- The R^2 statistic is a measure of the linear relationship between X and Y. Recall that $\text{Cor}(X, Y)$ is also a measure of the linear relationship between X and Y.

- It can be shown that in the simple linear regression setting, $R^2 = r^2$ where r is the sample correlation given by

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

For the Advertising data,

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Multiple Linear Regression

- Suppose we have an input vector $X^T = (X_1, X_2, ..., X_p)$ (we have p distinct predictors), and want to predict a real-valued output Y. The multiple linear regression model has the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$
$$= f(X) + \epsilon$$

where $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, X_j represents the jth predictor, and β_j (unknown coefficient) quantifies the association between that variable and the response.

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.

In the advertising example, the model becomes

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

Estimating the Regression Coefficients

- For a given set of training data $(x_1, y_1), \cdots, (x_n, y_n)$ (where each $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ is a vector of feature measurements for the ith case), the parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ are estimated using the least squares approach, in which we pick the coefficients to minimize the residual sum of squares

$$\begin{aligned} \text{RSS}(\beta) &= \sum_{i=1}^{n} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \end{aligned}$$

The values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

- We can write the residual sum of squares as

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

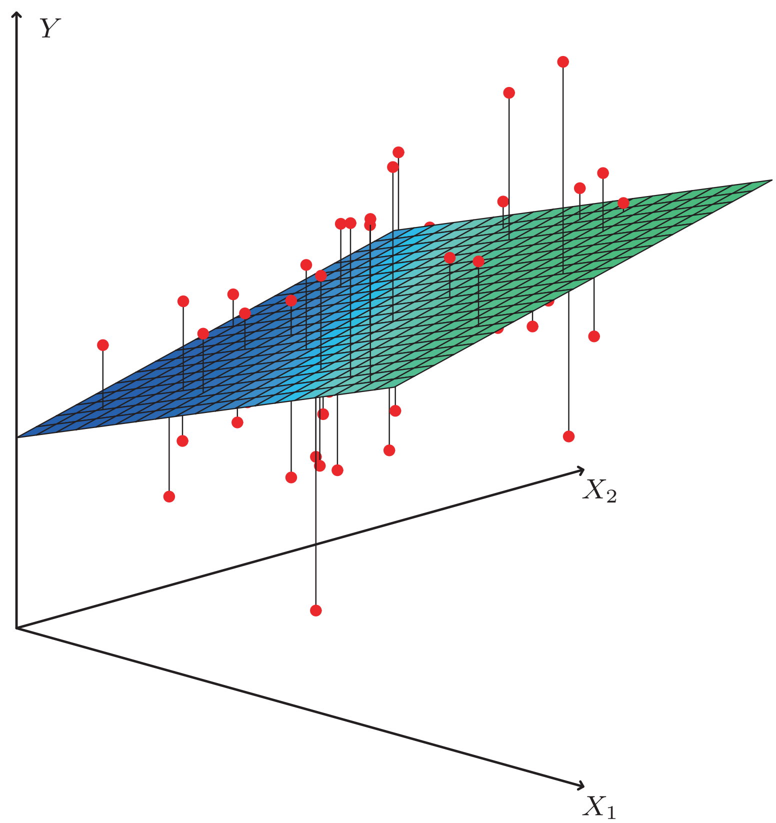
$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

$$= \|y - X\beta\|^2$$

where $\mathbf{X}$ the $n \times (p + 1)$ matrix with each row an input vector (with a 1 in the first position), and similarly let $\mathbf{y}$ be the n-vector of outputs in the training set.

- Assuming that $\mathbf{X}$ has full column rank, and hence $\mathbf{X}^T\mathbf{X}$ is positive definite, so we can show that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$



In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

- The predicted value at an input vector x_0 is given by $\hat{f}(x_0) = (1 : x_0)^T\hat{\beta}$; the fitted values at the training inputs are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

$(1 : x_0)^T = (1, x_{01}, x_{02}, \cdots, x_{0p})$

where $\hat{y}_i = \hat{f}(x_i)$.

Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, the predicted value at an input vector $x_0 = (x_{01}, x_{02}, \ldots, x_{0p})^T$ is given by

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_p x_{0p}$$

and the fitted values at the training inputs are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

Assessing the Accuracy of the Coefficient Estimates

- To draw inferences about the parameters and the model, additional assumptions are needed. We now assume that the linear model is the correct model for the mean; that is, the conditional expectation of Y is linear in $X_1, ..., X_p$. We also assume that the deviations of Y around its expectation are Gaussian. Hence

$$\begin{aligned} Y &= \mathrm{E}(Y|X_1, ..., X_P) + \epsilon \\ &= \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon \end{aligned}$$

  where the error ϵ is a Gaussian random variable with expectation zero and variance σ^2, written $\epsilon \sim N(0, \sigma^2)$.

- It can be shown that

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

  and

$$\hat{\sigma} = \sqrt{\mathrm{RSS}/(n-p-1)}$$

  In addition $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent.

For advertising data, in order to fit a multiple linear regression model using least squares, we again use the lm() function.

```
> lm_fit_full <- lm(sales ~ TV + radio + newspaper, data=Adv)
> summary(lm_fit_full)

Call:
lm(formula = sales ~ TV + radio + newspaper, data = Adv)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

Coefficients of the simple linear regression model for number of units sold on Top: TV advertising budget, Middle: radio advertising budget, and Bottom:

newspaper advertising budget.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.

Confidence Intervals

- We can isolate β_j in $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$ to obtain a $(1-\alpha)100\%$ confidence interval for β_j:

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2} v_j^{1/2} \hat{\sigma}$$

where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

For the advertising data,

```
> confint(lm_fit_full, level = 0.95)
                  2.5 %     97.5 %
(Intercept)  2.32376228 3.55401646
TV           0.04301371 0.04851558
radio        0.17154745 0.20551259
newspaper   -0.01261595 0.01054097
```

Prediction

The predicted value at an input vector x_0 is given by

$$\hat{\mu}_{Y|X_1,\ldots,X_p}(x_0) = \hat{y}_0 = (1 : x_0)^T \hat{\beta}$$

Example: What is the predicted sales for a market whose TV, radio, and newspaper advertising are 100, 30, and 40, respectively?

$\hat{y}_0 = 2.939 + 0.046\, x_{01} + 0.189\, x_{02} - 0.001\, x_{03}$

$\hat{y}_0 = 2.939 + 0.046 \times 100 + 0.189 \times 30 - 0.001 \times 40$

$= 13.169$ units

A $(1-\alpha)100\%$ confidence interval for a predicted response is

average response

$$\hat{y}_0 \pm t_{n-p-1,\alpha/2}\hat{\sigma}_{\hat{y}_0}$$

where $\hat{\sigma}_{\hat{y}_0} = \hat{\sigma}^2(1:x_0)^T(\mathbf{X}^T\mathbf{X})^{-1}(1:x_0)$

x_0 = vector we are predicting for

A $(1-\alpha)100\%$ prediction interval for a predicted response is

an individual response (future observation y_0 to be taken at x_0)

$$\hat{y}_0 \pm t_{n-p-1,\alpha/2}\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2_{\hat{y}_0}}$$

For the advertising data,

```
> predict(lm_fit_full,data.frame(TV=(c(20,50,100)), radio=(c(05,10,20)),
newspaper=(c(10,0,0))), interval ="confidence", level = 0.95)
        fit       lwr       upr
1  4.786457  4.273090  5.299825
2  7.112422  6.628395  7.596448
3 11.285954 10.859077 11.712832

> predict(lm_fit_full,data.frame(TV=(c(20,50,100)), radio=(c(05,10,20)),
newspaper=(c(10,0,0))), interval ="prediction", level = 0.95)
        fit      lwr       upr
```

```
1  4.786457 1.422984  8.149931
2  7.112422 3.753302 10.471542
3 11.285954 7.934592 14.637317
```

Hypothesis Testing

- To test the hypothesis that a particular coefficient $\beta_j = 0$:

$$H_0: \ \beta_j = 0$$

versus

$$H_a: \ \beta_j \neq 0,$$

that is,

$$H_0: \text{There is no relationship between } \ X_j \ \text{ and } \ Y$$

versus

$$H_a: \text{There is some relationship between } \ X_j \ \text{ and } \ Y$$

we compute a t-*statistic*, given by

$$t = \frac{\hat{\beta}_j - 0}{\hat{\sigma}\sqrt{v_j}} \quad , \ SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{v_j}$$

This will have a t-distribution with $n - p - 1$ degrees of freedom, assuming $\beta_j = 0$.
Using statistical software, it is easy to compute the p-value, the probability of observing any value equal to $|t|$ or larger.

- In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = ... = \beta_p = 0$. We test the null hypothesis,

$$H_0: \ \beta_1 = \beta_2 = ... = \beta_p = 0$$

versus

$$H_a: \ \text{at least one } \beta_j \text{ is non-zero}$$

This hypothesis test is performed by computing the F-statistic,

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

where as simple linear regression, $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Under the Gaussian assumptions, and the null hypothesis, the F statistic will have a $F_{p,n-p-1}$ distribution.

- When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, we expect F to be greater than 1.

- But how large does the F-statistic need to be before we can reject H_0 and conclude that there is a relationship?

It depends on the values of n and p.

when n is large, an F-statistic that is just a little larger than 1 might still provide evidence against H_0. In contrast, a larger F-statistic is needed to reject H_0 if n is small.

Alternatively, we can use the p-value associated with the F-statistic to make the decision.

For advertising data,

```
> summary(lm_fit_full)

Call:
lm(formula = sales ~ TV + radio + newspaper, data = Adv)
```