STAT 509: Statistics for Engineers

Chapters 11-12: Linear Regression

Dr. Dewei Wang
Associate Professor
Department of Statistics
University of South Carolina
deweiwang@stat.sc.edu

Chapters 11-12: Linear Regression

Learning Objectives:

1. Use linear regression for building empirical models to engineering and scientific data
2. Understand how the method of least squares is used to estimate the parameters in a linear regression model
3. Test statistical hypotheses and construct confidence intervals on regression model parameters
4. Use the regression model to predict a future observation and to construct an appropriate prediction interval on the future observation
5. Build regression models with polynomial terms
6. Use indicator variables to model categorical regressors
7. Analyze residuals to determine whether the regression model is an adequate fit to the data or whether any underlying assumptions are violated

Introduction

In previous chapters, we used the normal distribution $Y \sim N(\mu, \sigma^2)$ to model data. Suppose Y represents the salary of your first job after undergraduate study, $\mu = 35K$, and $\sigma = 5K$. It is saying that, about 95% students should expect a salary between $[25K, 45K]$. Now let x be your GPA. It is possible to think x could be an important factor determining your salary level. But the above normal model does not account for the factor x.

A simple improvement is to let μ be a linear function of x; i.e.,

$$Y|x \sim N(\mu_x, \sigma^2), \text{ where } \mu_x = \beta_0 + \beta_1 x,$$

or we write (simple linear regression)

$$Y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

If $\beta_0 = 10K$, $\beta_1 = 10K$, $\sigma = 3K$, then $Y|x = 1 \sim N(20K, 3K^2)$, $Y|x = 3 \sim N(40K, 3K^2)$. If x= 3, with $\approx 95\%$ chance, salary will be between $[34K, 46K]$. The ϵ accounts for a **non-deterministic** relationship between Y (salary) and x (GPA).

Deterministic Models

Many problems in engineering and the sciences involve a study or analysis of the relationship between two or more variables. For example, the velocity of water in an open channel is related to the width of the channel, and the displacement of a particle at a certain time is related to its velocity. In this last example, if we let d_0 be the displacement of the particle from the origin at time $x = 0$ and v be the velocity, the displacement at time x is $d_x = d_0 + vx$. This is an example of a **deterministic** linear relationship because (apart from measurement errors) the model predicts displacement perfectly.

Deterministic Models

However, in many situations, the relationship between variables is not deterministic. For example, the electrical energy consumption of a house (y) is related to the size of the house (x, in square feet), but it is unlikely to be a deterministic relationship. Similarly, the fuel usage of an automobile (y) is related to the vehicle weight x, but the relationship is not deterministic. In both of these examples, the value of y cannot be predicted perfectly from knowledge of the corresponding x. It is possible for different automobiles to have different fuel usage even if they weigh the same, and it is possible for different houses to use different amounts of electricity even if they are the same size.

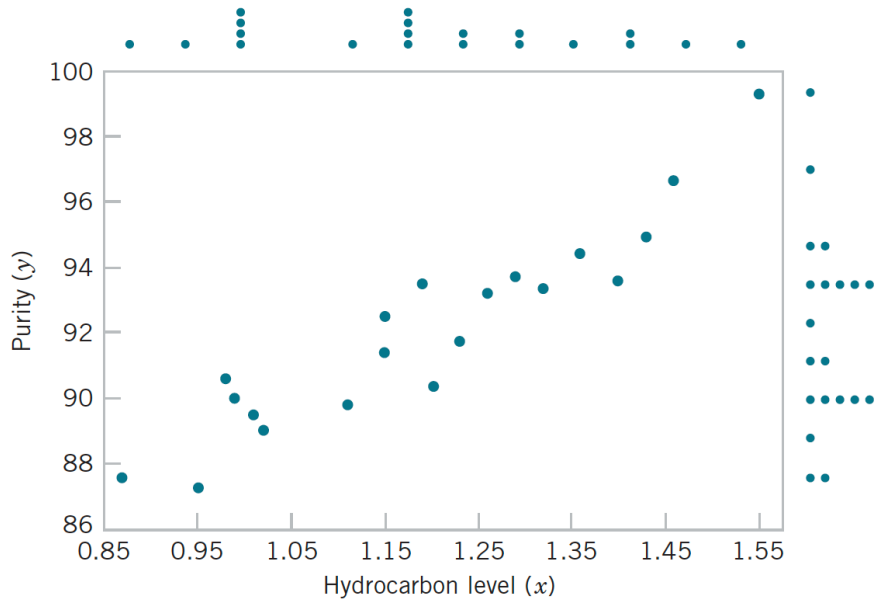
Empirical Models

The collection of statistical tools that are used to model and explore relationships between variables that are related in a **non-deterministic** manner is called **regression analysis**. Because problems of this type occur so frequently in many branches of engineering and science, regression analysis is one of the most widely used statistical tools.

For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature. Regression analysis can be used to build a model to predict yield at a given temperature level. This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes

Example 1

Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



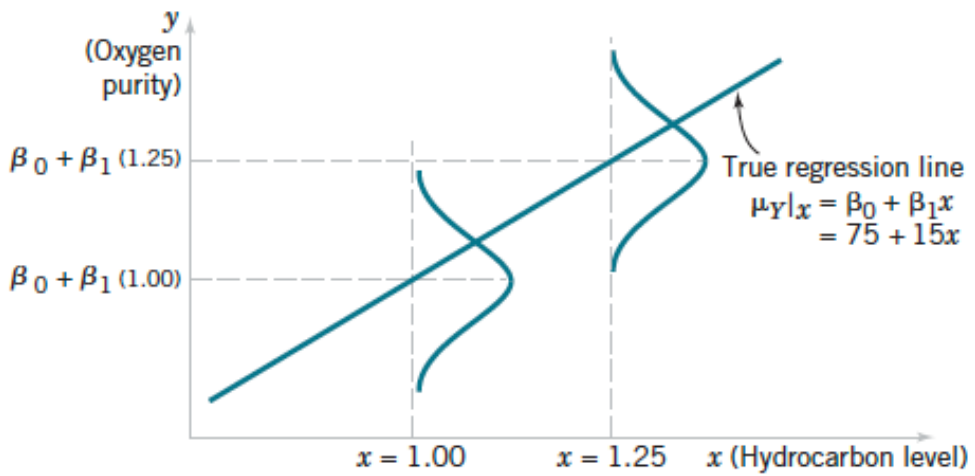
Simple Linear Regression

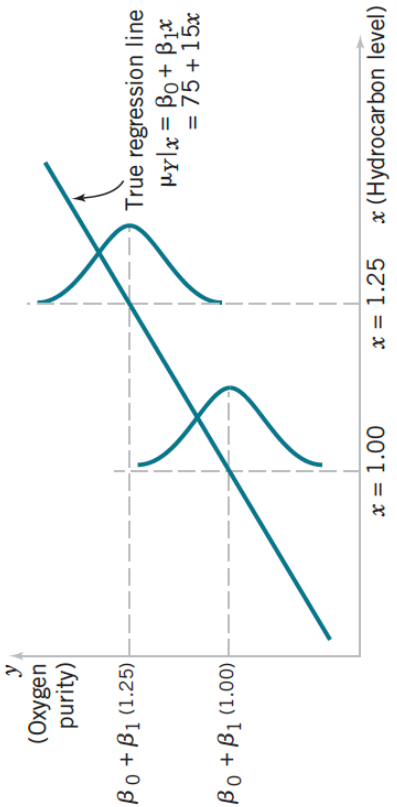
We explain the relationship between $Y = y$ and x by a empirical model (simple linear regression)

$$Y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

The ϵ is a random error term which collects all the variation that cannot be explained by the linear relationship $Y = \beta_0 + \beta_1 x$.

At a fixed x, the distribution of Y is $N(\beta_0 + \beta_1 x, \sigma^2)$.



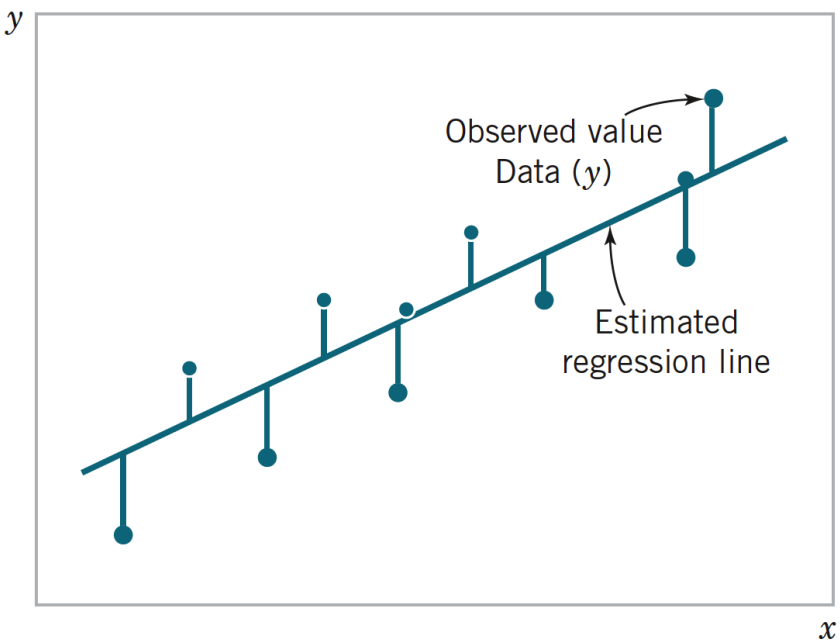


Least Squares Estimates

In simple linear regression, we have

$$Y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

There are three unknown parameters β_0 (intercept), β_1 (slope), and σ^2 (the variance of the random error) we wish to estimate based on n pairs of observations $(x_i, y_i), i = 1, \ldots, n$.
The estimation of β_0 and β_1 comes from the method called Least Squares (by German scientist Karl Gauss, 1777–1855); i.e.,



estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the sum of the squares of the deviations of the observations from the true regression line:

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

(Multiple) Linear Regression

In applications, a response y might depend on more than one factors. For example, the gasoline mileage performance (y) of a vehicle depends on the vehicle weight (x_1) and the engine displacement (x_2). Then we need a little bit more complex linear regression:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

Or maybe even a "nonlinear" structure such as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. If we let $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1 x_2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$, and $\beta_5 = \beta_{12}$, then we can rewrite the nonlinear regression as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ (though it is a nonlinear structure, as long as it can be represented in a linear form, we view it as a linear regression).

Linear Regression and Least SquaresEstimate

The general linear regression takes the form

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

Herein, Y is the response variable, and $x_1, \ldots, x_k$ are k predict variables (or regressor variables), and ϵ is the random error which accounts for all the variation that cannot be explained by the linear structure.

The observations are

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), i = 1, \ldots, n \text{ and } n > k.$$

Each observation satisfies

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, i = 1, \ldots, n. \quad (1)$$

Estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ come from the minimization of the sum of square

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2.$$

Linear Regression and Least SquaresEstimate

If we write things in matrix form:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Then we have (1) written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

and the Least Squaresestimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ has a closed from expression

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Fitted model and residuals

With the Least Squaresestimates, the fitted regression model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

or

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}, i = 1, \ldots, n.$$

And the residuals are

$$e_i = y_i - \hat{y}_i, i = 1, \ldots, n,$$

the difference between observed values and the fitted values. These residuals represent the variation of the observed data that cannot be explained by the linear regression model. We collect these unexplained variantion to a terms called the **error sum of squares**, which is the sum of squares of the residuals:

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Estimation of Variance

Our estimate of the unknown variance σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-(k+1)}.$$

Herein, p denotes the number of regression coefficients $(\beta_0, \ldots, \beta_k)$ we have already estimated; i.e., $p = k+1$.

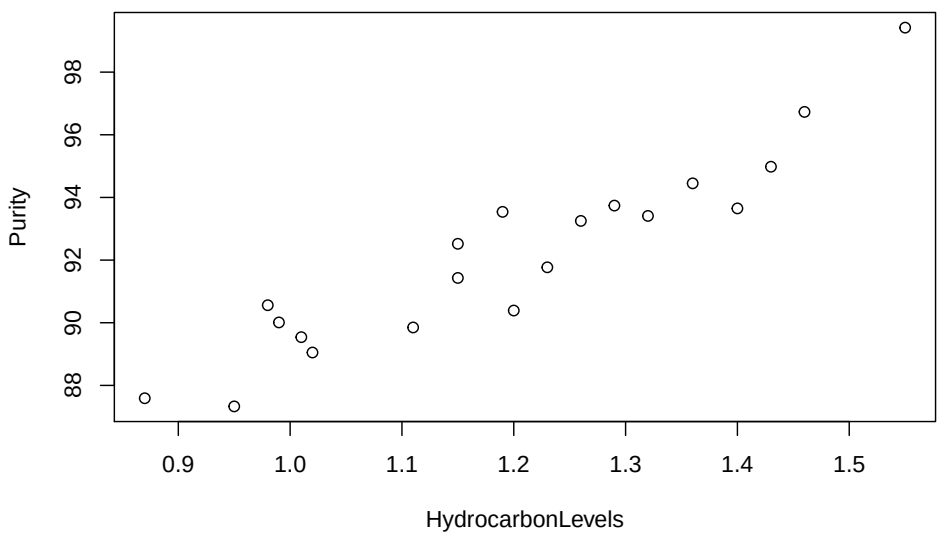
Summary

By now, we have estimated the entire linear regression model which contains $p = k+1$ regression coefficients $(\beta_0, \ldots, \beta_k)$ and the variance of the random error (σ^2). We note that all the estimators are unbiased: $E(\hat{\beta}_j) = \beta_j$ for $j = 0, 1, \ldots, k$, and $E(\hat{\sigma}^2) = \sigma^2$.

Calculation in R

Back to Example 1 on page 7, y is the purity of oxygen produced in a chemical distillation process, and x is the percentage of hydrocarbons present in the main condenser of the distillation unit. The data can be read and fitted as following.

```
> Example1=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
                    master/Data/HydrocarbonPurity.csv")
> head(Example1,2)
  HydrocarbonLevels Purity
1              0.99  90.01
2              1.02  89.05
> plot(Example1)
```



Calculation in R

Back to Example 1 on page 7, y is the purity of oxygen produced in a chemical distillation process, and x is the percentage of hydrocarbons present in the main condenser of the distillation unit. The data can be read and fitted as following.

```
> x=Example1$HydrocarbonLevels # make sure letter cases are matched exactly!
> y=Example1$Purity
> fit=lm(y~x); fit
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      74.28        14.95
> fit$fitted.values # compute the fitted values
> fit$residuals # compute all the residuals
> lm.est(fit)
       Estimate
beta0 74.283314
beta1 14.947480
sigma  1.086529
```

The estimated variance is $\hat{\sigma}^2 = 1.0865^2$. The estimated regression coefficients are $\hat{\beta}_0 = 74.28$ and $\hat{\beta}_1 = 14.95$. The fitted model is

$$\hat{y} = 74.28 + 14.95x_1.$$

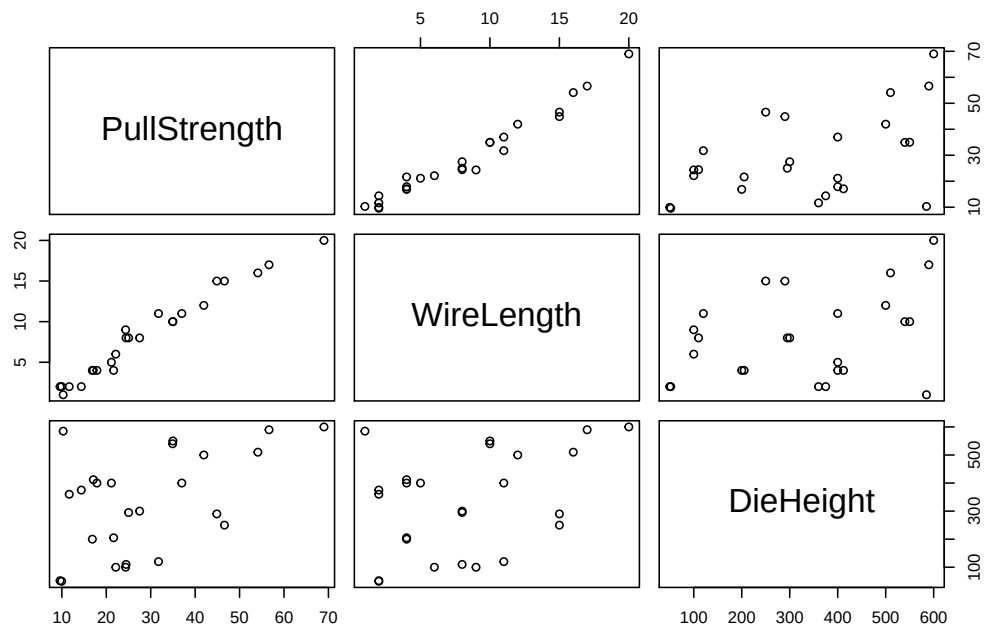
Calculation in R (continued)

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83029 -0.73334  0.04497  0.69969  1.96809

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.283      1.593   46.62  < 2e-16 ***
x             14.947      1.317   11.35 1.23e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared:  0.8774,Adjusted R-squared:  0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

Example 2

we used data on pull strength of a wire bond in a semiconductor manufacturing process, wire length, and die height to illustrate building an empirical model. The data can be loaded as

```
> Example2=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
                    master/Data/WireBond.csv")
> head(Example2,2)
  PullStrength WireLength DieHeight
1         9.95          2        50
2        24.45          8       110

> plot(Example2)
```



Example 2

```
> y=Example2$PullStrength # make sure letter cases are matched exactly!
> x1=Example2$WireLength
> x2=Example2$DieHeight
> fit=lm(y~x1+x2)
> fit
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
    2.26379      2.74427      0.01253

> fit$fitted.values # compute the fitted values
> fit$residuals # compute all the residuals
> lm.est(fit)
        Estimate
beta0 2.26379143
beta1 2.74426964
beta2 0.01252781
sigma 2.28804683
```

The estimated variance is $\hat{\sigma}^2 = 2.288^2$. The estimated regression coefficients are $\hat{\beta}_0 = 2.26379$, $\hat{\beta}_1 = 2.74427$, and $\hat{\beta}_2 = 0.01253$. The fitted model is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2.$$

Example 2 (continued)

```
> summary(fit)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
   Min     1Q Median     3Q    Max
-3.865 -1.542 -0.362  1.196  5.841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.263791   1.060066   2.136 0.044099 *
x1          2.744270   0.093524  29.343  < 2e-16 ***
x2          0.012528   0.002798   4.477 0.000188 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.288 on 22 degrees of freedom
Multiple R-squared:  0.9811,Adjusted R-squared:  0.9794
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

Confidence Intervals in Multiple Linear Regression

Let $\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ which is a $p \times p$ dimensional matrix. Denote by C_{jj} the jth diagnol entry of the matrix $\boldsymbol{C}$. Then the estimated standard error of $\hat{\beta}_j$ is

$$se(\hat{\beta}_j) = \hat{\sigma}\sqrt{C_{jj}}.$$

More importantly, we have

$$T_0 = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{C_{jj}}} \sim t(n-p).$$

Thus a two-tailed $100(1-\alpha)\%$ CI of β_j can be dervied from

$$-t_{n-p,\alpha/2} \leq T_0 = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \leq t_{n-p,\alpha/2}$$

which is

$$\left[\hat{\beta}_j \pm t_{n-p,\alpha/2} se(\hat{\beta}_j)\right].$$

Confidence Intervals in Multiple Linear Regression

A two-tailed $100(1-\alpha)\%$ confidence interval of β_j is $\left[\hat{\beta}_j \pm t_{n-p,\alpha/2} se(\hat{\beta}_j)\right]$.

The $100(1-\alpha)\%$ upper bound of β_j is $\hat{\beta}_j + t_{n-p,\alpha} se(\hat{\beta}_j)$.

The $100(1-\alpha)\%$ lower bound of β_j is $\hat{\beta}_j - t_{n-p,\alpha} se(\hat{\beta}_j)$.

Example 2 continued

Find a 95% two-tailed confidence interval of β_1. We have already fitted a (multiple) linear regression model saved in "fit".

```
> lm.coef.CI(fit,level=0.95)
Two-sided 95 % confidence intervals of regression coefficients are
            CI.lb      CI.ub
beta0 0.065348613 4.46223426
beta1 2.550313061 2.93822623
beta2 0.006724246 0.01833138
One-sided 95 % (lower and upper) confidence bounds are
      lower.bound upper.bound
beta0 0.443504657   4.0840782
beta1 2.583675700   2.9048636
beta2 0.007722522   0.0173331
```

Conclusion: based on the data, we are 95% confident that β_1 is between 2.5503 and 2.9382.

Hypothesis Tests in Multiple Linear Regression

We now consider test

$$H_0 : \beta_j = \beta_{j0} \text{ versus } H_1 : \beta_j \neq \beta_{j0}$$

at significance level α. The test statistic is

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{se(\hat{\beta}_j)}.$$

Reject H_0 if $|T_0| > t_{n-p,\alpha/2}$.

For one-tailed alternatives:

- $H_1 : \beta_j < \beta_{j0}$, reject H_0 if $T_0 < -t_{n-p,\alpha}$.
- $H_1 : \beta_j > \beta_{j0}$, reject H_0 if $T_0 > t_{n-p,\alpha}$.

We also have a confidence interval approach and a P-value approach. These three approaches can be done using StatEngin function `lm.coef.test`

Example 2 (continued)

We now want to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$.

```
> lm.coef.test(fit,alpha=0.05,H1="two")
      P_value t_alpha.2 T0.abs  CI.lb hypo.beta  CI.ub Reject
beta0  0.0441    2.0739 2.1355 0.0653         0 4.4622    Yes
beta1       0    2.0739 29.343 2.5503         0 2.9382    Yes
beta2   2e-04    2.0739 4.4767 0.0067         0 0.0183    Yes
```

This R program tests $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ at $\alpha = 0.05$ for **all** j's **separately**.

Conclusion: at $\alpha = 0.05$, the data provide sufficient evidence to reject $H_0 : \beta_1 = 0$. It means that x_1 is an important predictor variable for the response y.

For one-tailed alternatives:

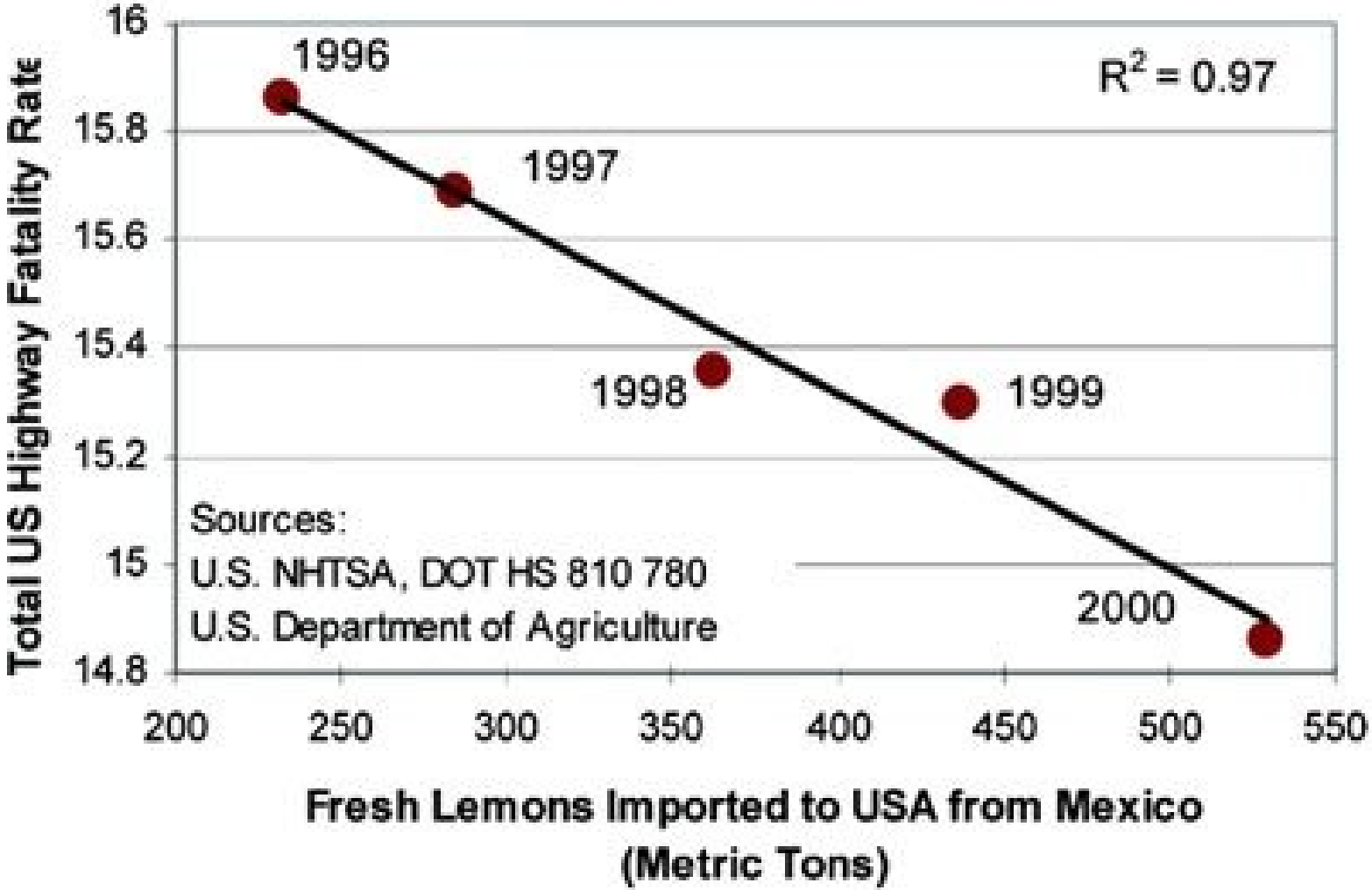
- `lm.coef.test(fit,alpha=0.05,H1="left")`
  tests $H_0 : \beta_j = 0$ versus $H_1 : \beta_j < 0$ at $\alpha = 0.05$ for **all** j's **separately**.
- `lm.coef.test(fit,alpha=0.05,H1="right")`
  tests $H_0 : \beta_j = 0$ versus $H_1 : \beta_j > 0$ at $\alpha = 0.05$ for **all** j's **separately**.

Attention 1

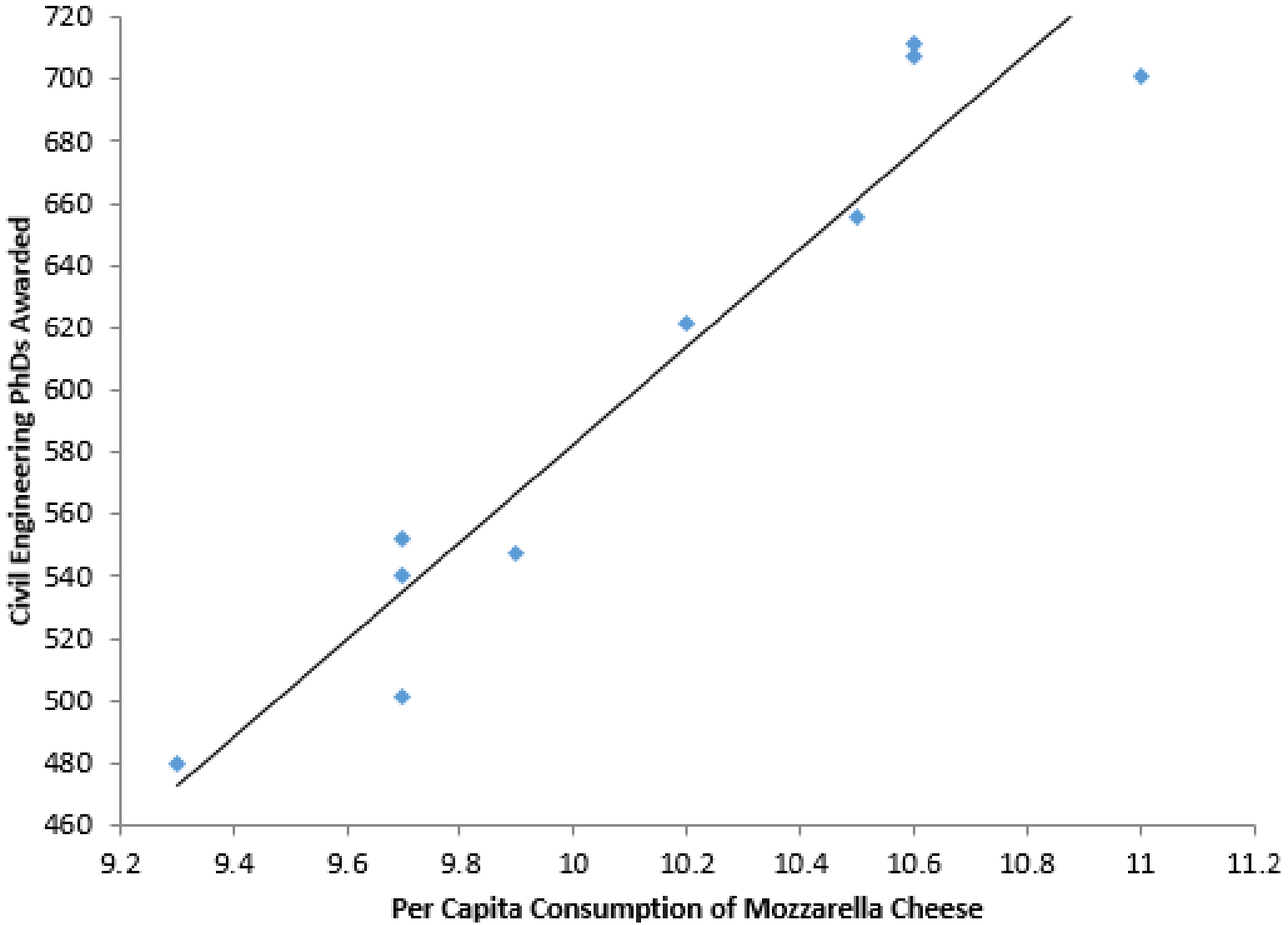
Abuses of Regression Regression is widely used and frequently misused; we mention several common abuses of regression briefly here. Care should be taken in selecting variables with which to construct regression equations and in determining the form of the model. It is possible to develop statistically significant relationships among variables that are completely unrelated in a **causal** sense.

For example, we might attempt to relate the shear strength of spot welds with the number of empty parking spaces in the visitor parking lot. A straight line may even appear to provide a good fit to the data, but the relationship is an unreasonable one on which to rely. We cannot increase the weld strength by blocking off parking spaces. A strong observed association between variables **does not necessarily imply** that a causal relationship exists between them. This type of effect is encountered fairly often in retrospective data analysis and even in **observational studies**. **Designed experiments** are the only way to determine cause-and-effect relationships.

Attention 1


Total US Highway Fatality Rate
16
15.8
15.6
15.4
15.2
15
14.8
1996
1997
1998
1999
2000
$R^2 = 0.97$
Sources:
U.S. NHTSA, DOT HS 810 780
U.S. Department of Agriculture
200
250
300
350
400
450
500
550
Fresh Lemons Imported to USA from Mexico
(Metric Tons)

Attention 1


720
700
680
660
640
620
600
580
560
540
520
500
480
460
Civil Engineering PhDs Awarded
9.2
9.4
9.6
9.8
10
10.2
10.4
10.6
10.8
11
11.2
Per Capita Consumption of Mozzarella Cheese

Example 2 (continued)

We now want to test $H_0 : \beta_1 = 2$ versus $H_1 : \beta_1 \neq 2$ at $\alpha = 0.05$.

```
> lm.coef.test(fit,alpha=0.05,H1="two",hypo.beta=c(0,2,0))
      P_value t_alpha.2 T0.abs  CI.lb hypo.beta  CI.ub Reject
beta0  0.0441    2.0739 2.1355 0.0653         0 4.4622    Yes
beta1       0    2.0739 7.9581 2.5503         2 2.9382    Yes
beta2   2e-04    2.0739 4.4767 0.0067         0 0.0183    Yes
```

Conclusion: at $\alpha = 0.05$ the data provide sufficient evidence to reject $H_0 : \beta_1 = 2$.

We now want to test $H_0 : \beta_0 = 1$ versus $H_1 : \beta_0 < 1$ at $\alpha = 0.05$.

```
lm.coef.test(fit,alpha=0.05,H1="left",hypo.beta=c(1,0,0))
      P_value     T0 t_alpha CI.lb hypo.beta  CI.ub Reject
beta0  0.8771 1.1922 -1.7171  -Inf         1 0.4435     No
beta1       1 29.343 -1.7171  -Inf         0 2.5837     No
beta2  0.9999 4.4767 -1.7171  -Inf         0 0.0077     No
```

Conclusion: at $\alpha = 0.05$ data do not provide sufficient evidence to reject $H_0 : \beta_0 = 1$.

Test $H_0 : \beta_2 = 0.02$ versus $H_1 : \beta_2 > 0.02$ at $\alpha = 0.05$.

```
lm.coef.test(fit,alpha=0.05,H1="right",hypo.beta=c(0,0,0.02))
      P_value t_alpha      T0  CI.lb hypo.beta CI.ub Reject
beta0   0.022  1.7171  2.1355 0.4435         0   Inf    Yes
beta1       0  1.7171  29.343 2.5837         0   Inf    Yes
beta2   0.993  1.7171 -2.6701 0.0077      0.02   Inf     No
```

At $\alpha = 0.05$ data do not provide sufficient evidence to reject $H_0 : \beta_2 = 0.02$.

Prediction of New Observations

Now we have a fitted model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k.$$

Suppose we have a new

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ \vdots \\ x_{0k} \end{bmatrix},$$

we want to predict the future observation of Y at $\boldsymbol{x} = \boldsymbol{x}_0$.

We know that at this $\boldsymbol{x} = \boldsymbol{x}_0$, the response Y has a distribution as

$$Y \sim N(\mu_{Y|\boldsymbol{x}_0}, \sigma^2)$$

where $\mu_{Y|\boldsymbol{x}_0} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k}$.

Prediction of New Observations

We have two types of prediction.

- Prediction of $\mu_{Y|\boldsymbol{x}_0}$: we predict the averaged Y-value at $\boldsymbol{x} = \boldsymbol{x}_0$. For this goal, we have a statistical inference called **Confidence Interval on the Mean Response**.
- Prediction of Y at $\boldsymbol{x} = \boldsymbol{x}_0$: we predict a single observation of Y at $\boldsymbol{x} = \boldsymbol{x}_0$. For this goal, the statistical inference is **Prediction Interval on a Single Future Response**.

Both of them are built on the fitted value $\hat{y}_0 = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}$, but account for different errors.

Prediction of New Observations

We have two types of prediction, where $\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$:

- Confidence Interval on the Mean Response ($\mu_{Y|\boldsymbol{x}_0}$) at $\boldsymbol{x} = \boldsymbol{x}_0$:

$$\hat{y}_0 - t_{n-p,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^\top \boldsymbol{C}\boldsymbol{x}_0} \leq \mu_{Y|\boldsymbol{x}_0} \leq \hat{y}_0 + t_{n-p,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^\top \boldsymbol{C}\boldsymbol{x}_0}$$

- Prediction Interval on a Single Future Response (Y_0) at $\boldsymbol{x} = \boldsymbol{x}_0$:

$$\hat{y}_0 - t_{n-p,\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^\top \boldsymbol{C}\boldsymbol{x}_0} \leq Y_0 \leq \hat{y}_0 + t_{n-p,\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^\top \boldsymbol{C}\boldsymbol{x}_0}$$

The prediction interval is always wider than the confidence interval. The confidence interval expresses the error in estimating the mean of a distribution, and the prediction interval expresses the error in predicting a future observation from the distribution at the point $\boldsymbol{x}_0$. This must include the error in estimating the mean at that point as well as the inherent variability in the random variable Y at the same value $\boldsymbol{x} = \boldsymbol{x}_0$.

Example 2 (continued)

Suppose that the engineer wishes to construct a 95% CI on the mean pull strength for a wire bond with wire length $x_1 = 8$ and the die height is $x_2 = 275$, and a 95% prediction interval on the wire bond pull strength when the wire length is $x_1 = 8$ and the die height is $x_2 = 275$.

Solution: continue to use "fit".

```
> fit
Call:
lm(formula = y ~ x1 + x2)
> predict.lm(fit,new=data.frame(x1=8,x2=275),interval="confidence")
      fit      lwr      upr
1 27.6631 26.66324 28.66296

> predict.lm(fit,new=data.frame(x1=8,x2=275),interval="prediction")
      fit      lwr      upr
1 27.6631 22.81378 32.51241
```

Based on the data, we are 95% confidence that the mean pull strength at $x_1 = 8$ and $x_2 = 275$ is between 26.6632 and 28.663, and the wire bond pull strength at $x_1 = 8$ and $x_2 = 275$ is between 22.8138 and 32.5124.

Attention 2

Regression relationships are valid for values of the regressor variable only within the range of the original data. The linear relationship that we have tentatively assumed may be valid over the original range of x, but it may be unlikely to remain so as we extrapolate—that is, if we use values of x beyond that range. In other words, as we move beyond the range for which data were collected, we become less certain about the validity of the assumed model. Regression models are not necessarily valid for extrapolation purposes.

Now this does not mean do not ever extrapolate. For many problem situations in science and engineering, extrapolation of a regression model is the only way to even approach the problem. However, there is a strong warning to be careful. A modest extrapolation may be perfectly all right in many cases, but **a large extrapolation will almost never** produce acceptable results.

Test for Significance of Regression

Testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ for some j is already one type of significance tests. It investigates whether the predictor x_j is significant to the response y. If H_0 is rejected, then it is; otherwise, x_j might not have a significant role to explain y.

Another significance test is (ANOVA test):

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ v.s. } H_1 : \beta_j \neq 0 \text{ for at least one } j.$$

Rejection of $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ implies that at least one of the regressor variables $x_1, \ldots, x_k$ contributes significantly to the model,

- but we do not know which one(s).
- This significance test is different than testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ for all j's separately and cannot be down using lm.coef.test

Test for Significance of Regression

Test statistic for ANOVA is

$$F_0 = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E}$$

- $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$, the regression sum of squares.
- $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, the error sum of squares
- MS_R and MS_E are called mean squares.

Let $SS_T = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$, the total corrected sum of squares. We have

$$SS_T = SS_R + SS_E.$$

At significance level α, we reject $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ if $F_0 > f_{r,n-p,\alpha}$ (always use P-value for this one).

Note: Rejection of H_0 does not necessarily imply that the linear relationship found is an appropriate model for predicting y as a function of $x_1, \ldots, x_k$. Further tests of model adequacy are required before we can be comfortable using this model in practice.

Example 2 (continued)

We test $H_0 : \beta_1 = \beta_2 = 0$ versus $H_1 : \beta_j \neq 0$ for at least one j at $\alpha = 0.05$.

Solution: Back to `summary(fit)`. Look for

```
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

This result tells that $F_0 = 572.2$, $r = 2$, $n - p = 22$, and the P-value for the ANOVA test is less than 2.2×10^{-16} (thus less than α, thus reject H_0).

Conclusion: at $\alpha = 0.05$, the data provide sufficient evidence to reject H_0.

Practical Interpretation: Rejection of H_0 does not necessarily imply that the relationship found is an appropriate model for predicting pull strength as a function of wire length and die height. Further tests of model adequacy are required before we can be comfortable using this model in practice.

R^2 and Adjusted R^2

We may also use the coefficient of multiple determination R^2 as a global statistic to assess the fit of the model. Computationally,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

In Example 2, back to `summary(fit)`. Look for

```
Multiple R-squared:  0.9811,Adjusted R-squared:  0.9794
```

This tells use $R^2 = 0.9811$.

Interpretation: the model accounts for about 98.11% of the variability in the pull strength response.

Drawback: The R^2 statistic is somewhat problematic as a measure of the quality of the fit for a multiple regression model because it never decreases when a variable is added to a model. Solely based on R^2 to select predictors could cuase **overfitting**. A better statistic to use is the Adjusted R^2.

```
> set.seed(100)
> summary(lm(y~x1+x2+rnorm(25)))
Multiple R-squared:  0.9813,  Adjusted R-squared:  0.9786
```

Adjusted R^2

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}.$$

Because $SS_E/(np)$ is the error or residual mean square and $SS_T/(n1)$ is a constant, R^2_{adj} will only increase when a variable is added to the model if the new variable reduces the error mean square.

In Example 2,

```
> summary(fit)
Multiple R-squared:  0.9811,  Adjusted R-squared:  0.9794
```

If we only include one predictor:

```
> summary(lm(y~x1))
Multiple R-squared:  0.964,  Adjusted R-squared:  0.9624
> summary(lm(y~x2))
Multiple R-squared:  0.2429,  Adjusted R-squared:   0.21
```

Therefore, we would conclude that adding x_2 or x_1 to the model does result in a meaningful reduction in unexplained variability in the response.

Partial F-test

We would like to determine whether a subset of regressor variables, say without loss of generality $x_1, x_2, \ldots, x_r$ $(r \leq k)$, as a whole contributes significantly to the regression model. This leads to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0 \text{ versus}$$
$$H_1 : \text{ at least one of } \beta_1, \ldots, \beta_r \text{ is nonzero}$$

We call this as a general regression test or a partial F-test. The ANOVE test is a special case when $r = k$.

How to do? The test statistic is also a F-statistic, call it F_0 again. Then we reject H_0 at α if $F_0 > f_{r,n-p,\alpha}$. There is also a P-value approach.
StatEngine: First fit a linear regression model using all the rest predictors not invovled in H_0, call the fit as `lmH0`. Then fit a full linear regression using all the predictors. Call the fit as `lmALL`.
`lm.partialFtest(fit.H0=lmH0,fit.ALL=lmALL,alpha=?)`

Example 2 (continued)

Consider the wire bond pull-strength data. We investigate the contribution of two new variables, $x_3 = x_1^2$ and $x_4 = x_2^2$, to the model using the partial F-test approach. That is we wish to test, at $\alpha = 0.05$,

$$H_0 : \beta_3 = \beta_4 = 0 \quad H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0.$$

```
> x3=x1^2
> x4=x2^2
> lmH0=lm(y~x1+x2)
> lmALL=lm(y~x1+x2+x3+x4)
> lm.partialFtest(fit.H0=lmH0,fit.ALL=lmALL,alpha=0.05)
                    F0  f_alpha  P_value Reject
PartialFtest 4.047007 3.492828 0.033432    Yes
```

Conclusion: we see the P-value is less than 0.05, thus conclude that at significance level $\alpha = 0.05$, the data provide sufficient evidence to reject H_0; i.e., at least one of the new variables contributes significantly to the model. Further analysis and tests will be needed to refine the model and determine *whether one or both of* x_3 *and* x_4 *are important*.

Example 2 (continued)

```
> summary(lm(y~x1+x2+x3))
Multiple R-squared:  0.9864,Adjusted R-squared:  0.9845
> summary(lm(y~x1+x2+x4))
Multiple R-squared:  0.9815,Adjusted R-squared:  0.9788
> summary(lm(y~x1+x2+x3+x4))
Multiple R-squared:  0.9866,Adjusted R-squared:  0.9839
```

Based on the Adjusted R^2, we see that the best choice is

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

or equivalently,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \epsilon \text{ (Polynomial Regression)}$$

Polynomial Regression

When data present a non-linear pattern, we could still use linear regression by adding high-degree polynomial term. For example, the second-degree polynomial regression in one variable is

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon,$$

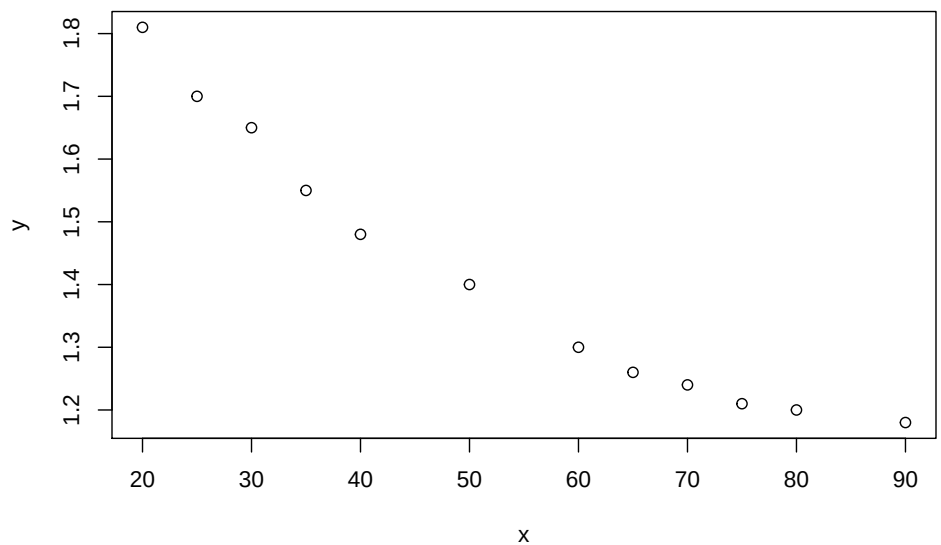
in two variables is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon.$$

Example 3: Airplane Sidewall Panels

Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit manufacturing cost varies with the production lot size. The following data give the average cost per unit (in hundreds of dollars) for this product (y) and the production lot size (x).

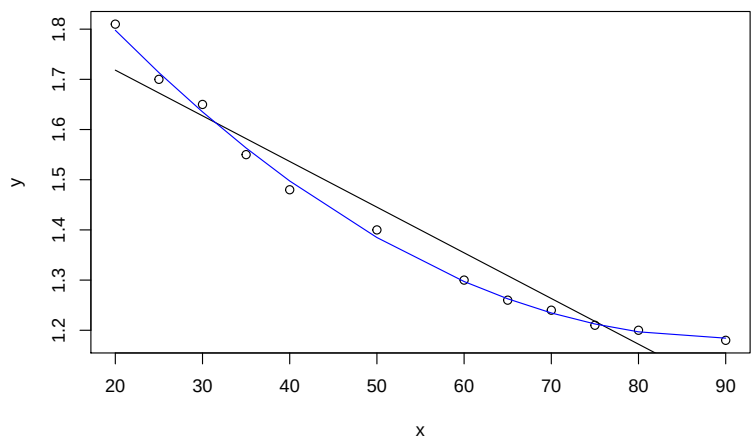
```
> Example3=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
                    master/Data/AirplaneSidewallPanels.csv")
> head(Example3,2)
> y=Example3$cost
> x=Example3$lotsize
> plot(x,y)
```



Example 3: Airplane Sidewall Panels (continued)

We first fit $Y = \beta_0 + \beta_1 x + \epsilon$. Result is the black line. Then we fit a second-order polynomial regression: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. Check the blue curve.

```
> plot(x,y)
> fit=lm(y~x)
> summary(fit)
Multiple R-squared:  0.9386,Adjusted R-squared:  0.9324
> lines(x,fit$fitted.values)
> x2=x^2
> fit.quad=lm(y~x+x2)
> summary(fit.quad)
Multiple R-squared:  0.9975,Adjusted R-squared:  0.9969
> lines(x,fit.quad$fitted.values,col="blue")
```



Categorical Regressors and Indicator Variables

To handle categorical predictor variables, we use indicator variables. For example, to introduce the effect of two different operators into a regression model, we could define an indicator (or dummy) variable as follows:

$$x = \begin{cases} 0 & \text{if the observation is from operator 1} \\ 1 & \text{if the observation is from operator 2} \end{cases}$$

Or more than two categories:

x_1	x_2	
0	0	if the observation is from operator 1
1	0	if the observation is from operator 2
0	1	if the observation is from operator 3

Example 4: Surface Finish

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown as

observation i	Surface Finish y_i	RPM x_{i1}	Type of Cutting Tool
1	45.44	225	302
11	33.50	224	416

x_1 is the lathe speed in revolutions per minute. We use

$$x_2 = \begin{cases} 0 & \text{for tool type 302} \\ 1 & \text{for tool type 416} \end{cases}$$

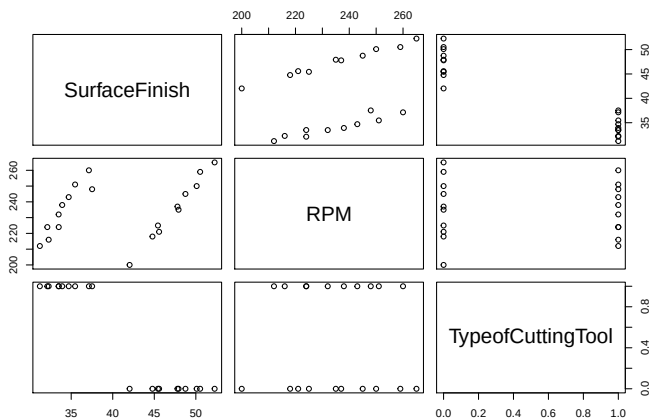
Then build

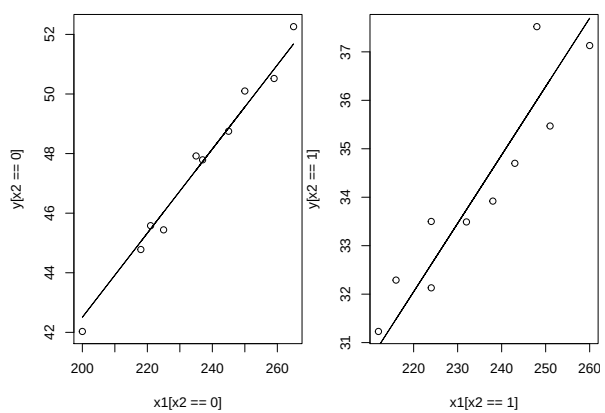
$$Y = \beta_0+\beta_1 x_1+\beta_2 x_2+\epsilon = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{for tool type 302} \\ \beta_0 + \beta_1 x_1 + \beta_2 + \epsilon & \text{for tool type 416} \end{cases},$$

two straight lines with different intercepts to account for the different tool types.

Example 4: Surface Finish

```
> Example4=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
                    master/Data/SurfaceFinishData.csv")
> head(Example4,2)
> plot(Example4)
> y=Example4$SurfaceFinish
> x1=Example4$RPM
> x2=Example4$TypeofCuttingTool
> fit=lm(y~x1+x2)
> summary(fit)
Multiple R-squared:  0.9924,Adjusted R-squared:  0.9915
> par(mfrow=c(1,2))
> plot(x1[x2==0],y[x2==0])
> lines(x1[x2==0],fit$fitted.values[x2==0])
> plot(x1[x2==1],y[x2==1])
> lines(x1[x2==1],fit$fitted.values[x2==1])
```





Example 4: Surface Finish

It is also possible to use indicator variables to investigate whether tool type affects both the slope and intercept. Let the model be Then build

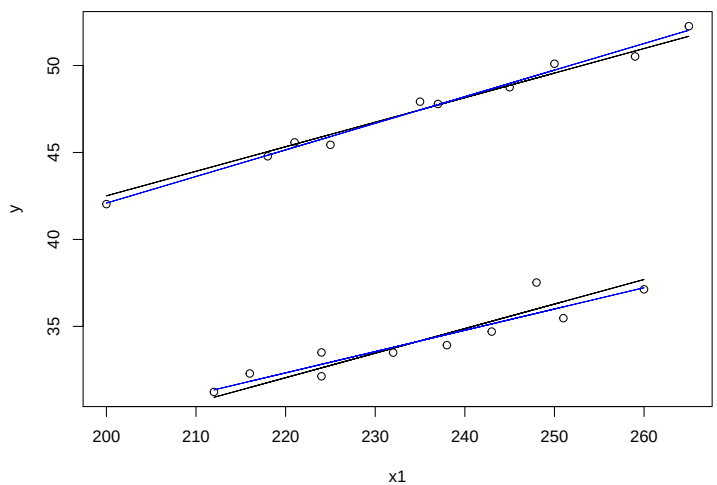
$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \epsilon \\
&= \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{for tool type 302} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon & \text{for tool type 416} \end{cases},
\end{aligned}
$$

two straight lines with different intercepts and slopes.

Example 4: Surface Finish

New model is in blue while the previous one is in black.

```
> x3=x1*x2
> fit2=lm(y~x1+x2+x3)
> summary(fit2)
Multiple R-squared:  0.9936,Adjusted R-squared:  0.9924
> par(mfrow=c(1,1))
> plot(x1,y)
> lines(x1[x2==0],fit$fitted.values[x2==0])
> lines(x1[x2==1],fit$fitted.values[x2==1])
> lines(x1[x2==0],fit2$fitted.values[x2==0],col="blue")
> lines(x1[x2==1],fit2$fitted.values[x2==1],col="blue")
```



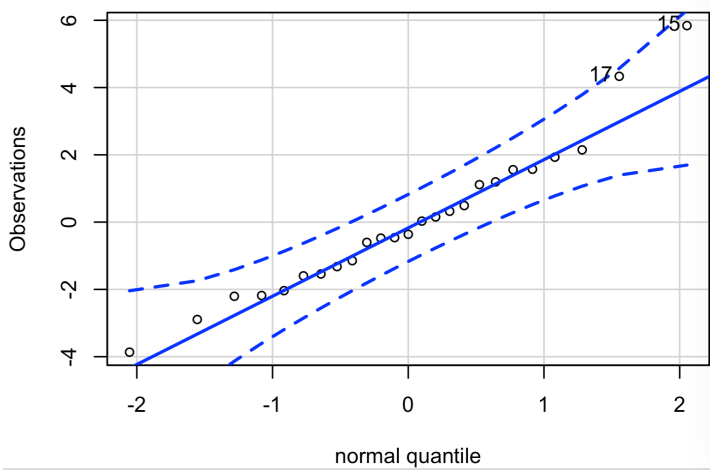
The Adjusted R^2 indicates the new (more complex) model is slightly better.

Model Adequacy Checking

The residuals from the multiple regression model, defined by $e_i = y_i - \hat{y}_i$, play an important role in judging model adequacy.

One model assumption is that $\epsilon \sim N(0, \sigma^2)$. Thus we can check the normality on e_i's by using QQ-plot. Back to Example 2, we have

```
> Example2=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
                     master/Data/WireBond.csv")
> y=Example2$PullStrength;x1=Example2$WireLength;x2=Example2$DieHeight
> fit=lm(y~x1+x2)
> data.summary(fit$residuals)
```



We see the normality seems to be satisfied thought we have two points near the boundary.

Model Adequacy Checking

The **standardized residuals**

$$d_i = \frac{e_i}{\hat{\sigma}}, \text{ for } i = 1, \ldots, n,$$

are often more useful than the ordinary residuals when assessing residual magnitude. If a $|d_i| > 3$, then we can think the ith observation might be an outlier.

It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to be alert to the possibility of **influential observations** and to take them into consideration when interpreting the results. Cook's Distance is a good measure of the influence of an observation (Using `cook.distance`). Let D_i be the Cook's Distance of the ith observation, if $D_i > 1$, then the ith observation might be an influencial observation.

Model Adequacy Checking

To obtain our least squares estimates $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$, it requires the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible. In many times, the predictors could present a very strong **multicollinearity** which makes $\boldsymbol{X}^\top \boldsymbol{X}$ nearly singular and leads to unreliable inferences.

To check whether strong **multicollinearity** exists, we often use **Variance Inflation Factor (VIF)**, which can be done using `vif` function.

Example 2

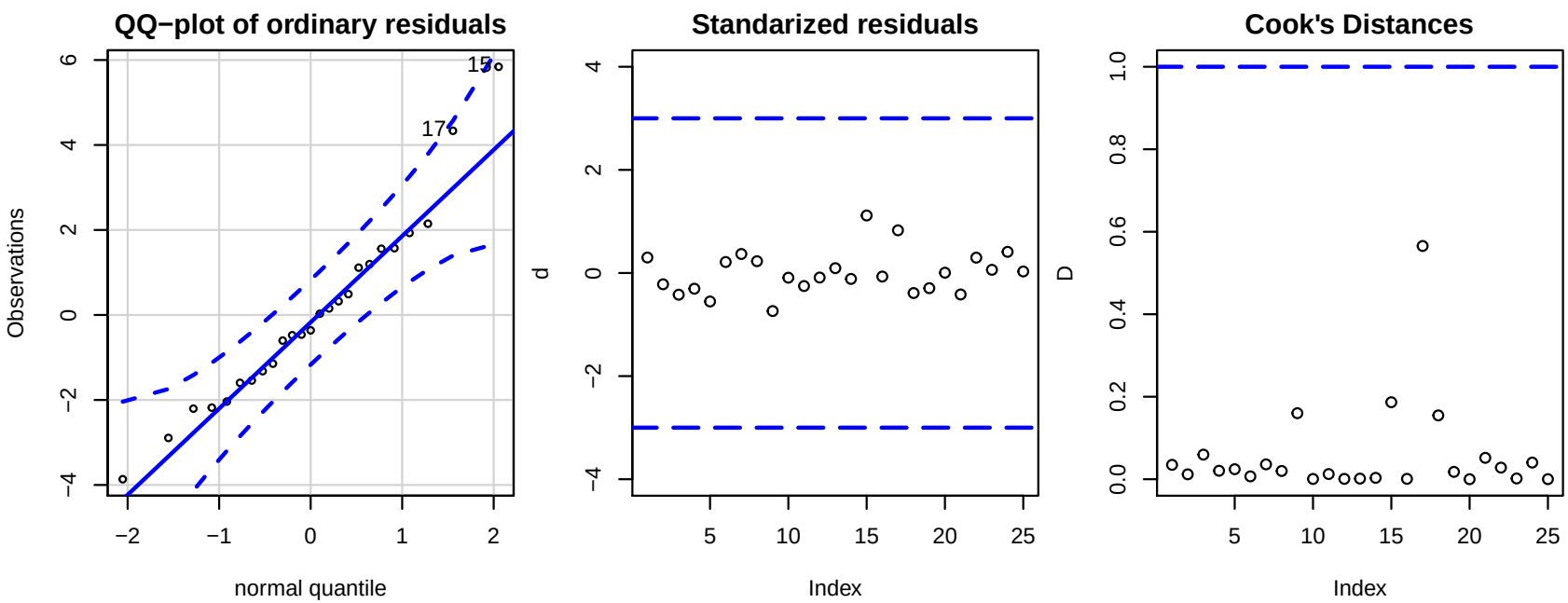
```
> fit=lm(y~x1+x2)
> vif(fit)
      x1       x2
1.167128 1.167128
```

For each predictor, we have a VIF, denoted by $VIF(\hat{\beta}_i)$ for $i = 1, \ldots, k$. A rule of thumb is that if $VIF(\hat{\beta}_i) > 10$ then multicollinearity is high (a cutoff of 5 is also commonly used). In Example 2, the VIFs are all less than 5, thus we do not need to worry about multicollinearity.

Example 2 (continued)

StatEngine has the function `lm.modelcheck` to compute the QQ-plot of ordinary residuals (e_i's), the standardized residuals (d_i's), and the cook's distances (D_i's)

```
> lm.modelcheck(fit)
VIFs are:
      x1       x2
1.167128 1.167128
```



Summary

First step, name y, $x_1, \ldots, x_k$ in R. Then fit the linear regression using `lm` function. Below is a summary of related R functions.

```
summary() # Summary of results (ANOVA test, R-square, and Adjusted R-square)
lm.est() # all the estimates
lm.coef.CI() # Confidence intervals on the regression coefficients
lm.coef.test() # Hypothesis testing on the regression coefficients
predict.lm() # Confidence interval and prediction interval at a new x
lm.partialFtest() # Partial F-test
lm.modelcheck() # Some residual analysis to check model adequacy
```