Midterm 1: Solutions

STAT 420/MATH 469   Section N1

February 25, 2014

Exercise 1 (9 Points)

Every summer, Sam's Cycle Shop sells X road bicycles and Y mountain bicycles. Assume X and Y jointly follow a bivariate normal distribution with parameters: $\mu_X = 200, \quad \sigma_X = 20, \quad \mu_Y = 140, \quad \sigma_Y = 19, \quad \rho = 0.8$

(a) $Y \sim \mathrm{N}(140, 19^2)$ and need to find $P(Y < 140)$.
$P(Y < 140) = P(Z < \frac{140-140}{19}) = P(Z < 0) = \mathbf{0.5}$

(b) Need to find $P(Y < 140|X = 215)$.
Given $x = 215$, we know that Y is normal with mean and variance

$$\mu^* = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = 140 + 0.8(19/20)(215 - 200) = 151.4$$
$$\sigma_*^2 = \sigma_Y^2(1 - \rho^2) = 19^2(1 - 0.8^2) = 129.96$$

So the solution is given by
$P(Y < 140|X = 215) = P(Z < \frac{140-151.4}{\sqrt{129.96}}) = P(Z < -1) = \mathbf{0.1587}$

(c) Need to find $P(X + Y > 414)$.
$(X + Y)$ is normally distributed.
$E(X + Y) = \mu_X + \mu_Y = 200 + 140 = 340$
$V(X + Y) = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y = 20^2 + 19^2 + 2(0.8)(20)(19) = 1369$
$P(X + Y > 414) = P(Z > \frac{414-340}{\sqrt{1369}}) = P(Z > 2) = \mathbf{0.0228}$

Exercise 2 (9 Points)

Suppose we have a random sample of $m = 9$ heights (in inches) from male students at the University of Illinois with sample statistics

$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = 70, \qquad s_x^2 = \tfrac{1}{8}\sum_{i=1}^{9}(x_i - \bar{x})^2 = 9$$

and a random sample of $n = 10$ heights (in inches) from female students at the University of Illinois with sample statistics

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 64.5, \qquad s_y^2 = \tfrac{1}{9}\sum_{i=1}^{10}(y_i - \bar{y})^2 = 8.5$$

Assume that male heights are normally distributed $X \sim \mathrm{N}(\mu_x, \sigma^2)$ and that female heights are normally distributed $Y \sim \mathrm{N}(\mu_y, \sigma^2)$.

(a) The 90% (two-sided) confidence interval for the average height of a male student at the University of Illinois is given by
$\bar{x} \pm t_{m-1}^{(\alpha/2)}\sqrt{s_x^2/m} = 70 \pm (1.859548)\sqrt{9/9} = [\mathbf{68.14045};\ \mathbf{71.85955}]$.

(b) To test $H_0 : \mu_x = 73$ versus $H_1 : \mu_x \neq 73$ using a significance level of $\alpha = 0.1$, you can use the confidence interval from part (a):
$73 \notin [68.14045;\ 71.85955] \Longrightarrow$ **Reject $\mathbf{H_0}$**

(c) To test $H_0 : \mu_x = \mu_y$ versus $H_1 : \mu_x \neq \mu_y$ using a significance level of $\alpha = 0.05$, use independent sample t test.

First, the pooled variance estimate is given by
$s_p^2 = \frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2} = \frac{(8)9+(9)8.5}{17} = 8.735294$
so the independent sample t test statistic is given by
$T^* = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_p^2}{m}+\frac{s_p^2}{n}}} = \frac{70-64.5}{\sqrt{\frac{8.735294}{9}+\frac{8.735294}{10}}} = 4.050125$

Comparing this to the critical t with 17 degrees-of-freedom:
$T^* = 4.050125 > t_{17}^{(.025)} = 2.109816 \Longrightarrow$ **Reject $\mathbf{H_0}$**

Exercise 3 (9 Points)

Suppose that the assembly time Y (in minutes) for a particular computer has a linear relationship with the number of custom specifications X. The below data represent a random sample of $n = 6$ assembly times corresponding to different numbers of custom specifications. Consider the simple linear regression model: $y_i = b_0 + b_1 x_i + e_i$ with $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$.

	x	y	x^2	y^2	xy
	0	12	0	144	0
	2	14	4	196	28
	4	18	16	324	72
	6	23	36	529	138
	8	34	64	1156	272
	10	55	100	3025	550
$\sum$	30	156	220	5374	1060

(a) The least-squares slope estimate is given by
$\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{1060 - (6)(5)(26)}{220 - (6)(5)(5)} = \frac{280}{70} = \mathbf{4}$
and the least-squares intercept estimate is given by
$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = 26 - 4(5) = \mathbf{6}$

(b) The fitted values and residuals are given in the table below:

	$6 + 4x = \widehat{\mathbf{y}}$	$y - \hat{y} = \widehat{\mathbf{e}}$
	6	**6**
	14	**0**
	22	**−4**
	30	**−7**
	38	**−4**
	46	**9**
$\sum$	156	0

(c) The mean-squared error is an unbiased estimate of σ^2:
$\hat{\sigma}^2 = \frac{1}{4} \sum_{i=1}^6 \hat{e}_i^2 = \frac{1}{4}(36 + 0 + 16 + 49 + 16 + 81) = \mathbf{49.5}$

Exercise 4 (9 Points)

Consider the simple linear regression model: $y_i = b_0 + b_1 x_i + e_i$ with $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. Suppose that a simple linear regression model was fit to sample of $n = 10$ observations and the statistics include

$$\sum_{i=1}^{n} x_i = 275, \qquad \sum_{i=1}^{n} x_i^2 = 9625, \qquad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 2062.5,$$

$$\hat{\sigma} = \sqrt{\frac{175.6}{8}}, \qquad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.46666667 & -0.0133333333 \\ -0.01333333 & 0.0004848485 \end{pmatrix},$$

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x = 2.2 + 0.5x$$

(a) To test $H_0 : b_0 = 0$ versus $H_1 : b_0 \neq 0$ using a significance level of $\alpha = 0.1$, calculate the t test statistic:
$T = \frac{2.2-0}{\sqrt{\frac{175.6}{8}(0.46666667)}} = 0.6873881 < t_{n-2}^{(\alpha/2)} = 1.859548 \Longrightarrow \textbf{Retain } \mathbf{H_0}$

(b) To form a 90% confidence interval for b_1, use:
$\hat{b}_1 \pm t_{n-2}^{(\alpha/2)} \hat{\sigma}_{\hat{b}_1} = 0.5 \pm (1.859548)\sqrt{\frac{175.6}{8}(0.0004848485)} = [\mathbf{0.3082};\ \mathbf{0.6918}]$

(c) To test $H_0 : E(Y|X = 10) = 12$ versus $H_1 : E(Y|X = 10) \neq 12$ using $\alpha = 0.1$, calculate the t test statistic.

First note that the variance of the prediction with $X = 10$ is given by:

$$\begin{aligned}
\hat{\sigma}^2_{Y|X=10} &= \frac{175.6}{8} \begin{pmatrix} 1 & 10 \end{pmatrix} \begin{pmatrix} 0.46666667 & -0.0133333333 \\ -0.01333333 & 0.0004848485 \end{pmatrix} \begin{pmatrix} 1 \\ 10 \end{pmatrix} \\
&= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = \frac{175.6}{8} \left(\frac{1}{10} + \frac{(10 - 27.5)^2}{2062.5} \right) \\
&= 5.454242
\end{aligned}$$

Next note that with $X = 10$ we would predict $\hat{y} = 2.2 + 0.5(10) = 7.2$, so the t test statistic is given by:
$T = \frac{\hat{y}-12}{\hat{\sigma}_{Y|X=10}} = \frac{7.2-12}{\sqrt{5.454242}} = -2.055294$
$T = -2.055294 < -t_{n-2}^{(\alpha/2)} = -1.859548 \Longrightarrow \textbf{Reject } \mathbf{H_0}$

Exercise 5 (9 Points)

Suppose that a simple linear regression model was fit using the below R code:

```
> mymod = lm(y ~ x)
> anova(mymod)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1 515.625  515.625  23.491 0.001277 **
Residuals  8 175.600   21.950
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

(a) To find the sample variance of y, i.e., $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, note that $n = 10$ (because df for residuals is $n - 2 = 8$) and

$$\begin{aligned}\sum_{i=1}^{10}(y_i - \bar{y})^2 &= \sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{10}(y_i - \hat{y}_i)^2 \\ &= 515.625 + 175.600 = 691.225\end{aligned}$$

which implies that the sample variance is given by:
$s_y^2 = \frac{1}{9}\sum_{i=1}^{10}(y_i - \bar{y})^2 = 691.225/9 = \mathbf{76.80278}$

(b) To test $H_0 : b_1 = 0$ versus $H_1 : b_1 \neq 0$ using a significance level of $\alpha = 0.01$, just report the ANOVA F test:
$F = \mathbf{23.491} \sim F_{1,8}$, p-value=**0.0013**, **Reject $\mathbf{H_0}$**.

(c) The proportion of variation in y_i that can be explained by the linear relationship with x_i is the model R^2:
$R^2 = SSR/SST = 515.625/691.225 = \mathbf{0.7459583}$

Exercise 6 (5 Points)

Use the below R code to answer this question:

```
> x=c(-9, -7, -5, -3, -1, 1, 3, 5, 7, 9)
> y=c(60, 40, 35, -15, 0, 5, -20, -25, -40, -60)
> mean(x)
1] 0
> mean(y)
[1] -2
> sum(x^2)
[1] 330
> sum((y+2)^2)
[1] 12860
> sum(x*y)
[1] -1950
```

(a) To test $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ using a significance level of $\alpha = 0.1$, calculate the t test statistic. First, note that

$$r = \frac{\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{10}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{10}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{10} x_i y_i}{\sqrt{\sum_{i=1}^{10} x_i^2}\sqrt{\sum_{i=1}^{10}(y_i + 2)^2}} \qquad \text{(because } \bar{x} = 0,\ \bar{y} = -2\text{)}$$

$$= \frac{-1950}{\sqrt{(330)12860}} = -0.9465796$$

is the sample correlation coefficient. So, the t test statistic is given by:
$T = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}} = \frac{\sqrt{8}(-0.9465796)}{\sqrt{1-(-0.9465796)^2}} = -8.302572$
$T = -8.302572 < t_8^{(.95)} = -1.859548 \Longrightarrow$ **Reject $\mathbf{H_0}$**

(b) Considering the SLR model: $y_i = b_0 + b_1 x_i + e_i$ with $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, the coefficient of determination is $R^2 = r^2 = (-0.9465796)^2 = \mathbf{0.896013}$