Tutorial 11: Solutions

April 4, 2018

Question 10.1.5, Page 419

Consider the following summary data on the modulus of elasticity (3×106 psi) for lumber of three different grades:

Grade	J	$\overline{x}_{i\cdot}$	s_i
1	10	1.63	0.27
2	10	1.56	0.24
3	10	1.42	0.26

Use this data and a significance level of 0.01 to test the null hypothesis of no difference in mean modulus of elasticity for the three grades.

We first construct the ANOVA table.

Source	df	SS	MS	F-Value
Grade	a	d	g	F
Error	b	e	h	
Total	c	i		

The Grand Mean, $\overline{x}_{\cdot\cdot} = \frac{1}{I}\sum_{i=1}^{3}\overline{x}_{i\cdot} = \frac{1}{3}(1.63+1.56+1.42) = \frac{4.61}{3}$. We leave this as fraction so that our future calculations are precise.

$$
\begin{aligned}
a &= I - 1 = 3 - 1 = 2 \\
b &= I(J-1) = 3(10-1) = 3(9) = 27 \\
c &= a + b = 2 + 27 = 29
\end{aligned}
$$

$$
\begin{aligned}
d &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2 = J\sum_{i=1}^{I}(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2 \\
&= 10\left(\left(1.63 - \frac{4.61}{3}\right)^2 + \left(1.56 - \frac{4.61}{3}\right)^2 + \left(1.42 - \frac{4.61}{3}\right)^2\right) \\
&= \frac{343}{1500}
\end{aligned}
$$

$$
\begin{aligned}
e &= \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \overline{x}_{i\cdot})^2 = \frac{J-1}{J-1}\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \overline{x}_{i\cdot})^2 = (J-1)\sum_{i=1}^{I}{s_i}^2 \\
&= (10-1)\left(0.27^2 + 0.24^2 + 0.26^2\right) = 1.7829
\end{aligned}
$$

$$
i = d + e = \frac{343}{1500} + 1.7829 = 2.012
$$

$$g = \frac{d}{a} = \frac{\frac{343}{1500}}{2} = \frac{343}{3000}$$

$$h = \frac{e}{b} = \frac{1.7829}{27}$$

$$F = \frac{g}{h} = \frac{\frac{343}{3000}}{\frac{1.7829}{27}} = 1.731$$

$H_0 : \mu_1 = \mu_2 = \mu_3$, H_A : At least one pair unequal

$F_{I-1,I(J-1),0.01} = F_{2,27,0.01} = 5.49$ (2 is the numerator degrees of freedom, 27 is the denominator degrees of freedom; make sure you know how to find these values on the F-table). We reject H_0 if $F > F_{2,27,0.01}$. Since $1.731 \not> 5.49$, we fail to reject H_0 at the 1% level of significance. We conclude that there is insufficient evidence that there is a difference in mean modulus of elasticity between the three grades.

Question 10.1.6, Page 419

An article reports the following data on total Fe for four types of iron formation (1 = carbonate, 2 = silicate, 3 = magnetite, 4 = hematite).

Category	Fe Amount									
1	20.5	28.1	27.8	27.0	28.0	25.2	25.3	27.1	20.5	31.3
2	26.3	24.0	26.2	20.2	23.7	34.0	17.1	26.8	23.7	24.9
3	29.5	34.0	27.5	29.4	27.9	26.2	29.9	29.5	30.0	35.6
4	36.5	44.2	34.1	30.3	31.4	33.1	34.1	32.9	36.3	25.5

Carry out an analysis of variance F-test at significance level 0.01, and summarize the results in an ANOVA table.

Take the same steps as the previous question, except you have to calculate all the sample means and standard deviations on your own. This question is not particularly difficult but somewhat time consuming, try it out!

Question 10.1.7, Page 419

An experiment was carried out to compare electrical resistivity for six different low-permeability concrete bridge deck mixtures. There were 26 measurements on concrete cylinders for each mixture; these were obtained 28 days after casting. Fill in the missing entries and test appropriate hypotheses.

Source	df	SS	MS	F-value
Mixture	a	d	g	F
Error	b	e	13.929	
Total	c	5664.415		

a: "six different mixtures" $\Longrightarrow I = 6 \Longrightarrow a = I - 1 = 5$.
b: "26 measurements for each mixture" $\Longrightarrow J = 26 \Longrightarrow b = I(J-1) = 6(26-1) = 150$.
$c = a + b = 5 + 150 = 155$
$\frac{e}{b} = \frac{e}{150} = 13.929 \Longrightarrow e = (13.929)(150) = 2089.35$
$d + e = d + 2089.35 = 5664.415 \Longrightarrow d = 5664.415 - 2089.35 = 3575.065$
$g = \frac{d}{a} = \frac{3575.065}{5} = 715.013$
$F = \frac{g}{13.929} = \frac{715.013}{13.929} = 51.33.$
$H_0 : \mu_1 = \mu_2 = \ldots = \mu_5 = \mu_6$, H_A : At least one pair unequal

$F_{5,150,0.05}$ = some value between 2.31 and 2.26. We reject H_0 if $F > F_{5,150,0.05}$. Since $51.33 > F_{5,150,0.05}$, we can reject H_0 at the 5% significance level. We conclude that the data suggests at least one pair of the means are not equal (or equivalently we can say that the means are not all identical).

Question 12.2.17, Page 507

A least squares analysis in studying how y - porosity (%), is related to x - unit weight(pcf) in concrete specimens. Consider the following representative data (note that the x value corresponds to the y value given immediately below it):

x	99.0	101.1	102.7	103.0	105.4	107.0	108.7	110.8
y	28.8	27.9	27.0	25.2	22.8	21.5	20.9	19.6
x	112.1	112.4	113.6	113.8	115.1	115.4	120.0	
y	17.1	18.9	16.0	16.7	13.0	13.6	10.8	

Relevant summary quantities are:

$$\sum_{i=1}^{n} x_i = 1640.1, \quad \sum_{i=1}^{n} y_i = 299.8, \quad \sum_{i=1}^{n} x_i y_i = 32,308.59$$

$$\sum_{i=1}^{n} {x_i}^2 = 179,849.73, \quad \sum_{i=1}^{n} {y_i}^2 = 6430.06$$

(a) Obtain the equation of the estimated regression line. Then create a scatterplot of the data and graph the estimated line. Does it appear that the model relationship will explain a great deal of the observed variation in y?

Remember that when you are referring to predicted variables that you remember to include the "hat" on top otherwise you will lose marks!

$$\begin{aligned}\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum {x_i}^2 - n\bar{x}^2} \\ &= \frac{32308.59 - (15)\left(\frac{1640.1}{15}\right)\left(\frac{299.8}{15}\right)}{179849.73 - (15)\left(\frac{1640.1}{15}\right)^2} \\ &= \frac{-471.542}{521.196} \\ &\approx -0.9047\end{aligned}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \left(\frac{299.8}{15}\right) - \left(\frac{-471.542}{521.196}\right)\left(\frac{1640.1}{15}\right)$$
$$\approx 118.91$$

The equation of our regression line is $\hat{y} = 118.91 - 0.9047x$. Note that for future calculations you should use the exact numbers obtained from the calculations of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ but I'm being lazy. Don't be me.

To perform the regression in `R`, first import the data into the variables `x` and `y`. Then:

```
fit <- lm(y ~ x)
plot(x, y)
lines(x, predict(fit))
```

We can also check the coefficients by calling:

```
summary(fit)
```

From the plot, we observe a very strong linear relationship. We expect that this model will explain much of the variability in y.

(b) Interpret the slope of the least squares line.

The slope is -0.9047. We expect a decrease in y - perosity (%) by 0.9047 per unit weight of concrete (x).

(c) What happens if the estimated line is used to predict porosity when unit weight is 135? Why is this not a good idea?

We first notice that the unit weight of 135 is already out of the scope of our given x values. We don't expect that plugging 135 into our regression equation will give us a good result. Plugging in $x = 135$ into our regression equation, we obtain $\hat{y} = -3.2245$ but we cannot have negative y values since y is a percentage and a negative percentage doesn't make sense in any context.

(d) Calculate the residuals corresponding to the first two observations.

$$e_1 = y_1 - \hat{y}_1 = 28.8 - (118.91 - (0.9047)(99)) = -0.5447$$
$$e_2 = y_2 - \hat{y}_2 = 27.9 - (118.91 - (0.9047)(101.1)) = 0.45517$$

Residuals are measuring the distance from the predicted point to your observed point. For example, e_1 being a negative value means that our observed value is less than our predicted value, so our regression equation overestimates at this point. Recall from the derivation of the least squares coefficient estimates that we are trying to minimize the sum of the square of these residuals, i.e. minimize the absolute distance between each observed point to the regression line.

(e) Calculate and interpret a point estimate of σ.

$$\text{SST} = S_{yy} = \sum {y_i}^2 - n\bar{y}^2$$
$$= 6430.06 - (15)\left(\frac{299.8}{15}\right)^2$$
$$= \frac{328543}{750}$$

$$
\begin{aligned}
\text{SSE} &= \sum (y_i - \hat{y}_i)^2 \\
&= S_{yy} - \widehat{\beta}_1 S_{xy} \\
&= \frac{328543}{750} - \left(\frac{-471.542}{521.196}\right)(-471.542) \\
&= 11.4388
\end{aligned}
$$

$$
\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2} = \frac{11.4388}{13} = 0.8799
$$

$$
\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} = 0.9388
$$

$\widehat{\sigma}$ represents the typical deviation from the regression line. Since the units of y is in %, we are typically within $\pm 0.9388\%$ from the true percentage.

(f) What proportion of observed variation in porosity can be attributed to the approximate linear relationship between unit weight and porosity?

$R^2 = 1 - \dfrac{\text{SSE}}{\text{SST}} = 1 - \dfrac{11.4388}{\frac{328543}{750}} = 0.974$

97.4% of the variability in the observed y can be explained by the regression model.

Question 12.3.31, Page 517

During oil drilling operations, components of the drilling assembly may suffer from sulfide stress cracking. An article reported on a study in which the composition of a standard grade of steel was analyzed. The following data on y - threshold stress (% SMYS), and x - yield strength (MPa), was read from a graph in the article (which also included the equation of the least squares line).

x	635	644	711	708	836	820	810	870	856	923	878	937	948
y	100	93	88	84	77	75	74	63	57	55	47	43	38

$$
\sum x_i = 10,576, \qquad \sum y_i = 894, \qquad \sum x_i y_i = 703,192
$$

$$
\sum {x_i}^2 = 8,741,264, \qquad \sum {y_i}^2 = 66,224
$$

(a) What proportion of observed variation in stress can be attributed to the approximate linear relationship between the two variables?

$$
\begin{aligned}
\text{SSE} &= S_{yy} - \widehat{\beta}_1 S_{xy} = S_{yy} - \frac{{S_{xy}}^2}{S_{xx}} \\
&= \sum {y_i}^2 - n\bar{y}^2 - \frac{(\sum x_i y_i - n\bar{x}\bar{y})^2}{\sum {x_i}^2 - n\bar{x}^2} \\
&= 66224 - (13)\left(\frac{894}{13}\right)^2 - \frac{\left(703192 - (13)\left(\frac{10576}{13}\right)\left(\frac{894}{13}\right)\right)^2}{8741264 - (13)\left(\frac{10576}{13}\right)^2} \\
&= 509.505
\end{aligned}
$$

$$
R^2 = 1 - \frac{\text{SSE}}{\text{SST}}
$$

$$= 1 - \frac{\text{SSE}}{S_{yy}}$$

$$= 0.8926$$

0.8926 is the proportion of observed variation in stress that can be attributed to the approximate linear relationship between two variables.

(b) Compute the estimated standard deviation $s_{\widehat{\beta}_1}$.

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

$$= \frac{509.505}{11} = 46.3187$$

$$\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$$

$$= \sqrt{46.3187} = 6.806$$

$$\text{se}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{S_{xx}}}$$

$$= 0.018368$$

(c) Calculate a confidence interval using confidence level 95% for the expected change in stress associated with a 1 MPa increase in strength. Does it appear that this true average change has been precisely estimated?

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{703192 - (13)\left(\frac{10576}{13}\right)\left(\frac{894}{13}\right)}{8741264 - (13)\left(\frac{10576}{13}\right)^2}$$

$$= -0.1756$$

$t_{n-2,\frac{\alpha}{2}} = t_{11,0.025} = 2.201$. Our 95% CI is:

$$\widehat{\beta}_1 \pm t_{11,0.025} \cdot \text{se}(\widehat{\beta}_1)$$
$$= -0.1756 \pm (2.201)(0.018368)$$
$$= (-0.21606, \ -0.13521)$$

It appears that this true average change associated to a unit increase in strength has been precisely estimated as our CI has a very narrow width.