

Updated solutions reflecting the format of the 5/01/09 exam 3 have been uploaded to the website.

Wednesday's lecture we finished up our discussion of one factor ANOVA (for the time being at least) by recalling the **Model equation for one way classification** :

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{for } i=1,2,\dots,k; \quad j=1,2,\dots,n_i$$

where the ϵ_{ij} are independent normals with zero means and common variance σ^2 . Here

$\mu_i = \mu + \alpha_i$ gives the mean of the i^{th} population. The null hypothesis in this formulation says that with $\alpha_i =$ the **effect** of the i^{th} treatment, all the effects are zero or

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0 \quad .$$

Johnson remarks that the best estimates of the parameters under a **least squares criterion** are

$$\hat{\mu} = \bar{y} = \text{grand mean} \quad , \quad \hat{\alpha}_i = \bar{y}_i - \bar{y} \quad , \quad \hat{\mu}_i = \bar{y}_i$$

We now turn to discuss what this least squares method means for the case of **linear regression** where for paired (X,Y) data that looks fairly linear we try to best fit a straight line to the data. Here we regard the given x-values as fixed , while the Y-values are random. While one can do least squares regression on any such linear looking data set, in order to analyze the data, make inferences and derive confidence intervals etc one typically makes for a

Model equation : a normal assumption of the form

$$Y = \alpha + \beta x + \epsilon \quad \text{where } \epsilon \sim N(0, \sigma^2)$$

is normal with mean 0 and variance σ^2 . Said differently, the random variable Y here depends upon x and is normally distributed with mean and variance

$$E[Y] = \alpha + \beta x \quad (\text{the } \mathbf{linear\ regression\ line}) \quad \text{and} \quad V[Y] = \sigma^2 \quad .$$

Whether the normal assumption holds or not we can try to best fit the data $(X_i, Y_i) \quad i=1,2,\dots,n$ to a straight line (assuming it looks linear) where best means that we minimize over all real number constants a and b , the sum of squared deviations (to obtain the **least squares estimates** a and b for α and β):

$$S(a, b) = \sum_{i=1}^n (Y_i - a - b x_i)^2 \quad .$$

The optimal a and b found in this fashion will be unbiased estimators of the actual constants

α and β appearing in the model . The plot of such data we call a **scatter plot** or **scatter diagram** summarizing the two related variables X and Y . At the beginning of the course we looked at a scatter diagram of the space shuttle challenger data of number of failed O-rings versus temperature but there the plot did not look linear so some sort of **non-linear regression** would be appropriate.

We know from calculus that to minimize such a quadratic function of a and b we want to set the partial derivatives of the function $S(a, b)$ with respect to a and b equal to 0 and solve. Using the chain rule we get the so called **normal equations**

$$\frac{\partial S}{\partial a} = \sum_{i=1}^n 2(Y_i - a - b x_i) \cdot (-1) = 0$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^n 2(Y_i - a - b x_i) \cdot (-x_i) = 0$$

the first which we write as $\sum_i y_i - n a - b \sum_i x_i = 0$ or $a = \bar{y} - b \bar{x}$ and the second as

$$\sum_i x_i y_i - a \sum_i x_i - b \sum_i x_i^2 = 0$$

Eliminating $a = \bar{y} - b \bar{x}$ from the second equation gives

$$b = \frac{S_{xy}}{S_{xx}} \quad \text{where} \quad S_{xy} = \sum_i x_i y_i - \frac{1}{n} \left(\sum_i x_i \right) \left(\sum_i y_i \right) \quad \text{and} \quad S_{xx} = \sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2$$

which by a little algebra can be written as the equivalent expressions

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad \text{and} \quad S_{xx} = \sum_i (x_i - \bar{x})^2 .$$

One checks that

$$E[Y_i] = \alpha + \beta x_i \quad \text{and} \quad E[\bar{Y}] = \alpha + \beta \bar{x} \quad \text{so that} \quad E[Y_i - \bar{Y}] = \beta(x_i - \bar{x})$$

from which it follows from the formula for b (since we regard the x -values as fixed constants) that

$$E[b] = \beta$$

$$\text{and hence} \quad E[a] = E[\bar{Y} - b\bar{x}] = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$$

so that a and b are unbiased estimators of the parameters α and β as claimed.

Thus we have our unbiased estimator

$$\hat{Y} = a + bx \quad \text{of} \quad E[Y] = \alpha + \beta x \quad \text{and hence of} \quad E[Y_i] = \alpha + \beta x_i \quad \text{for given } x_i \quad \text{using} \quad \hat{Y}_i = a + bx_i .$$

The differences

$$Y_i - \hat{Y}_i = \text{observation} - \text{fitted value} = y_i - a - bx_i$$

are called **residuals**. The **error sum of squares** or **residual sum of squares** is

$$SSE = \sum_{i=1}^n (y_i - a - bx_i)^2 = S_{yy} - S_{xy}^2 / S_{xx} \quad \text{where} \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

which may be seen by squaring out and summing over i the decomposition

$$(y_i - a - bx_i) = (y_i - \bar{y}) - b(x_i - \bar{x}) + (\bar{y} - a - b\bar{x}) ,$$

noting several cross terms vanish since $\sum_i (x_i - \bar{x}) = 0$ and $\sum_i (y_i - \bar{y}) = 0$, and using our formulas

for our least squares estimators a and b as done in Johnson.

Estimate of variance : Our estimate of σ^2 is then s_e^2 where

$$s_e^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - a - bx_i)^2}{n-2} = \frac{S_{yy} - S_{xy}^2 / S_{xx}}{n-2} .$$

Since $\sum_i (x_i - \bar{x}) = 0$ one has

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x})y_i$$

Then since $V[Y_i] = \sigma^2$ one finds

$$V[b] = \frac{\sigma^2}{S_{xx}}$$

using the independence of the Y_i 's . By a slight abuse of notation we use b to denote both the sample statistic value (involving the observed y_i 's) and the random variable involving the Y_i 's giving rise to it (found by substituting the Y_i 's in place of the y_i 's) . Similarly we have

$$a = \bar{y} - b\bar{x} = \sum_i \left(\frac{-(x_i - \bar{x})\bar{x}}{S_{xx}} + \frac{1}{n} \right) y_i$$

which being (the value of) a linear combination of independent normals is (the value of) a normal random variable. From $\sum_i (x_i - \bar{x}) = 0$ and since $V[Y_i] = \sigma^2$ one finds from the definition of

S_{xx} with the same said slight abuse of notation that

$$V[a] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

Using the standard deviations to standardize the normal random variables a and b and replacing the unknown σ^2 by s_e^2 gives the form of the t -statistics in Theorem 11.1 in Johnson :

Theorem 11.1 Under the assumptions that the regression is linear in x and that the n random variables Y_i are independent normally distributed with means of $E[Y_i] = \alpha + \beta x_i$ and common variance $V[Y_i] = \sigma^2$ the statistics

$$t = \frac{a - \alpha}{s_e} \sqrt{\frac{n S_{xx}}{S_{xx} + n(\bar{x})^2}}$$

$$t = \frac{(b - \beta)}{s_e} \sqrt{S_{xx}}$$

are values of random variables having the t -distribution with $n-2$ degrees of freedom.

Corresponding confidence intervals then follow as given in Johnson.

We have seen that $\hat{Y}(x_0) = a + bx_0$ is an unbiased estimator of $E[Y(x_0)] = E[a + bx_0] = \alpha + \beta x_0$

From our formulas for a and b and recalling $S_{xy} = \sum_i (x_i - \bar{x})y_i$ we find its variance is

$$V[a + bx_0] = V\left[\bar{y} + \frac{S_{xy}}{S_{xx}}(x_0 - \bar{x})\right] = V\left[\sum_{i=1}^n \left(\frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} + \frac{1}{n}\right) y_i\right]$$

$$= \sum_{i=1}^n \left(\frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} + \frac{1}{n}\right)^2 \sigma^2 = \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \sigma^2$$

where to get the last expression to the right of the rightmost equal sign we square out the proceeding expression to the left of the equal sign, and have used $\sum_i (x_i - \bar{x}) = 0$ so that the cross term vanishes,

the definition of $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ and the assumption that $V[Y_i] = \sigma^2$.

This gives for a t -random variable with $n-2$ degrees of freedom the value

$$t = \frac{a + b x_0 - (\alpha + \beta x_0)}{s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$$

which leads to Johnson's $100(1 - \alpha)$ % confidence interval for $\alpha + \beta x_0$ obtained by solving for $\alpha + \beta x_0$ for the above t the inequality $-t_{\alpha/2} < t < t_{\alpha/2}$. By similar reasoning, to predict a future value of $Y = Y(x_0)$ using $a + bx_0$ as its estimator we find that since both are independent of one another, and are normally distributed, so is their difference which has mean 0 and variance

$$V[Y(x_0) - (a + bx_0)] = V[Y(x_0)] + V[a + bx_0] = \sigma^2 + V[a + bx_0] = \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \sigma^2$$

so that using $s_e = \sqrt{\frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2}}$ to estimate σ we get for a t random variable with $n-2$ degrees of

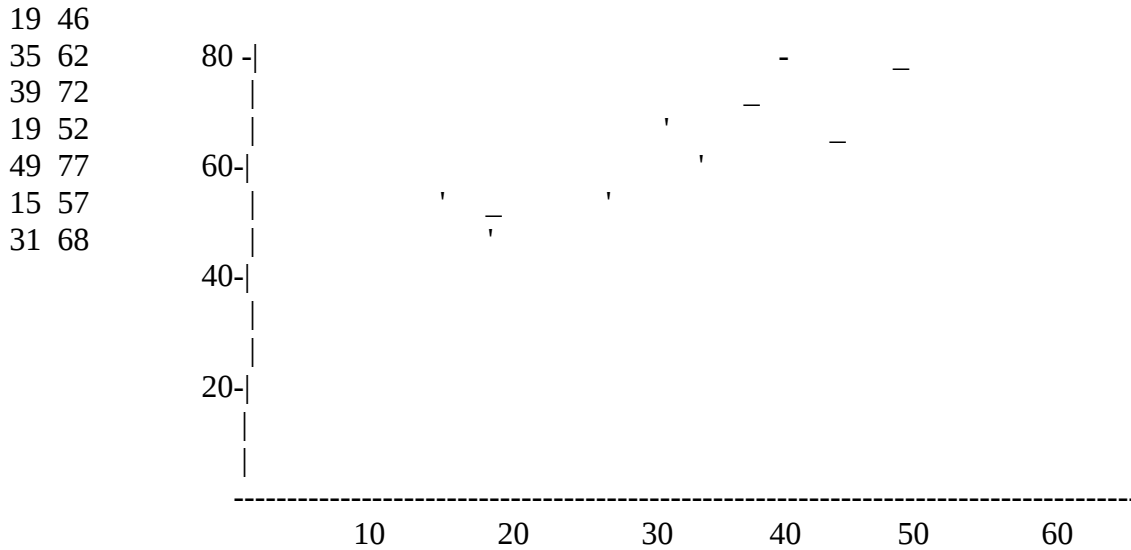
freedom the value $t = \frac{y - (a + bx_0)}{s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$ where $y = y(x_0) = y_{n+1}$ is the future value that we

wish to predict. This leads to the corresponding $100(1-\alpha)$ % prediction interval (prediction limits) given by Johnson obtained by solving for y the inequality $-t_{\alpha/2} < t < t_{\alpha/2}$.

EXAMPLE 1 modified Problem 11.3 of text (like HW problems 11.4, 11.5)

The extraction times x and extraction efficiencies y (a percentage %) are summarized by the table

x	y	
27	57	or in ascending order of x :
45	64	x : 15 19 19 27 31 35 39 41 45 49
41	80	y : 57 52 46 57 68 62 72 80 64 77
19	46	



a) Using my scatter plot of the data (which I have attempted to graph above) my eyeball estimate of $\hat{y}(35)$ is around 65 (the actual value obtained by least squares is about 65.792 or closer to 66) It is hard to graph this data in this fashion since there is not a lot of control over the vertical positioning

b) Using the least squares method we find with the $n = 10$ pairs of observations

$$\sum x_i = 320, \sum y_i = 635, \sum x_i y_i = 21275, \sum y_i^2 = 41395, \sum x_i^2 = 11490, S_{xy} = 955, S_{xx} = 1250$$

$$S_{yy} = 1072.5 \quad s_e = \sqrt{\frac{S_{yy} - S_{xy}^2 / S_{xx}}{n-2}} = \sqrt{42.86} = 6.54675$$

$$\text{or } b = S_{xy} / S_{xx} = .764, a = \bar{y} - b\bar{x} = 63.5 - (.764)32 = 39.052$$

so that

$$\hat{y} = 39.052 + .764x \text{ which gives } \hat{y}(35) = 39.052 + .764(35) = 65.792$$

as our estimate of the mean value of Y (the linear regression line value) when $x = 35$.

c) (like HW 11.5 a) Construct a 95% confidence interval for β , the increase in extraction efficiency per unit increase (one minute increase) in extraction time :

From $t_{.025,8} = 2.306$ for the t statistic with $n-2 = 8$ degrees of freedom

$$t = \frac{(b - \beta)}{s_e} \sqrt{S_{xx}} = \frac{.764 - \beta}{6.54675} \sqrt{1250}$$

we find the 95% confidence interval for β :

$$b \pm t_{.025} s_e / \sqrt{S_{xx}} = .764 \pm 2.306(6.54675) / \sqrt{1250} = .764 \pm .427$$

$$\text{or } [.3369975, 1.191]$$

d) Find a 95% prediction interval for a single future observation of extraction efficiency Y when the extraction time $x = 35$ minutes : Using (again with $t_{.025,8} = 2.306$)

$$t = \frac{y - (a + bx_0)}{s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} = \frac{y - 65.792}{6.54675 \sqrt{1 + \frac{1}{10} + \frac{(35 - 32)^2}{1250}}}$$

we arrive at the 95% prediction interval for the future value y of Y when $x = 35$ by solving for y the inequality $-t_{\alpha/2} < t < t_{\alpha/2}$ which gives for the P. I. :

$$(a + bx_0) \pm t_{.025,8} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 65.792 \pm (2.306) 6.54675 \sqrt{1 + \frac{1}{10} + \frac{(35 - 32)^2}{1250}}$$

or $65.792 \pm 15.885398 = [49.90660, 81.6774]$

Correlation In section 11.6 we now generalize the discussion of regression to the case where both the predictor variable X and the response variable Y can be random. When a scatter plot reveals points to scatter about a straight line we try to fit via least squares as before. Now the interpretation is that the conditional expectation of Y given $X = x$ gives the actual regression line (which we estimate by the least squares fitted line) :

$$E[Y | X = x] = \alpha + \beta x$$

When we studied expectation using the joint distribution of two random variables we discussed the **population correlation** ρ which is like a covariance but measured relative to the natural length scales given by the standard deviations of X and Y :

$$\rho = E \left[\left(\frac{X - \mu_1}{\sigma_1} \right) \left(\frac{Y - \mu_2}{\sigma_2} \right) \right] = \frac{Cov[X, Y]}{\sigma_X \sigma_Y} \text{ where } \sigma_1 = \sigma_X \text{ and } \sigma_2 = \sigma_Y$$

The **sample correlation** r used to estimate ρ is defined by

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} b \text{ since } s_x^2 = \frac{S_{xx}}{n-1} \text{ and } s_y^2 = \frac{S_{yy}}{n-1}$$

(the sample variances of X and Y). One has by the Cauchy- Schwarz inequality that

$$-1 \leq \rho \leq 1 \text{ and similarly } -1 \leq r \leq 1$$

Here $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \text{total sum of squares of } y = SST$. One has the

Decomposition of variance

SST	$=$	SSR	$+$	SSE
S_{yy}	$=$	S_{xy}^2 / S_{xx}	$+$	$S_{yy} - S_{xy}^2 / S_{xx}$
Total variability of y	$=$	variability explained by the linear relationship		residual or unexplained variability

or total sum of squares = regression sum of squares + the least squares error

where $SSR = \sum_i (a + bx_i - \bar{y})^2$.

That SSR can be written this way follows from the normal equations after squaring out and summing

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \text{ where } \hat{y}_i = a + bx_i$$

since by the normal equations the cross term vanishes.

The **correlation** r measures the strength of a linear relationship, being the strongest at 1 (all the points (X, Y) lie exactly on a line with positive slope) or at -1 (all points lie exactly on a line with negative slope). A value of r close to 0 indicates that a linear relationship if it exists is very weak.

$$r^2 = \frac{SSR}{SST} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \text{proportion of } y \text{ variability explained by the linear relationship}$$

To make further inferences we assume that the joint distribution of X and Y is given by the **bivariate normal distribution**

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]}$$

Inferences about the correlation ρ are based on the **Fisher Z transformation** :

$$\zeta = \frac{1}{2} \ln \frac{1+r}{1-r} \text{ is approximately normal with mean } \mu_\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \text{ and variance } \frac{1}{n-3}$$

$$\text{so } Z = \frac{\zeta - \mu_\zeta}{1/\sqrt{n-3}} \text{ is approximately standard normal and } r = \frac{e^\zeta - e^{-\zeta}}{e^\zeta + e^{-\zeta}} .$$

EXAMPLE 2 problem 11.49 of text (like HW 11.48)

For the air velocities (x) and evaporation coefficients (y) of the example on page 341 :

x : 20 60 100 140 180 220 260 300 340 380

y : .18 .37 .35 .78 .56 .75 1.18 1.36 1.17 1.65

a) Calculate r : $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{505.40}{\sqrt{132000(2.13745)}} = .95148137$ which gives $r^2 = .9053$

for the proportion of variation of y accounted for by the linear relationship.

b) Assuming necessary assumptions test

$$H_0: \rho = 0 \text{ against } H_a: \rho \neq 0 \text{ at significance level } \alpha = .05 .$$

We thus ignore the fact that the x data are not random and pretend that the pairs (X, Y) come from a bivariate normal distribution. Using the Fisher Z transform given on page 380 of the text, we find

$$Z = \sqrt{n-3} \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{\sqrt{7}}{2} \ln \frac{1+.95148}{1-.95148} = 4.887$$

is the value of an approximately standard normal random variable and hence

decision : we reject the null hypothesis

since this value lies outside the interval $-z_{.025} = -1.96 < Z < z_{.025} = 1.96$.

c) To get a 95% confidence interval for ρ we have

$$\zeta - \frac{z_{\alpha/2}}{\sqrt{n-3}} = 1.8471 - 1.96/\sqrt{7} = 1.10629 < \mu_\zeta < \zeta + \frac{z_{\alpha/2}}{\sqrt{n-3}} = 1.8471 + 1.96/\sqrt{7} = 2.58791$$

$$\text{where } \zeta = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{4.887}{\sqrt{7}} = 1.8471 \text{ so } r = \frac{e^\zeta - e^{-\zeta}}{e^\zeta + e^{-\zeta}} = \text{ and } \mu_\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

Inverting the relation $1.10629 < \mu_\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} < 2.58791$ gives our 95% confidence interval for ρ

:

$$\frac{e^{1.10629} - e^{-1.10629}}{e^{1.10629} + e^{-1.10629}} = .802747 < \rho < .98876 = \frac{e^{2.58791} - e^{-2.58791}}{e^{2.58791} + e^{-2.58791}} .$$

EXAMPLE 3 problem 11.72

Known concentration (x) versus measured concentration (y) for a calibration of measuring lead levels in water were

x : 0.00 0.00 1.25 1.25 2.50 2.50 2.50 5.00 10.00 10.00

y : .7 .5 1.1 2.0 2.8 3.5 2.3 5.3 9.1 9.4

a) A plot of measured concentration (y) versus known concentration (x) gives a linear looking scatter plot with a slope at least close to 1.

b) Fit the least squares line with these $n = 10$ observations :

$$\sum x_i = 35, \sum y_i = 36.7, \sum x_i^2 = 246.875, \sum y_i^2 = 230.59, \sum x_i y_i = 236.875$$

$$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n = 108.425, S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 124.375,$$

$$S_{yy} = \sum y_i^2 - (\sum y_i)^2/n = 95.901 \quad SSE = S_{yy} - S_{xy}^2/S_{xx} = 1.380553,$$

$$s_e^2 = SSE/(n-2) = .172569, s_e = .415414, b = S_{xy}/S_{xx} = .8717587,$$

$$a = \bar{y} - b\bar{x} = .618844221, \hat{y} = a + bx = .6188442 + .8717587x$$

c) Give the 95% confidence interval for the actual slope β :

$$b \pm t_{\alpha/2, n-2} s_e / \sqrt{S_{xx}} = .8717587 \pm (2.306)(.415414) / \sqrt{124.375}$$

$$= .8717587 \pm .0858963 = [.785862, .9576549]$$

d) Test $H_0: \beta = 1$ vs. $H_a: \beta \neq 1$ at significance level $\alpha = .05$

$$t = \frac{b-1}{s_e} \sqrt{S_{xx}} = -3.4428 \text{ falls outside of } \pm t_{.025, 8} = \pm 2.306 \text{ so reject } H_0 .$$

e) Test $H_0: \beta = 1$ vs. $H_a: \beta < 1$ at significance level $\alpha = .005$

$$t = \frac{b-1}{s_e} \sqrt{S_{xx}} = -3.4428 < -t_{.005, 8} = -3.355 \text{ so reject } H_0 .$$

f) What could go wrong if we tried to use the least square line to predict measured concentration when the known concentration is $x = 15$?

Since this falls outside of our range of observed data, it is possible that the linear relationship may no longer hold. We have no way of knowing.

g) problem 11.73 : for the proceeding problem find a 95% confidence interval for α the actual y intercept :

$$a \pm s_e t_{.025, 8} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} = .618844 \pm .415414(2.306) \sqrt{\frac{1}{10} + \frac{3.5^2}{124.375}} = .618844 \pm .426789 = [.1920, 1.046]$$

Remarks : There has to be some linear component to the data set for linear regression to be of use. One can imagine data organized into some curvilinear fashion but which has no preferred linear direction. The

model utility test tests whether there is evidence for a non-zero slope (linear dependence between X and Y) for the data. That is it tests the utility or usefulness of the hypothesis that a linear relationship exists :

$$H_0: \beta = 0 \text{ versus } H_a: \beta \neq 0 .$$

We have already seen that the t-statistic

$$t = \frac{(b-\beta)}{s_e} \sqrt{S_{xx}} = \frac{b}{s_e} \sqrt{S_{xx}} = \frac{\sqrt{n-2} S_{xy}}{\sqrt{S_{xx} S_{yy} - S_{xy}^2}} = \sqrt{\frac{SSR}{SSE/(n-2)}}$$

can be used in this context. Alternately, one can employ the F-statistic

$$F = \frac{SSR}{SSE/(n-2)} = F(1, n-2) \text{ where } F = t^2 \text{ and } t_{\alpha/2, n-2}^2 = F_{\alpha, 1, n-2}$$

We can summarize this discussion with an analysis of variance table for linear regression :

ANOVA table for simple linear regression :

Source of Variation	df	Sum of Squares	Mean Square	F
Regression	1	SSR	MSR = SSR	$\frac{SSR}{SSE/(n-2)}$
Error	n-2	SSE	MSE = $s_e^2 = \frac{SSE}{n-2}$	
Total	n-1	SST		

$$\alpha \text{ in } \hat{y} = \alpha + \beta x, \text{ namely } H_0: \alpha = 0 \text{ vs. } H_a: \alpha \neq 0 \text{ using } t = \frac{a}{s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} .$$

For a regression relating the number of units serviced X to the length of a service call Y in minutes typical computer output for the **minitab** software package when given some X data in column C1 and Y data in column C2 and given the command

> regress c2 1 c1

> predict 5 (i. e. predict the number of minutes Y when X = 5 units are serviced) gives :

MINUTES = 14.2 + 44.4 Serviced

Predictor	Coef	Stdev	t-ratio	p
Constant	14.187	4.230	3.35	.007
SERVICED	44.4139	.8474	52.41	.000

s = 5.804 R-sq = 99.6% R-sq (adj) 99.6%

Analysis of Variance

Source	DF	SS	MS	F	p
Regression	1	92547	92547	2747.18	.000
Error	10	337	34		
Total	11	92884			

Fit	St. Dev. Fit	95% C. I.	95% P.I.
236.26	1.71	(232.44, 240.07)	(222.77, 249.74)

The t-ratios (coef / stdev) and associated p-values are for the two model utility test mentioned above (the first testing if the constant y-intercept α is 0 and the second testing whether the slope β is 0).

The fitted value here is $\hat{y}(5) = 236.26 = a + b \cdot 5$, $s = s_e = 5.804$, $r^2 = 99.6\%$,

a) The number of data points used to determine the estimated regression equation is $N = 12$.

b) An estimate of the fixed setup time associated with a service call is $14.187 = \hat{\alpha} = a$.

c) An estimate of the time required per additional unit serviced is $44.4139 = \hat{\beta} = b$.

d) A measure of the percent of total variation explained by the least square line is

$$r^2 = 99.6\% = \frac{SSR}{SST} = \frac{92547}{92884} .$$

(The adjusted r-sq value is something that comes up in multiple regression when the response variable Y depends on more than one predictor variable)

e) A 95% confidence interval for the true time required per additional unit serviced is

$$b \pm t_{.025, 10} s_b = 44.4139 \pm (2.228)(.8474) = [42.5259, 46.3019] .$$

f) A 95% prediction interval for the time $y(5)$ required to complete an individual service call involving 5 servicings is [222.77, 249.74]

g) The estimated value of the error standard deviation σ is $s_e = s = 5.804$