Chapter 12 Class Notes – Linear Regression and Correlation

We'll skip all of §12.7 and parts of §12.8, and cover the rest.

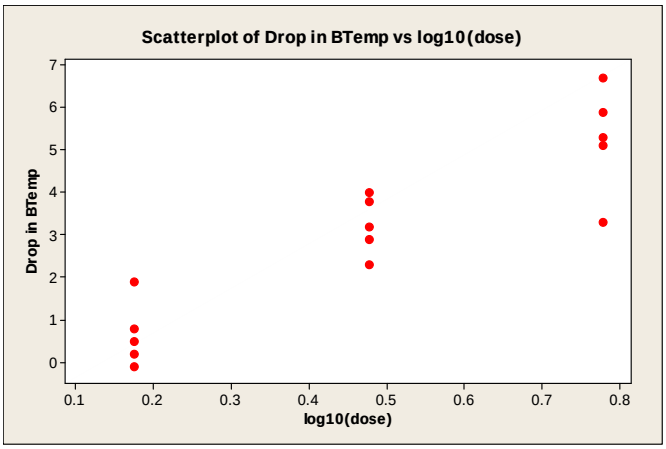
We'll consider the following two illustrations (graphs are below):

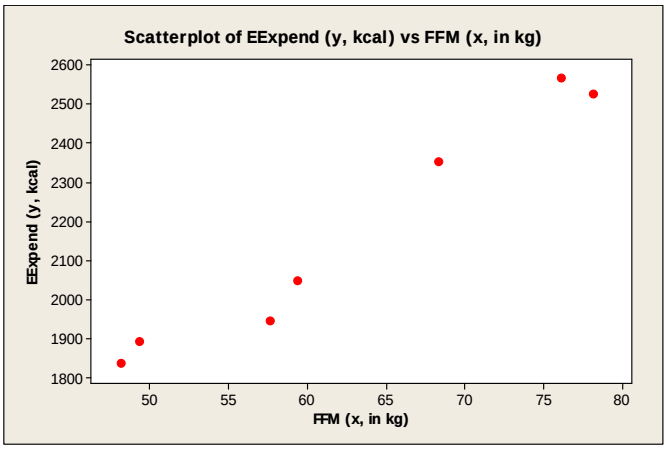
- **Example 1 (p.503 #12.3.2): y = drop in body temperature, x = $\log_{10}$(dose of ethanol)**
- **Example 2 (pp.491-2 #12.2.7): y = energy expenditure (EE, in kcal), x = fat-free mass (FFM, in kg)**

These illustrations demonstrate that there are two contexts for regression – both where x and y must be quantitative:

a. With y observed, and x is specified by the experimenter

b. With both x and y observed.





§12.2 <u>The Correlation Coefficient</u>: assesses how tightly the points hug to the best-fit line (discussed in the next sections) and contains the sign corresponds to the slope of that line. It is calculated via:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

For the FFM example:

Subj.	Fat-Free Mass x_i	Energy Expend. y_i	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
1	49.3	1894	-1.083	-0.890	0.963
2	59.3	2050	-0.256	-0.383	0.098
3	68.3	2353	0.488	0.601	0.293
4	48.1	1838	-1.182	-1.071	1.266
5	57.6	1948	-0.397	-0.714	0.283
6	78.1	2528	1.298	1.169	1.517
7	76.1	2568	1.132	1.299	1.470
means	$\bar{x} = 62.40$	$\bar{y} = 2168.4$			Sum = 5.89
SD's	$s_x = 12.10$	$s_y = 308$			

So, by the first formula, r = 5.89/6 = 0.981. Also, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 877.74$, $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 570,124$, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 21,953.7$, so by the second formula, we also get:

$$r = \frac{21,953.7}{\sqrt{877.74 \times 570,124}} = 0.9814$$

As mentioned above, the correlation coefficient assesses how tightly the points hug to the best-fit line and the sign corresponds to the slope of that line. So, in viewing the patterns in the scatter plots in the figure on p.485, it is easy to approximately guess the values of r in each of these cases.

Let ρ be the population (or "true") correlation coefficient. Then, we want to use our sample results to test hypotheses regarding ρ.

H_0: $\rho = 0$ (there is no linear relationship between X and Y)
H_A: $\rho \neq 0$ (there is a linear relationship between X and Y)
(could also make the alternative hypothesis be directional)

Test Statistic: $t_s = r\sqrt{\frac{n-2}{1-r^2}}$

Distribution: t_s has a t-distribution with $(n-2)$ degrees of freedom

Setting a 95% confidence interval (CI) for ρ is important, but involved:

Step 1: Find r and $z_r = \frac{1}{2}ln\left(\frac{1+r}{1-r}\right)$

Step 2: Find $a = z_r - \frac{1.96}{\sqrt{n-3}}$ and $b = z_r + \frac{1.96}{\sqrt{n-3}}$

Step 3: 95% CI for ρ is $\left(1 - \frac{2}{e^{2a}+1}, 1 - \frac{2}{e^{2b}+1}\right)$

For the FFM Example: n = 7 and r = 0.9814, so $t_s = 0.9814\sqrt{\frac{5}{1-0.9814^2}}$ = 11.43, df = 5, from Table 4: p < 0.001. Reject H_0: $\rho = 0$ and accept H_A: $\rho \neq 0$: these data indicate that there is a significant linear relationship between fat-free mass and energy expenditure in men.

For the 95% CI for ρ: (1) $z_r = 2.3342$, (2) a = 1.3542 and b = 3.3142, so (3) the CI is (0.8750,0.9974). We're 95% confident that the true correlation coefficient between fat-free mass and energy expenditure in men is between 0.8750 and 0.9974. R output confirms this:

```
FFM=c(49.3,59.3,68.3,48.1,57.6,78.1,76.1)
EE=c(1894,2050,2353,1838,1948,2528,2568)
cor.test(FFM,EE)
```

```
        Pearson's product-moment correlation
data:  FFM and EE
t = 11.427, df = 5, p-value = 8.988e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:  0.8749612  0.9973570
sample estimates:      cor = 0.9813868
```

Warnings about r: (1) correlation does NOT imply causation, (2) r only assesses linear correlation (not nonlinear, etc.)

§12.3 The Fitted Regression Line:

Our goal here is to find the equation of the best-fitting line in each of these two cases; we'll define what we mean by "best" below. The Fitted Regression Lines (see p.496) for these two cases are:

- $\hat{y} = -0.5921 + 7.6404x$
  (y = drop in body temperature and $x = log_{10}(dose)$)
- $\hat{y} = 607.7 + 25.012x$
  (y = energy expenditure in kcal and x = fat-free mass in kg)

In general, the fitted regression line is written $\hat{y} = b_0 + b_1 x$, and the equations to find these parameter estimates are

- **estimated slope:** $b_1 = r \times \frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}[(x_i - \bar{x})^2]}$
- **estimated intercept:** $b_o = \bar{y} - b_1 \bar{x}$

The second of these equations tells us that the regression line passes through the point of averages, $(\overline{x}, \overline{y})$. Another name for the fitted values (i.e., the $\hat{y}_i$) is the predicted values. The difference between the actual and the predicted values are the residuals. Thus the i^{th} residual (for $i = 1, 2, \ldots, n$) is

$$e_i = y_i - \hat{y}_i$$

An important measure used later is the residual sum of squares:

$$SS(resid) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

A very similar measure is the error sum of squares (SSE) and the minimum value of this SSE is $SS(resid)$; it is in this 'Least-Squares Criterion' sense that the Fitted Regression Line is "best".

We'll see below that there is an SD (and variance) associated with the regression line, and this SD is best estimated by the so-called Residual Standard Deviation, denoted s_e. We have

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{SS(resid)}{n-2}}$$

It is paramount to understand and appreciate the difference between (see the discussion on pp. 500-1):

- **the Residual SD = $s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$**

- the Sample SD = $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$

The difference between these highlights the importance of using SLR when it is appropriate: s_y measures the *variability about the sample mean* (horizontal line at $\bar{y}$), whereas s_e measures the variability about the fitted regression line.

Return to the FFM example (y = energy expenditure in kcal and x = fat-free mass in kg):

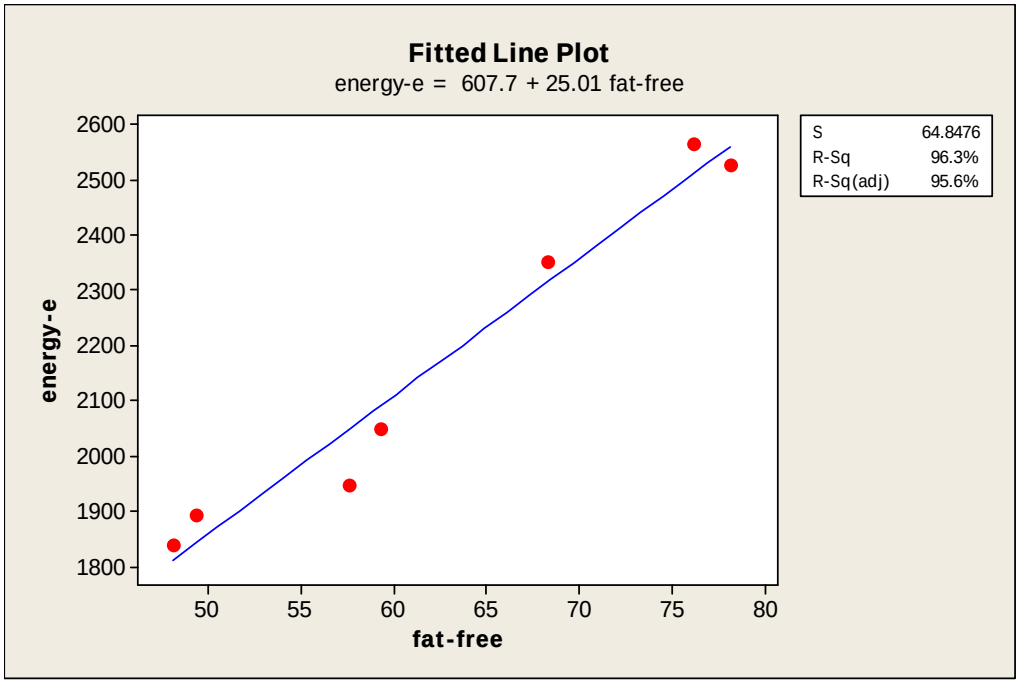
x_i	y_i	$(x_i-\bar{x})^2$	$(x_i-\bar{x})\times(y_i-\bar{y})$	$(y_i-\bar{y})^2$	$\hat{y}_i$	e_i	e_i^2
49.3	1894	171.61	3595.01	75311	1840.8	53.22	2832.8
59.3	2050	9.61	367.13	14025	2090.9	-40.89	1672.2
68.3	2353	34.81	1088.97	34067	2316.0	37.00	1369.2
48.1	1838	204.49	4725.13	109183	1810.8	27.24	741.9
57.6	1948	23.04	1058.06	48589	2048.4	-100.37	10074.7
78.1	2528	246.49	5645.27	129292	2561.1	-33.11	1096.3
76.1	2568	187.69	5474.13	159657	2511.1	56.91	3239.0
436.8	15179	877.74	21,953.7	570,124	----	0	21026.1

Here, n = 7, $\bar{x} = \frac{436.8}{7} = 62.4$, $\bar{y} = \frac{15179}{7} = 2168.43$, the slope estimate and intercept are:

- $b_1 = \frac{21953.7}{877.74} = 25.012$
- $b_o = 2168.43 - 25.012 \times 62.4 = 607.7.$

So, here the Fitted Regression Line is $\hat{y} = 607.7 + 25.012x$.

This line is plotted next with the original data.



This fitted linear regression equation is then used to find the predicted values, the residuals, and the squared residuals in the table (in the 3 right columns); note that here $SS(resid) = 21026.1$, so $s_e = \sqrt{21026.1/5} = 64.85$. To make the above comparison, note too that here $s_y = \sqrt{570124/6} = 308.25$: thus, although the variability measure (SD) of the Y's *about their mean* (2168.43) is 308.25, the variability measure *about the regression line* is only 64.85. The units of both of these variability measures here is Kcal (same as the units of the Y variable).

Finally, the interpretation of the estimated slope is the following: as fat-free mass (x) goes up by one kg, energy expenditure (y) goes up by 25.012 Kcal on average.

The Coefficient of Determination (r^2) is the proportion of the total variation "accounted for" or "explained by" the linear regression of Y on X. As proved in Appendix 12.2, the Coefficient of Determination is related to s_e^2 and s_y^2 by:

$$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Also $0 \leq r^2 \leq 1$; note that r^2 is indeed the square of r here.

For the FFM data (from the Handout p.2), the Coefficient of Determination is $1 - \frac{21026.1}{570124} = 0.9631$, so 96.31% of the variability in y = energy expenditure is explained by the linear regression of it on x = FFM. To approximate r^2, we use $r^2 \approx 1 - \frac{21026.1/5}{570124/6} = 0.955744$.

§12.3 (pp.505-11): The Linear Model

1. **Mean: each $Y = \mu_{Y|X} + \varepsilon$, where the mean is $\mu_{Y|X} = \beta_0 + \beta_1 x$ Thus, $Y = \beta_0 + \beta_1 x + \varepsilon$**
2. **Constant Variance (and SD): the (theoretical) SD, $\sigma_{Y|X}$ (denoted σ_ε), does not depend upon the value of x.**

See the informative graph on p.507. It's important to appreciate that β_0 is the true y-intercept and its estimate (from our small sample) is b_0, β_1 is the true slope and its estimate is b_1, and σ_ε is the true SD about the regression line and it is estimated by s_e – thus,

- **β_0, β_1, and σ_ε are parameters to be estimated**
- **Their respective sample estimates are b_0, b_1, and s_e.**

If we add subscripts above, our Linear Model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and the difference between the y_i's and the mean ($\beta_0 + \beta_1 x_i$) – i.e., the ε_i – are called the errors. The squared errors are essentially the vertical distances between the y_i's and the theoretical line, so our objective function is $SSE = \sum \varepsilon_i^2$. On p.4 of the class handout, we show that minimizing this SSE with respect to the parameters gives the equations for b_0 and b_1 given above. That is why this specific line is the 'best' line, and also called the least-squares line.

In the above,

$\mu_{Y|X}$ = the population mean Y value for a given X value

$\sigma_{Y|X}$ = the population SD of Y values for a given X value

Important assumption: simple linear regression requires both

(1) Normality (of Y given X) and

(2) Constant variances.

We view regression data as coming from the 'random subsampling model' (p.508): thus, we view the pair (x, y) as coming from the conditional distribution of Y given X = x. Using this approach and some mathematical statistics theory, we can show that then the mean of Y given X = x is linear in x, and the variance does not depend upon x (constant variance).

<u>Making Predictions</u>: often, we use regression to make predictions; this is done by simply plugging the new x value into the fitted regression line equation to get the predicted y.

For the FFM example from above:

- **we encounter a new man with FFM x = 55.0kg, and predict his energy expenditure to be $\hat{y}$ = 607.7 + 25.01(55.0) = 1983.4 Kcal**
- **we encounter another new man with FFM x = 45.0kg, should we predict his energy expenditure to be $\hat{y}$ = 607.7 + 25.01(45.0) = 1733.2 Kcal?**

The answer to the above question is that this can be dangerous since 45.0 is outside the range of the data. Predicting inside the range of the data is called interpolation (and this is usually fine), whereas predicting outside the range of the data is called extrapolation (and this should generally be avoided since it can be very risky).

§12.5. Inference About β_1

It was noted above that b_1 (the slope estimate) provides a good estimator for β_1 (the true slope) – but as with every estimator, it has variability associated with it. The estimated standard error of b_1 is

$$SE_{b_1} = \frac{s_e}{s_x\sqrt{n-1}} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

So, to set a 95% CI for β_1 (the true slope), we use:

$$b_1 \pm t_{0.025} \times SE_{b_1}$$

To test the hypothesis $H_0: \beta_1 = 0$, we use the TS $t_s = \frac{b_1}{SE_{b_1}}$ and the t-table. Since we have estimated two parameters in the mean model (the true intercept and slope), it's not surprising that the number of degrees of freedom (df) that we use here for the t-table is $n - 2$.

For the FFM example, $SE_{b_1} = \frac{64.8476}{\sqrt{877.74}} = 2.189$, so:

- **If we want the 95% CI for β_1, we obtain**
$$25.012 \pm 2.571 \times 2.189 \text{ or } (19.38, 30.64)$$
- **If we want to test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$, the test statistic is $t_s = \frac{b_1}{SE_{b_1}} = \frac{25.012}{2.189} = 11.43$; from Table 4 with df = 5, p-value < 2(0.0005) = 0.001. This test statistic is given on p.2 of the Handout.**

Using either of these methods and any reasonable α, we reject the claim that the true slope is zero, and we conclude that $\beta_1 \neq 0$. This means that – *assuming that the SLR model holds with all the related assumptions (normality, constant variance, etc.)* – x = FFM (fat-free mass) is a good linear predictor of y = energy expenditure. See the additional examples and discussion on pp.513-15.

§12.6. Guidelines for Interpreting Regression and Correlation

SLR modelling is based on important assumptions/requirements, notably the linear mean model and constant variance assumptions, and these can be assessed by examining the residual plot(s). Specifically, we examine the residual plots for (see pp.516-9):

- **a curvilinear pattern (violation of linear assumption)**
- **outliers (possible data input error)**
- **leverage and influential points (incorrect use of SLR)**
- **fanning pattern etc. (violation of constant variance assumption)**

When there are problems detected, sometimes a transformation can help – a case in point is the Ethanol example on p.1 of the Handout. Another example where a transformation helps is on p.524. Residual

plots for the illustrations in the Handout are on p.3, and both of these may show some potential anomalies, but clearly more data is required before we can declare a problem. Students should also work through exercise 12.6.6 on p.526.

Miscellaneous topics related to SLR – such as the connection between the two-sample t-test and SLR with a dummy variable, nonlinear regression, ANOCOV, and logistic regression – are discussed in §12.8; we discuss here the first two of these.

For the first of these, algebraic details are given on p.7 of the Handout, and the Toluene example (text pp.531-4) is analyzed in Minitab on p.8 of the Handout. Note that the (absolute value of) t-statistics and p-values for the match in both cases.

On pp.5-6 of the Handout, we look at the chromosome puffs example from p.558. We can fit the one-way ANOVA model:

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad (1)$$

($i = 1, 2, 3$ treatments, and $j = 1, 2, \ldots 40$ flies within each treatment)

Since the treatments are 'control', '30 minutes' and '60 minutes', we could think of the treatment variable as a continuous variable (x), and use linear regression instead (with x = 0, 30, 60). The plot of the data on p.5 (top right) looks quadratic, so we fit the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad (2)$$

Here $i = 1, 2, \ldots 120$ (flies), and this model is linear since it is linear in the parameters. **Interestingly, Equations (1) and (2) – and so the underlying models – are equivalent in this case!** This is verified in the output on p.5: in both cases, the SS (sums of squares) and the F statistics match. Here's the connection:

ANOVA: $H_0: \mu_1 = \mu_2 = \mu_3$
Regression: $H_0: \beta_1 = \beta_2 = 0$

With quadratic regression, sometimes researchers want to estimate the point on the x-axis where the maximum is reached. This occurs at $\gamma = -\frac{\beta_1}{2\beta_2}$ (see the calculus calculations on p.6). If we re-write Equation (2) to include γ instead of β_1 – we say the model has then been "reparameterized" – we obtain the model:

$$y_i = \beta_0 - 2\beta_2\gamma x_i + \beta_2 x_i^2 + \varepsilon_i \qquad (3)$$

This new model is *nonlinear* (in the parameters), so we need specialized software: see the SAS computer program and output on p.6. Here, γ is estimated to be 34.6552 and the 95% CI is (32.87, 36.44). Of course, finding the parameter estimate of γ is simple from p.5 since $-\frac{0.1954}{2(-0.002819)} = 34.66$, but setting the above CI is indeed challenging.

Topics such as nonlinear regression modelling, ANOCOV and logistic regression are discussed further in subsequent classes.