ENGG 319 CURE PROJECT

Exploring linear relationship between two variables by correlation analysis and simple linear regression analysis

---

MODEL

The simple linear regression model:
y = beta0 + beta1 (x) + error
x= independent or explanatory variable, y = dependent or response variable
beta0,beta1 = regression coefficients, beta0= y-intercept for population; beta1= slope for population

The sample linear regresson model:
yhat = beta0hat + beta1hat (x)
yhat= predicted value of y for a given value of x;
beta0,beta1 = least-squares coefficients
beta0hat= sample y-intercept = estimate of betanot, beta1hat = sample slope = estimate of beta1

EXAMPLE

How much noisier are streets where cars travel faster?
Data (Average speed in kmph vs noise in dB)
Source: Journal of Transportation Engineering, 1999: 152-159

Speed Noise
28.26 78.1
36.22 79.6
38.73 81.0
29.07 78.7
30.28 78.6
30.25 78.5
29.03 78.4
33.17 79.6
(a) Compute the least-squares line for predicting noise level (y) from speed (x)
(b) Compute the error standard deviation estimate s
(c) Construct a 95% confidence interval for the slope
(d) Find a 95% confidence level for the mean noise level for streets whose average speed is 30 kmph.
(e) Can you conclude that the mean noise level for streets whose average speed is 30 kmph is greater than 70 db?

SAMPLE DATASET
Sample size, n (greater than 2 but less than 21): 19

Please use the first n number of rows in the table below to input n pairs of observations.

x1:	1901	y1:	73022
x2:	1911	y2:	374295
x3:	1921	y3:	588454
x4:	1931	y4:	731605
x5:	1941	y5:	796169
x6:	1951	y6:	939501
x7:	1956	y7:	1123116
x8:	1961	y8:	1331944
x9:	1971	y9:	1627875
x10:	1976	y10:	1838035
x11:	1981	y11:	2237724
x12:	1986	y12:	2365830
x13:	1991	y13:	2545553
x14:	2001	y14:	2974807
x15:	2006	y15:	3290350
x16:	2011	y16:	3645257
x17:	2016	y17:	4067175
x18:	1969	y18:	1463203
x19:	1996	y19:	2696826
x20:		y20:	

Provide a x value for prediction of y : 2040

Note: The chosen value of x should be between the minimum and maximum value of x in the sample to avoid extrapolation

REGRESS

CORRELATION ANALYSIS

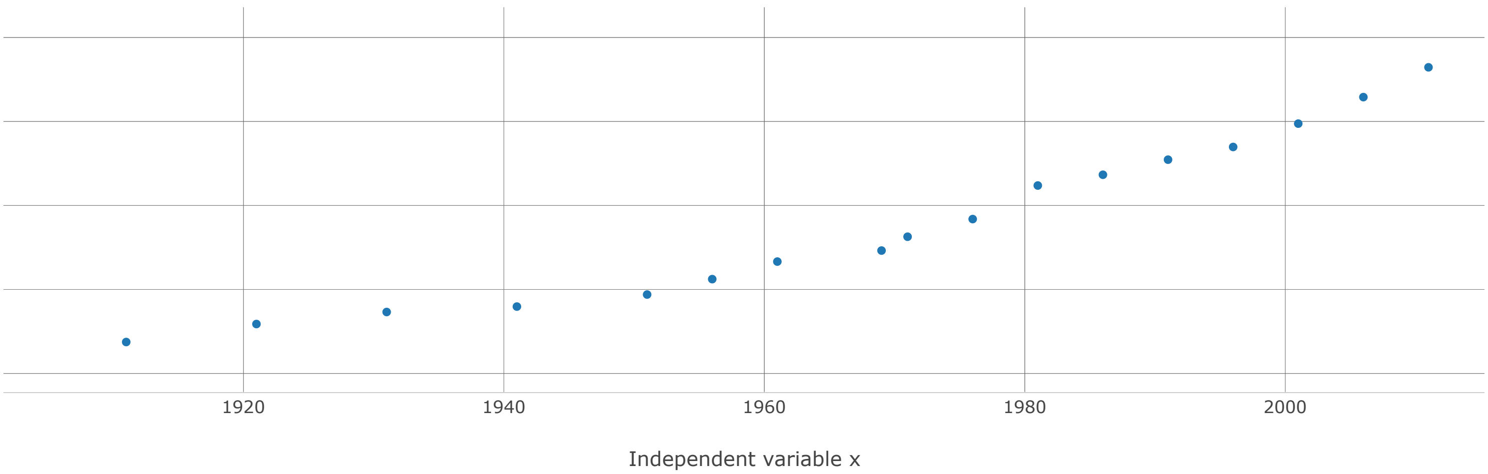
The sample correlation coefficient (r):
0.963

Note: Correlation coefficient, r, denotes the strength of linear association between x and y, see also FIG 1.:

The 95% upper confidence level for population correlation coefficient (rho):
0.9861

The 95% lower confidence level for population correlation coefficient (rho):
0.9055

FIG 1. Scatter plot, plot of y versus x

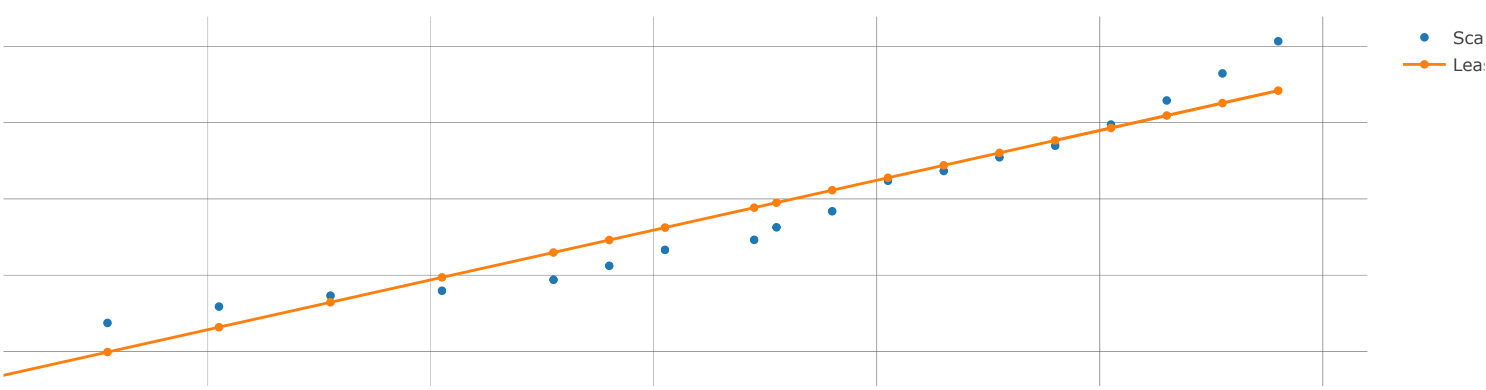


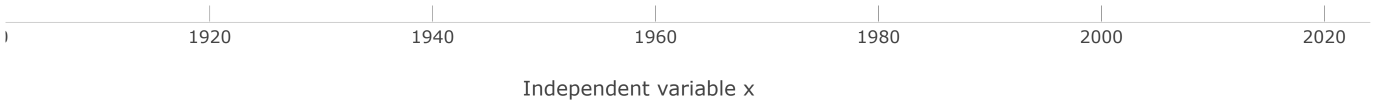
REGRESSION ANALYSIS

LEAST-SQUARES LINE (BEST FIT LINE)

Sample slope, beta1hat:
32642.31580708
Sample intercept, beta0hat:
-62387426.15375166
The coefficient of determination, r^2 :
0.92824

FIG 2. Scatter plot of y versus x. The least-squares line is superimposed.





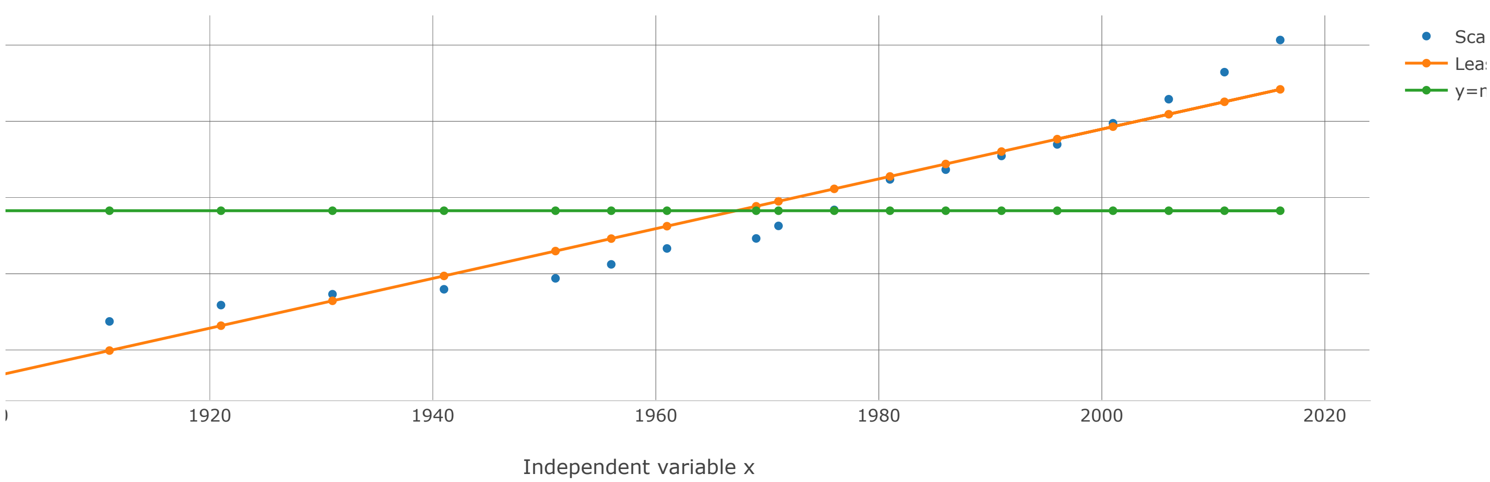
INFERENCES ON POPULATION SLOPE AND INTERCEPT

The 95% lower confidence level of the population slope(beta1):
27997.8108
The 95% upper confidence level of the population slope(beta1):
37286.8208
The 95% lower confidence level of the population intercept(beta0):
-71525454.9313
The 95% upper confidence level of the population intercept(beta0):
-53249397.3762
The t-statistic for testing the hypothesis: beta0 = 0:
-14.4055
The t-statistic for testing the hypothesis: beta1 = 0:
14.8294

GOODNESS-OF-FIT

The standard deviation of y vlaues (quantifies spread of sample data around the mean, see FIG 3.):
1159474.068
The standard error of estimate (quantifies spread of sample data around the regression line, see FIG 3):
319598.114
Note: The linear regression model has merit if the standard error of estiamte is less than the standard deviation of y values.
The coefficient of determination(SSR/SST, proportion of variation in y explained by regression):
0.9282

FIG 3. Scatter plot of y versus x. The least-squares line and the horizontal line y=y_mean are superimposed.



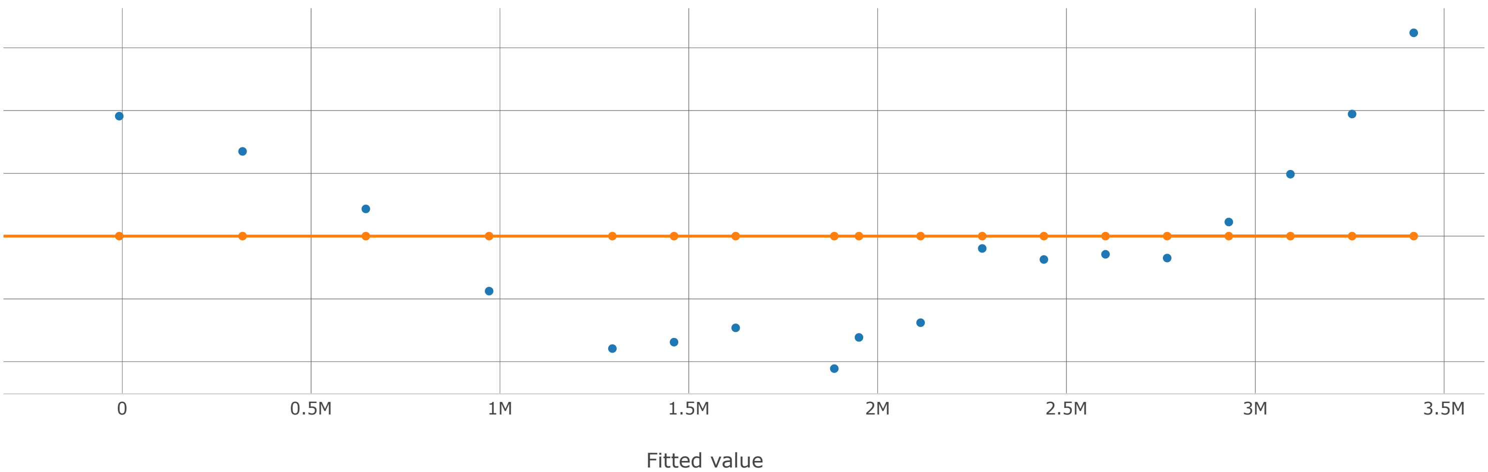
PREDICTION AND INFERENCES ON beta0+(beta1)(x)

The value of x chosen to predict the value of y:
2040
The predicted value of y:
4202898.0927
The standard deviation of yhat=beta0+(beta1)(x):
176202.8297
The 95% upper confidence level for beta0+(beta1)(x):
4574686.0633
The 95% lower confidence level for beta0+ (beta1)(x):
3831110.1221
The 95% upper prediction level for beta0+(beta1)(x):
4972948.0544
The 95% lower prediction level for beta0+ (beta1)(x):
3432848.1310

CHECKING ASSUMPTIONS
Assumptions: (i) The errors are random and independent

(ii) All errors have mean 0
(iii) The errors have same variance
(iv) The errors are normally distributed.
Note: See FIG 4. When the assumptions are satistied, the plot will show no substatial pattern (no curve).
The vertical spread of the points does not vary too much over the range of x values.

FIG 4. Plot of residuals(e_i) versus fitted/predicted values (yhat_i)



UNCERTAINTIES IN THE LEAST-SQUARES COEFFICIENTS
The standard deviation of beta0hat, $s_{beta0hat}$:
4330819.3258
The standard deviation of beta1hat, $s_{beta1hat}$:
2201.1872
Note: The smaller the uncertainties (standard deviations), the more precise are the estimation

---

ANALYSIS OF VARIANCE (ANOVA)
The total sum of squares (total variation)=SST = sum of squared deviations =$(y_i - \bar{y})^2$:
24198842058863.7930
The error sum of squares (unexplained variation)= SSE = sum of squared residuals =$(y_i - \hat{y}_i)^2$
1736430229225.2700
The regression sum of squares (explained variation) = SSR = SST-SSE = $(\hat{y}_i - \bar{y})^2$:
22462411829638.5234

F-statistic, a test-statistic from F-distribution with k=1 and n-k-1=n-2 degrees of freedom,
F = MSR/MSE, MSR=SSR/k, MSE= SSE/(n-k-1):
219.9115

SYMBOLS, FORMULAS, NOTES ON HYPOTHESIS TEST
Simple linear regression model: $y_i = \beta_0 x + \beta_1 x + \varepsilon_i$;
Sample linear regression model: $\hat{y}_i$ = (beta0hat)x + (beta1hat) x_i;
y_i = i-th observation of the dependent variable.
$\hat{y}_i$ =Predicted value of the i-th observation.
Residuals, $e_i = y_i - \hat{y}_i$.
Errors, $\varepsilon_i = y_i - (\beta_0 x + \beta_1 x)$.
Note: Residual = observed value - fitted value; error = observed value - true value.
Note: By 'error' we mean measurement error (as in ch 3), by 'residual' we mean prediction error (deficiency).
σ^2 = Error variance = Variance of the error ε_i;
To estimate error variance define, SSE = The sum of e_i^2
s^2 = A point-estimate of error variance = SSE/(n-2);
Note: 'beta0hat' is a point-estimate of β_0, 'beta1hat' is a point-estimate of β_1
Note: (beta0hat $-\beta_0)/s_{beta0hat}$ has t-distribution with dof = n-2.
Note: (beta1hat $-\beta_1)/s_{beta1hat}$ has t-distribution with dof = n-2.
Note: If H_1 has > sign, P-value is area to the right of observed t-value.
Note: If H_1 has < sign, P-value is area to the left of observed t-value.
Note: If H_1 has ≠ sign, P-value is areas under the tails cut-off by t and -t.
Note: If $P \leq \alpha$, the null hypothesis is rejected at the 100α% level.

Note: If $P \leq \alpha$, the test is statistically significant at the $100\alpha\%$ level.
