Multiple Regression I

Paweł Polak

October 16, 2017

Linear Regression Models - Lecture 7

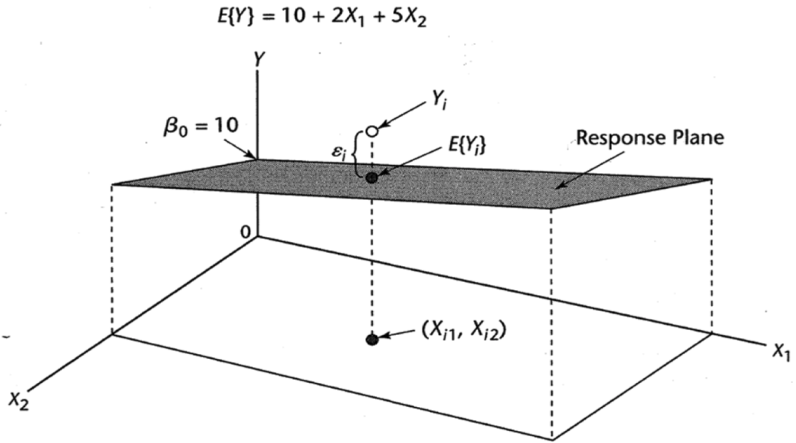
Multiple Regression

- Multiple regression is one of the most widely used tools in statistical analysis
- Matrix expressions for multiple regression are the same as for simple linear regression
- Often the response is best understood as being a function of multiple input quantities.
- Examples:
  - Spam filtering-regress the probability of an email being a spam message against thousands of input variables.
  - Revenue prediction - regress the revenue of a company against a lot of factors.

First-Order with Two Predictor Variables

- When there are two predictor variables X_1 and X_2 the regression model
$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$$
is called a first-order model with two predictor variables.
- A first order model is linear in the predictor variables.
- $X_{i,1}$ and $X_{i,2}$ are the values of the two predictor variables in the ith trial.
- Assuming noise equal to zero in expectation
$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
- The form of this regression function is of a plane, e.g.,
$$\mathbb{E}(Y) = 10 + 2X_1 + 5X_2$$

Example



$E\{Y\} = 10 + 2X_1 + 5X_2$
Y
Y_i
$\beta_0 = 10$
ε_i
$E\{Y_i\}$
Response Plane
0
(X_{i1}, X_{i2})
X_1
X_2

Meaning of the Coefficients

- β_0 is the intercept when both X_1 and X_2 are zero;
- β_1 indicates the change in the mean response $\mathbb{E}(Y)$ per unit increase in X_1 when X_2 is held constant
- β_2 -vice versa
- Example: fix $X_2 = 2$

$$\mathbb{E}(Y) = 10 + 2X_1 + 5(2) = 20 + 2X_1, \text{ where } X_2 = 2$$

intercept changes but clearly linear

Terminology & Comments

- When the effect of X_1 on the mean response does not depend on the level X_2 (and vice versa) the two predictor variables are said to have additive effects or not to interact.
- The parameters β_1 and β_2 are sometimes called *partial regression coefficients*. They represents the partial effect of one predictor variable when the other predictor variable is included in the model and is held constant.
- The response surface may not always be appropriate, but even when not it is often a good approximate descriptor of the regression function in "local" regions of the input space.
- The meaning of the parameters can be determined by taking partials of the regression function w.r.t. to each X_i.

First order model with > 2 predictor variables

Let there be $P-1$ predictor variables, then

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{P-1} X_{i,P-1} + \varepsilon_i$$

which can also be written as

$$Y_i = \beta_0 + \sum_{k=1}^{P-1} \beta_k X_{i,k} + \varepsilon_i$$

and if $X_{i,0} = 1$, then it also can be written as

$$Y_i = \sum_{k=0}^{P-1} \beta_k X_{i,k} + \varepsilon_i$$

where $X_{i,0} = 1$.

- In this setting the response surface is a hyperplane.
- This is difficult to visualize but the same intuitions hold.
  - Fixing all but one input variables, each β_k tells how much the response variable will grow or decrease according to that one input variable.

General Linear Regression Model

We have arrived at the general regression model. In general the $X_1, \ldots, X_{P-1}$ variables in the regression model do not have to represent different predictor variables, nor do they have to all be quantitative(continuous).

The general model is

$$Y_i = \sum_{k=0}^{P-1} \beta_k X_{i,k} + \varepsilon_i,$$

where $X_{i,0} = 1$. with response function when $\mathbb{E}(\varepsilon_i) = 0$ is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{P-1} X_{P-1}.$$

Qualitative(Discrete) Predictor Variables

- Until now we have (implicitly) focused on quantitative (continuous) predictor variables.
- Qualitative (discrete) predictor variables often arise in the real world.
- Examples:
  - Patient sex: male/female
  - College Degree: yes/no

Example

Regression model to predict the length of hospital stay (Y) based on the age (X_1) and gender(X_2) of the patient. Define gender as:

$$X_2 = \begin{cases} 1 & \text{if patient female} \\ 0 & \text{if patient male} \end{cases}$$

And use the standard first-order regression model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i.$$

where $X_{i,1}$ is patient's age, and $X_{i,2}$ is patient's gender If $X_2 = 0$, the response function is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1$$

Otherwise, it's

$$\mathbb{E}(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$$

Which is just another parallel linear response function with a different intercept.

Polynomial Regression

- Polynomial regression models are special cases of the general regression model.
- They can contain squared and higher-order terms of the predictor variables
- The response function becomes curvilinear.
- For example

  $$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

  which clearly has the same form as the general regression model (take $X_{i,1} = X_i$ and $X_{i,2} = X_i^2$).

General Regression

- Transformed variables

$$logY \quad \text{or} \quad 1/Y$$

- Interaction effects

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \varepsilon_i$$

- Combinations

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}^2 + \beta_3 X_{i,2} + \beta_4 X_{i,1} X_{i,2} + \varepsilon_i$$

- Key point-all linear in parameters, i.e., it is of the form

$$Y_i = c_{i,0}\beta_0 + c_{i,1}\beta_1 + c_{i,2}\beta_2 + c_{i,3}\beta_3 + \ldots + c_{i,P-1}\beta_{P-1} + \varepsilon_i$$

- An example of a nonlinear regression model is the following:

$$Y_i = \beta_0 e^{\beta_1 X_i} + \varepsilon_i$$

General Regression Model in Matrix Terms

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}_{N\times 1} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_N^T \end{pmatrix}_{N\times P} = \begin{pmatrix} X_{1,0} X_{1,1} \dots X_{1,P-1} \\ X_{2,0} X_{2,1} \dots X_{2,P-1} \\ \vdots \\ X_{N,0} X_{N,1} \dots X_{N,P-1} \end{pmatrix}_{N\times P}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_{P-1} \end{pmatrix}_{P\times 1} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}_{N\times 1}$$

Least Squares estimation

- The *linear model* is usually written as (in vector notation)

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_N^T \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{P-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix},$$

where $E(\varepsilon_i) = 0$, and $Cov(\varepsilon_i, \varepsilon_j) = 0$, if $i \neq j$.

- We can use the method of *least squares* to estimate $\boldsymbol{\beta}$.

Least Squares estimation

- The *least squares estimator* of β minimizes

$$Q = \|\mathbf{Y} - \mathbf{X}\beta\|^2, \text{ where } \|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \sum_{i=1}^{N} x_i^2$$

- In matrix notation we write this as:

$$\begin{aligned} Q &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}^T\mathbf{Y} - \beta^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta \\ &= \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \end{aligned}$$

- Derivative rules:

$$\frac{\partial}{\partial \mathbf{x}^T}(\mathbf{A}\mathbf{x}) = \mathbf{A}, \quad \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}^T) = \mathbf{A}^T, \quad \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x},$$

- To find the value that minimizes Q, we *differentiate* with respect to β and set the results equal to *zero*.

- Then, $\frac{\partial}{\partial \boldsymbol{\beta}} Q = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta$

Normal Equations

- The normal equations can be expressed in matrix notation as
$$\mathbf{X}^T\mathbf{X}\boldsymbol{b} = \mathbf{X}^T\mathbf{Y}.$$
- The least squares estimators are given by
$$\boldsymbol{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$
- We will show that this estimator coincides with the MLE when the errors are *normally* distributed.
- Thus the estimator $\boldsymbol{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ may be justified even when the normality assumption is uncertain.

Hat matrix

- The vector of the *fitted* values can be written:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

- We can re-express $\hat{\mathbf{Y}}$ as follows:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

  where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

- The matrix $\mathbf{H}$ is called the *hat matrix*.
- $\mathbf{H}$ is symmetric, i.e., $\mathbf{H}^T = \mathbf{H}$.
- $\mathbf{H}$ is *idempotent*, i.e., $\mathbf{H}\mathbf{H} = \mathbf{H}$.
- $(\mathbf{I} - \mathbf{H})$ is also idempotent.

Residuals

- The vector of *residuals* can be computed as follows:

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

- Expected value of $\mathbf{e}$:

$$\mathbb{E}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\ \mathbb{E}(\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\mathbf{X}\beta = \mathbf{X}\beta - \mathbf{X}\beta = 0$$

- Variance-covariance matrix of $\mathbf{e}$:

$$\sigma^2(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2(\mathbf{Y})(\mathbf{I} - \mathbf{H})^T = (\mathbf{I} - \mathbf{H})\sigma^2(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H})$$

Estimating σ^2

- As in the simple linear regression case, we can estimate σ^2 using the residuals, i.e.,

$$s^2 = \frac{\mathbf{e}^T\mathbf{e}}{N-P} = \frac{(\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}})}{N-P} = \frac{SSE}{N-P} = MSE$$

Degrees of Freedom

- The term $(N-P)$ in s^2 is the number of *degrees of freedom* associated with the estimate.
- To find s^2 we must first estimate P parameters, which results in a loss of P degrees of freedom.
- Using $(N-P)$ makes s^2 an *unbiased* estimate of σ^2.

ANOVA

- We can perform an equivalent ANOVA sums of square decomposition in multiple regression:

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2.$$

- Thus,

$$SST = SSE + SSR,$$

where

$$SST = \mathbf{Y}^T(\mathbf{I}-\mathbf{J}/N)\mathbf{Y}, \quad SSE = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}, \quad SSR = \mathbf{Y}^T(\mathbf{H}-\mathbf{J}/N)\mathbf{Y}.$$

- As usual SST has $(N-1)$ degrees of freedom associated with it.
- The term SSE has $(N-P)$ degrees of freedom.
- The term SSR has $(P-1)$ degrees of freedom.

F Test for Regression Relation

- To test whether there is a regression relation between the response variable Y and the set of X variables, i.e., to choose between the alternatives

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_{P-1} = 0$$
$$H_1: \quad \text{not all} \quad \beta_k \ (k = 1, \ldots, P-1) \text{ equal zero}$$

  we use the test statistic

$$F^* = \frac{MSR}{MSE}$$

  The corresponding decision rule is

$$\text{If } F^* \leq F(1-\alpha; P-1, N-P) \text{ conclude } H_0$$

$$\text{If } F^* > F(1-\alpha; P-1, N-P) \text{ conclude } H_1$$

- The existence of a regression relation by itself does not ensure that useful predictions can be made by using it.
- Note that when $P - 1 = 1$, this test reduces to the F test for testing in simple linear regression whether or not $\beta_1 = 0$.

Multiple Determination

- The coefficient of *multiple determination* is

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

- This is the proportion of the variability in Y that is explained by the explanatory variables in the multiple linear regression model.

- It provides a measure of how *well* the model fits the data.

- Adding *additional* explanatory variables to the regression model will *always* lead to an increase in the value of R^2.

- Since R^2 can be made large by including more (and sometimes unimportant) explanatory variables, it is sometimes *modified* to *adjust* for the *number of variables* included in the model.

- This allows us to balance model *parsimony* with explanatory power.

Adjusted R^2

- The *adjusted* coefficient of multiple determination, uses the mean squares instead of the sums of square, i.e.,

$$R_a^2 = 1 - \frac{MSE}{MST} = 1 - \left(\frac{N-1}{N-P}\right)\frac{SSE}{SST}.$$

- Since the term includes the number of model parameters, P, it *penalizes* for model complexity.

- The coefficient of multiple correlation R is the positive square root of R^2.

- When there is only one explanatory variable, R equals in absolute value the correlation coefficient r.

Multiple Linear Regression Model - Different Perspective

- The *multiple linear regression* model is usually written as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

(in vector notation) where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{P-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix},$$

where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, for $i = 1, 2, \ldots, N$.

- Unknown *parameter* vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{P-1})^T$, where $P < N$.

- We can look at the multiple linear regression model as collection of *independent responses* of the form $Y_i \sim N(\mu_i, \sigma^2)$, where

$$\mu_i = \mathbf{X}_i^T \boldsymbol{\beta}$$

for some known vector of *explanatory* variables $\mathbf{X}_i^T = (X_{i1}, \ldots, X_{ip})$.

Maximum Likelihood Estimation

- Sometimes this is written in the more compact notation

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_N).$$

- It is usual to assume that the $N \times P$ matrix $\mathbf{X}$ has *full rank* P, (i.e., lack of *multicollinearity in the predictor variables).*

- *The likelihood for* (β, σ^2) *is*

$$L(\beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \mathbf{X}_i^T\beta)^2\right]$$

Maximum Likelihood Estimation

- The log–likelihood for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2$$
$$= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \sum_{j=0}^{P-1}X_{ij}\beta_j\right)^2.$$

- The MLE $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ satisfies

$$0 = \frac{\partial}{\partial\beta_j}\ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{N}X_{ij}(Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\beta}}), \text{ for } j = 0, \ldots, P-1,$$
$$\sum_{i=1}^{N}X_{ij}X_i^T\hat{\beta}_j = \sum_{i=1}^{N}X_{ij}Y_i \text{ for } j = 0, \ldots, P-1,$$

so

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}.$$

Maximum Likelihood Estimation

- Since $\mathbf{X}^T\mathbf{X}$ is *non–singular* (it is a square and full rank) if $\mathbf{X}$ has rank P, we have
$$\hat{\beta} = \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$
- From $\frac{\partial}{\partial\sigma^2}\ell(\hat{\beta}, \hat{\sigma}^2) = 0$, it follows that
$$\hat{\beta} \sim N_P(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}),$$
$$\hat{\sigma}^2 = \frac{1}{N}\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbf{X}_i^T\hat{\beta})^2$$
and $\hat{\sigma}^2 \sim \frac{\sigma^2}{N}\chi^2_{N-P}$ (because $SSE = \mathbf{e}'\mathbf{e} \sim \sigma^2\chi^2_{N-P}$).
- We can also show that $\hat{\beta}$ and $\hat{\sigma}^2$ are *independent*.

Proof:
$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \varepsilon)\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\varepsilon\\ &= \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\end{aligned}$$

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

Maximum Likelihood Estimation

- These results can be used to obtain an exact $(1-\alpha)$–level confidence region for β: the distribution of $\hat{\beta}$ implies that

$$\frac{1}{\sigma^2}(\hat{\beta}-\beta)^T(\mathbf{X}^T\mathbf{X})(\hat{\beta}-\beta) \sim \chi^2_P.$$

- Let

$$MSE = \frac{1}{N-P}\|\mathbf{Y}-\mathbf{X}\hat{\beta}\|^2 \sim \frac{\sigma^2}{N-P}\chi^2_{N-P},$$

so that $\hat{\beta}$ and MSE are still *independent*.

- Let $F_{P,N-P}(\alpha)$ denote the upper α–point of the $F_{P,N-P}$ distribution,

$$1-\alpha = P_{\boldsymbol{\beta},\sigma^2}\left(\frac{\frac{1}{P}(\hat{\beta}-\beta)^T(\mathbf{X}^T\mathbf{X})(\hat{\beta}-\beta)}{MSE} \leq F_{P,N-P}(\alpha)\right)$$

Thus, a $(1-\alpha)$–level confidence set for β is

$$\left\{\beta \in \mathbb{R}^P : \frac{\frac{1}{P}(\hat{\beta}-\beta)^T(\mathbf{X}^T\mathbf{X})(\hat{\beta}-\beta)}{MSE} \leq F_{P,N-P}(\alpha).\right\}$$

Inference on individual coefficients

- When the *MSE* is substituted for σ^2 we obtain the *estimated* variance-covariance matrix of $\mathbf{b}$, i.e.,

$$s^2\{\mathbf{b}\} = MSE\ (\mathbf{X}^T\mathbf{X})^{-1}.$$

- From $s^2\{\mathbf{b}\}$ we can obtain the values $s^2\{b_0\}$, $s^2\{b_1\}$, etc. needed for inference.

- The *studentized* statistic for b_k is given by

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim t_{N-P}.$$

- Thus, $(1-\alpha)$ confidence limits for b_k are

$$b_k \mp t_{1-\alpha/2;N-P}\, s\{b_k\}.$$

Inference on individual coefficients

- To test the *significance* of *individual* regression coefficients, $H_0 : \beta_k = 0$, we use a test based on

$$t = \frac{b_k}{s\{b_k\}},$$

with P-values calculated from the t_{N-P} distribution.

- Tests on individual regression coefficients tell us whether there is a significant *improvement* in our ability to predict Y by adding X_k to a model which already *includes* the other explanatory variables.

- It does *not* tell us anything about whether X_k would be useful for *predicting* Y in a multiple regression model with a *different* set of explanatory variables.

Mean response

- We can now find the *mean response* at a given vector $\mathbf{X}_h$ of explanatory variables, i.e., $\mathbf{X}_h^T\beta$.
- The *estimated* mean response can then be expressed: $\hat{Y}_h = \mathbf{X}_h^T\boldsymbol{b}$.
- The estimator is *unbiased*, i.e., $\mathbb{E}(\hat{Y}_h) = \mathbf{X}_h^T\mathbb{E}(\boldsymbol{b}) = \mathbf{X}_h^T\beta$.
- Its variance can be written:

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h.$$

- The *estimated variance* is given by $s^2\{\hat{Y}_h\} = MSE\ \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h$.
- The $(1-\alpha)$ confidence limits for $\mathbb{E}(\hat{Y}_h)$ is

$$\hat{Y}_h \mp t_{1-\alpha/2;N-P}\ s\{\hat{Y}_h\}.$$

Prediction

- Suppose that there is another pair $(\mathbf{X}_{h(\text{new})}, Y_{h(\text{new})})$, independent of $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_N, Y_N)$, satisfying the relationship

$$Y_{h(\text{new})} = \mathbf{X}_{h(\text{new})}^T \beta + \varepsilon_{h(\text{new})}, \qquad \text{where } \varepsilon_{h(\text{new})} \sim N(0, \sigma^2).$$

- We suppose that $\mathbf{X}_{h(\text{new})}$ is *known*, and attempt to estimate $Y_{h(\text{new})}$.

- We may estimate $Y_{h(\text{new})}$ by $\hat{Y}_{h(new)} = \mathbf{X}_{h(new)}^T \hat{\beta}$.

- Notice that

$$\hat{Y}_{h(\text{new})} - Y_{h(\text{new})} = \mathbf{X}_{h(\text{new})}^T (\hat{\beta} - \beta) - \varepsilon_{h(\text{new})}$$

$$\sigma^2 \left\{ \hat{Y}_{h(\text{new})} - Y_{h(\text{new})} \right\} = \sigma^2 \left\{ \mathbf{X}_{h(\text{new})}^T (\hat{\beta} - \beta) - \varepsilon_{h(\text{new})} \right\} = \sigma^2 \left\{ \mathbf{X}_{h(\text{new})}^T (\hat{\beta} - \beta) \right\} + \sigma^2 \left\{ \varepsilon_{h(\text{new})} \right\}$$

- Writing $\tau^2 = \mathbf{X}_{h(\text{new})}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{h(\text{new})} + 1$, we see that

$$\hat{Y}_{h(\text{new})} - Y_{h(\text{new})} \sim N(0, \sigma^2 \tau^2).$$

- Therefore, $\frac{\hat{Y}_{h(new)} - Y_{h(new)}}{\sigma\tau} \sim N(0,1)$.

- Replacing the unknown σ^2 with the estimate MSE results in a student–t predictive distribution:

$$\frac{\hat{Y}_{h(\text{new})} - Y_{h(\text{new})}}{\sqrt{MSE}\ \tau} \sim t_{N-P}.$$

- Thus, a $(1-\alpha)$–level prediction interval for $Y_{h(\text{new})}$ is

$$\left[\hat{Y}_{h(\text{new})} - t_{N-P}(\alpha/2)\sqrt{MSE}\ \tau, \hat{Y}_{h(\text{new})} + t_{N-P}(\alpha/2)\sqrt{MSE}\ \tau\right].$$

- When estimating a mean response or predicting a new observation take care that the estimate not fall *outside* of the scope of the model.

Model diagnostics

- Model diagnostics play a key role in both the development and assessment of multiple regression models.
- Most diagnostic techniques carry over from simple regression.
- However, given more than one explanatory variables, one must also consider potential *relationships between variables*.
- Scatter plots of all *pair-wise* combinations of variables contained in the model can be summarized in a scatter plot matrix.
- We can also make 3D scatter plots.
- We can assess model assumptions by analyzing *residual* plots.
- In the multiple regression setting we should *plot the residuals* against *each* of the explanatory variables and against the fitted values.