prior to undertaking the test**.
Assume, again, $I = 5$ and we want to know which group accounts for the largest deviation of the data. **Remember**, $SST = SSG + SSE$ - this means that the total variance in the data set is either due to the variance between groups or within the groups themselves. If we have an $F > 1$ (and $p < \alpha$ given the sample size of each group is "large enough"), we know that it is attributable to variance between groups and now we want to know which particular group/s are different.

We have that,

$$\bar{x}_1 > \bar{x}_2 \qquad \text{Is group 1 different to group 2?} \qquad \begin{matrix} H_{01} : \mu_1 - \mu_2 = 0 \\ H_{a1} : \mu_1 - \mu_2 > 0 \end{matrix} \tag{36}$$

$$\text{coefficients: } (1, -1, 0, 0, 0) \tag{37}$$

$$\begin{matrix} \bar{x}_1 < \bar{x}_3 \\ \bar{x}_2 < \bar{x}_3 \end{matrix} \qquad \text{Are groups 1 and 2 different to group 3?} \qquad \begin{matrix} H_{02} : \frac{\mu_1+\mu_2}{2} - \mu_3 = 0 \\ H_{a2} : \frac{\mu_1+\mu_2}{2} - \mu_3 < 0 \end{matrix} \tag{38}$$

$$\text{coefficients: } \left(\frac{1}{2}, \frac{1}{2}, -1, 0, 0\right) \tag{39}$$

$$\bar{x}_5 > \max\{\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4\} \qquad \text{Is group 5 the most different?} \qquad \begin{matrix} H_{03} : \frac{\mu_1+\mu_2+\mu_3+\mu_4}{4} - \mu_5 = 0 \\ H_{a3} : \frac{\mu_1+\mu_2+\mu_3+\mu_4}{4} - \mu_5 < 0 \end{matrix} \tag{40}$$

$$\text{coefficients: } \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, -1\right) \tag{41}$$

We don't know the true values of $\mu_1, \ldots, \mu_5$, so we estimate with $\bar{x}_1, \ldots, \bar{x}_5$ to produce an estimated contrast value c and it's associated t statistic: $c = \sum_{i=1}^{I} a_i \bar{y}_i$ and $t = c/\text{SE}_c \sim t\,(n - I)$ (under H_0 the population contrast statistic is assumed to be zero). Let's look at an example:

Group	$\bar{y}_i$	s_i	n_i
1	33.31	3.63	9
2	28.72	2.91	15
3	32.46	3.98	9
Total	30.99	3.3	33

Table 4: My caption

Notice that group 2 has the lowest mean and sd, and the largest sample size? This means that we can infer already that group 2 is the most different. First, we should calculate the pooled standard deviation:

$$s_p = \sqrt{\frac{\sum_{i=1}^{3}(n_i - 1)s_i^2}{\sum_{i=1}^{3}(n_i - 1)}} = \sqrt{\frac{8 \times 3.63^2 + 14 \times 2.91^2 + 8 \times 3.98^2}{33 - 3}} = \sqrt{\frac{350.69}{30}} = \sqrt{11.69} = 3.42. \tag{42}$$

So we test,

$$H_{01} : \mu_2 = \mu_1 \tag{43}$$

$$H_{a1} : \mu_2 < \mu_1 \tag{44}$$

$$\implies c = -1 \times \bar{y}_1 + 1 \times \bar{y}_2 + 0 \times \bar{y}_3 = -4.59. \tag{45}$$

$$\implies \text{SE}_c = s_p \sqrt{\sum_{i=1}^{3} \frac{a_i^2}{n_i}} = 3.42 \times \sqrt{\frac{1}{9} + \frac{1}{15}} = 3.42 \times \sqrt{0.178} = 3.42 \times 0.422 = 1.44. \tag{46}$$

$$\implies t = \frac{c}{\text{SE}_c} = \frac{-4.59}{1.44} = -3.1875 \sim t\,(33 - 3)_{\alpha/2=0.025}\,. \tag{47}$$

Our critical $t^\star$ value for 32 df and (two-tailed) 5% significance is -2.042, which is larger than our t-statistic. So we must conclude to reject H_0 in favour of H_a; you can read the p-value from a table as around 0.003. If we were to construct a confidence interval about c:

$$\text{CI} = (c - t^\star \times \text{SE}_c, c + t^\star \times \text{SE}_c) \tag{48}$$

$$= (-3.1875 - 2.042 \times 1.44, -3.1875 + 2.042 \times 1.44) = (-6.13, -0.25)\,. \tag{49}$$

Note that the entire confidence interval is located to the left of zero? So, we certainly reject the null hypothesis, as even when it is assumed, we do not have zero contained in the interval!

Confidence intervals

The general equation for confidence intervals involving multiple comparisons is given by

$$\mathrm{CI}_{ij} = (\bar{y}_i - \bar{y}_j) \pm t^{**} \times s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{50}$$

For this point of the course, you have two options for your critical t^{**}:

1. **Bonferroni:** adjust the significance of each test to ensure that overall error rate is less than the specified α:

   $$k = \text{the number of tests} = \frac{I \times (I-1)}{2}, \qquad \text{where } I \text{ is the number of groups.} \tag{51}$$

   Let α^* be the significance of **each** of the k tests, then

   $$\alpha^* = \frac{\alpha}{k} \qquad \Longrightarrow t^{**} = t^*_{\substack{1-\alpha/2k \\ \nu=\mathrm{df}_E}}. \tag{52}$$

2. **Least significant differences (LSD):** we use this for $I = 3$ groups, otherwise use Bonferroni. This is because LSD method does not alter the significance, but only each individual test is improved, and $I = k$ for 3 groups.

   $$t^{**} = t^*_{\substack{1-\alpha/2 \\ \nu=\mathrm{df}_E}} \tag{53}$$

Important:

$$p_{\text{Bonferroni}} = \mathbb{P}\left(|T| > t_{1-\alpha/2k}\right) = \frac{\mathbb{P}\left(|T| > t_{1-\alpha/2}\right)}{k} = \frac{p_{\text{LSD}}}{k} \tag{54}$$

The p-values are scaled with respect to the number of tests performed.

Kruskal-Wallis procedure

ANOVA assumptions are normality, homoskedasticity and independence, and if they are severely violated then we turn to the Kruskal-Wallis procedure (non-parametric ANOVA). The null hypothesis is similar to ANOVA, however the violation of assumptions leads us to a new direction: "the **distribution** is the same in all groups". The alternative is that the scores in some groups are systematically larger.
Begin by ordering all n scores from lowest to highest, assigning rank 1 to the lowest. If some scores are equal, they receive the mean of the ranked score, e.g. there are two scores of 9 and it is the fifth lowest score (taking up spaces 5 and 6) then each of the '9"s receive a rank of $(5+6)/2 = 5.5$.
Now that each score has a rank, you need to separate the ranked scores back into their respective groups. Sum the rankings in each group to yield R_i. Furthermore,

1. if the sample sizes n_i are small (<5), then we use the following test statistic in ANOVA I:

   $$H = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{n(n+1)} \mathrm{SSG}_{\text{ranks}}. \tag{55}$$

2. if the sample sizes are not too small, say $n_i \geq 5$, then use the following approximation: $H \sim \chi^2(I-1)$.

For example, you are investigating the effects of exercise on depression. So you have three groups: no exercise, 20 minutes of jogging per day, and 60 minutes of jogging per day. In order to simplify, we assume that each participant is equivalently depressed, then at the end of the month you ask each participant how depressed they feel as a score out of one hundred, where 1 is totally miserable and 100 is ecstatically happy.
The method of testing is self-recorded and the scores (ordinal) given are non-parametric, so in order to draw conclusion we will need to use the Kruskal-Wallis procedure.

Table 5: This table contains the self-recorded ratings of 1 to 100 from 27 depressed people, where 1 is totally miserable and 100 is ecstatically happy. The ratings were ask for after performing daily jogging for a month, in groups of 20 and 60 minutes per day. The third group is the control group, who were asked to not exercise every day.

$i = 1, 2, 3$; $n_i = 8$	**no exercise**	**20 min/day**	**60 min/day**
	23	22	59
	26	27	66
	51	39	38
	49	29	49
	58	46	56
	37	48	60
	29	49	56
	44	65	62
$\bar{\mathbf{x}}_\mathbf{i}$	**39.63**	**40.63**	**55.75**
$\mathbf{s_i}$	**12.85**	**14.23**	**8.73**

Looking at Table 5, the minimum rating given was 22 and the maximum was 66, so 22 is given rank 1 and 66 is given the lowest rank. Tied scores get the average of the rankings they would've received. I've summarised the next few steps in the following ranked table:

Table 6: This table contains the rankings of the scores given, as well as their sums per group.

$\mathbf{i = 1, 2, 3}$; $\mathbf{n_i = 8}$	**no exercise**	**20 min/day**	**60 min/day**
	2	1	20
	3	4	24
	16	9	8
	14	5.5	14
	19	11	17.5
	7	12	21
	5.5	14	17.5
	10	23	22
$\bar{\mathbf{x}}_\mathbf{i}$	**39.63**	**40.63**	**55.75**
$\mathbf{s_i}$	**12.85**	**14.23**	**8.73**
$\mathbf{R_i}$	**76.5**	**79.5**	**144**

You can see in Table 5 that there is a rating of 49/100 in each column, which is the 14th, 15th and 16th lowest ranking. So they each get the ranking of $(14 + 15 + 16)/3 = 14$. As each $n_i = 8 > 5$, we assume that the H test statistic is approximated by the χ^2 distribution with 2 degrees of freedom. Calculating H:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} - 3(n+1) \tag{56}$$

$$= \frac{12}{24(24+1)} \sum_{i=1}^{3} \frac{R_i^2}{8} - 3(24+1) \tag{57}$$

$$= \frac{12}{24 \times 25} \left(\frac{76.5^2}{8} + \frac{79.5^2}{8} + \frac{144^2}{8} \right) - 3 \times 25 \tag{58}$$

$$\approx 7.271 \sim \chi^2(2) \tag{59}$$

Looking at the p-values for a $\chi^2(2)$, we can surmise that the probability of obtaining an H statistic more extreme than what we calculated is between 5% and 2.5%:

$$0.025 \leq \mathbb{P}\left(H > 7.271\right) \leq 0.05 \tag{60}$$

Given the small p-value for the H statistic, we can conclude that there is some difference between the three groups, i.e. the effects of exercise on depression is evident. Looking back at Table 6, you may notice that

the mean of group 3 is relatively larger than the others, and it's standard deviation is comparatively smaller. Meaning that we may infer that group three is the "different" group: 60 minutes of medium exercise (e.g. jogging) has the greatest impact on depression (in terms of improvement), compared to 20 minutes or no exercise (check this with the next section on effect size ☺).

Post-hoc methods

Post-hoc multiple comparisons: tests, using confidence intervals, for differences between **all** pairs of means:

$$\text{CI}_{ij} \text{ for comparing the means of groups } i \text{ and } j\text{: } (\bar{y}_i - \bar{y}_j) \pm t^{**}_{\text{df}_E} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{61}$$

Note the use of t^{**}? This refers you to the section on LSD and Bonferroni!
All this is about, is two main things: do your CI's overlap; is your assumed value inside the CI.

1. So, if we are comparing groups (comparing means), then we would like our assumed value $\mu_i - \mu_j = 0$ (H_0 : all groups are the same, i.e. all means are the same) to be inside the confidence interval. If it does not, we might need to look at our test for any errors (e.g. low effect size, small sample size, chance captilisation). However, if there are no errors present (e.g. large effect size, $n \geq 30$, Bonferroni method used, testing procedure is widely regarded as effective, etc.) then you would **reject** H_0. Why? You assume something, then you construct your test in such a way that there is no possible room for error, however the result shows that $(1-\alpha)\%$ of all samples lead to an interval that does not cover the unknown parameter. So our assumed parameter must be incorrect ($\mu_i - \mu_j \neq 0$). If our samples collected are large enough, the CLT allows us to conclude that true parameter lies somewhere closer to the estimated value $a = \bar{y}_i - \bar{y}_j$.

2. What does it mean if the CI's overlap? Referring back to Figure 2, you can see the boxplots represent variation within the groups, and that the red and blue boxplots overlap whilst the other box stands alone. Rather than adding another figure here, imagine that the boxplots are displaying 95% confidence intervals. So you have that the confidence intervals of groups one and two overlap, whilst group three's confidence interval is completely separated. This means that it is likely that the population means for groups one and two are equal, and the pop.mean of group three is likely greater. In relation to (61), if more than two CI's overlap sequentially (e.g. the upper bound of the CI for $\mu_1 - \mu_2$ is contained in the CI for comparing $\mu_2 - \mu_3$, and the upper bound for the CI comparing $\mu_2 - \mu_3$ is contained in the CI for comparing $\mu_3 - \mu_4$, ...), then you might assume that there it is likely that the groups are all (relatively) the same. How much they overlap is indicative of how likely it is that the groups are all the same. For example, if the upper bound of the CI for $\mu_1 - \mu_2$ is more than the estimate for comparing $\mu_2 - \mu_3$, namely the midpoint $\bar{y}_2 - \bar{y}_3$, then you might assume that these groups all differ by the same amount (more than relatively). If the CI's (almost) coexist, then you can be sure that these groups certainly differ by the same amount - if zero is included in all the CI's, then the groups are the same (no difference).

Effect size

What is "effect size" in the relation of statistics? It's an objective and standardised way to determine if there is indeed an observable effect in the data. The usual method is by way of ratio, so if the ratio is on the large end of the scale then you can say that the researcher will notice an effect by the factors/groups in the data. Another way to think about it, is to regard effect size as the statistical 'yard stick'; "relative to the size of my hand, how big is this ant? So, will I notice it (see it) crawling onto my hand?".
Whilst writing this, I came across a really great website that explains effect size in lay-mans terms. I think it is helpful to read, in order to give a 'layman's response' to a psychological research question. It may also help you to visualise effect size: `https://www.theanalysisfactor.com/effect-size/`.

Cohen's d

Cohen's d is used to measure the standardised difference between means in

1. one random sample drawn from a normally distributed population $\left(y \sim \mathcal{N}\left(\mu, \sigma^2\right)\right)$;

$$d = \frac{\bar{y} - \mu}{\sigma} \tag{62}$$

   This is a z-score, and the old "68-95-99.7" rule gives an indication of how we might view this: if you have a z-score of less than 1, then you know that your estimate based on the data, $\bar{y}$, lies in the middle 68%.

Let's consider only $|d|$ and assume $|d| \leq 0.2$ (small effect size - see slide 15 from lecture 2) - checking the z-table we conclude that there is less than 16% of the population, which is less extreme than our sample estimate ($\mathbb{P}(|D| < 0.2) = 0.15852$). The key point here is that this is a **standardised score**, so you can use it to measure the small thing across different groups of populations. The simplest example is comparing the group means for different classes within a school for the same test. If you know the results for the test is normally distributed with mean μ and standard deviation σ, you can compare whether the means from different classes are noticeably different from what you expect (μ).

2. two random samples drawn from normally distributed populations

$$d = \frac{\bar{y}_1 - \bar{y}_2}{s_p}. \tag{63}$$

N.B.: the use of pooled variance implies that we assume homoskedasticity; if this is true, then s_p is the best approximation for the true standard deviation σ.
Similar to the above, however we assume that $\mu_1 = \mu_2$ and so $\mu_1 - \mu_2 = 0$ is "invisible" in the equation. This is almost a t-statistic (larger standard error due to the omission of $\sqrt{1/n_1 + 1/n_2}$ in the denominator), so it gives the standardised difference of sample means between two (assumed equal) groups. This will tell you if the standardised difference (effect size) is is small (≤ 0.2), medium (≈ 0.5) or large (≥ 0.8).

Eta squared

η^2 is the proportion of the total sample variance explained by the effect $(A, B, A \times B, \dots)$.

$$\eta^2 = \frac{\mathrm{SS}_{\text{effect}}}{\mathrm{SST}} \tag{64}$$

Again, it's basically a ratio: if the variation in the data is due mostly to the effect, then this ratio is near (or more) 0.14. In ANOVA II, "effect" can be factor A, interaction effect $A \times B$, etc. In ANOVA I, the only "effect" we consider is group variation ($\mathrm{SS}_{\text{effect}}$ =SSG) so $\eta^2 = R^2$, where $R^2 = [\mathrm{Cov}(\hat{y}, y)]/s_{\hat{y}}^2 s_y^2$ is the squared ratio of covariance and variance of the data y and the predicted model $\hat{y}$ - **percentage of variance explained by the model** (see the section on regression). So for ANOVA I (not considering regression models), $\eta^2 \times 100$ is the percentage of variation in the data as explained by the variation between groups.
Advantages:

- effects are additive for balanced groups, i.e. $n_1 = n_2 = \cdots = n_I$:

$$\sum_{\text{all effects}} \mathrm{SS}_{\text{effect}} = \mathrm{SSM} \tag{65}$$

Disadvantages:

- η^2 depends on the number and size of the remaining effects. So if you have a lot of factors/interactions, or if one factor/interaction contributes the greatest, then you might have an η^2 which is small, despite the effect size being (actually) medium, for instance:

$$\begin{aligned}
\eta^2 &= \frac{\mathrm{SS_A}}{\mathrm{SST}} && (66)\\
&= \frac{\mathrm{SS_A}}{\sum_{\text{all effects}} \mathrm{SS}_{\text{effect}} + \mathrm{SS_E}} && (67)\\
&= \frac{\mathrm{SS_A}}{\mathrm{SS_A} + \mathrm{SS_B} + \mathrm{SS_C} + \cdots + \mathrm{SS_{A \times B}} + \cdots + \mathrm{SS_{A \times B \times C}} + \cdots + \mathrm{SS_E}} && (68)\\
& && (69)
\end{aligned}$$

- η^2 does not estimate the proportion of variance accounted for in the **population** - it is a **biased estimator**. It always overestimates the explained variance in the population, despite being 'good' for the sample.

Partial eta squared

This tries to solve the disadvantages of η^2, by restricting the ratio to only the effect you are interested in. The disproportional affect of adding effects to the denominator are cancelled out, however there is still the same problem of it being a biased estimator. Additionally, like η^2, $\eta_p^2 = R^2$ in ANOVA I.

$$\eta_p^2 = \frac{\text{SS}_{\text{effect}}}{\text{SS}_{\text{effect}} + \text{SSE}} \tag{70}$$

Advantages:

- η_p^2 does not depend on the remaining effects, like η^2 does. This is because the denominator omits the explained variation from other effects, and focuses solely on the ratio of explained variation (of the effect in question) vs. the combination of explained and unexplained, but restricted to a particular effect.

Disadvantages:

- Effects are no longer additive for balanced designs (see the section on η^2 for further explanation of this).
- η_p^2 only estimates the proportion of variance accounted for in the sample, and not in the population. Thus, it overestimates the population effects.

N.B.: SPSS only outputs η_p^2, so in the exam if you see an SPSS output with η^2, you can be sure which it is ☺.

Omega squared

ω^2 aims to **exactly** estimate population effects, rather than sample effects, so it is an unbiased estimator.

$$\omega^2 = \frac{\text{SS}_{\text{effect}} - \text{df}_{\text{effect}} \times \text{MSE}}{\text{MSE} + \text{SST}} \tag{71}$$

Advantages:

- ω^2 does not overestimate population effects, like η^2 and η_p^2 do.

Disadvantages:

- Effects are no longer additive for balanced designs (see the section on η^2 for further explanation of this). Furthermore, ω^2 can be negative! What does this mean?

  $$\omega^2 < 0 \iff \text{SS}_{\text{effect}} - \text{df}_{\text{effect}} \times \text{MSE} < 0 \tag{72}$$

  I should note here that the symbol $\iff$ means 'if and only if' and denotes equivalency. So in this case, 'ω^2 is negative' is equivalent to:

  $$\text{SS}_{\text{effect}} < \text{df}_{\text{effect}} \times \text{MSE} \iff \frac{\text{SS}_{\text{effect}}}{\text{df}_{\text{effect}}} = \text{MS}_{\text{effect}} < \text{MSE} \iff F = \frac{\text{MS}_{\text{effect}}}{\text{MSE}} < 1. \tag{73}$$

I found a great website which is written in a fun and bitchy tone about the benefits of which effect size to use: `http://daniellakens.blogspot.com/2015/06/why-you-should-use-omega-squared.html`. I hope you have a laugh, too!

Power

We will run quickly over statistical power: **power is how well your test works**. You, as a psychologist/statistician, want to ensure that you only make necessary changes to society and do so at the right time. What is the right time? When the data says you have to!
Consider that you are running a test on whether a new psychological disorder should be included in the new DSM. You structure your test as a hypothesis test at a significance level of $\alpha = 0.015$ (one-tailed, two-sample t-test), and find a p-value which is smaller than your acceptance level.

$$H_0 : \mu = \mu_1 - \mu_2 = 0 \qquad\qquad H_a : \mu > 0 \tag{74}$$

If you have already rejected your null hypothesis, that means your sample statistic t is more extreme than the accepted value for a particular significance level **and** degrees of freedom:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \qquad \implies \bar{x}_1 - \bar{x}_2 > \mu_0 + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{75}$$

Suppose you later find out a closer estimate that $\mu_0 = 0$ for the difference in means: $\mu_a > 0$. What is probability that you correctly reject H_0, given that you know H_a?

$$\implies \mathbb{P}\left(\bar{x}_1 - \bar{x}_2 > \mu_0 + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \,\middle|\, \mu_a\right) = \mathbb{P}\left(\bar{x}_1 - \bar{x}_2 - \mu_a > \mu_0 - \mu_a + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) \tag{76}$$

$$= \mathbb{P}\left(\frac{\bar{x}_1 - \bar{x}_2 - \mu_a}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > \frac{\mu_0 - \mu_a + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) \tag{77}$$

$$= \mathbb{P}\left(T > \frac{\mu_0 - \mu_a}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}}\right) \tag{78}$$

$$= 1 - \beta = \text{power}. \tag{79}$$

You can find the power of the test using any population, if you know μ_a (or a value which is extremely close to it - can never be 100% certain in science!).

Correlation and regression

Remember! correlation and regression are two different things, despite being heavily interrelated. Correlation is the measurement of the association between two (or more) variables, whereas regression uses the information provided by the association to predict or extrapolate data.

Pearson's r

Pearson's ρ is the measure of a linear association **in a population** between two random variables X and Y, given by

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \times \sigma_Y}, \tag{80}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively. Researchers do not always have the luxury of knowing the magnitude or direction of a linear relationship in a population, so they must estimate it using sample data. Pearson's r approximates ρ, so that we may draw inferences about the population.

$$r_{x,y} = \frac{\text{Cov}(x,y)}{s_x \times s_y} \tag{81}$$

We use subscript x, y to denote the variables which r is measuring the relationship of. We substitute in the equation for sample covariance:

$$r_{x,y} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{n-1}}{s_x \times s_y} \tag{82}$$

We can rearrange the denominators so that the standard deviations for x and y are grouped with their respective expressions:

$$r_{x,y} = \frac{\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\times\left(\frac{y_i-\bar{y}}{s_y}\right)}{n-1} \tag{83}$$

You may notice the familiar expressions for calculating the z scores for x and y for a particular i:

$$r_{x,y} = \frac{\sum_{i=1}^{n} z_{x_i} \times z_{y_i}}{n-1} \tag{84}$$

So, we have that Pearson's r is the expected value of the product of z-scores, where the degrees of freedom is $n-1$. If we want to see this equation in another way, we can return to the first expression, and elaborate:

$$r_{x,y} = \frac{\mathrm{Cov}(x,y)}{s_x \times s_y} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{n-1}}{s_x \times s_y} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}} \tag{85}$$

The $n-1$ in all of the denominators equate out to 1 ('cancel out'):

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{86}$$

Next, we expand the brackets in the denominator:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{87}$$

Using that $n\bar{x} = \sum x_i$, and similarly for y, results in the following expression for r:

$$r_{x,y} = \frac{\sum_{i=1}^{n} x_i y_i - n \times \bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{88}$$

r is bounded between -1 and 1, where an $r = 1$ represents a positive correlation between x and y: $x \propto y$ they are proportional, e.g. $y = a + bx$ (a, b are constants with $b > 0$); alternatively, $r = -1$ represents a negative correlation between x and y: $\frac{1}{x} \propto y$ they are inversely proportional, e.g. $y = a - bx$.
If $r \approx 0$, then there is no linear relationship between your variables x_i and y_i: **they are linearly independent**.

Simple linear regression (SLR)

Now that you know what Pearson's r is, you can begin learning about simple linear regression and how they interrelate. You collect a data sample from a population, and after calculating Pearson's r for your sample you conclude that x and y are correlated (later you will want to know if they are correlated in the population - Fisher Z transformation). Now, you wish to predict future outcomes associated with your data sample, and so you construct a model :

$$\underbrace{y_i}_{\text{data}} = \underbrace{\beta_0 + \beta_1 x_i}_{\substack{\mu_{y_i} \\ \text{model}}} + \underbrace{\varepsilon_i}_{\text{error}}, \tag{89}$$

for your **population** which predicts the value y_i given your input x_i, based on the constants β_0 and β_1. Given your input x, the y values are normally distributed with population mean μ_y and variance σ^2, and the error term $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is independent of x_i. This is an incredibly important assumption, as you do not reasonably expect that your model can capture/test all possible contributions to the data. What you cannot test/capture is ε_i - typical examples may be infant life experiences or innate ability. Also, you expect that your model is 'good', so you assume that the mean of ε_i is zero.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \mu_{y_i} + \varepsilon_i \sim \mathcal{N}(\mu_{y_i}, \sigma^2) \tag{90}$$

Notice that the mean of y_i also has subscript i? This is because the mean for a particular value y_i is dependent on the input x_i, so it is really $\mu_{y \text{ given } x_i}$. If we rearrange this expression in favour of ε_i, the population error term, we find that it is normally distributed with mean zero and the same variance as y:

$$\implies y_i - \mu_{y_i} = \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{91}$$

You estimate β_0 and β_1 (population coefficients) using your sample, where b_0 and b_1 are the **ordinary least squares estimates** (OLS estimates) of β_0 and β_1. They are called OLS estimators because they minimise the sum of squared errors between the actual data y_i and the predicted data $\hat{y}_i$.

$$\min\left(\sum_{i=1}^{n} e_i^2\right) = \min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) = \min\left[\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2\right] \tag{92}$$

The b_0 and b_1 which satisfy this are:

$$b_0 = \bar{y} - b_1\bar{x}; \qquad b_1 = r_{x,y} \times \frac{s_y}{s_x} = \frac{\mathrm{Cov}(x,y)}{s_x^2} \tag{93}$$

where $r_{x,y}$ is the Pearson's r for your data sample, s_y is the standard deviation of the sample for output y, s_x is the standard deviation of the sample for the input x, and $\mathrm{Cov}(x,y)$ is the covariance of x and y.

b_0 is your intercept, i.e. the value of y_i when $x_i = 0$, so this just tells you the minimum of the range for your predicted output. Under H_0, you assume β_0 to be zero, which is what you will test: is my sample coefficient b_0 significantly different from zero? Note that if your b_0 is forced to be equal to zero, then the regression line does not run through the centre mass point of $(x_i, y_i) = (\bar{x}, \bar{y})$:

$$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1 \bar{x}) + b_1 x_i = \bar{y} + b_1 (x_i - \bar{x}) = \bar{y}. \tag{94}$$

If we have a predicted output $\hat{y}_i$ equal to the sample mean $\bar{y}$, there are two possibilities; either,

$$\begin{cases} b_1 = 0: & \text{This implies that } r_{x,y} = 0, \text{ and so } x \text{ and } y \text{ are} \\ & \text{linearly independent.} \\ x_i = \bar{x}: & \text{This implies that the regression line passes} \\ & \text{through the point } (\bar{x}, \bar{y}) \text{ if and only if } b_0 \text{ is} \\ & \text{not equal to zero.} \end{cases} \tag{95}$$

It is also cool to note:

$$\hat{y}_i = (\bar{y} - b_1 \bar{x}) + b_1 x_i = \bar{y} + b_1 (x_i - \bar{x}) = \bar{y} + \left(r_{x,y} \times \frac{s_y}{s_x} \right) \times (x_i - \bar{x}) \tag{96}$$

$$\implies \underbrace{\frac{\hat{y}_i - \bar{y}}{s_y}}_{z\text{-score for } \hat{y}_i} = r_{x,y} \times \underbrace{\frac{x_i - \bar{x}}{s_x}}_{z\text{-score for } x_i} \tag{97}$$

b_1 is your slope coefficient: when my x_i increases, does my y_i increase or decrease and at what rate? You can see from (97) that $r_{x,y}$ is the slope of the regression line of the standardized data points. If your variables are strongly linearly correlated, such as $r = 1$, then $b_1 \approx {}^{s_Y}/_{s_X}$ which is the ratio of standard deviations between your independent and dependent variables; if $r = -1$, then $b_1 \approx -{}^{s_Y}/_{s_X}$, meaning that $\hat{y}_i$ will decrease as x_i increases, respective to the rate ${}^{s_Y}/_{s_X}$. It is important to note that the sign of b_1 (i.e. negative or positive) is strictly governed by the sign of the covariance as s_y and s_x are always strictly greater than zero (equal to zero if and only you have a single data point for your sample). So r not only tells you about the observed relationship in the sample, but what to expect of your predictive regression line.
So, now you have your prediction model $\hat{y}_i = b_0 + b_1 x_i + e_i$, you can calculate the (squared) **standard error of the estimate**, which estimates σ^2, the variance of the population y:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{\underbrace{n-2}_{\text{df}}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \tag{98}$$

If you are asked, "what is the percentage of explained variance ..." then you know that they are asking you for R^2 (or adjusted R^2). If the regression model is linear, then $R^2 = r^2$ and the percentage of explained variance is $R^2 \times 100\%$.

$$R^2 = 1 - \frac{\text{SS}_{\text{residuals}}}{\text{SS}_{\text{total}}} \tag{99}$$

Recall that,

$$\text{SS}_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{SS}_{\text{regression}} + \text{SS}_{\text{residuals}}. \tag{100}$$

$$\implies R^2 = \frac{\text{SS}_{\text{total}} - \text{SS}_{\text{residuals}}}{\text{SS}_{\text{total}}} = \frac{\text{SS}_{\text{regression}}}{\text{SS}_{\text{total}}} = \text{VAF} \tag{101}$$

If you have more than one independent variable, e.g. $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$, then you consider adjusting for the increase in parameters:

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-p-1} = 1 - \frac{\text{SS}_{\text{residuals}}}{\text{SS}_{\text{total}}} \cdot \frac{n-1}{n-p-1} = 1 - \frac{\text{SS}_{\text{residuals}}/(n-p-1)}{\text{SS}_{\text{total}}/(n-1)} \tag{102}$$

Comparing the equation for R^2 and $\bar{R}^2$ (Wherry's adjusted R^2), the 'adjustment' is the degrees of freedom for the residuals. R^2 assumes that $\text{df}_{\text{residuals}} = n - 1$, which leads to a biased estimate for the population variance of the residuals. By adjusting the degrees of freedom to $n - p - 1$, where p is the number of independent variables (not including constant β_0), we gain an unbiased estimate. This will be further explored during multivariate regression.

If you are asked, "what is the variability about the line of regression", then you know that they are asking you for the standard error of the estimate $s = \sqrt{\sum e_i^2/(n-2)}$.
If you are asked to construct a $(1-\alpha)\,\%$ CI for the population regression parameter β_0 (or β_1), then you know that they are asking you for a t-statistic based CI (see Table 7):

Table 7

	β_0	β_1
Estimate	$b_0 = \bar{y} - b_1\bar{x}$	$b_1 = r_{x,y} \times \frac{s_y}{s_x}$
Mean of estimate	$\mu_{b_0} = \beta_0$	$\mu_{b_1} = \beta_1$
CI	$b_0 \pm t^*_{n-2} \times \text{SE}_{b_0}$, where SE_{b_0} approximates σ_{b_0} and is computed by SPSS	$b_1 \pm t^*_{n-2} \times \text{SE}_{b_1}$, where SE_{b_1} approximates σ_{b_1} and is computed by SPSS
Test	H_0: $\beta_0 = 0$ vs H_a: $\beta_0 \neq 0$	H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$
Statistic	$t = \frac{b_0 - \beta_0^{=0, \text{ under } H_0}}{\text{SE}_{b_0}} \sim t(n-2)$	$t = \frac{b_1 - \beta_1^{=0, \text{ under } H_0}}{\text{SE}_{b_1}} \sim t(n-2)$
Assumptions	$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2_{b_0}\right)$	$b_1 \sim \mathcal{N}\left(\beta_1, \sigma^2_{b_1}\right)$
	$\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$	

In general, we are most interested in the results of testing done on the true parameter β_1: if we do not reject H_0, then we conclude that x and y do **not** have a linearly correlated relationship in the population (could be independent variables or another relationship is more appropriate, e.g. semi-logarithmic).

SLR pop-quiz

1. Why do we perform (linear) regression?
2. Based on an SPSS output, what inferences can we make?
3. When might we examine the residual plot, and what inferences might we draw from it?
4. What is the "point of centre mass" on a scatter plot?
5. What tests might we perform on the output of a (linear) regression?
6. What is R^2, and why/when do we adjust it?
7. If we compute the Pearson's correlation coefficient, what can we infer from this statistic? Think about:
   - scatterplot (x_i, y_i);
   - residual plot $(y_i, \hat{y}_i)$;
   - roles in regression equation $\hat{y}_i = b_0 + b_1 x_i$.

Fisher Z-transformation

In the previous sections, we looked at the population model and tested goodness of fit for a linear model. In order to further investigate the existance of a linear relationship between x and y in the population, we look distinctly at the population correlation coefficient, Pearson's ρ, and its sample counterpart, Pearson's r. The following t statistic is used to test H_0: $\rho = 0$ (linear independence) against H_a: $\rho \neq 0$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \tag{103}$$

The problem we face is when we reject the null hypothesis, and consider population correlation coefficients which are not equal to zero - H_0: $\rho = a$ where a is a non-zero constant against H_a: $\rho \neq a$. Under this assumption, r is not normally distributed, so how do we construct a confidence interval around r for ρ? The answer is the Fisher Z transformation to r_z (approximately normal):

$$r_z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \sim \mathcal{N}\left(\rho_z = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \sigma^2_{r_z} = \frac{1}{n-3}\right) \tag{104}$$

This says that our transformed sample correlation coefficient, r_z, is normally distributed with mean ρ_z, which is the transformed population correlation coefficient (just insert ρ in place of r in the equation for r_z), and standard deviation $1/\sqrt{n-3}$.

$$Z = \frac{r_z - \rho_z}{1/\sqrt{n-3}} \sim \mathcal{N}(0,1) \tag{105}$$

If we want to find out if r_z is significant, i.e. the probability of achieving a more extreme value, we can do a z-test:

$$\mathbb{P}\left(|Z| > \frac{r_z - \rho_z}{1/\sqrt{n-3}}\right) = p \leftarrow \text{look this up from the } z\text{-table.} \tag{106}$$

The above z-test is two tailed, because our alternative hypothesis is that $\rho \neq a$, so $\rho < a$ or $\rho > a$. Now we can construct a $(1-\alpha)\%$ confidence interval about r_z for ρ_z, using the critical z-value.

$$\implies \text{CI for } \rho_z: \quad r_z \pm z^* \frac{1}{\sqrt{n-3}} \tag{107}$$

So we have an upper and lower bound now for our confidence interval for ρ_z:

$$\text{LB}_{\rho_z} = r_z - z^* \frac{1}{\sqrt{n-3}} \qquad \text{UB}_{\rho_z} = r_z + z^* \frac{1}{\sqrt{n-3}} \tag{108}$$

This does not tell us anything about ρ, only ρ_z, so in order to have a confidence interval for ρ we can transform r_z back to r using:

$$r_z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \qquad \implies 2r_z = \log\left(\frac{1+r}{1-r}\right) \qquad \implies e^{2r_z} = \frac{1+r}{1-r} \qquad \implies r = \frac{e^{2r_z}-1}{e^{2r_z}+1} \tag{109}$$

The above is the Inverse Fisher Z transformation. We can now construct a confidence interval for our population correlation parameter.

$$\implies \text{CI for } \rho: \quad \left(\frac{e^{2\text{LB}_{\rho_z}}-1}{e^{2\text{LB}_{\rho_z}}+1}, \frac{e^{2\text{UB}_{\rho_z}}-1}{e^{2\text{UB}_{\rho_z}}+1}\right) \tag{110}$$

What do we know about confidence intervals? Well if we have constructed them well (effect size, sample size n, etc.) and the hypothesised population parameter is not contained in the interval **then we reject the null hypothesis**. So if we get a CI for ρ, and ρ_0 is not included in the interval, then we must reject $\rho = \rho_0$, and look for an alternative value for ρ.

$$\begin{array}{ccc} r & \Rightarrow & b_0 \text{ and } b_1 \\ \Downarrow & & \\ r_z & \Rightarrow & \text{CI for } \rho_z \\ \Downarrow & & \Downarrow \\ p\text{-value for } r_z & & \text{CI for } \rho \end{array} \tag{111}$$

Multiple Linear Regression

The two regression "sentences" that I want you to keep inside of your head for the rest of time are:

$$y = \overbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}^{\mu_y} + \varepsilon \sim \mathcal{N}\left(\mu_y, \sigma^2\right); \qquad \varepsilon = y - \mu_y \sim \mathcal{N}\left(0, \sigma^2\right).$$

These sentences are dense in information about the assumptions of (multiple) linear regression.

Firstly, it is important to imagine your data in matrix-vector form: given a sample of size n (of the dependent variable).

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + \cdots + \beta_p X_{1,p} \\ \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + \cdots + \beta_p X_{2,p} \\ \vdots \\ \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + \cdots + \beta_p X_{n,p} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

I have placed parentheses around the vectors and square brackets around the matrices so that it is easy to see the difference. We can rewrite the vector containing the β's and X's to a matrix containing only the X's multiplied by a vector containing only the β's.

$$\implies \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{bmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

It is also important to know how covariance/variance is calculated in vector matrix form, with regards to the following assumptions. Given a sample of data in a vector form, e.g. $(x_1, x_2, \ldots, x_n)$, you can display the variance of each and the pair-wise covariance between data in a matrix form which is denoted as Σ, which is the capital version of σ. As $\operatorname{Var} x_j = \operatorname{Cov}(x_j, x_j)$, along the diagonal of Σ are the variances, and every other cell contains the covariance.

$$\Sigma = \begin{bmatrix} \operatorname{Var} x_1 & \operatorname{Cov}(x_1, x_2) & \operatorname{Cov}(x_1, x_3) & \ldots & \operatorname{Cov}(x_1, x_n) \\ \operatorname{Cov}(x_2, x_1) & \operatorname{Var} x_2 & \operatorname{Cov}(x_2, x_3) & \ldots & \operatorname{Cov}(x_2, x_n) \\ \operatorname{Cov}(x_3, x_1) & \operatorname{Cov}(x_3, x_2) & \operatorname{Var} x_3 & \ldots & \operatorname{Cov}(x_2, x_n) \\ \vdots & & & \ddots & \vdots \\ \operatorname{Cov}(x_n, x_1) & \operatorname{Cov}(x_n, x_2) & \operatorname{Cov}(x_n, x_3) & \ldots & \operatorname{Var} x_n \end{bmatrix}$$

1. **Independence of observations**

   This assumptions says that the sample of dependent variable data values $y_1, y_2, \ldots, y_n$ were collected or observed with imposed or assumed independence. It might be that the data collection is anonymous or your method of sampling may ensure independence, e.g. stratified, simple random, etc. In essence, you are stating that the covariance between any dependent data points is not significantly different from zero, for example $\operatorname{Cov}(y_1, y_2) \approx 0$. With reference to the covariance matrix Σ above, the covariance matrix of the sample $y_1, y_2, \ldots, y_n$ has the same value σ^2 along the diagonal and zeros everywhere else.

   $$\begin{bmatrix} \sigma^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma^2 & 0 & \ldots & 0 \\ 0 & 0 & \sigma^2 & \ldots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sigma^2 \end{bmatrix}$$

   In terms of the data matrix below,

   $$\begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \ldots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \ldots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \ldots & X_{n,p} \end{bmatrix}$$

   this assumption states that the rows are not linear combinations of each other, e.g. that row 1 doesn't equal a multiple of row 2. This is why you can never copy-paste data in order to increase your n!!

2. **Linearity**

   This assumption just states the relationship between the dependent variable y and the independent variables X's is linear in the regression equation, e.g. $y = \beta_0 + \beta_1 X + \beta_2 \log X + \varepsilon$; we do not want β^2's in the model. You can think of this as "linear in parameters".

3. **Normality**

   As frequentist statistics works because of the miracle of the Central Limit Theorem, we need to conform to it's rule of normality. The distributions we use to test (most) of our hypotheses all come from the normal distribution, whether directly with the z-test or indirectly like the F-test.

   You can confirm this assumption using QQ-plots of the variables, or of the residuals.

4. **Homoskedasticity**

   Linear relationships $y = x$ are the main focus however there are other relationships (non-linear), such as quadratic $y = x^2$, log-linear $y = \log x$, and exponential $y = e^x$, which require transformation before regressing.

   It is best to view scatter plots of the residuals v.s. dependent variable or the predicted v.s. dependent variable to confirm whether the homoskedasticity assumption is met. If the relationship between the variables is non-linear, there will be a great difference between the predicted and observed values of the dependent variable. You might see that the residuals increase or decrease, which displays heteroskedasticity. You might see that the best fitting line for predicted v.s. observed has a curve in it, which displays that you might need to include an X_1^2 or $X_1 X_2$, etc. to your regression equation.

   NO FAN IN, NO FAN OUT, UNIFORM DISTRIBUTION OF RESIDUALS AROUND THE ZERO LINE.

We expect the residuals to be zero, but accept some margin of error and additionally expect that this margin of error is uniform over all values of y. The expected value is also called the mean, hence $\mu_\varepsilon = 0$, and that the variance is constant, hence $\operatorname{Var}\varepsilon = \sigma^2$ and not the Σ matrix. If we have the Σ matrix, then some of the covariances are not zero anymore, and there is some fan-in or -out displayed in the residuals plot. If you have only σ^2, then you know that the variance for each residual is the same and there is no covariances between the residuals.

5. **No perfect multicollinearity**

   This simply states that the correlation between the independent variables is zero and is tested using the VIF or tolerance. In terms of the data matrix below,

$$\begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \dots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \dots & X_{n,p} \end{bmatrix}$$

   this assumption states that the columns are not linear combinations of each other, e.g. that column 1 doesn't equal a multiple of column 2, and that if you were to perform a regression of one independent variable on the others, then you would not have too high an R^2. For example, if $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ you might calculate the R^2 from $X_1 = \alpha_0 + \alpha_2 X_2 + \delta$ and if that $R^2 > 0.75$ you would say that there is strong correlation between X_1 and the other independent variables. Therefore you might consider adding an interaction variable to the y regression model; in this example, add $\beta_3 X_1 X_2$ if positive correlation or $\beta_4 X_1/X_2$ if negative correlation.

Variance inflation factor (VIF)

The variance inflation factor is a measure of multicollinearity, and is related to how tolerant we are of covariance between the regressors. In the regression equation

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

we assume that the regressors $X_1, \dots, X_p$ are dependent on y but independent of each other. In order to be sure of this, for each $j = 1, 2, \dots, p$ we calculate the R^2 of the following regression equation:

$$X_j = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \cdots + \alpha_p X_p + \delta.$$

This gives us R_j^2, which is the R^2 when we regress all independent variables (excluding X_j)

$X_{-j} = \{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p\}$ on X_j. The rule of thumb: if VIF<4 or tolerance>0.25, then we do not suspect any issue with multicollinearity.

$$\text{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{tolerance}} < 4 \qquad \text{which is equivalent to} \qquad R_j^2 < 0.75.$$

This rule of thumb tells us that we accept that there might be some multicollinearity present in the data, but we are only tolerant of a certain level, namely if the Variance (of X_j) Accounted For (VAF) by the other independent variables X_{-j} is less than 75%.

For example, if we have two independent variable and one dependent variable $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ and then we can calculate the VIF's by first forming a second regression equation $X_1 = \alpha_0 + \alpha_2 X_2 + \delta$, which involves only the X's. As we only have two independent variables, we can rewrite the second regression equation so that it has X_2 on the left of the equals sign and everything else on the right and so $R_1^2 = R_2^2$. In other words, the VIF's are equal if you have only two independent variables as the R_j^2 just becomes r^2, where r is the Pearson's correlation coefficient between X_1 and X_2.

What, when and why (assumptions)

The usual question that statisticians ask is, "what formula do I need to use for this?" and hopefully Tables 8 to 10 will help to clear that up.

Usually, we start by talking about some random variable y which has an approximately normal population distribution, i.e. $y \sim \mathcal{N}(\mu, \sigma^2)$. If the population of y is **not** distributed normally, the Central Limit Theorem allows us to conclude that the sampling distribution of the mean of y (many many samples) is! So even if the distribution of the population of y is skewed, or has no observable pattern, the CLT states that the mean of all possible samples (i.e. large n) has an observable pattern, namely normal. So if y has mean μ and standard deviation σ, then $\bar{y} \sim \mathcal{N}(\mu, \sigma^2/n)$ is the sampling distribution of the mean of y.

Table 8

Number of groups (I)	Assumptions	Name	Statistic	Confidence interval
1	Normality CLT μ known σ known independence	z-score standardised score	$z = \frac{y-\mu}{\sigma} \sim \mathcal{N}(0,1)$	
1	CLT μ known σ **unknown** independence	One-sample independent t-test for estimating population mean	$t = \frac{\bar{y}-\mu}{s/\sqrt{n}} \sim t(n-1)$	CI for μ: $\bar{y} \pm t^* \frac{s}{\sqrt{n}}$
2	Normality CLT independence μ_1, μ_2 unknown σ_1, σ_2 unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 \approx \sigma_2$ $H_0 : \mu = \mu_1 - \mu_2 = \mu_0 = 0$	Two-sample independent t-test for comparing means (assumed equal variance)	$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n-2)$	CI for $\mu = \mu_1 - \mu_2$: $(\bar{y}_1 - \bar{y}_2) \pm t^* s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
2	Normality CLT **dependence** μ_1, μ_2 unknown σ_1, σ_2 unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 \approx \sigma_2$ Construct a new variable $d_i = y_i - x_i$ (difference) from the dependent sample (x_i, y_i) Equal sample sizes $n_x = n = n_y$ $H_0 : \mu_d = \mu_0 = 0$	Two-sample **dependent** t-test for comparing means (**paired data**, e.g. before and after a treatment)	$\left.\begin{array}{l} \bar{d} = \frac{\sum_{i=1}^n (y_i - x_i)}{n} = \bar{y} - \bar{x} \\ s_d^2 = \frac{\sum_{i=1}^n d_i^2 - \bar{d}^2/n}{n-1} \end{array}\right\} \Longrightarrow t = \frac{\bar{d}-\mu_d}{s_d/\sqrt{n}} \sim t(n-1)$	CI for $\mu_d = \mu_y - \mu_x$: $\bar{d} \pm t^* \frac{s_d}{\sqrt{n}}$
2	Normality CLT independence μ_1, μ_2 unknown σ_1, σ_2 unknown **Homogeneity of variances violated** $\sigma_1 \neq \sigma_2$ $H_0 : \mu = \mu_1 - \mu_2 = \mu_0 = 0$	Two-sample independent t-test for comparing means (assumed unequal variance)	$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t(k)$, where k is approximated by a computer	CI for $\mu = \mu_1 - \mu_2$: $(\bar{y}_1 - \bar{y}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

Table 9

ANOVA I	
Number of groups	3+
Assumptions	Normality CLT independence μ_i, $i = 1, 2, \ldots, I$, unknown σ_i, $i = 1, 2, \ldots, I$, unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 = \sigma_2 = \cdots = \sigma_I$ $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ all of the groups are the same H_a: one of the groups is different from the rest
Name	ANOVA I "is the observed variance in the data attributable to the variation between the groups, or within the groups" Compare the means of 3 or more groups; if 2 groups, then t-test suffices
Statistic	$i = 1, 2, \ldots, I$ indexes the **group number** $j = 1, 2, \ldots, n_i$ indexes the **score** within a group $\mathrm{SST} = \mathrm{SSG} + \mathrm{SSE}$ $\mathrm{df}_T = \mathrm{df}_G + \mathrm{df}_E$ $\mathrm{MST} = \dfrac{\mathrm{SST}}{\mathrm{df}_T} = \mathrm{Var}\, y$ $\mathrm{SST} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$ $\mathrm{df} = n - 1$ $\mathrm{MSG} = \dfrac{\mathrm{SSG}}{\mathrm{df}_G}$ $\mathrm{SSG} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2$ $\mathrm{df}_G = I - 1$ $\mathrm{MSE} = \dfrac{\mathrm{SSE}}{\mathrm{df}_E} = s_p^2$ $\mathrm{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{I} (n_i - 1)\, s_i^2$ $\mathrm{df}_E = n - I$ $\implies F = \dfrac{MSG}{MSE} \sim F(\mathrm{df}_G, \mathrm{df}_E)$
Confidence interval	CI for mean of group i: (**homoskedasticity met**): $\bar{y}_i \pm t^*_{n-1} \frac{s_p}{\sqrt{n_i}}$ (**homoskedasticity violated**): $\bar{y}_i \pm t^*_{n_i-1} \frac{s_i}{\sqrt{n_i}}$

Table 10

ANOVA II	
Number of groups	3+ or 2+ factors
Assumptions	Normality CLT independence μ_i, $i = 1, 2, \ldots, I$, unknown σ_i, $i = 1, 2, \ldots, I$, unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 = \sigma_2 = \cdots = \sigma_I$ $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ all of the groups are the same H_a: one of the groups is different from the rest
Name	ANOVA II "is the observed variance in the data attributable to the variation between the groups (across factors), or within the groups" Compare the means of 3 or more groups, where group membership is defined by 2 factors, A and B; if 2 groups, then t-test suffices
Statistic	$i = 1, 2, \ldots, I$ indexes **factor** A $j = 1, 2, \ldots, J$ indexes **factor** B $k = 1, 2, \ldots, n$ indexes the individual score $\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}$ $\text{df}_T = \text{df}_A + \text{df}_B + \text{df}_{A\times B} + \text{df}_E$ $\text{MST} = \dfrac{\text{SST}}{\text{df}_T} = \text{Var}\, y$ $\text{SST} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ijk} - \bar{y}\right)^2$ $\text{df} = n - 1$ $\text{MSA} = \dfrac{\text{SSA}}{\text{df}_A}$ $\text{SSA} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{y}_i - \bar{y}\right)^2 = \sum_{i=1}^{I} n \times J \times \left(\bar{y}_i - \bar{y}\right)^2$ $\text{df}_A = I - 1$ $\text{MSB} = \dfrac{\text{SSB}}{\text{df}_B}$ $\text{SSB} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{y}_j - \bar{y}\right)^2 = \sum_{j=1}^{J} n \times I \times \left(\bar{y}_j - \bar{y}\right)^2$ $\text{df}_B = J - 1$ $\text{MSAB} = \dfrac{\text{SSAB}}{\text{df}_{A\times B}}$ $\text{SSAB} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}\right)^2$ $\text{df}_{A\times B} = (I - 1) \times (J - 1)$ $\text{MSE} = \dfrac{\text{SSE}}{\text{df}_E} = s_p^2$ $\text{SSE} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ijk} - \bar{y}\right)^2$ $\text{df}_E = n - I \times J$ $\Longrightarrow \begin{cases} F_A = \dfrac{MSA}{MSE} \sim F\left(\text{df}_A, \text{df}_E\right) \\ F_B = \dfrac{MSB}{MSE} \sim F\left(\text{df}_B, \text{df}_E\right) \\ F_{A\times B} = \dfrac{MSAB}{MSE} \sim F\left(\text{df}_{A\times B}, \text{df}_E\right) \end{cases}$
Confidence interval	CI_{ij} for comparing the means of groups i and j: $\left(\bar{y}_i - \bar{y}_j\right) \pm t^{**}_{\text{df}_E} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ (See the section on post-hoc methods)

Exercises

Fisher Z Transformation

1. A regression model $Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j$ has $n = 11$ observations. The sample correlation between X and Y is 0.60. We test the null hypothesis $H_0 : \rho = 0$ (the true correlation between the X and Y variables is zero).

   (a) What is the t-value to test the null hypothesis?

   (b) What is the p-value to test the null hypothesis? Summarise your results of the test.

   (c) What can you say about the results of the test with respect to the sample correlation coefficient?

2. A linear regression with 11 data points has an estimated β_1 of 4.5 and a sample correlation between the X and Y values of 0.60.

   (a) What is the t-value to test the null hypothesis that the correlation ρ is zero? Summarise your results of the test.

   (b) What is the t-value to test the null hypothesis that β_1 is zero? Summarise your results of the test.

   (c) What is the standard error of the estimate of β_1? What does this tell you about the reliability of the t-test you performed, and how might you improve the test?

   (d) How are these two tests similar/different?

3. X and Y are a bivariate normal distribution from which a sample of 40 observations is taken. The sample correlation between X and Y is 0.833. We test the null hypothesis $H_0 : \rho = 0.750$. The alternative hypothesis is $H_a : \rho_0 > 0.750$.

   (a) What is the Fisher transform r_z of the random variable r of the correlation r between X and Y?

   (b) What is the Fisher transform of the observed correlation?

   (c) What is the distribution of r_z?

   (d) What is the Fisher transform of the correlation ρ_0 assumed in the null hypothesis?

   (e) What is the z-value to test this null hypothesis?

   (f) What is the p-value for this test of the null hypothesis?

   (g) What is the 95% confidence interval for the true value of the Fisher transform of the correlation?

   (h) What is the 95% confidence interval for the true value of the correlation?

   (i) Summarise your results of the previous parts.

4. Using the data set Album Sales (JASP>data library>regression>album sales), we have the following outputs:

Table 11: Descriptive Statistics

	sales	adverts
Valid	200	200
Missing	0	0
Mean	193.200	614.412
Std. Deviation	80.699	485.655
Minimum	10.000	9.104
Maximum	360.000	2271.860

Table 12: Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.578	0.335	blue!25	65.991

Table 13: Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	134.140	7.537		blue!25	< .001
	adverts	0.096	0.010	blue!25	blue!25	< .001

(a) Fill in the missing values (highlighted lilac). Explain the steps you take to calculate the appropriate values.

(b) Perform a hypothesis test for correlation (there are two ways of doing this). State the null and alternative hypotheses, compute your statistic and summarise your findings.

(c) Compute the confidence interval for the correlation. Explain all intermediary steps.

Solutions

1. Recall:
$$t = \frac{\text{estimate} - \text{hypothesised value}}{\text{standard error of the estimate}}.$$
For this question, we have the hypothesised value $\rho_0 = 0$, estimate $r = 0.60$, and the standard error of r is $se_r = \sqrt{\frac{1-r^2}{n-2}}$. Assume that the alternative hypothesis is $H_A : \rho > 0$.
$$\implies t = \frac{(r-\rho_0)\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.60\sqrt{9}}{\sqrt{0.64}} = \frac{0.6 \cdot 3}{0.8} = 2.25 \sim t_9 = 2.262 \implies p = \mathbb{P}(t_9 > 2.262) = 0.025502.$$
We reject the null hypothesis at a 5% level, i.e. the population correlation coefficient between variables X and Y is significantly greater than zero, and they are dependent.

2. The t-statistic for $H_0 : \beta_1 = 0$ is identical to the t-statistic for $H_0 : \rho = 0$; this question was meant to trick you. Sorry, not sorry ☺
You can use the answer to the previous question to determine the standard error for β_1:
$$t = \frac{b_1 - \beta_1}{se_{\beta_1}} \implies 2.25 = \frac{4.5 - 0}{se_{\beta_1}} \implies se_{\beta_1} = \frac{4.5}{2.25} = 2.$$
We know that the standard error is always decreased by increasing our sample size.

3.
$$r_z = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2} = \frac{\ln\left(\frac{1.833}{0.167}\right)}{2} = 1.198 \tag{112}$$
$$r_z \sim \mathcal{N}\left(\frac{\ln\left(\frac{1+\rho}{1-\rho}\right)}{2}, \frac{1}{n-3}\right) = \mathcal{N}\left(\frac{\ln\left(\frac{1.750}{0.250}\right)}{2}, \frac{1}{40-3}\right) = \mathcal{N}(0.973, 0.027) \tag{113}$$
$$\rho_{0z} = \frac{\ln\left(\frac{1.750}{0.250}\right)}{2} = 0.973 \tag{114}$$
$$z = (r_z - \rho_{0z})\sqrt{n-3} = (1.198 - 0.973)\sqrt{37} = 1.368 \sim \mathcal{N}(0,1) \tag{115}$$
$$p = \mathbb{P}(z > 1.368) = 0.085656 \tag{116}$$
$$\text{CI for } \rho_z\text{: } r_z \pm \frac{z^*}{\sqrt{n-3}} = 1.198 \pm \frac{1.645}{\sqrt{37}} = (0.928, 1.468) \tag{117}$$
$$\text{CI for } \rho\text{: } \left(\frac{e^{2\cdot \text{LB}_{\rho_z}} - 1}{e^{2\cdot \text{LB}_{\rho_z}} + 1}, \frac{e^{2\cdot \text{UB}_{\rho_z}} - 1}{e^{2\cdot \text{UB}_{\rho_z}} + 1}\right) = \left(\frac{e^{2\cdot 0.928} - 1}{e^{2\cdot 0.928} + 1}, \frac{e^{2\cdot 1.468} - 1}{e^{2\cdot 1.468} + 1}\right) = (0.730, 0.899) \tag{118}$$
We do not reject the null hypothesis at the 5% level as $p > 0.05$ and 0.75 is contained in the 90% confidence interval for ρ.

4. Stein's adjusted R^2
$$\bar{R}^2_S = 1 - \left(\frac{n-1}{n-p-1}\right)\left(\frac{n-2}{n-p-2}\right)\left(\frac{n+1}{n}\right)\left[1 - R^2\right] \tag{119a}$$
Wherry's adjusted R^2
$$\bar{R}^2_W = 1 - \left(\frac{n-1}{n-p-1}\right)\left[1 - R^2\right] \tag{119b}$$
$$\bar{R}^2_S = 1 - \left(\frac{199}{198}\right)\left(\frac{198}{197}\right)\left(\frac{201}{200}\right)\left[1 - 0.578^2\right] = 0.324 \tag{120}$$
$$\bar{R}^2_W = 1 - \left(\frac{199}{198}\right)\left[1 - 0.578^2\right] = 0.331 \tag{121}$$
$$t_{b_0} = \frac{143.14}{7.537} = 18.99; \qquad t_{b_1} = \frac{0.096}{0.010} = 9.6. \tag{122}$$
The standardised slope coefficient in simple linear regression is equal to $r = 0.578$. The p-value for the slope coefficient indicates that the variables are significantly correlated, i.e. $H_0 : \rho = 0$ is rejected as this is equal to $H_0 : \beta_1 = 0$. The 95% CI for ρ (two-sided alternative hypothesis) is $(0.478, 0.664)$.

Regression - ANOVA analysis

1. The "Healthy Breakfast" dataset contains, among other variables, the Consumer Reports ratings of 77 cereals, the number of grams of sugar contained in each serving, and the number of grams of fat contained in each serving.

   Considering "Sugars" as the explanatory variable and "Rating" as the response variable generated the following regression line:

   Rating = 59.3 - 2.40 Sugars

Source	DF	SS	MS	F	p
Regression	1	8654.7	8654.7	102.35	0.000
Error	75	6342.1	84.6		
Total	76	14996.8	194.76		

Table 14: Analysis of Variance - rating ~ sugar

   As a simple linear regression model, we previously considered "Sugars" as the explanatory variable and "Rating" as the response variable.

   The regression line generated by the inclusion of "Sugars" and "Fat" is the following:

   Rating = 61.1 - 2.21 Sugars - 3.07 Fat

Source	DF	SS	MS	F	p
Regression	2	9325.3	4662.6	60.84	0.000
Error	74	5671.5	76.6		
Total	76	14996.8	194.76		
Source	DF	Seq SS			
Sugars	1	8654.7			
Fat	1	670.5			

Table 15: Analysis of Variance - rating ~ sugar + fat

   (a) Define the population regression model using table 15. If two cereals have the same fat content but different sugar content, what can you say about the rating?

   (b) What does VIF stand for? Compute the VIF using table 15.

   (c) What does VAF stand for? Compute the VAF using tables 14 and 15.

   (d) How do the ANOVA results change when "FAT" is added as a second explanatory variable?

   (e) Formulate appropriate hypotheses, make a decision and explain your reasoning.

2. Answer the following questions using the tables and graphs below.

Table 16: Descriptive Statistics

	sales	adverts	airplay	attract
Valid	200	200	200	200
Missing	0	0	0	0
Mean	193.200	614.412	27.500	6.770
Std. Error of Mean	5.706	34.341	0.868	0.099
Std. Deviation	80.699	485.655	12.270	1.395
Minimum	10.000	9.104	0.000	1.000
Maximum	360.000	2271.860	63.000	10.000

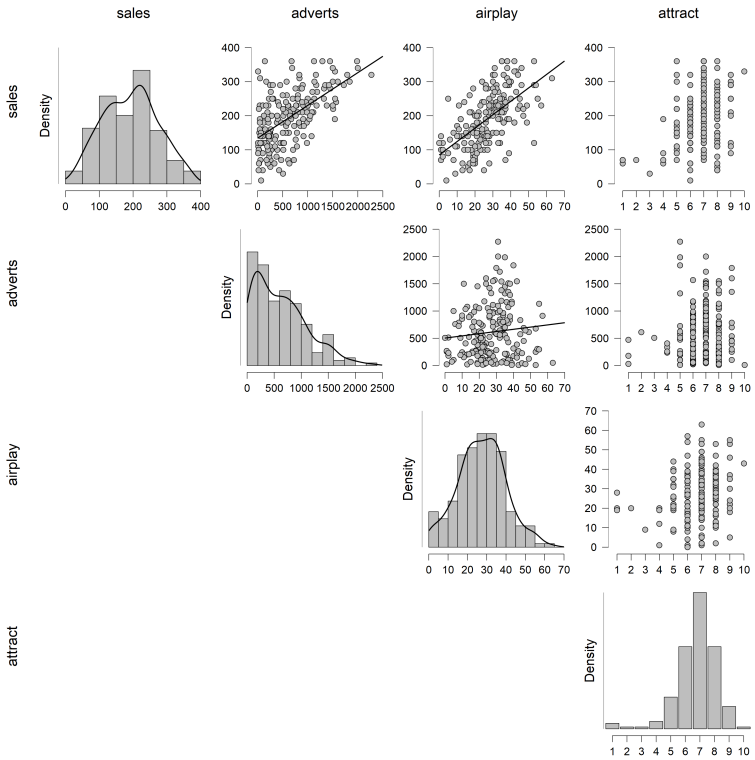


Figure 3

This fictional data set, "Album Sales", provides factors that may influence album sales Variables:

adverts Amount (in thousands of pounds) spent promoting the album before release.

sales Sales (in thousands of copies) of each album in the week after release.

airplay How many times songs from the album were played on a prominent national radio station in the week before release.

attract How attractive people found the band's image (1 to 10).

Table 17: Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
0	0.578	0.335	0.331	65.991	0.335	99.587	1	198	< .001
1	0.815	0.665	0.660	47.087	0.330	96.447	2	196	< .001

Table 18: Coefficients

							95% CI		Collinearity Statistics	
Model		Unstandardized	Standard Error	Standardized	t	p	Lower	Upper	Tolerance	VIF
0	(Intercept)	134.140	7.537		17.799	< .001	119.278	149.002		
	adverts	0.096	0.010	0.578	9.979	< .001	0.077	0.115	1.000	1.000
1	(Intercept)	-26.613	17.350		-1.534	0.127	-60.830	7.604		
	adverts	0.085	0.007	0.511	12.261	< .001	0.071	0.099	0.986	1.015
	airplay	3.367	0.278	0.512	12.123	< .001	2.820	3.915	0.959	1.043
	attract	11.086	2.438	0.192	4.548	< .001	6.279	15.894	0.963	1.038

Table 19: ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	433687.833	1	433687.833	99.587	< .001
	Residual	862264.167	198	4354.870		
	Total	1.296e+6	199			
1	Regression	861377.418	3	287125.806	129.498	< .001
	Residual	434574.582	196	2217.217		
	Total	1.296e+6	199			

(a) What is the population regression equations for Model 0 and 1?

(b) Describe the regression equations you wrote above, in words (1-2 sentences each model). What is the point of the comparison?

(c) Summarise the findings of table 20a and compare with fig. 3.

(d) Write the null and alternative hypothesis based on the regression equations you wrote in part (a). Now, describe these hypothese in words (do not refer to beta coefficients, just use plain language - like you're informing a friend). What does table 21a inform you about your hypotheses?

(e) Under Model 0, what is the expected output if the explanatory variable input has value 600? Compare this with output with the output from Model 1 under the same conditions. Explain the difference in your results.

(f) Explain the fourth and seventh columns of table 21a.

(g) Table 21c provides you with the VIF for both models. Interpret the results without making too many references to the exact value of the VIF, i.e. what do these values mean?

(h) Provide the standardised regression equations for both models.

(i) Use table 21b to make your decision about your hypotheses. Explain your reasoning.

Solutions

1. (a) The population regression model used in table 15 is rating = β_0 + β_1 sugars + β_2 fat + ε, where β_j's are approximated by b_j's such that $\vec{b} = (61.1, -2.21, -3.07)^T$. If variable fat is kept constant, then the marginal difference in rating is -2.21 per gram of sugar. This says that the rating of the breakfast cereal will decrease by 2.21 points per additional gram of sugar, under the condition that fat content is kept constant.

   (b) VIF stands for variance inflation factor, and is given by $\text{VIF}_j = 1/\left(1 - R_j^2\right)$, where R_j^2 is the coefficient of determination of the regression equation $X_j = \alpha_0 + \alpha_1 X_{-j} + \delta$ (regress the explanatory variables on the others). The square root of the VIF indicates how much larger the standard error increases compared to if that variable had 0 correlation to other predictor variables in the model. For example, if the variance inflation factor of a predictor variable were 5.27 ($\sqrt{5.27} = 2.3$), this means that the standard error for the coefficient of that predictor variable is 2.3 times larger than if that predictor variable had 0 correlation with the other predictor variables. Rule of thumb: $VIF_j > 10$ indicates multicollinearity in the model, i.e. explanatory variables are dependent on each other.

   It is not possible to compute the VIF using table 15, as we need the partial SS information.

   (c) VAF stands from variance accounted for and is given by the R^2 coefficient for linear regression, where $R_2 = SSR/SST$. From table 14, R^2 = 8654.7/14996.8 = 0.577, and from table 15, R^2 = 9325.3/14996.8 = 0.622.

   (d) Comparing the VAF's tells us that the model is improved with the addition of fat to the model. This is further shown by the column Seq SS, which shows that fat reduces the SSE by 670.5, which in turn reduces the MSE, indicating less deviation between the observed and fitted values.

   (e) $H_0 : \beta_2 = 0$ and $H_A : \beta_2 \neq 0$. F is significant with $p < 0.05$, i.e. reject H_0 in favour of H_A and conclude that fat is in the population model.

2. See the attached JASP file.

   (a) Model 0: sales = β_0 + β_1 adverts + ε.
   Model 1: sales = β_0 + β_1 adverts + β_2 airplay + β_3 attract + ε.

   (b) Model 0: The sales (in thousands of copies) of each album in the week after release depends only on the amount (in thousands of pounds) spent promoting the album before release.

   Model 1: The sales (in thousands of copies) of each album in the week after release depends on the amount (in thousands of pounds) spent promoting the album before release, how many times songs from the album were played on a prominent national radio station in the week before release, and how attractive people found the band's image (on a scale from 1 to 10).

   The point is to see whether album sales depend only on how much is spent on advertising, or if there is a "organic" component to the music industry, and that album sales still depend on radio airtime and fans. It is clear to all that the music industry has progressed from this organic state to a hyper-commercialised money-making machine, and we wish to test if this formula prescribed by the big music producers is actually what drives sales (i.e. money in their pockets), or if people still care about the artists as people and that the music is subjectively good.

   (c) Table 20a gives us the descriptive statistics of all four variables. We can use this information to create confidence intervals for the population means (of each variable). Figure 3 shows the correlations between all four variables, as well as the distribution of each variable. It is evident from the first row that sales is correlated with adverts, airplay and attract, and that the latter three are uncorrelated with each other. The row for adverts shows that adverts is slightly right skewed, however we can be sure of any violations of normality by viewing the QQ-plots - we do not have enough information from the density plots to make a decision.

   (d) $H_0 : \beta_2 = 0$ and $\beta_3 = 0$.
   $H_1 : \beta_2 \neq 0$ and $\beta_3 \neq 0$.

   Note: these hypotheses talk about whether airplay and attract are **jointly significant** in the model, meaning that we must conduct an F-test. If we reject H_0, we do not know anything about the values of β_2 and β_3 in the population model, **only that they are both not zero**. It could be that $\beta_2 = 0$ and $\beta_3 \neq 0$ (attract is **individually significant** in the model, but airplay isn't), or vice versa, or maybe even that they are both **individually significant** in the model i.e. both coefficients not significantly zero in the model. In order to ascertain which of the two (or both) are non-zero, we would need to perform multiple t-tests however this presents an issue with a too large Type-I error (Bonferroni). This is why we perform an F-test first, to see whether it is worth our time to perform any t-tests.

If we find sufficient evidence in the data, we can conclude that album sales depends on the amount of money spent on advertising, the radio airtime, and the band's public image. However if there isn't sufficient evidence, then we must conclude that the album sales solely depend on the amount of money spent on advertising.

table 21a tells us that both Model 0 and 1 are "good" as the F-test shows (see the p-value), however Model 1 is "better" as shown by the R or R^2 columns. Furthermore, the adjusted R^2 column shows that there is no issue of "over-fitting", i.e. we haven't included too many predictors in the model. The R^2 change column tells us that adding airplay and attract to the model increases the VAF by 33% - the F change column is an F-test done on the VAF, which shows significant difference. All of this just says that Model 0 is ok, but Model 1 is better.

(e) Under Model 0,

$$\widehat{\text{sales}}\,(\text{adverts} = 600) = 134.140 + 0.096 \times 600 = 191.74,$$

or equivalently "spending 6,000 pounds on advertising yields an estimated 191,740 in album sales". Under Model 1,

$$\widehat{\text{sales}}\,(\text{adverts} = 600) = -26.613 + 0.085 \times 600 + 3.367\text{airplay} + 11.086\text{attract} = 24.387 + 3.367\text{airplay} + 11.086\text{attract},$$

or equivalently "spending 6,000 pounds on advertising yields an estimated 24,387 in album sales, keeping other factors constant". The value of the coefficient for adverts in either model are similar which shows that the marginal contributions are the same (confirm this with the standardised coefficient, i.e. the r between sales and adverts), however the intercepts are very different so keeping other factors constant does not provide an accurate estimate in Model 1. In Model 1, airplay and attract capture some of the variation in sales and so the marginal contribution of adverts to sales is smaller in Model 1 than in Model 0.

(f) See part (d).

(g) The VIF tells us whether adverts, airplay and attract are correlated with each other. If, for example, adverts and attract were correlated, then the R_j^2 produced from the following regression

attract $= \alpha_0 + \alpha_1$ adverts $+ \delta$

would be closer to 1 than to zero. Then, instead of including attract in the model, we would reformulate Model 1 as follows:

$$\begin{aligned}\text{sales} &= \beta_0 + \beta_1 \text{ adverts} + \beta_2 \text{ airplay} + \beta_3 (\alpha_0 + \alpha_1 \text{ adverts} + \delta) + \varepsilon \\ &= (\beta_0 + \beta_3\alpha_0) + (\beta_1 + \beta_3\alpha_1) \text{ adverts} + \beta_2 \text{ airplay} + (\varepsilon + \beta_3\delta).\end{aligned}$$

We have a new intercept term $(\beta_0 + \beta_3\alpha_0)$, a new slope term for adverts $(\beta_1 + \beta_3\alpha_1)$, and a new error term $(\varepsilon + \beta_3\delta)$. The new intercept and slope are interesting however the new error term is not interesting for us - we can rename it as $\tilde{\varepsilon}$.

sales $= (\beta_0 + \beta_3\alpha_0) + (\beta_1 + \beta_3\alpha_1)$ adverts $+ \beta_2$ airplay $+ \tilde{\varepsilon}$.

VIF is calculated as $1/(1-R_j^2)$ and the rule of thumb is that if VIF>10, then there is multicollinearity between X_j and the other predictors - a VIF>10 corresponds with an $R_j^2 > 0.9$, i.e. strong correlation. The VIF columnn in table 21c tells us that the there is no issue of multicollinearity.

(h) Model 0: sales = 0.578×adverts.

Model 1: sales = 0.511×adverts + 0.512×airplay + 0.192×attract.

(i) Table 21b tells us that the MSE is smaller in Model 1, when comparing it to the MSE in Model 0, which tells us that the addition of the predictors airplay and attract to the model has had a positive effect on the reliability of future predictions of album sales based on the data. The p-values say that the MSE is low enough for both models for us to conclude that linear regression is appropriate, however the MSE cells tell us that Model 1 is better.

College success: Linear regression and ANOVA

Description:

This data set, "College Success", provides high school grades, SAT scores, and Grade Point Average of 224 university students.

Variables:

id Participant ID.

gpa Grade Point Average (GPA) after three semesters in college.

hsm Average high-school grade in mathematics.

hss Average high-school grade in science.

hse Average high-school grade in English.

satm SAT score for mathematics.

satv SAT score for verbal knowledge.

sex Gender (labels not available).

We will examine which variables best predict GPA. First, we will fit a model predicting GPA by high school grades. Then, we will use a model that predicts GPA by SAT scores. Finally, we will fit a model that uses both high school grades and SAT scores to predict GPA.

Table 20

(a) Descriptive Statistics

	gpa	hsm	hss	hse	satm	satv
Valid	224	224	224	224	224	224
Missing	0	0	0	0	0	0
Mean	2.635	8.321	8.089	8.094	595.286	504.549
Std. Deviation	0.779	1.639	1.700	1.508	86.401	92.610
Minimum	0.120	2.000	3.000	3.000	300.000	285.000
Maximum	4.000	10.000	10.000	10.000	800.000	760.000

(b) Correlation Table

		gpa	hsm	hss	hse	satm	satv
gpa	Pearson's r	–					
	p-value	–					
	Spearman's rho	–					
	p-value	–					
hsm	Pearson's r	0.436	–				
	p-value	< .001	–				
	Spearman's rho	0.463	–				
	p-value	< .001	–				
hss	Pearson's r	0.329	0.576	–			
	p-value	< .001	< .001	–			
	Spearman's rho	0.395	0.587	–			
	p-value	< .001	< .001	–			
hse	Pearson's r	0.289	0.447	0.579	–		
	p-value	< .001	< .001	< .001	–		
	Spearman's rho	0.316	0.483	0.598	–		
	p-value	< .001	< .001	< .001	–		
satm	Pearson's r	0.252	0.454	0.240	0.108	–	
	p-value	< .001	< .001	< .001	0.106	–	
	Spearman's rho	0.272	0.446	0.237	0.110	–	
	p-value	< .001	< .001	< .001	0.102	–	
satv	Pearson's r	0.114	0.221	0.262	0.244	0.464	–
	p-value	0.087	< .001	< .001	< .001	< .001	–
	Spearman's rho	0.155	0.218	0.271	0.274	0.403	–
	p-value	0.021	0.001	< .001	< .001	< .001	–

What information can be gleamed from table 20a?

1. Do you have to omit any data entries; why/why not?
2. The coefficient of variation (CV) is a standardised measure of dispersion, and often expressed as a percentage. What is the interpretation of c_v = 29.56% for GPA? Compute, interpret, and compare the CV for all six variables. Can you do this for all six variables; why/why not?
3. What does the phrase "no meaningful zero" mean? Explain this in context with regards to one or more of the variables. Does this influence your answer from the previous question?

The correlations in table 20b were produced by JASP using which formula? Why is it important to study correlation in regression? Refer to this table for the following questions.

4. Provide a brief description of the four statistics reported in table 20b.

5. What is the difference between the two correlation statistics reported in the table? When is one more appropriate to use than the other?

6. What can you tell me about the relationship between all six variables (note: you will have to make 15 comparisons[1]).

7. Are there any relationships which you find concerning, and why are you concerned? Provide a reasonable method to solving this problem.

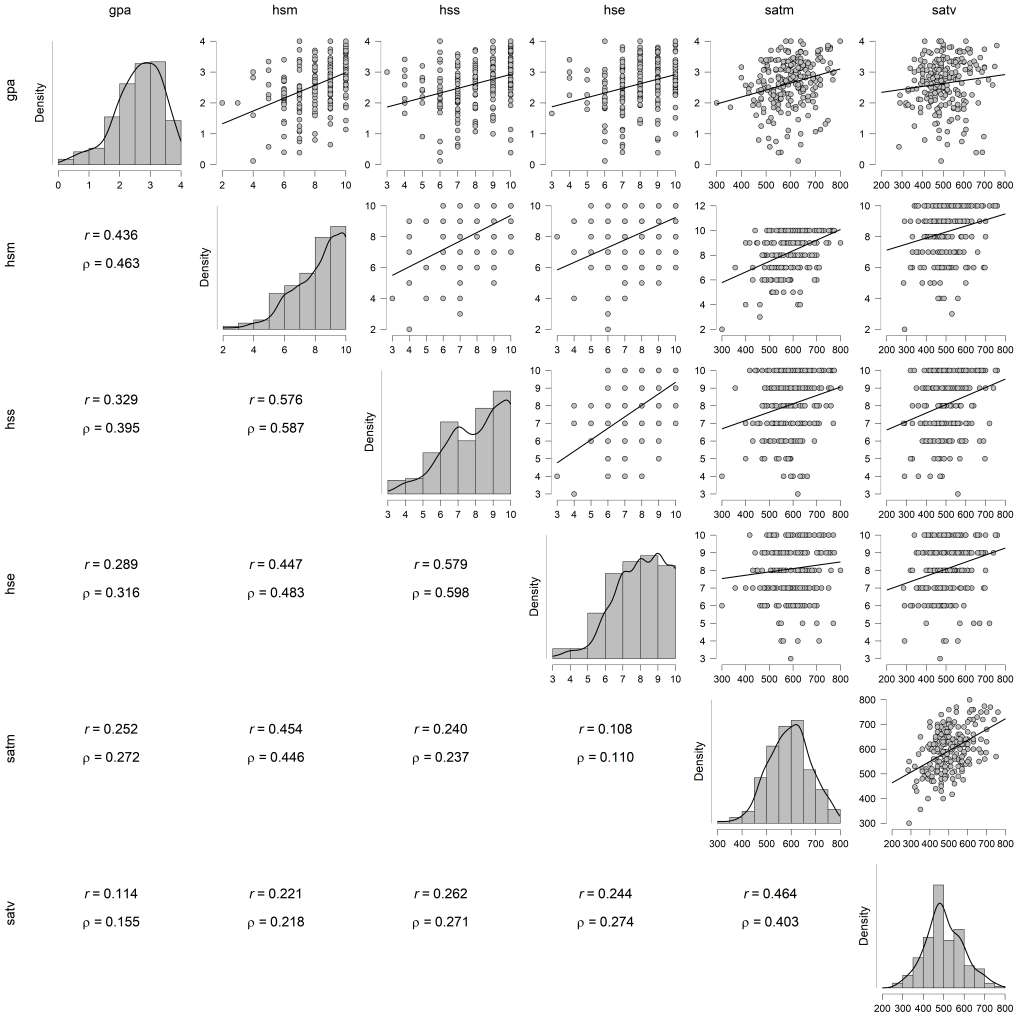


Figure 4

In fig. 4, you are provided with two point estimates and three different graphs in a matrix format, where the lower tri-diagonal contains the point estimates, and the upper tri-diagonal and diagonal contains the graphs.

8. What are the two point estimates and three graph types? Explain the difference.

9. Along the diagonal of the matrix, do you see any graphs which lead to assume there might be violations of linear regression assumptions? If yes, then which graphs might violate which assumptions?

10. With reference to your answer to the previous part, what might be a reason for the observed pattern which leads to a violation? How could you transform your data to solve these problems?

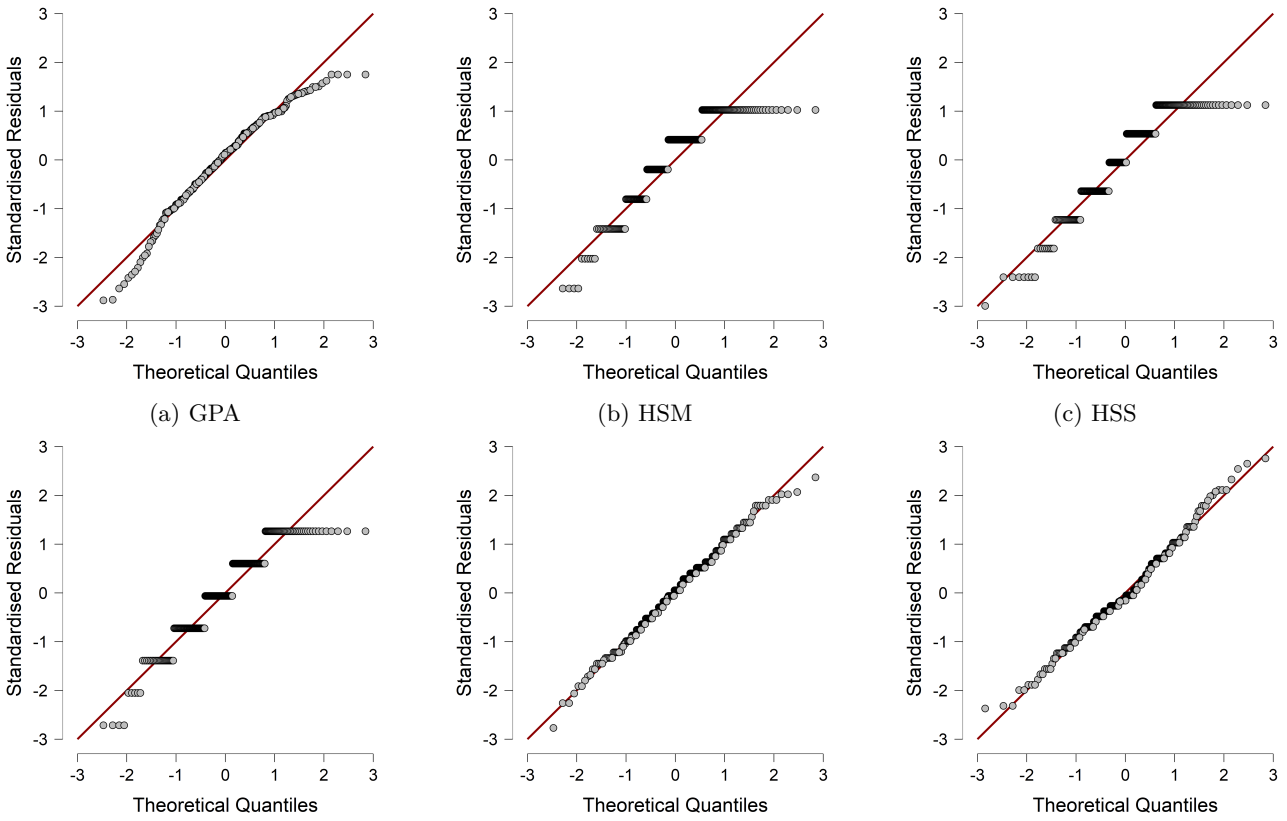
11. Compare the point estimates in fig. 4 to those in table 20b. Do these values agree?

12. Compare the point estimates on the lower tri-diagonal to the graphs on the upper tri-diagonal, and discuss. (Hint: make reference to direction and strength of relationships.)

[1]You have to make 15 comparisons because you have 6 variables and you're going to choose 2 each time to calculate the correlation, i.e. 6 choose 2 = 6!/2!*(6-2)! = 15.

13. Do the graphs on the upper tri-diagonal lead you to believe that there might be a violation of assumption/s? Explain.

14. If you were to standardise the variables, how might the graphs on the upper tri-diagonal change? Why might it be benefit to standardise your variables?



(a) GPA (b) HSM (c) HSS

(d) HSE (e) SATM (f) SATV

Figure 5

Figure 5 displays the Q-Q plots for all six variables. What does "Q-Q" mean? Answer the following questions with reference to this graph.

15. What can you infer about the six variables from fig. 5? How does this compare to your answer given in question 9?

16. What assumption/s are you checking for with a Q-Q plot? Why is this important in inferential frequentist statistics?

17. Write down two equations which represent the assumption/s you presented in the previous question.

Table 21

(a) Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.452	0.205	0.194	0.700

(b) ANOVA

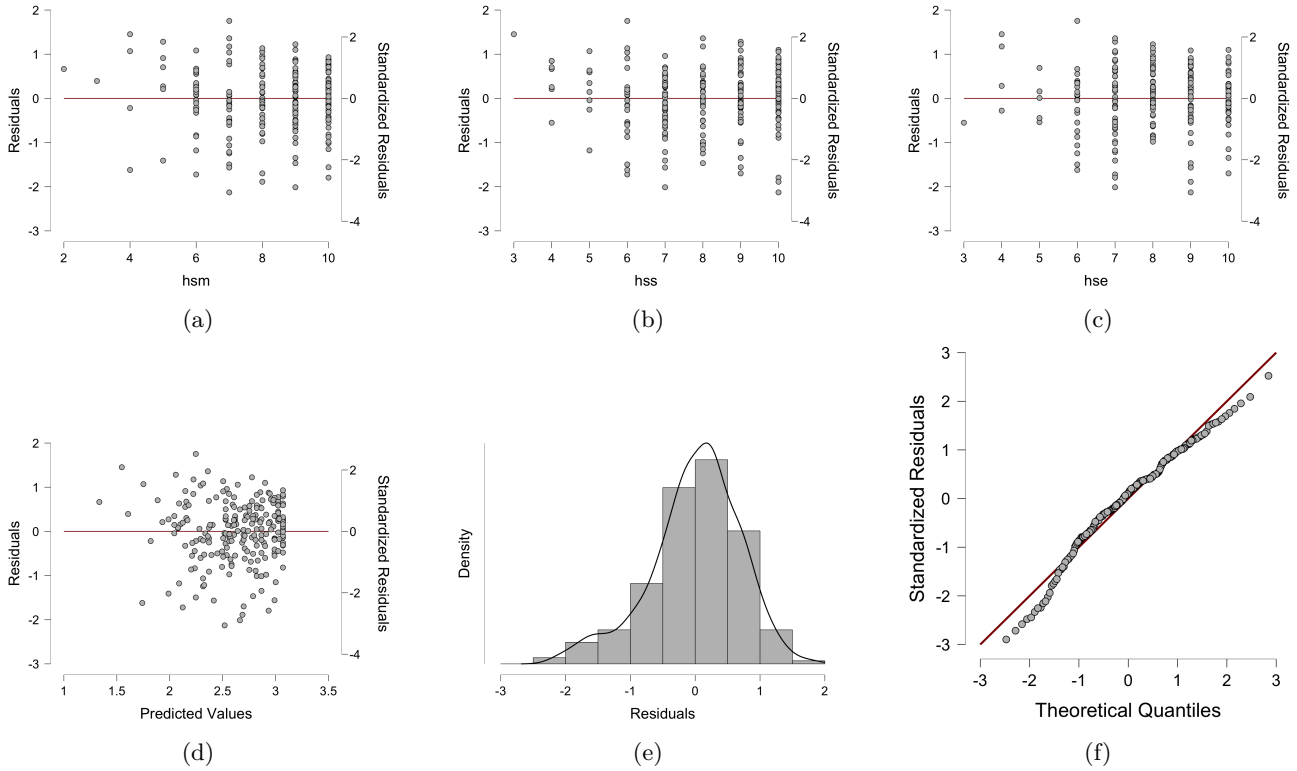
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	27.712	3	9.237	18.861	< .001
	Residual	107.750	220	0.490		
	Total	135.463	223			

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
1	(Intercept)	0.590	0.294		2.005	0.046		
	hsm	0.169	0.035	0.354	4.749	< .001	0.649	1.540
	hss	0.034	0.038	0.075	0.914	0.362	0.539	1.855
	hse	0.045	0.039	0.087	1.166	0.245	0.645	1.550

(c) Coefficients

Table 21 shows the results of the regression of high-school grades on a dependent variable. What is the population regression model?

18. Explain the four point estimates given in table 21a.
19. Using table 21b, compute the VAF and adjust accordingly. Compare with table 21a.
20. What null hypothesis might you test using table 21b? State the hypothesis/es and make a decision.
21. Explain how the values in each column of table 21b are calculated.
22. Using your population regression model and table 21c, what is the prediction equation? Compute the predicted output for a student with an average high-school mathematics grade of 70%, and explain your answer in words.
23. What can you say about the relevance of the variables in the population model that you have defined? Define hypotheses, test and explain your findings. What might you change/keep the same in the population model and why?
24. What does the VIF column of table 21c tell us? Compare these values with your responses in question 6.
25. How are the columns VIF and Tolerance related? What is meant by the term “tolerance”?
26. In the previous questions, I asked you to investigate individual and joint significance in the population model. Which type responds to which question, and what is the difference?



(a) (b) (c)

(d) (e) (f)

Figure 6

Explain the graphs in fig. 6.

27. You can use figs. 6a to 6c to check which assumption/s? What do you conclude from these graphs?

28. In fig. 6d, the standardised residuals are plotted against the predicted values. Why are the residuals standardised? What can conclude about the assumption of homoskedasticity using this graph?

29. Compare figs. 6e and 6f. How are they different/the same? What assumption are we checking for in this graph? Write the population regression equation which relates to these graphs.

30. With reference to the previous question, are you able to conclude anything about this assumption without any further information?

Now, we include also the SAT scores. Specifically, we include the high school grades in the 'null model'. Then, we add the SAT scores to the model to test whether SAT scores contribute to the prediction of GPA over and above the high-school grades.

Table 22

(a) Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
0	0.452	0.205	0.194	0.700	0.205	18.861	3	220	< .001
1	0.460	0.211	0.193	0.700	0.007	0.950	2	218	0.388

(b) ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	27.712	3	9.237	18.861	< .001
	Residual	107.750	220	0.490		
	Total	135.463	223			
1	Regression	28.644	5	5.729	11.691	< .001
	Residual	106.819	218	0.490		
	Total	135.463	223			

Note. Null model includes hsm, hss, hse

							Collinearity Statistics	
Model		Unstandardized	Standard Error	Standardized	t	p	Tolerance	VIF
0	(Intercept)	0.590	0.294		2.005	0.046		
	hsm	0.169	0.035	0.354	4.749	< .001	0.649	1.540
	hss	0.034	0.038	0.075	0.914	0.362	0.539	1.855
	hse	0.045	0.039	0.087	1.166	0.245	0.645	1.550
1	(Intercept)	0.327	0.400		0.817	0.415		
	hsm	0.146	0.039	0.307	3.718	< .001	0.531	1.884
	hss	0.036	0.038	0.078	0.950	0.343	0.532	1.878
	hse	0.055	0.040	0.107	1.397	0.164	0.617	1.620
	satm	9.436e-4	6.857e-4	0.105	1.376	0.170	0.626	1.597
	satv	-4.078e-4	5.919e-4	-0.048	-0.689	0.492	0.731	1.367

(c) Coefficients

Answer the following questions using table 22.

31. What is the population regression equation for Model 1, and what is it estimated to be?

32. Explain table 22a. Making reference to the equation you defined in the previous question, what is your null and alternative hypotheses? What kind of test do you perform?

33. Explain why some values in table 22b are similar/the same, and explain why some values are different. Why does the value of F change between the models; what does this tell us?

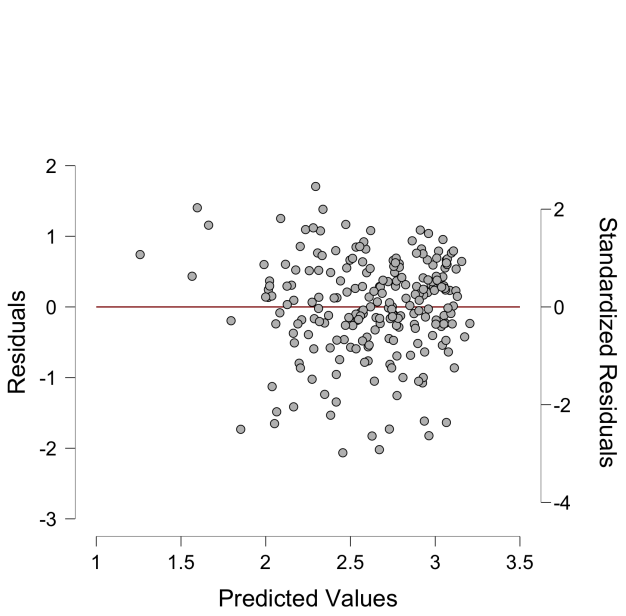
34. The estimated beta coefficients are given in table 22c. Explain the difference in the intercept and slope values.

35. What do you conclude about the individual significance of the beta coefficients in Model 1? How many tests are required for this comparison?

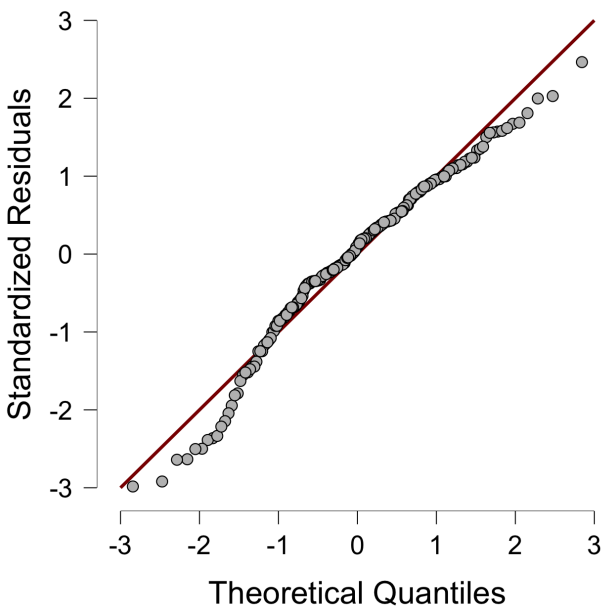
36. One of the standardised beta coefficients is negative. Why is that?

37. Are there any violations of multicollinearity present in either model? Explain.

38. Is the addition of SAT scores to the population model warranted? Why/ why not?



(a)



(b)

Figure 7

Write 4-5 lines describing why we need to examine figs. 8a and 8b and what you conclude (4-5 lines for each graph).

Solutions

1. No; all data points are valid.

2. CV shows the extent of variability in relation to the mean of the population. You do not interpret a CV alone, rather use it to compare with other variables.

$$c_v^{HSM} = \frac{1.639}{8.321} = 19.70\%; \qquad c_v^{HSS} = \frac{1.700}{8.089} = 21.02\%; \qquad c_v^{HSE} = \frac{1.508}{8.094} = 18.63\%;$$
$$c_v^{SATM} = \frac{86.401}{595.286} = 14.51\%; \qquad c_v^{SATV} = \frac{92.610}{504.549} = 18.36\%.$$

   The variables `hss`, `hse`, and `hsm` are expressed on an integer scale from 1 to 10 (interval scale), and `gpa`, `satm`, and `satv` are expressed on different ratio scales. Therefore it only makes sense to compare the c_v's of `gpa`, `satm`, and `satv`. Clearly, `gpa` has more variability than `satv`, and nearly twice as much as `satm`.

3. This phrase means that there is a value attainable by the variable which indicates the state of nothing. In this case, `satm` and `satv` have a meaningful zero point which indicates receiving no points on the exams, and `gpa` has a meaningful zero point indicating a grade average of zero. The variables `hss`, `hse`, and `hsm` do not have a meaningful zero point, as the grades start at 1.

4. Pearson's r gives the linear correlation coefficient which displays direction and strength of the relationship between two variables. The associated p-value gives the probability of observing r under the assumption that the two variables are independent. If the p-value is less than our prescribed error minimum, then we must accept that the variables are dependent.

   Spearman's ρ is the non-parametric version of Pearson's r which relies solely on the rank of paired data, i.e. if most pairs increase together then $\rho \approx 1$, however if most pairs move in opposite directions to each other then $\rho \approx -1$. However, if the data pairs do not have a majority moving together or apart, and the "movement" is random, then $\rho \approx 0$. The association p-value has the same interpretation.

5. See above. Pearson's for quantative data and Spearman's for qualitative data.

6. `gpa` is significantly (at the 5% level) correlated with all variables (look at Pearson's for `satm` and `satv` and Spearman's for `hsm`, `hss`, and `hse`), except `satv` with a p-value of ~ 8% and therefore is significant at the 10% level. `hsm` and `hss` are significantly correlated with all other variables, and `hse` fails to be significantly correlated with `satm` at the 10% level.

7. We are concerned about multicollinearity in the predictors, so it may be better to regress `hse` on `hsm` and `hss`, and then regress `satm` on `satv` so that the model is now `gpa` = β_0 + β_1 `hse` + β_2 `satm` + ε. This will also provide a partial solution to the low correlation present between `gpa` and `satv`.

8. Pearson's r and Spearman's ρ (see above). Along the diagonal are the distributions of the variables. On the upper tri-diagonal are the scatter plots between the two variables. It is clear to see that `hsm`, `hss` and `hse` have interval scale type graphs, and `gpa`, `satm` and `satv` have ratio scale type graphs. The interval scale graphs show the data points in groups of straight lines, whereas the ratio scale type graphs show a cloud of data points.

9. `gpa` seems slightly skewed to the left, however `hsm`, `hss` and `hse` are greatly skewed to the left. This seems appropriate as the grades are bounded above and grade averages tend toward the upper bound (students try to get top grades). The graphs of `satm` and `satv` seem relatively normal, however `satv` appears to have slightly positive kurtosis (sharp peak). The normality of these two graphs makes sense as these are the scores of a single test, rather than an average of grades, and therefore an even distribution is expected. The slight positive kurtosis might be best explained by the graph itself: the bar indicating the mode in the histogram is almost double in length than the other proceeding bars, i.e. the number of people who got the median score is about double the number of people for any other score. This might be due to most people having a below-average level of reading and writing (in the US).

10. See above. Instead of performing tests on means and standard deviations, you could test on medians and IQR's, or use Bayesian methods. You could try to normalise the skewed variables and ease the kurtosis with JASP or SPSS.

11. Yes.

12. All positively correlated. Steep slope indicates strong linear correlation, and flat slope indicates weak linear correlation.

13. We want the slopes of the graphs on the top line to be steep, and the graphs on the lines below to be quite flat. This is not what is observed so there may not be a strong linear relationship between the dependent and independent variables, and there might a violation of the multicollinearity assumption.

14. Standardising the variables would result in the slope of the graph being exactly Pearson's r, which allows a much quicker picture of the linear relationship than the unstandardised variables. This may also resolve slight skewing and kurtosis on the diagonal (not much, but a little).

15. Q-Q plots compare the distributions of two variables using theoretical quantiles. If the two variables have the same distribution, then the data points line up along the $y = x$ line. By assumption, the standardised residuals have the standard normal distribution, so we are determining if the variables are roughly normally distributed. We confirm our suspicions about the skewing but conclude that it is not terrible (difference is about 1). Similarly, we confirm that `satm` is normal and the kurtosis in `satv` is negligible.

16. Checking for normality, as we use the Central Limit Theorem to provide inferences about the population (assumed normal) using sample data.

17. We assume that the dependent variable is normally distributed in the population (CLT) with some mean μ_Y and some variance σ_Y^2, and that the data is sampled in a way which ensures independence between the Y_i's. We assume that the independent variables are uncorrelated with each other (multicollinearity) and uncorrelated with the error term. We assume that the the error term is normally distributed with mean zero (we expect no errors) and with the same variance as Y (variance of error does not depend on independent variables: homoskedasticity), and that the error terms are not correlated with each other.

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{\mu_Y} + \varepsilon; \qquad Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right); \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma_Y^2\right).$$

18. R is actually Pearson's r for the regression equation, i.e. combined linear correlation coefficient. R^2 is the variance accounted for (VAF) by the independent variables (of the dependent variable). Adjusted R^2 is the VAF adjusted for the number of predictors; adding predictors to the model increases R^2 whether there is any true benefit to the MSE, so R^2 is adjusted downwards relative to the number of predictors added (Wherry's). RMSE (root mean squared error) is the standard error for the estimate of R.

19. $R^2 = SSR/SST = 27.712/135.463 = 0.205$. $\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot (n-1)/(n-p-1) = 1 - (1-0.205) \cdot (223)/(223-4) = 0.194$.

20. $\texttt{gpa} = \beta_0 + \beta_1\ \texttt{hsm} + \beta_2\ \texttt{hss} + \beta_3\ \texttt{hse} + \varepsilon$.

    $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ and $\beta_3 = 0$, i.e. `hsm`, `hss`, and `hse` are jointly insignificant in the model.

    $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$, i.e. `hsm`, `hss`, and `hse` are jointly significant in the model.

    Notice the use of "and" and "or" in the hypotheses - very important. The null says that all beta coefficients are zero in the population model, i.e. dependent is uncorrelated with the independent variables. The alternative says that at least one of the beta coefficients is non-zero, i.e. at least one of the predictors is correlated with the dependent variable.

    The value of $F(3, 220) = 18.861$ is significant $(p < 0.001)$ therefore we reject the null hypothesis.

21. SS column is additive: SST = SSR + SSE. df column is additive: dfT = $n-1$ = dfR + dfE = $(p-1)+(n-p)$. MSR = SSR / dfR and MSE = SSE / dfE, and then F = MSR / MSE. $p = \mathbb{P}\left(F_{3,220} > 18.861\right)$ found using software or an F-table.

22. $\widehat{\texttt{gpa}} = 0.59 + 0.169\ \texttt{hsm} + 0.034\ \texttt{hss} + 0.045\ \texttt{hse}$. If `hsm` = 7, then we can input this into the prediction equation and set `hss` and `hse` to meaningful zeros, i.e. to their respective means: $\widehat{\texttt{gpa}} = 0.59 + 0.169 * 7 + 0.034 * 8.089 + 0.045 * 8.089 = 2.412$. We can interpret this as, on average a student with a high-school math grade of 7 out of 10 has a GPA in university of 2.412.

23. Look at the p-values: only `hsm` is individually significant.

24. Rule of thumb (SL 17; L06): VIF < 4 is good. Previously we were concerned about multicollinearity however all VIF's are below 2.

25. VIF = 1 / Tolerance, where Tolerance = 1 - R_j^2 is percentage of variance of X_j which is not explained by the other predictors.

26. Individual is t-tests on single beta coefficients. Joint is an F-test on all beta coefficients (or a range).

27. Homoskedasticity and uncorrelated error terms. No violations as the pattern of the data points is random and bounded by parallel lines.

28. The residuals are standardised so you can get a visual estimate of the variance of the residuals. The left axis gives the unstandardised residual values. Homoskedasticity assumption not violated as the pattern of the data points is random and bounded by parallel lines.

29. See the answer to question 17.

30. Yes, we conclude that the assumptions of normality (for both the dependent variable and the error term) and equal variance are not violated.

31.

$$\texttt{gpa} = \beta_0 + \beta_1\texttt{hsm} + \beta_2\texttt{hss} + \beta_3\texttt{hse} + \beta_4\texttt{satm} + \beta_5\texttt{satv} + \varepsilon$$
$$\widehat{\texttt{gpa}} = 0.327 + 0.146\texttt{hsm} + 0.036\texttt{hss} + 0.055\texttt{hse} + 0.0009346\texttt{satm} + 0.0005919\texttt{satv}$$

32. $H_0 : \beta_4 = 0$ and $\beta_5 = 0$. $H_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$. This is an F-test for joint significance. We look at the F change column and note that we do not reject our null hypothesis.

33. The SSR's are quite similar, however the dfR's are different and therefore result in different MSR's. As the SSR's are similar (and for obvious reasons the SST's are the same), the SSE's are similar and despite the difference in dfE's, the MSE's are equal. The different MSR's and equal MSE's result in different F statistics. This says that the addition of `satv` and **satm** to the model does not decrease the MSE significantly, which agrees with the answer given in the previous question.

34. The slope values do not vary greatly which suggests that the predictors in Model 0 are better than those which were included in Model 1. The intercept values change by approx. 0.2 as the inclusion of the variables has altered the slope coefficients slightly, moreover the model now has variables with zero-valued meaningful zeros (whereas Model 0's meaningful zeros are the means of the variables).

35. We require 5 t-tests for individual significance, and it is clear that only `hsm` is individually significant.

36. The linear relationship between `gpa` and `satv` is negative, i.e. those with a higher SAT verbal score usually do worse in hard-science subjects, such as maths, and high maths grades is a stronger positive contributor to university GPA. The negative correlation coefficient is quite near zero, and table 20b stated insignificant correlation between `gpa` and `satv`.

37. No. All VIF < 2.

38. No. Reject null hypothesis.

See the answer to question 17.

Semi-partial and partial correlation

Description:

This fictional data set, "Exam Anxiety", provides questionnaire scores by students prior to an exam (the variables are anxiety, preparedness, and grade).

`Revise` Time spent studying for the exam (in hours).

`Exam` Performance in the exam (percentages).

`Anxiety` Anxiety prior to the exam as measured by the Exam Anxiety Questionnaire.

(a) Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.457	0.209	0.193	23.306

(b) ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	14321.514	2	7160.757	13.184	< .001
	Residual	54315.690	100	543.157		
	Total	68637.204	102			

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	87.833	17.047		5.152	< .001
	Revise	0.241	0.180	0.169	1.339	0.184
	Anxiety	-0.485	0.191	-0.321	-2.545	0.012

(c) Coefficients

Table 23

We wish to know which independent variable (hours of revision or anxiety score prior to the exam) best describes the dependent variable (exam performance).

For convenience, Y = `exam`, X_1 = `revise`, and X_2 = `anxiety`, then $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ is the population regression equation. In order to compute the partial and semi-partial correlations, we need *other* regression equations. First, we regress X_1 and X_2 (separately) on Y, and also regress X_1 and X_2 on each other:

$$Y = \alpha_0 + \alpha_1 X_1 + e_{Y.X_1}; \qquad Y = \gamma_0 + \gamma_1 X_2 + e_{Y.X_2};$$
$$X_1 = \delta_0 + \delta_1 X_2 + e_{X_1.X_2}; \qquad X_2 = \kappa_0 + \kappa_1 X_1 + e_{X_2.X_1}.$$

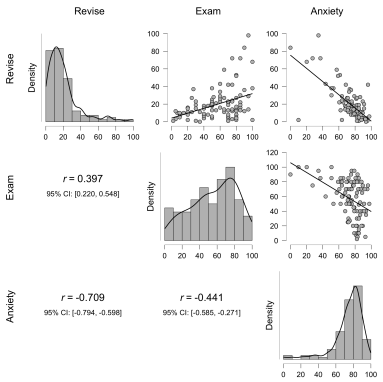


Figure 9

Then we can calculate the partial correlation coefficients as

$$pr_1 = \mathrm{Cor}\left(e_{Y.X_2}, e_{X_1.X_2}\right) = \frac{r_{Y,X_1} - r_{Y,X_2} \cdot r_{X_1,X_2}}{\sqrt{\left(1 - r^2_{Y,X_2}\right) \cdot \left(1 - r^2_{X_1,X_2}\right)}}$$
$$pr_2 = \mathrm{Cor}\left(e_{Y.X_1}, e_{X_2.X_1}\right) = \frac{r_{Y,X_2} - r_{Y,X_1} \cdot r_{X_1,X_2}}{\sqrt{\left(1 - r^2_{Y,X_1}\right) \cdot \left(1 - r^2_{X_1,X_2}\right)}}$$

Using fig. 9, calculate the partial correlation coefficients for `revise` and `anxiety`. Describe what these values represent.

The semi-partial correlation coefficients are calculated as

$$sr_1 = pr_1 \cdot \sqrt{1 - r^2_{Y,X_2}} = \frac{r_{Y,X_1} - r_{Y,X_2} \cdot r_{X_1,X_2}}{\sqrt{1 - r^2_{X_1,X_2}}}; \qquad sr_2 = pr_2 \cdot \sqrt{1 - r^2_{Y,X_1}} = \frac{r_{Y,X_2} - r_{Y,X_1} \cdot r_{X_1,X_2}}{\sqrt{1 - r^2_{X_1,X_2}}}$$

Using fig. 9, calculate the semi-partial correlation coefficients for `revise` and `anxiety`. Describe what these values represent.

If we square the partial and semi-partial correlation coefficients, we can use the Ballentine Venn Diagram to compute their values.

$$pr^2_1 = \frac{R^2 - r^2_{Y,X_2}}{1 - r^2_{Y,X_2}}; \qquad pr^2_2 = \frac{R^2 - r^2_{Y,X_1}}{1 - r^2_{Y,X_1}};$$
$$sr^2_1 = pr^2_1 \cdot \left(1 - r^2_{Y,X_2}\right) = R^2 - r^2_{Y,X_2}; \qquad sr^2_2 = pr^2_2 \cdot \left(1 - r^2_{Y,X_1}\right) = R^2 - r^2_{Y,X_1}.$$

The circles represent the variances of each variable, and how they overlap. As none of our variables are uncorrelated, all circles overlap.

In words, what do the letters *a, b, c, e* represent?
Calculate the values of the squared partial and semi-partial correlations, and explain their meaning.
Use the values *a, b, c, e* to represent the squared partial and semi-partial correlations.
Which is a better predictor of the variable `exam`, `revision` or `anxiety`? Summarise your findings in a one or two sentences.

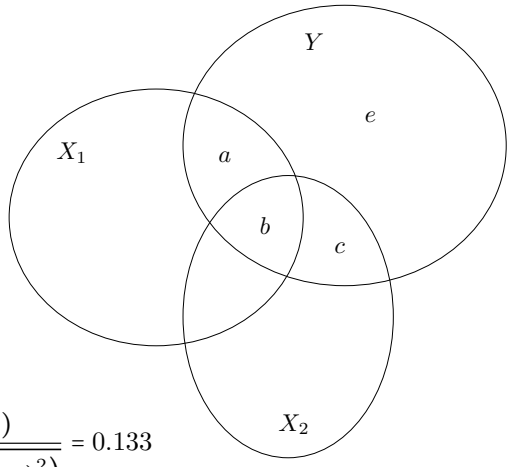


Figure 10

Solution.

$$pr_1 = \frac{0.397 - (-0.441)\cdot(-0.709)}{\sqrt{\left(1-(-0.441)^2\right)\cdot\left(1-(-0.709)^2\right)}} = 0.133$$

$$pr_2 = \frac{-0.441 - (0.397)\cdot(-0.709)}{\sqrt{\left(1-(0.397)^2\right)\cdot\left(1-(-0.709)^2\right)}} = -0.247$$

When partialing out the effect of anxiety on the other two variables, the correlation between exam results and the time spent revising is 0.133. Similarly, the correlation between exam results and pre-exam anxiety scores is -0.247 after partialing out the effect of revising on the other two variables.

$$sr_1 = 0.133\cdot\sqrt{1-(-0.441)^2} = \frac{-0.441 - (0.397)\cdot(-0.709)}{\sqrt{1-(-0.709)^2}} = 0.199$$

$$sr_2 = -0.247\cdot\sqrt{1-(0.397)^2} = \frac{-0.441 - (0.397)\cdot(-0.709)}{\sqrt{1-(-0.709)^2}} = -0.226$$

After partialling out the effect of anxiety on the time spent revising, the correlation of revision with exam results is 0.199. Similarly, the correlation between exam results and the part of anxiety unexplained by revision is -0.226.

a and b are the portions of the variance of Y which are solely explained by X_1 and X_2, respectively. b is the portion of the variance of Y explained dually by X_1 and X_2. e is the portion of the variance of Y which is not explained by X_1 nor X_2. Therefore, $R^2 = a + b + c$ is the variance of Y accounted for by the regression (VAF) and $e = 1 - R^2$ is the fraction of variance unaccounted for by the regression (FVU).

$$pr_1^2 = \frac{0.209 - (-0.441)^2}{1-(-0.441)^2} = 0.0180 = \frac{a}{a+e}; \qquad pr_2^2 = \frac{0.209 - (0.397)^2}{1-(0.397)^2} = 0.0610 = \frac{c}{c+e};$$

$$sr_1^2 = 0.0177\cdot\left(1-(-0.441)^2\right) = 0.0145 = a; \qquad sr_2^2 = 0.0610\cdot\left(1-(0.397)^2\right) = 0.0514 = c.$$

`Anxiety` solely explains 6.1% of the variance of `exam` (with `revision` partialled out), where as `revision` solely explains 1.8%. Therefore `anxiety` is a better predictor of `exam` than `revision`. We can calculate b and e in the following way: $b = R^2 - a - c = 0.209 - 0.0145 - 0.0514 = 0.1431$ and $e = 1 - R^2 = 1 - 0.209 = 0.791$, and then the percentage of variance of Y explained dually by X_1 and X_2 is $b/(b+e) = 0.1431/(0.1431 + 0.791) = 0.1532$.

First partial exam - first sample

1. What is the purpose of a simple linear regression?
   (a) To predict scores on an independent variable from scores on a single dependent variable.
   (b) To predict scores on an independent variable from scores on multiple dependent variables.
   (c) To assess whether there is a significant difference between repeated measures.
   (d) To assess whether there is a significant difference between independent groups.
   (e) To predict scores on a dependent variable from scores on multiple independent variables.
   (f) To predict scores on a dependent variable from scores on a single independent variable.
2. What is the purpose of a multiple regression?
   (a) To assess whether there is a significant difference between repeated measures.
   (b) To assess whether there is a significant difference between independent groups.
   (c) To predict scores on an independent variable from scores on multiple dependent variables.
   (d) To predict scores on a dependent variable from scores on a single independent variable.
   (e) To predict scores on a dependent variable from scores on multiple independent variables.
   (f) To predict scores on an independent variable from scores on a single dependent variable
3. What does the Adjusted R squared value tell you?
   (a) The Adjusted R squared value tells you if there is a positive relationship.
   (b) The Adjusted R squared value tells you if there is a negative relationship.
   (c) The Adjusted R squared value tells you if there is a significant difference.
   (d) The Adjusted R squared value tells you how much of the variance in the dependent variable can be accounted for by the independent variable.
   (e) The Adjusted R squared value tells you if there is a significant relationship.
   (f) None of these.
4. Which of the following points are not true when conducting a multiple regression?
   (a) Data must be free from outliers for a multiple regression.
   (b) Data must be homogeneous for a multiple regression.
   (c) The assumption of multicollinearity must be met for a multiple regression.
   (d) Multiple regression can be used to assess linear relationships.
   (e) Data must be normally distributed for multiple regression.
   (f) Multiple regression can be used to assess quadratic relationships
5. Which of these points reflect the assumption of multicollinearity?
   (a) An independent variable cannot be a combination of other independent variables.
   (b) Data must be normally distributed and not skewed.
   (c) There must not be any extreme scores in the data set.
   (d) The relationship between your independent variables must not be above $r = 0.7$.
   (e) The variance across your variables must be equal.
   (f) None of these.
6. What are residuals?
   (a) Residuals are the differences between the observed and expected dependent variable scores.
   (b) Extreme scores.
   (c) Confidence intervals.
   (d) Uncontrolled variables.
   (e) Serendipitous findings.
   (f) Left over scores

7. The assumption that the variance of the residuals about the predicted dependent variable scores should be the same for all predicted scores reflects which assumption?

   (a) Singularity.
   (b) Multicollinearity.
   (c) Normality.
   (d) Homoscedasticity.
   (e) Homogeneity.
   (f) All of these.

8. What do you report in a multiple regression to say whether your model was significant or not?

   (a) ANOVA.
   (b) Correlation.
   (c) R squared.
   (d) Chi-squared.
   (e) Beta.
   (f) Adjusted R squared.

9. What degrees of freedom do you report in a multiple regression?

   (a) Error and residual degree of freedom.
   (b) Regression and residual degrees of freedom.
   (c) Adjusted R squared and regression degrees of freedom.
   (d) Residual degree of freedom.
   (e) Regression degree of freedom.
   (f) None.

10. What does a beta of 0.478 mean?

    (a) That one model is a better predictor than another.
    (b) That the relationship between the independent and dependent variables is not linear.
    (c) This means that for every unit increase in your independent variable, your dependent variable increases by 0.478 units.
    (d) That there is no predictive power in your independent variable.
    (e) That the regression is not significant.
    (f) That the correlation is significant.

11. What is the correct format for reporting the ANOVA in a multiple regression?

    (a) N = 23, P = 0.000, F = 963.
    (b) R (12) = -78.97, $p > 0.001$.
    (c) R2 = 78%, F = 278, $p > 0.05$.
    (d) T (18) = +8.90, $p < 0.05$.
    (e) F (3, 89) = 789.34, $p < 0.001$.
    (f) None of these.

12. In a multiple regression problem involving two independent variables, what can you say about their relationship if b1 = 2.0?

    (a) The relationship between X1 and Y is significant.
    (b) The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X1, holding X2 constant.
    (c) The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X1, without regard to X2.
    (d) The estimated average value of Y is 2 when X1 equals zero.

13. What does the coefficient of multiple determination measure?

    (a) It measures the variation around the predicted regression equation.
    (b) It measures the proportion of variation in Y explained by X1 and X2.
    (c) It measures the proportion of variation in Y that is explained by X1 holding X2 constant.
    (d) It will have the same sign as b1.

14. What formula would you use to calculate the coefficient of multiple determination?

    (a) SSR/SST
    (b) SSE/SST
    (c) SSR/SSE
    (d) (SSR+SSE)/SST

15. What is adjusted r2 "adjusted" for?

    (a) The number of predictors only.
    (b) The sample size only.
    (c) The number of predictors and the sample size.
    (d) None of the above.

16. Which of the following is not a plot of residuals typically used in multiple regression analysis?

    (a) Residuals versus time.
    (b) Residuals versus X1.
    (c) Residuals versus X2.
    (d) Residuals versus correlation coefficients.

17. What is the formula for the F statistic for testing the entire regression model?

    (a) SSR/SSE.
    (b) MSE/MSR.
    (c) MSR/MSE.
    (d) MSR/SST.

18. What test would you use to test for the significance of individual regression coefficients in a multiple regression model with more than two explanatory variables?

    (a) The Z test.
    (b) The t test.
    (c) The F test.
    (d) None of the above.

19. How are the degrees of freedom associated with the multiple regression model when running a t test for the individual coefficients determined?

    (a) n-p.
    (b) n-1.
    (c) n-p-1.
    (d) n-p+1.

20. Which of the following is correct regarding the value of the adjusted r2 in a multiple regression model?

    (a) It can be negative.
    (b) It has to be positive.
    (c) It has to be larger than the coefficient of multiple determination.
    (d) It can be larger than 1.

21. How are the degrees of freedom determined for SST?

    (a) k
    (b) n-k-1
    (c) n-1
    (d) None of the above.

22. Besides the estimated regression coefficient and appropriate t statistic, what else is needed to construct a confidence interval for a regression coefficient?

    (a) The standard error of the regression coefficient.
    (b) The F statistic.
    (c) The standard error of the estimate.
    (d) The coefficient of determination.

23. In least squares regression, which of the following is not a required assumption about the error term ε?

    (a) The expected value of the error term is one.
    (b) The variance of the error term is the same for all values of x.
    (c) The values of the error term are independent.
    (d) The error term is normally distributed.

24. Larger values of r^2 imply that the observations are more closely grouped about the

    (a) average value of the independent variables
    (b) average value of the dependent variable
    (c) least squares line
    (d) origin

25. In a regression analysis if $r^2 = 1$, then

    (a) SSE must also be equal to one
    (b) SSE must be equal to zero
    (c) SSE can be any positive value
    (d) SSE must be negative

26. If the correlation coefficient is 0.8, the percentage of variation in the response variable explained by the variation in the explanatory variable is

    (a) 0.80%
    (b) 80%
    (c) 0.64%
    (d) 64%

27. If the correlation coefficient is a positive value, then the slope of the regression line

    (a) must also be positive
    (b) can be either negative or positive
    (c) can be zero
    (d) can not be zero

28. When the error terms have a constant variance, a plot of the residuals versus the independent variable x has a pattern that

    (a) fans out
    (b) funnels in
    (c) fans out, but then funnels in
    (d) forms a horizontal band pattern
    (e) forms a linear pattern that can be positive or negative

29. Consider the following regression equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. What does β_1 imply?

    (a) β_1 measures the marginal effect of x_1 on x_2.
    (b) β_1 measures the marginal effect of y on x_1.
    (c) β_1 measures the marginal effect of x_1 on y.
    (d) β_1 measures the marginal effect of x_1 on ε.

30. If the explained sum of squares is 35 and the total sum of squares is 49, what is the residual sum of squares?

    (a) 10
    (b) 12
    (c) 18
    (d) 14

31. Which of the following is true of R2?

    (a) R2 is also called the standard error of regression.
    (b) A low R2 indicates that the Ordinary Least Squares line fits the data well.
    (c) R2 usually decreases with an increase in the number of independent variables in a regression.
    (d) R2 shows what percentage of the total variation in the dependent variable, Y, is explained by the explanatory variables.

32. If an independent variable in a multiple linear regression model is an exact linear combination of other independent variables, the model suffers from the problem of

    (a) perfect collinearity
    (b) homoskedasticity
    (c) heteroskedasticty
    (d) omitted variable bias

33. Exclusion of a relevant variable from a multiple linear regression model leads to the problem of

    (a) misspecification of the model
    (b) multicollinearity
    (c) perfect collinearity
    (d) homoskedasticity

34. High (but not perfect) correlation between two or more independent variables is called

    (a) heteroskedasticty
    (b) homoskedasticty
    (c) multicollinearity
    (d) micronumerosity

35. Find the degrees of freedom in a regression model that has 10 observations and 7 independent variables.

    (a) 17

    (b) 2

    (c) 3

    (d) 4

36. True or False:

    (a) The term "linear" in a multiple linear regression model means that the equation is linear in parameters.

    (b) The key assumption for the general multiple regression model is that all factors in the unobserved error term be correlated with the explanatory variables.

    (c) The coefficient of determination (R2) decreases when an independent variable is added to a multiple regression model.

    (d) A larger error variance makes it difficult to estimate the partial effect of any of the independent variables on the dependent variable.

37. The normality assumption implies that:

    (a) the population error ε is dependent on the explanatory variables and is normally distributed with mean equal to one and variance σ^2.

    (b) the population error ε is independent of the explanatory variables and is normally distributed with mean equal to one and variance σ.

    (c) the population error ε is dependent on the explanatory variables and is normally distributed with mean zero and variance σ.

    (d) the population error ε is independent of the explanatory variables and is normally distributed with mean zero and variance σ^2.

38. A normal variable is standardized by:

    (a) subtracting off its mean from it and multiplying by its standard deviation.

    (b) adding its mean to it and multiplying by its standard deviation.

    (c) subtracting off its mean from it and dividing by its standard deviation.

    (d) adding its mean to it and dividing by its standard deviation.

39. Which of the following is a statistic that can be used to test hypotheses about a single population parameter?

    (a) F statistic

    (b) t statistic

    (c) χ^2 statistic

    (d) Durbin Watson statistic

40. Consider the equation, $Y = \beta_1 + \beta_2 X_2 + \varepsilon$. A null hypothesis, $H_0 : \beta_2 = 0$ states that:

    (a) X_2 has no effect on the expected value of β_2.

    (b) X_2 has no effect on the expected value of Y.

    (c) β_2 has no effect on the expected value of Y.

    (d) Y has no effect on the expected value of X_2.

41. The significance level of a test is:

    (a) the probability of rejecting the null hypothesis when it is false.

    (b) one minus the probability of rejecting the null hypothesis when it is false.

    (c) the probability of rejecting the null hypothesis when it is true.

    (d) one minus the probability of rejecting the null hypothesis when it is true.

42. The general t statistic can be written as:

    (a) t = hypothesized value / standard error

    (b) t = estimate - hypothesized value

    (c) t = (estimate - hypothesised value) / variance

    (d) t = (estimate - hypothesised value) / standard error

43. Which of the following statements is true of hypothesis testing?

    (a) The t test can be used to test multiple linear restrictions.

    (b) A test of single restriction is also referred to as a joint hypotheses test.

    (c) A restricted model will always have fewer parameters than its unrestricted model.

    (d) OLS estimates maximize the sum of squared residuals.

44. Which of the following statements is true?

    (a) If the calculated value of F statistic is higher than the critical value, we reject the alternative hypothesis in favor of the null hypothesis.

    (b) The F statistic is always non-negative as SSR_0 is never smaller than SSR_1.

    (c) Degrees of freedom of a restricted model is always less than the degrees of freedom of an unrestricted model.

    (d) The F statistic is more flexible than the t statistic to test a hypothesis with a single restriction.

45. True or False:

    (a) If the calculated value of the t statistic is greater than the critical value, the null hypothesis, H0 is rejected in favor of the alternative hypothesis, H1.

    (b) $H_1 : \beta_j \neq 0$, where β_j is a regression coefficient associated with an explanatory variable, represents a one-sided alternative hypothesis.

    (c) Standard errors must always be positive.

Solution. 1f, 2e, 3d, 4f, 5d, 6a, 7d, 8a, 9b, 10c, 11e, 12b, 13b, 14a, 15c, 16d, 17c, 18b, 19c, 20a (independence, too many p, or small n), 21c, 22a, 23a, 24c, 25b, 26d, 27a, 28d 29c, 30d, 31d, 32a, 33b, 34c, 35b, 36(T, F, F, T), 37d, 38c, 39b, 40b, 41c, 42d, 43c, 44b, 45(T, F, T)

20a: $\bar{R} \leq 0 \implies \begin{cases} R^2 \leq \dfrac{p}{n-1} \\ p \geq R^2(n-1) \\ n \leq 1 + \dfrac{p}{R^2} \end{cases}$

First partial exam - second sample

1. The strength (degree) of the correlation between a set of independent variables and a dependent variable is measured by
   (a) Coefficient of Correlation
   (b) Coefficient of Determination
   (c) Standard error of estimate
   (d) All of the above
2. The percent of total variation of the dependent variable explained by the set of independent variables is measured by
   (a) Coefficient of Correlation
   (b) Coefficient of Skewness
   (c) Coefficient of Determination
   (d) Standard Error of Estimate
   (e) Multicollinearity
3. A coefficient of correlation is computed to be -0.95 means that
   (a) The relationship between two variables is weak
   (b) The relationship between two variables is strong and positive
   (c) The relationship between two variables is strong and but negative
   (d) Correlation coefficient cannot have this value
4. Let the coefficient of determination computed to be 0.39 in a problem involving one independent variable and one dependent variable. This result means that
   (a) The relationship between two variables is negative
   (b) The correlation coefficient is 0.39 also
   (c) 39% of the total variation is explained by the independent variable
   (d) 39% of the total variation is explained by the dependent variable
5. Relationship between correlation coefficient and coefficient of determination is that
   (a) both are unrelated
   (b) The coefficient of determination is the coefficient of correlation squared
   (c) The coefficient of determination is the square root of the coefficient of correlation
   (d) both are equal
6. Multicollinearity exists when
   (a) Independent variables are correlated less than -0.70 or more than 0.70
   (b) An independent variables is strongly correlated with a dependent variable
   (c) There is only one independent variable
   (d) The relationship between dependent and independent variable is non-linear
7. If “time” is used as the independent variable in a simple linear regression analysis, then which of the following assumption could be violated
   (a) There is a linear relationship between the independent and dependent variables
   (b) The residual variation is the same for all fitted values of the dependent variable
   (c) The residuals are normally distributed
   (d) Successive observations of the dependent variable are uncorrelated

8. In multiple regression, when the global test of significance is rejected, we can conclude that

   (a) All of the net sample regression coefficients are equal to zero
   (b) All of the sample regression coefficients are not equal to zero
   (c) At least one sample regression coefficient is not equal to zero
   (d) The regression equation intersects the Y-axis at zero.

9. A residual is defined as

   (a) $y_i - \hat{y}_i$
   (b) Error sum of square
   (c) Regression sum of squares
   (d) Type I Error

10. What test statistic is used for a global test of significance?

    (a) Z test
    (b) t test
    (c) Chi-square test
    (d) F test

11. In multiple regression analysis, the correlation among the independent variables is termed

    (a) homoscedasticity
    (b) linearity
    (c) multicollinearity
    (d) adjusted coefficient of determination

12. In a multiple regression model, the error term e is assumed to

    (a) have a mean of 1
    (b) have a variance of zero
    (c) have a standard deviation of 1
    (d) be normally distributed

13. In order to test for the significance of a regression model involving 14 independent variables and 50 observations, the numerator and denominator degrees of freedom (respectively) for the critical value of F are

    (a) 13 and 48
    (b) 13 and 49
    (c) 14 and 48
    (d) 14 and 35
    (e) none of the above

14. A multiple regression analysis includes 4 independent variables results in sum of squares for regression of 1400 and sum of squares for error of 600. The VAF will be:

    (a) 0.300
    (b) 0.700
    (c) 0.429
    (d) 0.084
    (e) none of the above

15. There are situations where a set of explanatory variables forms a logical group. The test to determine whether the extra variables provide enough extra explanatory power to warrant inclusion in the equation is the:

    (a) complete F-test
    (b) reduced F-test
    (c) partial F-test
    (d) reduced t-test
    (e) none of the above

16. In the example of explaining a person's height by means of his/her right and left foot length, how would you treat for multicollinearity?

    (a) Eliminate the right foot variable
    (b) Eliminate the left foot variable
    (c) Eliminate either foot variable
    (d) Eliminate both feet variables
    (e) None of the above

17. Determining which variables to include in regression analysis by estimating a series of regression equations by successively adding or deleting variables according to prescribed rules is referred to as:

    (a) elimination regression
    (b) logical regression
    (c) forward regression
    (d) backward regression
    (e) stepwise regression

18. In Regression Analysis $\sum \hat{Y}$ is equal to

    (a) 0
    (b) $\sum Y$
    (c) b_0
    (d) $b_1 \sum X$
    (e) None

19. In the Least Square Regression Line, $\sum (Y - \hat{Y})^2$ is always

    (a) Negative
    (b) Zero
    (c) Non-Negative
    (d) Fractional
    (e) None

20. Which one is equal to explained variation divided by total variation?

    (a) Sum of squares due to regression
    (b) Coefficient of Determination
    (c) Standard Error of Estimate
    (d) Coefficient of Correlation

21. The best fitting trend is one for which the sum of squares of error is

    (a) Zero
    (b) Minimum (Least)
    (c) Maximum
    (d) None

22. If a straight line is fitted to data, then

    (a) $\sum Y = \sum \hat{Y}$
    (b) $\sum Y > \sum \hat{Y}$
    (c) $\sum Y < \sum \hat{Y}$
    (d) $\sum (Y - \hat{Y})^2 = 0$

23. In Regression Analysis two regression lines intersect at the point

    (a) (0, 0)
    (b) (b_0, b_0)
    (c) (X, Y)
    (d) $(\overline{X}, \overline{Y})$
    (e) None

24. In the Least Square Regression line the quantity $\sum (Y - \hat{Y})$ is always

    (a) Negative
    (b) Zero
    (c) Positive
    (d) Fractional
    (e) None

Solution. 1d, 2c, 3c, 4c, 5b, 6a, 7d, 8c, 9a, 10d, 11c, 12d, 13e, 14b, 15c, 16c, 17e, 18b, 19c, 20b, 21a, 22d, 23d, 24b, 25