Formula sheet for the exam in Statistical modeling in logistics (MMS075), March 16, 2020

Simple linear regression

Model equation for simple linear regression:
$Y = \beta_0 + \beta_1 X + \varepsilon$

Predicted response at a given value x of predictor X:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Observed data in form of (predictor, response) pairs:
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

i-th predicted response, residual and residual squared:
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad e_i = y_i - \hat{y}_i, \quad e_i^2 = (y_i - \hat{y}_i)^2$

Residual sum of squares and residual standard error:
$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2, \quad \mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2}$

Approximate confidence intervals for the coefficients when sample size $n \geq 30$:
$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$ and $\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$

Total sum of squares:
$\mathrm{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$, where $\bar{y}$ is the average of the observed responses

Proportion of variability in the response that is explained by the predictor:
$R^2 = \frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$

Multiple linear regression

Model equation for multiple linear regression with $p \geq 2$ predictors:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$

Predicted response at given values $x_1, x_2, \ldots, x_p$ of predictors $X_1, X_2, \ldots, X_p$:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$

Sample of n observations, containing values for each predictor and the response:
$(x_{1,1}, x_{1,2}, \ldots, x_{1,p}, y_1), (x_{2,1}, x_{2,2}, \ldots, x_{2,p}, y_2), \ldots, (x_{n,1}, x_{n,2}, \ldots, x_{n,p}, y_n)$

i-th predicted response:
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}$

i-th residual:
$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_p x_{i,p}$

Residual sum of squares and residual standard error:
$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2, \quad \mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$

Total sum of squares:
$\mathrm{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$, where $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$

Proportion of variability in the response explained by the model: $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$

Test the null hypothesis that all coefficients are zero:
$H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$
H_a : at least one β_j is not zero
Compute F-statistic: $F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$

Test relationship of X_j with the response in the presence of all other predictors:
$H_0 : \beta_j = 0$
$H_a : \beta_j \neq 0$
Compute t-statistic: $t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$

Variance inflation factor for predictor j: denoting the R^2 value for the linear regression model predicting X_j using all other predictors by $R^2_{X_j|X_{-j}}$,
VIF$= \frac{1}{1-R^2_{X_j|X_{-j}}}$

Threshold for the identification of high leverage points: $2(p+1)/n$

Logistic regression

Model equation for binomial logistic regression with $p \geq 1$ predictors:
$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$
where $p(X) = \Pr(Y = 1|X)$.

Predicted probability of a "case" at values $x_1, x_2, \ldots, x_p$ of predictors $X_1, X_2, \ldots, X_p$:
$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_p x_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_p x_p}}$

Multinomial logistic regression

Model equation for multinomial logistic regression with $p \geq 1$ predictors:
$\log\left(\frac{\Pr(Y=k|X)}{\Pr(Y=s|X)}\right) = \beta_{0,k,s} + \beta_{1,k,s} X_1 + \beta_{2,k,s} X_2 + \cdots + \beta_{p,k,s} X_p$

Training error and test error

Training mean squared error for numerical response:
$MSE = \frac{(y_1-\hat{f}(x_1))^2+(y_2-\hat{f}(x_2))^2+\ldots+(y_n-\hat{f}(x_n))^2}{n}$
Test mean squared error for new data $(x_1^{\text{new}}, y_1^{\text{new}}), (x_2^{\text{new}}, y_2^{\text{new}}), \ldots$:
$\text{Average}\left[(y^{\text{new}} - \hat{f}(x^{\text{new}}))^2\right]$

Training error rate for categorical response:
$\text{Error rate} = \frac{I(y_1 \neq \hat{y}_1)+I(y_2 \neq \hat{y}_2)+\ldots+I(y_n \neq \hat{y}_n)}{n}$
Test error rate for new data $(x_1^{\text{new}}, y_1^{\text{new}}), (x_2^{\text{new}}, y_2^{\text{new}}), \ldots$:
$\text{Average}\left[I(y^{\text{new}} \neq \hat{y}^{\text{new}})\right]$