Handout 2

Review of Regression

Simple linear regression

Body Fat data: For a random sample of 18 individuals, we have records of 'measured body fat' (Y, in percent) and 'measured dietary fat' (X, in percent). Here Y=dependent variable and X=independent variable.
Scatterplot of the data indicates that there is a relation between dietary fat and body fat. The goal is to relate these two variables using a simple linear regression method. The model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, ..., n = 18,$$

where β_0 and β_1 are the intercept and slope of the regression line, and $\{\varepsilon_i\}$ are independent and follow a normal distribution with mean zero and variance σ^2. Note that X_i and Y_i are the dietary fat and body fat, respectively, for the i^{th} individual in the sample. Estimates of β_0 and β_1 are given by

$$\hat{\beta}_1 = S_{XY}/S_{XX},\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

where $\bar{X}$ is the average of X-values and $\bar{Y}$ is the average of Y-values, and

$$S_{XX} = \sum (X_i - \bar{X})^2,\ S_{YY} = \sum (Y_i - \bar{Y})^2,\ S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

For the 'Body Fat' data, a summary of the data and the fitted regression line are

$$\begin{aligned}
\bar{X} &= 25.8333, \bar{Y} = 10.3167, \\
S_{XX} &= 1010.50, S_{YY} = 36.2650, S_{XY} = 117.450, \\
\hat{\beta}_0 &= 7.3141, s(\hat{\beta}_0) = 1.006, \hat{\beta}_1 = 0.1162, s(\hat{\beta}_1) = 0.374, \\
\hat{Y} &= 7.3141 + 0.1162X.
\end{aligned}$$

For an individual with dietary fat $X = 30$, our estimate of the body fat is

$$\hat{Y} = 7.3141 + 0.1162(30) = 10.801.$$

A measure of linear relationship between X and Y (in the population) is called the correlation coefficient which is defined as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

In order to know ρ we will need to know the $Cov(X,Y)$, $Var(X)$ and $Var(Y)$, and we clearly we do not know these quantities. However we can estimate each of these using our data set. Here are the estimates

$$\widehat{Cov(X,Y)} = S_{XY}/(n-1), \widehat{Var(X)} = S_{XX}/(n-1), \widehat{Var(Y)} = S_{YY}/(n-1).$$

Plugging in these estimates we can get a formula for obtaining an estimate of ρ

$$\hat{\rho} = \frac{\widehat{Cov(X,Y)}}{\sqrt{\widehat{Var(X)}\ \widehat{Var(Y)}}} = \frac{S_{XY}/(n-1)}{\sqrt{[S_{XX}/(n-1)][S_{YY}/(n-1)]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.$$

For our data, we have $\hat{\rho} = 0.6135$.

Fitted values and residuals:

The fitted Y-values and residuals are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ \hat{\varepsilon}_i = Y_i - \hat{Y}_i,\ i = 1, ..., n.$$

Note that $\{\hat{Y}_i\}$ are the fitted Y-values for the X-values in the sample using the estimated regression line. It is also important to note that the residuals $\hat{\varepsilon}_i$'s are the estimates of ε_i's.

Sums of squares and mean squares.

There are three sums of squares: total sum of squares (SSTO), regression sum of squares (SSR) and the residual (or error) sum of squares (SSE). They are given by

$$\begin{aligned} SSTO &= \sum (Y_i - \bar{Y})^2,\ SSR = \sum (\hat{Y}_i - \bar{Y})^2, \\ SSE &= \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{\varepsilon}_i^2. \end{aligned}$$

Associated with the sums of squares, there are concepts of degrees of freedom (df). They are

$$\begin{aligned} df(SSTO) &= n - 1, \\ df(SSR) &= \text{ number of beta parameters estimated} - 1 = 2 - 1 = 1, \\ df(SSE) &= n - \text{ number of beta parameters estimated} = n - 2. \end{aligned}$$

The sums of squares and their degrees of freedom satisfy the following identities

$$\begin{aligned} SSTO &= SSR + SSE, \\ df(SSTO) &= df(SSR) + df(SSE). \end{aligned}$$

The mean square errors are defined as

$$\begin{aligned} MSR &= SSR/df(SSR) = SSR/1, \\ MSE &= SSE/df(SSE) = SSE/(n-2), \\ MSTO &= SSTO/df(SSTO) = SSTO/(n-1). \end{aligned}$$

The quantity $MSTO$ is rarely used. **An estimate of σ^2 (the common variance of ε's) is given by MSE** and this is a consequence of the following fact

Fact: $E(MSE) = \sigma^2$.

Another important measure of association:

Coefficient of determination: proportion of variability in Y that can be explained by its regression on X is given by

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

We should note that $0 \leq R^2 \leq 1$ and $R^2 = \hat{\rho}^2$.

For the "Body Fat" data, $R^2 = 0.376$. Thus we can say that about 37.6% of the variability in body fat (Y) can be be explained by its regression on dietary fat (X).

There is another measure that is a little better than R^2. It is called the adjusted R^2 and it is given by

$$R^2_{adj} = 1 - MSE/MSTO = 0.337.$$

Adjusted R^2_{adj} has the same interpretation as R^2. It is always true that $R^2_{adj} \leq R^2$. Please note that, as a measure of linear association, R^2_{adj} is generally preferrred over R^2.

Multiple regression

Consider the Electric Bill data where we have $n = 34$ households. For the i^{th} household, $i = 1, \ldots, n = 34$, we have

Y_i = monthly electric bill (in dollars), X_{i1} =monthly income (in dollars), X_{i2} = number of persons, X_{i3} = living area (in square feet).

The goal is to relate Y to X_1, X_2 and X_3 by a linear regression method. The model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \; i = 1, ..., n = 34,$$

where $\{\varepsilon_i\}$ are independent, normally distributed with zero mean and common variance σ^2. Unlike in the simple linear regression case (the "Body Fat" case), we cannot have simple expressions for the estimates of the beta parameters. Matrix-vector notations need to be used. The regression model above can be re-expressed as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix},$$

$$or, \;\; \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Estimates of the beta parameters, fitted Y-values and residuals are now given in vector-matrix notations:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \; \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \; \hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}},$$

where for any matrix or vector "$'$" denotes its transpose. The concepts of $SSTO$, SSR and SSE remain the same as in the case of simple linear regression. Thus we have

$$SSTO = \sum (Y_i - \bar{Y})^2, \; SSR = \sum (\hat{Y}_i - \bar{Y})^2,$$
$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{\varepsilon}_i^2.$$

The degrees of freedom (df) are

$$\begin{aligned} df(SSTO) &= n-1, \\ df(SSR) &= \text{ number of beta parameters estimated} - 1 = 4 - 1 = 3, \\ df(SSE) &= n - \text{ number of beta parameters estimated} = n - 4 = 30. \end{aligned}$$

As in the case with simple linear regression, the sum of squares and their degreees of freedom satisfy the identities

$$SSTO = SSR + SSE,$$
$$df(SSTO) = df(SSR) + df(SSE).$$

If we denote the number of beta parametrs estimated by p (here $p = 4$), then

$$\begin{aligned} MSR &= SSR/df(SSR) = SSR/(p-1), \\ MSE &= SSE/df(SSE) = SSE/(n-p), \\ MSTO &= SSTO/df(SSTO) = SSTO/(n-1).. \end{aligned}$$

Estimate of σ^2 is given by MSE..

As usual, MSE estimates σ^2.
Estimate of the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$s^2(\hat{\boldsymbol{\beta}}) = MSE \; (\mathbf{X}'\mathbf{X})^{-1}.$$

Note that $s^2(\hat{\boldsymbol{\beta}})$ is a matrix whose diagonal elements are $s^2(\hat{\beta}_0)$, $s^2(\hat{\beta}_1)$,..... These can can be used for constructing confidence intervals for β_0, β_1 etc. They can also be used to decide if a particular variable can be dropped from the regression model. For the electric bill data we have $\hat{\beta}_1 = 0.0751$ and $s(\hat{\beta}_1) = 0.1361$, So a 95% confidence interval for β_1 is given by

$$\begin{aligned} &\hat{\beta}_1 \pm t_{1-\alpha/2;n-p} s(\hat{\beta}_1), \; i.e., \hat{\beta}_1 \pm t_{0.975;30} s(\hat{\beta}_1), \; i.e, \; 0.0751 \pm (2.042)(0.1361), \\ &i.e., 0.0751 \pm 0.2779, \; i.e., (-0.203, 0.353). \end{aligned}$$

If we want to check if a particular variable, say X_1, can be dropped from the regression model, we need to carry out the test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at level of significance $\alpha = 0.05$. The t-statistic is $t^* = \hat{\beta}_1/s(\hat{\beta}_1) = 0.55$. Since $|t^*|$ is larger that the cut-off point (also called critical value) $t_{1-\alpha/2;n-p} = t_{0.975;30} = 2.042$, we cannot reject H_0. So the conclusion is: we may drop variable X_1 from the full model. Alternatively, we may simply look at the p-value given in the output in the next page. This p-value is 0.585

and it is larger than $\alpha = 0.05$, and hence we conclude that variable X_1 may be dropped. Another way of carrying out the test is to note that zero is inside the 95% confidence interval constructed above.

Remark: Note that the decision to retain or drop a variable is equivalent to testing if the corresponding beta coefficient is zero or not. It is important to keep in mind that, when building a model using a backward elimination method (or more generally backward stepwise procedure), variables are dropped one at a time using a preselected α till no deletion is possible. In many cases, one may wish to build a model by employing a foward selection procedure (or more generally a forward stepwise procedure) whereby one starts with no independent variable and then adds one variable at a time till no inclusion is possible. In any of these methods (forward or backward), it is not advisable to add or drop more than one variable at a time.

Sums of squares and measures of association.

The definitions of the sums of squares, degrees of freedom, mean squares, R^2 and adjusted R^2 remain the same as in the case of simple linear regression.

For the "Electric Bill" data, $R^2 = \frac{SSR}{SSTO} = 0.851$. Thus we can say that about 85.1% of the variability in electricity bill can be expalined by its regression on income (X_1), number of persons (X_2) and living area (X_3).

There is one more measure of association between Y and the $X's$ which is especially useful in the multiple regression case, and it is the the concept of "multiple correlation". Multiple correlation R is the positive square root of R^2. For the Electric Bill data, the multiple correlation is $R = \sqrt{R^2} = \sqrt{0.851} = 0.923$.

Here is another important fact

$$R = Corr(Y, \hat{Y}).$$

So if the multiple correlation R is close to 1, it means that the values of $\hat{Y}$'s are close to Y's, i.e., the regression function is very effective in guessing the Y-values.

Body fat data: X =dietary fat (in percent), Y = body fat (in percent).

Y	9.8	11.7	8.0	9.7	10.9	7.8	9.7	11.6	8.6	11.2	12.3	10.2	12.0	11.6	10.4
X	22	22	14	21	32	26	30	21	17	35	35	24	24	36	20

Y	10.8	11.5	7.9
X	37	35	14

The regression equation is: $\hat{Y} = 7.314 + 0.1162X$,

$Predictor$	$Coef$	SE	T	P
$Constant$	7.314	1.006	7.27	0.000
$DietaryFat$	0.11623	0.0374	3.11	0.007

$S = \sqrt{MSE} = 1.18885, R^2 = 0.376, R^2_{adj} = 0.337.$

Analysis of Variance

$Source$	df	SS	MS	F	P
$Regression$	1	13.651	13.651	9.66	0.007
$Error$	16	1622.614	1.413		
$Total$	17	36.265			

Electric Bill data.

Y	228	156	648	528	552	636	444	144	744	1104	204	420	876
X_1	3220	2750	3620	3940	4510	3990	2430	3070	3750	4790	2490	3600	5370
X_2	2	1	1	1	3	4	1	1	2	5	1	3	1
X_3	11602	1080	1720	1840	2240	2190	830	1150	1570	2660	900	1680	2550

Y	840	876	276	1236	372	276	540	1044	552	756	636	708	960
X_1	3180	5910	3020	5920	3520	3720	4840	4700	3270	4420	4480	3820	5740
X_2	7	2	2	3	2	1	1	6	2	2	2	4	2
X_3	1770	2960	1190	3130	1560	1510	2190	2620	1350	1990	2070	1850	2700

Y	1080	480	96	1272	1056	156	396	768
X_1	5600	3950	2290	5580	5820	3160	2880	3780
X_2	3	2	3	5	2	2	4	3
X_3	3030	1700	890	3270	2660	1330	1280	1950

The regression equation is
$\hat{Y} = 358.4 + 0.0571X_1 + 55.09X_2 + 0.2811X_3$.

Predictor	*Coef*	*SE*	*T*	*P*
Constant	-358.4	198.7	-1.80	0.081
Income	0.0751	0.1361	0.55	0.585
Person	55.09	29.05	1.90	0.068
Area	0.2811	0.2261	1.24	0.223

$S = \sqrt{MSE} = 135.421,\ R^2 = 0.851, R^2_{adj} = 0.837.$

Analysis of Variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	3	3151504	1050502	57.28	0.000
Error	30	550163	18339		
Total	33	3701667			

Figure 1: Body Fat Data

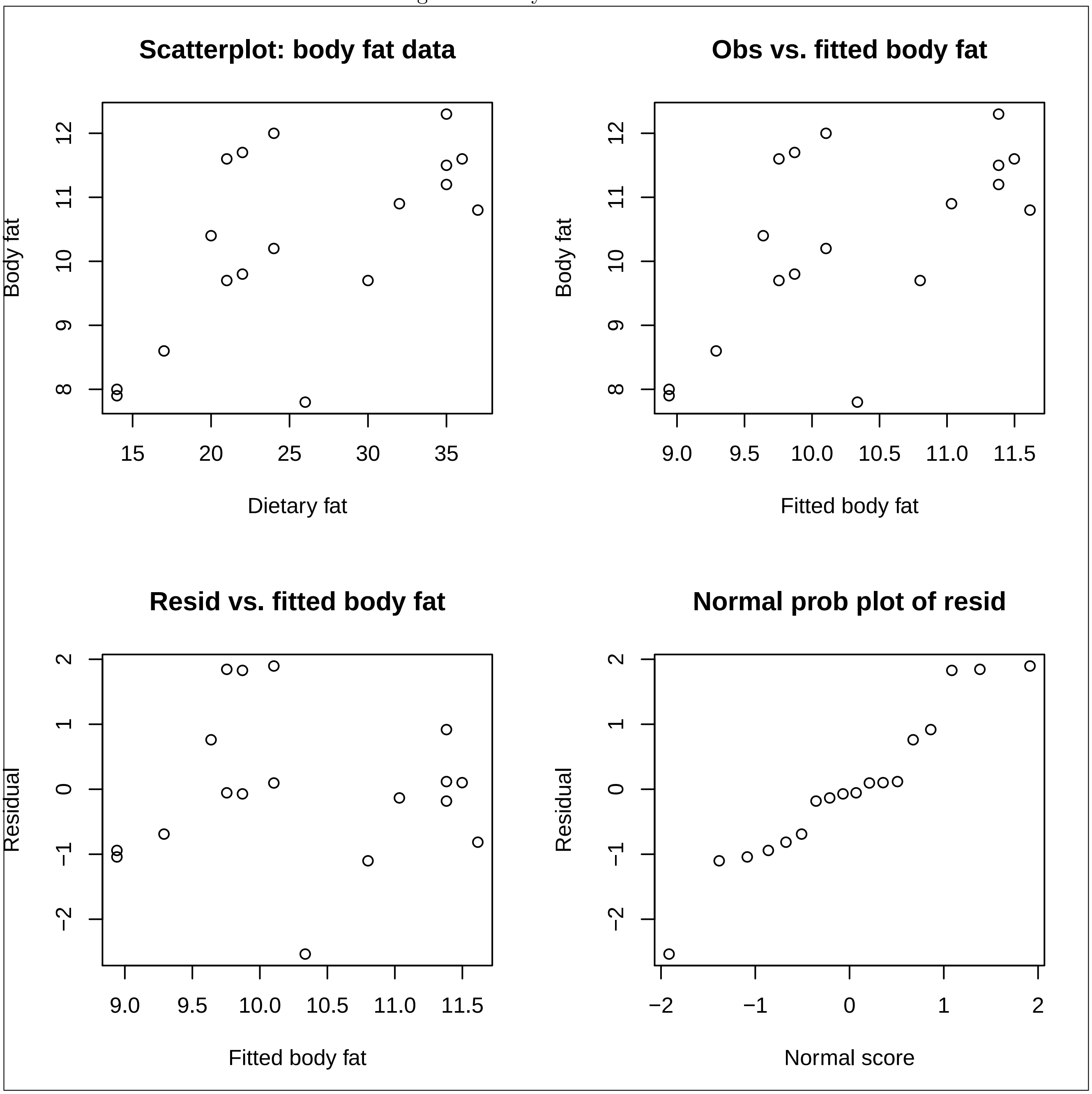

Figure 2: Electric Bill Data