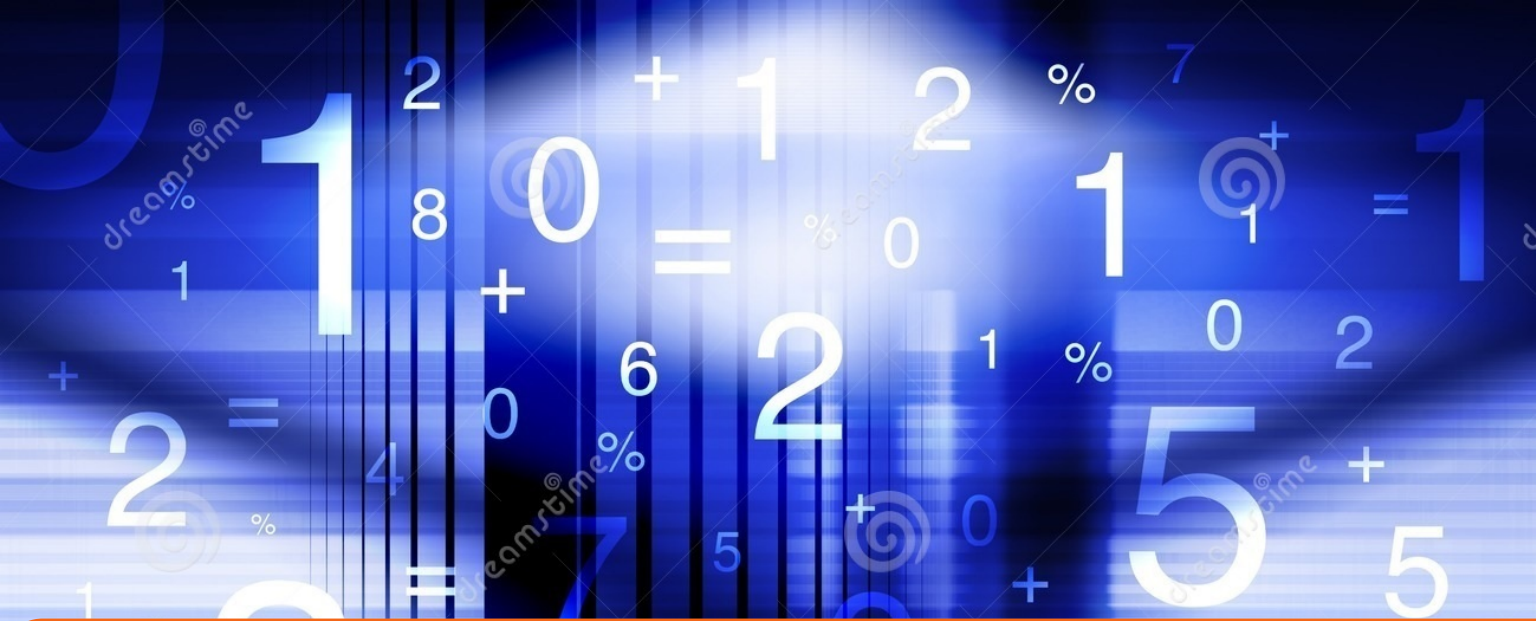
Notes on Probability and Statistics



Contentsmkboth CONTENTSCONTENTS

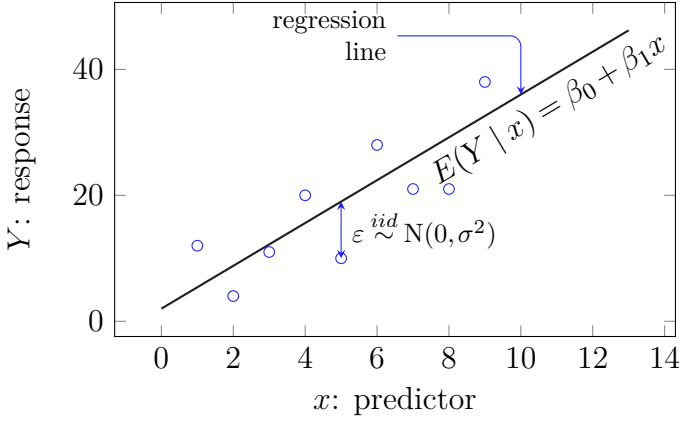
Statistics

7. Linear regression

7.1 The regression model

A **simple** linear regression model takes the form

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{7.1}$$



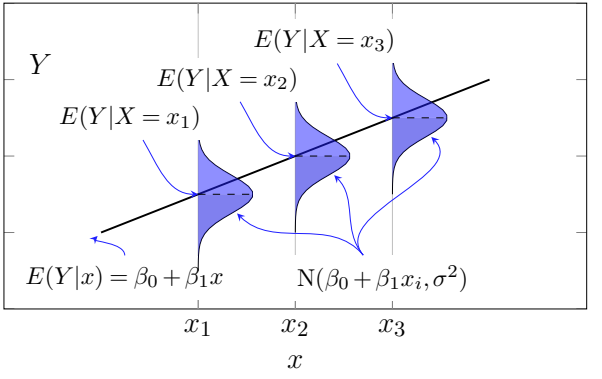
For a sample $\{(X_1,Y_1),(X_2,Y_2)\dots(X_n,Y_n)\}$ a regression model takes the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\dots,n, \tag{7.2}$$

where $Y_1, Y_2, \dots, Y_n$ are observable rv's **conditional** on $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ and

$$\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$$

are non-observable random variables.



Terminology :
x_i is called an *explanatory variable* or *independent variable* or *predictor* or *factor*;
Y_i is the *response* or *dependent variable*;
ε_i is the *error* random variable, whose realizations are called *residual*.

Note: x_i **is not considered a random variable** in linear regression because it is a realization the random variable X_i , i.e. we take the values of x_i as "given", and the term Y_i should be interpreted as a conditional:

$$Y_i \leftrightarrow Y_i \mid X_i = x_i$$

The assumption of the regression model is:

$$\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$$

This means:

1. Normality: $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, n$.
2. Independence of the errors: $\varepsilon_1, \ldots, \varepsilon_n$ are independent.
3. Homoscedasticity: $\mathrm{V}(\varepsilon_i) = \sigma^2$, with σ^2 constant for all $i = 1, \ldots, n$.

Therefore,

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2) \tag{7.3}$$

"Linear" model means that it is *linear in the unknown parameters* $\beta = \beta_0, \beta_1, \beta_2 \ldots$, and **not** in x. For example, the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{7.4}$$

is a linear regression model because it is linear in $\beta_0, \beta_1, \beta_2$.

7.2 Matrix notation

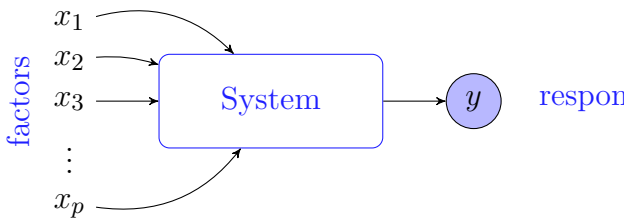
The regression model (7.2) is a set of simultaneous equations which can be written more concisely as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.5}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Multiple regression model.
With p **explanatory variables**, the model takes the form:



$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad (7.6)$$

In matrix form it still reads as (7.5) defining: $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and $\mathbf{X}$ is a $n \times (1+p)$ matrix, called the **design matrix**:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Interpretation of β_j. In the model

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

we can see that

$$\frac{\partial Y}{\partial x_j} = \beta_j, \qquad j = 1, 2, 3.$$

which means that the parameter β_j represents the marginal change in Y due to a change in x_j. **One should always verify that the sign of β_j accords with intuition.**

A quadratic model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

can be written in the form of Equations (7.5) by defining

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Main effects and interactions give the following regression function:

$$Y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{main effects}} + \underbrace{\beta_3 x_1 x_2}_{\text{interaction term}} + \varepsilon$$

can be written in the form of Equations (7.5) by defining

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Example application: Annual income model Let:

Y is annual income (\$1000/year),
x_1 is educational level (number of years of schooling),
x_2 is number of years of work experience, and
x_3 is gender ($x_3 = 0$ is male, $x_3 = 1$ is female),

Suppose we estimated the following model

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \varepsilon$$

and obtained (using statistical software),

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 = 0.8 \\ \hat{\beta}_1 = 0.8 \\ \hat{\beta}_2 = 0.5 \\ \hat{\beta}_3 = -3.0 \end{pmatrix} \qquad \text{and} \qquad \hat{\sigma} = 9.$$

We can answer questions like: "what is the probability that a female with 16 years education and no work experience will earn more than \$40,000/year?"
Recall that

$$Y \sim \mathrm{N}(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3, \sigma^2)$$

The mean for such a person is 24.8, so standardizing yields the probability:

$$\begin{aligned} P(Y > 40) &= P((Y - 24.8)/9 > (40 - 24.8)/9) \\ &= P(Z > 1.69) \\ &\approx 0.05. \end{aligned}$$

The gender variable x_3 is an **indicator variable**, since it only takes on the values 0/1 (as opposed to x_1 and x_2 which are quantitative).

7.3 The method of ordinary least squares (OLS)

To estimate the β_j's we minimize the **sum of squared errors, SSE**:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

over all possible values of the intercept and slopes. To minimize $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we differentiate with respect to $\boldsymbol{\beta}$ and equating to $\mathbf{0}$:

$$2\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}, \qquad \rightarrow \qquad \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}.$$

which is a set of linear simultaneous equations called the *normal equations* for the linear model.

Fact 7.8 — OLS estimators. Provided $\mathbf{X}^T\mathbf{X}$ is non-singular, the OLS estimators are:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{Y}.$$

The matrix $\mathbf{C}$. For convenience, let:

$$\mathbf{C} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \quad \rightarrow \quad \hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{X}^T\mathbf{Y}.$$

Fact 7.9 — Properties of the OLS estimators. The OLS estimators have useful properties:

a) $\hat{\boldsymbol{\beta}}$ is unbiased: $\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$

b) $\hat{\boldsymbol{\beta}}$ is a linear transformation of $\mathbf{Y}$, so it has the (multivariate) normal distribution

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$, and

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{C}.$$

This result implies that

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}) \tag{7.7}$$

where c_{ij} is the (i, j) element of $\mathbf{C}$, so

$$\mathrm{V}\left(\hat{\beta}_i\right) = \sigma^2 c_{ii}$$

$$\mathrm{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right) = \sigma^2 c_{ij}$$

But the true variance σ^2 is unknown, and therefore has to be estimated.

Fact 7.10 — An unbiased estimator for σ^2.

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} \quad \text{is unbiased for } \sigma^2.$$

Furthermore,

$$\frac{(n-p-1)\,\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Finally, $\hat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma}^2$.

We conclude that the estimator of the covariance matrix is $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2 \mathbf{C}$.

The standard errors of the coefficient estimates $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\}$ are

$$\sqrt{\mathrm{V}\left(\hat{\beta}_i\right)} = \hat{\sigma}\sqrt{c_{ii}}$$

7.4 Testing the significance of coefficients

Since

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}) \quad \text{then} \quad Z_i = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1).$$

Using the result that

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

and remembering the definition of a t-distribution we conclude that

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\mathrm{V}\left(\hat{\beta}_i\right)}} \sim t_{n-p-1} \tag{7.8}$$

This enables us to carry out hypothesis tests or calculate confidence intervals for coefficients.

Significance test for β_i : $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$

Let $t_0 = \hat{\beta}_i / \sqrt{\mathrm{V}\left(\hat{\beta}_i\right)}$, then we reject H_0 if

$$|t_0| > t_{\alpha/2} \tag{7.9}$$

where $t_{\alpha/2}$ is the critical values from the distribution t_{n-p-1}, which is typically ≈ 2. This test is important because **if we cannot reject H_0 it means that the variable x_i does not help explain Y** and therefore should be removed from model.

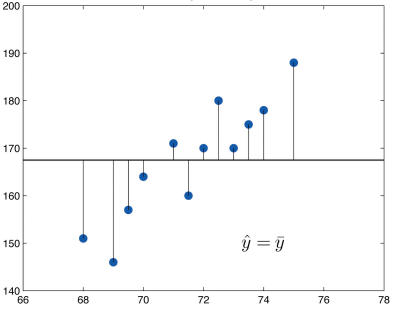
Recall that instead of this T-test we can also use the p-value, if available, and reject H_0 if $p < \alpha$.

7.5 Goodness-of-fit: R^2

The Coefficient of determination R^2 measures the proportion of the total variations in Y that can be explained by the linear model and it is defined as

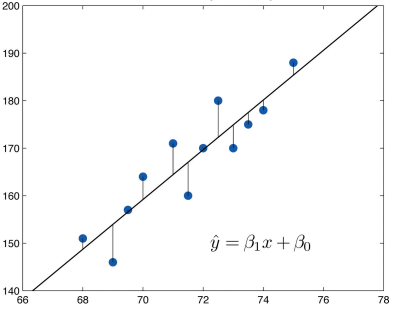
$$R^2 = 1 - \frac{SSE}{SST} \tag{7.10}$$

(a): $\sum_{i=1}^n \varepsilon_i^2 = SST$,



(a): Set I

(b): $\sum_{i=1}^n \varepsilon_i^2 = SSE$



(b): Set II

It quantifies the reduction in variability of the response variable as a result of the linear relationship with X.

R is called the sample correlation coefficient and $\approx \rho_{X,Y}$ as $n \to \infty$, and therefore measures the strength of the linear relationship.

7.5.1 Adjusted R^2

The problem with R^2 is that it cannot decrease when additional explanatory variables are added to the model, even if they have no significant effect on Y. An alternative measure, computed by most econometrics packages, is the so-called 'Adjusted R^2" :

$$\bar{R}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s_Y^2} \tag{7.11}$$

where the numerator and denominator of R^2 are divided by their respective degrees of freedom.

7.5.2 One-way ANOVA

We can decompose the total variance as follows:

$$\begin{aligned} SST &= SSR + SSE \\ \sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y})^2 \end{aligned}$$

where SSR = regression sum of squares, and the *predicted values* or fitted values is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots$. It turns out that under the no hypotheses $H_0 : \beta_1 = \beta_2 = \cdots = 0$ the statistic:

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \tag{7.12}$$

follows an F_{ν_1,ν_2} distribution, where $\nu_1 = p$ and $\nu_2 = n-p-1$. As usual, we will reject the no hypothesis if the observed F statistic is greater than the critical value from the **F probability tables**.

7.6 Assessing the model

The first step in looking at the adequacy of a model is to check the assumptions on which it is based:

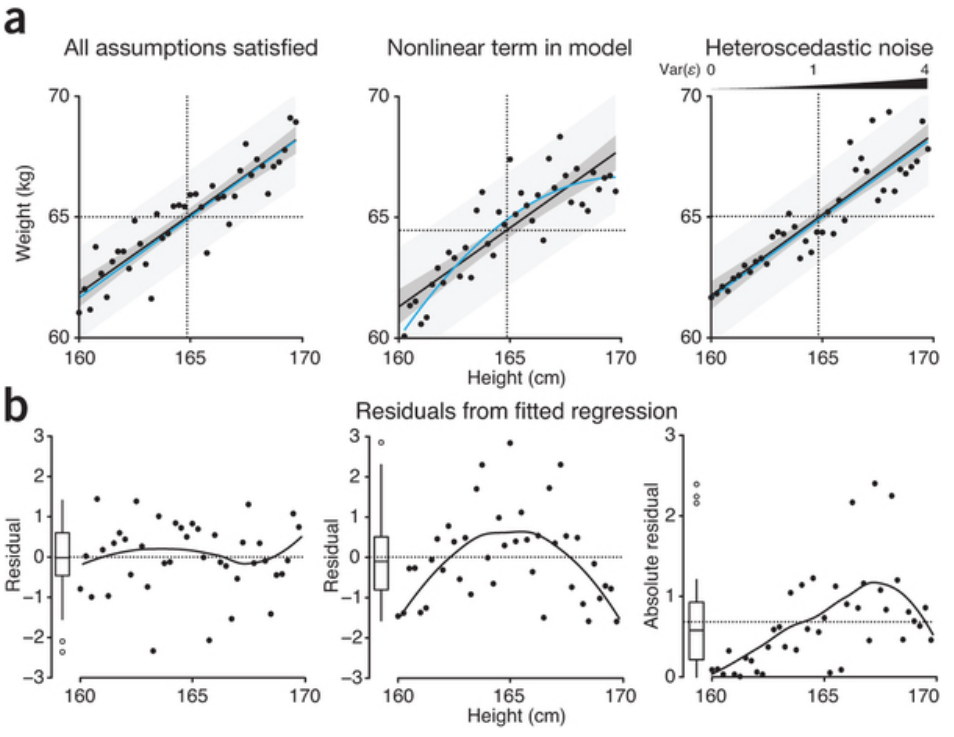
The assumption of the regression model is:

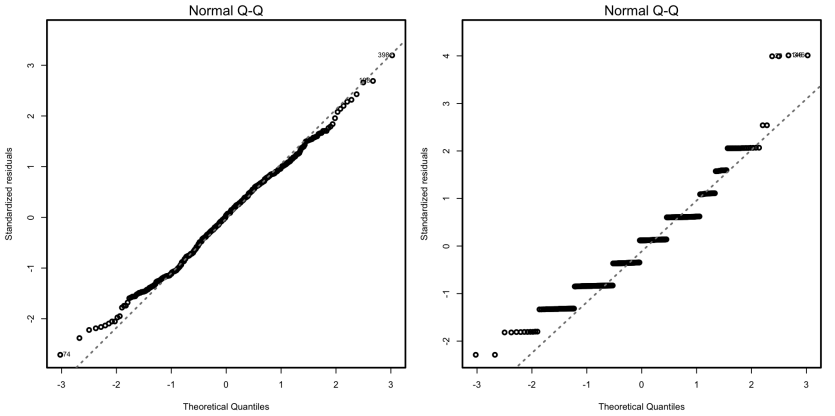
$$\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$$

This means:

1. Normality: $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, n$.
2. Independence of the errors: $\varepsilon_1, \ldots, \varepsilon_n$ are independent.
3. Homoscedasticity: $\mathrm{V}(\varepsilon_i) = \sigma^2$, with σ^2 constant for $i = 1, \ldots, n$.

Residual plots (residuals versus the fitted values) are the main tool for checking model assumptions. **No discernible pattern** in this plot is a good sign of independence, linearity and constant variance. **The Q-Q plot** (normal probability plot) is a scatterplot between the observed quantiles of residuals vs. the theoretical quantiles from the normal distribution. **A linear relationship** indicates good agreement with the normal distribution.


a
All assumptions satisfied
Nonlinear term in model
Heteroscedastic noise
Var(ε) 0
1
4
Weight (kg)
70
65
60
160
165
170
Height (cm)
b
Residuals from fitted regression
Residual
Absolute residual
3
2
1
0
−1
−2
−3

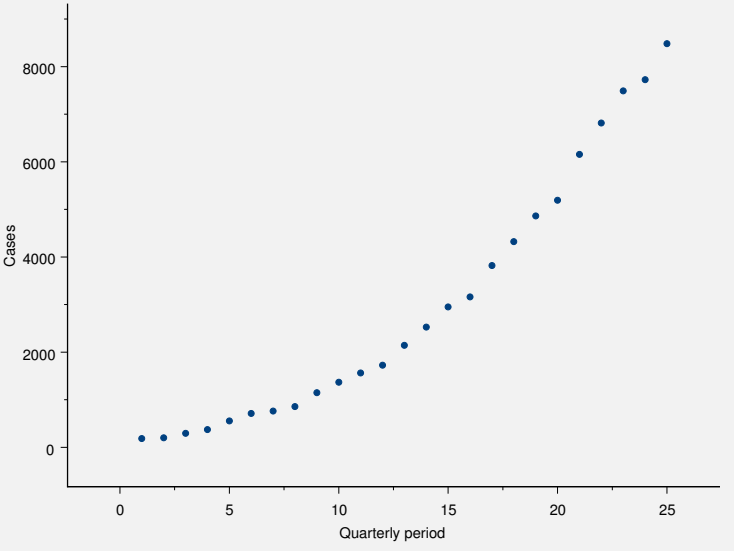

Normal Q-Q
Standardized residuals
Theoretical Quantiles
74
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Normal Q-Q
Standardized residuals
Theoretical Quantiles

QQ-plots for datasets respecting (left column) and violating (right column) the normality assumption. Source.

Example 1. — AIDS data for the USA These data are for AIDS incidence in the USA, adjusted for reporting delays. The data are taken from Rosenberg, P.S. and Gail, M.H. (1991): Backcalculation of flexible linear models of the Human Immunodeficiency Virus infection curve. *Applied Statistics*, **40**, 269-282.

Newly reported cases are recorded quarterly and the variable *Time* therefore counts 3-monthly periods, staring with the first quarter in 1982.

The scatterplot shows that the trend is not linear.



Incidence of AIDS cases in the USA against time

Let's fit the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

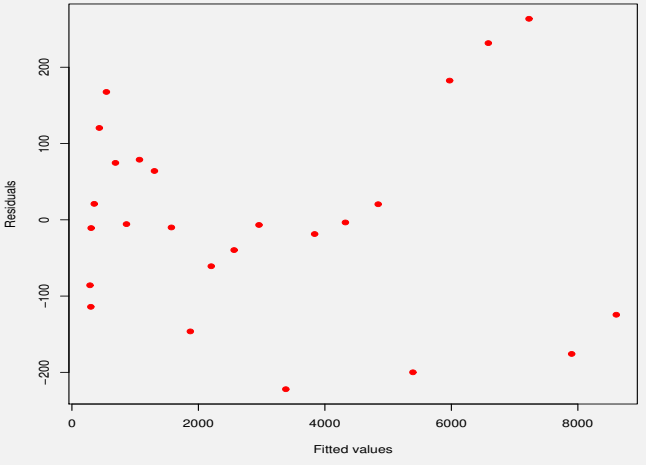
and test the coefficients β_1, β_2. We can use a standard computer package to carry out the regression.

Variable	Coefficient		s.e.	t-value	p-value
Intercept	β_0	343.5913	87.7446	3.9158	0.0007
Time	β_1	−60.1380	15.5514	−3.8671	0.0008
Time^2	β_2	15.6277	0.5806	26.9158	0.0000
$R^2 = 0.9976$	$d.f. = 22$	$\hat{\sigma} = 134.7$	$SSE = 399155$		

The fitted model is

$$Cases = 343.5913 - 60.1380 \times Time + 15.6277 \times Time^2.$$

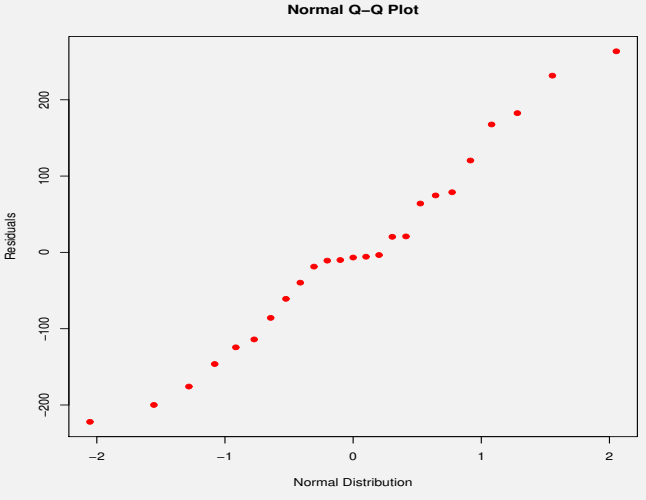
The value $R^2 = 0.9976$ above means that if it is excellent. But can this model be relied upon?



Plot of residuals against fitted values for AIDS data

- The points to the right seem a little more spread out but there is no indication that spread is a function of fitted value.

- There does seem to be some curvature − re-think the assumption about the model?
- We also need to check up on the third assumption.



Normal probability plot of AIDS residuals

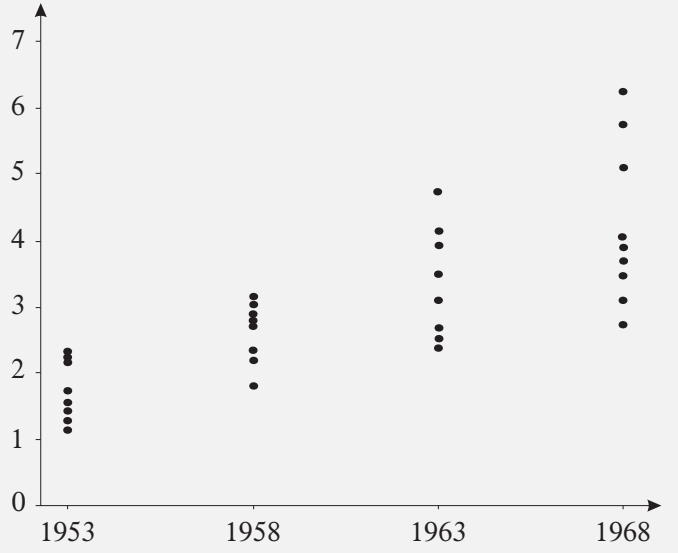
> This doesn't look too bad. We can therefore conclude that we have a reasonable model which could perhaps be improved, but which fits the data pretty well for the most part.

Example 2. — Testicular cancer The table below comprises data from Lee, Hitosugi and Peterson (1973): Rise in mortality from tumors of the testis in Japan, 1947-70. *J. Nat. Cancer Inst.*, **51**, 1485-90. It gives the populations and numbers of deaths from testicular cancer in 5-year age groups and 5-year periods in Japan. The ages refer to the lowest age in each group and the populations are expressed in millions of persons.

Table 5.4 Deaths in Japan from testicular cancer

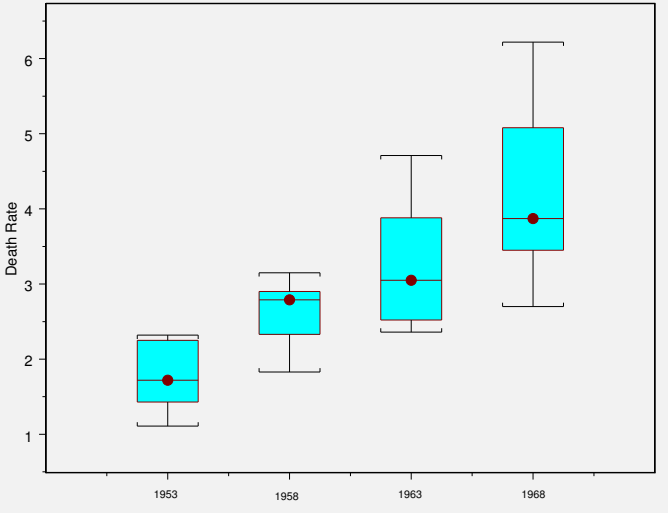
	1951-55		1956-60		1961-65		1966-70	
Age	Popn.	Dths	Popn.	Dths	Popn.	Dths	Popn.	Dths
20	20.4	27	21.3	39	22.2	56	24.0	83
25	17.2	40	20.0	58	20.6	97	21.8	125
30	12.6	18	17.1	54	19.9	77	20.8	129
35	11.7	13	12.5	36	17.0	70	19.9	101
40	11.5	26	11.5	32	12.2	29	16.8	67
45	10.3	16	11.2	26	11.1	34	12.0	37
50	9.3	16	9.8	27	10.7	27	10.7	29
55	7.6	17	8.7	19	9.2	32	10.1	39
60	5.9	13	6.8	21	7.9	21	8.4	31

The scatterplot shows that the mean death rate from cancer of the testis in Japan has been rising steadily since 1951.



Death rate against year for cancer of the testis in Japan

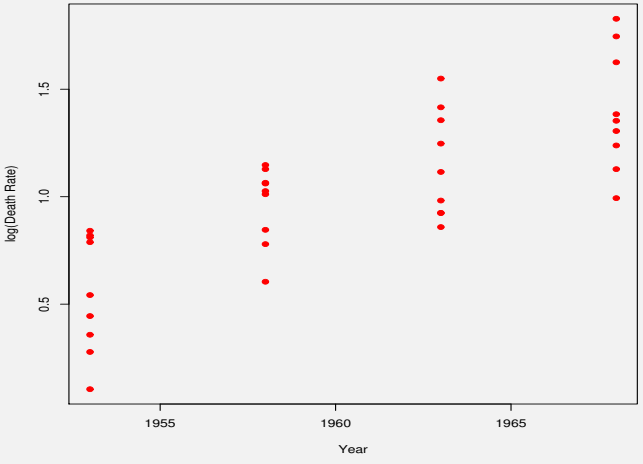
Note that there is no variability in the explanatory variable, but there is marked variability in the response. This is clearly shown in the boxplots.



Boxplots of death rates

It is clear that the variability in the data increases as the year variable increases. What should one do about this?

The answer is to look for a transformation which will stabilise the variance. Here we need a transformation which compresses large values of the response more than it compresses smaller values; something like a square root or a cube root or possibly even a log transformation. Taking the log of the death rate results in the next scatterplot.



Testicular cancer: plot of log(Death rate) against year

This looks reasonable and we could now go ahead and fit a model of the form

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

With the log transformed data, the fitted model turns out to be as given below.

Linear Model
Response: log(Death rate)

Variable		Coefficient	s.e.	t-value	p-value
Intercept	β_0	-105.9198	14.4887	-7.3105	0.0000
Year	β_1	0.0545	0.0074	7.3808	0.0000

$r^2 = 0.6157$ $d.f. = 34$ $s = 0.2479$ $SSE = 2.0891$

Of course stable variance alone is not enough because the residuals also need to be normally distributed. This can be checked with a normal q-q plot of the residuals.

Normal probability plot of testicular cancer residuals

The plot seems to show a very rough straight line, but it is not entirely convincing.

7.7 Model selection

Question: With a large number of potential predictors, how do we choose the predictors to include in the model?

> **Occam's Razor (The law of parsimony):** *"Among competing hypotheses, the one with the fewest assumptions should be selected."*
> $\rightarrow$ choose the model with the fewest number of parameters that explains the data.

Good fit Criteria

A good model would give **yes** to the following questions:

- are the assumptions met?
- good Adj-R^2?
- good t-statistics for the β_j's ($|t_0| > 2$) ?
- the sign of β_j's accords with intuition?

Example 3. — * Housing values in suburbs of Boston. Home values for 506 Boston suburbs with potential influential factors is shown below. (Source: Belsley D. A., Kuh, E. and Welsch, R. E. (1980) Regression Diagnostics.).

X_i	Description
1	Per capita crime rate by town
2	Proportion of residential land zoned for lots over 25000 square feet
3	Proportion of non-retail business acres per town
4	Charles River dummy variable (1 if tract bounds river, 0 otherwise)
5	Nitrogen oxide concentration (parts per 10 million)
6	Average number of rooms per dwelling
7	Proportion of owner-occupied units built prior to 1940
8	Weighted mean of distances to five Boston employment centers
9	Index of accessibility to radial highways
10	Full-value property-tax rater per $10000
11	Pupil-teacher ratio by town
13	Lower status of the population (percent)

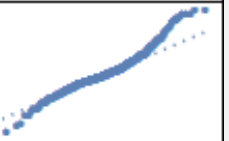
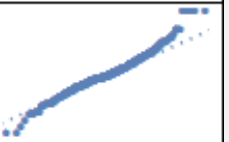
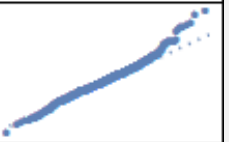
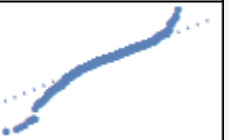
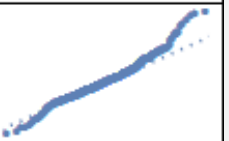
The models below are the output of an automated model selection procedure, for the response variable $Y =$ **median value of owner-occupied homes in \$1000s**, and for models that only consider main effects:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

The parameter table gives the parameter values in the second row and the T-statistic in the third row in bold.

a) Explain how the parameter values should be used to assess the adequacy of a model; give 3 example.
b) decide which model is the “best" and carefully explain why.
c) decide which model is the “worst" and carefully explain why.

Y = Median value of owner-occupied homes in $1000s

Model	Adj-R^2	Parameter Table	Residuals	Q-Q plot
M-1	0.7	β_0 β_4 β_9 β_{11} β_{13} 4.3 0.15 -0.0019 -0.039 -0.04 **46. 3.9 -1.4 -7.4 -25.**		
M-2	0.39	β_0 β_1 β_2 β_5 β_7 3.6 -0.018 0.0021 -0.8 -0.0012 **41. -9.9 2.9 -4.2 -1.6**		
M-3	0.36	β_0 β_2 β_8 β_{10} 3.6 0.0049 -0.026 -0.0013 **53. 5.9 -2.5 -13.**		
M-4	0.49	β_0 β_1 β_3 β_4 β_9 β_{11} 4.3 -0.016 -0.02 0.2 0.0033 -0.055 **35. -8.4 -8.5 3.9 1.5 -7.9**		
M-5	0.5	β_0 β_2 β_3 β_6 1.4 0.00058 -0.02 0.29 **11. 0.88 -8.5 14.**		
M-6	0.72	β_0 β_1 β_3 β_9 β_{11} β_{13} 4.3 -0.01 -0.0018 0.0043 -0.043 -0.037 **49. -6.8 -0.92 2.6 -8.4 -21.**		

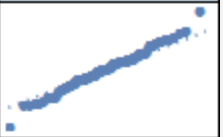
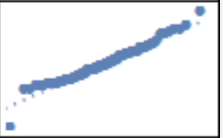
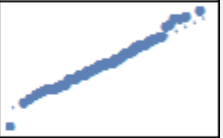
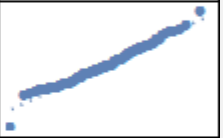
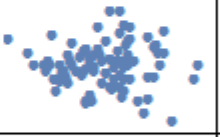
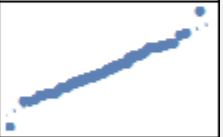
Example 4. — * Data from 93 cars on sale in the USA in 1993. Data from 93 cars, selected at random, on sale in the US in 1993 with 27 variables. Source: Lock, R. H. (1993) 1993 New Car Data. Journal of Statistics Education 1(1).

X_i	Description
1	Minimum Price (in $1,000): price for a basic version
2	Price (in $1,000): average of Min.Price and Max.Price.
3	Maximum Price (in $1,000): price for 'a premium version'
4	City MPG (miles per US gallon by EPA rating).
5	Highway MPG.
6	Number of cylinders (missing for Mazda RX-7, which has a rotary engine).
7	Engine size (litres).
8	Horsepower (maximum).
9	RPM (revs per minute at maximum horsepower)
10	Engine revolutions per mile (in highest gear).
11	Fuel tank capacity (US gallons).
12	Passenger capacity (persons).
13	Length (inches).
14	Wheelbase (inches).
15	Width (inches).
16	U-turn space (feet).
17	"Rear" seat room (inches) (missing for 2-seater vehicles).
18	Luggage capacity (cubic feet) (missing for vans).
19	Weight (pounds).

The models below are the output of an automated model selection procedure, for the response variable Y = the (log of) price of the basic car, and for models that only consider the **main effects**. The parameter table gives the parameter values in the second row and the T-statistic in the third row. You can ignore the column labeled "BIC".

a) Explain how the parameter values should be used to assess the adequacy of a model; give an example.
b) decide which model is the "best" and carefully explain why.
c) decide which model is the "worst" and carefully explain why.

Y = Log of Minimum Price (in $1,000): price for a basic version

Model	Adj-R^2	BIC	Parameter Table	Residuals	Q-Q plot
M-1	0.76	19.	β_0 β_5 β_8 β_{13} β_{15} β_{16} 2.9 -0.025 0.0058 0.014 -0.035 -0.011 **4. -3.8 9. 4.6 -2.4 -0.79**		
M-2	0.75	19.	β_0 β_8 β_{12} β_{15} β_{19} 3.2 0.0039 -0.068 -0.04 0.00066 **4.5 3.9 -1.7 -2.9 5.1**		
M-3	0.73	21.	β_0 β_4 β_8 2.6 -0.028 0.0055 **12. -4.5 8.1**		
M-4	0.75	20.	β_0 β_7 β_8 β_{15} β_{19} 3.9 0.071 0.0047 -0.051 0.0005 **4.3 1.3 6.3 -3.2 4.9**		
M-5	0.77	17.	β_0 β_4 β_8 β_{11} β_{14} β_{15} 2.6 -0.022 0.006 0.01 0.024 -0.043 **3.4 -2.8 8.5 0.62 3.7 -3.3**		
M-6	0.77	15.	β_0 β_8 β_{11} β_{13} β_{15} β_{19} 3. 0.0052 0.022 0.01 -0.061 0.00032 **4.3 7.5 1.3 3.2 -4.2 2.6**		

> More examples like this here.

Solution:

a) Explain how the parameter values should be used to assess the adequacy of a model; give an example.

What is important about the parameter value is its sign, in that it has to make intuitive sense. For example in model M-3 the parameter β_8, which is the parameter for horsepower, is positive. This makes sense because the price of a car should increase with its horsepower.

b) decide which model is the “best" and carefully explain why.

The residuals and future plots for all models are very similar and not particularly faulty, so the basic assumptions are met fairly well by all models.
By the principle parsimony, the best model is the one with the fewest number of parameters, all of them significant (good t-test) and a good Adj-R^2. I would choose M-3, because the closest competitor, M-7, has only a slightly better Adj-R^2 but at the high price of one additional variable.

c) decide which model is the “worst" and carefully explain why.

I would choose model M-8 because it has six parameters, two of them not significantly different from zero. Model M-9 is very similar but has slightly better T-statistics.

□

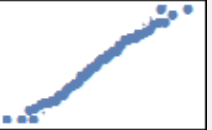
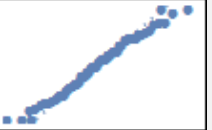
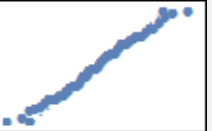
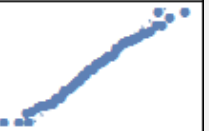
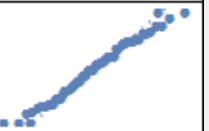
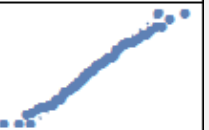
Example 5. — * Climate change This data series from 1959 - 2016 includes the annual global mean surface temperature (Temp) and two possible explanatory variables, the year and the annual average fraction of CO2 contained in the earth' s atmosphere (CO2). Source

	Description
X_1	Year
X_2	CO2 atmospheric composition is defined as the number of molecules of carbon dioxide divided by the number of molecules of dry air multiplied by one million (ppm).
Y	The annual average Temperature is measured in units of 1/100 of a degree centigrade increase above the 1950-1980 mean, often referred to as the global surface temperature anomaly.

The models below are the output of an automated model selection procedure, for the response variable Y. The parameter table gives the factor that the parameter multiplies in the first row ("1." means intercept), the parameter values in the second row and the T-statistic in the third row. You can ignore the column labeled "BIC".

In the models below, decide which model is the best and explain why.

Y = Temp

Model	Adj-R^2	BIC	Parameter Table	Residuals	Q-Q plot
M-1	0.92	-107.	1. / x_1 / x_2 / $x_1 x_2$ -158. / 0.068 / 0.92 / -0.00043 **-3.6** / **3.4** / **3.9** / **-3.8**		
M-2	0.92	-107.	1. / x_2 / x_1^2 / $x_1 x_2$ -94. / 0.94 / 0.00002 / -0.00044 **-3.7** / **3.9** / **3.4** / **-3.8**		
M-3	0.92	-107.	1. / x_2 / x_2^2 / $x_1 x_2$ -33. / 0.54 / -0.0001 / -0.00021 **-4.1** / **3.9** / **-3.2** / **-3.8**		
M-4	0.92	-103.	1. / x_2 / x_1^2 / x_2^2 / $x_1 x_2$ -118. / 1.1 / 0.00003 / 0.00005 / -0.00053 **-1.2** / **1.7** / **0.89** / **0.27** / **-1.4**		
M-5	0.92	-103.	1. / x_1 / x_2 / x_2^2 / $x_1 x_2$ -197. / 0.09 / 1. / 0.00004 / -0.0005 **-1.1** / **0.89** / **1.8** / **0.22** / **-1.5**		
M-6	0.92	-103.	1. / x_1 / x_2 / x_1^2 / $x_1 x_2$ -390. / 0.31 / 0.86 / -0.00007 / -0.0004 **-0.28** / **0.21** / **2.** / **-0.17** / **-1.9**		
M-7	0.92	-103.	1. / x_1 / x_2 / x_1^2 / x_2^2		

Solution: Solved in class. □

7.8 Making predictions

Once the coefficients have been estimated and the assumptions verified, the fitted equation can be used to obtain predictions for Y for any given values $\mathbf{x_0} = (1, x_{01}, \ldots, x_{0p})$ of the explanatory variables $\mathbf{x} = (1, x_1, \ldots, x_p)$. There are two type of predictions that we can do: prediction of the mean response and prediction of a particular realization.

7.8.1 Prediction of the mean response at $\mathbf{x_0}$, $\mathbf{E}(Y \mid \mathbf{x_0}) = \mathbf{x_0\beta}$.

The **point estimate** of $\mathrm{E}(Y \mid \mathbf{x_0}) = \mathbf{x_0\beta}$ is:

$$\begin{aligned}\hat{Y}_0 &= \mathbf{x_0}\hat{\boldsymbol{\beta}}, \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_p x_{0p}\end{aligned}$$

and is an unbiased estimator of $\mathbf{x_0\beta}$. For the **confidence interval** for $\mathrm{E}(Y \mid \mathbf{x_0}) = \mathbf{x_0\beta}$, we note that $\hat{Y}_0$ is a linear combination of the random vector $\hat{\boldsymbol{\beta}}$ and therefore must be normally distributed with

$$\mathrm{E}(\hat{Y}_0) = \mathbf{x_0\beta}, \quad \mathrm{V}(\hat{Y}_0) = \mathbf{x_0\Sigma x_0}^T = \sigma^2 \mathbf{x_0 C x_0}^T.$$

Thus,

$$T = \frac{\mathbf{x_0}\hat{\boldsymbol{\beta}} - \mathbf{x_0\beta}}{\hat{\sigma}\sqrt{\mathbf{x_0 C x_0}^T}} \sim t_{n-p-1}$$

and an inequality can be constructed and re-arranged for $\mathbf{x_0\beta}$ in the usual way:

$(1-\alpha)\%$ **Confidence interval for** $\mathbf{E}(Y \mid \mathbf{x_0})$**:**

$$\mathbf{x_0}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\ \hat{\sigma}\sqrt{\mathbf{x_0}\mathbf{C}\mathbf{x_0}^T}$$

7.8.2 Prediction of a particular realization of $Y_0 = \mathbf{x_0}\boldsymbol{\beta} + \varepsilon_0$

The **point estimate** of $Y_0 = \mathbf{x_0}\boldsymbol{\beta} + \varepsilon_0$, with $\varepsilon_0 \sim N(0, \sigma^2)$ is also:

$$\hat{Y}_0 = \mathbf{x_0}\hat{\boldsymbol{\beta}}$$

and is an unbiased estimator of Y_0. For the **confidence interval**, we know that $Y_0 \sim N(\mathbf{x_0}\boldsymbol{\beta}, \sigma^2)$, and therefore $Y_0 - \hat{Y}_0$ has a normal distribution with mean

$$\mathrm{E}\left(Y_0 - \hat{Y}_0\right) = \mathbf{x_0}\boldsymbol{\beta} - \mathbf{x_0}\boldsymbol{\beta} = 0$$

$$\begin{aligned} \mathrm{V}\left(Y_0 - \hat{Y}_0\right) &= \mathrm{V}(y_0) + \mathrm{V}(\mathbf{x_0}\hat{\boldsymbol{\beta}}) \\ &= \sigma^2 + \sigma^2 \mathbf{x_0}\mathbf{C}\mathbf{x_0}^T \\ &= \sigma^2(1 + \mathbf{x_0}\mathbf{C}\mathbf{x_0}^T). \end{aligned}$$

since Y_0 and $\hat{\boldsymbol{\beta}}$ are independent. Then,

$$T = \frac{Y_0 - \mathbf{x_0}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{1 + \mathbf{x_0}\mathbf{C}\mathbf{x_0}^T}} \sim t_{n-p-1}. \tag{7.13}$$

$(1-\alpha)\%$ **Confidence interval for a particular realization of** $Y \mid \mathbf{x_0}$**:**

$$\mathbf{x_0}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1+\mathbf{x_0}\mathbf{C}\mathbf{x_0}^T}$$

7.9 Simple linear regression

In this case:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}, \quad \mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$\mathbf{C} = \frac{\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}}{n\sum(x_i - \bar{x})^2}$$

It will be convenient to define :

$$\begin{aligned} S_{xy} &= \sum(x_i - \bar{x})(Y_i - \bar{Y}) \\ &= \sum x_i Y_i - n\bar{x}\bar{Y} \\ &\approx n\mathrm{Cov}(X, Y) \end{aligned} \quad \left\| \quad \begin{aligned} S_{xx} &= \sum(x_i - \bar{x})^2 \\ &= \sum x_i^2 - n\bar{x}^2 \\ &= (n-1)S_X^2 \end{aligned} \quad \right\| \quad \begin{aligned} S_{yy} &= \sum(Y_i - \bar{Y})^2 \\ &= \sum Y_i^2 - n\bar{Y}^2 \\ &= (n-1)S_Y^2 \end{aligned}$$

where $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ and, $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ are the unbiased variance estimators of previous chapters.

so that

$$\mathbf{C} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \mathbf{C}\mathbf{X}^T\mathbf{Y} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i \\ -\sum x_i \sum Y_i + n\sum x_i Y_i \end{pmatrix} = \frac{1}{nS_{xx}} \begin{pmatrix} n\bar{Y}\sum x_i^2 - n\bar{x}\sum x_i Y_i \\ nS_{yy} \end{pmatrix}.$$

Fact 7.17 The OLS expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Fact 7.18 — The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\mathrm{V}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}^2}{S_{xx}}, \quad \mathrm{V}\left(\hat{\beta}_0\right) = \hat{\sigma}^2 \frac{\sum x_i^2}{nS_{xx}}, \quad \text{and} \quad \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\hat{\sigma}^2 \frac{\sum x_i}{nS_{xx}}$$

where $\hat{\sigma}^2 = \frac{SSE}{n-2}$ and $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$

$\rightarrow$ Since S_{xx} is proportional to $\mathrm{V}(X)$, a more precise estimate the slope is obtained for x-values that are **more spread out**.

Significance test for β_1 : $H_0 : \beta_1 = 0$ against $H_1 : \beta_i \neq 0$

$\rightarrow$ Reject H_0 if

$$|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}}| > t_{\alpha/2}.$$

This test is important because if we can't reject H_0 it means that the variable X_i does not help explain Y and therefore should be removed from model.

The regression line passes by $(\bar{x}, \bar{Y})$

$$\begin{aligned} \mathrm{E}(Y_i) &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\ &= \bar{Y} + \hat{\beta}_1 (x_i - \bar{x}) \end{aligned}$$

We can also estimate β_0 and β_1 without matrices by differentiating

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

and solving:

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{\partial SSE}{\partial \beta_1} = -2 x_i \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

for β_0 and β_1. We would get the same result.

SSE, R^2

$$\begin{aligned} SSE &= S_{yy} - \hat{\beta}_1 S_{xy} \\ &= S_{yy} - S_{xy}^2/S_{xx}. \quad \underline{(proof)} \end{aligned}$$

Since $SST = S_{yy}$ and $SSE = S_{yy} - S_{xy}^2/S_{xx}$, we can see that

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

$R \approx \rho_{X,Y}$, the correlation between X and Y. Recall

$$\begin{aligned}
\rho(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{V}(X)\,\mathrm{V}(Y)}} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])\,(Y-\mathrm{E}[Y])\right]}{\sqrt{\mathrm{E}\left[(X-\mathrm{E}[X])^2\right]\mathrm{E}\left[(Y-\mathrm{E}[Y])^2\right]}} \\
&\approx \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2\frac{1}{n}\sum(y_i-\bar{y})^2}} \\
&= \frac{\frac{1}{n}S_{xy}}{\sqrt{\frac{1}{n}S_{xx}\frac{1}{n}S_{yy}}} \\
&= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = R
\end{aligned}$$

Connection between the slope and the correlation coefficient:

$$\hat{\beta}_1 = \frac{S_Y}{S_X}R.$$

$\rightarrow$ Since $\hat{\rho} = R$, the test $H_0 : \beta_1 = 0$ is similar to $H_0 : \rho = 0$.
To see this,

$$\begin{aligned} R^2 &= \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{S_{xy}^2}{S_{xx}^2}\frac{S_{xx}}{S_{yy}} = \hat{\beta}_1^2\frac{S_{xx}}{S_{yy}} \\ &= \hat{\beta}_1^2\frac{S_X^2}{S_Y^2}. \end{aligned}$$

and the result follows.

7.9.1 Predictions

Suppose we wish to predict $E(Y)$ at the point x_0. Then $\mathbf{x_0} = (1, x_0)$ so

$$\mathrm{V}\left(\hat{Y}_0\right) = \hat{\sigma}^2 \mathbf{x_0} \mathbf{C} \mathbf{x_0}^T = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

And we obtain:

$(1-\alpha)\%$ **Confidence interval for** $E(Y \mid \mathbf{x_0})$**:**

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2}\ \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$(1-\alpha)\%$ **Confidence interval for a particular realization of** Y**:**

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2}\ \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

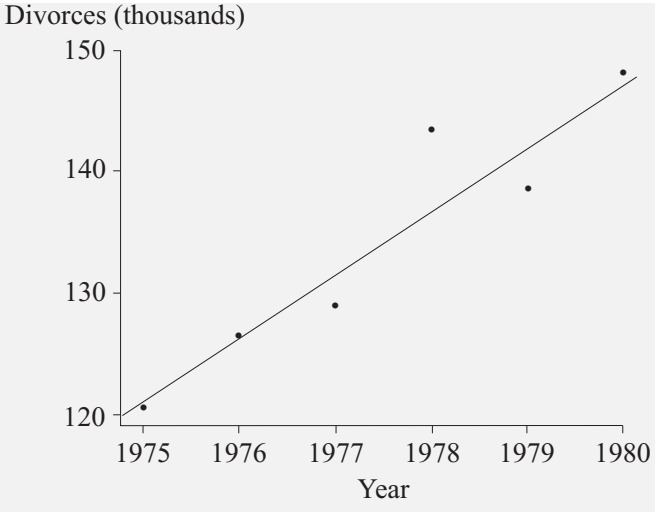
Example 6. — Divorces in England and Wales The table below gives the data and summary for:

Y = the annual number of divorces recorded in England and Wales between 1975 and 1980.
x = years since 1974 ($x = 1$ means year 1975).

							Total
x_i	1	2	3	4	5	6	21
y_i	120.5	126.7	129.1	143.7	138.7	148.3	807.
$x_i y_i$	120.5	253.4	387.3	574.8	693.5	889.8	2919.3
x_i^2	1	4	9	16	25	36	91
y_i^2	14520.3	16052.9	16666.8	20649.7	19237.7	21992.9	109120.

We therefore obtain

n	$\bar{x}$	$\bar{y}$	S_{xy}	S_{xx}	S_{yy}	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\sigma}^2$	R^2
6.	3.5	134.5	94.8	17.5	578.72	5.417	115.54	16.294	0.887



Divorces in England and Wales with fitted line

Our estimate of the rate of increase of divorces is $\hat{\beta}_1 = 5.417$ and we would like to answer the

question "Is the divorce rate changing?" In other words, we would like to test the null hypothesis

$$H_0 : \beta_1 = 0$$

Under H_0

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \sim t(4)$$

giving

$$t = \frac{5.417}{\sqrt{16.294/17.5}} = 5.614.$$

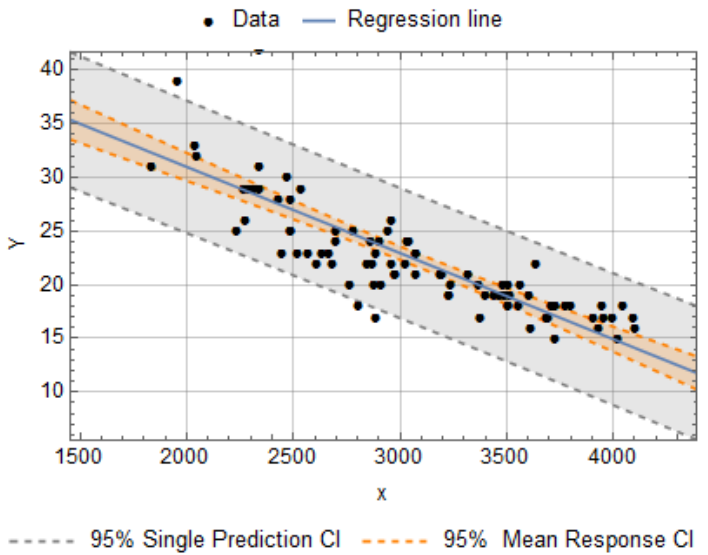
We therefore have strong evidence to reject the null hypothesis that the divorce rate is not changing; that is, there is strong evidence of an increasing divorce rate.

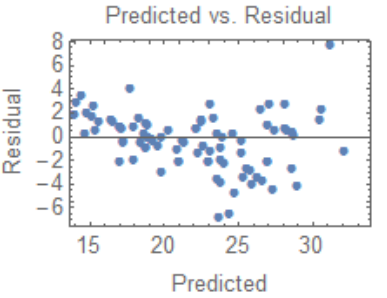
Example 7. — **US Cars** :

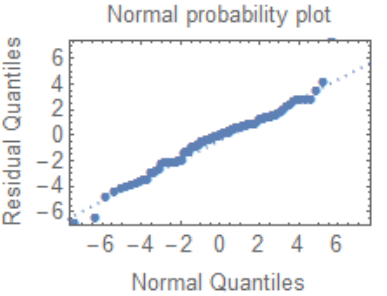
Data from 93 cars on sale in the USA in 1993.

Data from 93 cars, selected at random, on sale in the US in 1993 with 27 variables Source: Lock, R. H. (1993) 1993 New Car Data. Journal of Statistics Education 1(1).

```
Y = City MPG (miles per US gallon by EPA rating).
x = Weight (pounds).
```







Data summary:

> More examples here.

Example 8. You recorded the speed of 7 individual vehicles on a highway segment with posted speed limit of 55 mph, Y, in mph, and the rainfall, X, at the time of each particular measurement, in millimeters per hour, mm/h. The following descriptive statistics were obtained:

$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$	n
125	377.44	5846.55	3012.5	21,580.36	7

(a) Find the linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and interpret the meaning of the estimated parameters in this case.
(b) Conduct a hypothesis test (5% significance level) to determine whether rainfall is a useful linear predictor of vehicle speeds.
(c) Use hypotheses testing to assess whether or not the average speed in this highway exceeds the speed limit by 10 mph on non-rainy days.
(d) Consider the confidence interval for the speed of a particular vehicle. At what level of rainfall is this interval the narrowest? Calculate this interval and interpret its meeting.
(e) Test the hypotheses that the true variance of the regression model is at least 100.
(f) Estimate the probability that an individual driver will be traveling at 10 mph over the speed limit.

Solution:

$$
\begin{aligned}
S_{xx} &= \sum (x_i - \bar{x})^2 \\
&= \sum x_i^2 - n\bar{x}^2 \\
&= 780.36 \\
S_{xy} &= \sum (x_i - \bar{x})(Y_i - \bar{Y}) \\
&= \sum x_i Y_i - n\bar{x}\bar{Y} \\
&= -893.45 \\
S_{yy} &= \sum (Y_i - \bar{Y})^2 \\
&= \sum Y_i^2 - n\bar{Y}^2 \\
&= 1228.80 \\
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
&= -1.1449 \\
\hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x} \\
&= 74.365 \\
SSE &= S_{yy} - \hat{\beta}_1 S_{xy} \\
&= 205.89 \\
\hat{\sigma}^2 &= \frac{SSE}{n-2} \\
&= 41.178
\end{aligned}
$$

(a)

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= 74.365 - 1.1449\hat{x}\end{aligned}$$

In this straight line, which shows the relationship between vehicle speed (Y) and rainfall (X), the intercept $\hat{\beta}_0$ is average speed when a non-rainy day, and the slope $\hat{\beta}_1$ represents the change in vehicle speed due to a unit change in the rainfall.

(b) We test $H_0 : \beta_1 = 0$ against $H_1 \neq 0$

$$\begin{aligned}|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}}| &= |\frac{-1.1449}{\sqrt{41.178}/\sqrt{780.36}}| \\ &= 4.98 > t_{0.025,5} = 2.5706\end{aligned}$$

so we reject H_0 which means rainfall is a useful linear predictor of vehicle speeds.

(c) We test $H_0 : \beta_0 + \beta_1 x = 65$ against $H_1 : \beta_0 + \beta_1 x > 65$, but on a non-rainy day, $x = 0$, so

we are testing $H_0 : \beta_0 = 65$ against $H_1 : \beta_0 > 65$

$$\begin{aligned} Var(\hat{\beta}_0) &= \hat{\sigma}^2 \frac{\sum x_i^2}{nS_{xx}} \\ &= 41.178 \times \frac{3012.5}{7 \times 780.36} \\ &= 22.709 \\ \frac{\hat{\beta}_0 - 65}{\sqrt{Var(\hat{\beta}_0)}} &= \frac{74.365 - 65}{\sqrt{22.709}} \\ &= 1.9652 < t_{0.05,5} = 2.0150 \end{aligned}$$

so we accept H_0, which means vehicle speeds do not exceed the speed limit by 10 mph on non-rainy days.

(d) The confidence interval is narrowest when rainfall level equals its average value: $x_0 = \bar{x} = 17.86$: the 95% confidence interval is

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2,n-2}\ \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} &= 74.365 - 1.1449 \times 17.86 \pm 2.0150 \times 6.85 \\ &= 53.92 \pm 13.82 \\ &= (40.10, 67.74) \end{aligned}$$

so the confidence interval for vehicle speed when rainfall level is 17.86 mm/h is (40.10, 67.74) mph. As with any confidence interval, the interpretation is that the method we use to compute this interval will contain a realization of a particular vehicle speed 95% of the time; that is, if the experiment of taking a sample and computing this interval is repeated *many* times, then on average 95% of those intervals will contain a realization of a particular vehicle speed when $x_0 = \bar{x}$.
We *cannot* say: "This means, when rainfall level is 17.86 mm/h, the probability that vehicle speed is between 40.10 and 67.74 mph is 0.95."
(e)

$$\begin{aligned} H_0 &: \sigma^2 = 100 \\ H_1 &: \sigma^2 > 100 \\ C^2 &= \frac{(7-2)\hat{\sigma}^2}{100} \sim \chi^2_{n-2} \\ &= 2.0589 \end{aligned}$$

Because $2.0589 < \chi^2_{df=5,\alpha=0.05} = 11.07$, we cannot reject H_0.
(f) Since rainfall is not specified, we assume $x_0 = \bar{x}$:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ &= 74.365 - 1.1449 \times 17.86 = 53.92 \end{aligned}$$

Since Y is normally distributed with this mean and variance: $\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} = 6.85$, we get

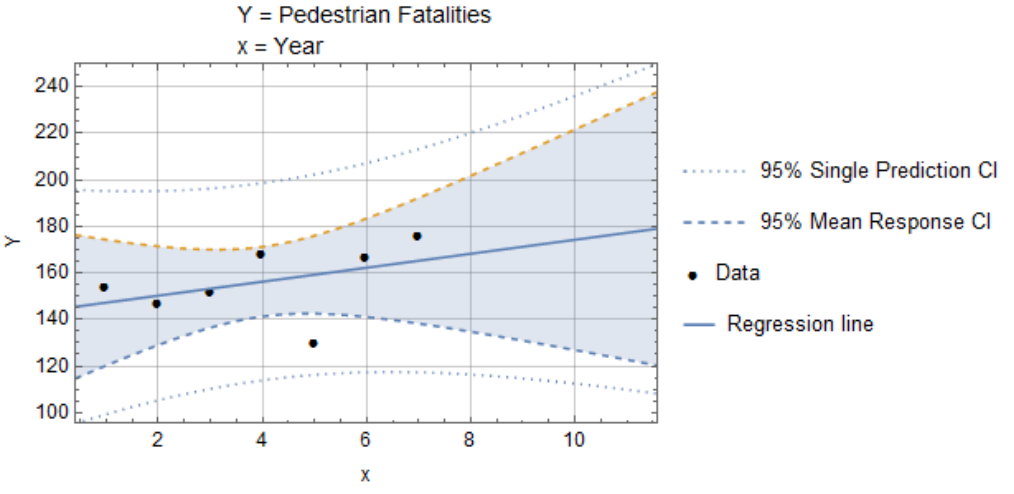
$$\begin{aligned} P(Y > 75) &= 1 - \Phi(\frac{75-53.92}{6.85}) \\ &= 0.053 \end{aligned}$$

and the probability that an individual driver will be traveling at 10 mph over the speed limit is 5.3%.

□

7.10 Problems

Problem 7.1 — Pedestrian Fatalities in Georgia. The number of yearly pedestrian fatalities in the state of Georgia (Y) is presented below for the years 2007-2013 ($x = 1$ means year 2007). Source. Given the data summary, answer the following questions.



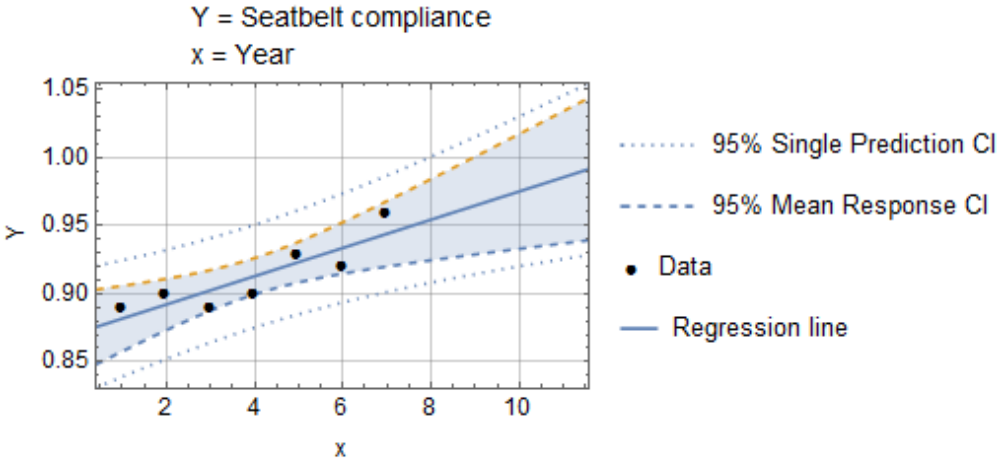
Data summary:

n	x̄	ȳ	$\Sigma x_i y_i$	Σx_i^2	Σy_i^2
7.	4.	156.286	4460.	140.	172418.

a) Find the linear regression model $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and interpret the meaning of the estimated parameters in this case.

b) Calculate $\hat{\sigma}^2$ and R^2 and interpret their meaning.
c) Would you say there is statistical evidence suggesting an increasing trend in pedestrian fatalities?
d) Use the model to test the hypotheses that the expected number of pedestrian fatalities in 2017 (x=11) will be less than 120.
e) Use the model to test the hypotheses that the true variance of the regression model is less than 200.

Problem 7.2 — Seatbelt compliance in Georgia. Data for seatbelt compliance is presented below for the years 2007-2013 ($x = 1$ means year 2007). Source. Given the data summary, answer the following questions.



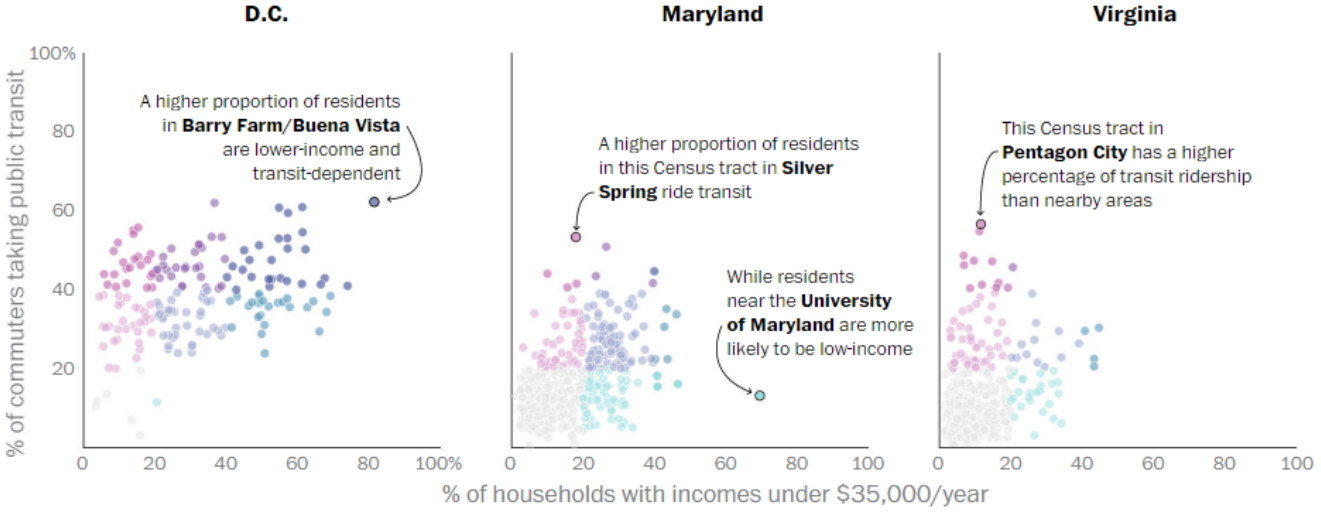
Data summary:

n	x	y	$\Sigma x_i y_i$	Σx_i^2	Σy_i^2
7.	4.	0.913	25.85	140.	5.837

a) Find the linear regression model for seatbelt compliance $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and interpret the meaning of the estimated parameters in this case.

b) Calculate $\hat{\sigma}^2$ and R^2 and interpret their meaning.
c) Would you say there is statistical evidence suggesting an increasing trend in seatbelt compliance ?
d) Use the model to test the hypotheses that the expected seatbelt compliance in 2017 (x=11) will be less than 0.95.
e) Use the model to test the hypotheses that the true variance of the regression model is less than 200.

Problem 7.3 — Washington low-income, transit-reliant residents. This article (https://goo.gl/omGFCd) appeared recently on the Washington Post, analyzes the relationship between income and transit usage in the D.C. area. The data can be summarized in the following figures.



The author of the article, however, did not give any statistical foundations to his observations and conclusions, the most prominent of which is that “D.C. has a higher concentration of low-income, transit-reliant residents than nearby counties in Virginia and Maryland."

You are asked to use the DC data and Virginia data to fill this gap by using the statistical techniques that you deem appropriate to verify/disprove the claims in this article. It is expected that you use

at least two techniques and that you compare and comment the results.

Problem 7.4 The figure shows the average global temperature (relative to the year 1921) from 1880 to 2005. Although it may seem obvious from the figure that the temperature is increasing at a higher rate since the 70's, many people believe that such an increase can be explained by random fluctuations. The factor X denotes the number of years since 1879; i.e., 1880 corresponds to $x = 1$.

Sample	$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$	n
1970-2005	3815	9.87	1180	419405	6.78	35
1880-1969	4186	-21.7	-693	255346	9.35	91
1880-2005	8001	-11.83	486	674751	16.13	126

a) Using the whole sample, test the hypotheses that global warming can be explained by statistical fluctuations around a constant mean that does not grow in time.

b) Test the hypotheses that the global warming rate has increased since the 70's (1970-2005) compared to the rate in (1880-1969).

c) Provide a 95% confidence interval prediction for the global temperature for the year where this interval would be the narrowest, and interpret the meaning of this interval.

d) It is believed that vehicle emissions, Z, are proportional to the square of the global temperature due to the increased use of air conditioning. Estimate the probability that emissions in the year 2017 will double the levels observed in year 2000.

Example 9. From the same survey on the previous question, here we fit a simpler model with only one factor:

$$Y_i = \beta_0 + \beta_1 x_i + + \varepsilon_i$$

Based on the model estimation results summarized below,

a) What would you say are the problems of the fitted model? Clearly justify your answer in each case.

b) As you can see from the scatter plot, the 95 percent confidence interval for Y when $x = 40$ is between 5 and 10. How can we interpret this interval?

c) Do you agree with the slope of the regression line? Why?

Housing values in suburbs of Boston.

Home values for 506 Boston suburbs with potential influential factors. Source: Belsley D. A., Kuh, E. and Welsch, R. E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

```
Y = Median value of owner-occupied homes in $1000s
x = Per capita crime rate by town
```