REGRESSION AND CORRELATION – PART 2

INFERENCES ABOUT β_1

SECTION 12.3 (DEVORE & BERK 2012)

EXAMPLE 12.11

Reconsider Example 12.5, where $x = CO_2$ and $y =$ tree growth mass. There are 8 observations, 2 at each of the 4 x values. We assume that the parameters of the true regression line are $\beta_1 = .0085$, $\beta_0 = -2.35$, and $\sigma = .5$.

Using R, we proceeded to generate a sample of random deviations $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_8$ from a normal distribution with mean 0 and standard deviation .5.

We added $\tilde{\varepsilon}_i$ to $\beta_0 + \beta_1 x_i$ to obtain 8 corresponding y values. We then estimated slope, intercept, and standard deviation using regression, and repeated the process a total of 20 times.

EXAMPLE 12.11

Table 12.1 Simulation results for Example 12.11

	$\hat{\beta}_0$	$\hat{\beta}_1$	s
1	−2.606	0.0086	0.312
2	−3.639	0.0104	0.345
3	−3.316	0.0100	0.530
4	−3.042	0.0093	0.475
5	−3.400	0.0103	0.441
6	−3.932	0.0107	0.328
7	−2.533	0.0090	0.423
8	−2.862	0.0100	0.676
9	−2.152	0.0081	0.401
10	−2.975	0.0093	0.409
11	−2.255	0.0084	0.639
12	−3.003	0.0095	0.437
13	−3.187	0.0093	0.587
14	−2.424	0.0087	0.598
15	−1.490	0.0073	0.735
16	−1.812	0.0074	0.332
17	−1.845	0.0079	0.552
18	−4.080	0.0107	0.520
19	−2.958	0.0090	0.718
20	−1.670	0.0072	0.574

EXAMPLE 12.11

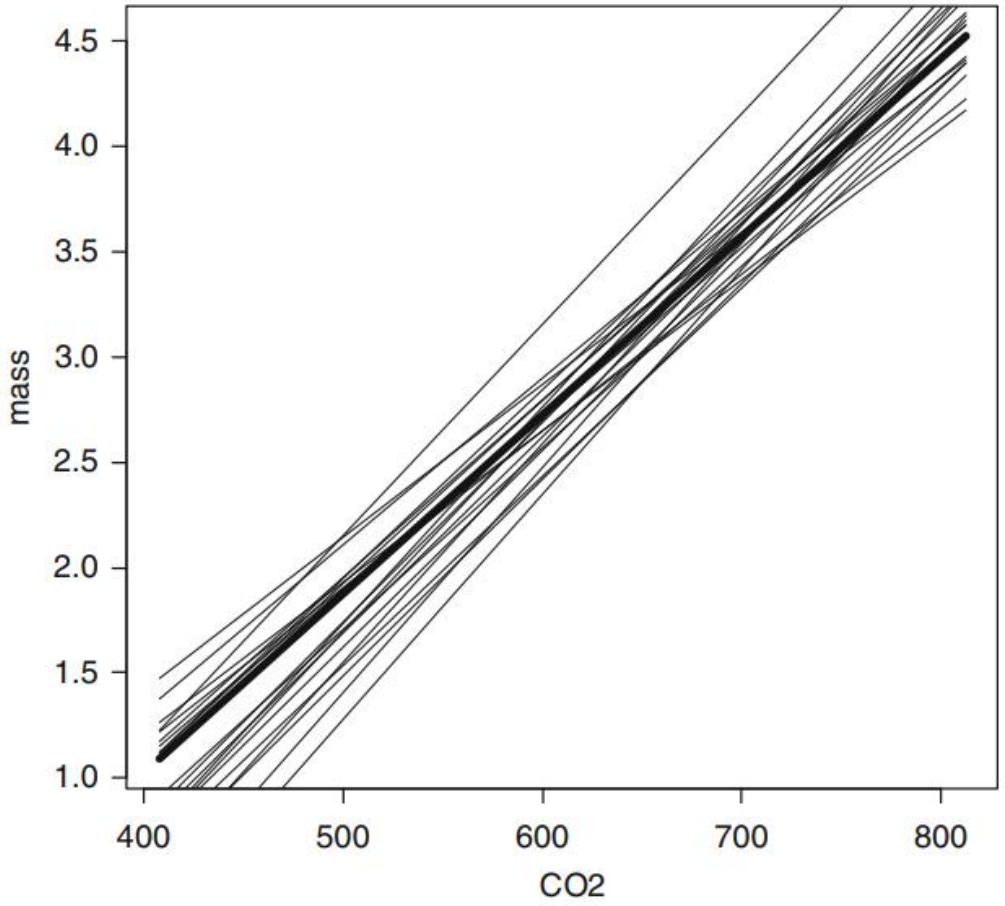


Figure 12.15 Simulation results from Example 12.11: graphs of the true regression line and 20 least squares lines (from R)

PROPERTIES OF $\hat{\beta}_1$

1. The mean value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator of β_1 (the distribution of $\hat{\beta}_1$ is always centered at the value of β_1).
2. The variance and standard deviation of $\hat{\beta}_1$ are

$$V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

   Replacing σ by its estimate s gives an estimate for $\sigma_{\hat{\beta}_1}$: $s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$.
3. The estimator $\hat{\beta}_1$ has a normal distribution (because it is a linear function of independent normal random variables).

INFERENCES CONCERNING β_1

Theorem: the assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

has a *t* distribution with $n - 2$ df.

- T-values are reported by every statistical software!
- in R you access to df and t-values by using the function `summary()`, as

```
summary(lm(y~x))
```

CONFIDENCE INTERVAL FOR β_1

Definition: A $100(1-\alpha)\%$ CI for the slope β_1, of the true regression line is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$$

- The CI is provided in some statistical software
- The CI is centered at the point estimate of the parameter and extends out to each side by an amount that depends on the confidence level and on the extent of variability in the estimator $\hat{\beta}_1$.

HYPOTHESIS-TESTING

- The null hypothesis in a test about β_1 will be an equality statement.
- We denote the null value with β_{10} (“beta one naught”)
- The test statistic has a t distribution with $n - 2$ df when H_0 is true, so the type I error probability is controlled at the desired level α by using an appropriate t critical value.
- Most common pair of hypotheses: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$.
- Unless n is quite small, H_0 will be rejected and the utility of the model confirmed precisely when r^2 is reasonably large.

HYPOTHESIS TESTING

Null hypothesis: $H_0: \beta_1 = \beta_{10}$

Test statistic value: $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Alternative Hypothesis	Rejection Region for level α Test
$H_a: \beta_1 > \beta_{10}$	$t \geq t_{\alpha,n-2}$
$H_a: \beta_1 < \beta_{10}$	$t \leq t_{\alpha,n-2}$
$H_a: \beta_1 \neq \beta_{10}$	Either $t \geq t_{\alpha,n-2}$ or $t \leq t_{\alpha,n-2}$

The model utility test is the test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, in which case the test statistic value is the t ratio $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$.

REGRESSION AND ANOVA

- The splitting of the total sum of squares SST in SSE – which measures unexplained variation – and SSR – which measures variation explained by the linear relationship – is strongly reminiscent of one-way ANOVA.
- The null hypothesis $H_0: \beta_1 = 0$ can be tested against $H_a: \beta_1 \neq 0$ by constructing an ANOVA table and rejecting H_0 if $f \geq F_{\alpha,1,n-2}$

Source of variation	Df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

ANOVA IN R

- The function we use in R is `anova()`
- If we want to compare $H_0: \beta_1 = 0$ against $H_a: \beta_1 \neq 0$ for a model `m1`, we simply write

```
anova(m1)
```

- If we want to compare two different models `m1` and `m2` which have different amount of predictors, then we write

```
anova(m1,m2)
```

`R script on screen`

INFERENCES ABOUT $\mu_{Y \cdot x^*}$

SECTION 12.4 (DEVORE & BERK 2012)

Let x^* be a specified value of the independent variable x; then, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ can be regarded either as a point estimate of $\mu_{y \cdot x^*}$ or as a prediction of the *y* value that will result from a single observation made when $x = x^*$.

This by itself gives no information concerning how precisely $\mu_{y \cdot x^*}$ has been estimated or *y* has been predicted. To remedy this problem, we can develop:

- A confidence interval for $\mu_{y \cdot x^*}$
- A prediction interval for a single *y* value.

SOME REMARKS

Remark: Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are subject to sampling variability: both values will vary from sample to sample.

Example: Assume that $\beta_0 = 50$ and $\beta_1 = 2$; then, two different samples of (x, y) pairs may give $\hat{\beta}_0 = 52.35, \hat{\beta}_1 = 1.895$ and $\hat{\beta}_0 = 46.52, \hat{\beta}_1 = 2.056$ respectively.

Remark: In the same way that a confidence interval for β_1 was based on properties of the sampling distribution of $\hat{\beta}_1$, a confidence interval for a mean y value in regression is based on properties of the sampling distribution of the statistic $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Substitution of the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ into $\hat{\beta}_0 + \hat{\beta}_1 x^*$ followed by some algebraic manipulation leads to the representation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as a linear function of the y_i's:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} \right] y_i = \sum_{i=1}^{n} d_i y_i$$

The coefficients $d_1, d_2, \ldots, d_n$ in this linear function involve the x_i's and x^*, all of which are fixed.

Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ where x^* is some fixed value of x. Then

1. $E(\hat{y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$
   Thus, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $\beta_0 + \beta_1 x^*$ (i.e., for $\mu_{y \cdot x^*}$)

2. $V(\hat{y}) = \sigma^2_{\hat{y}} = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum x_i^2 - (\sum x_i)^2/n}\right] = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$
   and the standard deviation $\sigma_{\hat{y}}$ is the square root of this expression. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$, denoted by $s_{\hat{y}}$ or $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$, results from replacing σ by its estimate s: $s_{\hat{y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$

3. $\hat{y}$ has a normal distribution (because the y_i's are normally distributed and independent)

ABOUT THE PROPERTIES

The variance of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is smallest when $x^* = \bar{x}$ and increases as x^* moves away from $\bar{x}$ in either direction. Thus, the estimator of $\mu_{y \cdot x^*}$ is more precise when x^* is near the center of the x_i's than when it is far from the x values where observations have been made. This implies that both CI and PI are narrower for an x^* near $\bar{x}$ than for an x^* far from $\bar{x}$.

INFERENCES CONCERNING $\mu_{Y \cdot x^*}$

Just as inferential procedures for β_1 were based on the t variable obtained by standardizing $\hat{\beta}_1$, a t variable obtained by standardizing $\hat{\beta}_0 + \hat{\beta}_1 x^*$ leads to a CI and test procedures here.

Theorem: The variable

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{y} - (\beta_0 + \beta_1 x^*)}{s_{\hat{y}}}$$

has a t distribution with $n - 2$ df.

CONFIDENCE INTERVAL

Definition: A $100(1-\alpha)\%$ CI for $\mu_{y\cdot x^*}$, the expected value of *y* when $x = x^*$, is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_0+\hat{\beta}_1 x^*} = \hat{y} \pm t_{\frac{\alpha}{2},n-2} \cdot s_{\hat{y}}$$

where $s_{\hat{y}}$ is as defined in slide 16.

- The CI is centered at the point estimate for $\mu_{y\cdot x^*}$ and extends out to each side by an amount that depends on the confidence level and on the extent of variability in the estimator on which the point estimate is based
- The range/width of the CI is varying with the distance of x^* to $\bar{x}$.

CONFIDENCE VS PREDICTION INTERVAL

Note: A CI refers to a parameter, or population characteristic, whose value is fixed but unknown to us. In contrast, a future value of *Y* is not a parameter but instead a random variable. For this reason we refer to an interval of plausible values for a future *Y* as a prediction interval rather than a confidence interval.

Confidence Interval	Prediction Interval
We use the error of estimation: $\beta_0 + \beta_1 x^* - \left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) =$ Fixed unknown term – random variable	We use the error of prediction: $(\beta_0 + \beta_1 x^* + \varepsilon) - \left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) =$ Random variable – random variable
with ε, there is more uncertainty in prediction $\Rightarrow$ PI is wider than CI	

$$V\left[Y - \left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right)\right] = \text{variance of prediction error}$$
$$= V(Y) + V\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right)$$
$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$
$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

Furthermore, because $E(Y) = \beta_0 + \beta_1 x^*$ and $E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \beta_0 + \beta_1 x^*$, $E\left(Y - \hat{\beta}_0 + \hat{\beta}_1 x^*\right) = 0$. Then, the standardized variable

$$T = \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

has a t distribution with $n - 2$ df.

Definition: A $100(1-\alpha)\%$ PI for a future Y observation made when $x = x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} =$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \hat{y} \pm t_{\frac{\alpha}{2},n-2} \cdot \sqrt{s^2 + s^2_{\hat{y}}}$$

- If $s^2 = 0$, then PI=CI.
- In the long run the interval will actually contain y values $100(1-\alpha)\%$ of the time
- As $n \to \infty$ the width of the CI approaches 0, whereas the width of the PI approaches $2z_{\alpha/2}\sigma$ (because even with perfect knowledge of β_0 and β_1, there will still be uncertainty in prediction).

INTERVALS – IMPLEMENTATION IN R

- Use the function `lm` to fit the model, then apply the function `predict` / `predict.lm` to the fitted model
- One argument is necessary for specifying the interval type

  `interval = "confidence“` or `interval = “prediction“`
- The argument `newdata` allows to determine for which data the prediction is carried out

Example: `- script on screen -`

CORRELATION

SECTION 12.5 (DEVORE & BERK 2012)

Definition: Pearson's sample correlation coefficient for the n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is defined by $r = \frac{S_{xy}}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$.

1. The value of r does not depend on which of the two variables is labeled x and which one is labeled y
2. The value of r is independent of the units in which x and y are measured
3. $-1 \leq r \leq 1$
4. $r = 1 \Leftrightarrow$ all (x_i, y_i) pairs lie on a straight line with positive slope, and $r = -1 \Leftrightarrow$ all (x_i, y_i) pairs lie on a straight line with negative slope
5. The square of r gives the value of the coefficient of determination that would result from fitting the simple linear regression model, i.e. $(r)^2 = r^2$.

When can it be said that there is a strong correlation between the variables, and when is the correlation weak?

Rule of thumb: The correlation is weak if $0 \leq |r| \leq 0.5$, is strong if $0.8 \leq |r| \leq 1$ and moderate otherwise

Note: if $r = 0.5 \Rightarrow r^2 = 0.25$, so only 25% of observed y variation would be explained by a linear model where x predicts y.

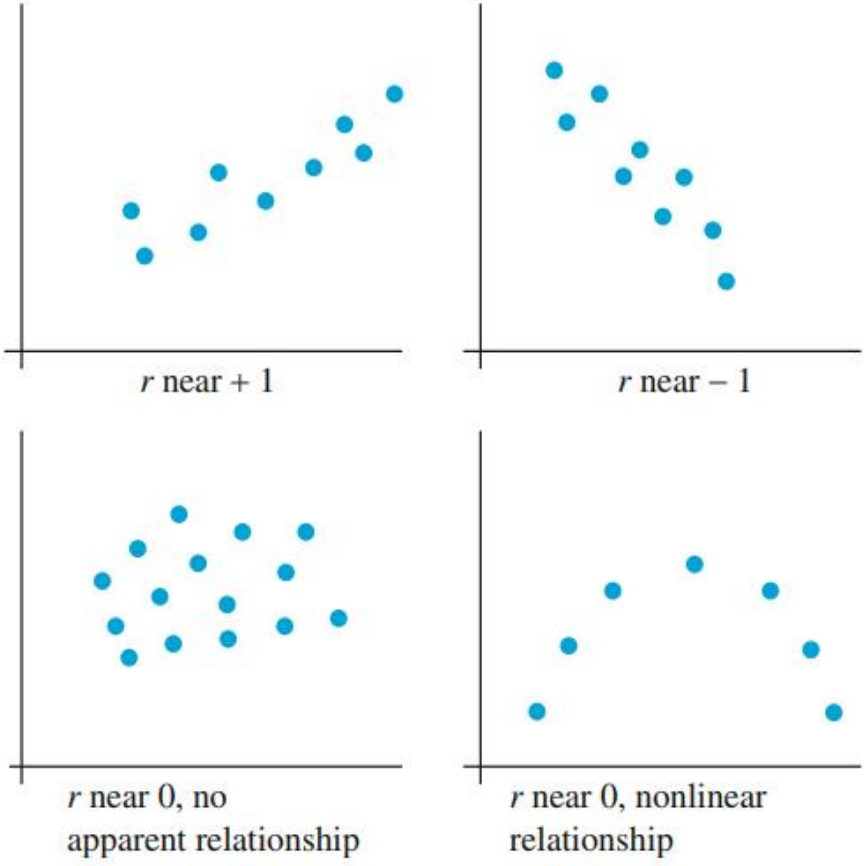


Figure 12.24 Data plots for different values of r

THE POPULATION CORRELATION COEFFICIENT ρ

In order to carry out statistical inference, we define the population correlation coefficient ρ as

$$\rho(X,Y) = \frac{\text{Cov(X, Y)}}{\sigma_X \sigma_Y}$$

where

$$\text{Cov(X, Y)} = \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y) \text{ for discrete}(X,Y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y) dx dy \text{ for continuous}(X,Y) \end{cases}$$

OBSERVATIONS ABOUT CORRELATIONS

- The population correlation coefficient ρ is a parameter or population characteristic, just as μ_X, μ_Y, σ_X, and σ_Y are
- We can use the sample correlation coefficient to carry out various inferences about ρ.
- In particular, a commonly used estimator for ρ is sample correlation coefficient *r* given by

$$\hat{\rho} = R = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

To test hypotheses about ρ, we must assume that

1. both *X* and *Y* are random,
2. with joint distribution given by the bivariate normal pdf.

Remarks

- Recall: for given $X = x$, the conditional distribution of *Y* is normal with mean $\mu_{Y \cdot x} = \mu_2 + ((\rho\sigma_2)/\sigma_1)(x - \mu_1)$ and variance $(1 - \rho^2)\sigma_2^2$.
  This corresponds to the model used in simple linear regression with $\beta_0 = \mu_2 - (\rho\mu_1\sigma_2)/\sigma_1, \beta_1 = \rho\sigma_2/\sigma_1$, and $\sigma^2 = (1 - \rho^2)\sigma_2^2$ independent of x.

- An implication of this finding is the following: if the observed pairs (x_i, y_i) are actually drawn from a bivariate normal distribution, then the simple linear regression model is an appropriate way of studying the behavior of *Y* for fixed x
- If $\rho = 0$, then $\mu_{Y \cdot x} = \mu_2$ independent of x; in fact, the joint pdf could be factored into a part involving x only and a part involving y only.
  This implies that *X* and *Y* are independent variables, i.e., here independence follows from non-correlation

PARTIAL CHECK AND TESTING FOR MULTIVARIATE NORMALITY

There is no completely satisfactory or unique way to check the plausibility of the bivariate normal distribution.

- Partial check: bivariate normality implies that the marginal distribution of both *X* and *Y* are normal. Hence, construct two separate normal probability plots, one for the sample x_i's and another for the sample y_i's.
  Moreover, a scatter plot of the (x_i, y_i)'s should show a roughly elliptical shape.
- Formal tests: several tests exist for testing for multivariate normality, such as Mardia's , Henze-Zirkler's, Royston's, Doornik-Hansen's, or the Energy test. Each of them has its particular advantages and disadvantages.

Note: the following procedure should not be used if the sample size *n* is very small or if multivariate normality cannot be assumed based on visual inspection or formal tests.

TESTING FOR THE ABSENCE OF CORRELATION

When $H_0: \rho = 0$ is true, the test statistic

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

follows a t distribution with $n-2$ df. Consequently, a p-value based on a t distribution with $n-2$ df can be calculated as already described before.

Alternative Hypothesis	Rejection Region for level α Test
$H_a: \rho > 0$	$t \geq t_{\alpha,n-2}$
$H_a: \rho < 0$	$t \leq t_{\alpha,n-2}$
$H_a: \rho \neq 0$	Either $t \geq t_{\alpha,n-2}$ or $t \leq t_{\alpha,n-2}$

IMPLEMENTATION IN R

- Use the function `cor.test`
- As for the many other tests, the argument `alternative` allows to specify a one- or two-sided test
- The argument `method` permits to evaluate different types of correlation

Example: `- script on screen -`

ASSESSING MODEL ADEQUACY

SECTION 12.6 (DEVORE & BERK 2012)

Motivation

- A very basic idea to test a model adequacy is to superimpose a graph of the best-fit function on the scatter plot of the data.
  However, any tilt or curvature of the best-fit function may obscure some aspects of the fit that should be investigated.
- A different, more effective approach is to compute the fitted or predicted values $\hat{y}_i$ and the residuals $e_i = y_i - \hat{y}_i$. Then, one can plot various functions of these computed quantities.
  To derive some properties of the residuals, let's assume that the simple linear regression model is correct (so $y = \hat{\beta}_0 + \hat{\beta}_1 x$), and interpret the i^{th} residual as random variable given by $e_i = Y_i - \hat{Y}_i$.

PROPERTIES OF RESIDUALS

In the linear regression setting holds:

1. $E(e_i) = E\left(Y_i - \hat{Y}_i\right) = E(Y_i) - E\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i)$
   $= 0$

2. $V(Y_i - \hat{Y}_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right]$

Replacing σ^2 by s^2 and taking the square root in the second equation gives the estimated standard deviation of a residual.

STANDARDIZED RESIDUALS

In order to analyze the residuals, we need to standardize them by subtracting the mean value (zero) and then dividing by the estimated standard deviation.

Definition: The standardized residuals are given by

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}, \qquad i = 1, \ldots, n$$

STANDARDIZED RESIDUALS

Remarks:

- In general, the variances of the residuals differ from one another! However, if n is reasonably large, the term $1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}$ will be approximately 1. Therefore, some sources use e_i / s as the standardized residual.
- Several other ways exist for defining standardized residuals. These may vary depending on the software and / or package used

DIAGNOSTIC PLOTS

Some plots that can be used for an assessment of model validity and usefulness:

1. y_i on the vertical axis versus x_i on the horizontal axis (standard plot)
2. y_i on the vertical axis versus $\hat{y}_i$ on the horizontal axis (prediction plot)
3. e_i^* (or e_i) on the vertical axis versus x_i on the horizontal axis (residual plot against the independent variable)
4. e_i^* (or e_i) on the vertical axis versus $\hat{y}_i$ on the horizontal axis (residual plot against the fitted value)
5. A normal probability plot of the standardized residuals (or residuals)

DIAGNOSTIC PLOTS

- If the prediction plot (2.) yields points close to the 45° line ($y = x$), then the estimated regression function gives accurate predictions of the values actually observed. Hence, this plot provides a visual assessment of model effectiveness in making predictions.
- If the model is correct, neither residual plot should exhibit distinct patterns. They allow to check the assumptions of both homoscedasticity and linearity.
- Plot 4 is a combination of Plot 2 and Plot 3, showing implicitly both how residuals vary with x and how fitted values compare with observed values.
- Plot 5 allows to assess the plausibility of the normality of ε.

SOME DIFFICULTIES

Some typical problems you may encounter when fitting (linear) models are:

1. A nonlinear probabilistic relationship between x and y is appropriate.
2. The variance of ε (and of Y) is not a constant σ^2 but depends on x (or y).
3. The selected model fits the data well except for a very few discrepant or outlying data values, which may have greatly influenced the choice of the best-fitting function.
4. The error term ε does not have a normal distribution.
5. When the subscript i indicates the time order of the observations, the ε_i's exhibit dependence over time.
6. One or more relevant independent variables have been omitted from the model.

FIGURE 12.28

Nonlinear relationship

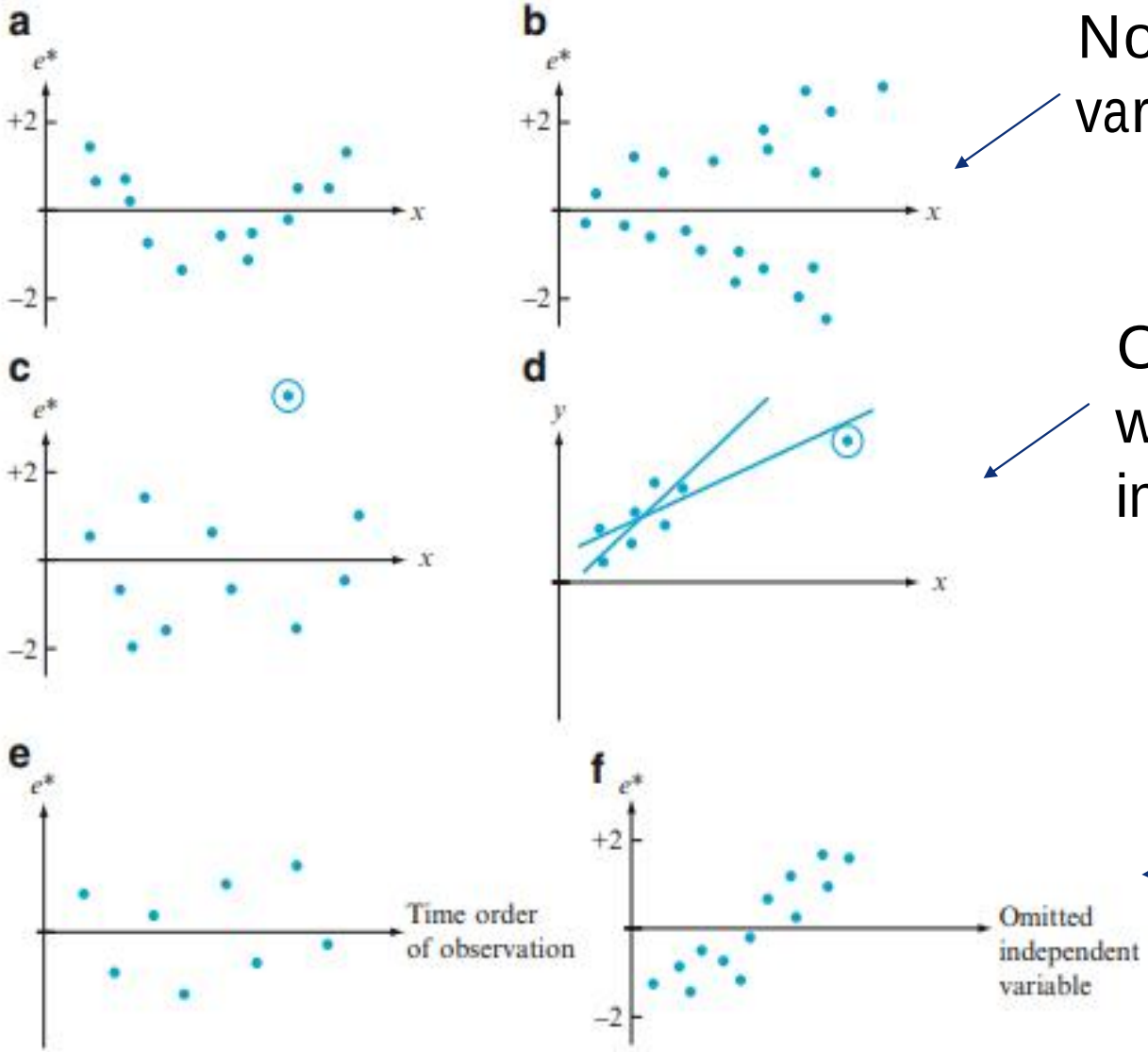
Non-constant variance

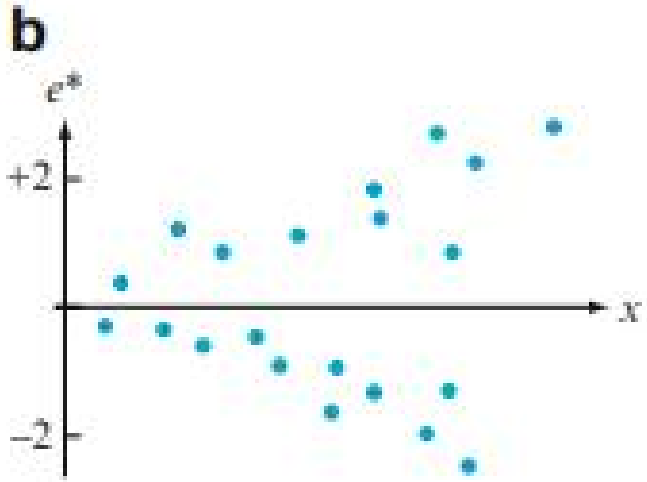
Discrepant observation

Observation with large influence

Dependence in errors

Variable omitted





Note: in Figure b the points gets more and more distant from the x axis. This suggests that, although a linear relationship may be reasonable, the assumption $V(Y_i) = \sigma$ for every i is probably not adequate.

HETEROSCEDASTICITY AND WEIGHTED LEAST SQUARES

Assume that ε satisfies the independence and homoscedasticity assumptions (normality is not needed) for the simple linear regression model.
Then, it can be shown that among all unbiased estimators of β_0 and β_1, the ordinary least squares estimators have minimum variance. These estimators give equal weight to each (x_i, Y_i).

However, what to do in case of heteroscedasticity? If the variance of Y increases with x then Y_i's for large x_i should be given less weight than those with small x_i.

This suggests that β_0 and β_1 should be estimated by minimizing

$$f_w(b_0, b_1) = \sum w_i[y_i - (b_0 + b_1 x_i)]^2$$

Where the w_i's are weights that decrease with increasing x_i. Minimization of the above expression yields weighted least squares estimates.

Example: Assume that the standard deviation of Y is proportional of x for $x > 0$ – that is, $V(Y) = kx^2$ - then it can be shown that the weights $w_i = 1/x_i^2$ yield a minimum variance estimators of β_0 and β_1

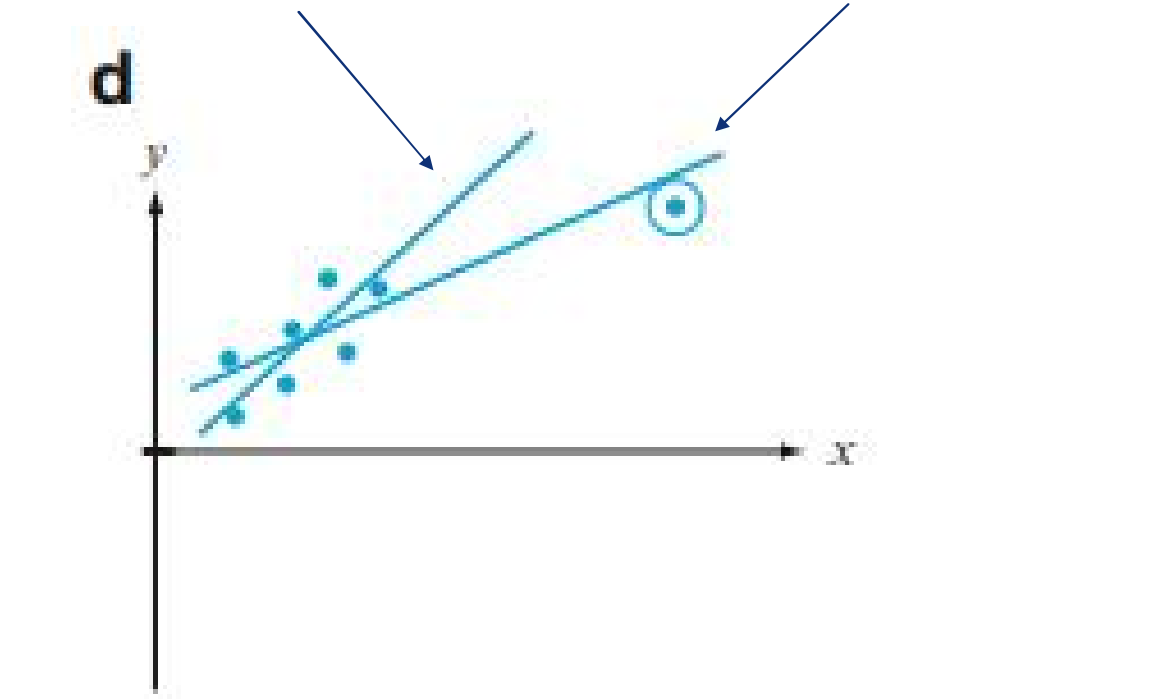
IMPLEMENTATION IN R

- Use the function `lm`
- Specify the argument `weights`

Example: `- script on screen -`

OUTLIERS

When the data set contains outliers or points having large influence on the resulting fit, one approach is to omit these points and re-compute the estimated regression equation. This is a good approach if the outliers are the result of errors in recording the data values, or experimental errors.



OUTLIERS - MAD

Another approach results in using an estimation principle which puts relatively less weight on outlying values that does the principle of least squares.

- One such principle is MAD (minimize absolute deviations), which selects $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize $\sum |y_i - (b_0 + b_1 x_i)|$.
- There are many other approaches, leading to robust regression.

Note: There are no formulas for the MAD estimates: their values must be found by using an iterative computational procedure. The same holds true for many other so-called robust regression approaches.

NON-NORMALITY

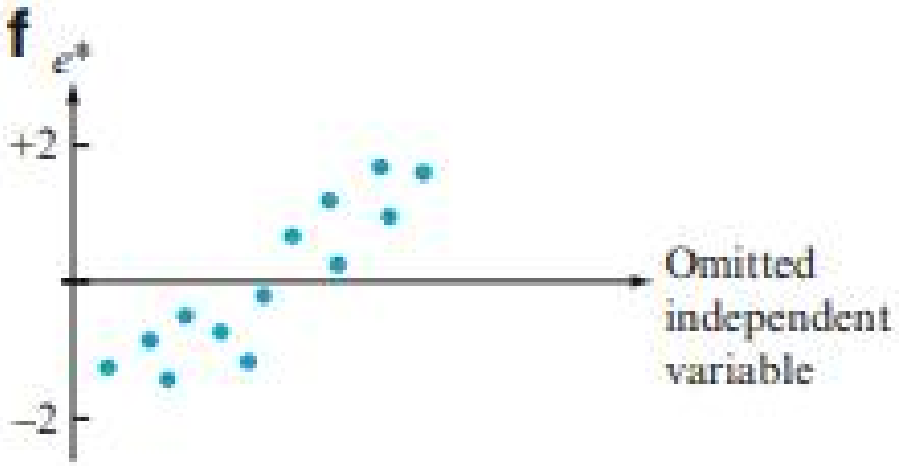
Non-normality of the residuals may also occur, e.g., the ε_i's may be

- Heavy-tailed,
- Skewed,
- (contain outliers,)
- …

In such cases, robust regression procedures may also be chosen. Robust means that these procedures produce reliable estimates for a wide variety of underlying error distributions. Least squares estimators are not robust in the same way that the sample mean $\bar{X}$ is not a robust estimator for μ.

TIME DEPENDENCY AND OMITTED VARIABLES

When a plot suggests time dependence in the error terms, we may need to transform the y's or else a model explicitly including a time variable.



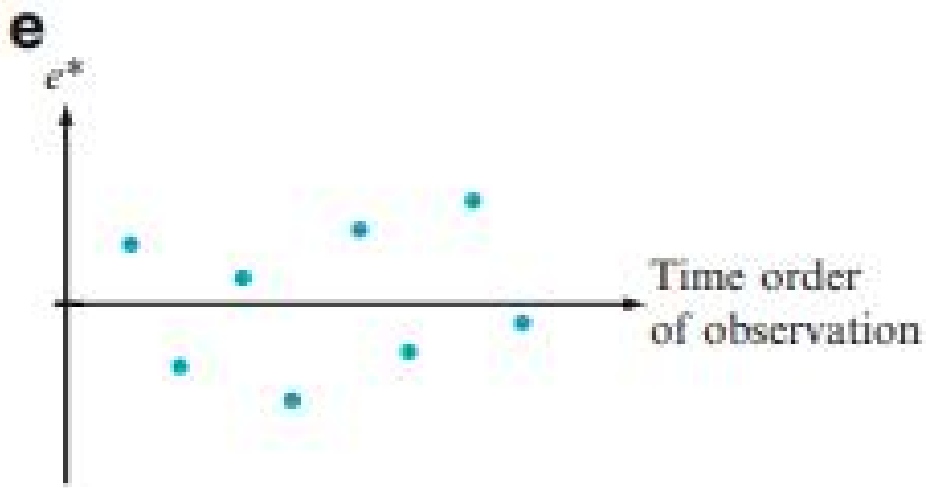


Figure f shows a pattern in the residuals when plotted against an omitted variable. It suggests considering a model that includes the omitted variable.