Simple Linear Regression

(Find the Least-Square Line).

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\left[\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i\right]$$

Computational Formulas for Sum of Squares

$$SSE = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i$$

$$SSE = SSTO - SSR$$

$$SSR = \hat{\beta}_1^2 \cdot \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] = \hat{\beta}_1^2 \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum y_i)^2}{n}$$

Coefficient of Determination

$$R^2 = \frac{SSR}{SSTO}$$

Note: $0 \le R^2 \le 1$

"large", "moderate", "low".

Coefficient of Correlation

$r = \sqrt{R^2}$ if $\hat{\beta}_1 > 0$;

or $-\sqrt{R^2}$ if $\hat{\beta}_1 < 0$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \times \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Note: $-1 \le r \le 1$ and $|r| > R^2$

if $r = 0 \rightarrow$ Only there is no linear relationship between x and y. That doesn't imply that there is no relationship between these variables.

Correlation does not imply Causation

F-test under ANOVA for simple linear regression model.

$H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

Test statistic: $F = \frac{MSR}{MSE}$

Decision rule:

Reject H_0 if $F > F_\alpha$, where F_α is based on $(1, n-2)$ degrees of freedom.

Error term: ε_i

The random errors are independent random variables that are normally distributed with mean 0 and variance σ^2.

Underlying Assumptions of the Model.

1. Linearity: Y and X are linearly related.
2. Normality: Random errors are normally distributed.
3. Constant variance: Error variance is constant across all values of the covariate X.
4. Independence: Errors associated with the observations are mutually independent.

Heteroscedasticity

- When the requirement of a constant variance is violated we have a condition of heteroscedasticity.

Examining the residual plots: Detection of nonlinearity and heteroscedasticity.

- The residual plot fans out or fans in: Variance is not constant.
- Fans out: the variance is increasing as a function of the covariate X.
- Fans in: the variance is a decreasing function of X.

Detection of nonnormality:

- Examining the normal probability plot of the residuals.

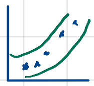
Cook's distance

- Measure of influence of a data point.
- $D_i > 1$: Have high influence.

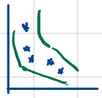
Remedial Measures (Power Transformations)

Transformations	on dep. var.	on ind. var.
Square Root	$\sqrt{Y}$	$\sqrt{X}$
Log	$\log(Y)$	$\log(X)$
Inverse	$1/Y$	$1/X$

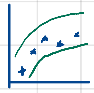
Selection of transformation to remedy non-linearity



Try transforming Y first.



Try transforming X and/or Y. but choose the variable with wider range first.



Try transforming X first.

Remedial measure for non-constant Variance: Must involve a transformation of Y. (Effective: $\sqrt{Y}$ or $\log(Y)$).

Remedial to Non-Nonlinearity

- Usually non-normality and non-constant variance occur together.
- So just use measures for non-constant variance.

Test for Linear Correlation

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

One-tail tests

$H_0: \rho = 0 \quad H_a: \rho > 0$

$H_0: \rho = 0 \quad H_a: \rho < 0$

Two-tailed Test

$H_0: \rho = 0 \quad H_a: \rho \neq 0$

Analysis of Variance (ANOVA)

- Variation in y = SSR + SSE

V = V. explained by the regression line + Unexplained variation (error).

$$\sigma^2 = MSE = \frac{SSE}{df_{residual}}$$

t test for SLR

$H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - 0}{SE_{\beta_1}}$$

Other forms: Simple Linear Regression

$$\beta_1 = \frac{SSxy}{SSxx}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$SSE = SSyy - \beta_1 \cdot SSxy = \sum (y_i - \hat{y}_i)^2$$

$$SSxy = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$SSxx = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Interpretation of the β_1 and β_0 in SLR.

$\hat{\beta}_0$: Represents the predicted value of y when $x = 0$.

$\hat{\beta}_1$: Represents the increase (or decrease) in y for every 1-unit increase in x. (Valid only when x-values within the range of sample data).

Interpretation of s, the Estimated Standard Deviation of σ.

We expect most ($\approx 95\%$) of the observed y values to lie within 2s of their respective least squares predicted values, $\hat{y}$.

Sampling Distribution of $\hat{\beta}_1$

$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SSxx}}$, $s_{\hat{\beta}_1}$ as the estimated standard error of the least squares slope $\hat{\beta}_1$.

Figures we need to know to construct a least squares:

1. Sum of all x-values: $\sum x$: "= sum(A1:An)"
2. Sum of all y-values: $\sum y$: "= sum(B1:Bn)"
3. Sum of the product of x and y: $\sum xy$: "C1 = A1·B1" "= sum(C1:Cn)"
4. Sum of the squares of the x: $\sum x^2$
5. Number of data points (n).

ANOVA Table (In SLR Context)

- df(regression) = 1
- df(residual) = $n-2$
- df(total) = $n-1 = 1+(n-2)$
- $SSR = \sum (n_i \cdot (\bar{x}_i - \bar{x}))^2$
- $SSE = \sum (x_i - \hat{x}_i)^2$
- $SSTO = \sum (x_i - \bar{x})^2$
- $MSR = \frac{SSR}{df_{regression}}$ · $MSE = \frac{SSE}{df_{error}}$
- $F = \frac{MSR}{MSE}$

y_i: actual value

$\hat{y}_i$: predicted value

$\bar{y}$: mean of y.

$$SSR = R^2 \times SSTO = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Source of Variation	df	SS	MS	F
Regression	1	$\sum (n_i(\bar{x}_i - \bar{x}))^2$	$\frac{SSR}{df_{ssr}}$	$\frac{MSR}{MSE}$
Residual	$n-2$	$\sum (x_i - \hat{x}_i)^2$	$\frac{SSE}{df_{sse}}$	
Total	$n-1$	$\sum (x_i - \bar{x})^2$		

$$s^2 = \frac{SSE}{n-2}$$

$$SSyy = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$$

Residuals: $e_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$

Error Sum of Squares (SSE):

$$\sum_{i=1}^{n} e_i^2$$

Estimation of σ^2 (Mean Squared Error)

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

Computational formula for σ^2

$$\sigma^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \cdot \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i\right]$$

Inference about model parameters.

- **Inference about β_1**

To test $H_0: \beta_1 = b$

Test statistic: $t = \dfrac{\hat{\beta}_1 - b}{s(\hat{\beta}_1)}$,

where $s(\hat{\beta}_1)$ is the standard error of the estimate $\hat{\beta}_1$ and is given by:

$$s^2(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

Decision Rule:

$H_0: \beta_1 = b$ vs. $H_a: \beta_1 > b$

Reject H_0 if $t > t_\alpha$

$H_0: \beta_1 = b$ vs. $H_a: \beta_1 < b$

Reject H_0 if $t < -t_\alpha$

$H_0: \beta_1 = b$ vs. $H_a: \beta_1 \neq b$

Reject H_0 if $t < -t_{\frac{\alpha}{2}}$ or if $t > t_{\frac{\alpha}{2}}$

- t_α and $t_{\frac{\alpha}{2}}$ values are based on $(n-2)$ df.
- $H_0: \beta_1 = 0$ → No linear relationship between the dependent variable (y) and covariate (x).
- $H_a: \beta_1 \neq 0$ → Suggest the existence of a linear relationship.

Inference about β_1: Confidence Interval:

- **Confidence interval for β_1 with confidence coefficient $(1-\alpha)$**

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{SS_{xx}}}$$

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} \cdot s(\hat{\beta}_1)$$

df: $n-2$

Inference about β_0.

- May not be meaningful unless data is available at or near $x = 0$.

Confidence interval for β_0

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}} \cdot s(\hat{\beta}_0)$$

where $s(\hat{\beta}_0)$ is the standard error of $\hat{\beta}_0$ given by:

$$s^2(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]}$$

Test of hypothesis for β_0

To test $H_0: \beta_0 = a$

Test statistic: $t = \dfrac{\hat{\beta}_0 - a}{s(\hat{\beta}_0)}$

Decision Rule:

$H_0: \beta_0 = a$ vs. $H_a: \beta_0 > a$

Reject H_0 if $t > t_\alpha$

$H_0: \beta_0 = a$ vs. $H_a: \beta_0 < a$

Reject H_0 if $t < -t_\alpha$

$H_0: \beta_0 = a$ vs. $H_a: \beta_0 \neq a$

Reject H_0 if $t < -t_{\frac{\alpha}{2}}$ or if $t > t_{\frac{\alpha}{2}}$

- t_α and $t_{\frac{\alpha}{2}}$ values are based on $(n-2)$ df.

Common method for significance testing in regression model: "To test if the slope is different from 0."

Inference about the dependent variable.

- Two types of inference:
  - Inference about the mean of y (expected value of y or $E(y)$) at a given value of $x = x_m$.
  - Prediction of an individual value of y at a given value of $x = x_p$.

Interval estimation of $E(y)$ at $x = x_m$.

Point Estimate of $E(y)$: $\hat{y}_m = \hat{\beta}_0 + \hat{\beta}_1 x_m$

Confidence Interval: $\hat{y}_m \pm t_{\frac{\alpha}{2}} \cdot s(\hat{y}_m)$

where $s(\hat{y}_m)$ is the standard error of the estimate $\hat{y}_m$ given by:

$$S(\hat{y}_m) = \hat{\sigma} \cdot \sqrt{\left[\frac{1}{n} + \frac{(x_m - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

Sample: Construct a 95% confidence interval for $E(y)$ at $x = 5$. This is an estimate of the average annual salary of all people with 5 years of experience.

$\hat{y}_m = -0.675 + 14.229 \times (5) = 70.47$

$s(\hat{y}_m) = 19.61 \times \sqrt{\left[\frac{1}{12} + \frac{(5-5.5)^2}{59}\right]} = 5.80$

↳ $70.47 \pm 2.228 \times 5.80 \rightarrow (57.55, 83.39)$.

Prediction interval of Y at $x = x_p$

Best prediction of Y: $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$

Prediction Interval: $\hat{y}_p \pm t_{\frac{\alpha}{2}} s(\hat{y}_p)$

Standard Error of $\hat{y}_p$:

$$S(\hat{y}_p) = \hat{\sigma} \cdot \sqrt{\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

$$S^2(\hat{y}_p) = \hat{\sigma}^2 + S^2(\hat{y}_m)$$

Sample: Construct a 95% prediction interval of Y at $x = 5$. This prediction interval provides a salary estimate of an individual with 5 years of experience.

$\hat{y}_p = -0.675 + 14.229 \cdot (5) = 70.47$

$S(\hat{y}_p) = 19.61 \sqrt{\left[1 + \frac{1}{12} + \frac{(5-5.5)^2}{59}\right]} = 20.45$

↳ $70.47 \pm 2.228 \times 20.45 \rightarrow (24.91, 116.03)$

Note: The prediction interval for Y at $x = 5$ is significantly wider than the confidence interval for $E(y)$ at $x = 5$.

Calculator:
- Inverse T: invt(0.975, 6) → 2.44691
- Inverse F: invF(0.95, 1, 10) → 4.9646