2. Regression Analysis

Isabel M. Rodrigues

2.1. Multiple linear regression.

- Multiple linear regression represents a generalization, to more than a single explanatory variable, of the simple linear regression model.
- The method is used to investigate the relationship between a continuous dependent variable, y, with a p number of continuous explanatory variables.
- Note in particular that the explanatory variables are, strictly, not regarded as random variables at all so that multiple regression is essentially a univariate technique with the only random variable involved being the response, y.

2.1. Multiple linear regression.

Multiple Linear Regression Model:

$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$ for $i = 1, \ldots, n$.

- n is total number of observations.
- y is the response variable.
- p is the number of explanatory variables (covariates, including x_0); number of betas's.
- x_{i0} is associated with the interceptor, β_0, and is usually 1.
- The regression coefficients $\beta_1, \beta_2, \ldots, \beta_{p-1}$ give the amount of change in the response variable associated with a unit change in the corresponding explanatory variable, conditional on the other explanatory variables in the model remaining unchanged, i.e., $\frac{\partial E[y]}{\partial x_k} = \beta_k$.
- The errors terms are assumed $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$,

2.1. Multiple linear regression.

1. This implies that, for given values of the explanatory variables, the response variable is normally distributed with a mean that is a linear function of the explanatory variables and a variance that is not dependent on these variables. Consequently an equivalent way of writing the multiple regression model is as $Y \sim N(\mu, \sigma^2)$ where $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$.

2. The "linear" in multiple linear regression refers to the parameters rather than the explanatory variables, so the model remains linear if, for example, a quadratic term for one of these variables is included. (An example of a non-linear model is $y = \beta_0 e^{\beta_1 x_{i1}} + \beta_2 e^{\beta_4 x_{i2}} + \epsilon_i)$.

3. The aim of multiple regression is to achieve a set of values for the regression coefficients that makes the values of the response variable predicted from the model as close as possible to the observed values.

2.2. Least squares estimation of the parameters.

- The least-squares procedure is used to estimate the parameters in the multiple regression model ($\boldsymbol{\beta}$ and σ^2).
- The resulting estimators are most conveniently written with the help of some matrices and vectors.
- Define the vector of responses, $\mathbf{y}$, and matrix of explanatory variables, X-called the design matrix, as:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix}.$$

- We will consider that $\mathbf{x}_0^{\mathrm{T}} = \mathbf{1}_n$.

2.2. Least squares estimation of the parameters.

- Defining the vector of regression coefficients and the vector of errors as:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

- In matrix form, the regression model partitions the response into a non-random $\mathbf{X}\,\boldsymbol{\beta}$ and a random component $\boldsymbol{\varepsilon}$ as follows:

$$\boxed{\mathbf{y}=\mathbf{X}\,\boldsymbol{\beta}+\boldsymbol{\varepsilon}}, \quad \mathbf{X}\,\boldsymbol{\beta} \text{ (is the matrix-vector product)}$$

  where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\operatorname{COV}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I_n}$.

- As a consequence, we have $E(\mathbf{y}) = \mathbf{X}\,\boldsymbol{\beta}$ and $\operatorname{COV}(\mathbf{y}) = \sigma^2\mathbf{I}$.

Specific individual observation

- For a specific observation i, define the row vector of observed explanatory variables by:

$$\mathbf{x}_i^{\mathrm{T}} = (1, x_{i,1}, \ldots, x_{i,p-1}).$$

- Thus, we see that the regression model for this specific observation can be written as:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i.$$

2.2. Least squares estimation of the parameters.

Least squares estimates

- In order to estimate $\boldsymbol{\beta}$, we take a least squares approach that is analogous to what we did in the simple linear regression case. That is, we want to find $\hat{\boldsymbol{\beta}}$ that minimize:

$$
\begin{aligned}
SSE = & \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 && (1)\\
= & \mathbf{e}^{\mathrm{T}}\mathbf{e} = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) && (2)\\
= & \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}. && (3)
\end{aligned}
$$

- Note that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a symmetric matrix.

Recall-Vector differentiation:

Let $f(\mathbf{x})$ be a scalar function of the vector $\mathbf{x}$.

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \cdots \\ \frac{\partial f(\mathbf{x})}{\partial x_p} \end{pmatrix},$$

$$\frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_1} \\ \cdots \\ \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_p} \end{pmatrix} = \mathbf{a},$$

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x},$$

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}, \text{ if } \mathbf{A} \text{ is symmetric},$$

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\mathbf{x}, \text{ if } \mathbf{A} \text{ is not symmetric}.$$

2.2. Least squares estimation of the parameters.

- Differentiating (3) and then setting to zero, we have the normal equations:
- $\dfrac{\partial SSE}{\partial \hat{\boldsymbol{\beta}}} = 0 \Leftrightarrow -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \Leftrightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$
- Provided $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, we have: $\boxed{\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}$

2.2. Least squares estimation of the parameters.

- The vector of fitted values:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \left[\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]\mathbf{y} = \mathbf{H}\mathbf{y},$$

  is the orthogonal projection of $\mathbf{y}$ onto the estimation space and the projection matrix $\mathbf{H}$ (referred as the "hat matrix") satisfies

$$\mathbf{H}\mathbf{H} = \mathbf{H} \text{ (idempotent)} \quad \text{and} \quad \mathbf{H}^{\mathrm{T}} = \mathbf{H} \text{ (symmetric)}$$

- The residual vector is defined by $\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H})\mathbf{y}$, and it corresponds to the orthogonal projection of $\mathbf{y}$ onto the subspace orthogonal to the estimation space (i.e., $\mathbf{H}\mathbf{e} = \mathbf{0}$).

2.2. Least squares estimation of the parameters.

- Residual sum squares:

$$SSE = \mathbf{e}^{\mathrm{T}}\mathbf{e} = \mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{H}\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}.$$

Variance estimator

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p},$$

where MSE is the mean square for the residuals.

$\hat{\beta}$ Estimator

- Unbiased Estimator:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

- Covariance Matrix:

$$\mathrm{COV}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2\mathbf{C}$$

- Variance estimator of $\hat{\beta}_k$:

$$\hat{\mathrm{VAR}}(\hat{\beta}_k) = \hat{\sigma}^2 c_{k+1,k+1}, \quad k = 0, \ldots, p-1,$$

where $c_{k+1,k+1}$ is the $(k+1)$-*th* diagonal entry of $\mathbf{C}$.

2.4. Tests and confidence intervals for the parameters.

- Supposition: $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}) \Rightarrow \mathbf{y} \sim N_n(\mathbf{X}\,\boldsymbol{\beta}, \sigma^2\mathbf{I})$.
- $\hat{\boldsymbol{\beta}} \sim N_p(\,\boldsymbol{\beta}, \sigma^2\mathbf{C})$.

Confidence Intervals and tests for individual Slope Coefficients

- Pivotal quantity: $\boxed{\dfrac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 c_{k+1,k+1}} \sim t_{(n-p)},}}\ k = 0, \ldots, p-1.$
- $C.I._{\cdot(1-\alpha)\times 100\%}(\beta_k) = \left(\hat{\beta}_k \pm t_{1-\frac{\alpha}{2}(n-p)}\sqrt{\hat{\sigma}^2 c_{k+1,k+1}}\right)$
- Test on individual coefficients (variable x_k is significant?)

$H_0 : \beta_k = 0 \quad \text{vs} \quad H_1 : \beta_k \neq 0$ (or $H_1 : \beta_k > 0$ or $H_1 : \beta_k < 0$)

- We should reject the null hypothesis

2.4. Tests and confidence intervals for the parameters.

Confidence Intervals for the mean response:

$$E[Y|\mathbf{x}_0] = \mu_{Y|\mathbf{x}_0} = \beta_0 + \beta_1 x_{0,1}, \ldots, \beta_{p-1} x_{0,p-1}$$

- Estimator: $\hat{E}[Y|\mathbf{x}_0] = \hat{\mu}_{Y|\mathbf{x}_0}$
- Unbiased Estimator: $E\left[\hat{\mu}_{Y|\mathbf{x}_0}\right] = \mu_{Y|\mathbf{x}_0}$
- $\mathrm{VAR}(\hat{\mu}_{Y|\mathbf{x}_0}) = \mathbf{x}_0^{\mathrm{T}} \mathrm{COV}(\hat{\boldsymbol{\beta}}) \mathbf{x}_0 = \sigma^2 \mathbf{x}_0^{\mathrm{T}} \mathbf{C} \mathbf{x}_0$
- Pivotal quantity: $\boxed{\dfrac{\hat{\mu}_{Y|\mathbf{x}_0} - \mu_{Y|\mathbf{x}_0}}{\sqrt{\hat{\sigma}^2 \mathbf{x}_0^{\mathrm{T}} \mathbf{C} \mathbf{x}_0}} \sim t_{(n-p)}}$
- $C.I._{(1-\alpha)\times 100\%}(\mu_{Y|\mathbf{x}_0}) = \hat{\mu}_{Y|\mathbf{x}_0} \pm t_{1-\frac{\alpha}{2}(n-p)} \sqrt{\hat{\sigma}^2 \mathbf{x}_0^{\mathrm{T}} \mathbf{C} \mathbf{x}_0}$

Overall Test-Test for Significance of Regression

- Does the entire set of variables $(x_1, x_2, \ldots, x_{p-1})$ explain significantly the variations of Y?
- We should reject the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

when at least one explanatory variable is correlated with Y, that is when:

$$H_1 : \exists^1 \beta_k \neq 0, \ k = 1, \ldots, p-1$$

2.4. Tests and confidence intervals for the parameters.

- Very similar to what was done in the simple linear regression, we can decompose the total variability into variability due to the regression and variability due to the residuals.
- $\boxed{SST = SSE + SSR}$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{y}^{\mathrm{T}}\mathbf{y} - n\bar{y}^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - n\bar{y}^2$$

$$\boxed{MST = \frac{SST}{n-1}} \qquad \boxed{MSR = \frac{SSR}{p-1}}$$

2.4. Tests and confidence intervals for the parameters.

- This decomposition allows us to account for these variabilities into an ANOVA table.

Source	SS	$d.f.$	MS	$F-$ratio
Regression	SSR	$(p-1)$	MSR	$F = \frac{MSR}{MSE}$
Residuals	SSE	$(n-p)$	MSE	
Total	SST	$(n-1)$	MST	

- Under H_0, the F_0 statistics is "small" and distributed as a Fisher variable with $(p-1, n-p)$ degrees of freedom. Then, we will reject the null hypothesis (with a significance level of α) when $F_0 > F^{-1}_{F_{1-\alpha(p-1,n-p)}}$.
- Obs: $R^2 = \frac{SSR}{SST} \Rightarrow F_0 = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$, if F_0 is big, then the regression is "working" $\Rightarrow R^2 \to 1$.

2.4. Tests and confidence intervals for the parameters.

- Rejection of the null does not mean all of the explanatory variables are useful, just that at least one of them is. If the null is rejected, we can then use the individual t-tests on each coefficient to determine which of the explanatory variables is statistically helpful in explaining the variation in y. The F-test (overall test) is a test on the entire model.

2.4. Tests and confidence intervals for the parameters.

Test-F Partial-The question that will be addressed is:
"Is the full model significantly better than the reduced model at explaining variation in y?"

- Some β_k are zero?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0 \textbf{ vs } H_1 : \exists^1 \beta_k \neq 0,\ k = 1, \ldots, r < (p-1)$$

- Considering

$$\boldsymbol{\beta}_1^{\mathrm{T}} = (\beta_1, \beta_2, \ldots, \beta_r) \quad \text{and} \quad \boldsymbol{\beta}_2^{\mathrm{T}} = (\beta_{r+1}, \beta_{r+2}, \ldots, \beta_{p-1})$$

we have that:

- $\boxed{SSR(\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2) = SSR(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - SSR(\boldsymbol{\beta}_2)}$

2.4. Tests and confidence intervals for the parameters.

- Under H_0, the $F_0 = \frac{SSR(\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2)}{rMSE}$ statistics is "small" and distributed as a Fisher variable with $(r, n-p)$ degrees of freedom. Then, we will reject the null hypothesis (with a significance level of α) when $F_0 > F^{-1}_{F_{1-\alpha(r,n-p)}}$.

- If the null is accepted we should use the reduced model.

- If the null is rejected, the full model is significantly better than the reduce model.

Prediction interval for a new or future observation of Y: Y_0

- $Y_0 = Y|_{\mathbf{x}_0} = \mathbf{x}_0^{\mathrm{T}}\,\boldsymbol{\beta} + \epsilon = E(Y_0) + \epsilon$ and $Y_0 \sim N(\mathbf{x}_0^{\mathrm{T}}\,\boldsymbol{\beta}, \sigma^2)$, with $E[Y_0]$ estimated by $\hat{E}[Y_0] = \hat{Y}_0 = \mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.
- $\hat{Y}_0 - Y_0 = \mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{x}_0^{\mathrm{T}}\,\boldsymbol{\beta} + \epsilon$ with:
  $E[\hat{Y}_0 - Y_0] = 0$ and $\mathrm{VAR}(\hat{Y}_0 - Y_0) = \mathrm{VAR}(\mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + \sigma^2$
  $= \sigma^2\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0 + \sigma^2 = \sigma^2(1 + \mathbf{x}_0^{\mathrm{T}}\mathbf{C}\mathbf{x}_0)$
- $\hat{Y}_0 - Y_0 \sim N(0, \sigma^2(1 + \mathbf{x}_0^{\mathrm{T}}\mathbf{C}\mathbf{x}_0))$
- Pivotal quantity: $\boxed{\dfrac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0^{\mathrm{T}}\mathbf{C}\mathbf{x}_0)}} \sim t_{(n-p)}}$

2.5.Prediction. Model adequacy checking.

Forecasting: Then a $(1-\alpha)100\%$ prediction interval for a future observation of Y_0 is given by:

- $P.I._{(1-\alpha)\times 100\%}(Y_0) = \hat{Y}_0 \pm t_{1-\frac{\alpha}{2}(n-p)}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0^{\mathrm{T}}\mathbf{C}\mathbf{x}_0)}$

Model adequacy checking: Coefficient of determination

- $$\boxed{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},}$$

  interpreted as the fraction of variability in Y explained by the set of explicative variables $(x_1, x_2, \ldots, x_{p-1})$. If no linear dependency exists then R^2 lies near 0; in the case of a strong linear dependency it lies near 1.

- However, this coefficient artificially increases with the number of explicative variables consider in the model. This is because SSR will rise and SST never changes.
- Thus the R^2 needs to be adjusted to account for the correct degrees of freedom.

Adjusted Coefficient of determination

- The adjusted R^2 is calculated as:

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{MSE}{MST}$$

- Unimportant variables will no longer cause the R^2_{adj} is always increase with the number of explanatory variables. The R^2_{adj} may actually decrease with additional explanatory variables. This simply means that the new variables add little to help explain the variation in Y.

Adjusted Coefficient determination

- Because of the adjustment, R^2_{adj} can no longer be represented as the fraction of variability of Y accounted by the regression. However, this measure is useful when comparing two regressions with different number of explanatory variables. If the model with more explanatory variables has a lower R^2_{adj} that simply means that the additional variables add little to explain the variation in Y.

Example: yield of a chemical process

The yield (Y) of a chemical process is supposed to be related to the reagent concentration (x_1) and the operating temperature (x_2). To investigate the relationship between the variables a sample of 8 chemical processes was observed:

y	81	89	83	91	79	87	84	90
x_1	1	1	2	2	1	1	2	2
x_2	150	180	150	180	150	180	150	180

The linear regression model, $y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, for $i = 1, \ldots, 8$, with the usual assumptions was assumed. The following matrices were obtained from the raw data:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{C} = \begin{bmatrix} 16.375 & -0.75 & -0.09166667 \\ -0.75 & 0.5 & 0.0 \\ -0.09166667 & 0.0 & 0.00055556 \end{bmatrix},$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{bmatrix} 684 \\ 1032 \\ 113310 \end{bmatrix} \text{ and } \mathbf{y}^{\mathrm{T}}\mathbf{y} = 58618.$$

Example: yield of a chemical process

a) Get the estimated regression equation.

b) Get an pontual estimate of the mean yield to the values: $\mathbf{x}_0 = (1, 1, 150)$, $\mathbf{x}_0^* = (1, 2, 170)$ and $\mathbf{x}_0^{**} = (1, 3, 150)$.

c) Test the significance of this regression model. Assume that $\alpha = 0.01$.

d) Calcule the coefficient of determination and the adjusted coefficient of determination.

e) Test the hypothesis of the temperature (x_2) not be important in the explanation of the expected value of Y.

f) Calculate the 99% confidence interval for the mean yield for $x_1 = 1$ and $x_2 = 150$. Calculate the 99% prediction interval for the yield value with the same values of x_1 and x_2.

Multicollinearity

- If several explanatory variables are highly correlated (0.90 and above), some of the diagonal components of the inverse matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{C}$ will be very large. Consequently, the confidence intervals for some of the coefficients (say the β_j) tend to be very wide. This problem is called multicollinearity.

- For example, let suppose that we have just two explanatory variables x_1 and x_2. Now, suppose that the model fits well to the the data set. Then the overall test should reject $H_0 : \beta_1 = \beta_2 = 0$.

Multicollinearity

- Nevertheless, if x_1 and x_2 are highly correlated, the individual tests may lead to accept $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. This paradoxical situation can occur because both variables convey essentially the same information. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot.

- Multicollinearity occurs because the variables contain redundant information. If one of the variables doesn't seem logically essential to your model, removing it may reduce or eliminate multicollinearity.

Model checking

- Standardized residuals (d_i): residuals: $e_i = y_i - \hat{y}_i$ with $E(e_i) = 0$ and sample variance $\frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = MSE = \hat{\sigma}^2$
- $\boxed{d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}}$, $i = 1, \ldots, n$ with $E(d_i) = 0$ and $\mathrm{VAR}(d_i) \approx 1$
- If the fitted regression model is adequate, we expected the standardized residuals to look like independent draws from an $N(0, 1)$.
- Studentized residuals(r_i): Since $\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H})\mathbf{y}$, $\mathrm{COV}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$ and $\mathrm{VAR}(e_i) = \sigma^2(1 - h_{ii})$, $i = 1, \ldots, n$.
- $\boxed{r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, \quad h_{ii} = \mathbf{x}_i^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{x}_i}$

Model checking

- **Univariate Diagnostics and Graphs:** The first step in a regression analysis is generally to examine all of the variables in the model.
- One dimensional graphs such as histograms or boxplots are also very useful to see if there are any outliers, a point which deviates from the model, in the covariates or response.
- Outliers in a covariate can often indicate a point that will have very high influence on the fitted model. Outliers in the response are often points for which the model will not fill well. In some cases it may be wise to omit them from the analysis.

Model checking-QQ-Plot

- QQ-plot of e_i to check the normality assumption and possible outliers. In statistics, a QQ-plot ("Q" stands for quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). If the two distributions being compared are similar, the points in the QQ-plot will approximately lie on the line $y = x$.

Model checking

- Residual plots: The residuals e_i or standardized residuals $d_i = e_i/\sqrt{\hat{\sigma}^2}$ are used to obtain various residual plots for model checking:
  1. Plot e_i against the fitted model $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$, where $\mathbf{x}_i$ is the i-th row of the design matrix $\mathbf{X}$. This plot can be used to check the constant variance of ϵ_i. If the model is appropriate for the data the plot should show an even scatter. Any discernible pattern in the plot means that the regression equation does not describe the data correctly, since pattern forms when the residuals are unevenly distributed about the regression line. Outliers may also get identified in such a plot.
  2. In addition one also needs scatter plots with e_i or d_i on the vertical axis and each predictor variable, by turn, on the horizontal axis. These should show the same amount of variation in the residuals for all the predictors.

Model checking

- However, in simple linear regression, an observation that is unconditionally unusual in either its y or x value is called a univariate outlier, but it is not necessarily a regression outlier.
- While it is relatively easy to find outliers in univariate datasets and in simple regression, it is harder in multiple regression. An outlier in a regression setting may not be an outlier in any of the individual variables.
- Generally, regression outliers will have a large standardized residual in absolute value. Typically we will examine a point with $|d_i| > 2$ as a possible outlier.

High leverage points and influential observations

- It is helpful to distinguish leverage points and influential cases.
- **Leverage:** refers to the influence of an observation because it is outlying in the x-direction. Leverage statistics are based on the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, we have that:

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j.$$

  In addition, it can be shown that $0 < h_{ii} < 1$ for all i and $tr(\mathbf{H}) = p$.
- Thus, if h_{ii} is large relatively to other h_{ij} (in magnitude), then y_i will be a major contributor to the fitted value $\hat{y}_i$. Large leverage points have been defined as those bigger than $2p/n$.

High leverage points and influential observations

- The leverage h_{ii} has another interpretation. It measures the distance of $\mathbf{x}_i$ to the center of the explanatory variables, where $\mathbf{x}_i$ is the ith data point of the design matrix $\mathbf{X}$. For instance, consider the simple linear regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$. It can be shown that:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Consequently, if the i-th data point of the explanatory variables is far away from the center, then it has a high leverage and pulls the model fit toward itself. It is, therefore, useful in linear regression analysis to check the high leverage points.

High leverage points and influential observations

- **Influential observations** of a linear regression model are defined as those points that significantly affect the inferences drawn from the data. Methods for assessing the influence are often derived from the change in the $\hat{\boldsymbol{\beta}}$ if the observations are removed from the data.

- The well-known statistics for assessing influential observations is the Cook's distance. The Cook's distance for the i-th observation is defined as:

$$\text{Cook's distance: } D_i = \frac{e_i^2}{pMSE}\frac{h_{ii}}{(1-h_{ii})^2} = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})},$$

since

$$r_i = e_i/\sqrt{MSE(1-h_{ii})}$$

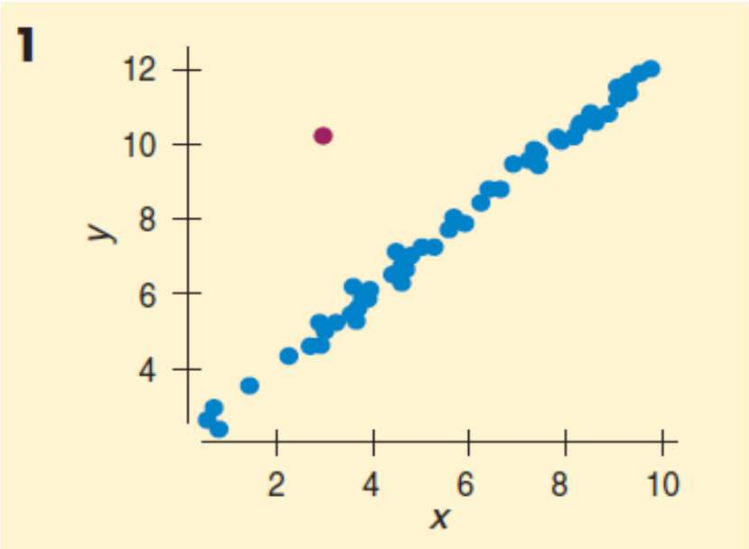
.

High leverage points and influential observations

- Cook's distance can thus be understood to depend on three quantities:
  1. The number of variables, p;
  2. A component reflecting how well the model fits y, r_i^2;
  3. A measure of how much an observation is discrepant from the rest of the data in the independent variables, h_{ii}.
- Large D_i indicate influential data points, was suggest that points with $D_i > \frac{4}{n-p}$ are influential point.

Outliers, Leverage and Influence

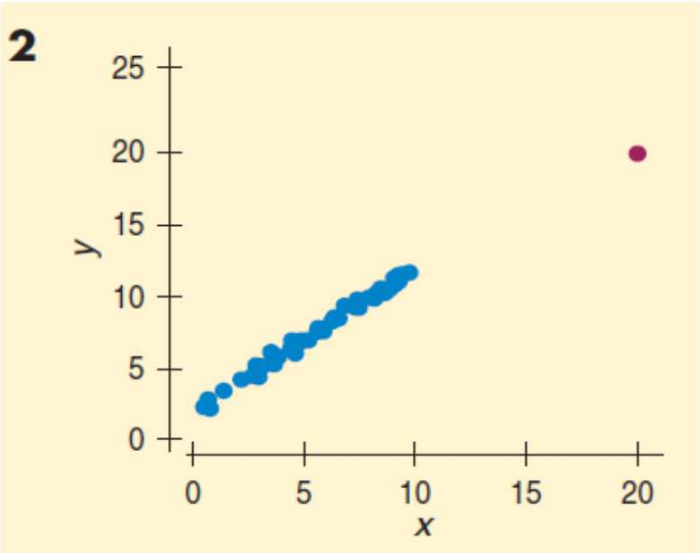
Tell whether the point is a high-leverage point, if it has a large residual and if it is influential.



- Not high-leverage
- Large residual
- Not very influential

Outliers, Leverage and Influence

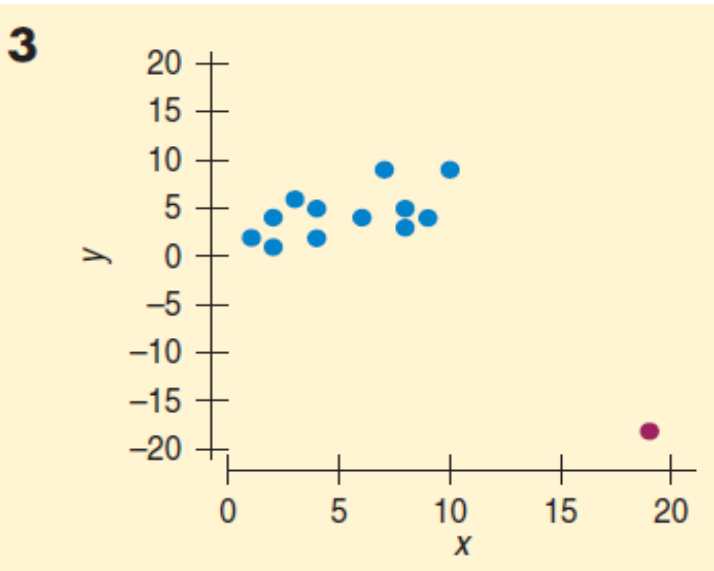
Tell whether the point is a high-leverage point, if it has a large residual and if it is influential.



- High-leverage
- Small residual
- Not very influential

Outliers, Leverage and Influence

Tell whether the point is a high-leverage point, if it has a large residual and if it is influential.



- High-leverage
- Medium/Large residual
- Very influential (omitting the red point will change the slope dramatically!)

High leverage points, influential observations and outliers

To summarize:

- An observation might be an outlier in either the x or the y direction;
- Outliers in x are called leverage points. These are diagnosed with the leverage statistic, h_{ii};
- Outliers in the y direction can be diagnosed with studentized residuals;
- An outliers may not be influential in the regression results,
- High leverage points are necessary but not sufficient conditions for influential observations;
- Observations are influential when the regression results change a lot as a consequence of leaving an observation out of the analysis. These are diagnosed with the Cook's distance statistics, D_i.

Quantitative and Qualitative Predictor Variables in Regression

- The response variable y must be quantitative.
- Each independent predictor variable can be either a quantitative or a qualitative variable, whose levels represent qualities or characteristics and can only be categorized.
- We can allow a combination of different variables to be in the model, and we can allow the variables to interact.
- A quantitative variable x can be entered as a linear term, x, or to some higher power such as x^2 or x^3.
- We could use the first-order model:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Quantitative and Qualitative Predictor Variables in Regression

- We can add an interaction term and create a second-order model:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Qualitative predictor variable are entered into a regression model through dummy or indicator variables.

Quantitative and Qualitative Predictor Variables in Regression

- For example, suppose each employee included in a study belongs to one of three ethnic groups: A, B or C, we can enter the qualitative variable ethnicity into the model using two dummy variables:

$$x_1 = \begin{cases} 1 & \text{if group is B} \\ 0 & \text{if not} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if group is C} \\ 0 & \text{if not} \end{cases}$$

- The model allows a different average response for each group.

Quantitative and Qualitative Predictor Variables in Regression

- $\hat{\beta}_1$ measures the difference in the average responses between groups B and A, while $\hat{\beta}_2$ measures the difference between groups C and A.
- When a qualitative variable involves k categories, $(k-1)$ dummy variables should be added to the regression model.

2.7. Selection of variables and model building.

- Model Selection in linear regression attempts to suggest the best model for a given purpose. Recall that the two main purposes of linear regression models are:
  1. Estimate the effect of one or more covariates while adjusting for the possible confounding effects of other variables.
  2. Prediction of the outcome for the next set of similar subjects.

- In variable selection there are two, often competing, criteria to be considered:
  1. The selected model should fit the data well;
  2. A simpler model (fewer covariates) is preferred over a more complex model. This is called the Principle of Parsimony.

2.7. Selection of variables and model building.

- Some Approaches:
  - Backwards selection
  - Forwards selection
  - All subsets selection
  - Stepwise Regression
  - R^2 criterion
  - Adjusted R^2 criterion.

Backwards Selection:

- First run a model with all covariates included in the model. Then, check to see which of the covariates has the largest p-value, and eliminate it from the model, leaving $p-2$ independent variables left in the model. Repeat this procedure with those that are left, continually dropping variables until some stopping criterion is met. A typical criterion is that all p-values are above some threshold.
- Backward elimination requires at most p regressions.

Forwards Selection:

- First run a model with no covariates, included in the model, i.e., intercept only. Then, run $p-1$ separate models, one for each of the possible independent variables, keeping track of the p-values each time. At the next step, consider a model with a single variable in it, the one with the lowest p-values at the first step. Repeat this procedure, so that at the second step, consider all models that have two parameters in it, the one selected at the first step, and all others, one at a time, and create the second model as the one where the second value has the smallest p-value, and so on. Continue to add variables until some stopping criterion is met. A typical criterion is that all p-values left at some stage are above some threshold, so no more new parameters are added.

- Forward selection requires fitting at most $1 + p(p-1)/2$ regressions.

2.7. Selection of variables and model building.

All subsets regression:

- Alternative to backwards/forwards procedures, a generic term which describes the idea of calculating some fit criterion over all possible models. In general, if there are $p-1$ potential predictor variables, there will be 2^{p-1} possible models.

Stepwise Regression:

- As with Forward Selection, start with the null model and add a variable. At subsequent stages, however, we will also consider the possibility of dropping one of the variables before adding another. If a variable in the current model is insignificant then delete it, otherwise consider adding a variable. May require more fits than forward selection but generally results in a simpler model.

2.7. Selection of variables and model building.

- **R^2 criterion:** Choose the model with largest R^2. In general, this model will simply be the largest model, so not a very useful criterion. Can be helpful in choosing among models with the same numbers of included parameters.

- **Adjusted R^2 criterion:** As above, but Adjusted R^2 penalizes for numbers of parameters, so largest model not necessarily always best. Generally selects models that are too large, because "penalty" is too small.

Some Comments.

Procedure for developing a multiple regression model:

- Select the predictor variables to be included in the model.
- Use the analysis of variance F, R^2 and adjusted R^2 to determine how well fit the model fits the data.
- Check the t tests for the partial regression coefficients to see which ones are contributing significant information in the presence of the others.
- Use residual plots to check for violation of the regression assumptions: normality and inequality of variances.

Some Comments.

- Confidence intervals can be generated by computer to estimate the average value of y, $E(y)$, for given value of $\mathbf{x}$. Prediction intervals can be used to predict a particular observation y for given value of $\mathbf{x}$. For given $\mathbf{x}$, prediction intervals are always wider than confidence intervals. Be careful with extrapolation!

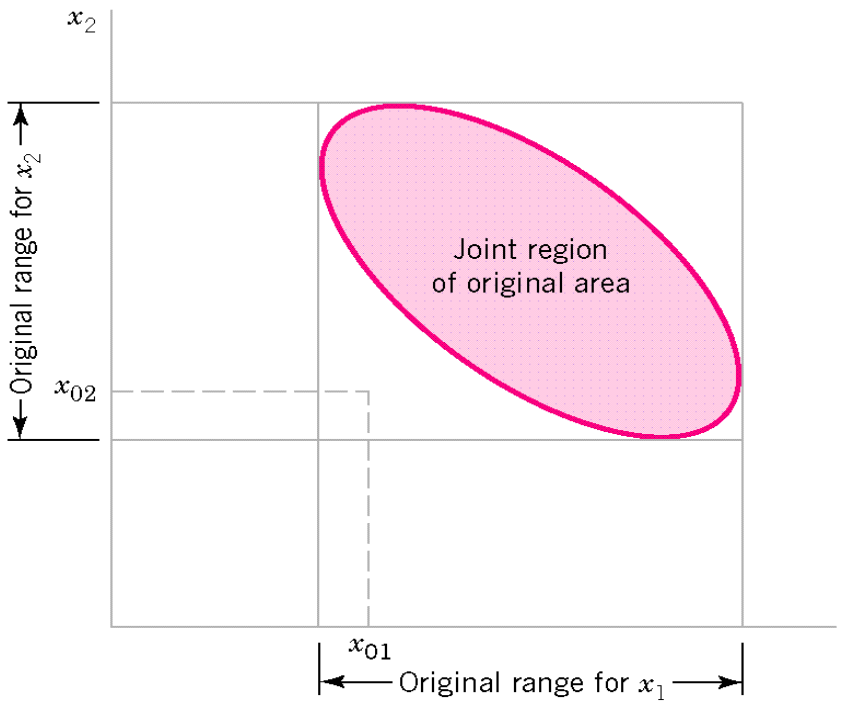


Figure 6-11 An example of extrapolation in multiple regression.

Example

Prestige data set with 

```
Prestige                  package:car                  R Documentation

Prestige of Canadian Occupations

Description:

     The 'Prestige' data frame has 102 rows and 6 columns. The
     observations are occupations.

Usage:

     Prestige

Format:

     This data frame contains the following columns:

     education Average education of occupational incumbents, years, in
          1971.

     income Average income of incumbents, dollars, in 1971.

     women Percentage of incumbents who are women.

     prestige Pineo-Porter prestige score for occupation, from a social
          survey conducted in the mid-1960s.

     census Canadian Census occupational code.

     type Type of occupation.  A factor with levels (note: out of
          order): 'bc', Blue Collar; 'prof', Professional, Managerial,
          and Technical; 'wc', White Collar.
```

Example

Prestige data set with 

```
> library(car)
> head(Prestige)
                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
physicists              15.64  11030  5.13     77.6   2113 prof
```

Example

```
> reg1 <-lm(prestige ~ education + log2(income) + women, data=Prestige)
> summary(reg1)

Call:
lm(formula = prestige ~ education + log2(income) + women, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-17.364  -4.429  -0.101   4.316  19.179

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -110.9658    14.8429  -7.476 3.27e-11 ***
education       3.7305     0.3544  10.527  < 2e-16 ***
log2(income)    9.3147     1.3265   7.022 2.90e-10 ***
women           0.0469     0.0299   1.568     0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.093 on 98 degrees of freedom
Multiple R-squared:  0.8351,    Adjusted R-squared:   0.83
F-statistic: 165.4 on 3 and 98 DF,  p-value: < 2.2e-16
```

Example

```
> prestige_hat<-fitted(reg1) # predicted values
> head(as.data.frame(prestige_hat))

                    prestige_hat
gov.administrators      65.07260
general.managers        71.50702
accountants             60.16243
purchasing.officers     54.21544
chemists                65.55434
physicists              72.70790
```

Example

```
> prestige_resid<-residuals(reg1) # residuals
> head(as.data.frame(prestige_resid))

                    prestige_resid
gov.administrators        3.727401
general.managers         -2.407019
accountants               3.237568
purchasing.officers       2.584560
chemists                  7.945657
physicists                4.892102
```

Example

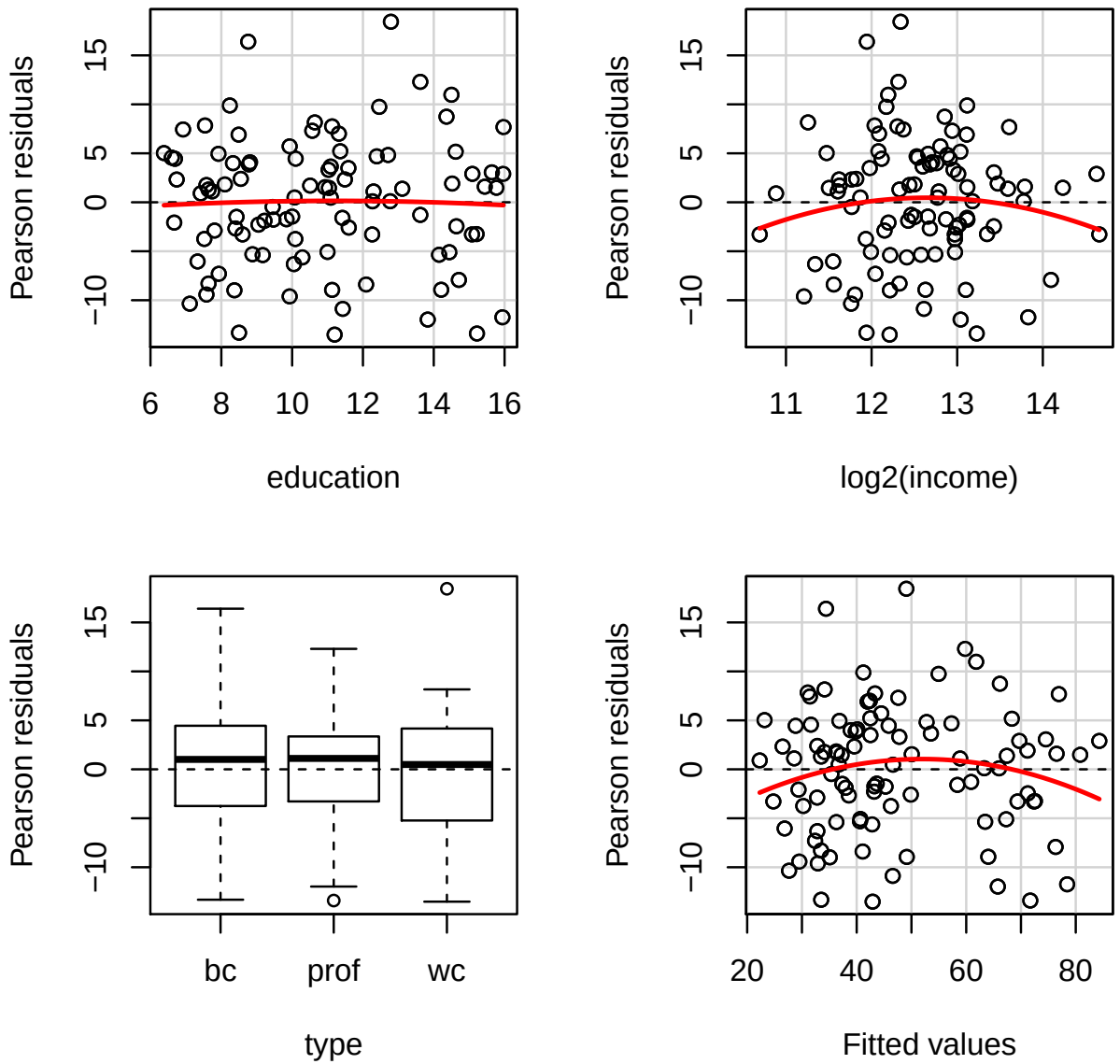
NOTE: “type” is a categorical or factor variable with three options:
bc(blue collar), prof(professional, managerial, and technical) and wc(white collar).
R automatically recognizes it as factor and treat it accordingly.

```
> reg2 <-lm(prestige ~ education + log2(income) + type, data = Prestige)
> summary(reg2)

Call:
lm(formula = prestige ~ education + log2(income) + type, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-13.511  -3.746   1.011   4.356  18.438

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -81.2019    13.7431  -5.909 5.63e-08 ***
education      3.2845     0.6081   5.401 5.06e-07 ***
log2(income)   7.2694     1.1900   6.109 2.31e-08 ***
typeprof       6.7509     3.6185   1.866   0.0652 .
typewc        -1.4394     2.3780  -0.605   0.5465
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.637 on 93 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.8555,    Adjusted R-squared:  0.8493
F-statistic: 137.6 on 4 and 93 DF,  p-value: < 2.2e-16
```

Example

```
residualPlots(reg2)
              Test stat Pr(>|t|)
education        -0.237    0.813
log2(income)     -1.044    0.299
type                 NA       NA
Tukey test       -1.446    0.148
```

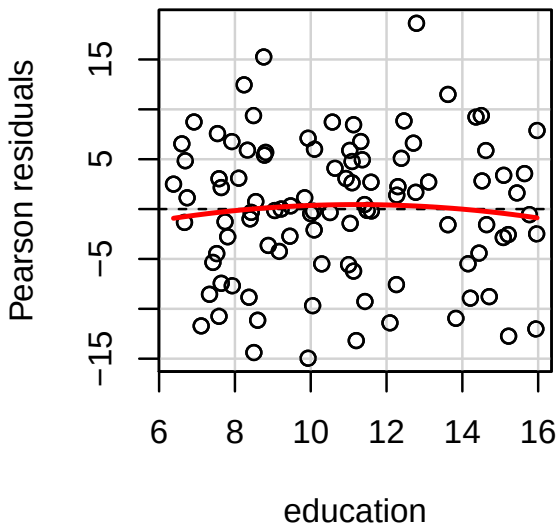
Example

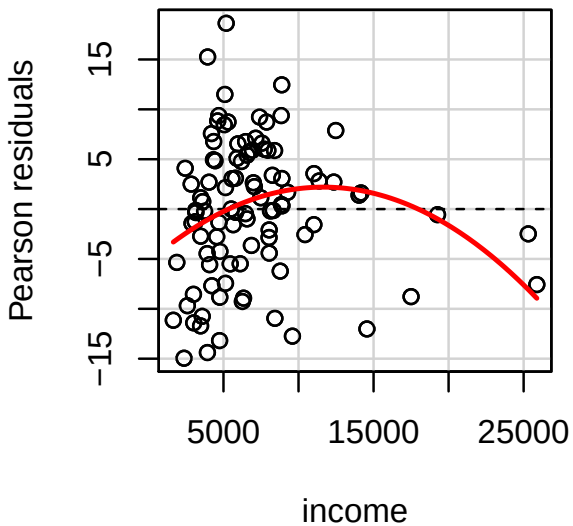

Pearson residuals
15
5
0
−10
6
8
10
12
14
16
education
Pearson residuals
15
5
0
−10
11
12
13
14
log2(income)
Pearson residuals
15
5
0
−10
bc
prof
wc
type
Pearson residuals
15
5
0
−10
20
40
60
80
Fitted values

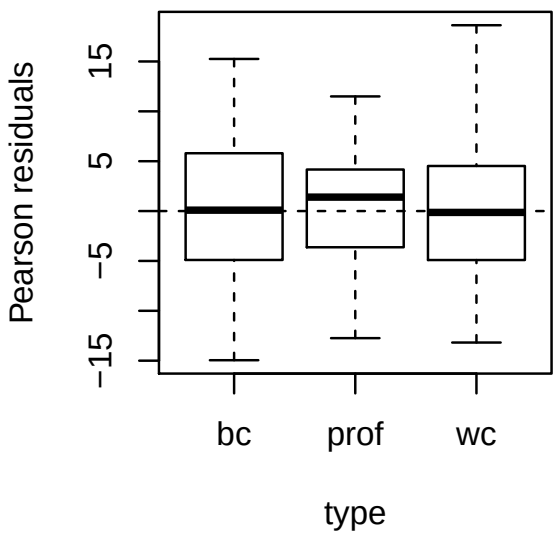
Example

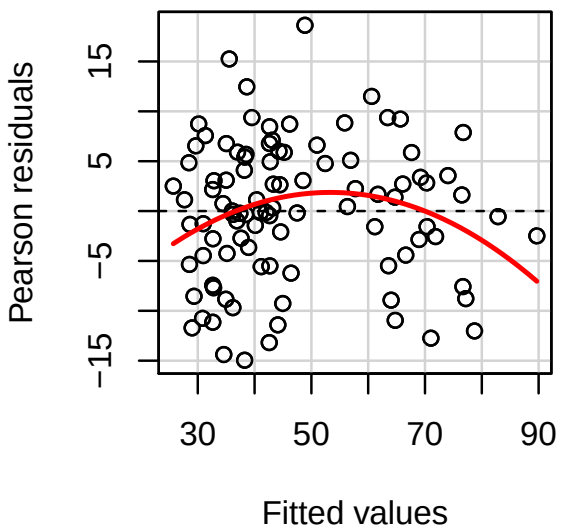
```
> reg3 <-lm(prestige ~ education + income + type, data = Prestige)
> residualPlots(reg3)
           Test stat Pr(>|t|)
education     -0.684    0.496
income        -2.886    0.005
type             NA       NA
Tukey test    -2.610    0.009
```

Example


Pearson residuals
15
5
−5
−15
6
8
10
12
14
16
education


Pearson residuals
15
5
−5
−15
5000
15000
25000
income

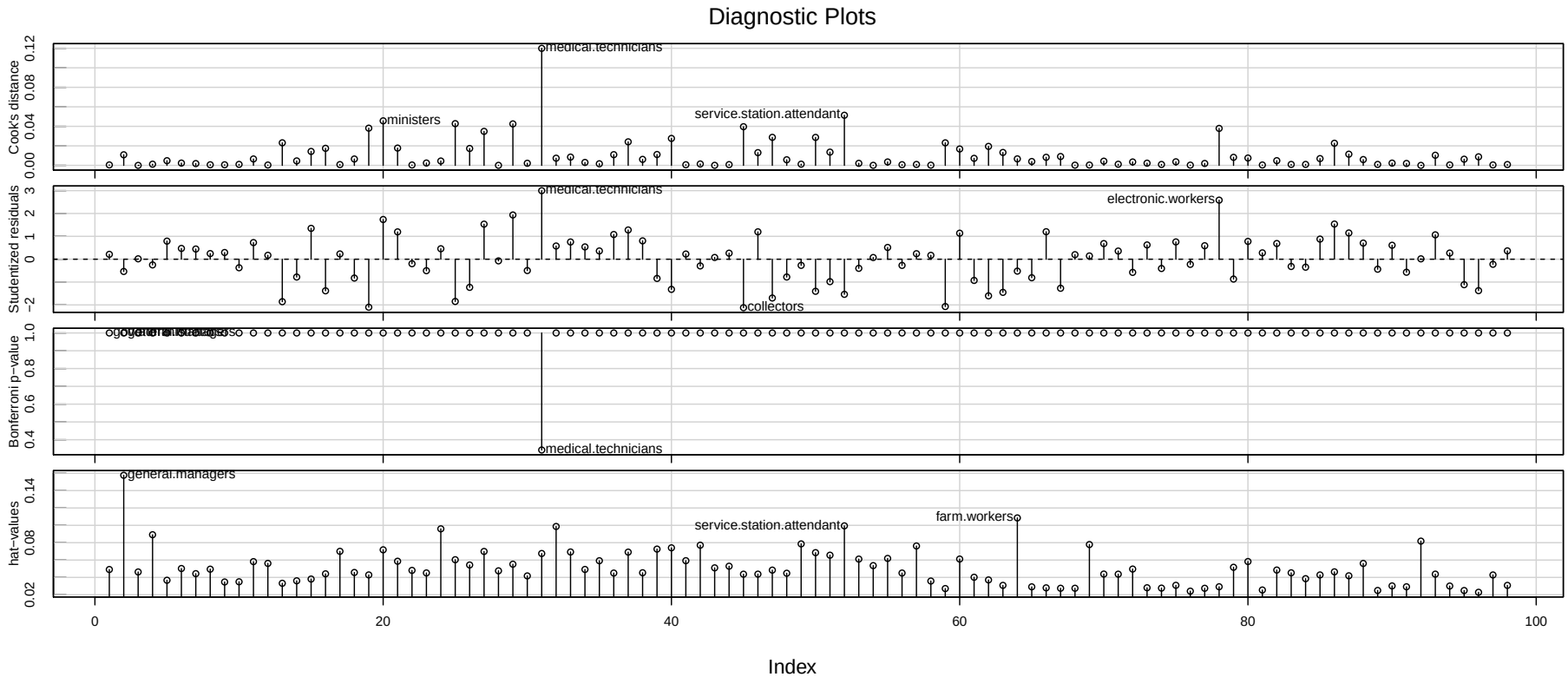

Pearson residuals
15
5
−5
−15
bc
prof
wc
type


Pearson residuals
15
5
−5
−15
30
50
70
90
Fitted values

Example

```
> influenceIndexPlot(reg2, id.n=3)


# Cook's distance measures how much an observation influences the overall
model or predicted values

# Studentized residuals are the residuals divided by their estimated
standard deviation as a way to standardized

# Bonferronitest to identify outliers

# Hat-points identify influential observations (have a high impact on the
predictor variables)
```

Example


Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p−value
hat−values
medical.technicians
ministers
service.station.attendant
electronic.workers
collectors
general.managers
farm.workers
0.00
0.04
0.08
0.12
−2
0
1
2
3
0.4
0.6
0.8
1.0
0.02
0.08
0.14
0
20
40
60
80
100
Index

Example

influencePlot(reg2)

Creates a bubble-plot combining the display of Studentized residuals, hat-values and Cook's distance (represented in the circles).