APPLIED STATISTICS, MATH10096

Workshop 1

1. In the following, assume, where appropriate, either observations $\{y_i\}_{i=1}^n$ or $\{(y_i, x_i)\}_{i=1}^n$.

   (a) For the model
   $$y_i = \beta + \epsilon_i, \quad i = 1, \ldots, n,$$
   determine the least squares estimate of β and show that the model residuals will sum to zero.

   (b) For the model
   $$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$
   determine the least squares estimates of α and β and show that the model residuals will sum to zero.

   (c) For the model
   $$y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$
   determine the least squares estimate of β. Will the residuals generally sum to zero for the least squares estimated model?

   (d) Bearing in mind the results from the previous parts of this question, deduce the condition under which the residuals of a linear model will sum to zero.

2. Which, if any, of the following common linear model assumptions are required for $\hat{\boldsymbol{\beta}}$ to be unbiased: (i) the error terms are independent, (ii) the error terms all have the same variance, (iii) the error terms are normally distributed?

3. In Section 4.1 of the notes, when introducing the QR decomposition, it was claimed that the squared Euclidean length of the vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is unchanged if multiplied by $\mathcal{Q}^\mathsf{T}$ (with $\mathcal{Q}$ being an orthogonal matrix of dimension $n \times n$). Proof this claim. **Note**: This should take approximately one minute.

4. Suppose that you fit a linear model to some response data, $\mathbf{y}$ and obtain residuals, $\hat{\boldsymbol{\epsilon}}$. Now suppose that you fit exactly the same model but with this $\hat{\boldsymbol{\epsilon}}$ vector treated as the response. What would the fitted value vector be for the second model fit?

5. Sections 4.1 to 4.3 of the notes derive a number of important linear model related results: (i) the form of the general least squares estimator, $\hat{\boldsymbol{\beta}}$; (ii) that $\hat{\boldsymbol{\beta}}$ is unbiased; (iii) the distribution of $\hat{\boldsymbol{\beta}}$; (iv) the t-statistic distributional result and (v) the F-ratio distributional result.

   If the constant variance assumption of the linear model does not hold, which of these results would still hold? Explain your answer.

6. In section 4.3.1 of the notes it was shown that
   $$\hat{\sigma}^2 = \frac{\|\mathbf{r}\|^2}{n - p}$$
   is an unbiased estimator of the residual variance σ^2. The proof there used the fact that the r_i are i.i.d. $N(0, \sigma^2)$ random variables, but it is not actually necessary to assume normality. As an alternative, show that $\hat{\sigma}^2$ is unbiased by considering only $\mathbb{E}(r_i^2)$.

7. In class we have analysed the Hubble constant data and, based on physical grounds, we did not include an intercept term in the linear model. Refit the model, now with an intercept, and check if the statistical evidence aligns with the physics/cosmology evidence.

8. The `grain` data available in Learn are from a 2007 study investigating the relationship between alcohol yields from distillation and the nitrogen content of the wheat grain used in the distillation, based on wheat grown at 4 sites in the UK. This is a subject of considerable interest to the whisky industry and to the biofuel industry. The `grain` data frame has 3 columns: `nitrogen` is % nitrogen (by weight), `alcohol` is alcohol yield in Litres per Tonne, and `location` indicates location of growing site.

   Whether or not yield really changes with nitrogen content can be investigated with the aid of the following linear model.

   $$\texttt{alcohol}_i = \alpha + \beta \texttt{nitrogen}_i + \epsilon_i$$

   Imagine that you are the consultant statistician asked, by the experimenters, to analyse the data.

   (a) What assumptions would you usually make about the ϵ_i?

   (b) Plot alcohol yield against nitrogen in R. **Hint**: You can easily read the data in `R` by typing

   ```
   grain <- read.table("grain.dat").
   ```

   Do not forget to set your working directory properly (in this case to the place on your computer where you have saved the data).

   (c) Using R, fit this model, check that it fits the data and that the assumptions are reasonable.

   (d) If the model is reasonable, find a 95% confidence interval for β. Is there evidence that yield depends on nitrogen?

   (e) By how much would you expect yield to change if nitrogen content increased by .1%?

   (f) Produce a scatter plot showing the yield against nitrogen, with the best fit line from your model overlaid on it.

   (g) Write a concise report (no more than 2 paragraphs), for the experimenters, giving the model fitted (mathematically, with the assumptions on the ϵ_i included), a brief explanation of how it was fitted (the underlying statistical ideas, not the R code), whether it is a plausible fit, and what can be concluded about the relationship between yield and nitrogen. If possible, include your plot from the previous part, as illustration of the model.

9. A 95% frequentist confidence interval is supposed to include the true value of a parameter with probability 0.95, where the probability is taken over an infinite series of replications of the data gathering and inference process. For a correct linear model, with p parameters and n observations, a 95% interval for a parameter β_i is $\hat{\beta}_i \pm t_{n-p}(.975)\hat{\sigma}_{\hat{\beta}_i}$, where $\hat{\beta}_i$ is the parameter estimate, and $\hat{\sigma}_{\hat{\beta}_i}$ is its estimated standard error. $t_{n-p}(.975)$ denotes the value below which a t_{n-p} random variable lies with probability 0.975. This question examines the coverage probability of such intervals by simulation. That is we simulate data with known true parameter values, and then see how well our statistical methods do at making inferences about them.

(a) The following code simulates data from the model $y_i = 0.5 + x_i + 10x_i^2 + \epsilon_i$ where the x_i are uniformly distributed predictor variables, and $\epsilon_i \sim N(0, 0.3^2)$.

```
n <- 100 # sample size
b.true <- c(0.5, 1, 10) # true beta parameter values
ct <- qt(.975, df = n-3) # critical points for CIs
cp <- b.true*0 #coverage probability array
n.rep <- 1000  #number of replicates to run
set.seed(1) #ensure results are reproducible
for (i in 1:n.rep) {
  x <- runif(n) #simulated covariate
  mu <- b.true[1] + b.true[2]*x + b.true[3]*x^2 #mean structure
  y <- mu + rnorm(n)*0.3 #simulated data
  m1 <- lm(y ~ x + I(x^2))  # fit model to this replicate
  b <- coef(m1)  # extract parameter estimates
  sig.b <- sqrt(diag(vcov(m1))) # and standard errors
  # accumulate count of how often intervals include
  # true value...
  cp <- cp + as.numeric(b - ct*sig.b <= b.true &
                        b + ct*sig.b >= b.true)
}
cp/n.rep ## empirical coverage probability
```

Read through the code to make sure that you understand what each line is doing. Then run the code to see how close the observed coverage probability is to the nominal coverage of 0.95.

(b) What happens if the y_i data have the same dependence of the expected response on x, but the y_i are Poisson deviates? That is we simulate y_i, not as above, but using `y <- rpois(n,mu)`. Modify your coverage probability loop so that the response is Poisson, but the confidence intervals and their coverage is computed as before. How does the observed coverage probability compare to the nominal level of 0.95 now? Why do you think this might be?