Linear Regression Models

Babak Shahbaba, Ph.D.

Associate Professor, Department of Statistics
University of California, Irvine

Irvine, CA

Overview

- A brief review of estimation and hypothesis testing
- Simple linear regression models with one binary explanatory variable
- Statistical inference using linear regression models
- Simple linear regression models with one numerical explanatory variable
- Model assessment and diagnostics
- Multiple linear regression models
- Linear regression models with interaction terms
- Variable transformation

Estimation

Parameter estimation

- In many scientific problems, we are interested in finding the **population mean**, denoted as μ, of a random variable (blood pressure, BMI).
- This is **unknown** in general.
- We refer to unknown quantities as **parameters**.
- We refer to the process of using the observed data to guess these unknown values as **parameter estimation**.

Convention

- We denote the unknown population mean as μ.
- We use $X_1, X_2, \ldots, X_n$ to denote n possible values of X obtained from a sample randomly selected from the population.
- We treat $X_1, X_2, \ldots, X_n$ themselves as n random variables because their values can change depending on which n individuals we sample.
- We assume the samples are **independent and identically distributed** (IID).
- We use $x_1, x_2, \ldots, x_n$ as the specific set of values we have observed in our sample.

Point estimation vs. interval estimation

- Sometimes we only provide a single value as our estimate.
- This is called **point estimation**.
- We use $\hat{\mu}$ to denote the point estimates for μ.
- Point estimates do not reflect our **uncertainty**.
- To address this issue, we can present our estimates in terms of a range of possible values (as opposed to a single value).
- This is called **interval estimation**.

Estimating population mean

- Given n observed values, $X_1, X_2, \ldots, X_n$, from the population, we can estimate the population mean μ with the sample mean:

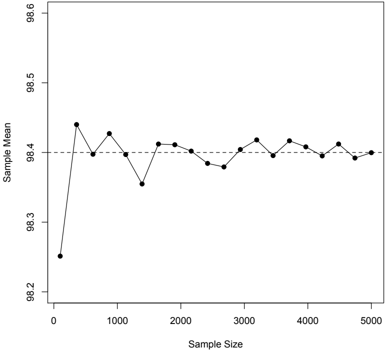
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

- In this case, we say that $\bar{X}$ is an **estimator** for μ.
- The estimator itself is considered as a random variable since it value can change.
- We usually have only one sample of size n from the population $x_1, x_2, \ldots, x_n$.
- Therefore, we only have one value for $\bar{X}$, which we denote $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Law of Large Numbers (LLN)

- Suppose the true population mean for normal body temperature is 98.4F.



- Here, the estimate of the population mean is plotted for different sample sizes.

Confidence intervals for the population mean

- It is common to express our **point estimate along with its standard error** to show how much the estimate could vary if different members of population were selected as our sample.
- Standard error is calculated as

$$\mathrm{SE} = s/\sqrt{n}$$

  where s is the standard deviation and n is the sample size.
- Alternatively, we can use the point estimate and its standard deviation to express our estimate as a range (interval) of possible values for the unknown parameter.
- This is known as **confidence interval**.

Confidence intervals for the population mean

- The confidence interval for the population mean at c confidence level is
- In other words,

$$\left[\bar{x} - t_{\text{crit}} \times SE,\ \bar{x} + t_{\text{crit}} \times SE\,\right].$$

- For 95% confidence interval (i.e., $c = 0.95$), we have $t_{\text{crit}} \approx 2$,

$$\left[\bar{x} - 2 \times SE,\ \bar{x} + 2 \times SE\,\right].$$

Confidence intervals for the population mean

- We can use R or R-Commander to find t_{crit} and consequently confidence intervals (details in the book).
- Alternatively, we can use `Statistics` → `Means` → `Single-sample t-test` to obtain the confidence intervals directly from our data.
- For two-sided confidence intervals, make sure the option `Population mean != mu0` is selected.
- Notice that in this approach, the confidence interval is provided as a part of hypothesis testing; you can ignore the hypothesis testing part for now.

Hypothesis testing

Relationship Between a Numerical Variable and a Binary Variable

- Very often, we try to explain how the population mean changes.
- Here, we focus on problems where we are investigating changes in the population mean between two groups, represented by a binary indicator.
- We treat the binary variable (a.k.a., factor) as the explanatory variable in our analysis.

Relationship Between a Numerical Variable and a Binary Variable

- In general, we can denote the means of the two groups as μ_1 and μ_2.
- The null hypothesis indicates that the population means are equal, $H_0 : \mu_1 = \mu_2$.
- In contrast, the alternative hypothesis is one the following:

$H_A : \mu_1 > \mu_2$ if we believe the mean for group 1 is greater than the mean for group 2.

$H_A : \mu_1 < \mu_2$ if we believe the mean for group 1 is less than the mean for group 2.

$H_A : \mu_1 \neq \mu_2$ if we believe the means are different but we do not specify which one is greater.

Relationship Between a Numerical Variable and a Binary Variable

- We can also express these hypotheses in terms of the **difference** in the means: $H_A : \mu_1 - \mu_2 > 0$, $H_A : \mu_1 - \mu_2 < 0$, or $H_A : \mu_1 - \mu_2 \neq 0$.
- Then the corresponding null hypothesis is that there is no difference in the population means, $H_0 : \mu_1 - \mu_2 = 0$.
- We use a two-sample *t*-test to examine our hypothesis.
- For this, we can use R or R-Commander: `Statistics` → `Means` → `Independent-samples t-test`.

Observed significance level

- We usually decide whether we should reject the null hypothesis (or fail to reject it) based on the **observed significance level** of the test.
- The observed significance level for a test is the probability of values as or more extreme than the observed value, based on the null distribution in the direction supporting the alternative hypothesis.
- This probability is also called the *p*-**value** and denoted p_{obs}.

Interpretation of *p*-value

- The *p*-value is the conditional probability of extreme values (as or more extreme than what has been observed) of the test statistic assuming that the null hypothesis is true.
- When the *p*-value is small, say 0.01 for example, it is rare to find values as extreme as what we have observed (or more so).
- As the *p*-value increases, it indicates that there is a good chance to find more extreme values (for the test statistic) than what has been observed.
- Then, we would be more reluctant to reject the null hypothesis.
- A common **mistake** is to regard the *p*-value as the **probability of null** given the observed test statistic: $P(H_0|\bar{x})$.

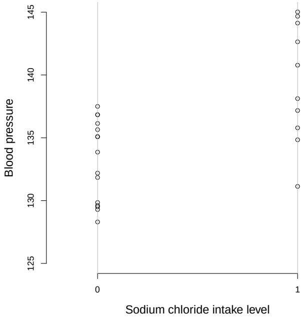
Regression Analysis

Introduction

- We now discuss **linear regression models** for either testing a hypothesis regarding the relationship between one or more **explanatory variables** and a response variable, or **estimating** (predicting) unknown values of the response variable using one or more **predictors**.
- We use X to denote explanatory variables and Y to denote response variables.
- We start by focusing on problems where the explanatory variable is binary. As before, the binary variable X can be either 0 or 1.
- We then continue our discussion for situations where the explanatory variable is numerical.

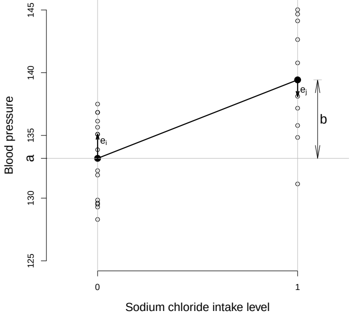
One Binary Explanatory Variable

- Suppose that we want to investigate the relationship between sodium chloride (salt) consumption (low vs. high consumption) and blood pressure among elderly people (e.g., above 65 years old).



One Binary Explanatory Variable

- The following figure shows the dot plot along with sample means, shown as black circles, for each group.
- We connect the two sample means to show the overall pattern for how blood pressure changes from one group to another.



Regression Line

- Using the **intercept** a and **slope** b, we can write the equation for the straight line that connects the estimates of the response variable for different values of X as follows:

$$\hat{y} = a + bx.$$

- The above equation specifies a straight line called the **regression line**.
- The regression line captures the linear relationship between the response variable (here, blood pressure) and the explanatory variable (here, "low" versus "high" sodium chloride diet).

Regression Line

- For this example,

$$\hat{y} = 133.17 + 6.25x.$$

- We **expect** that on average the blood pressure increases by 6.25 units for **one unit increase** in X.
- In this case, one unit increase in X from 0 to 1 means moving from low to high sodium chloride diet group.

Estimation (Prediction) of the Response Variable

- For an individual with $x = 0$ (i.e., low sodium chloride diet), the estimate (expected value) of blood pressure according to the above regression line is

$$\begin{aligned} \hat{y} &= a + b \times 0 = a \\ &= \hat{y}_{x=0}, \end{aligned}$$

  which is the sample mean for the first group.

- For an individual with $x = 1$ (i.e., high sodium chloride diet), the estimate according to the above regression line is

$$\begin{aligned} \hat{y} &= a + b \times 1 = a + b \\ &= \hat{y}_{x=1}. \end{aligned}$$

Residual

- We refer to the difference between the observed and estimated values of the response variable as the **residual**.
- For individual i, we denote the residual e_i and calculate it as follows:

$$e_i = y_i - \hat{y}_i.$$

- For instance, if someone belongs to the first group, her estimated blood pressure is $\hat{y}_i = a = 133.17$.
- Now if the observed value of her blood pressure is $y_i = 135.08$, then the residual is

$$e_i = 135.08 - 133.17 = 1.91.$$

Residual Sum of Squares (RSS)

- As a measure of discrepancy between the observed values and those estimated by the line, we calculate the **Residual Sum of Squares** (RSS):

$$RSS = \sum_{i}^{n} e_i^2.$$

- Among all possible straight lines we could have drawn, the linear regression line provides the smallest value of RSS.
- Therefore, the above approach for finding the regression line is called the **least-squares** method, and the resulting line is called the **least-squares regression line**.

Activity 1

Statistical Inference Using Simple Linear Regression Models

- We discussed fitting a regression line to observed data with one numerical response variable and one binary explanatory variable.
- As usual, we would like to extend our findings to the entire population, i.e., **perform statistical inference**.
- More specifically, we want to **predict** the unknown value of the response variable in the population, **estimate** regression parameters, and **test hypotheses** regarding the relationship between the response and explanatory variables.

Statistical Inference Using Simple Linear Regression Models

- We start by extending our regression model to the whole population.
- Recall that

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ \hat{y}_i &= a + bx_i \end{aligned}$$

- Based on this line, we can write the value of the response variable for individual *i* in terms of the above regression line and the residual:

$$y_i = a + bx_i + e_i.$$

- For the whole population we write the model as follows:

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

Simple Linear Regression Models

- We refer to the above equation as the **linear regression model**.
- More specifically, we call it the **simple linear regression model** since there is only one explanatory variable.
- We refer to α and β as the **regression parameters**. More specifically, β is called the **regression coefficient** for the explanatory variable.
- ϵ is called the **error term**, representing the difference between the estimated (based on the regression line for the entire population) and the actual values of Y in the population.

Estimating Regression Parameters

- The slope *a* and the intercept *b* of the regression line provide **point estimates** for regression parameters α and β.
- Point estimates, however, do not reflect the extent of our uncertainty. Therefore, we find **interval estimates** based on confidence intervals.
- Finding confidence intervals for regression parameters is quite similar to the finding confidence intervals for the population mean.

$$[b - t_{\mathrm{crit}} \times SE_b, b + t_{\mathrm{crit}} \times SE_b].$$

- For simple linear regression models, the standard error SE_b is

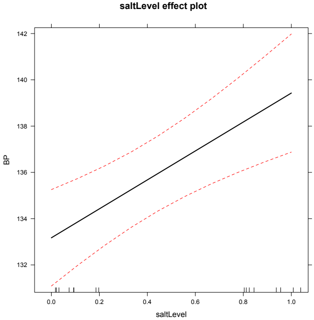
$$SE_b = \frac{\sqrt{RSS/(n-2)}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Estimating Regression Parameters

- We can simply use R-Commander to obtain the confidence intervals.
- Click `Models` → `Confidence intervals` and set the confidence level. (The default is 0.95.)
- R-Commander provides the point estimates along with the confidence intervals for α and β in the output window.

Effect plot

- As our estimates for α and β change, the least-squares regression line changes.
- We can obtain confidence intervals for the regression line and predictions we obtain based on this line.
- Click `Models` → `Graphs` → `Effect plots`.



Hypothesis Testing with Simple Linear Regression Models

- Linear regression models can be used for testing hypotheses regarding possible linear relationship between the response variable and the explanatory variable.
- The null hypothesis stating no [linear] relationship between the two variable can be written as $H_0 : \beta = 0$.
- Similar to the two sample t-test, we first find the t-score.
- Then, we find the p-value (i.e., the observed significance level) by calculating the probability of as or more extreme values than t-score under the null hypothesis.

Hypothesis Testing with Simple Linear Regression Models

- For linear regression models, the t-score is

$$t = \frac{b}{SE_b}.$$

- We find the corresponding p-value as follows:

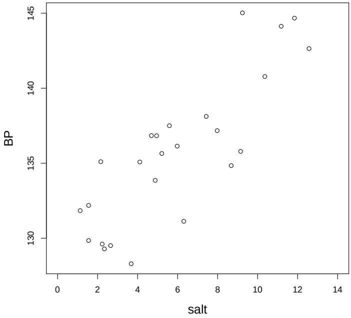
$$\begin{array}{ll} \text{if } H_A : \beta < 0, & p_{\text{obs}} = P(T \le t), \\ \text{if } H_A : \beta > 0, & p_{\text{obs}} = P(T \ge t), \\ \text{if } H_A : \beta \ne 0, & p_{\text{obs}} = 2 \times P\big(T \ge |t|\big), \end{array}$$

- T has the t-distribution with $n - 2$ degrees of freedom.
- When we use R-Commander to fit a linear regression model, the output provides the t-score and its corresponding p-value.

Activity 2

One Numerical Explanatory Variable

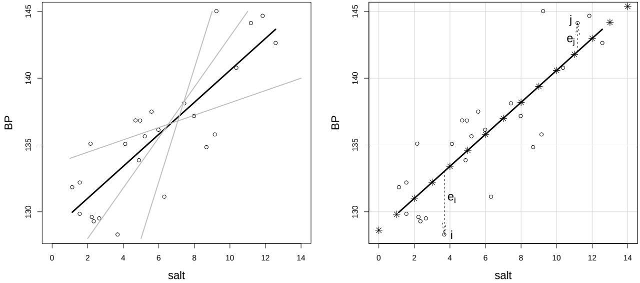
- We now discuss simple linear regression models (i.e., linear regression with only one explanatory variable), where the explanatory variable is numerical.



- As before, we want to find a straight line that captures the relationship between the two variables.

One Numerical Explanatory Variable

- Similar to what we discussed before, among all possible lines we can pass through the data, we choose the **least-squares regression line**, which is the one with the **smallest sum of squared residuals**.



One Numerical Explanatory Variable

- First, we find the slope of regression line using the sample correlation coefficient, r, and the sample standard deviation of Y of X, denoted as s_y and s_x respectively,

$$b = r\frac{s_y}{s_x}.$$

- After finding the slope, we find the intercept as follows:

$$a = \bar{y} - b\bar{x},$$

  where $\bar{y}$ and $\bar{x}$ are the sample means

- For our example, $b = 1.2$ and $a = 128.6$.

Residual

- Given x, we can find the expected value of y for each subject.
- For one individual in our sample, the amount of daily sodium chloride intake is $x_i = 3.68$.
- The estimated value of the blood pressure for this person is

$$\hat{y}_i = 128.60 + 1.20 \times 3.68 = 133.02.$$

- The actual blood pressure for this individual is $y_i = 128.3$. The residual therefore is

$$e_i = y_i - \hat{y}_i = 128.3 - 133.02 = -4.72.$$

Prediction

- We can also use our model for *predicting* the unknown values of the response variable (i.e., blood pressure) for all individuals in the target population.
- For example, if we know the amount of daily sodium chloride intake is $x = 7.81$ for an individual, we can predict her blood pressure as follows:

$$\hat{y} = 128.60 + 1.20 \times 7.81 = 137.97.$$

Interpretation

- The interpretation of the intercept a and the slope b is similar to what we had before.
- $a = 128.6$: the *expected* value of blood pressure is 128.6 for subjects with zero sodium chloride diet.
- $b = 1.2$: the *expected* value of blood pressure increases by 1.2 points corresponding to one unit increase in the daily amount of sodium chloride intake.

Confidence Interval

- As mentioned above, a and b are the point estimates for the regression parameters α and β,

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

- Finding confidence intervals for regression parameters α and β also remains as before.
- More specifically, the confidence interval for regression coefficient is obtained as follows:

$$[b - t_{\mathrm{crit}} \times SE_b, b + t_{\mathrm{crit}} \times SE_b].$$

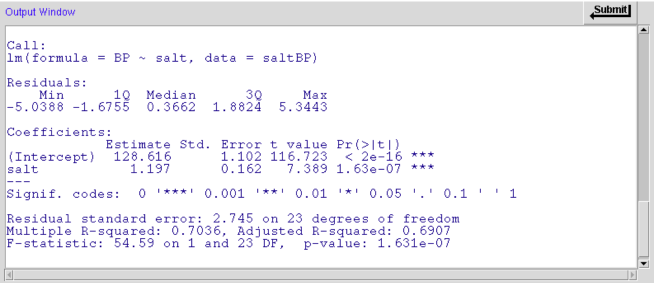
Hypothesis Testing

- The steps for performing hypothesis testing regarding the linear relationship between the response and explanatory variables also remain the same.
- The null hypothesis is $H_0 : \beta = 0$, which indicates that the two variables are not linearly related.
- To evaluate this hypothesis, we need to find the t-score first,

$$t = \frac{b}{SE_b}.$$

Statistical inference using regression models

- As before, we can use R or R-Commander to find the least-squares regression line, obtain confidence intervals, and perform hypothesis testing.



```
Call:
lm(formula = BP ~ salt, data = saltBP)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0388 -1.6755  0.3662  1.8824  5.3443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.616      1.102 116.723  < 2e-16 ***
salt           1.197      0.162   7.389 1.63e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.745 on 23 degrees of freedom
Multiple R-squared: 0.7036, Adjusted R-squared: 0.6907
F-statistic: 54.59 on 1 and 23 DF,  p-value: 1.631e-07
```

Activity 3

Goodness of Fit

- We now want to examine how well the regression line represents the observed data; in other words, how well the regression model **fits** the data.
- In statistics, we use **goodness-of-fit** measures for this purpose.
- The residual sums of squares (RSS) can be interpreted as the **unexplained variation** or **lack of fit**.
- The **total variation** in the response variable is measured by the **Total Sum of Squares** (TSS),

$$TSS = \sum_{i}^{n} (y_i - \bar{y})^2.$$

Goodness of Fit

- The fraction RSS/TSS can be interpreted as the percent of total variation that was not explained by the regression model.
- In contrast, $1 - RSS/TSS$ is fraction of total variation explained by the model.
- This fraction is R^2, which measures the goodness of fit for the regression model,

$$R^2 \quad = \quad 1 - \frac{RSS}{TSS}.$$

- For *simple linear regression* models with one numerical explanatory variable, R^2 is equal to the square of the correlation coefficient r.

Model Assumptions and Diagnostics

- The typical assumptions of linear regression models are

  1. Linearity
  2. Independent observations
  3. Constant variance and normality of the error term

$$\epsilon \sim N(0, \sigma^2).$$

- The first two assumptions are justified by our domain knowledge, our study design, and simple visualization of data.
- To investigate the validity of the third assumptions, click `Models` → `Graphs` → `Basic diagnostic plots`.

Activity 4

Model Assumptions and Diagnostics

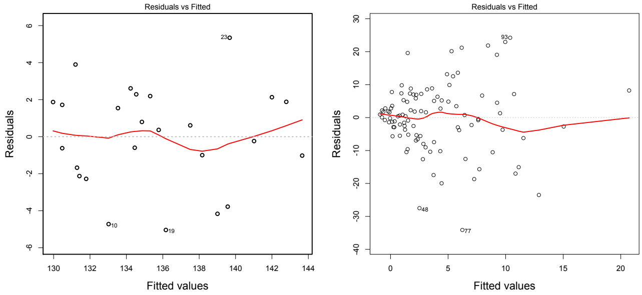


Figure: *Left panel:* The residual plot for the blood pressure example *Right panel*: An illustrative example, where the constant variance assumption is violated.

Multiple Linear Regression

- So far, we have focused on linear regression models with only one explanatory variable.
- In most cases, however, we are interested in the relationship between the response variable and multiple explanatory variables.
- Such models with multiple explanatory variables or predictors are called **multiple linear regression** models.
- For example, we might want to examine the relationship between the birthweight of babies and the smoking status of their mothers during pregnancy.

Multiple Linear Regression

- A multiple linear regression model with p explanatory variables can be presented as follows:

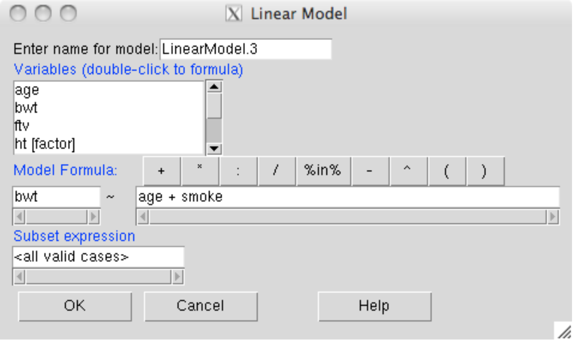
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

- We use the least-squares method as before to estimate the model parameters,

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p.$$

- For this, we can use R-Commander as before.

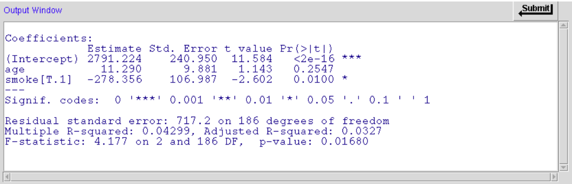
Multiple Linear Regression



Enter name for model: LinearModel.3

bwt ~ age + smoke

Subset expression: <all valid cases>



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2791.224    240.950  11.584   <2e-16 ***
age           11.290      9.881   1.143   0.2547
smoke[T.1]  -278.356    106.987  -2.602   0.0100 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 717.2 on 186 degrees of freedom
Multiple R-squared: 0.04299, Adjusted R-squared: 0.0327
F-statistic: 4.177 on 2 and 186 DF,  p-value: 0.01680
```

Interpretation

- The intercept in multiple linear regression model is the expected (average) value of the response variable when all the explanatory variables in the model are set to zero simultaneously.
- In the above example, the intercept is $a = 2791$, which is obtained by setting age and smoking to zero.
- We might be tempted to interpret this as the average birthweight of babies for nonsmoking mothers (`smoke=0`) with age equal to zero.
- In this case, however, this is not a reasonable interpretation since mother's age cannot be zero.

Interpretation

- We interpret b_j as our estimate of the expected (average) change in the response variable associated with a unit increase in the corresponding explanatory variable X_j *while all other explanatory variables in the model remain fixed*.
- For the above example, the point estimate of the regression coefficient for age is $b_1 = 11$, and the estimate of the regression coefficient for `smoke` is $b_2 = -278$.
- We expect that the birthweight of babies increase by 11 grams as the mother's age increases by one year among mothers with the same smoking status.
- The expected birthweight decreases by -278 grams associated with one unit increase in the value of the variable `smoke` among mothers with the same age.

Activity 5

Additivity

- In multiple linear regression models, we usually assume that the effects of explanatory variables on the response variable are **additive**.
- This means that the expected change in the response variable corresponding to one unit increase in one of the explanatory variables remains the same regardless of the values of other explanatory variables in the model.

Additivity

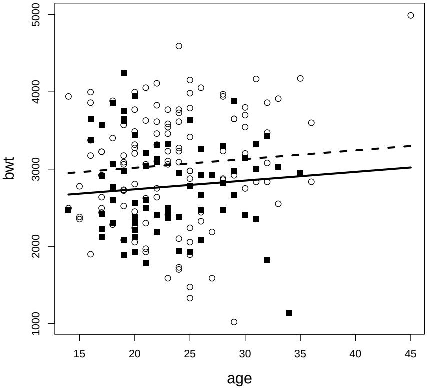


Figure: Nonsmoking mothers are shown as *circles*, while smoking mothers are shown as *squares*. The *dashed line* shows the regression line among nonsmoking mothers, and the *solid line* shows the regression line among the smoking mothers

Interaction

- We might believe that the effects are not additive.
- That is, the effect of one explanatory variable X_1 on the response variable depends on the value of another explanatory variable X_2 in the model).
- We can still use linear regression models by including a new variable $X_3 = X_1 X_2$,

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon.$$

- The term $X_1 X_2$ is called the **interaction term**.
- We refer to β_1 and β_2 as the **main effects**, and refer to β_{12} as the **interaction effect**.

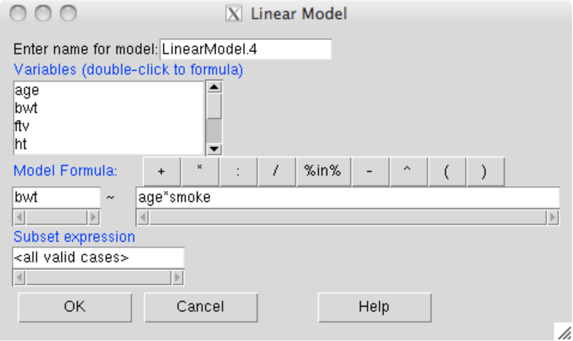
Interaction

- As before, we use the least-squares method to estimate the model parameters,

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2.$$

- In R and R-Commander, to fit models with interaction terms, we use "*" instead of "+" to separate variables.
- Note that when we include an interaction term in our model, we should be cautious about how we interpret model parameters.

Interaction



Enter name for model: LinearModel.4

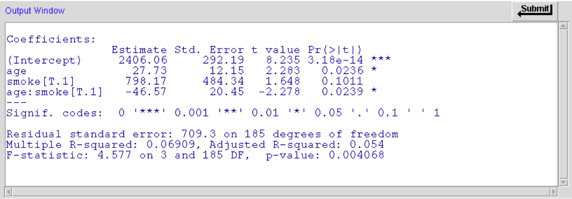
age
bwt
ftv
ht

Model Formula:

bwt ~ age*smoke

Subset expression

<all valid cases>



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2406.06     292.19   8.235 3.18e-14 ***
age               27.73      12.15   2.283   0.0236 *  
smoke[T.1]       798.17     484.34   1.648   0.1011    
age:smoke[T.1]   -46.57      20.45  -2.278   0.0239 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 709.3 on 185 degrees of freedom
Multiple R-squared: 0.06909,	Adjusted R-squared: 0.054
F-statistic: 4.577 on 3 and 185 DF,  p-value: 0.004068
```

Interaction

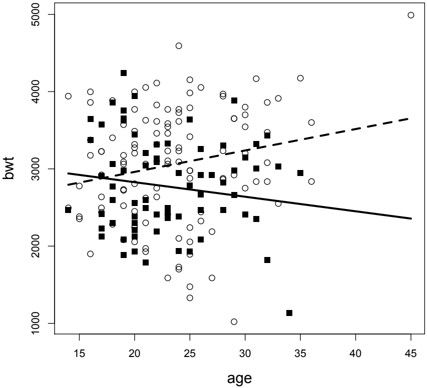


Figure: Nonsmoking mothers are shown as *circles*, while smoking mothers are shown as *squares*. The *dashed line* shows the regression line among nonsmoking mothers, and the *solid line* shows the regression line among the smoking mothers

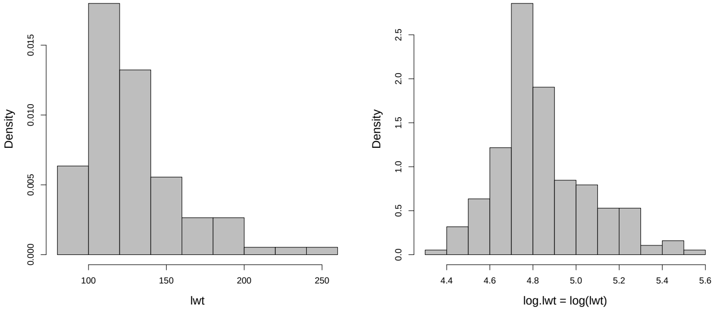
Activity 6

Variable transformation

- Occasionally, we rely on data transformation techniques (i.e., applying a function to variables) to stabilize variance, reduce noise, and control the influence of extreme values in our analysis.
- Two of the most commonly used transformation functions for this purpose are *logarithm* and *square root*.
- The logarithm function, $\log(x)$, is usually used to transform right-skewed variables with positive values.
- The square root function is usually used for right-skewed count variables.

Variable transformation

- For example, consider the `lwt` variable in the `birthwt` data set.
- As shown in the left panel of the following figure, the variable is right-skewed.



- As shown in the right panel, the resulting variable is less skewed compared to the original variable.

Variable transformation

- To do this in R-Commander, click `Data` → `Manage variables in active data set` → `Compute new variable`.
- Under `New variable name` enter `log.lwt`, and under `Expression to compute` enter `log(lwt)`.
- This creates a new variable, `log.lwt` whose values are the natural logarithm of `lwt`.
- If we want to use the square root transformation. we use `sqrt` instead of `log`.

Caution!

- When such data transformations are used for explanatory and/or response variables in regression models, you have to be very cautions about interpreting the results.
- For example, if you use $\log_2$ transformation for an explanatory variable, its corresponding coefficient represents the average change in the response variable for one unit increase in the $\log_2$ scale, which in turn corresponds to doubling the explanatory variable in the original scale.
- If you use log-transformation for the response variable, your interpretation should be in terms of its average in the log-scale, and this is not the same as the log of its average.

Creating new variables based on existing ones

- Sometimes, we need to create a new variable based on two or more existing variables.
- For this, we can follow steps similar to those for data transformation described above.
- For example, if a data set includes the variable `weight` (in pounds) and the variable `height` (in inches) for each person in the sample, we can calculate the value of BMI as follows:

$$BMI = \frac{weight \times 703}{(height)^2}$$

- To create a new variable for BMI, we follow the above steps, set `New variable name` to `BMI`, and under `Expression to compute` we enter the above expression.

Discretization

- Another common preprocessing technique is to create categorical variables based on numerical variables.
- This could help us to see the patterns more clearly and identify relationships more easily.
- For example, according to the Centers for Disease Control and Prevention (CDC), the standard weight status categories associated with BMI ranges for adults are as follows:

BMI	**Weight Status**
Below 18.5	Underweight
18.5-24.9	Normal
25.0-29.9	Overweight
30.0 and Above	Obese

Discretization

- In R-Commander, we can divide `bmi` (from the `Pima.tr`) into four groups: Underweight, Normal, Overweight, and Obese.
- Click `Data` → `Manage variables in active data set` → `Recode variables`. Select `bmi` as the `Variable to recode` and enter "weight.status" as the `New variable name`. Then in the `Enter recode directives` box, type

```
0:18.5 = "Underweight"
18.5:24.9 = "Normal"
25.0:29.9 = "Overweight"
30.0:100 = "Obese"
```

- The newly created variable `weight.status` is added to the data set.

Discretization

- This variable is categorical. More specifically, it is an ordinal variable.
- To specify the order of categories in R-Commander, click `Data` → `Manage variables in active data set` → `Reorder factor levels`. Then select `weight.status`.
- R-Commander will open a window to reorder levels of the categorical variable.
- Note that the default order is alphabetical.

Activity 7