

# Simple Linear Regression

## Parameters

The simple model is

$$y_i = \beta_0 + \beta_1 x_i$$

with  $E(y_i) = \beta_0 + \beta_1 x_i$ ,  $\text{Var}(y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \epsilon) = \sigma^2$

Estimates for  $\beta_0, \beta_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n k_i y_i = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^n y_i (x_i - \bar{x})$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^n k_i^2$$

Estimation on  $\sigma^2$

$$SSE = \sum_{i=1}^n e_i^2 = (y_i - \hat{y}_i)^2$$

$SSE$  has  $n - 2$  degrees of freedom.

$$\hat{\sigma}^2 = \frac{SSE}{n - 2} = MSR$$

## Hypothesis Testing

Testing on the slope for a constant  $\beta$

$$H_0 : \hat{\beta}_1 = \beta, H_1 : \hat{\beta}_1 \neq \beta$$

If  $\sigma^2$  is known,

$$Z_0 = \frac{\hat{\beta}_1 - \beta}{\sqrt{\sigma^2 / S_{xx}}}$$

If  $\sigma^2$  is unknown,

$$t_0 = \frac{\hat{\beta}_1 - \beta}{\sqrt{MSE / S_{xx}}}$$

We reject the null hypothesis  $|t_0| > t_{\alpha/2, n-2}$ . We test the intercept similarly,

$$H_0 : \beta_0 = \beta, H_1 : \beta_0 \neq \beta$$

$$t_0 = \frac{\beta_0 - \beta}{se(\hat{\beta}_0)}$$

where  $se^2(\hat{\beta}_1) = \frac{MSE}{S_{xx}}$ ,  $se^2(\hat{\beta}_0) = MSE(1/n + \bar{X}^2 / S_{xx})$

## Significance of Regression

We test significance with

$$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$$

Using the same statistic,  $|t_0| > t_{\alpha/2, n-2}$ .

Source	SS	DF
Regression	$SSR = \hat{\beta}_1^2 \sum (X_i - \bar{X})^2$	$p - 1$
Error	$SSE = \sum (Y_i - \hat{Y}_i)^2$	$n - p$
Total	$SSTO = \sum (Y_i - \bar{Y})^2$	$n - 1$

MS=SS/df	E(MS)	F
MSR	$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$	$MSE/MSR$
MSE	$\sigma^2$	

## Confidence Intervals

The confidence interval on the slope  $\beta_1$ ,

$$\hat{\beta}_1 - t_{\alpha/2, n-2} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} se(\hat{\beta}_1)$$

For the intercept  $\beta_0$ ,

$$\hat{\beta}_0 - t_{\alpha/2, n-2} se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} se(\hat{\beta}_0)$$

$$\text{For } \sigma^2, \quad \frac{(n-2)MSE}{\chi_{\alpha/2, n-2}^2} \leq \sigma^2 \leq \frac{(n-2)MSE}{\chi_{1-\alpha/2, n-2}^2}$$

## Interval Estimation on Mean Response

An unbiased estimator for  $E(y|x_0)$  for a value of regressor  $x = x_0$  is

$$E(\hat{y}|x_0) = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

The variance is

$$\text{Var}(\hat{\mu}_{y|x_0}) = \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}}$$

The sampling distribution for

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MSE(1/n + (x_0 - \bar{x})^2 / S_{xx})}} \sim t_{n-2}$$

So the confidence interval is then

$$\left[ \hat{\mu}_{y|x_0} \pm t_{\alpha/2, n-2} \sqrt{MSE \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

## Correlation

The coefficient of determination is

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

The pearson correlation coefficient is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}$$

When applied to a sample,

$$\begin{aligned} r &= b_1 \left( \frac{S_{xx}}{SST} \right)^{\frac{1}{2}} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ &= \frac{S_{xy}}{(S_{xx} SST)^{1/2}} \end{aligned}$$

If we want to test  $\rho = 0$ ,

$$H_0 : \rho = 0, H_1 : \rho \neq 0$$

We use the  $t$  statistic,

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

We reject the null hypothesis  $H_0 : \rho = 0$  if  $|t_0| > t_{\alpha/2, n-2}$ . To test  $\rho = \rho_0$ ,

$$H_0 : \rho = \rho_0, H_1 : \rho \neq \rho_0$$

Use the  $Z$  statistic,

$$Z = \text{arctanh } r = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) \sim N \left( \mu_z, \frac{1}{n-3} \right)$$

where

$$\mu_z = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right)$$

Then we standardize it to test

$$Z_0 = (\text{arctanh}(r) - \text{arctanh}(\rho_0)) \sqrt{n-3}$$

We can obtain our confidence interval with

$$\left[ \tanh \left( \text{arctanh}(r) \pm \frac{Z_{\alpha/2}}{\sqrt{n-3}} \right) \right]$$

where  $\tanh(u) = (e^u - e^{-u}) / (e^u + e^{-u})$ . We reject  $H_0 : \rho = \rho_0$  if  $|Z_0| > Z_{\alpha/2}$ .

# Multiple Linear Regression

## Model

We write the multiple linear regression model as

$$Y = X\beta + \epsilon$$

where (Note  $p = k + 1$ .)

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$Y$  is  $n \times 1$ ,  $X$  is  $n \times p$ ,  $\beta$  is  $p \times 1$ , and  $\epsilon$  is  $n \times 1$ . In matrix form, we get the fitted line

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$$

$H = X(X'X)^{-1}X'$  is the **hat matrix**.

### Properties of the Hat Matrix

- $H$  is a projection matrix, so it is idempotent and symmetric  $HH = H$ ,  $H' = H$ .
- The matrix  $H$  is orthogonal to the matrix  $I - H$ , so  $(I - H)H = H - HH = 0$ . Moreover,  $(I - H)$  is idempotent and a project matrix as well.
- The vector of residuals is

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

- $Y$  is projected onto a space spanned by the columns of  $H$ , and the residuals are in an orthogonal space.

$$Y = HY + (I - H)Y$$

### Estimation of $\sigma^2$

Residual sum of squares is

$$SSE = \sum_{i=1}^n e_i^2 = e'e = Y'Y - \hat{\beta}'X'Y$$

$SSE$  has  $n - p$  degrees of freedom, then  $MSE$  is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n - p}$$

## Estimation and Hypothesis Testing

### Testing for Significance

We test for significance with

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0, H_1 : \beta_j \neq 0$$

Rejecting the null hypothesis means at least one regressor contributed significantly. We use an  $F$  statistic

$$F_0 = \frac{SSR/k}{SSE/(n - p)} = \frac{MSR}{MSE} \sim F_{k, n-p}$$

We reject the null hypothesis when  $F_0 > F_{\alpha, k, n-k-1}$ .

The **total sum of squares** is

$$SST = \sum_{i=1}^n Y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n Y_i \right)^2 = Y'Y - \frac{1}{n} \left( \sum_{i=1}^n Y_i \right)^2$$

The **regression sum of squares** is

$$SSR = \hat{\beta}'X'Y - \frac{1}{n} \left( \sum_{i=1}^n Y_i \right)^2$$

The **residual sum of squares** is

$$SSE = Y'Y - \hat{\beta}'X'Y = Y'(I - H)Y$$

We can also write  $SST$  and  $SSR$  in terms of the  $J_n$ , and  $n \times n$  matrix with 1's.

$$SST = Y' \left( I - \frac{1}{n} J_n \right) Y$$

$$SSR = Y' \left( H - \frac{1}{n} J_n \right) Y$$

### Tests on Individual Coefficients

To test an individual coefficient  $\beta_j$ , we use

$$H_0 : \beta_j = 0, H_1 : \beta_j \neq 0$$

The test statistic is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Where  $C_{jj}$  is the diagonal entry of  $(X'X)^{-1}$ . We reject  $H_0$  when  $|t_0| > t_{\alpha/2, n-p}$ .

If we fail to reject the null hypothesis, we can remove the corresponding regressor  $x_j$  from the model.

## Extra Sum Of Squares

We want to partition  $r$  of the  $k$  regressors to test

$$H_0 : \beta_2 = 0, H_1 : \beta_2 \neq 0$$

$Y = X\beta + \epsilon$ , where  $Y$  is  $n \times 1$ ,  $X$  is  $n \times p$ ,  $\beta$  is  $p \times 1$ , and  $\epsilon$  is  $n \times 1$  with  $p = k + 1$ .

### Full Model

$$Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

$X_1$  is  $n \times (p - r)$ ,  $X_2$  is  $n \times r$ .

$$\hat{\beta} = (X'X)^{-1}X'Y, SSR(\beta) = \hat{\beta}'X'Y$$

which has  $k = p - 1$  degrees of freedom.

### Reduced Model

To test regressors in  $\beta_2$ , fit the model assuming  $H_0 : \beta_2 = 0$  is true.

$$Y = X_1\beta_1 + \epsilon$$

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y, SSR(\beta_1) = \hat{\beta}_1'X_1'Y$$

which has  $k - r = p - 1 - r$  degrees of freedom. The sum of squares due to  $\beta_2$  given that  $\beta_1$  is already in the model is

$$SSR(\beta_2|\beta_1) = SSR(\beta) - SSR(\beta_1)$$

The null hypothesis  $\beta_2 = 0$  can be tested with (partial  $F$ -test)

$$F_0 = \frac{SSR(\beta_2|\beta_1)/r}{MSE}$$

If  $F_0 > F_{\alpha, r, n-p}$ , then we reject the null hypothesis and conclude that at least one regressor in  $X_2$  contributes.

## Lack of Fit

**Pure Error Sum of Squares:**

$$SS_{PE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

**Sum of Squares Due to Lack of Fit:**

$$SS_{LOF} = \sum_{i=1}^m n_i (\bar{y} - \hat{y}_i)$$

**$F$ -Statistic:**

$$F^* = \frac{SS_{LOF}/(m - 2)}{SS_{PE}/(n - m)} = \frac{MS_{LOF}}{MS_{PE}}$$

### Testing Lack of Fit

If the regression is linear, then  $E(y_i) = \beta_0 + \beta_1 x_i$ ,

$$H_0 : E(y_i) = \beta_0 + \beta_1 x_i, H_1 : E(y_i) \neq \beta_0 + \beta_1 x_i$$

Reject the null hypothesis when  $F^* > F_{\alpha, m-2, n-m}$ .

## Anova Table for Lack of Fit

Source	Sum of Squares	DF
Regression	$SSR = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	1
Residuals	$SSE(R) = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	$n - 2$
Lack of Fit	$SS_{LOF} = \sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2$	$m - 2$
Pure Error	$SS_{PE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$n - m$
Total	$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	$n - 1$

Source	Mean Square = SS/df	F-Statistic
Regression	$SSR/1$	$MSR/MSE$
Residuals	$SSE(R)/n - 2$	
Lack of Fit	$SS_{LOF}/m - 2$	$MS_{LOF}/MS_{PE}$
Pure Error	$SS_{PE}/n - m$	

## Model Adequacy

### Normality

- **Using a boxplot:** Box plot of residuals should be symmetric around a median of 0.
- **Histogram:** Should be of the shape of a normal distribution.
- **QQ-Plot:** Plot  $E_k = \sqrt{MSE} \cdot \Phi^{-1} \left( \frac{k-0.375}{n+0.25} \right)$  vs the residuals  $e_{(k)}$ , should be a straight line.

### Constant Variance

Studentize the residuals, and plot  $\sqrt{e_i^*}$  vs  $\hat{Y}_i$ .

$$e_i^* = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

- Plot should show a random distribution of points. Otherwise, signs of non-constant variance.
- Residuals lie in a narrow band around 0  $\Rightarrow$  no need of correction.
- Residuals are increasing or decreasing  $\Rightarrow$  variance is non constant.
- Double-bow pattern  $\Rightarrow$  variance in the middle is larger than the variance at the extremes.
- Quadratic relationship (parabola shape)  $\Rightarrow$  maybe a nonlinear relationship

## Confidence Intervals

### Confidence Intervals on Regression Coefficients

To construct a confidence interval on  $\beta_j$ , use the statistic

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}$$

The CI is then

$$\hat{\beta}_j - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}}$$

Recall  $C_{jj}$  is the  $j$ th diagonal entry of  $(X'X)^{-1}$  the standard error is

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

### Confidence Interval on Mean Response

To construct confidence intervals at points  $x_{01}, x_{02}, \dots, x_{0k}$ , define

$$x_0 = [1 \quad x_{01} \quad x_{02} \quad \dots \quad x_{0k}]^T$$

The fitting value is then

$$\hat{y}_0 = x_0' \hat{\beta}$$

This is an unbiased estimator,  $E(y|x_0) = x_0' \beta = E(\hat{y}_0)$ , and  $\text{Var}(\hat{y}_0) = \sigma^2 x_0' (X'X)^{-1} x_0$ . The CI is then

$$[\hat{y}_0 \pm t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 x_0' (X'X)^{-1} x_0}]$$

### Simultaneous Confidence Interval

**Theorem** (Bonferroni Inequality). For two events  $A_1, A_2$ , we have that

$$P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$$

From DeMorgan's identity, we also have

$$P(A_1^c \cap A_2^c) = 1 - P(A_1 \cup A_2) \geq 1 - P(A_1) - P(A_2)$$

If we define the events

$$A_1^c: \hat{\beta}_0 \pm t_{1-\alpha/2, n-2} s(\hat{\beta}_0)$$

$$A_2^c: \hat{\beta}_1 \pm t_{1-\alpha/2, n-2} s(\hat{\beta}_1)$$

From Bonferroni's Inequality, if we have  $P(A_1) = P(A_2) = \alpha$ , then

$$P(A_1^c \cap A_2^c) \geq 1 - P(A_1) - P(A_2) = 1 - 2\alpha$$

In general, if we have  $p$  parameters and each confidence interval has confidence,  $1 - \frac{\alpha}{p}$ , then

$$P\left(\bigcap_{i=1}^p A_i^c\right) \geq 1 - p \frac{\alpha}{p} = 1 - \alpha$$

## Transformations and Weighting

### Variance Stabilizing Transformations

- **Poisson** ( $\mu = \sigma^2$ ):  $y \sim \text{Poisson}(\lambda) \Rightarrow \sqrt{y}$  is nearly normal and has variance 1/4 if  $\lambda$  is large.
- **Binomial**:  $y \sim \text{Bin}(n, p)$  with mean  $m = np$ , then

$$y' = \sin^{-1} \left( \sqrt{\frac{y+c}{n+2c}} \right)$$

The optimal value of  $c$  is 3/8 when  $m$  and  $n-m$  are large. The variance is approximately  $\frac{1}{4} (n + \frac{1}{2})^{-1}$ .

### Transformations to Linearize Models.

- **Exponential**:  $\beta'_0 = \ln \beta_0$ ,  $\epsilon' = \ln \epsilon$ ,  
 $y = \beta_0 e^{\beta_1 x} \epsilon \rightarrow y' = \ln y = \beta'_0 + \beta_1 x + \epsilon'$
- **Reciprocal**:  $x' = \frac{1}{x}$ ,  
 $y = \beta_0 + \beta_1 \frac{1}{x} + \epsilon \rightarrow y = \beta_0 + \beta_1 x' + \epsilon$   
 $\frac{1}{y} = \beta_0 + \beta_1 x + \epsilon \rightarrow y' = \frac{1}{y}$
- **Two Step Reciprocal**:  $y' = \frac{1}{y}$ ,  $x' = \frac{1}{x}$ ,  
 $y = \frac{x}{\beta_0 + \beta_1 x} \rightarrow y' = \beta_0 x' + \beta_1$

### Box-Cox Transformations

When data is not normally distributed, can apply a power transformation

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda y^{\lambda-1}} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}, \quad \hat{y} = \ln^{-1} \left( \frac{1}{n} \sum_{i=1}^n \ln y_i \right)$$

We want a value for  $\lambda$  that minimizes  $SSE$ , this value is found by trial and error.

### Weighted Least Squares

$$W = \begin{bmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_n \end{bmatrix}$$

$$X_W = \begin{bmatrix} 1\sqrt{w_1} & \dots & x_{1k}\sqrt{w_1} \\ 1\sqrt{w_2} & \dots & x_{2k}\sqrt{w_2} \\ \vdots & \ddots & \vdots \\ 1\sqrt{w_n} & \dots & x_{nk}\sqrt{w_n} \end{bmatrix}, \quad Y_W = \begin{bmatrix} y_1\sqrt{w_1} \\ y_2\sqrt{w_2} \\ \vdots \\ y_n\sqrt{w_n} \end{bmatrix}$$

**New Weighted Model:**  $Y_w = X_w \beta + \epsilon_w$ , estimate becomes

$$\hat{\beta} = (X_W' X_W)^{-1} X_W' Y_W = (X' W X)^{-1} X' W Y$$

Weighted mean square error is

$$MSE_W = \frac{\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2}{n - p} = \frac{\sum_{i=1}^n w_i e_i^2}{n - p}$$