Tentative solutions to TMA4255 Applied Statistics, June 6, 2011

Problem 1 Body mass index

a) The goal of the one-way ANOVA is to find out whether data from several groups have a common mean.

One-way ANOVA model:
Let Y_{ij} be the BMI or ln(BMI) for the jth individual within the ith genotype group. The one-way ANOVA model can be written as:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where μ is the overall mean and α_i is the deviation from the mean of group i to the overall mean. We also assume that the error terms, ε_{ij} are independent and normally distributed with mean 0 and variance σ^2. This also implies that the variances are the same for the different groups.

Choose to model BMI or ln(BMI)?
I would choose to model ln(BMI) since this is more in accordance with the one-way ANOVA model than choosing BMI. The reason lies mainly in the normality of the studentized residuals. We can see this from the symmetry of the distribution for ln(BMI) for each genotype. The BMI boxplots are slightly skewed to the right, since the mean is larger than the median for all genotypes. We also only see observations larger than the upper whisker and not smaller than the lower whisker. The ln(BMI) boxplots show more symmetric distributions, where means are nearly equal to medians and observations are observed both larger than the upper whisker and smaller than the lower whisker. The studentized residuals from the BMI model does not look normally distributed. The studentized residuals from the ln(BMI) model look normally distributed.

There is no obvious part of the ANOVA printout that can guide us in our choice.

b) The null hypothesis tested in a one-way ANOVA is

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k \text{ vs. } H_1 : \text{at least one } \alpha_i \text{ different}$$

Here $i = 1, ..., k$ are the groups, and in our ln(BMI) data we have $k = 3$ groups.

The letter `F` and the `DF` columns:

$$F = \frac{\frac{\text{SSA}}{k-1}}{\frac{\text{SSE}}{N-k}}$$

where SSA is the sum-of-squares for FTO, and SSE is Error in the printout. Here F is the test statistic, and under H_0 it has a Fisher distribution with two parameters. These two parameters are the degrees of freedom for the FTO-grouping $k - 1 = 3 - 1 = 2$ and the Error degrees of freedom $n - k = 4435 - 3 = 4432$. For our ln(BMI) data the null hypothesis is rejected since the p-value for the test is in the order of 10^{-5}. We conclude that the mean ln(BMI) is not the same for all the FTO-genotype groups.

The one-way ANOVA model assumes that the ε_{ij} are independent and normally distributed with mean 0 and variance σ^2. This means that $\text{Var}(Y_ij) = \sigma^2$ and the same variance is assumed for all groups. This can be tested with Bartlett's test or Levene's test of equal variances. The null hypotheis tested is:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \text{ vs. } \sigma_i^2 \neq \sigma_l^2 \text{ for at least one pair } (i, l)$$

From the residual plots we have seen that we are close to normality, and we can focus on using Bartlett's test. With a p-value of 0.25 it means that the null hypothesis will not be rejected and we find that there is not reason to believe that the variances for the different FTO genotype groups differ.

c) Individual standard deviations for the groups, s_i, and S from ANOVA:
In the one-way ANOVA we assume a common variance for the ln(BMI) in all groups. This means that we may pool together the individual empirical variances to estimate a common variance for all groups. The weighing factor is the number of observations minus one (for the group mean) in each group. Here S_i^2 is the estimator for σ^2 based on the ith group (as given in the exam text).

$$S^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{n - k} = \frac{\text{SSE}}{n - k}$$

Numerically:

$$S = \sqrt{(1678 - 1) \cdot 0.1579^2 + (2068 - 1) \cdot 0.1577^2 + (689 - 1) \cdot 0.1656^2} / \sqrt{4435 - 3} = 0.159$$

95% confidence interval for the difference in ln(BMI) between the 0 and 1 FTO genotype group:
We assume that Y_{ij} is observation j from group i and that $Y_{ij} \sim N(\mu_i, \sigma^2)$, where all observations are independent. We look at group $i = 0$ and $i = 1$ and use estimators $Y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ for μ_i and S from the ANOVA output for σ. Then

$$\frac{(Y_{0.} - Y_{1.}) - (\mu_0 - \mu_1)}{\sqrt{(\frac{1}{n_0} + \frac{1}{n_1})}S}$$

follows a t-distribution with the same number of degrees of freedom as S^2. S^2 has $n - k = 4432$ degrees of freedom. The $t_{0.025,4432}$ critical value is close to the standard normal $z_{0.025} = 1.96$ for 4432 degrees of freedom. The 95% CI then becomes:

$$[(y_{0.} - y_{1.}) \pm t_{0.025,4432}\sqrt{(\frac{1}{n_0} + \frac{1}{n_1})S}] =$$
$$[(3.1809 - 3.1920) \pm 1.96 \cdot \sqrt{(1/1678 + 1/2068)} \cdot 0.159] = [-0.021, -0.0009]$$

We may conclude that it seems that these two grops have different means (since the CI does not contain 0).

When using «one-way multiple comparisons» with Tukey's method we aim at holding the coverage probability for all three CIs (group 0 vs 1, group 0 vs 2 and group 1 vs 2) at level 95%. The critical value used for the individual confidence intervals is found theorethically using the "Studentized Range distribution" , and the Tukey method assumes the worst by focusing on the distribution of the largest difference for our three comparisons. We now see that the CI for the difference between group 0 and 1 covers 0 and may conclude that we do not have strong enought evidence to conclude that these two groups have different means. The printout from MINITAB further shows that the means from group 0 and 2 may be considered different, and also the means from group 1 and 2.

d) We use the first order Taylor approximation to find the estimated mean and standard deviation for the BMI = exp(ln(BMI)) from the estimated mean and standard deviation for the ln(BMI).

Let $X = \ln(\text{BMI})$ with $\text{E}(X) = \mu$ and $Y = \text{BMI} = \exp(X)$, so that $g(X) = \exp(X)$. The first order Taylor approximation:

$$g(X) \approx g(\mu) + g'(\mu)(X - \mu)$$

With

$$\text{E}(g(X)) \approx g(\mu)$$
$$\text{Var}(g(X)) \approx [g'(\mu)]^2\text{Var}(X)$$

Here $g'(X) = \exp(X)$.

With estimates: we have $\hat{\mu} = 3.2151$ and $\hat{\sigma} = 0.1656$.

$$\widehat{\text{E(BMI)}} \approx \exp(\hat{\mu}) = \exp(3.2151) = 24.9$$
$$\widehat{\text{SD(BMI)}} \approx \exp(\hat{\mu}) \cdot \hat{\sigma} = \exp(3.2151) \cdot 0.1656 = 4.1243$$

Problem 2 We use the χ^2 goodness of fit test, based on calculated expected frequencies using the distribution under the null hypothesis.

We need to calculate the expected frequency for each grade, which again is based on calculating the probability for each grade under the null hypothesis. Define:

- p_i expected probability for class i.
- o_i observed count in class i.
- e_i expected count in class i.

Let $X \sim N(46.3, 13.2)$.

$$\begin{aligned}
P(X \leq 68.5) &= P(Z \leq \frac{68.5 - 46.3}{13.2}) = \Phi(1.68) = 0.9535 \\
P(X \leq 58.5) &= P(Z \leq \frac{58.5 - 46.3}{13.2}) = \Phi(0.92) = 0.8212 \\
P(X \leq 44.5) &= P(Z \leq \frac{44.5 - 46.3}{13.2}) = \Phi(-0.14) = 0.4443 \\
P(X \leq 36.5) &= P(Z \leq \frac{36.5 - 46.3}{13.2}) = \Phi(-0.74) = 0.2296 \\
P(X \leq 31.5) &= P(Z \leq \frac{31.5 - 46.3}{13.2}) = \Phi(-1.12) = 0.1314
\end{aligned}$$

$$\begin{aligned}
p_A &= P(X > 68.5) = 1 - P(X \leq 68.5) = 1 - 0.9535 = 0.0465 \\
p_B &= P(X \leq 68.5) - P(X \leq 58.5) = 0.9535 - 0 - 8212 = 0.1323 \\
p_C &= P(X \leq 58.5) - P(X \leq 44.5) = 0.8212 - 0.4443 = 0.3769 \\
p_D &= P(X \leq 44.5.5) - P(X \leq 36.5) = 0.4443 - 0.2296 = 0.2147 \\
p_E &= P(X \leq 36.5) - P(X \leq 31.5) = 0.2296 - 0.1314 = 0.0982 \\
p_F &= P(X \leq 31.5) = 0.1314
\end{aligned}$$

The expected value for each grade is found as $e_i = n \cdot p_i$, with $n = 577$.

Table of probabilies, expected and observed frequencies.

Grades	A	B	C	D	E	F
Probability under the null	0.0465	0.1323	0.3769	0.2147	0.0982	0.1314
Expected	26.8	76.3	217.5	123.9	56.7	75.8
Observed	38	80	193	131	86	49

The test statistic is

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

which is approximately χ^2 distributed with $k-1$ degrees of freedom, where k is the number of groups used. We have used $k = 6$.

$$X^2 = 4.65 + 0.18 + 2.75 + 0.41 + 15.19 + 9.49 = 32.7$$

Critical value in the χ^2 distribution with 5 degrees of freedom is for $\alpha = 0.05$ equal to 11.070.

This means that we reject the null hypothesis and do not believe that the total scores are normally distributed with mean 46.3 and standard deviation 13.2.

Only assuming normality:
Estimate μ and σ^2 from data (maximum likelihood). Then X^2 is approximately χ^2 distributed with $k-2-1$ degrees of freedom. In our case we have a χ^2 distributed with 3 degrees of freedom, which has critical value for $\alpha = 0.05$ equal to 7.815. The estimated mean and standard deviation from the data are the given values 46.3 and 13.2, which means that our conclusion would not change.

Problem 3 Happiness

a) The regression coefficient for `work` is 0.4761. Assume that we look at two individuals that have scored the same values for `sex`, `love` and `money`. Further assume that one of the individuals has reported `work` to have value 1 (seeking other employment) and the other has 2 (inbetween seeking other employment and OK). Then, on average, we would expect that the happiness for the last individual is 0.4761 higher than for the first individual. Keeping the other variables fized, the effect of `work` on happiness is that happiness increases on average with 0.4761 units for every one unit increase in the `work` variable.

We can perform a t-test to see if `work` is significant in the full model (given that the other variables are present). Test statistic:

$$t = \frac{\hat{\beta}_4}{\widehat{\mathrm{Var}(\hat{\beta}_4)}} = \frac{0.4761}{0.1994} = 2.39$$

which under the null hypothesis that $\beta_4 = 0$ (vs. the two-sided alternative that $\beta_4 \neq 0$) is referred to a t-distribution with $n - k - 1 = 34$ degrees of freedom ($n = 39$ is the number of observations and $k = 4$ is the number of explanatory variables, and 1 for the intercept). We use significance level 0.05 and the critical value at level 0.025 (since two-sided test) in the t_{34} distribution is approximately 2.03 (found for 35 in the table), which means that we reject the null hypothesis.

To test if the regression is significant, that is, not all coefficients are zero, we look at the ANOVA table and the $F = 20.83$ value. Using a significance level of 0.05 this is referred to a critical value for the Fisher distribution with 4 and 34 degrees of freedom: approximately 2.64 (with 35 from the table). This means that the regression is highly significant.

Residual plots: We see no clear trend in the plot of residuals vs. fitted values, which is good. The quantile-quantile plot shows no clear deviation from normality, but at least one outlier is identified.

b) In a multiple regression the least squares estimates for the regression coefficients are found by solving a set of equations. When the explanatory variables are not orthogonally selected the value for one explanatory variable will influence the estimate of the regression coefficient for the other. In a design of experiments where explanatory variables are chosen so that they are independent of eachother (orthogonal columns) the normal equations will become uncoupled and the regression coefficient estimate for the explanatory variables will not influence eachother.

We will test a set of two regression coefficients with the aim to compare two models (model A and model B).This can be done by looking at the difference in sums of squares of regression for the two models and relate this to the error sums of squares in the largest model. It is important that model B is a reduced version of model A (i.e. model A and B both contain the explanatory variables x_3 and x_4).

Formally: let $SSR(modelA)$ be the regression sums of squares for model A and $SSR(modelB)$ be the regression sums of squares for model B. Further, $SSE(modelA)$ is the error sums of squares for the full model A. The difference in number of parameters between model A and B is $m = 2$ and $n-k-1 = 39-4-1 = 34$ is the degrees of freedom for $SSE(modelA)$. Under the null hypothesis the test statistic F follows a Fisher distribution with m and

$n-k-1$ degrees of freedom.

$$F = \frac{\frac{SSR(modelA)-SSR(modelB)}{m}}{\frac{SSE(modelA)}{n-k-1}} = \frac{\frac{93.349-89.484}{2}}{\frac{38.087}{34}} = 1.73$$

The critical value in the Fisher distribution is 3.27 at level 0.05 and the null hypothesis is not rejected. This means that model B is preferred to model A.

c) Let SSE be the sum-of-squares of error, SSR be the regression sum-of-squares, and SST be the total som of squares. Then R^2: coefficient of multiple determination is defined as

$$1 - SSE/SST = SSR/SST$$

and is interpreted as the amount of variability in the data that is accounted for by the regression. R^2 will increase when a regressors are added to the model, even if the new regressors are independent of the response. The R^2_{adj} is constructed to also include information about the number of parameters estimated and the number of observations in the data set.

$$R^2_{\text{adj}} = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}}$$

It is possible that R^2_{adj} takes on a negative value if $\frac{SSE}{SST} > \frac{n-k-1}{n-1}$. R^2_{adj} can not be given an easy interpretation, other than a penalized version of R^2. It is not wise to base model selection on R^2, but R^2_{adj} can be used. Then we may choose the model with the largest R^2_{adj}.

In the table with results from fitting the 15 models the model with the largest R^2_{adj} is in row 13, where `money`, `love` and `work` are included in the regression model.

Other things to consider if you were to analyse these data:

- Include interaction terms in the model, maybe the effect of `sex` is different for the different levels of `love`?
- Look at scatter plots, or crosstabulations, beween the regressors to assess the correlation structure.
- Look at other model selection criteria.
- Evaluate model fit using residual plots.

From the table we see that the best model (according to R^2_{adj}) with one regressor is `love`. If we are to include two regressors the best model is the one including both `love` and `work`. So, following this path of reasoning `love` and `work` are the important factors. But, this data set is only of size 39 and are based on employed MBA students from Chicago. Using this sample from a population of employed MBA students from Chicago to draw conclusions about a general population (e.g. students in Trondheim) might be a questionable strategy.