Homework 12

Stat061-F23

Prof Amanda Luby

---

1. For the simple linear regression case ($y = \beta_0 + \beta_1 x + \epsilon$), show that $\hat{\beta}_1 = r\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$.
2. Assuming the standard multiple linear model ($Y = X\beta + \epsilon$, where X is an $n \times p$ design matrix):
   (a) Show that $\sigma^2 I = \Sigma_{\hat{y}} + \Sigma_{\hat{\epsilon}}$
   (b) Using (a), conclude that $n\sigma^2 = \sum Var(\hat{Y}_i) + \sum Var(\hat{\epsilon}_i)$
3. Consider a multiple linear regression problem with design matrix $\mathbf{X}$ and observations $\mathbf{Y}$. Let $\mathbf{X}_1$ be the matrix remaining when at least one column is *removed* from $\mathbf{X}$. (So $\mathbf{X}_1$ is the design matrix for a linear regression on $\mathbf{Y}$ but with fewer predictors). Show that R^2 (non-adjusted) for the regression model calculated using design matrix $\mathbf{X}$ is *at least as large* as the R^2 for the regression model using design matrix $\mathbf{X}_1$.
4. ~~Problem from Monday~~
5. Suppose we observe data $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$, where each Y_i represents a count and has mean μ_i. (e.g. the answer to "How many devices do you own that can access the internet?" or "How many children do you expect to have in your lifetime?). Since counts are always positive, we often use a *log-linear model* to model the mean of the Y_i's: $\log \mu_i = \beta_0 + \beta_1 x_i$. The Poisson loglinear model additionally assumes that the counts are independent poisson random variables: $Y_i \sim Pois(\mu_i)$.
   (a) Show that the log-likelihood (in terms of the Poisson parameters μ_i) is $l(\mu) = \sum[y_i \log(\mu_i) - \mu_i - log(y_i!)]$
   (b) Substitute the linear model component to show that part (a) is equivalent to $l(\beta) = \beta_0 \sum y_i + \beta_1 \sum y_i x_i - \sum \exp(\beta_0 + \beta_1 x_i) - \sum \log(y_i)!$.
   (c) Explain why the sufficient statistics for the model parameters (β_0 and β_1) are $\sum y_i$, $\sum x_i y_i$.
   (d) Show that the likelihood equation solutions have the form $\sum y_i = \sum \mu_i$ and $\sum y_i x_i = \sum \mu_i x_i$
   (e) We could also think about simply transforming the Y variables and fitting a linear regression for $\log(Y)$. The transformed-data approach uses a linear predictor for $E(\log(Y))$ whereas the GLM approach uses a linear predictor for $\log E(Y)$. Explain why these are not the same, and state an advantage of using the GLM approach if we are truly interested in modeling $E(Y)$.
6. Please fill out two "exit surveys" before the final exam. I like to split these up into two components:
   (a) A self-reflection which is *not anonymous*: https://forms.gle/CQfvjj8RYy6Pf1pY7 , which give you a chance to reflect on your growth and challenges this semester. I read these before assigning final participation grades in the course, so it is also a chance to share anything you think that I should be aware of.
   (b) A course evaluation which is *anonymous*: https://forms.gle/CnNTRwNhansEoHZTA . This contains a few department-wide assessment questions, along with questions about Stat61 and questions I include after every class. Your opinion is important to me as it helps me consider ways to implement changes, or things to keep in my next iteration of Stat061.