7.3 Extensions of Simple Linear Regression

➢ Power Laws: $Y = \alpha X^\beta$

- $\beta > 1$
- $0 < \beta < 1$
- $\beta < 0$
If a scatterplot exhibits evidence of a monotonic nonlinear trend, then it may be possible to improve the regression model by first transforming the data according to one of the power functions above, depending on its overall shape.

If \( Y = \alpha X^\beta \), then it follows that \( \log(Y) = \log(\alpha) + \beta \log(X) \)

\[
V = \beta_0 + \beta_1 U.
\]

[See Appendix > Basic Reviews > Logarithms for properties of logarithms.]

That is, if \( X \) and \( Y \) have a power law association, then \( \log(X) \) and \( \log(Y) \) have a linear association. Therefore, such \((X, Y)\) data are often replotted on a log-log \((U, V)\) scale in order to bring out the linear trend. The linear regression coefficients of the transformed data are then computed, and backsolved for the original parameters \( \alpha \) and \( \beta \). Algebraically, any logarithmic base can be used, but it is customary to use natural logarithms “\( \ln \)” – that is, base \( e = 2.71828… \)

Thus, if \( Y = \alpha X^\beta \), then \( V = \beta_0 + \beta_1 U \), where \( V = \ln(Y), U = \ln(X) \), and the parameters \( \beta_0 = \ln(\alpha) \) and \( \beta_1 = \beta \), so that the scale parameter \( \alpha = e^{\beta_0} \), and the shape parameter \( \beta = \beta_1 \). However…

Comment: This description of the retransformation is not quite complete. For, recall that linear regression assumes the true form of the response as \( V = \beta_0 + \beta_1 U + \varepsilon \). (The random error term \( \varepsilon \) is estimated by the least squares minimum \( \text{SS}_{\text{Error}} = \sum_{i=1}^{n} e_i^2 \).

Therefore, exponentiating both sides, the actual relationship between \( X \) and \( Y \) is given by \( Y = \alpha X^\beta e^\varepsilon \). Hence (see section 7.2), the conditional expectation is \( E[Y \mid X] = \alpha X^\beta E[e^\varepsilon] \), where \( E[e^\varepsilon] \) is the mean of the exponentiated errors \( e_i \), and is thus estimated by the sample mean of the exponentiated residuals \( e_i \). Consequently, the estimate of the original scale parameter \( \alpha \) is more accurately given by

\[
\hat{\alpha} = e^{\hat{\beta}_0} \times \frac{1}{n} \sum_{i=1}^{n} e_i^{\hat{\varepsilon}_i}.
\]

(The estimate of the original shape parameter \( \beta \) remains \( \hat{\beta} = \hat{\beta}_1 \).)

In this context, the expression \( \frac{1}{n} \sum_{i=1}^{n} e_i^{\hat{\varepsilon}_i} \) is called a smearing factor, introduced to reduce bias during the retransformation process. Note that, ideally, if all the residuals \( e_i = 0 \) – i.e., the model fits exactly – then (because \( e^0 = 1 \)) it follows that the smearing factor \( = 1 \). This will be the case in most of the “rigged” examples in this section, for the sake of simplicity. The often-cited reference below contains information on smearing estimators for other transformations.

Example: This example is modified from a pharmaceutical research paper, *Allometric Scaling of Xenobiotic Clearance: Uncertainty versus Universality* by Teh-Min Hu and William L. Hayton, that can be found at the URL http://www.aapsj.org/view.asp?art=ps030429, and which deals with different rates of metabolic clearance of various substances in mammals. (A *xenobiotic* is any organic compound that is foreign to the organism under study. In some situations, this is loosely defined to include naturally present compounds administered by alternate routes or at unnatural concentrations.) In one part of this particular study, $n = 6$ mammals were considered: mouse, rat, rabbit, monkey, dog and human. Let $X$ = “body weight (kg)” and the response $Y$ = “clearance rate of some specific compound.” Suppose the following “ideal” data were generated (consistent with the spirit of the article’s conclusions):

<table>
<thead>
<tr>
<th>$X$</th>
<th>.02</th>
<th>.25</th>
<th>2.5</th>
<th>5</th>
<th>14</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>5.318</td>
<td>35.355</td>
<td>198.82</td>
<td>334.4</td>
<td>723.8</td>
<td>2420.0</td>
</tr>
</tbody>
</table>
Solving for the least squares regression line yields the following standard output.

Residuals:
1 2 3 4 5 6
-102.15 -79.83 8.20 59.96 147.60 -33.78

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 106.800 | 49.898 | 2.14 | 0.099 |
| X | 33.528 | 1.707 | 19.64 | 3.96e-05 |

Residual standard error: 104.2 on 4 degrees of freedom
Multiple R-Squared: 0.9897, Adjusted R-squared: 0.9872
F-statistic: 385.8 on 1 and 4 degrees of freedom, p-value: 3.962e-05

Residuals vs Fitted
The residual plot, as well as a visual inspection of the linear fit, would seem to indicate that model improvement is possible, despite the high \( r^2 \) value. The overall shape is suggestive of a power law relation \( Y = \alpha X^\beta \) with \( 0 < \beta < 1 \). Transforming to a log-log scale produces the following data and regression line.

\[
\begin{align*}
U &= \ln X \\
V &= \ln Y
\end{align*}
\]

\[
\begin{array}{cccccc}
-3.912 & -1.386 & 0.916 & 1.609 & 0.264 & 4.248 \\
\end{array}
\]

\[
\hat{V} = 4.605 + 0.75 U
\]

Residuals:

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
-2.469e-05 & -1.944e-06 & -1.938e-06 & 6.927e-05 & 2.244e-05 & -6.313e-05
\end{array}
\]

Coefficients:

\[
\begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & t value & \text{Pr(>|t|)} \\
(Intercept) & 4.605e+00 & 2.097e-05 & 219568 & <2e-16 *** \\
U & 7.500e-01 & 7.602e-06 & 98657 & <2e-16 *** \\
\end{array}
\]

Residual standard error: 4.976e-05 on 4 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 9.733e+009 on 1 and 4 degrees of freedom, p-value: 0
The residuals are all within \(10^{-6}\) of 0; this is clearly a much better fit to the data. Transforming back to the original \(X, Y\) variables from the regression line

\[
\ln(\hat{Y}) = 4.605 + 0.75 \ln(X),
\]

we obtain…

\[
\hat{Y} = e^{4.605 + 0.75 \ln(X)} = e^{4.605} e^{0.75 \ln(X)} = 100 X^{0.75}.
\]

That is, the variables follow a power law relation with exponent \(\frac{3}{4}\), illustrating a result known as Kleiber’s Law of “quarter power scaling.” See Appendix > Regression Models > Power Law Growth for more examples and information.
Logarithmic Transformation: \( Y = \alpha e^{\beta X} \) (Assume \( \alpha > 0 \)).

In some systems, the response variable \( Y \) grows (\( \beta > 0 \)) or decays (\( \beta < 0 \)) exponentially in \( X \). That is, each unit increase in \( X \) results in a new response value \( Y \) that is a constant multiple (either \( > 1 \) or \( < 1 \), respectively) of the previous response value. A typical example is unrestricted cell division where, under ideal conditions, the number of cells \( Y \) at the end of every time period \( X \) is twice the number at the previous period. (The resulting explosion in the number of cells helps explain why patients with bacterial infections need to remain on their full ten-day regimen of antibiotics, even if they feel recovered sooner.) The half-life of a radioactive isotope is a typical example of exponential decay.

In general, if \( Y = \alpha e^{\beta X} \), then \( \ln(Y) = \ln(\alpha) + \beta X \), i.e.,

\[
V = \beta_0 + \beta_1 X.
\]

That is, \( X \) and \( \ln(Y) \) have a linear association, and the model itself is said to be log-linear. Therefore, the responses are often replotted on a semilog scale – i.e., \( \ln(Y) \) versus \( X \) – in order to bring out the linear trend. As before, the linear regression coefficients of the transformed data are then computed, and backsolved for estimates of the scale parameter \( \alpha = e^{\beta_0} \) and shape parameter \( \beta = \beta_1 \).

Also see Appendix > Regression Models > Exponential Growth and Appendix > Regression Models > Example - Newton’s Law of Cooling.

Comment: Recall that the square root and logarithm functions also serve to transform positively skewed data closer to being normally distributed. Caution: If any of the values are \( \leq 0 \), then add a constant value (e.g., +1) uniformly to all of the values, before attempting to take their square root or logarithm!!!
Multiple Linear Regression

Suppose we now have \( k - 1 \) independent explanatory variables \( X_1, X_2, \ldots, X_{k-1} \) (numerical or categorical) to predict a single continuous response variable \( Y \). Then the regression setup “Response = Model + Error” becomes:

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{k-1} X_{k-1} + \epsilon
\]

\[ \leftarrow \text{main effect terms} \]

\[
+ \beta_{11} X_1^2 + \beta_{22} X_2^2 + \ldots + \beta_{k-1,k-1} X_{k-1}^2
\]

\[ \leftarrow \text{quadratic terms (if any)} \]

\[
+ \beta_{25} X_2 X_5 + \beta_{68} X_6 X_8 + \ldots
\]

\[ \leftarrow \text{two-way interaction terms (if any)} \]

\[
+ \beta_{147} X_1 X_4 X_7 + \ldots
\]

\[ \leftarrow \text{three-way interaction terms (if any)} \]

\[
+ \epsilon
\]

For simplicity, first consider the general additive model, i.e., main effects only.

**Question 1:** How are the estimates of the regression coefficients obtained?

**Answer:** Least Squares Approximation (LS), which follows the same principle of minimizing the residual sum of squares \( \text{SS}_{\text{Error}} \). However, this leads to a set of complicated normal equations, best formulated via matrix algebra, and solved numerically by a computer. See figure below for two predictors.

![Diagram of multiple linear regression](image)
Question 2: Which predictor variables among $X_1, X_2, \ldots, X_{k-1}$ are the most important for modeling the response variable? That is, which regression coefficients $\beta_j$ are statistically significant?

Answer: This raises the issue of model selection, one of the most important problems in the sciences. There are two basic stepwise procedures: forward selection (FS) and backward elimination (BE) (as well as widely used hybrids of these methods (FB)). The latter is a bit easier to conceptualize, and the steps are outlined below.

**Model Selection: Backward Elimination (BE)**

**Step 0.** In a procedure that is extremely similar to that for multiple comparison of $k$ treatment means (§6.3.3), first conduct an overall $F$-test of the full model $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1}$, by constructing an ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS = SS/df</th>
<th>Test Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$k-1$</td>
<td>$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$</td>
<td>MS_{Reg}</td>
<td>$F = \frac{MS_{Reg}}{MS_{Err}}$</td>
<td>$F_{k-1, n-k}$, $0 \leq p \leq 1$</td>
</tr>
<tr>
<td>Error</td>
<td>$n-k$</td>
<td>$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$</td>
<td>MS_{Err}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td>$\sum_{i=1}^{n} (y_i - \bar{y})^2$</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Null Hypothesis $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0 \iff$ “There is no linear association between the response $Y$ and any of the predictors $X_1, \ldots, X_k$.”

Alternative Hyp. $H_A$: $\beta_j \neq 0$ for some $j \iff$ “There is a linear association between the response $Y$ and at least one predictor $X_j$.”

If – and only if – the null hypothesis is (hopefully) rejected, it then becomes necessary to determine which of the predictor variables correspond to statistically significant regression coefficients. (Note that this is analogous to determining the mean of which of the $k$ treatment groups are significantly different from the others, in multiple comparisons.)
Example ~

**Step 0.** Conduct an overall $F$-test of significance (via ANOVA) of the **full model**.

**Step 1.** $t$-test

- $H_0: \beta_1 = 0$
- $H_0: \beta_2 = 0$
- $H_0: \beta_3 = 0$
- $H_0: \beta_4 = 0$

$p$-values:
- $p_1 < .05$
- $p_2 < .05$
- $p_3 > .05$
- $p_4 < .05$

Reject $H_0$ Reject $H_0$ Accept $H_0$ Reject $H_0$

**Step 2.** Are all the $p$-values significant (i.e., $< .05 = \alpha$)? If not, then...

**Step 3.** Delete the predictor with the **largest** $p$-value, and recompute new coefficients. Repeat Steps 1-3 as necessary, until all $p$-values are significant.

**Step 4.** Check feasibility of the final **reduced model**, and interpret.
Comment: The steps outlined above extend to much more general models, including interaction terms, binary predictors (e.g., in women’s breast cancer risk assessment, let \( X = 1 \) if a first-order relative – mother, sister, daughter – was ever affected, \( X = 0 \) if not), binary response (e.g., \( Y = 1 \) if cancer occurs, \( Y = 0 \) if not), multiple responses, etc. The overall goal is to construct a parsimonious model based on the given data, i.e., one that achieves a balance between the level of explanation of the response, and the number of predictor variables. A good model will not have so few variables that it is overly simplistic, yet not too many that its complexity makes it difficult to interpret and form general conclusions. There is a voluminous amount of literature on regression methods for specialized applications; some of these topics are discussed below, but a thorough treatment is far beyond the scope of this basic introduction.
## Interaction Terms

Consider the following example. We wish to study the effect of two continuous predictor variables, say $X_1 = \text{“Drug 1 dosage (0-10 mg)”}$ and $X_2 = \text{“Drug 2 dosage (0-20 mg)”}$, on a response variable $Y = \text{“systolic blood pressure (mm Hg)”}$. Suppose that, based on empirical data using different dose levels, we obtain the following additive multilinear regression model, consisting of main effects only:

$$
\hat{Y} = 120 + 0.5 X_1 + 0.25 X_2, \quad 0 \leq X_1 \leq 10, \quad 0 \leq X_2 \leq 20.
$$

Rather than attempting to visualize this planar response surface in three dimensions, we can better develop intuition into the relationships between the three variables by projecting it into a two-dimensional interaction diagram, and seeing how the response varies as each predictor is tuned from “low” to “high.”

First consider the effect of Drug 1 alone on systolic blood pressure, i.e., $X_2 = 0$. As Drug 1 dosage is increased from a low level of $X_1 = 0$ mg to a high level of $X_1 = 10$ mg, the blood pressure increases linearly, from $\hat{Y} = 120$ mm Hg to $\hat{Y} = 125$ mm Hg. Now consider the effect of adding Drug 2, eventually at $X_2 = 20$ mg. Again, as Drug 1 dosage is increased from a low level of $X_1 = 0$ mg to a high level of $X_1 = 10$ mg, blood pressure increases linearly, from $\hat{Y} = 125$ mm Hg to $\hat{Y} = 130$ mm Hg. The change in blood pressure remains constant, thereby resulting in two parallel lines, indicating no interaction between the two drugs on the response.
However, suppose instead that the model includes a statistically significant (i.e., $p$-value < $\alpha$) **interaction term**:

$$\hat{Y} = 120 + 0.5X_1 + 0.25X_2 + 0.1X_1 X_2 \quad 0 \leq X_1 \leq 10, \ 0 \leq X_2 \leq 20.$$ 

This has the effect of changing the response surface from a plane to a “hyperbolic paraboloid,” shaped somewhat like a saddle.

- **Change in response with respect to $X_2$ depends on $X_1$.**
  - $X_1 = 0$: $\Delta Y = 5$ mm Hg
  - $X_1 = 10$: $\Delta Y = 25$ mm Hg

**Interaction between $X_1$, $X_2$ on $Y$!**

Again, at the Drug 2 low dosage level $X_2 = 0$, systolic blood pressure linearly increases by 5 mm Hg as Drug 1 is increased from $X_1 = 0$ to $X_1 = 10$, exactly as before. But now, at the Drug 2 high dosage level $X_2 = 20$, a different picture emerges. For as Drug 1 dosage is increased from a low level of $X_1 = 0$ mg to a high level of $X_1 = 10$ mg, blood pressure linearly increases from $\hat{Y} = 125$ mm Hg to a hypertensive $\hat{Y} = 150$ mm Hg, a much larger difference of 25 mm Hg! Hence the two resulting **lines are not parallel**, indicating a significant **drug-drug interaction** on the response.

**Exercise:** Draw the interaction diagram corresponding to the model

$$\hat{Y} = 120 + 0.5X_1 + 0.25X_2 - 0.1X_1 X_2.$$

**Comment:** As a rule, if an explanatory variable $X_j$ is not significant as a main effect, but is a factor in a statistically significant interaction term, it is nevertheless retained as a main effect in the final model. This convention is known as the **Hierarchical Principle**.
These ideas also appear in another form. Consider the example of constructing a \textbf{simple linear} regression model for the response variable “$Y = \text{height (in.)}$” on the single predictor variable “$X = \text{weight (lbs.)}$” for individuals of a particular age group. A reasonably positive correlation might be expected, and after obtaining sample observations, the following scatterplot may result, with accompanying least squares regression line.

However, suppose it is the case that the sample is actually composed of two distinct subgroups, which are more satisfactorily modeled by separate, \textit{but parallel}, regression lines, as in the examples shown below.
It is possible to fit both parallel lines to a single multiple linear model simultaneously, by introducing a binary variable that, in this case, codes for gender. Let $M = 1$ if Male, and $M = 0$ if Female. Then the model

$$
\hat{Y} = 48 + 0.1 X + 4 M
$$

incorporates both the (continuous) numerical variable $X$, as well as the (binary) categorical variable $M$, as predictors for the response.

However, if the simple linear regression lines are not parallel, then it becomes necessary to include an interaction term, just as before. For example, the model

$$
\hat{Y} = 48 + 0.1 X + 4 M + 0.2 M X
$$

becomes $\hat{Y}_1 = 48 + 0.1X$ if $M = 0$, and $\hat{Y}_2 = 52 + 0.3X$ if $M = 1$. These lines have unequal slopes (0.1 and 0.3), hence are not parallel.

More generally then, categorical data can also be used as predictors of response, by introducing dummy, or indicator variables in the model. Specifically, for each disjoint category $i = 1, 2, 3, \ldots, k$, let $I_i = 1$ if category $i$, and 0 otherwise. For example, for the $k = 4$ categories of blood type, we have

$$
I_1 = \begin{cases} 
1, & \text{if Type A} \\
0, & \text{otherwise} 
\end{cases}
$$

$$
I_2 = \begin{cases} 
1, & \text{if Type B} \\
0, & \text{otherwise} 
\end{cases}
$$

$$
I_3 = \begin{cases} 
1, & \text{if Type AB} \\
0, & \text{otherwise} 
\end{cases}
$$

$$
I_4 = \begin{cases} 
1, & \text{if Type O} \\
0, & \text{otherwise.} 
\end{cases}
$$

Note that $I_1 + I_2 + \ldots + I_k = 1$, so there is collinearity among these $k$ variables; hence – just as in multiple comparisons – there are $k - 1$ degrees of freedom. (Therefore, only this many indicator variables should be retained in the model; adding the last does not supply new information.) As before, a numerical response $Y_i$ for each of the categories can then be modeled by combining main effects and possible interactions of numerical and/or indicator variables.

But what if the response $Y$ itself is categorical, e.g., binary?
Logistic Regression

Suppose we wish to model a *binary* response variable $Y$, i.e., $Y = 1$ ("Success") with probability $\pi$, and $Y = 0$ ("Failure") with probability $1 - \pi$, in terms of a predictor variable $X$. This problem gives rise to several difficulties, as the following example demonstrates.

**Example:** “If you live long enough, you will need surgery.” Imagine that we wish to use the continuous variable “$X = \text{Age}$” as a predictor for the binary variable “$Y = \text{Ever had major surgery (1 = Yes, 0 = No)}$.” If we naively attempt to use simple linear regression however, the resulting model contains relatively little predictive value for the response (either 0 or 1), since it attains all continuous values from $-\infty$ to $+\infty$; see figure below.

This is even more problematic if there are several people of the same age $X$, with some having had major surgery (i.e., $Y = 1$), but the others not (i.e., $Y = 0$). Possibly, a better approach might be to replace the response $Y$ (either 0 or 1), with its *probability* $\pi$, in the model. This would convert the binary variable to a continuous variable, but we still have two problems. First, we are restricted to the finite interval $0 \leq \pi \leq 1$. And second, although $\pi$ is approximately normally distributed, its variance is not constant (see §6.1.3), in violation of one of the assumptions on least squares regression models stated in 7.2.

\[
\hat{\pi} = \hat{\beta}_0 + \hat{\beta}_1 X
\]
One solution to the first problem is to transform the probability \( \pi \), using a continuous link function \( g(\pi) \), which takes on values from \(-\infty\) to \(+\infty\), as \( \pi \) ranges from 0 to 1. The function usually chosen for this purpose is the log-odds, or logit (pronounced “low-jit”): 

\[
g(\pi) = \ln \left( \frac{\pi}{1 - \pi} \right)
\]

Thus, the model is given by…

\[
\ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = b_0 + b_1 X \quad \Leftrightarrow \quad \hat{\pi} = \frac{1}{1 + e^{b_0 - b_1 X}}
\]

\( \pi = P(Y = 1) \)

This reformulation does indeed put the estimate \( \hat{\pi} \) between 0 and 1, but with the constant variance assumption violated, the technique of least squares approximation does not give the best fit here. For example, consider the following artificial data:

<table>
<thead>
<tr>
<th>( X )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi )</td>
<td>0.01</td>
<td>0.01</td>
<td>0.50</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Least squares approximation gives the regression parameter estimates \( b_0 = -5.514 \) and \( b_1 = 2.757 \), resulting in the dotted graph shown. However, a closer fit is obtained by using the technique of Maximum Likelihood Estimation (MLE) – actually, a generalization of least squares approximation – and best solved by computer software. The MLE coefficients are \( b_0 = -7.072 \) and \( b_1 = 3.536 \), resulting in the solid graph shown.
Comments:

- This is known as the “S-shaped,” “sigmoid,” or **logistic curve**, and appears in a wide variety of applications. See Appendix > Regression Models > Logistic Growth for an example involving restricted population growth. (Compare with unrestricted exponential growth, discussed earlier.)

- It is often of interest to determine the **median** response level, that is, the value of the predictor variable $X$ for which a 50% response level is achieved. Hence, if $\hat{p} = 0.5$, then $b_0 + b_1 X = \ln \left( \frac{0.5}{1 - 0.5} \right) = 0$, so $X = -\frac{b_0}{b_1}$.

**Exercise:** Prove that the median response corresponds to the point of inflection (i.e., change in concavity) of any general logistic curve.

- Other link functions sometimes used for binary responses are the **probit** (pronounced “pro-bit”) and **tobit** (pronounced “toe-bit”) functions, which have similar properties to the logit. The technique of using link functions is part of a larger regression theory called **Generalized Linear Models**.

- Since the method of least squares is not used for the best fit, the traditional “coefficient of determination” $r^2$ as a measure of model fitness does not exist! However, several analogous **pseudo-$r^2$** formulas have been defined (Efron, McFadden Cox & Snell, others…), but must be interpreted differently.

- Another way to deal with the nonconstant variance of proportions, which does not require logistic regression, is to work with the **variance-stabilizing transformation** $\arcsin \sqrt{p}$, a technique that we do not pursue here.

- To compare regression models: **Wald Test**, **Likelihood Ratio Test**, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC).

- **Polytomous regression** is used if the response $Y$ has more than two categories.
The **logit** function can be modeled by more than one predictor variable via **multilinear logistic regression**, using selection techniques as described above (except that MLE for the coefficients must be used instead of LS). For instance,

\[
\ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_{k-1} X_{k-1}.
\]

In particular, suppose that one of these variables, say \(X_1\), is binary. Then, as its category level changes from \(X_1 = 0\) to \(X_1 = 1\), the right-hand amount changes exactly by its coefficient \(b_1\). The corresponding amount of change of the left side is equal to the difference in the two log-odds which, via a basic property of logarithms, is equal to the logarithm of the **odds ratio** between the two categories. **Thus, the odds ratio itself can be estimated by** \(e^{b_1}\).

**Example:** Suppose that, in a certain population of individuals 50+ years old, it is found that the probability \(\pi = P(\text{Lung cancer})\) is modeled by

\[
\ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = -6 + 0.05 X_1 + 4.3 X_2
\]

where the predictors are \(X_1 = \text{Age (years)}\), and \(X_2 = \text{Smoker (1 = Yes, 0 = No)}\); note that \(X_1\) is numerical, but \(X_2\) is binary. Thus, for example, a 50-year-old *nonsmoker* would correspond to \(X_1 = 50\) and \(X_2 = 0\) in the model, which yields \(-3.5\) on the right hand side for the “log-odds” of this group. (Solving for the actual probability itself gives \(\hat{\pi} = 1 / (1 + e^{3.5}) = .03\).) We can take this value as a **baseline** for the population. Likewise, for a 50-year-old *smoker*, the only difference would be to have \(X_2 = 1\) in the model, to indicate the change in smoking status from baseline. This would yield \(+0.8\) for the “log-odds” (corresponding to \(\hat{\pi} = 1 / (1 + e^{-0.8}) = 0.69\). Thus, taking the difference gives

\[
\text{log-odds}_{\text{Smokers}} - \text{log-odds}_{\text{Nonsmokers}} = 0.8 - (-3.5)
\]

i.e.,

\[
\log \left( \frac{\text{odds}_{\text{Smokers}}}{\text{odds}_{\text{Nonsmokers}}} \right) = 0.8 + 3.5
\]

Recall that…

\[
\log A - \log B = \log \left( \frac{A}{B} \right)
\]

or

\[
\log(\hat{OR}) = 4.3, \quad \text{so that} \quad \hat{OR} = e^{4.3} = 73.7, \quad \text{quite large.}
\]

That the exponent 4.3 is also the coefficient of \(X_2\) in the model is not a coincidence, as stated above. Moreover, \(\hat{OR} = 73.7\) here, for *any* age \(X_1\)!
**Pharmaceutical Application: Dose-Response Curves**

**Example:** Suppose that, in order to determine its efficacy, a certain drug is administered in subtoxic dosages $X$ (mg) of 90 mg increments to a large group of patients. For each patient, let the binary variable $Y = 1$ if improvement is observed, $Y = 0$ if there is no improvement. The proportion $\pi$ of improved patients is recorded at each dosage level, and the following data are obtained.

<table>
<thead>
<tr>
<th>$X$ (mg)</th>
<th>90</th>
<th>180</th>
<th>270</th>
<th>360</th>
<th>450</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$</td>
<td>0.10</td>
<td>0.20</td>
<td>0.60</td>
<td>0.80</td>
<td>0.90</td>
</tr>
</tbody>
</table>

The logistic regression model (as computed via MLE) is

$$
\ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = -3.46662 + 0.01333 X \quad \Leftrightarrow \quad \hat{\pi} = \frac{1}{1 + e^{3.46662 - 0.01333 X}},
$$

the following graph is obtained.

The **median dosage** is $X = \frac{3.46662}{0.01333} = 260.0$ mg. That is, above this dosage level, more patients are improving than not improving.