Linear Regression

Formula	Explanation
$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ where $\text{SS}_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$ $\text{SS}_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$	Least squares estimates of β's
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	Least-squares line
$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \text{SS}_{yy} - \hat{\beta}_1 \text{SS}_{xy}$ where $\text{SS}_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$	Sum of squared errors
$s^2 = \frac{\text{SSE}}{n-2}$	Estimated variance of σ^2 of ϵ
$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\text{SS}_{xx}}}$	Estimated standard error of $\hat{\beta}_1$
$T = \frac{\hat{\beta}_1 - \beta_1(0)}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$	Test statistic for H_0: $\beta_1 = 0$ i.e. # of SDs from the student's T distribution mean
$\hat{\beta}_1 \pm (t_{\alpha/2}) s_{\hat{\beta}_1}$	$(1-\alpha)100\%$ confidence interval for β_1
$r^2 = \frac{\text{SS}_{yy} - \text{SSE}}{\text{SS}_{yy}}$	Coefficient of determination
$r = \frac{\text{SS}_{xy}}{\sqrt{\text{SS}_{xx}\text{SS}_{yy}}} = \pm\sqrt{r^2}$ (same sign as $\hat{\beta}_1$)	Coefficient of correlation
$\hat{y} \pm (t_{\alpha/2}) s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\text{SS}_{xx}}}$	$(1-\alpha)100\%$ confidence interval for $E(y)$ when $x = x_p$
$\hat{y} \pm (t_{\alpha/2}) s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\text{SS}_{xx}}}$	$(1-\alpha)100\%$ prediction interval for y when $x = x_p$

Linear Regression Example

X	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1	420	408.16	11.84	139.7124
2	410	422.36	-12.36	153.2644
3	437	436.56	0.44	0.1764
4	467	450.76	16.24	263.0884
5	448	464.96	-16.96	288.3204
6	460	479.16	-19.16	261.7924
7	507	493.36	13.64	185.5044
8	514	507.56	6.44	41.2164
$\sum X = 36$	$\sum Y = 3,663$	$\sum \hat{Y} = 3662.88$	$\sum(Y - \hat{Y}) = 0.12$	$\sum(Y - \hat{Y})^2 = 1439.1569$

$$
\begin{aligned}
\sum xy &= 1 \cdot 420 + 2 \cdot 410 + 3 \cdot 437 + 4 \cdot 467 + 5 \cdot 448 + 6 \cdot 460 + 7 \cdot 507 + 8 \cdot 514 = 17080 \\
\mathrm{SS}_{xy} &= 17080 - \frac{36 \cdot 3663}{8} = 596.05 \\
\sum x^2 &= 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 8^2 = 204 \\
\mathrm{SS}_{xx} &= 204 - \frac{36^2}{8} = 42 \\
\hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} = \frac{597.04}{42} = 14.20238 \\
\bar{x} &= 36/8 = 4.5 \\
\bar{y} &= 3663/8 = 457.875 \\
\hat{\beta}_0 &= 457.875 - 14.20238 \cdot 4.5 = 393.96 \\
\hat{y} &= 393.96 + 14.20x \\
\mathrm{SSE} &= \sum (y_i - \hat{y}_i)^2 = 1439.1569 \\
s^2 &= \frac{\mathrm{SSE}}{n-2} = 239.8595 \\
s &= \sqrt{239.8595} = 15.4874 \\
s_{\hat{\beta}_1} &= \frac{15.4874}{\sqrt{42}} = 2.389757326 \\
T &= \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{14.20238}{2.389757326} = 5.943021848 \qquad (1)
\end{aligned}
$$

Looking at the table for the students T distribution for 6 degrees of freedom, we see that at the 95% confidence level, in order to accept the null hypothesis we must be within $\pm$ 2.44691185 standard deviations of the mean error. Our results show that we are +5.943021848 standard deviations from the mean. Therefore we reject the null hypothesis and the values of X and Y are correlated and NOT random.

If we use the p calculator for student's T distribution for 5.943021848 and 6 degrees of freedom, we get a p value of 0.0010140, which also leads us to the conclusion that we reject the null hypothesis.

The 95% confidence interval for β_1 is $\hat{\beta}_1 \pm (t_{\alpha/2}) s_{\hat{\beta}_1} = 14.20238 \pm 2.44691185 \cdot 2.389757326 =$ (8.3525, 20.0475)

$r^2 = \frac{8468.879 - 1439.1569}{8468.879} = 0.8548$ where $\mathrm{SS}_{yy} = 8468.879$

$r = \sqrt{r^2} = 0.9245$

Confidence interval of the mean value of y when $x = 2$

$$422.36 \pm 2.44691185 \cdot 15.4874 \cdot \sqrt{1/8 + \frac{(2-4.5)^2}{42}} = (402.5301, 442.1899)$$

Confidence interval of the predicted value of y when $x = 2$

$$422.36 \pm 2.44691185 \cdot 15.4874 \cdot \sqrt{1 + 1/8 + \frac{(2-4.5)^2}{42}} = (379.5890314, 465.1310)$$