Solutions to Practice Problems

Math 430, Winter 2017

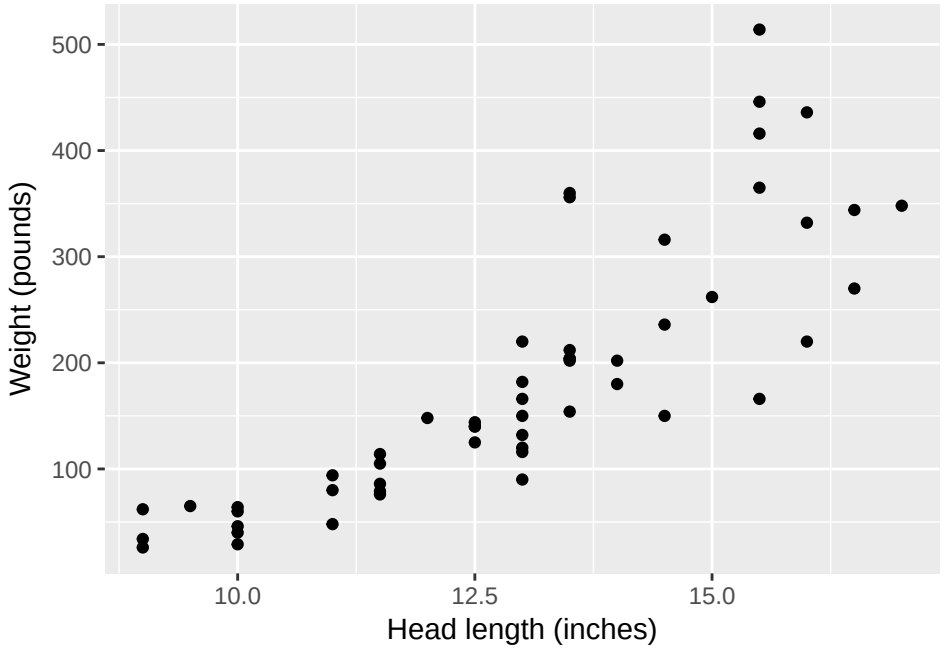
Problem 1

The weight of a bear is an important measure of how well it is doing. Weighing a bear in the wild is difficult. It is a lot easier to measure the length of various parts of the bear's body. The following data were collected in an attempt to find simpler measures that could adequately predict bear weight. 54 bears were located in the wild (assume this is a random sample of bears from this location). Each was anesthetized, weighed, and measured.

We will consider the data from three X variables: `chest`, `headlen`, and `neck`. These are the girth of the chest, the length of the head and the length of the neck. All are measured in inches. The goal is to predict the weight (Y) of the bear, measured in pounds. For the purpose of this assignment, use only these four variables (`weight`, `chest`, `headlen`, and `neck`). Do not worry about the rest of the variables in the data set. Also, do not worry about assumptions. We'll assume that the model is linear and that the errors have equal variances.

(a)

Below is a plot of weight vs. head length for these bears. Describe the relationship. Does this relationship make sense (biologically)?



As the head-length increases, the size and the weight of the bear increases. The relationship between weight and head length doesn't appear linear; the increment is larger for higher values of head length—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.

(b)

Use the below R output to estimate the value of β_H in the SLR model $Y = \beta_0 + \beta_H x_{\text{headlength}} + e$. Interpret your estimate of β_H in the context of the problem.

```
##
## Call:
## lm(formula = weight ~ headlen, data = bears)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -137.557  -35.062   -7.972   20.086  210.443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -430.982     57.031  -7.557 6.41e-10 ***
## headlen       47.390      4.345  10.908 4.75e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.81 on 52 degrees of freedom
## Multiple R-squared:  0.6959, Adjusted R-squared:   0.69
## F-statistic:   119 on 1 and 52 DF,  p-value: 4.752e-15
```

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, the heavier the bear.

(c)

Use the below R output to estimate the value of β_H in the MLR model $Y = \beta_0 + \beta_H x_{\text{headlength}} + \beta_C x_{\text{chest}} + \beta_N x_{\text{neck}} + e$. Interpret β_H in the context of the problem. Does this make sense?

```
##
## Call:
## lm(formula = weight ~ headlen + chest + neck, data = bears)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.174 -18.890  -1.933  10.828  98.639
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -239.295     30.553  -7.832 3.03e-10 ***
## headlen       -4.791      4.390  -1.091  0.28036
## chest          9.599      1.321   7.269 2.28e-09 ***
## neck           6.904      2.378   2.903  0.00549 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.2 on 50 degrees of freedom
## Multiple R-squared:  0.9381, Adjusted R-squared:  0.9344
## F-statistic: 252.7 on 3 and 50 DF,  p-value: < 2.2e-16
```

The estimated average weight of a bear decreases by 4.8 pounds for a 1-inch increment in the length of the head, holding chest girth and neck length constant.

Use the use the above models for parts (d)-(g).

(d)

Test $H_0 : \beta_H = 0$, using a t-test. What is the p-value? Write a one-sentence conclusion.

For the multiple regression model we find a t-statistic of -1.091 with an associated p-value of 0.28036. You can write many possible conclusions. Some focus on the parameter estimate, others on predictions from the model.

- There is no evidence that head length of a bear is associated with its average weight, after adjusting for neck and chest size.
- Adding head length to a model with chest size and neck size does not significantly improve the prediction of bear weight.
- There is no evidence that head length needs to be included in a model that predicts weight from chest and neck measurements.

(e) Not yet

Test $H_0 : \beta_H = 0$, using an F-test (this is also known as the model comparison approach). What is the p-value? Write a one-sentence conclusion.

(f) Not yet

Construct a test of $H_0 : \beta_H = \beta_C = 0$. What is the p-value (at least approximately)? Write a one-sentence conclusion. Hint: thinking about the models that are being compared may help you construct this test and write a conclusion.

(g)

Construct a test of $H_0 : \beta_H = \beta_N = \beta_C = 0$. What is the p-value? Write a one-sentence conclusion.

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 the F-test that R provides in the summary of the full model.

For parts (h) and (i) use the following model:

```
bear_mod3 <- lm(weight ~ chest + neck, data = bears)
summary(bear_mod3)
```

```
##
## Call:
## lm(formula = weight ~ chest + neck, data = bears)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.453 -20.347  -1.682  15.495 103.363
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -267.065     16.942 -15.764  < 2e-16 ***
## chest          9.292      1.292   7.189 2.74e-09 ***
## neck           5.769      2.143   2.692  0.00958 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 31.25 on 51 degrees of freedom
## Multiple R-squared:  0.9366, Adjusted R-squared:  0.9342
## F-statistic:   377 on 2 and 51 DF,  p-value: < 2.2e-16
```

(h)

Given the below R output, predict the weight for the following 3 bears:

bear	chest	neck
A	35 in	20 in
B	55 in	30 in
C	50 in	15 in

Using the prediction equation $\widehat{y} = -267.065 + 9.292 * x_C + 5.769 * x_N$ we find the following predictions:

bear	chest	neck	prediction
A	35 in	20 in	173.5 lbs
B	55 in	30 in	417.1 lbs
C	50 in	15 in	284.1 lbs

(i)

Calculate the standard error of the predicted average (i.e. the SE of the line) for each of the three bears in the previous part.

The SE for the predicted weight is given by:

$$\sqrt{MSE(\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})}$$

where $\mathbf{x}' = (1\ 35\ 20)$, $(1\ 55\ 30)$, or $(1\ 50\ 15)$.

Adding this to the above table we get

bear	chest	neck	prediction	se
A	35 in	20 in	173.5 lbs	4.3 lbs
B	55 in	30 in	417.1 lbs	10.3 lbs
C	50 in	15 in	284.1 lbs	30.3 lbs

Problem 2

Pharmaceuticals usually have a shelf life. One reason is that the concentration of active ingredient decreases over time. One common model for this is the exponential decay model

$$Y_i = C_0 e^{-\lambda x_i} \times e_i$$

where Y_i is the concentration at time i, C_0 is the initial concentration (at time of manufacture), x_i is the time since manufacture, and λ is the decay rate. This model can be fit using linear regression by log transforming both sides to get

$$Y_i^* = \log(Y_i) = \beta_0 + \lambda x_i + e_i$$

Data were collected from a randomized experiment. Twenty packages of a particular drug were randomly assigned to be sampled at one of five times, 0 months, 2 months, 4 months, 8 months, and 12 months after manufacture. 4 packages were assigned to each sampling time. Each package was measured once and only once, at the assigned sampling time.

(a)

Four different models were fit to the data. Some unimportant details of each model are left out. The residual sums of squares for each are:

Model	# parameters	RSS
$Y_i^* = \beta_0$	1	10.020
$Y_i^* = \beta_0 + \lambda x_i$	2	0.5051
$Y_i^* = \beta_0 + \lambda x_i + \beta_2 x_i^2$	3	0.5050
$Y^* = \mu + \tau_i$	5	0.2550

(i) Please test whether $\lambda = 0$. Report the test statistic, the appropriate d.f. and an approximate p-value. If this is not possible from the information provided, indicate what additional information you need.

This can be answered using an F-test comparing models 1 and 2 (which is the typical F-test for SLR). The trick here is to recall that RSS for model 1 is just SST. After that, it's easy to calculate SSreg: $SST - RSS = 10.020 - 0.5051 = 9.5149$. Now, the F-test is

$$F = \frac{9.5149}{0.5051/18} = 339.1$$

If you had R it would be easy to find the p-value as the upper-tail area past 339.1. On the exam, you could sketch a picture of the F distribution and make it clear what the p-value would be in terms of area under the curve.

(ii) Not yet Please test whether $\beta_2 = 0$. Report the test statistic, the appropriate d.f. and an approximate p-value. If this is not possible from the information provided, indicate what additional information you need.

(b)

The drug company has a large batch of the drug (many packages) that has been in storage for four months. They will use the regression to predict the mean concentration of drug in this batch. They also want to report the precision of that estimate. There are three possible formulae that could be used:

1) $\sqrt{S^2\left(\frac{1}{n} + \frac{(4-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}$

2) $\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(4-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}$

3) $\sqrt{S^2/4}$

Which is the most appropriate formula to calculate standard error? Explain (briefly) why your choice is the most appropriate.

Equation 1 is most appropriate, as the drug company wants to predict the **mean** of a large number of batches using regression.

Problem 3

Set up the design matrix, $\mathbf{X}$, and the parameter vector, $\boldsymbol{\beta}$, for each of the following regression models (assume $i = 1, \ldots, 5$).

1) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + e_i$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{11}^2 \\ 1 & x_{21} & x_{21}^2 \\ 1 & x_{31} & x_{31}^2 \\ 1 & x_{41} & x_{41}^2 \\ 1 & x_{51} & x_{51}^2 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

2) $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 \log(X_{i2}) + e_i$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \log(x_{12}) \\ 1 & x_{21} & \log(x_{22}) \\ 1 & x_{31} & \log(x_{32}) \\ 1 & x_{41} & \log(x_{42}) \\ 1 & x_{51} & \log(x_{52}) \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Problem 4 Not yet

The following regression model is being considered in a market research study:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + e_i$

Write the equation of the reduced regression models used in a nested F-test for testing whether or not:

1) $\beta_1 = \beta_3 = 0$
2) $\beta_0 = 0$
3) $\beta_1 = \beta_2$