Exam 1

Subrata Paul

9/23/2020

Problem 1 (5 points)

Write down a linear regression model with assumptions.

Answer

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_p X_p + \epsilon$$

where, - Y is the response variable - $X_1, \dots X_p$ are predictor variables - $\beta_0, \beta_1, \beta_p$ are the regression coefficient - ϵ is unobserved random error

Assumptions:

```
- Linearity: $E[Y|X] = \beta_0 + \beta_1X_1 + \dots + \beta_pX_p$
    -that is the relationship between the predictor variables and the mean of the response variable is 
- Assumption on error: $\epsilon\sim N(0,\sigma^2)$
    1. The variance of the error term is constant for all observations
    2. Observations are independent to each other
    3. For any fixed value of the predictor variables, the response is normally distributed
```

Problem 2 (5 points)

In fitting a simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$, it was found that observation Y_i fell directly on the fitted regression line. If this case were deleted, would the least square regression line fitted on the remaining $n - 1$ cases be changed? [Hint: try to use the function that we minimize in the least square procedure.]

Answer

The OLS estimate will not change since the residual sum of squares that we minimize does not change by removing the observation for which $y = \hat{y}$.

Problem 3(a) (10 points)

In this problem, you will simulate data with 5000 observations. About 50% of them are male. Use a binomial distribution to choose the number of males randomly (Hint: You are flipping a fair coin and counting the number of heads). Call the variable `Gender`. Diastolic blood pressure of male and female follows a normal distribution with mean $\mu_{\text{male}} = 82$, $\mu_{\text{female}} = 80$ mmHg and standard deviation $\sigma_{\text{male}} = \sigma_{\text{female}} = 10.5$. Total cholesterol in blood follows a normal distribution with a mean of 5.69 and variance 1.31. Glucose follows a normal distribution with a mean 5.12 and a standard deviation of 1.24. Gender, cholesterol, and glucose are predictor variables. The response variable is BMI. The error term, ϵ $N(0, 9)$ accounts for the randomness and effect of other factors that affect BMI. The mean BMI while all the predictors are zero is 23. Simulate BMI so that the effect sizes (regression coefficients) of Gender (Male), blood pressure, cholesterol, and blood glucose are 0.01, 0.07, 0.1, and -0.1, respectively. Run a multiple linear regression and discuss if the regression model could identify the simulated relationship. You should also discuss if you find an estimated coefficient for a variable that is much different than the parameter used in the simulation.

```
set.seed(123)
n = 5000
gen = sample(size = n, prob = c(0.5,0.5), x = c(0,1),replace = T)
blood = rep(NA, n)
blood[gen==0] = rnorm(sum(gen==0), 80, 10.5)
blood[gen==1] = rnorm(sum(gen==1), 82, 10.5)
chol = rnorm(n, 5.69, sd = sqrt(1.31))
glu = rnorm(n, 5.12, 1.24)
dat = data.frame(Gender = gen, bp = blood, chol = chol, glu = glu)
dat$bmi = 23 + 0.01*dat$Gender+ 0.07*dat$bp + 0.1 * dat$chol - 0.1*dat$glu + rnorm(5000, mean = 0, sd =
dat$Gender = as.factor(dat$Gender)
summary(lm(bmi~Gender + bp + chol+ glu, data = dat))
```

```
##
## Call:
## lm(formula = bmi ~ Gender + bp + chol + glu, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3770  -2.0119   0.0161   1.9709  11.6812
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.081490   0.432088  51.104  < 2e-16 ***
## Gender1      0.043594   0.085272   0.511 0.609211
## bp           0.078840   0.004066  19.389  < 2e-16 ***
## chol         0.137420   0.036676   3.747 0.000181 ***
## glu         -0.117182   0.034044  -3.442 0.000582 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.998 on 4995 degrees of freedom
## Multiple R-squared:  0.07521,    Adjusted R-squared:  0.07447
## F-statistic: 101.6 on 4 and 4995 DF,  p-value: < 2.2e-16
```

Problem 3(b) (2 points)

Run the same analysis as problem 3(a) multiple times. Do you get the same or different estimates? Why?

Answer

The estimates will change slightly.

```
run_lm<-function(){
  n = 5000
gen = sample(size = n, prob = c(0.5,0.5), x = c(0,1),replace = T)
blood = rep(NA, n)
blood[gen==0] = rnorm(sum(gen==0), 80, 10.5)
blood[gen==1] = rnorm(sum(gen==1), 82, 10.5)
chol = rnorm(n, 5.69, sd = sqrt(1.31))
glu = rnorm(n, 5.12, 1.24)
dat = data.frame(Gender = gen, bp = blood, chol = chol, glu = glu)
dat$bmi = 23 + 0.01*dat$Gender+ 0.07*dat$bp + 0.1 * dat$chol - 0.1*dat$glu + rnorm(5000, mean = 0, sd =
```

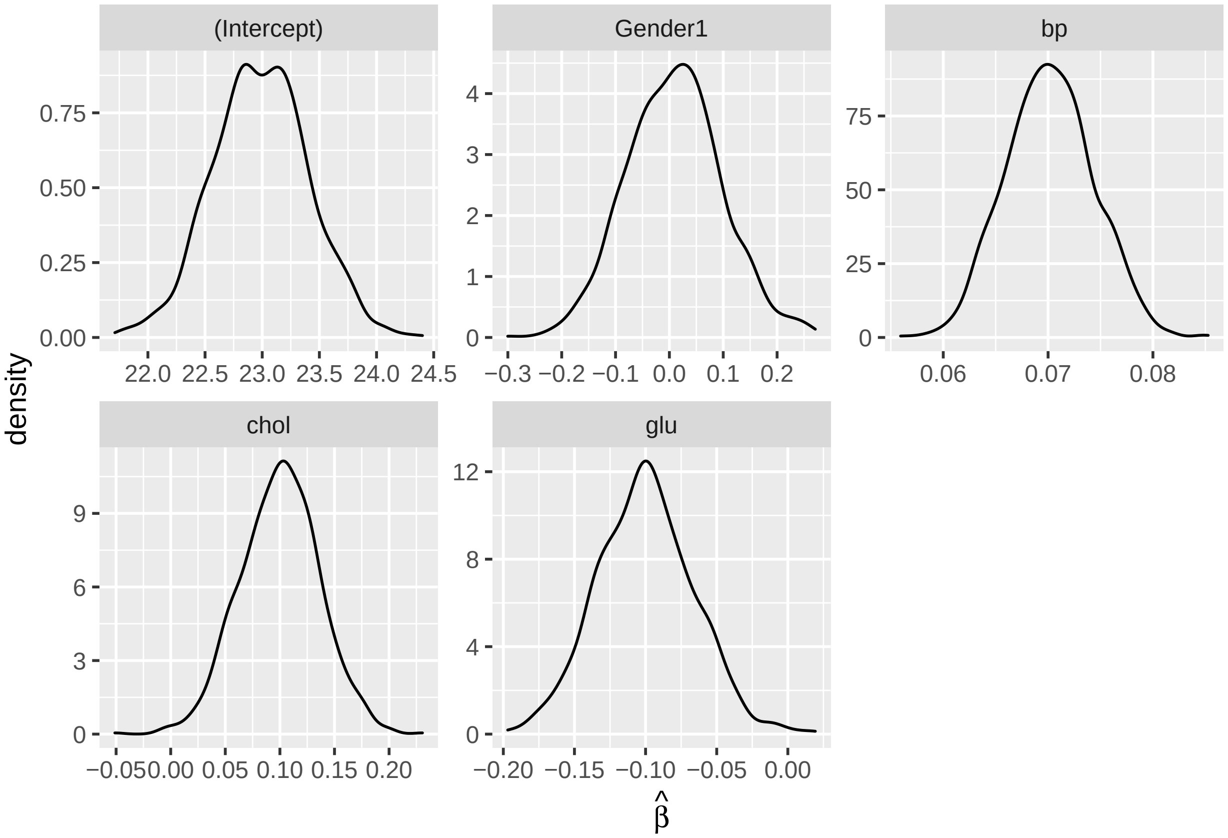
```r
dat$Gender = as.factor(dat$Gender)
coefs = data.frame(coef(lm(bmi~., data = dat)))
return(coefs)
}

n_sim = 1000
coefs <- data.frame()
for(i in 1:n_sim){
  coefs <- rbind(coefs, t(run_lm()))
}
row.names(coefs) = 1:n_sim
head(coefs)
```

```
##   (Intercept)     Gender1         bp       chol         glu
## 1    22.47907 -0.03692055 0.07357643 0.14030352 -0.09247463
## 2    22.53854 -0.19281750 0.07346449 0.12150629 -0.05923224
## 3    22.59409  0.10724695 0.07472367 0.10616658 -0.11349930
## 4    22.70365  0.09055622 0.07230327 0.14522831 -0.13341983
## 5    23.13740  0.08716673 0.07012052 0.04571838 -0.07221209
## 6    23.32271 -0.05148559 0.06685178 0.12032925 -0.11922812
```

```r
plot_dat = reshape2::melt(coefs)
```

```
## No id variables; using all as measure variables
```

```r
ggplot(plot_dat, aes(x = value))+
  geom_density()+
  facet_wrap(vars(variable), scale = 'free')+
  xlab(expression(hat(beta)))
```



Problem 4 (10 points)

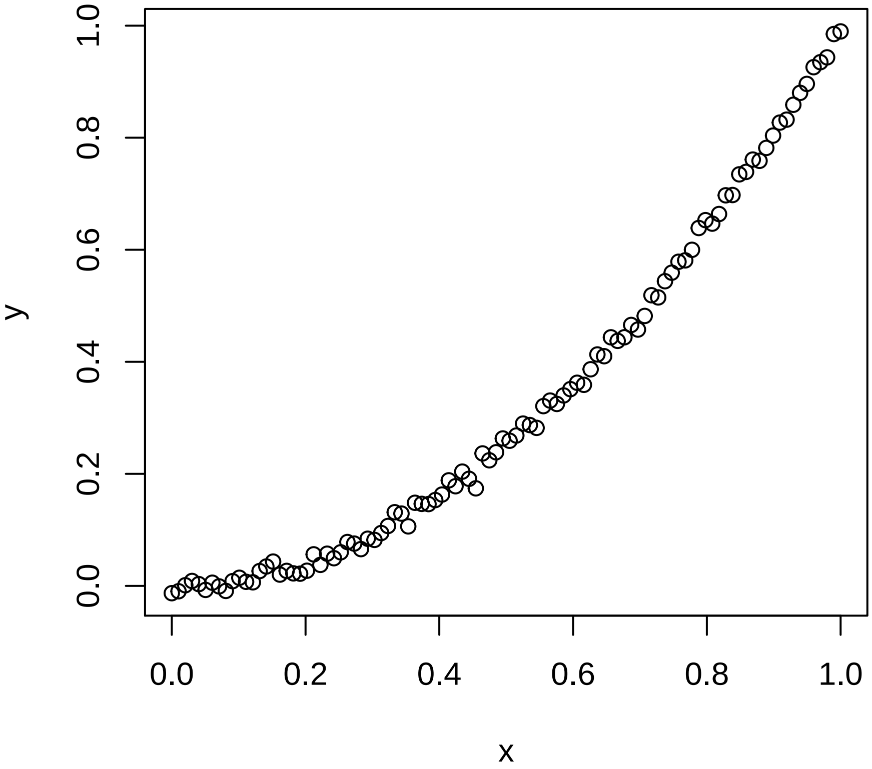
a. What does it mean for a regression model to be a linear model? (Specifically, explain what linear model means in the context of a regression model.)

Linear in terms of the coefficients

b. Consider a setting where there are four observations ($n = 4$) and two predictors ($p = 3$). Construct a 4×3 design matrix X that would lead to an unidentifiable model but where no two columns are identical.

Consider the figure below for parts (c) and (d) of this question.

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ -1 & 1 & 0 \\ 5 & 6 * 11 & \end{bmatrix}$$



(c) Is the relationship between x and y linear? Why?

Does not looks like linear rather quadratic

(d) Explain how the relationship between y and x can be approximated reasonably well by a linear model.

Square root transformation of the predictor is suggested

Problem 5 (8 points)

Consider the model

$$\log(\text{ppgdp}) = \beta_0 + \beta_1 \text{fertility} + \beta_2 \log(\text{pctUrban}) + \epsilon$$

You can find the description of the data and the variables using `?alr4::UN11`. Fit the model, print the summary of the model and, interpret the coefficient of `pctUrban`.

You will not get full credit for using generic terms or variable names like fertility or pctUrban. Clearly indicate what these variables are measuring/representing.

```
data('UN11', package = 'alr4')
lmod = lm(log(ppgdp) ~ fertility + log(pctUrban), data = UN11)
summary(lmod)
```

```
##
## Call:
## lm(formula = log(ppgdp) ~ fertility + log(pctUrban), data = UN11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8678 -0.6270 -0.1053  0.6857  2.8836
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.51457    0.72579   6.220 2.94e-09 ***
## fertility      -0.56088    0.05798  -9.674  < 2e-16 ***
## log(pctUrban)   1.38993    0.15758   8.820 6.25e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9178 on 196 degrees of freedom
## Multiple R-squared:  0.6564, Adjusted R-squared:  0.6529
## F-statistic: 187.2 on 2 and 196 DF,  p-value: < 2.2e-16
```

Keeping the fertility rate i.e. number of children per women constant, an 1% increase in the percentage of urbanization leads to 1.0139263 percent increase in per capita gross domestic product in US dollars.

Problem 6 (10 points)

Assume that the observations for the response variable are correlated i.e. $\text{cov}(y_i, y_j) \neq 0$. So the variance-covariance matrix $Var(\epsilon) \neq \sigma^2 I$, where σ is a constant and I is the identity matrix. Instead assume that $Var(\epsilon) = \sigma^2 I + \gamma^2 K$, where K is not a diagonal matrix. How does this phenomena effects the estimates $\hat{\beta}$. (More specifically is $E[\hat{\beta}]$ and $Var(\hat{\beta})$ in this case and how they vary from that under usual linear regression model assumption?)

Answer

The OLS estimate is still unbiased because we have $E[\epsilon] = 0$

The variance of $\hat{\beta}$:

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} + \gamma^2 (X^T X)^{-1} X^T K X (X^T X)^{-1}$$

According to OLS the esimate of variance of $\hat{\beta}$ is $\hat{\sigma}^2 (X^T X)^{-1}$ which will be a biased estimate under the assumption of correlation between the observations. Whether the variance of $\hat{\beta}$ be overestimated or underestimated depends on the sign in the second term of the above equation. If the variance of $\hat{\beta}$ is overestimated, we will loose in power but in the other case we will get false positive association.

Problem 7 (10 points)

Consider a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

with usual notations and assumptions.

a. How does the parameter β_0 and β_1 chaanges if we center the predictor variable X (i.e. substract $\overline{X}$ from X).

β_1 does not change but β_0 changes by $\beta_1 X$

b. How do the parameters changes if we scale the predictor variable X (i.e. divide X by its standard deviation?)

β_0 does not change but β_1 will be a multiple of the standard deviation of X.

c. If X and Y are uncorrelated what can be said about β_0 and β_1?

$\beta_0 = \overline{Y}$ and $\beta_1 = 0$

Problem 8 (10 points)

```
n=5000
dat = data.frame(row.names = seq(1,n))
dat$x1 = abs(rnorm(n, mean = 5, sd = 2))
dat$x2 = 1000*rgamma(n, 1, 20)
dat$x3 = 100*rbeta(n,1,5)
dat$y = 20 - sqrt(dat$x1) + 2*log(dat$x2)+ 0.5*dat$x3 + rnorm(n, 0, 4)
lmod = lm(y~x1+x2+x3, data = dat)
summary(lmod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1110  -2.7636   0.1039   2.9192  14.8338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.484420   0.193947  121.09  < 2e-16 ***
## x1          -0.177658   0.030953   -5.74 1.01e-08 ***
## x2           0.040152   0.001226   32.75  < 2e-16 ***
## x3           0.497003   0.004374  113.62  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.318 on 4996 degrees of freedom
## Multiple R-squared:  0.7392, Adjusted R-squared:  0.739
## F-statistic:  4719 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```
#write.table(dat, '../data/simu_exam1.txt', sep = '\t', row.names = F, col.names = T, quote = F)
```

Download the data `simu_exam1.txt` from the canvas. Fit a multiple linear regression model with Y as the response variable and $x1, x2, x3$ as predictors (just one model with three predictors). Perform model diagnostic for structure. If there are issues, suggest a model that is more appropriate for the data. Give the coefficients of the final model and interpret them.

Most of you did a really good job here to find out the correct transformation of X_2. The transformation of X_1 was not obvious on the diagnostic plots but some of you picked it up.

Problem 9 (20 points)

Select data from http://archive.ics.uci.edu/ml/datasets.php. On the left sidebar select `Regression`, `Numerical`, and `Multivariate`. You can choose any data from the list that has `Default Task = Regression` and the number of instances more than 500. Do not select data that has `Time Series` in the `Default Task` column. You will describe the data, run an appropriate multiple linear regression model, perform diagnostic for model structure, and transform variable if appropriate, and at the end interpret your result.

Problem 10 (MATH 5387 only)

Consider the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j + \cdots + \beta_{p-1} X_{p-1}$$

Show that the OLS linear fit to the data in an added variable plot for predictor x_j will have slope β_j and intercept 0.