Vinay Maturi

Homework 7, STAT 135

1] (1) $mean_{egglength}$: 31.38954545 mm

$mean_{chickweight}$: 6.14545 grams

$SD_{egglength}$: 1.10089 mm

$SD_{chickweight}$: 0.41659 grams

Correlation: 0.67614

(β_0) intercept: −1.77018 grams

(β_1) slope: 0.25217 $\frac{grams}{mm}$

computed in attached code.

$cw_i = \beta_0 + \beta_1 el_i + \epsilon_i$, ϵ_i normally distributed.

where β_0, β_1 are values above; cw = chick weight

el = egglength

please see code for plot.

(2) lm() produced same slope and intercept as part (1).

see code

Residual plot shows homoscedastic errors, zero mean in errors, independence.

Normal - QQ plot shows theoretical & sample quantiles match well meaning normality is a good assumption.

(3) (intercept): $H_0: \beta_0 = 0$; $H_A: \beta_0 \neq 0$

with a t-statistic of −1.329 and a p-value for this t-statistic of 0.191, we fail to reject the H_0.

There is not enough evidence to prove the intercept is non-zero.

e1: $H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$

with a t-statistic of 5.947, and a p-value of 4.73×10^{-7}, we reject the H_0, there is strong evidence that the slope is non-zero; (coeff. is statistically significant).

2] (1) 95% confidence interval for mean weight of Snowy Plover chicks that hatch from eggs weighing 8.5 grams:

see attached code for computations.

$[5.957077, 6.141139]$

$\hat{Y} \pm t_{n-2}(\alpha/2) \cdot se(\hat{Y}) = 6.05 \pm t_{42}(0.025) \cdot 0.04560$

(2) 95% prediction interval:

$\hat{Y} \pm t_{n-2}(\alpha/2) \cdot se(Y-\hat{Y}) = 6.05 \pm t_{42}(0.025) \cdot 0.47199$

$= [5.096586, 7.001631]$

see attached code for computations.

(3) for egg weight 12 grams,

95% CI: $[7.997698, 9.130126]$

95% prediction interval: $[7.459635, 9.668189]$

see attached code for computations

However, we should be skeptical of this result, as maximum egg weight in data is 9.9 grams, so we are extrapolating regression model to data it has not fit before, hence leading to this prediction and intervals related being likely biased/erroneous.

3 (1) Both coefficients for egg length and egg breadth are individually (t-test) and jointly (F-test) statistically significant as evidenced by t-tests, F-tests, and related pvalues.

The problem (2) regression is better than the regression in this problem and the regression in problem 1. That's because it has a higher R^2, lower residual standard error, and higher F-statistic for both. This implies that the problem (2) model has more statistical power in predicting chick weight than either of the other 2 models.

The residual plot has noticeable heteroscedasticity.

See code for regressions + residual plot.

(2) This regression model predicts egg weight really well, as evidenced by the R^2 of 0.9506. All coefficients are highly statistically significant, as evidenced by the individual t-test and overall F-test results. This implies egg length and egg breadth have a high amount of statistical power in predicting egg weight both individually and jointly.

See code for regression + VIF computation

VIF:

e1	e6
1.193629	1.193629

(3) The F-test is comparing your model against an intercept-only model. It has the following null hypotheses:

$H_0: \beta_1 = \beta_2 = \quad = \beta_n = 0$

H_A: for at least one value of j, $\beta_j \neq 0$.

see code for regression + VIF computation.

or

H_0: intercept-only model $[Y_i = \beta_0 + \epsilon_i]$

H_A: H_0 is false.

In this problem, the F-stat is 34.98 on 3 + 40 DF with a p-value = 2.903e-11.
We can interpret this to mean that the model has at least one non-zero β_j and that the intercept-only model is rejected.
However, as shown by the t-tests, none of the coefficients have strong evidence that they are non-zero.
This contrasting finding between overall significance (F-test) and individual significance (t-test) can be reconciled by saying that we can't identify which coefficient is non-zero, but we know at least one of them is. [I suspect a collinearity issue, as evidenced by a VIF of 20.249946 for EW.]

Hence, this regression is not as impressive, as there is a collinearity issue potentially, and none of the coefficients individually can be proven to be non-zero: the regressions I compared in (1) have none of these issues, and thus are more impressive.

(4) See code for performing all possible regressions of chick weight.

The regressions of chick weight on

(1) egg weight

(2) egg length and egg breadth

are better than the others because of their high r-squared values and the fact that all coefficients are statistically significant, as confirmed by individual t-tests and the overall F-test. High R^2 means they predict chick weight well, but additionally having all coefficients be statistically significant is a sign of good model specification, and causal explanations of factors affecting chick weight. With statistical significance, it is easier to make causal claims. (simplistically speaking of course.)

4] (1) $X_1, \ldots, X_n \overset{iid}{\sim} N(\beta, \sigma^2)$

~~[redacted]~~ prior distribution on θ: $N(\mu, \tau^2)$

$\rightarrow$ assumption: β is θ in this problem.

$$P(\theta \mid \text{data}) = P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta) P(\beta)}{P(\text{data})}$$

$$\underset{\beta}{\propto} P(\text{data} \mid \beta) P(\beta)$$

$$= P(\beta) \prod_{i=1}^{n} P(X_i \mid \beta) \propto e^{-\frac{1}{2\tau^2}(\beta-\mu)^2} \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(X_i-\beta)^2}$$

$$= e^{\left(-\frac{1}{2\tau^2}(\beta-\mu)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\beta)^2\right)}$$

$$= e^{\left(-\frac{1}{2\tau^2}(\beta^2 - 2\beta\mu + \mu^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i^2 - 2X_i\beta + \beta^2)\right)}$$

$$= e^{\left(\left[-\frac{1}{2\tau^2} - \frac{n}{2\sigma^2}\right]\beta^2 + \left(\frac{\mu}{\tau^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i\right)\beta + \left(-\frac{\mu^2}{2\tau^2} - \frac{\sum_{i=1}^{n} X_i^2}{2\sigma^2}\right)\right)}$$

$$= e^{\left(-\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\left(\beta^2 - 2\frac{\left(\frac{\mu}{\tau^2} + \frac{\sum_{i=1}^{n} X_i}{\sigma^2}\right)}{\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)}\beta + m^2\right) + c'\right]}$$

denote as c a constant

m is equal to cross term here

c' and m are constants

$$= e^{\left[-\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)(\mu - m)^2 + c'\right]}$$

$$= e^{\left[-\frac{1}{2\sigma_n^2}(\mu - m)^2 + c'\right]}$$

where $\sigma_n^2 = \dfrac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$, $m = \sigma_n^2\left(\frac{\mu}{\tau^2} + \frac{\sum_{i=1}^{n} X_i}{\sigma^2}\right)$

$$p(\theta|D) \propto e^{\left(-\frac{1}{2\sigma_n^2}(\mu-m)^2\right)}$$

$$\Rightarrow \theta|D \sim N(m, \sigma_n^2)$$

where $D = X_n$ now form m, σ_n^2 into requested expessions...

$$\Rightarrow m = \frac{\sigma_n^2}{\tau^2}\mu + \frac{\sigma_n^2}{\sigma^2} \cdot n \cdot \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{\sigma_n^2}{\tau^2}\mu + \frac{n\sigma_n^2}{\sigma^2}\bar{x}_n$$

simplify σ_n^2:

$$\sigma_n^2 = \frac{1}{\frac{\sigma^2 + n\tau^2}{\tau^2\sigma^2}} = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} \quad \leftarrow \mathrm{var}(\theta|X_n)$$

plug in σ_n^2 simplified into μ:

$$m = \frac{\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}}{\tau^2}\mu + \frac{n\left[\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\right]}{\sigma^2}\bar{x}_n$$

$$m = \frac{\sigma^2}{\sigma^2+n\tau^2}\mu + \frac{n\tau^2}{\sigma^2+n\tau^2}\bar{x}_n$$

$E(\theta|X_n)$

$$m = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}\mu + \frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{x}_n$$

(2) $\lim_{n\to\infty} Var(\theta|X_n) = \dfrac{\sigma^2\tau^2}{\sigma^2+n\tau^2} = 0$

$$\lim_{n\to\infty} E(\theta|X_n) = \frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{X}_n + \frac{\sigma^2/n}{\sigma^2/n+\tau^2}\mu$$

$= \bar{X}_n$

$\boxed{\text{The posterior distribution will become a point mass at } \bar{x}_n \text{ as } n\to\infty \text{ with no variance.}}$

(3) $$\lim_{\tau^2\to\infty} E(\theta|X_n) = \frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{X}_n + \frac{\sigma^2/n}{\sigma^2/n+\tau^2}\mu$$

$= \bar{X}_n$

This is consistent with the intuition, as I show that as the prior gets vague $[\tau^2\to\infty]$, $E(\theta|X_n)$ relies more on the sample information [in this case of the limit, solely reliant on sample information.]

5] (1) $X_1, \ldots, X_n$ are iid Poisson(λ) rvs.

$\lambda \sim \text{Gamma}(\alpha, \beta)$

prior $\longrightarrow$ $\Longrightarrow \pi(\lambda) = \dfrac{\beta^{\alpha} \lambda^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\lambda}, \quad \lambda > 0$

sample likelihood

$$f(\underline{X}_n | \lambda) = \prod_{i=1}^{n} f(x_i | \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}$$

posterior distribution.

$$\pi(\lambda | \underline{X}_n) \propto f(\underline{X}_n | \lambda)\, \pi(\lambda)$$

$$= \left[\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}\right] \left[\frac{\beta^{\alpha} \lambda^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\lambda}\right]$$

$$= \left[\lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-n\lambda - \beta\lambda}\right] \frac{\beta^{\alpha}}{\Gamma(\alpha)}$$

this looks like a Gamma pdf.

$$\propto \lambda^{\alpha^* - 1} e^{-\beta^* \lambda}$$

where $\alpha^* = \sum_{i=1}^{n} x_i + \alpha$

$\beta^* = n + \beta$

Hence,

$$\pi(\lambda | \underline{X}_n) \sim \text{Gamma}(\alpha^*, \beta^*)$$

(2) posterior mean: $E(\lambda | \underline{X}_n) = \dfrac{\alpha^*}{\beta^*} = \dfrac{\sum_{i=1}^{n} x_i + \alpha}{n + \beta}$

$$= \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i + \alpha\right)}{\frac{1}{n}(n + \beta)} = \frac{\bar{X}_n + \frac{\alpha}{n}}{1 + \frac{\beta}{n}} = \frac{\bar{X}_n}{1 + \frac{\beta}{n}} + \left(\frac{\frac{\alpha}{n}}{1 + \frac{\beta}{n}}\right) \frac{\frac{1}{\beta}}{\frac{1}{\beta}}$$

$$\boxed{E(\lambda \mid \bar{X}_n)} = \left(\frac{1}{1+\frac{\beta}{n}}\right)\bar{x}_n + \left(\frac{\frac{1}{n}}{\left(\frac{1}{\beta}+\frac{1}{n}\right)}\right)\frac{\alpha}{\beta}$$

posterior variance

$$\rightarrow Var(\lambda \mid \bar{X}_n) = \frac{\alpha^*}{\beta^{*2}} = \frac{\sum_{i=1}^{n} x_i + \alpha}{(n+\beta)^2}$$

$$Var(\lambda \mid \bar{X}_n) = \frac{\sum_{i=1}^{n} x_i + \alpha}{n^2 + 2\beta n + \beta^2}$$

(3) $\boxed{\text{Yes,}}$ the posterior mean is again a linear combination of the prior mean $\left[\frac{\alpha}{\beta}\right]$ and sample mean $[\bar{X}_n]$ as shown by the quantity for posterior mean in (2).