1. Regression Design

a) The s.e. of the slope = σ / sqrt($\sum(x_i\text{-xbar})^2$) . When you are using *n* copies of the same set of X's, $\sum(x_i\text{-xbar})^2$ for the entire data set can be written as $n\sum(x_i\text{-xbar})^2$ where the sum is only over 1 of each of the unique densities. I gave you that sum, $\sum(x_i\text{-xbar})^2$ = 0.4557. So, for *n* copies of the 9 densities, s.e. of the slope = σ / sqrt(*n* 0.4557) I gave you values for σ =0.25 and the desired s.e. = 0.075. The rest is solving for *n*: *n* = σ^2 / (se^2 0.4557) = 24.4. Since we want se < 0.075, we need *n* = 25.

b) Same ideas, but with a different configuration of x's. Here, $\sum(x_i\text{-xbar})^2$ for one copy of the two densities (0.001 and 0.686) is (0.001 – 0.3435)^2 + (0.686 – 0.3435)^2 = 0.2346. Hence, *n* = σ^2 / (se^2 0.2346) = 47.3, i.e. use *n* = 48. .

* You need more replicates per density in the two-density design, but notice that you need many fewer blocks total. In the first design (9 densities), you use a total of 25*9 = 225 blocks. In the second design (2 densities), you use a total of 48*2 = 96 blocks.

c) Not graded, since this was a 'think about the issues' question, for which there are many reasonable responses.

I would be concerned that the two density design assumes that the relationship is a straight line. It provides no data to assess that, either graphically or formally. You can always draw a line through 2 points, even if the true response is very non-linear.

I might recommend the 2 density design if I knew that the response was a straight line (perhaps from previous studies or the physics of the relationship between density and log gain. It needs fewer blocks to achieve the same precision, so it is the 'cheaper' experiment.

I might recommend the 9 density design when I needed to figure out an appropriate model for the mean response.

d) The lack of fit SS = 0.416 – 0.095 = 0.321 with 7 d.f. I get F = (0.321 / 7) / (0.095 / 81) = 39.1. This is much larger than the 0.999 quantile of the F 7,60 distribution (or the F 7,120) distribution, so p < 0.001. There is very strong evidence of lack of fit.

e) The MSE from 4d is 0.0.00117, which is a lot smaller than the MSE from the original data, 0.062). The two values are measuring different variabilities. The smaller value from part 4d is the variability between measurements on the same block. The larger value from the original data is the variability between measurements on different blocks.

I would never recommend the design in 4d unless I knew there was absolutely no variability between two blocks of the same density.

Another way to think about these issues is in terms of observational and experimental units. This is not a randomized experiment, but you could imagine that density is 'assigned' to a block. So, the block is an experimental unit. In the original data, each of the 90 blocks is measured once, so a block is the observational unit. The assumption of independence is ok, because o.u. = e.u.. In the part 4d study, the o.u. is a measurement, not the block. There are 90 measurements, but only 9 blocks. There is a problem: the o.u. is not the same as the e.u. Part 4d illustrates 'cluster effects'. Each block is a cluster.

Either (or any other reasonable) explanation is acceptable.

2. Inference on correlations - fuel efficiency

(a) We have two continuous random variables where neither variable could be thought of as a response. We want a measure of association and are not trying to predict either variable. Pearson's correlation coefficient seems to be an adequate measure of linear association.

(b) r = 0.942. The *t* statistic is t = $r\sqrt{n-2}/\sqrt{1-r^2}$ = 9.31 with 11 d.f.. The p-value is < 0.001.

* You can also get the p-value directly from the SAS output.

(c) A 99% CI for $\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$ is

$$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \pm z_{0.995}\sqrt{\left(\frac{1}{n-3}\right)} = 1.753 \pm 2.576(0.316) = (0.938, 2.568)$$

A 99% CI for ρ is $\left(\frac{e^{2*0.938}-1}{e^{2*0.938}+1}, \frac{e^{2*2.568}-1}{e^{2*2.568}+1}\right) = (0.734, 0.988)$

(d) $H_0 : \rho = 0.90 \text{ vs } H_a : \rho \neq 0.90$

$$z = \sqrt{n-3}\left(Z_r - \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)\right) = \sqrt{10}\left(1.753 - \frac{1}{2}\log\left(\frac{1+0.9}{1-0.9}\right)\right) = 0.888$$ p-value = 0.38

There is no evidence that the population correlation differs from 0.90.

(e) The test in (d) is a one-sample test, comparing a sample correlation against a hypothesized value for the population correlation. To compare two population correlations based on two samples, we need to account for the variability in the estimate of the second sample correlation. The two samples (luxury and compact cars) are independent so we should use a two sample test. Let ρ_1 and ρ_2 stand for the correlations of the two populations respectively, then

$$Z_{r1} \sim N\left(\frac{1}{2}\log\frac{1+\rho_1}{1-\rho_1}, \frac{1}{n_1-3}\right),\text{ and } Z_{r2} \sim N\left(\frac{1}{2}\log\frac{1+\rho_2}{1-\rho_2}, \frac{1}{n_2-3}\right)$$

Under $H_0 : \rho_1 = \rho_2$, $Z_{r1} - Z_{r2} \sim N\left(0, \frac{1}{n_1-3} + \frac{1}{n_2-3}\right)$. Therefore an appropriate statistic would be

$$z = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{2}{13-3}}} = 0.623$$ p-value = 0.53

There is no evidence of a difference in correlations.

3. Review of matrix operations

For both parts, the starting point is the definition of $H = X (X^TX)^{-1} X^T$

(a) H is symmetric if $H = H^T$. Let's see: $H^T = [X (X^TX)^{-1} X^T]^T = (X^T)^T [(X^TX)^{-1}]^T (X)^T = X (X^TX)^{-1} X^T = H$.

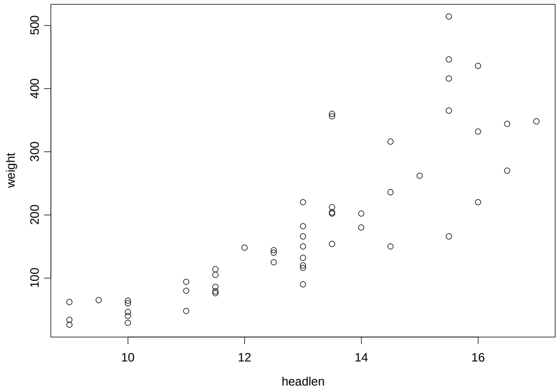
yes, H is symmetric.

Note: $[(X^TX)^{-1}]^T = (X^TX)^{-1}$ because (X^TX) is symmetric (given in the problem).

(b) H is idempotent if H H = H. Let's see: $X (X^TX)^{-1} X^T X (X^TX)^{-1} X^T = X (X^TX)^{-1} I X^T = X (X^TX)^{-1} X^T = H$.

4. Weighing bears.

(a) As the head-length increases, the size and the weight of the bear increases. The relationship between weight vs headlen doesn't appear linear; the increment is larger for higher values of headlen . It might be quadratic.



- You may notice other possible difficulties with the models we've been considering (e.g. the apparent increase in variability at large head lengths). If you log transformed weight, then replotted the data, you would probably log transform all the X variables. We'll talk a lot more about these issues later.

(b) $\hat{\beta}_H = 47.3896$ pounds per inch. The estimated average weight of a bear increases by 47.4 pounds for a 1-inch increment in the length of the head. People with biological background might make an assessment about 47 pounds being a "reasonable" value or not, but even without that background we can say that having a positive number makes sense, the bigger the head, the bigger the bear, and therefore heavier.

(c) $\hat{\beta}_H = -4.7914$ pounds per inch. Something about the relationship between the explanatory variables is making the coefficient for headlen to be negative, which does not make sense. The smaller the head, the heavier the bear?

(d) No. If headlen was uncorrelated with the other predictor variables then $\hat{\beta}_H$ would be the same, but if some correlation with the other predictors exists then the coefficient would change. In particular, the correlation between headlen and chest is 0.86, and the correlation between headlen and neck is 0.88. (Correlation between chest and neck is 0.93).

(e) t-statistic = -1.0913 with p-value = 0.2804. You can write many possible conclusions. Some focus on the parameter estimate, others on predictions from the model.

1) There is no evidence that head length of a bear is associated with its average weight, after adjusting for neck and chest size.

2) Adding head length to a model with chest size and neck size does not significantly improve the prediction of bear weight.

3) No evidence that head length needs to be included in a model that predicts weight from chest and neck measurements.

(f) Compare the reduced model $weight = \beta_0 + \beta_C chest + \beta_N neck + \varepsilon$ with the full model $weight = \beta_0 + \beta_H headlen + \beta_C chest + \beta_N neck + \varepsilon$.

$$F = \frac{\left(SSE_{reduced} - SSE_{full}\right)/r}{MSE_{full}} = \frac{\left(49815.41 - 48656.42\right)/1}{973.1} = 1.19$$, p-value = 0.28

- I didn't ask for a conclusion here. This is the same test as in 2e, so the same conclusions are appropriate.

(g) Compare the reduced model $weight = \beta_0 + \beta_C chest + \varepsilon$ with the full model.

$$F = \frac{\left(SSE_{reduced} - SSE_{full}\right)/r}{MSE_{full}} = \frac{\left(56895.06 - 48656.42\right)/2}{973.1} = 4.23$$, p-value = 0.02

Again, there is more than one conclusion; either is appropriate

1) Predictions of the average weight of a bear are improved by including either (or both) of the neck and headlength variables.

2) One or more of the regression slopes for neck and headlength are non-zero.

(h) F-statistic: 252.7 on 3 and 50 degrees of freedom, the p-value is < 0.0001. At least one of these three variables is useful to predict the weight of a bear.

* This is one of the default tests provided by SAS. You could also get it by comparing models.

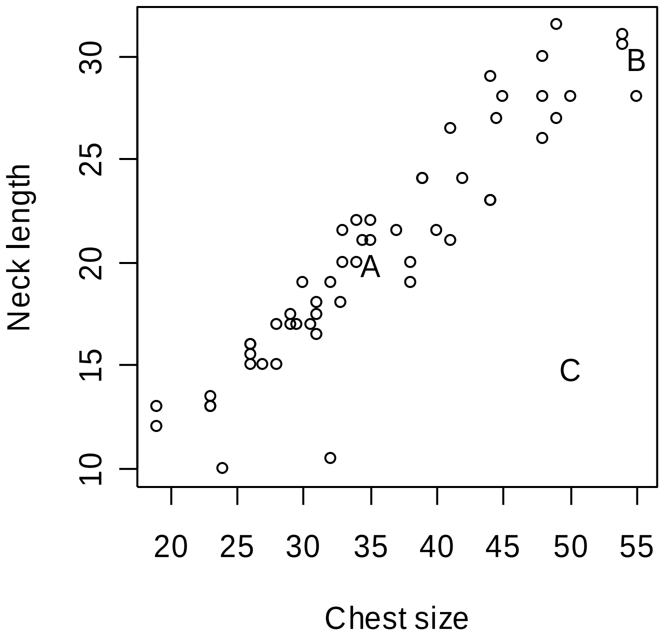
(i) The estimated coefficients are: β_0 = -267, β_1 = 9.29, β_0 = 5.77. Predicted weights (and s.e.'s for next part) are in the table:

Bear	Chest	Neck	Predicted weight	s.e.
A	35	20	173 lbs	4.3 lbs
B	55	30	417 lbs	10.3 lbs
C	50	15	284 lbs	30.3 lbs

(j) The s.e. for the predicted weight is given by $\sqrt{MSE(\mathbf{x}_n^T(\mathbf{X}^T X)^{-1}\mathbf{x}_n}$ where $\boldsymbol{x_n}^{\mathbf{T}}$ **=** [1 35 20], [1 55 30], or [1 50 15].

(k) The explanation is obvious when you plot the chest size and neck length and superimpose the prediction locations (plot on next page). Bear C is much farther from the center of the data than bear B. Even though each dimension is individually closer to the mean, bear C has an unusually small neck for the chest size (or an unusually large chest for the neck size). That puts it a long way from the mean and increases the s.e.

* This is one of the reasons why it is hard to identify X outliers in multiple regression. A multivariate outlier (here, point C) may not be a univariate outlier.


Neck length
Chest size
A
B
C