HW2

Monday, May 19, 2025 11:48 AM

Problem 1. Regression through the origin. We will consider a special case of the simple linear regression model, where the intercept term is assumed to be zero from the outset (this is often assumed in the calibration of certain measuring devices). Let

$$Y_i = x_i\beta + \epsilon_i,$$

where $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$, and the ϵ_is are uncorrelated. In what follows we will treat the x_is as known constants.

(a) Find the least squares estimate of β, call it b; i.e., find β that minimizes $Q(\beta) = \sum_{i=1}^n (Y_i - x_i\beta)^2$.

(b) Show that $E(b) = \beta$.

(c) Show the $V(b) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$.

(d) Investigate whether b is a BLUEs estimator of β.

a) $\frac{\partial}{\partial\beta} Q(\beta) = \sum_{i=1}^n (Y_i - x_i\beta)(-x_i)$

$= -2\sum_{i=1}^n (Y_i x_i - x_i^2\beta)$

$= 0 \Rightarrow \beta = \frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}$

b) $E(b) = E\left[\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right] = E\left[\frac{\sum_{i=1}^n (x_i\beta + \Sigma_i)x_i}{\sum_{i=1}^n x_i^2}\right]$

$= E\left[\beta\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} + \frac{\sum_{i=1}^n \Sigma_i x_i}{\sum_{i=1}^n x_i^2}\right]$

$= \beta + \frac{\sum_{i=1}^n x_i E[\Sigma_i]}{\sum_{i=1}^n x_i^2}$ ($E[\Sigma_i] = 0$)

$= \beta$

c) $V(b) = V\left[\frac{\sum_{i=1}^n (x_i\beta + \Sigma_i)x_i}{\sum_{i=1}^n x_i^2}\right] = V\left[\beta + \frac{\sum_{i=1}^n \Sigma_i x_i}{\sum_{i=1}^n x_i^2}\right]$

$= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 V(\Sigma_i)$

$= \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$

d) b is unbiased and is from a linear model

WTS $V(b) \le V(\tilde{b})$ $\forall$ unbiased linear estimators of β

Let $\tilde{b} = \sum_{i=1}^n w_i Y_i$ be unbiased

$= \sum_{i=1}^n w_i(\beta x_i + \Sigma_i)$

$\Rightarrow \sum_{i=1}^n w_i x_i = 1$ since $E[\tilde{b}] = \beta$

It follows

$V[\tilde{b}] = \sum_{i=1}^n w_i^2\sigma^2$ since $COV(Y_i, Y_j) = \sigma^2\delta_{ij}$ by uncorrelation of Σ_is

Thus

$V[b]\sum_{i=1}^n x_i^2 = \sigma^2 = \sigma^2(\sum_{i=1}^n w_i x_i)^2 \overset{\text{Cauchy-Schwartz}}{\le} \sigma^2\sum_{i=1}^n w_i^2\sum_{j=1}^n x_i^2 = V[\tilde{b}]\sum_{i=1}^n x_i^2$

$\therefore$ b is a BLUEs estimator of β

Problem 2. Suppose a sample of 10 types of compact cars reveals the following one-day rental prices (in dollars) for Hertz and Thrifty, respectively:

	Car Type									
Renter	A	B	C	D	E	F	G	H	I	J
Hertz	37.16	14.36	17.59	19.73	30.77	26.29	30.03	29.02	22.63	39.21
Thrifty	29.49	12.19	15.07	15.17	24.52	22.32	25.30	22.74	19.35	34.44

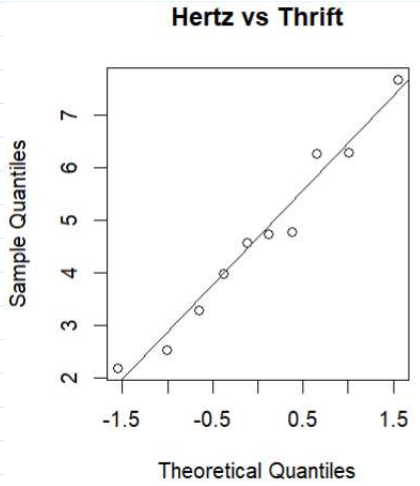
(a) Explain why this is a paired-sample problem.

(b) Use a graph to determine whether the assumption of normality is reasonable.

(c) Using a p-value, test at $\alpha = 0.05$ whether Thrifty has a lower true mean rental rate than Hertz via a t-test.

a) Because both sets of prices are dependent on which car is being rented and we are interested on the difference in price.

b)
```
#Question 2b
hertz<-c(37.16,14.36,17.59,19.73,30.77,26.29,30.03,29.02,22.63,39.21)
thrift<-c(29.49,12.19,15.07,15.17,24.52,22.32,25.30,22.74,19.35,34.44
dif<-hertz-thrift
qqnorm(dif,main="Hertz vs Thrift")
qqline(dif)
```



Since the data runs roughly linearly with the qqline, the normal assumption is reasonable

c)
```
> #Question 2c
> nd<-length(dif)
> pd<-pt(0.95,nd-1)
> ts<-mean(dif)/(sd(dif)/sqrt(nd))
> if(ts>pd){cat("Reject")}else{cat("Fail to reject")}
Reject
```

Since dif=Hertz-Thrift, $H_0: \mu_d \leq 0$ and $H_a: \mu_d > 0$ which is a right tail test.

Problem 3. Supposed our observed data follow a normal error simple linear regression model,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Now suppose we have a sample of size n, where n is an even number, and we divide the sample up into non-overlapping pairs, $(i_1, j_1), (i_2, j_2), \ldots, (i_{n/2}, j_{n/2})$; i.e., we arrange our sample as $\{(x_{i_1}, y_{i_1}), (x_{j_1}, y_{j_1})\}, \ldots, \{(x_{i_{n/2}}, y_{i_{n/2}}), (x_{j_{n/2}}, y_{j_{n/2}})\}$. Consider the following estimator of β_1:

$$\tilde{\beta}_1 = \frac{2}{n} \sum_{k=1}^{n/2} \frac{Y_{i_k} - Y_{j_k}}{x_{i_k} - x_{j_k}}.$$

For instance, if we had $n = 6$ observations and paired the indices as $(1, 6), (2, 5), (3, 4)$, the estimator would be

$$\tilde{\beta}_1 = \frac{1}{3}\left[\frac{Y_1 - Y_6}{x_1 - x_6} + \frac{Y_2 - Y_5}{x_2 - x_5} + \frac{Y_3 - Y_4}{x_3 - x_4}\right].$$

(a) Show that $\tilde{\beta}_1$ is an unbiased estimator of β_1.

(b) Without any additional calculations, what can you say about the variance of $\tilde{\beta}_1$ as it compares to the variance of the usual least squares estimator $\hat{\beta}_1$?

a) $E[\tilde{\beta}_1] = \frac{2}{n}\sum_{k=1}^{n/2} \frac{E[Y_{ik} - Y_{jk}]}{X_{ik} - X_{jk}} = \frac{2}{n}\sum_{k=1}^{n/2} \frac{E[\beta_1(X_{ik} - X_{jk}) + \varepsilon_i - \varepsilon_j]}{X_{ik} - X_{jk}}$

$= \frac{2}{n}\sum_{k=1}^{n/2} \beta_1$

$= \frac{2}{n}\frac{n}{2}\beta_1 = \beta_1$

b) $V(\tilde{\beta}_1) \geq V(\hat{\beta}_1)$ because $\hat{\beta}_1$ is a BLUEs estimator

Problem 4. An economist compiled data on the productivity improvements last year for a sample of firms producing electronic computing equipment. The firms were classified according to the level of their average expenditures for research and development in the past three years (Low and Moderate). The results of the study follows:

Rating	1	2	3	4	5	6	7	8	9	10	11	12
Low	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0			
Moderate	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4

(a) Use R to prepare side-by-side box plots for the two samples. Do the spreads seems to differ across samples?

(b) Use a graph to determine whether the assumption of normality is reasonable.

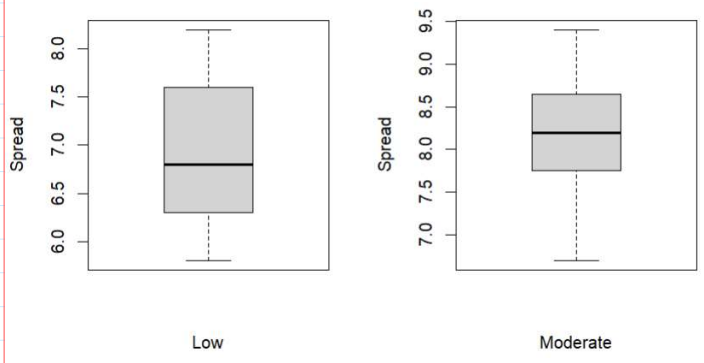
(c) Using R, obtain the p-value from the test of $H_0 : \sigma_1^2 = \sigma_2^2$. What do you conclude here? Note that you can only use a cumulative distribution function in R.

(d) Using a p-value from a t-test, test at $\alpha = 0.05$ whether the firms rated Moderate have a significantly higher mean productivity improvement than those rated Low. Note that you can only use a cumulative distribution function in R. Also note that you may want to consider what you have discovered in parts (a)-(c))

(e) Obtain and interpret a 95% CI for the difference in mean productivity improvement between firms rated Moderate and those rated Low.

a)

```
#Question 4a
low<-c(7.6,8.2,6.8,5.8,6.9,6.6,6.3,7.7,6.0)
mod<-c(6.7,8.1,9.4,8.6,7.8,7.7,8.9,7.9,8.3,8.7,7.1,8.4)
par(mfrow = c(1,2))
boxplot(low,xlab="Low",ylab="Spread")
boxplot(mod,xlab="Moderate",ylab="Spread")
```



The moderate data seems more clustered than the low data

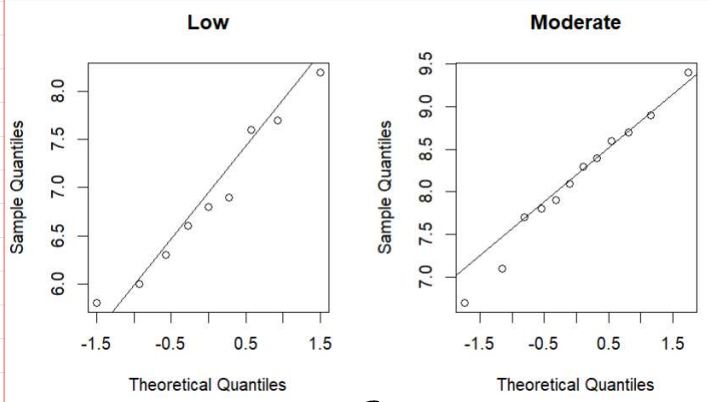
b)

```
#Question 4b
par(mfrow = c(1,2))
qqnorm(low,main="Low")
```

These graphs make it seem

b)
```
#Question 4b
par(mfrow = c(1,2))
qqnorm(low,main="Low")
qqline(low)
qqnorm(mod,main="Moderate")
qqline(mod)
```

These graphs make it seem like a reasonable assumption



c)
```
#Question 4c
lsd<-sd(low)
msd<-sd(mod)
nl<-length(low)
nm<-length(mod)

pp<-pf(lsd^2/msd^2,nl-1,nm-1)
p<-2*min(pp,1-pp)
print(p)
] 0.8035586
```

This p-value is large enough to fail to reject $\sigma_1^2 = \sigma_2^2$

d)
```
> #Question 4d
> sp2<-((nl-1)*lsd^2+(nm-1)*msd^2)/(ntot-2)
> ts<-(mean(mod)-mean(low))/sqrt(sp2/(ntot-2))
> pv<-pt(ts,ntot-2)
> if(1-pv<0.05){cat("Reject")}else{cat("Fail to reject")}
Reject
```

We reject $H_0: \mu_{mod} = \mu_{low}$

e)
```
> #Question 4e
> lowerbound<-mean(mod)-mean(low)-qt(0.975,ntot-2)*sqrt(sp2/(ntot-2))
> upperbound<-mean(mod)-mean(low)+qt(0.975,ntot-2)*sqrt(sp2/(ntot-2))
> cat("CI=(",lowerbound,",", upperbound,")")
CI=( 0.880331 , 1.63078 )
```

Problem 5. Refer again to the salt concentration data. Assume a model of the form

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

is appropriate, where Y is the salt concentration and x is the roadway area. Do the following **by hand**. (You can use the following results: $\bar{x} = 0.824$, $\bar{y} = 17.135$, $\sum_i x_i^2 = 17.2502$, $\sum_i x_i y_i = 346.793$.)

(a) Find $\hat{\beta}_0$ and $\hat{\beta}_1$ and 95% confidence intervals about β_0 and β_1. Give the estimated regression equation. You may use the fact that, for this model fit, $MSE = 3.21$.

(b) Estimate the mean salt concentration when 1.5 percent of the watershed area consists of paved roads. Provide the 95% confidence interval. You may use the fact that, for this model fit, $MSE = 3.21$.

(c) Predict what salt concentration we would actually observe when 1.5 percent of the watershed area consists of paved roads. Give the 95% prediction interval. You may use the fact that, for this model fit, $MSE = 3.21$.

Salt	Area	
3.8	0.19	1
5.9	0.15	2
14.1	0.57	3
10.4	0.4	4
14.6	0.7	5
14.5	0.67	6
15.1	0.63	7
11.9	0.47	8
15.5	0.75	9
9.3	0.6	10
15.6	0.78	11
20.8	0.81	12
14.6	0.78	13
16.6	0.69	14
25.6	1.3	15
20.9	1.05	16
29.9	1.52	17
19.6	1.06	18
31.3	1.74	19
32.7	1.62	20

a) $\frac{b_1 - \beta_1}{\sqrt{MSE / \sum (x_i - \bar{x})^2}} \sim t_{n-2}$

$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$

Note
```
> qt(0.975,18)
[1] 2.100922
```

So $\frac{b_1}{CI} = b_1 \pm t_{0.975,18} \sqrt{MSE / (\sum x_i^2 - n\bar{x}^2)}$

$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum x_i^2 - n\bar{x}^2}$

$= \frac{346.793 - 20(0.824)(17.135)}{17.2502 - 20(0.824)^2} \pm 2.100922 \sqrt{\frac{3.21}{17.2502 -}}$

$b_1 = \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}$

$= \frac{\sum x_i y_i - y_i\bar{x} - x_i\bar{y} + \bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$

$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$

$= \frac{346.793 - 20(0.824)(17.135)}{17.2502 - 20(0.824)^2} \pm 2.100922\sqrt{\frac{3.21}{17.2502 - 20(0.824)^2}}$

$= (15.582, 19.511)$

$b_0 = \bar{y} - b_1\bar{x}$

Estimated regression

$Y = 2.677 + 17.547x$

b_0

$CI = b_0 \pm t_{0.975,18}\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)}$

$= 17.135 - \frac{346.793 - 20(0.824)(17.135)}{17.2502 - 20(0.824)^2}0.824 \pm 2.100922\sqrt{3.21\left(\frac{1}{20} + \frac{(0.824)^2}{17.2502 - 20(0.824)^2}\right)}$

$= (0.852, 2.677)$

b) $E[Y_{1.5}] = 2.677 + 17.547(1.5/100)$

$= 2.940$

$CI = 2.940 \pm 2.100922\sqrt{3.21\left(\frac{1}{20} + \frac{(0.015 - 0.824)^2}{17.2502 - 20(0.824)^2}\right)}$

$= (1.141, 4.738)$

c) $CI = 2.940 \pm 2.100922\sqrt{3.21\left(\frac{1}{20} + \frac{(0.015 - 0.824)^2}{17.2502 - 20(0.824)^2} + 1\right)}$

$= (-1.232, 7.111)$

negative numbers don't make sense so $(0, 7.111)$

Problem 6. Discussion questions: Provide a brief response summarizing your thoughts.

(a) The regression function relating the production output (Y) by an employee taking a training program for (X) hours, where X ranges from 40 to 100. An observer concludes that the training program does not increase productivity since β_1 is less than 1. Comment.

(b) Evaluate the statement "For the least squares method to be fully valid, it is required that the distribution of Y be normal." Comment.

(c) Recall, we saw that $\sum_{i=1}^n e_i = 0$ and we have used e_i as a surrogate for the unobserved ϵ_i, thus does this imply that $\sum_{i=1}^n \epsilon_i = 0$. Comment.

a) It depends on if $\beta_1 < 0$, not if $\beta_1 < 1$ since β_1 is the slope. $0 \leq \beta_1 \leq 1$ still has positive growth

b) The only thing we assumed was that $M_{\varepsilon_i} = 0$ and $Cov(\varepsilon_i, \varepsilon_j) = \sigma^2\delta_{ij}$. So the normal assumption was not needed for the least squares approach.

normal assumption was not needed for the least squares approach.

c) The e_i's are observed values
However, since ε_i's are RV's
we cannot expect $\sum \varepsilon_i = 0$ all the time