Simple Linear Regression

* In simple linear regression we are concerned about the relationship between two variables, X and Y.

* There are two components to such a relationship.
  1. The strength of the relationship.
  2. The direction of the relationship.

* We shall also be interested in making inferences about the relationship.

* We will be assuming here that the relationship between X and Y is linear (or has been linearized through transformation).

Covariance

* Suppose that (X, Y) is a bivariate random vector.

* An important characteristic of the joint distribution of this random vector is the covariance.

* The theoretical definition of the covariance is

$$\text{Cov}(X, Y) \;=\; \text{E}\Big((X - \mu_X)(Y - \mu_Y)\Big)$$

where $\mu_X = \text{E}(X)$ and $\mu_Y = \text{E}(Y)$.

* It is easy to see that an alternative equivalent expression is

$$\text{Cov}(X, Y) \;=\; \text{E}(XY) - \text{E}(X)\text{E}(Y).$$

This form is more often used in practice.

Covariance

* If X and Y are related in such a way that large values of Y tend to occur with large values of X then Cov(X, Y) will be positive.

* Similarly if large values of Y occur with small values of X then Cov(X, Y) will be negative.

* In this sense, the sign of Cov(X, Y) tells us the direction of the relationship between X and Y.

* If X and Y are independent random variables then Cov$(X, Y) = 0$ (but the reverse is NOT true).

* Note that Cov$(X, X) = \mathsf{E}\left((X - \mu_X)^2\right) = \mathsf{Var}(X)$.

Covariance of Linear Combinations

* The following theorem will be used later and gives the covariance between two linear combinations of the same set of independent random variables.

Theorem 2

Suppose $Y_1, \ldots, Y_n$ *are independent random variables. Let* $a_1, \ldots, a_n$ *and* $b_1, \ldots, b_n$ *be two sets of real constants and define the two linear combinations*

$$U_a = \sum_{i=1}^{n} a_i Y_i \qquad U_b = \sum_{i=1}^{n} b_i Y_i$$

Then U_a *and* U_b *have covariance*

$$\text{Cov}(U_a, U_b) = \sum_{i=1}^{n} a_i b_i \text{Var}(Y_i)$$

Correlation

* A problem with the covariance is that it's value depends on the units in which X and Y are measured and so is not a good measure of strength of the relationship.

* To remove the effect of units we standardise the variables.

$$Z_X = \frac{X - \mu_X}{\sqrt{\text{Var}(X)}} \qquad Z_Y = \frac{Y - \mu_Y}{\sqrt{\text{Var}(Y)}}$$

* Both Z_X and Z_Y have mean 0 and variance 1.

* The covariance between Z_X and Z_Y is called the correlation.

$$\rho_{X,Y} = \text{Cov}(Z_X, Z_Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

Correlation (ctd)

* It can be shown that

$$-1 \leqslant \rho_{X,Y} \leqslant 1.$$

* The sign of ρ is the same as the sign of the covariance and so gives the direction of the relationship.

* When the relationship is perfectly linear then $|\rho| = 1$.

* If the two variables are independent then $\rho = 0$.
  NOTE The inverse of this does not hold

* The strength of the relationship between the variables can be assessed by $|\rho|$ (or ρ^2).

The Sample Correlation

* The sample estimate of ρ based on the random sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is

$$r_{X,Y} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{Sxy}{\sqrt{Sxx\ Syy}}.$$

* This sample quantity also satisfies that $-1 < r < 1$.

* It is important to note that r only measures the strength of the linear relationship between the two observed variables.

* When a linear relationship between X and Y is plausible, r gives us a very good indication of the strength and direction of that relationship.

A test for $\rho = 0$

* When $\rho = 0$ and the joint distribution of (X, Y) is bivariate normal it can be shown that

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

has a Student's t distribution with $n-2$ degrees of freedom.

* We can therefore test

$$H_0 : \rho = 0 \quad \text{V} \quad H_1 : \rho \neq 0$$

by calculating the observed value of t (call this t_{obs}).

* A p-value for the test can then be found as

$$p = 2\text{P}\left(t_{n-2} \geqslant |t_{\text{obs}}|\right)$$

Simple Linear Regression

* Correlation is an attribute of the joint distribution of (X, Y).

* Cov(X, Y) = Cov(Y, X) and $\rho_{X,Y} = \rho_{Y,X}$.

* Regression is based on the conditional distribution of Y given X.

* Regression is generally not symmetric so it matters which variable is called Y (the response) and which is called X (the covariate).

* Since X is considered fixed in regression, it is not necessary that it be a random variable at all.

The Simple Linear Model

* The assumptions of the model are
  1. A linear relationship $Y = \beta_0 + \beta_1 X + \varepsilon$ exists between X and Y.
  2. $\mathrm{E}(\varepsilon \mid X = x) = 0$ and $\mathrm{Var}(\varepsilon \mid X = x) = \sigma^2$ for every x.
  3. $\varepsilon_1, \ldots, \varepsilon_n$ is a random sample from a $\mathrm{N}(0, \sigma^2)$ distribution.

* In terms of Y this means that the conditional distribution of Y given $X = x$ is normal

$$Y \mid X = x \quad \sim \quad \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$$

* Note that the marginal (unconditional) distribution of Y may not be normal and this is not required for our model. All that is required is that the conditional distribution is normal for every x under consideration.

Fitted Values and Errors

* Suppose that we have a dataset $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$.

* Our interest is in using our model to predict values of Y for any given value of $X = x$.

* If we know the values of β_0 and β_1 then the fitted value for the observation y_i would be $\beta_0 + \beta_1 x_i$.

* The error in the fitted value can be measured by the vertical distance

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Estimating the Parameters

* We wish to find a line which makes the smallest total vertical error.

* Since we do not want negative errors to cancel out positive errors we use the sum of squared errors

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

to be an overall measure of the fit of the line.

* The Method of Least Squares is an estimation method which estimates β_0 and β_1 as those values which minimize $S(\beta_0, \beta_1)$.

Least Squares Estimates

Theorem 3

Suppose we have a dataset

$$(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$$

for which the simple linear model holds. Then the least squares estimates of β_1 *and* β_0 *are given by*

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum(y_i - \overline{y})(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} \quad = \quad \frac{Sxy}{Sxx} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} \end{aligned}$$

The Least Squares Regression Line is then

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

Residuals

* For each observation in our dataset we can compute the fitted value

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \quad i = 1, \ldots, n.$$

* The vertical distance from the observed y_i to the fitted value is called the residual

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \quad i = 1, \ldots, n.$$

* The residuals can be thought of as predicted values of the unknown errors $\varepsilon_1, \ldots, \varepsilon_n$.

Properties of the Least Squares Line

* The least squares line always passes through the point $(\overline{x}, \overline{y})$.

* The estimated slope $\widehat{\beta}_1$ always has the same sign as the sample correlation between $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$.

* The sum of the residuals is 0.

* The sum of the squares of the e_i's is called the Residual Sum of Squares or Sum of Squared Errors (SSE).

* An unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\text{SSE}}{n-2}.$$

Theoretical Properties

Theorem 4

Suppose that the linear model holds and that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least squares estimators given in Theorem 3. Then

1.
$$\mathrm{E}\left(\widehat{\beta}_1 \mid x_1, \ldots, x_n\right) = \beta_1$$
$$\mathrm{Var}\left(\widehat{\beta}_1 \mid x_1, \ldots, x_n\right) = \frac{\sigma^2}{Sxx}$$

2.
$$\mathrm{E}\left(\widehat{\beta}_0 \mid x_1, \ldots, x_n\right) = \beta_0$$
$$\mathrm{Var}\left(\widehat{\beta}_0 \mid x_1, \ldots, x_n\right) = \left[\frac{1}{n} + \frac{\overline{x}^2}{Sxx}\right]\sigma^2$$

3.
$$\mathrm{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1 \mid x_1, \ldots, x_n\right) = -\frac{\overline{x}}{Sxx}\sigma^2.$$

Standard Errors

* The variances in Theorem 4 depend on the unknown value of σ^2.

* If we replace σ^2 with the unbiased estimator $\widehat{\sigma}^2$ and take the square root we get the standard errors of the estimators.

$$\text{se}(\widehat{\beta}_0) = \sqrt{\left[\frac{1}{n} + \frac{\overline{x}^2}{Sxx}\right]\widehat{\sigma}^2}$$

$$\text{se}(\widehat{\beta}_1) = \sqrt{\frac{\widehat{\sigma}^2}{Sxx}}$$

* The software will always report both the estimates and their standard errors.

Sampling Distributions

* The following theorem will not be proven in this course but the result is very important.

Theorem 5
Suppose that the linear model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ *holds and* $\varepsilon_1, \ldots, \varepsilon_n$ *are* iid $N(0, \sigma^2)$. *Then*

$$\frac{\widehat{\beta}_0 - \beta_0}{\text{se}(\widehat{\beta}_0)} \sim t_{n-2}$$

$$\frac{\widehat{\beta}_1 - \beta_1}{\text{se}(\widehat{\beta}_1)} \sim t_{n-2}$$

Confidence Intervals

* We can use Theorem 5 to find confidence intervals for β_0 and β_1.

* Let $t_{(n-2,\ \alpha/2)}$ be the $(1-\alpha/2)$ percentile of the t distribution with $n-2$ degrees of freedom.

* Table A.2 in your textbook gives the values of this percentile for various values of $\alpha/2$ and degrees of freedom $n-2$.

* Confidence intervals for β_0 and β_1 are then

$$\widehat{\beta}_0 \pm t_{(n-2,\alpha/2)} \text{se}(\widehat{\beta}_0)$$

$$\widehat{\beta}_1 \pm t_{(n-2,\ \alpha/2)} \text{se}(\widehat{\beta}_1)$$

Hypothesis Testing

* Suppose that we wish to test the hypotheses

$$H_0 : \beta_1 \;=\; \beta_1^0 \quad \vee \quad H_1 : \beta_1 \neq \beta_1^0.$$

* From Theorem 5 we see that, if H_0 is true then

$$T_1 \;=\; \frac{\widehat{\beta}_1 - \beta_1^0}{\mathrm{se}(\widehat{\beta}_1)} \;\sim\; t_{n-2}.$$

* We can therefore find the observed value, t_1, of T_1 and calculate the p-value,

$$p_1 \;=\; 2\mathsf{P}\left(t_{n-2} \geqslant |t_1|\right)$$

. When p_1 is very small we reject H_0.

* We can do similar to test hypotheses about the intercept β_0.

Hypothesis Testing (ctd)

* Of particular interest are the tests of

$$H_0^0 : \beta_0 = 0 \;\; \vee \;\; H_1^0 : \beta_0 \neq 0$$

$$H_0^1 : \beta_1 = 0 \;\; \vee \;\; H_1^1 : \beta_1 \neq 0$$

* The test statistics for these two tests are the ratios of the estimates to their standard errors.

* Software generally includes the values of these test statistics as well as the associated p-values.

The ANOVA Table

* Another part of the software output is the ANOVA table.
* For example you might have

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27420	27420	943.20	<.0001
Error	12	348.84837	29.07070		
Corrected Total	13	27768			

Sums of Squares

* The sums of squares are

$$\begin{aligned} \text{Total Sum of Squares (SST)} &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\ \text{Error Sum of Squares (SSE)} &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ \text{Model Sum of Squares (SSR)} &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 \end{aligned}$$

* We can show that

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

The ANOVA Table (ctd)

* The Total degrees of freedom are always $n-1$.
* For simple regression the degrees of freedom used by the model is 1.
* Hence the Error degrees of freedom is $n-1-1=n-2$.
* The Mean Square column is the Sum of Squares divided by the degrees of freedom.
* Note that the Error Mean Square (MSE) is $\hat{\sigma}^2$.

The F test for the Model

* The F Value in the ANOVA table is the ratio of the Model Mean Square over MSE.

* If the model is not useful in predicting Y then this ratio has an F distribution with 1 and $n-2$ degrees of freedom.

* We can get a p-value for the test that the model is not useful by looking at the tail probability that an $F_{1,n-2}$ random variable is greater than the observed F value.

* In simple regression, this is equivalent to the test of $\beta_1 = 0$ against $\beta_1 \neq 0$.

The Coefficient of Determination

* The total sum of squares is a measure of the variability in $y_1, \ldots, y_n$ without taking the covariate into account.

* The error sum of squares is the amount of variability left after fitting a linear regression for the covariate.

* The model sum of squares is the amount of variability explained by the model.

* The proportion of the variability explained by the model is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

* In simple regression R^2 is the square of the sample correlation between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.

Predictions

* Suppose that x_0 is some new value of x for which we want to do prediction.

* There are two types of prediction that are of interest to us
  1. Estimation of $\mu_0 = \mathsf{E}(Y \mid X = x_0)$.
  2. Prediction of a Y value for an individual with $X = x_0$.

* We can use our fitted regression model to do both of these.

Estimation of μ_0

* By the linear model the true value of μ_0 is $\beta_0 + \beta_1 x_0$.

* An obvious estimator of μ_0 is then

$$\widehat{\mu}_0 \;=\; \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

* From Theorem 4 we see that $\widehat{\mu}_0$ is unbiased and

$$\text{Var}(\widehat{\mu}_0) \;=\; \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{Sxx}\right)\sigma^2.$$

* A confidence interval for μ_0 is

$$\widehat{\mu}_0 \;\pm\; t_{(n-2,\ \alpha/2)}\text{se}(\widehat{\mu}_0)$$

Predicting an Individual Value

* The value of Y for an individual with $X = x_0$ is

$$Y_0 \;=\; \beta_0 + \beta_1 x_0 + \varepsilon_0$$

* We can plug in the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and take ε_0 to be equal to its mean ($\mathsf{E}(\varepsilon_0) = 0$) to get the predicted value

$$\widehat{y}_0 \;=\; \widehat{\beta}_0 + \widehat{\beta}_1 x_o \;=\; \widehat{\mu}_0$$

* The Variance of $\widehat{Y}_0$, however is

$$\mathsf{Var}(\widehat{Y}_0) \;=\; \mathsf{Var}(\widehat{\mu}_0) + \mathsf{Var}(\varepsilon_0) \;=\; \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{Sxx}\right)\sigma^2.$$

* A prediction interval for an individual with $x = x_0$ is then

$$\widehat{y}_0 \;\pm\; t_{(n-2,\ \alpha/2)}\mathsf{se}(\widehat{y}_0)$$