MA331 Intermediate Statistics

Lecture 09 Linear Regression Analysis [1]

Xiaohu Li

Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, New Jersey 07030

Weeks 11-12

[1] Based on Chapters 3, 10 and 11.

0. Topics to be covered

While ANOVA studies the association between a quantitative variable (response) and categorical variables (factors), Linear Regression focuses on modelling the association between a quantitative variable and other quantitative variables. Topics to be covered include

- Response and covariates
- Simple linear regression model
- Inference for simple linear regression
- Multiple linear regression model
- Inference for multiple linear regression
- ANOVA test

1. Review and outlook

☞ Two-way table analysis checks the association between two categorical r.v.'s, and ANOVA studies the association between a quantitative variable and categorical variables. In practice, some categorical variables (e.g., income level) are actually just discretised versions of quantitative variables (e.g., income).

☞ In applied science and engineering, it is always of interest to study and then model the association between a quantitative r.v., called as response, dependent, output variable, and some other quantitative ones, called as covariates, explanatory, independent, input variables.

☞ Two examples

- Is there a strong correlation between the binge-drinking rate and the average price for bottled beer at establishments within a 2-mile radius of campus?
- What is the relationship between the sale price of the used car and its mileage and age?

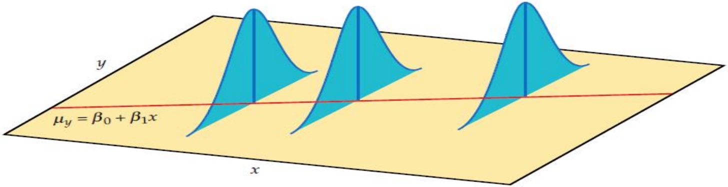
2. Simple linear regression

✍ Simple linear regression models the relationship between a response variable Y and a single explanatory variable X.

- The explanatory variable X can take many different values, and different values of X are expected to produce different mean responses.
- Given $X = x$ the response variable Y is of normal distribution with the mean

$$\mu_Y = \beta_0 + \beta_1 x.$$

- All these normal distributions have the same variance σ^2.



3. Data for simple linear regression

✍ The data for a linear regression are observed pairs (y_i, x_i)'s of (Y, X).

- The model takes each x_i to be a fixed known quantity.
- The response Y corresponding to a given x is random. So, y_i is the observed value of Y given $X = x_i$, $i = 1, \cdots, n$.

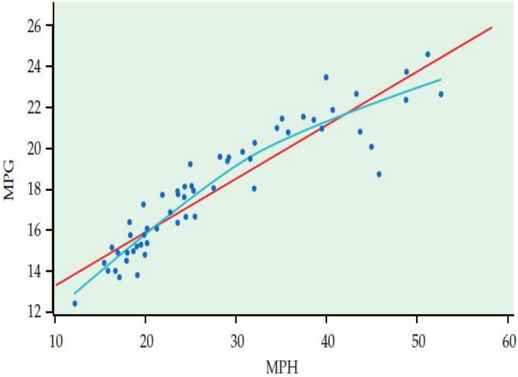
The linear regression model describes the mean and standard deviation of r.v. Y, and these unknown parameters must be estimated from the data.

✍ An example on the relationship between speed driven and fuel efficiency

- Record MPG (miles per gallon) and MPH (miles per hour) each time the gas tank is filled up. Draw a random sample of size 60 from 262 observations.
- We want know how does the speed at which the vehicle is driven affect the fuel efficiency?
- By the regression model we can predict the fuel efficiency from speed for other similar cars.

4. A graphical study of the data

✍ To better understand the relationship between two variables, we always starting with the scatter plot (x_i, y_i)'s – a visual displaying.



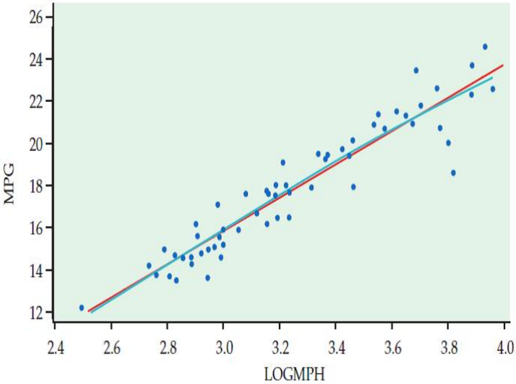
✍ **General guidelines**

- It makes no sense to do statistical inference if the data set does not approximately meet the assumption.
- At this point, we may confine our interest to speeds that are 30 mph or less, a region appears to be a good linear fit to the data.
- Also we may do some mathematical transformation to make the relationship approximately linear for the entire data.

5. Transformation making a linear relationship

✍ The scatter plot of $(\ln x_i, y_i)$'s tells us that the logarithm looks similar to the smooth-function fit of the data.

- the smooth function and the line are quite close, and
- approximately the relationship between $\ln MPH$ and MPG is linear for this data set.



✍ As a consequence, we will instead examine the effect of transforming speed by taking the natural logarithm.

✍ The statistical model assumes that these MPG's are of normally distribution with a mean μ_Y that depends upon $\ln x$ in a linear way, i.e., $\mu_Y = \beta_0 + \beta_1 \ln x$.

6. Simple linear regression model

✍ Given observations of the explanatory variable X and the response variable Y,

$$(x_1, y_1), \cdots\cdots, (x_n, y_n).$$

✍ The statistical model for simple linear regression states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \cdots, n.$$

- $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$, and
- the deviations $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$'s are independent.

✍ The parameters of the model are β_0, β_1 and σ^2. We will do inference about

- the slope β_1 and the intercept β_0 of the regression line,
- the mean response μ_Y for a given value of x, and
- an individual future response – prediction y for a given value of x.

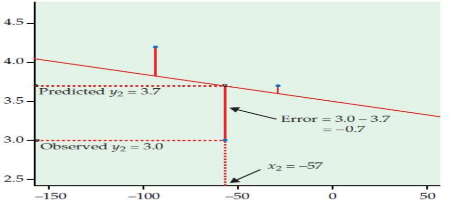
7. Sum of squared errors of the regression line

✍ Error due to the regression equation

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

is measured by

$$y_i - \hat{y}_i, \qquad i = 1, \cdots, n.$$

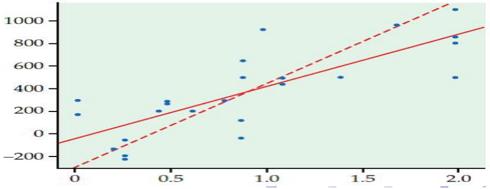


✍ The Sum of Squared Errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

measures the goodness-of-fit of the regression line $y = \beta_0 + \beta_1 x$.

✍ So, the best fit should be the line that achieves the minimum of SSE, viz., to find the best fit it suffices to minimize the above SSE.



8. Least Square Estimation of regression parameters

✍ Setting $\frac{\partial SSE}{\partial \beta_j} = 0$ for $j = 0, 1$, we get equations

$$\begin{aligned}\sum_{i=1}^{n} y_i - n \cdot \beta_0 - \sum_{i=1}^{n} x_i \cdot \beta_1 &= 0,\\ \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \beta_0 - \sum_{i=1}^{n} x_i^2 \cdot \beta_1 &= 0.\end{aligned}$$

✍ By using matrix and determinant we solve them as follows:

LSE of regression parameters

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \cdot \frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

❐ Query 1: Does $\hat{\beta}_1$ always exist? What will happen if we observe $x_1 = \cdots = x_n$?

9. LSE of regression parameters —— interpretation

✍ Recall that based on the sample $((X_1, Y_1), \cdots, (X_n, Y_n))$ we have

- the sample variance

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

- the sample correlation coefficient

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}}.$$

✍ $\hat{\beta}_1$ should have something to do with the Pearson's correlation coefficient.

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = r_{x,y}\frac{s_y}{s_x}. \end{aligned}$$

10. LSE of regression parameters —— example

✍ The body weight Y as a function of the body height x for American women:

x (m)	1.47	1.50	1.52	1.55	1.57	1.60	1.63	1.65
	1.68	1.70	1.73	1.75	1.78	1.80	1.83	
Y (kg)	52.21	53.12	54.48	55.84	57.20	58.57	59.93	61.29
	63.11	64.47	66.28	68.10	69.92	72.19	74.46	

✍ The Pearson's correlation coefficient $r_{x,y} = 0.9945$.

✍ In view of

$$\sum_{i=1}^{15} x_i = 24.76, \quad \sum_{i=1}^{15} y_i = 931.17,$$

$$\sum_{i=1}^{15} x_i^2 = 41.05, \qquad \sum_{i=1}^{15} x_i y_i = 1548.25, \qquad \sum_{i=1}^{15} y_i^2 = 58498.54,$$

we get the estimations

$$\hat{\beta}_1 = 61.272, \qquad \hat{\beta}_0 = -39.062.$$

✍ The linear regression equation is thus

$$Y = -39.062 + 61.272 \cdot X + \varepsilon.$$

11. Simple linear regression —— matrix form

✍ For the data set (x_i, y_i), $i = 1\cdots, n$, denote

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

✍ The simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \cdots, n,$$

may be rephrased as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

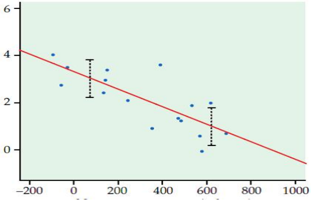
✍ If $\boldsymbol{X}'\boldsymbol{X}$ is non-degenerate (i.e., $\boldsymbol{X}$ is of full column rank), $\boldsymbol{\beta}$ gets the LSE

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}.$$

❐ QUERY 2: The generalized inverse $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is used instead if $\boldsymbol{X}'\boldsymbol{X}$ is degenerate.

12. Prediction based on the regression line

✍ With the regression model, for a given value x^* of the explanatory variable X the corresponding response variable Y is predicted as the point on the regression line



$$y^* = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*,$$

which is an unbiased estimator of the mean response μ_Y when $X = x^*$.

✍ Since $Y = \beta_0 + \beta_1 x + \varepsilon \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$, for x^* there will a deviation between the prediction y^* and the corresponding y to be observed. To get the knowledge of the deviation we have to estimate σ^2.

✍ Denote the residual $e_i = y_i - \hat{y}_i$, $i = 1, \cdots, n$. It can be proved that

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

serves as a nice estimate for σ^2.

13. Distribution of LSE for regression parameters

✍ Note that

- $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \ldots, n$, are mutually independent, and
- $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}$.

✍ Since the linear combination of normal random variable is also of normal distribution, by routine algebra we can prove the following two statements:

- $\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma_0^2)$ with
$$\sigma_0^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\sigma^2;$$
- $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma_1^2)$ with
$$\sigma_1^2 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\sigma^2.$$

✍ Obviously, both σ_0^2 and σ_1^2 are functions of the unknown variance σ^2.

14. Confidence intervals for regression parameters

✍ Replacing σ^2 with the estimation

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

we obtain the following estimations for σ_0 and σ_1,

$$\begin{aligned} SE_{\hat{\beta}_0} &= \sqrt{S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}, \\ SE_{\hat{\beta}_1} &= \sqrt{S^2\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \end{aligned}$$

✍ So, we can construct the following t-intervals with confidence level $1-\alpha$:

- for β_0, $\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2)SE_{\hat{\beta}_0}$,
- for β_1, $\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)SE_{\hat{\beta}_1}$.

15. Inference on regression parameters – example

✍ Body weight Y as a function of body height X for American women:

x (m)	1.47	1.50	1.52	1.55	1.57	1.60	1.63	1.65
	1.68	1.70	1.73	1.75	1.78	1.80	1.83	
Y (kg)	52.21	53.12	54.48	55.84	57.20	58.57	59.93	61.29
	63.11	64.47	66.28	68.10	69.92	72.19	74.46	

LSE of regression parameters are $\hat{\beta}_1 = 61.272$ and $\hat{\beta}_0 = -39.062$.

✍ The model's variance σ^2 gets the estimation

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i)\right]^2 = 0.5762.$$

✍ Consequently, the standard errors of the LSE of β_0 and β_1

$$SE_{\hat{\beta}_0} = \sqrt{S^2\left(\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} = \sqrt{3.15}, \quad SE_{\hat{\beta}_1} = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{8.63}.$$

✍ So, the 95% confidence intervals of β_0 and β_1

$$\hat{\beta}_0 \pm t_{.975}(13)SE_{\hat{\beta}_0} = [-45.4, -32.7], \qquad \hat{\beta}_1 \pm t_{.975}(13)SE_{\hat{\beta}_1} = [57.4, 65.1].$$

16. Significance test for regression parameters

✍ A slop parameter $\beta_1 \neq 0$ reveals the linear association between the response Y and the regressor X.

✍ Based on the sample we obtain the estimate $\hat{\beta}_1$ for β_1. If $\hat{\beta}_1$ is observed to be quite close to zero, one may naturally wonder whether $\beta_1 = 0$ in actual. So, we need to test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0.$$

✍ Note that $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma_1^2)$.

- Under the null H_0, it holds that

$$T = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} \sim t_{n-2}.$$

- Hence, the testing rule is to reject H_0 if the above T is observed as t having p-value

$$\mathrm{P}(|T| > |t|) = 2\mathrm{pt}(|t|, n-2) - 1 < \alpha.$$

17. Confidence intervals for mean response

✍ Plugging $X = x^*$ into the regression equation we obtain

$$\hat{\beta}_0 + \hat{\beta}_1 x^*,$$

which serves as the estimated mean response μ_Y corresponding to x^*, i.e.,

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x^* = (1, x^*)\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = (1, x^*)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Then, the variance is evaluated as

$$\sigma^2_{\hat{\mu}_y} = \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\sigma^2.$$

✍ As a result, the standard error of $\hat{\mu}_y$ is

$$SE_{\hat{\mu}_y} = \sqrt{\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)S^2}.$$

✍ Accordingly, corresponding to $X = x^*$, μ_Y gets the $1 - \alpha$ confidence interval

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{1-\alpha/2}(n-2) \cdot SE_{\hat{\mu}_y}.$$

18. Prediction intervals for mean response

✍ Also, we can consider $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as the predicted response Y^* corresponding to x^*. Then, the variance of

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^* + \varepsilon$$

can be evaluated as

$$\sigma^2_{\hat{y}} = \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sigma^2,$$

✍ Thus, the standard error is

$$SE_{\hat{y}} = \sqrt{\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)S^2}.$$

✍ Accordingly, the $1 - \alpha$ confidence interval of $\hat{Y}$ corresponding to $X = x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{1-\alpha/2}(n-2) \cdot SE_{\hat{y}}.$$

☞ It is evident that the predicted interval is wider than the confidence interval.

19. Decomposition of sum of squares

✍ Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \cdots, n$. Then, it CAN be proved that

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

✍ Since

- SST has the degree of freedom $n - 1$, and
- SSE has the degree of freedom $n - 2$,

Sum of squares due to model (SSM) gets the degree of freedom $(n-1)-(n-2) = 1$.

✍ So, it holds that

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = S^2.$$

20. ANOVA for regression

✍ To test $H_0 : \beta_1 = 0$ (i.e., Y is not linearly related to X), we can resort to F statistic

$$F = \frac{MSM}{MSE} = \frac{SSM/1}{SSE/(n-2)}$$

with degree of freedom $(1, n-2)$.

✍ Testing rule: Reject H_0 if F observed as f gets the p-value

$$1 - \mathrm{pf}(f, 1, n-2) < \alpha, \qquad \text{the significance level.}$$

✍ Again, because the coefficient of determination

$$R^2 = \frac{SSM}{SST}$$

gives the fraction of variation explained by the regression equation, it is usually used to evaluate the performance of the linear regression model – larger R^2 is an indicator of better performance.

21. Correlation coefficient

✍ The correlation coefficient (population version)

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}(Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}$$

tells the degree and direction of linear correlation between Y and X.

✍ Pearson's correlation coefficient (sample version)

$$r_{x,y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

measures the degree and direction of linear correlation within sample (x_i, y_i), $i = 1, \cdots, n$.

✍ $r_{x,y}$ serves as the estimate of $\rho_{X,Y}$. Actually, it can be proved that

- $\mathrm{E}[r_{x,y}] = \rho_{X,Y}$, and approximately,
- $\frac{r_{x,y}}{\sqrt{(1-r_{x,y}^2)/(n-2)}}$ has the Student's t distribution with degree of freedom $n-2$.

22. Inference for correlation

✍ To test $H_0 : \rho_{X,Y} = 0$ (no correlation), we utilize the Student's t statistic

$$T = \frac{r_{x,y}}{\sqrt{(1 - r_{x,y}^2)/(n-2)}} = \frac{r_{x,y}\sqrt{n-2}}{\sqrt{1 - r_{x,y}^2}}$$

with degree of freedom $n - 2$.

✍ So, for the significance level α, when T is observed as t, we reject H_0 if

- $H_a : \rho_{X,Y} \neq 0$ and the p-value

$$2[1 - \mathrm{pt}(|t|, n-2)] < \alpha,$$

- $H_a : \rho_{X,Y} > 0$ and the p-value

$$1 - \mathrm{pt}(t, n-2) < \alpha,$$

- $H_a : \rho_{X,Y} < 0$ and the p-value

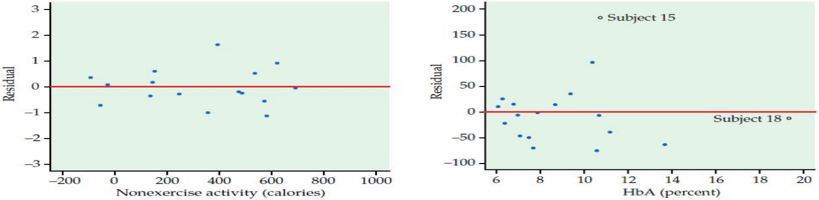
$$\mathrm{pt}(t, n-2) < \alpha.$$

23. Diagnosis based on the residual plot

✍ For the data (x_i, y_i)'s, the residual $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ records the estimated random errors ε_i, $i = 1, \cdots, n$.

✍ It is reasonable to check the performance of the model by the residual plot (x_i, e_i), $i = 1, \cdots, n$.

✍ Since ε_i's are assumed to be mutually independent and have a common normal distribution, uniformly scattered points around the real axis suggest nothing weird, and a pattern of the scatter plot tells us that the regression model can be improved by taking the pattern into account.



(a) Good fit without pattern

(b) Poor fit due to pattern and outliers

24. Multiple linear regression

✍ Multiple linear regression models the relationship between a response variable Y and several explanatory variables $X_1, \cdots, X_p$.

- Different values of $(X_1, \cdots, X_p)$ expect to produce different mean responses.
- Given $(X_1, \cdots, X_p) = (x_1, \cdots, x_p)$ the response variable Y is of normal distribution with variance σ^2 and mean

$$\mu_Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

✍ The data for multiple regression are independent observations of $(X_1, \cdots, X_p, Y)$:

$$(x_{i,1}, \cdots, x_{i,p}, y_i), \qquad i = 1, \cdots, n.$$

✍ For example, to study the fuel efficiency (MPG), except for the speed driven (MPH) it is reasonable to take the temperature (TEMP) into account. So, a better model should be achieved by characterizing the relationship between MPG and (MPH,TEMP).

25. Multiple linear regression models

✍ Given observations of the explanatory variable $(X_1, \cdots, X_p)$ and the response variable Y,

$$(x_{i,1}, \cdots, x_{i,p}, y_i), \qquad i = 1, \cdots, n.$$

✍ The statistical model for multiple linear regression states that corresponding to the value $(x_{i,1}, \cdots, x_{i,p})$ of explanatory variables the response is observed as

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \qquad i = 1, \cdots, n,$$

where random errors ε_i's are mutually independent and of $\mathcal{N}(0, \sigma^2)$.

✍ Regression parameters $\beta_0, \cdots, \beta_p$ and the variance σ^2 are all unknown, and we will do statistical inference about them:

- Least Square Estimate (LSE) and confidence intervals;
- Hypothesis tests for the significance of each parameter and the model.

26. Multiple linear regression – Least squared estimation

✍ For the data set $(x_{i,1}, \cdots, x_{i,p}, y_i)$, $i = 1 \cdots, n$, denote

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

✍ The multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i, \qquad i = 1, \cdots, n,$$

may be rephrased as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

✍ If $\boldsymbol{X}'\boldsymbol{X}$ is non-degenerate (i.e., $\boldsymbol{X}$ is of full column rank), LSE of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}.$$

27. Understanding the regression equation

✍ Based on LSE's of β_j, $j = 0, 1, \cdots, p$, we have the regression equation

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p,$$

which is also called as the fitted regression line.

✍ Based on the residuals

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_p x_{i,p}, \quad i = 1, \cdots, n,$$

σ^2, the variance of random error ε_i, is estimated as

$$S^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1}. \qquad (p + 1 \text{ equations solved to estimate } \beta_0, \cdots, \beta_p)$$

✍ It is reasonable to check the performance of the model by the residual plot

$$(i, e_i), \qquad i = 1, \cdots, n.$$

28. Inference on regression parameters

✍ Due to $\hat{\boldsymbol{\beta}} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}$, each $\hat{\beta}_j$ is of normal distribution with mean β_j and the variance proportional to σ^2, for $j = 0, 1, \cdots, p$.

✍ The t confidence interval can be constructed as

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-p-1) \cdot SE_{\hat{\beta}_j}, \quad j = 0, 1, \cdots, p.$$

✍ For $j = 1 \cdots, p$, to test for $H_0 : \beta_j = 0$ (the explanatory variable X_j is not correlated with the response Y), we utilize

$$T_j = \frac{\hat{\beta}_j}{SE_{\hat{\beta}_j}}, \qquad \text{with degree of freedom } n-p-1$$

and reject H_0 if T_j observed as t_j having the p-value

$$\mathrm{P}(|T_j| > |t_j|) = 2[1 - \mathrm{pt}(|t_j|, n-p-1)] < \alpha.$$

29. Decomposition of sum of squares

✍ Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}$, $i = 1, \cdots, n$. Then,

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

✍ Since

- SST has the degree of freedom $n - 1$, and
- SSE has the degree of freedom $n - p - 1$,

SSM gets the degree of freedom $(n - 1) - (n - p - 1) = p$.

✍ So, it holds that

$$MSE = \frac{SSE}{n - p - 1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p - 1} = S^2.$$

30. ANOVA test for significance of the regression model

✍ To check the significance of the model, we consider

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{versus} \quad H_a : \beta_j \neq 0 \text{ holds for at least one } j$$

- The null claims that all X_i's are not significant predictors for Y, and
- the alternative states that at least one of X_i's is a good predictor of Y.

✍ ANOVA suggests that under H_0 the testing statistics

$$F = \frac{MSM}{MSE} = \frac{SSM/p}{SSE/(n-p-1)} \sim \mathcal{F}_{p,n-p-1}.$$

So, we reject H_0 if F observed as f gets the p-value

$$\mathrm{P}(F > f) = 1 - \mathrm{pf}(f, p, n-p-1) < \alpha.$$

✍ Again, a larger R^2 given below implies a better performance of the regression,

$$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

and $R = \sqrt{R^2}$ is called as the multiple correlation coefficient.

31. One-way ANOVA – a linear model

✍ Say k new drugs $D_1, \ldots, D_k$ are designed for some disease. One want to know whether they are of the same performance or at least one of them is different from (superior to) the others. That is, we are interested in testing

$$H_0: \quad \mu_1 = \cdots = \mu_k \qquad \text{versus} \qquad H_1: \quad \mu_i \neq \mu_l \quad \text{for some } 1 \leq i \neq l \leq p.$$

✍ To accomplish this job, each drug is distributed to n_i patients and the following effects are collected,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1, \cdots, k, \ j = 1, \cdots, n_i.$$

- μ represents the overall effect,
- α_i represents the effect of the i-th drug, and
- ε_{ij} is the unexplained random variation of the j-th patient taking drug D_i, and all ε_{ij}'s are mutually independent and have a common normal distribution with variance σ^2.

32. One-way ANOVA – matrix form

✍ Denote $\mathbf{1}_{n_i} = (\underbrace{1, \cdots, 1}_{n_i})', \mathbf{0}_{n_i} = (\underbrace{0, \cdots, 0}_{n_i})', \quad i = 1, \cdots, k,$

$$\boldsymbol{Y} = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ \vdots \\ Y_{k,1} \\ \vdots \\ Y_{k,n_k} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \cdots & \mathbf{1}_{n_k} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,n_1} \\ \vdots \\ \varepsilon_{k,1} \\ \vdots \\ \varepsilon_{k,n_k} \end{pmatrix}$$

✍ In matrix form, one-way ANOVA model may be formulated as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{X}$ has the column rank $k < k + 1$ and thus $\boldsymbol{X}'\boldsymbol{X}$ is not invertible any more.

✍ So, $H_0 : \mu_1 = \cdots = \mu_k$ v.s. $H_a : \mu_i \neq \mu_l$ for some $1 \leq i \neq l \leq n$ is equivalent to

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0 \quad \text{v.s.} \quad H_a : \alpha_i \neq 0 \quad \text{for some } i = 1, \cdots, k.$$

33. Concluding remarks

✍ Two-way ANOVA can also be represented as a linear model with the closed matrix form.

✍ Both ANOVA and linear regression are special cases of the linear model, which is one of the most well-developed and useful branch of statistics.

✍ By using matrix theory and probability theory, an unified mathematical expression for the statistical inferences (parameter estimation and hypothesis testing) can be obtained.

✍ Based on the accurate expressions one can have a better understanding on the statistical models and thus effectively apply them to practical problems.

✍ No doubt, industrial enterprises and IT companies are in favor of computer guys with a better background in statistics.