Solutions MVE190-MSG500 Linear Statistical Models, 17/01/2019

1. (notice, the following are based on computations made using 3 decimal digits as provided in the exam script, hence results are approximate compared to using a software)

   (i) An unbiased estimator is $s^2 = \sum_{i=1}^{n} e_i^2/(n-p)$. Here $p = 3$ and $n = 12$ and the sum of residuals is given in Table 1, hence $s^2 = 139.913/(12-3) = 15.547$.

   (ii) $SE(\hat{\beta}_j) = s\sqrt{(\boldsymbol{X'X})^{-1}_{jj}}$. Where $j = 1, 2, 3$. Hence by taking the square root of the second and third element on the diagonal of $(X'X)^{-1}$, we have $SE(\hat{\beta}_2) = 3.943 * \sqrt{0.008} = 0.353$ and $SE(\hat{\beta}_3) = 3.943 * \sqrt{0.002} = 0.176$.
   Confidence intervals (CI) are given by $[\hat{\beta}_j \pm t_{\alpha/2,n-p} SE(\hat{\beta}_j)] = [\hat{\beta}_j \pm t_{1-\alpha/2,n-p} SE(\hat{\beta}_j)]$. We have that the 95% CI for β_2 is $[0.196 \pm 2.262 \cdot 0.353] = [-0.602, 0.994]$. Similarly for β_3 we have $[0.191 \pm 2.262 \cdot 0.176] = [-0.207, 0.589]$. Both intervals include the zero. The model does not consider the effect of height and weight jointly significant in explaining the response. It might be that one or the other are separately relevant, but not when they are jointly in the model. To be discussed in later questions.

   (iii) Let's call "raw residuals" the usual e_i. The plot of raw residuals vs i only points to two observations that seem to be outliers, though this can better be diagnosed via studentised residuals; otherwise the plot does not point to anything problematic. The plot of raw residuals vs the fitted responses seem to show a larger variation for observations having a smaller fitted response. However the small size of the dataset is not helping in deciding whether there is any systematic trend to fix.

   Leverage values show two observations that stand out (2 and 6), these being large according to the empirical criterion $h_{ii} > 2p/n$, likely because child 2 has large height and child 6 because it is tall compared to the weight. Other three observations are also standing out (8, 10, and 12), but given that the dataset is small we should perhaps not to try to over-interpret this too much.

   Studentised residuals shows that there are two observations that are fitted badly, namely observations 8 and 11 standing outside the values [-2,2]. Child 8 has considerably low weight, low height and low "length" (compared to the others) and child 11 also has low weight, compared to other kids having the same height and "length".

   All in all the Cook's distance (that depends on leverage and standardised residuals) shows that the observations that are most influential are child 6 and 8.

   Finally, from the pairs plot we observe a worringly high correlation among the predictors. One of the two has to go away. These are causing the large standard errors that are making the two effects of the covariates not-significant (as seen in ii). Clearly, by looking at the plot, both are separately informative for the response.

   (iv) We run a partial F test. All ingredients are available, since we need to compute $Q = (SS(Error)_{reduced} - SS(Error)_{full})/k/(SS(Error)_{full}/(n-p))$. The reduced model is the one with height and intercept, while the full model has also the weight covariate. So $SS(Error)_{reduced} = 160.665$ and $SS(Error)_{full} = 139.913$. Here $k = 1$ and hence $Q = 1.335$ which has to be compared with the quantile $F_{k,n-p} = F_{1,9} = 161.448$ and $Q < F_{1,9}$. We thus fail to reject the smaller model (corresponding to

testing $H_0 : \beta_3 = 0$), that is we do not need the information carried by weight, when the height is already in the model. This was expected because the pairs plot show that each of those is highly informative for the response, but also that the covariates are highly correlated. Therefore one of the two is redundant.

2. (i) The hourly wage is 67 cents (0.67 dollars) higher per year of education, conditionally on all other variables kept as constant. Work experience is associated with an increase of 9 cents in the hourly wage per year of experience, everything else being the same. Union members earn \$1.52 more per hour than non-union members, everything else being the same.

   (ii) There is a quite strong gender gap unfortunately. Female workers earn on average \$1.85 per hour less than males with the same education, work experience, union membership, region of residence and occupation. This figure is highly significant. Regarding occupation: all estimated coefficients are to be interpreted in terms of comparison with occupations of type "managerial". For example, compared to managerial positions, those working into sales earn about 4 dollars less per hour (significant figure); those working as clerks earn about 3.3 dollars less per hour (significant), etc. Those having a job of type "professional" we cannot say they earn significantly less than managers as the coefficient is not significant at 5% level. That is, data do not suggest that there is a statistically significant difference in hourly wage between professional and management jobs.

   (iii) There appears to be the "trumpet" shape denoting heteroskedasticity. That is the assumption of costant variance of the error term is markedly violated, with larger variance corresponding to larger estimated wages. This has consequences on OLS inference (which relies on the assumption of homoscedasticity), thus making the standard errors wrong, the estimated SE is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on, which means results discussed in (ii) could be false. In addition, we notice a quite large outlier.

   (iv) we are fitting $E(\log(\text{wages})) = \beta_0 + \beta_1 \cdot \text{education} + \cdots + \beta_{10} \cdot \text{female}$.
   Regarding the effect of education: all other variables held constant, we have that, when we increase the education by 1 year, we have $E(\log(\text{wages})|education + 1) = \beta_0 + \beta_1 \cdot (\text{education} + 1) + \cdots + \beta_{10} \cdot \text{female}$, which means that

$$\begin{aligned} E(\text{wages}|education + 1) &= e^{\beta_1} \cdot e^{\beta_0 + \beta_1 \cdot \text{education} + \cdots + \beta_{10} \cdot \text{female}} \\ &= e^{\beta_1} \cdot E(\text{wages}|education). \end{aligned}$$

   Therefore

$$\frac{E(\text{wages}|education + 1)}{E(\text{wages}|education)} = e^{\beta_1}$$

   Hence, since $\hat{\beta}_1 = 0.069474$ and $\exp(0.069474) = 1.072$ we conclude that each additional year of education increases the hourly wage by about 7.2%. A similar reasoning shows that each additional year of work experience increases the hourly wage by $\exp(0.0106) = 1.0106$, i.e. by about 1.06%.

3. (i) We conduct a Wald test (we could also construct a confidence interval) at significance level of 0.05 to test $H_0 : \beta_{black} = 0$ versus $H_1 : \beta_{black} \neq 0$. So we have that, under H0 assumed true, $Z = (\hat{\beta}_{black} - 0)/SE(\hat{\beta}_{black}) \approx N(0, 1)$. The observed value of Z is $-0.306/0.099 = -3.091$. Since $|Z|$ is larger than the quantile $|z_{\alpha/2}| = 1.96$ we reject H0. This implies that there exists a difference in the number of visits to the doctor for people of black ethnicity compared to white ones. Namely black subjects visit the doctor less often (negative coefficient) and specifically $\exp(-0.306) = 0.736$ implying that they report 27% fewer visits than white subjects.

(ii) We construct a likelihood ratio test for comparing a reduced vs a larger model (the concerned model are nested).The likelihood ratio test is based on a difference between deviances. The summary functions for both models report that $2\log L = -1.9658.633$ for the larger model and $2\log L = -19681.19$ for the smaller one. We take the difference of the two deviances and we obtain $D_{diff} = 19681.19 - 19658.633 = 22.577$. The two models differ by one parameter hence $D_{diff} \sim \chi^2_1$ (asymptotically as n goes to infinity), which has quantile $\chi^2_{1-\alpha,1} = \chi^2_{0.95,1} = 3.841$. We have $22.577 > 3.841$ hence we reject $H_0 : \beta_{income} = 0$ and the income covariate is a useful addition to a model having the stated covariates.

(iii) By using the provided definition of negative binomial pmf, we have that when $y = 0$ (zero doctor visits) we obtain

$$P(Y = 0) = \frac{\Gamma(\theta)}{\Gamma(1)\Gamma(\theta)} \cdot \frac{1}{(1 + \mu/\theta)^\theta} = (1 + \mu/\theta)^{-\theta}$$

since for x integer we have $\Gamma(x) = (x-1)!$ and since we assume $0! = 1$. Therefore we just need to evaluate $\mu = \exp(\beta_0 + \beta_{private} + \beta_{female} + 10 \cdot \beta_{income})$, which after fitting becomes $\hat{\mu} = \exp(-0.201 + 0.809 + 0.544 + 10 \cdot 0.0037) = 3.284$.

Since $\hat{\theta} = 0.588$ we finally have that the estimated probability evaluates to $(1 + 3.284/0.588)^{-0.588} = 0.330$, that is 33%.

Q 4 (i)

FROM LECTURES NOTES 2 (PAGE 3)

WE HAVE THAT $\hat{\beta}_1$ CAN BE WRITTEN AS (THIS IS JUST ONE OF THE POSSIBLE WAYS)

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_{i=1} (x_i - \bar{x})^2}$$

AND IN THE NOTES WE PROVE THAT

$$E(\hat{\beta}_1) = \beta_1 \quad \forall \beta_1 \quad \text{(UNBIASEDNESS)}$$

$$\text{VAR}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$$

THEN, $\hat{\beta}_1$ IS CLEARLY A LINEAR COMBINATION OF GAUSSIANS (SINCE EVERYTHING INVOLVING THE X VALUES IS DETERMINISTIC, WHILE $Y_i | X_i$ IS GAUSSIAN) HENCE $\hat{\beta}_1$ HAS GAUSSIAN DISTRIBUTION

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 / \sum_j (x_j - \bar{x})^2\right)$$

THE PRECISION OF THIS ESTIMATOR CAN BE AFFECTED IN MULTIPLE WAYS:

FOR EXAMPLE IT IS MORE PRECISE IN ESTIMATING β_1 WHEN ITS VARIANCE IS SMALL:

THIS HAPPENS IF E.G. σ^2 IS SMALL (MORE INFORMATIVE DATA)

IT ALSO HAPPEN WHE THE DATA SIZE m INCREASES, SINCE IN THAT CASE $\sum_j (x_j - \bar{x})^2$ INCREASES.

ALSO, $V(\hat{\beta}_1)$ DECREASES WHEN THE SPREAD FOR X (i.e. $\sum_i (x_j - x)^2$ INCREASES (X DISTRIBUTED OVER SPREAD VALUES)

Q4 (ii):

AN HYPOTHETICAL FUTURE RESPONSE Y_0 AS SUCH CONTAINS A NOISE TERM ϵ_0, WHICH WE ASSUME $\epsilon_0 \sim N(0, \sigma^2)$.

WE CAN ESTIMATE Y_0 BY FITTING THE CURRENTLY AVAILABLE DATA, OBTAIN $\hat{\beta}$, AND WRITE

$$\hat{Y}_0 = X_0^T \hat{\beta} + \epsilon_0$$

SINCE $\hat{\beta}$ IS GAUSSIAN AND ϵ_0 IS GAUSSIAN, Y_0 IS ALSO GAUSSIAN WITH $E(\hat{Y}_0) = X_0^T E(\hat{\beta}) + E(\epsilon_0)$

$$= X_0^T \hat{\beta} + 0 = \hat{X}_0^T \hat{\beta}$$

WE HAVE THEN:

$$\mathrm{Var}(\hat{Y}_0) = \mathrm{Var}(X_0 \hat{\beta} + \epsilon_0)$$

$$= X_0 \mathrm{Var}(\hat{\beta}) X_0^T + \mathrm{Var}(\epsilon_0) \quad (\mathrm{Var}(\epsilon_0) = \sigma^2)$$

$$\mathrm{Var}(\hat{\beta}) = (\text{SEE SLIDES\_5 OR NOTES}) = \sigma^2 (X^T X)^{-1}$$

HENCE $\mathrm{Var}(\hat{Y}_0) = \sigma^2 X_0 (X^T X)^{-1} X_0 + \sigma^2 = \sigma^2 \left(1 + X_0 (X^T X)^{-1} X_0\right)$

$$\hat{Y}_0 \sim N\left(X_0^T \beta,\ \sigma^2 \left(1 + X_0 (X^T X)^{-1} X_0\right)\right)$$

PI AT SIGNIF. LEVEL α: $PI = \left(\hat{Y}_0 \pm t_{\alpha/2,\, n-p}\, SE(\hat{Y}_0)\right) =$

$$= X_0 \hat{\beta} \pm t_{\alpha/2,\, n-p} \cdot \sqrt{1 + X_0 (X^T X)^{-1} X_0}$$

A PI REPRESENTS THE UNCERTAINTY OF PREDICTING A SINGLE FUTURE RESPONSE, USING INFERENCE BASED ON CURRENT OBSERVATIONS. IT PREDICTS THAT THE INTERVAL SHOULD CONTAIN WITH FREQUENCY $(1-\alpha)\cdot 100$ FUTURE POTENTIAL OBSERVATIONS.

INSTEAD A CI FOR THE MEAN $E(Y|x_0)$ GIVES AN UNCERTAINTY FOR THE ESTIMATION OF THE MEAN OF THE THEORETICAL POPULATION (WHEN $X=x_0$), NOT A FUTURE OBSERVATION (WHICH CONTAINS ADDITIONAL NOISE) FROM SAID POPULATION.

Q4 (iii):

$$R^2 = \frac{SS(\text{REGRS})}{SS(\text{TOTAL})} \in [0,1] \qquad \text{ALSO } R^2 = 1 - \frac{SS(\text{Error})}{SS(\text{TOT})}$$

$$SS(\text{TOT}) = \sum_i (y_i - \bar{y})$$

$$SS(\text{REGR}) = \sum_i (\hat{y}_i - \bar{y})$$

$$SS(\text{Error}) = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

R^2 GIVES THE FRACTION OF THE TOTAL OBSERVED RESPONSE VARIABILI... THAT IS "EXPLAINED" BY THE REGRESSION MODEL. THE LARGER THE BETTER.

GÖTEBORGS UNIVERSITET

A SMALL R^2 DOES NOT MEAN THAT THERE IS NO RELATIONSHIP BETWEEN Y AND THE ASSUMED MODEL. ONLY THAT THERE IS NO LINEAR RELATIONSHIP.

IN FACT IT IS ALSO TRUE THAT $R^2 = \text{Corr}^2(Y, \hat{Y})$

WHERE $\text{Corr}(\cdot, \cdot)$ IS THE PEARSON'S (LINEAR) CORRELATION COEFFICIENT.