Review of Regression

(1) Body fat data: Simple linear regression

For a random sample of $n = 18$ individuals, records of 'measured body fat' (Y, in percent) and 'measured dietary fat intake' (X, in percent) were obtained. Here Y is the dependent variable and X is the independent variable. The scatter plot of the data in the top left panel of Figure 1 indicates that there is a relation between X and Y, so the goal is to relate the variables to each other through a simple linear regression.

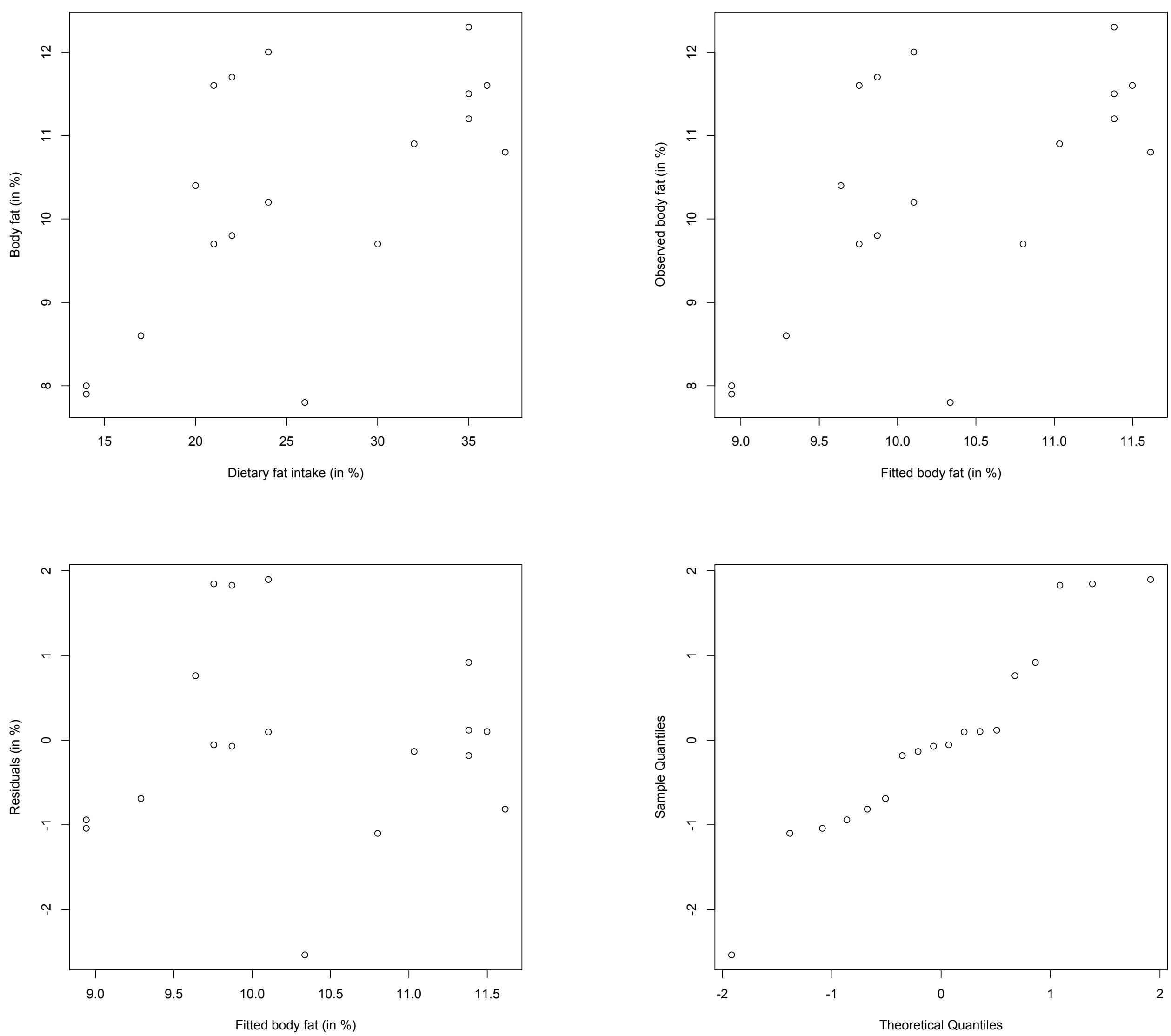

Figure 1: Scatter plot (top left), plot of observed versus fitted values (top right), plot of residuals versus fitted values (bottom left) and residual qq-plot (bottom right) for the body fat data.

If X_j and Y_j are the dietary fat intake and body fat in percent for the jth individual in the sample, the model is

$$Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j, \qquad j = 1, \dots, n = 18,$$

where β_0 and β_1 are the intercept and slope of the regression line and $\varepsilon_1, \dots, \varepsilon_{18}$ are independent, identically distributed (normal) with mean zero and variance σ^2. Using the simple linear regression framework, the parameters β_0 and β_1 can be estimated by

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X},$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the X and Y values,

$$S_{XY} = \sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y}) \qquad \text{and} \qquad S_{XX} = \sum_{j=1}^{n}(X_j - \bar{X})^2.$$

If one computes these estimates for the body fat data with the function `lm` in R, then the estimates $\bar{X} = 25.83$ and $\bar{Y} = 10.32$ are obtained, further $\hat{\beta}_1 = 0.12$ and $\hat{\beta}_0 = 7.31$. Assuming that the data are stored in `bf`, the following shows the R output:

```
Call:
lm(formula = bf[, 2] ~ bf[, 1])

Residuals:
      Min         1Q    Median        3Q       Max
 -2.53604  -0.78342  -0.06301  0.60048  1.89642

Coefficients:
              Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept)    7.3141       1.0059    7.271  1.87e-06 ***
 bf[, 1]        0.1162       0.0374    3.108   0.00677 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  1.189 on 16 degrees of freedom
Multiple R-squared:  0.3764, Adjusted R-squared:  0.3375
F-statistic:  9.659 on 1 and 16 DF, p-value:  0.006768
```

With these estimates at hand, the *fitted regression line* becomes $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 7.31 + 0.12X$. The scatter plot of the observed values Y_j versus the *fitted values* $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$ is shown in the top right panel of Figure 1. The *residuals* are then defined as the difference between observed and fitted values, that is, $\hat{\varepsilon}_j = Y_j - \hat{Y}_j$. Note that $\hat{\varepsilon}_j$ is an estimator of the innovations ε_j. The plot of residuals versus fitted values is given in the bottom left and the qq-plot of the residuals in the bottom right panel of Figure 1.

To measure the strength of the linear relationship between X and Y on the population level, one may use the *correlation coefficient*

$$\rho = \frac{\gamma_{XY}}{\sigma_X \sigma_Y},$$

where $\gamma_{XY} = \text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$, $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Y^2 = \text{Var}(Y)$. Since ρ is a population quantity and therefore unknown, it needs to be estimated from the data. This can be done using the *sample correlation coefficient*

$$\hat{\rho} = \frac{\hat{\gamma}_{XY}}{s_X s_Y} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}},$$

where $\hat{\gamma}_{XY} = (n-1)^{-1} S_{XY}$, $s_X^2 = (n-1)^{-1} S_{XX}$ and $s_Y^2 = (n-1)^{-1} S_{YY}$ with $S_{YY} = \sum_{j=1}^n (Y_j - \bar{Y}_j)^2$. For the body fat data, $\hat{\rho} = 0.61$.

Next, note the decomposition of variation given by

$$\text{SST} = \sum_{j=1}^n (Y_j - \bar{Y})^2 = \sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2 + \sum_{j=1}^n (Y_j - \bar{Y}_j)^2 = \text{SSR} + \text{SSE},$$

where SST, SSR and SSE are called the *total sum of squares* the *regression sum of squares* and the *residual* or *error sum of squares*, respectively. Associated with the sums of squares are concepts of *degrees of freedom (df)*. They are

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

For the body fat data, $\text{SST} = 17$, $\text{SSR} = 2 - 1 = 1$ and $\text{SSE} = 18 - 2 = 16$. Moreover, with sums of squares and degrees of freedom one defines the *mean squared errors*

$$\text{MST} = \frac{\text{SST}}{\mathit{df}(\text{SST})}, \qquad \text{MSR} = \frac{\text{SSR}}{\mathit{df}(\text{SSR})}, \qquad \text{and} \qquad \text{MSE} = \frac{\text{SSE}}{\mathit{df}(\text{SSE})}.$$

The quantity MST is not used very often. Since $E[\text{MSE}] = \sigma^2$, MSE is an unbiased estimate of σ^2, the innovation variance. After introducing these quantities, another important measure of association can be defined, namely the (adjusted) proportion of variability in Y that can be explained by its regression on X. First define the *coefficient of determination* as

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

Note that $R^2 = \hat{\rho}^2$. Typically one does not use R^2 but prefers its adjusted value

$$R^2_{\text{adj}} = 1 - \frac{\text{MSE}}{\text{MST}},$$

which has the same interpretation as R^2. It is always true that $R^2_{\text{adj}} \le R^2$. From the R output, $R^2_{\text{adj}} = 0.34$, more than 10% smaller than $R^2 = \hat{\rho}^2 = 0.38$.

(2) Electricity bill data: Multiple linear regression

For a random sample of $n = 34$ households, the monthly electricity bill (Y, in \$), the monthly income (X_1, in \$), the number of people in the household (X_2) and the size of the living area (X_3, in square feet) were obtained. The goal is to relate Y to X_1, X_2 and X_3 via a linear regression method. The model is

$$Y_j = \beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \beta_3 X_{j3} + \varepsilon_j, \qquad j = 1, \ldots, n = 34,$$

where $\beta_0, \ldots, \beta_3$ are the regression parameters and $\varepsilon_1, \ldots, \varepsilon_n$ are independent, identically distributed (normal) random variables with zero mean and variance σ^2. Unlike in simple linear regression cases, there are no easy expressions for the estimates of the parameters. Note that the regression model can be re-expressed as

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or, more compactly, as

$$Y = X\beta + \varepsilon.$$

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

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

The estimate of σ^2 is given by

$$\text{MSE} = \frac{1}{df(\text{SSE})} \sum_{j=1}^{n} (Y_j - \bar{Y})^2 = \frac{1}{n-p} \sum_{j=1}^{n} (Y_j - \bar{Y})^2,$$

where p is the number of beta parameters to be estimated, so that $df(\text{SSE}) = n - p$. Assuming the data has been stored in `eb`, the following shows the R output:

```
Call:
lm(formula = eb[, 1] ~ eb[, 2] + eb[, 3] + eb[, 4])

Residuals:
     Min      1Q  Median      3Q     Max
 -223.60  -91.82  -15.18  79.22  327.21

Coefficients:
               Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept)  -554.9446     98.57474   -5.630  3.94e-06 ***
 eb[, 2]        0.24758      0.02351   10.531  1.35e-11 ***
 eb[, 3]       82.10514     16.28270    5.042  2.07e-05 ***
 eb[, 4]       -0.01444      0.01407   -1.027     0.313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  140.7 on 30 degrees of freedom
Multiple R-squared:  0.8396, Adjusted R-squared:  0.8236
F-statistic:  52.36 on 3 and 30 DF, p-value:  4.92e-12
```

Figure 2 contains some supporting plots. [It can be seen that the living area of household 1 is an outlier. My best guess is that the square footage has been incorrectly recorded. More likely it should have been 1602 and not 11602. If one drops household 1 and plots living area against income, one can see a strong linear relationship. In fact, excluding household 1, the regression of income on living area produces $R^2_{\text{adj}} = 0.92$. This helps explain why living area is not significant in the above output; it's contribution may have been absorbed by X_1.]

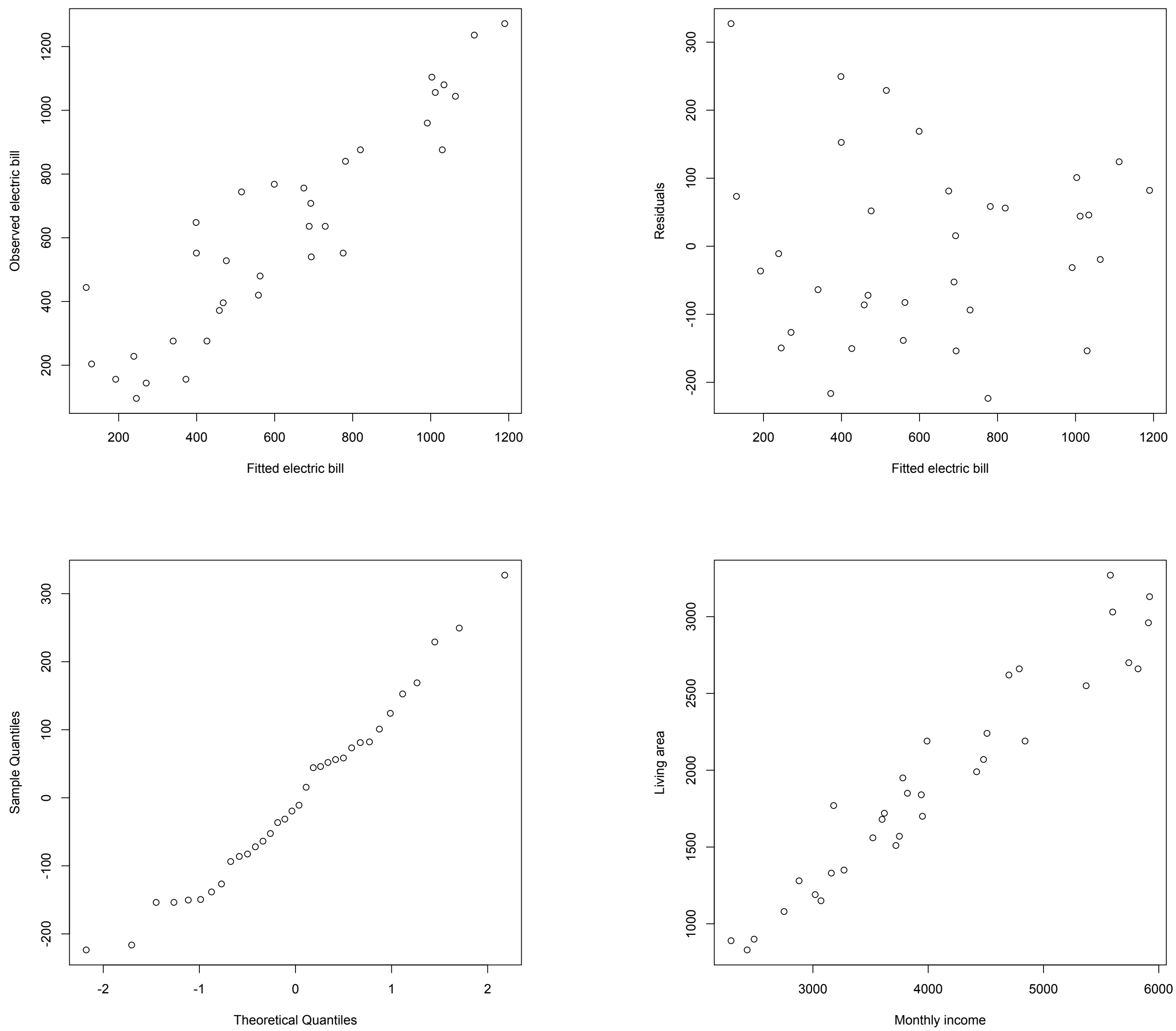

Figure 2: Plot of observed versus fitted values (top left), plot of residuals versus fitted values (top right), residual qq-plot (bottom left) and scatter plot of living area and monthly income after one outlier has been removed (bottom right) for the electricity bill data.

Recall from STA 108 that the variance-covariance matrix of the estimator $\hat{\beta}$ is given by

$$s^2(\hat{\beta}) = \mathrm{MSE}(X'X)^{-1}$$

and note that $s^2(\hat{\beta})$ is a $(p+1) \times (p+1)$ matrix whose diagonal elements are $s^2(\hat{\beta}_j)$. These can for example be used for constructing confidence intervals for β_j, but also to decide if a particular variable can be dropped from the regression model. For the electricity bill data, the R output shows that the parameter estimate for the living area variable is given by $\hat{\beta}_3 = -0.01444$ and that $s(\hat{\beta}_3) = 0.01407$. A 95% confidence interval for β_1 is therefore given by

$$\hat{\beta}_3 \pm t_{0.975,30} s(\hat{\beta}_3) = -0.01444 \pm (2.042)(0.01407) = (-0.04317, 0.01429).$$

Since 0 is contained in the confidence interval, the variable X_3 is not significant. Formally, one would carry out a test $H_0\colon \beta_3 = 0$ against $H_A\colon \beta_3 \neq 0$ at significance level $\alpha = 0.05$. The t-statistics for this case is

$$t^* = \frac{\hat{\beta}_3 - 0}{s(\hat{\beta}_3)} = -1.027.$$

[Check the value in the R output!] But this is larger than the critical value $t_{0.975,30} = 2.042$, so H_0 cannot be rejected and X_3 is dropped from the model. Alternatively, one looks at the p-value which is equal to 0.313. [Check the value in the R output!] This is larger than $\alpha = 0.05$ and therefore the same conclusions are reached. [Note that the decision whether a variable is retained or dropped is equivalent to testing whether the corresponding β coefficient is equal to zero.]

The definition of sums of squares, degrees of freedom, mean squares, R^2 and R^2_{adj} remain the same as in the case of a simple linear regression. However, there is an additional measure of association between Y and the X_j for multiple regression. It is based on the concept of *multiple correlation*. Multiple correlation is the positive square root of R^2. For the electrical bill data this value is 0.8396.

Finally note the important fact

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

that is, if the multiple correlation R is close to 1, then the fitted values $\hat{Y}_j$ are close to the observed values Y_j and the regression function is very effective in guessing the response variable values.