STAT425 Homework #1
Fall 2015 Sep 3, 2015

Homework 1

Due: Sep 11, 2015

Problem 1. (8pt) Consider the simple linear regression model with no intercept

$$y_i = \beta x_i + e_i, \quad i = 1, \dots, n,$$

with $\mathbb{E}[e_i] = 0, \mathrm{Var}(e_i) = \sigma^2$, and $\mathrm{Cov}(e_i, e_j) = 0$ for $i \neq j$.

(a) Find the least squares estimator of β.

(b) Find the mean and variance of $\hat{\beta}$.

(c) Let $r_i = y_i - \hat{y}_i$ denote the residuals. Show that $\sum_i r_i x_i = 0$.

(d) Show that $\sum_i y_i^2 = \sum_i r_i^2 + \sum_i y_i^2$. (This decomposition is used to define R-square for regression without an intercept.)

Problem 2. (12pt) The following are outputs from R regarding a simple linear regression model. Some outputs have been removed and you are asked to recover those outputs based on the provided information.

```
> lm.out = lm(y˜x, data=mydata)
> summary(lm.out)
Call:
lm(formula = y ˜ x, data = mydata)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0550      ????     ????     ????
x             1.0413      ????     ????     ???? *

Residual standard error: 4.596 on 18 degrees of freedom
Multiple R-squared:   ????,Adjusted R-squared:  ????
> anova(lm.out)
Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
x          ?   ????   ????  ????      ????  *
Residuals ??   ????   ????  ????

> var(mydata$y)
[1] 28.7474
> xbar=mean(mydata$x)
[1] 5.128502
> xvar=var(mydata$x)
[1] 8.05311
```

(a) What's the std. error for the intercept (β_0) and what's the corresponding t-value?

(b) What's the std. error for β_1 (the coefficient in front of x) and what's the corresponding p-value?

(c) What percentage of variation in y is explained by x?

(d) What's the adjusted R-square?

Problem 3. (14pt) Download the data set `coffee.Rdata` from Compass. It gives the cooling temperatures (in Celsius) of a freshly brewed cup of coffee after it is poured from the brewing pot into a serving cup.

The relationship between y (temperature) and x (waiting time) is often modeled with an exponential function of the following form:

$$y \approx A \times B^x.$$

Take the logarithm of both sides, and we have

$$\ln y = \ln A + x \ln B,$$

then we can apply LS on the transformed data $(x, \ln y)$ to obtain estimates of $\ln A$ and $\ln B$ and transform them back to A and B.

(a) Report $\hat{A}$ and $\hat{B}$, as well as their 95% CIs.

(b) Graph the fitted exponential line, $y = \hat{A} \times \hat{B}^x$, on the scatter plot of the original data points (i.e., use Temp as the y-coordinate, instead of log of Temp).

(c) On average, how long do we need to wait till the coffee is not hotter than 40 °C?

(d) What is the predicted temperature of the coffee after 1 hour?

(e) What is the predicted temperature of the coffee after 1 hour and 15 minutes, if the temperature in the room is 24 °C?

(f) In 1992, a woman sued McDonald's for serving coffee at a temperature of 180 °F that caused her to be severely burned when the coffee spilled. An expert witness at the trial testified that liquids at 180 °F will cause a full thickness burn to human skin in two to seven seconds. It was stated that had the coffee been served at 155 °F, the liquid would have cooled and avoided the serious burns. The woman was awarded over 2.7 million dollars. As a result of this famous case, many restaurants now serve coffee at a temperature around 155 °F. How long should restaurants wait (after pouring the coffee from the pot) before serving coffee, to ensure that the coffee is unlikely to be hotter than 155 °F (here *unlikely* means 155 is outside the 95% prediction interval of the temperature at that time)?

Note that the question is stated in Ferinheight, but the temperature in the data is in Celsius.

Problem 4. The file `gift.csv`, downloaded from Google Trends, shows the weekly search interest for two terms, "gift girlfriend" (abbreviated as "GF") and "gift boyfriend" (abbreviated as "BF"), in the year of 2014.

(a) (6pt) Provide two graphic displays of the data. In one figure, display search interest versus time (wks), color the data for the two terms differently, and connect the data points to produce a line for each term. In the other figure, produce a scatter plot of GF vs BF.

Comment on any interesting patterns in the data. For example, why are there two bumps in the first figure?

(b) (4pt) For counts data, we usually apply a log-transformation. To check whether log-transformation can improve the model fitting, calculate the R-squares without/with log transformation. Do you want to do the log-transformation?

(c) (10pt) Analyze the data using what you have learned on SLR. Report your results. Especially, form some hypotheses, e.g., is one term significantly searched more then the other? Also state how you test your hypotheses and what your conclusions are.

Note: Remember to submit a copy of all necessary `R` code on Compass. See "Assignments".