Review of the expected value, covariance, correlation coefficient, mean, and variance.

**Random variable.** A variable that takes on alternative values according to chance. More specifically, a random variable assumes different values, each with probability less than or equal to 1.

A *continuous* random variable may take on any value on the real number line. A *discrete* random variable may take only a specific number of values.

**Probability mass function (pmf) and probability density function (pdf).** The process that generates the values of a random variable. It lists all possible outcomes and the probability that each will occur.

**Expected values.** The mean, or the expected value, of a random variable $X$ is a weighted average of the possible outcomes, where the probabilities of the outcomes serve as the weights. Expectation operator denoted $E$, mean of $X$ denoted $\mu_X$. In the discrete case,

$$\mu_X = E(X) = p_1X_1 + p_2X_2 + \cdots + p_NX_N = \sum_{i=1}^{N} p_iX_i,$$

where $p_i = P(X = X_i)$ and $\sum_{i=1}^{N} p_i = 1$.

In the continuous case, $\mu_X = E(X) = \int_{\Omega} f(x)xdx$, where $\Omega$ is the support of $X$, and $f(x)$ is the pdf of $X$.

The sample mean of a set of outcomes on $X$ is denoted by $\bar{X}$.

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

where $x_1, \ldots, x_n$ are realizations of $X$. 

1
**Variance.** The variance of a random variable provides a measure of the spread, or dispersion, around the mean. \( \text{Var}(X) = \sigma_X^2 = E[(X - E(X))^2] \). In the discrete case,

\[
\text{Var}(X) = \sigma_X^2 = \sum_{i=1}^{N} p_i [X_i - E(X)]^2.
\]

The positive square root of the variance is called the **standard deviation**.

**Joint distribution.** In the discrete case, joint distributions of \( X \) and \( Y \) are described by a list of probabilities of occurrence of all possible outcomes on both \( X \) and \( Y \).

The **covariance** of \( X \) and \( Y \) is \( \text{Cov}(X, Y) = E[X - E(X)(Y - E(Y))] \). In the discrete case,

\[
\text{Cov}(X, Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} [X_i - E(X)][Y_j - E(Y)],
\]

where \( p_{ij} \) represents the joint probability of \( X = X_i \) and \( Y = Y_j \) occurring.

The covariance is a measure of the linear association between \( X \) and \( Y \). If both variables are always above and below their means at the same time, the covariance will be positive. If \( X \) is above its mean when \( Y \) is below its mean and vice versa, the covariance will be negative.

The value of covariance depends upon the units in which \( X \) and \( Y \) are measured. The **correlation coefficient** \( \rho_{XY} \) is a measure of the association which has been normalized and is **scale-free**.

\[
\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},
\]

where \( \sigma_X \) and \( \sigma_Y \) represent the standard deviation of \( X \) and \( Y \) respectively. The correlation coefficient is always between -1 and 1. A positive correlation indicates that the variables move in the same direction, while a negative correlation implies that they move in opposite directions.

**Properties.** Suppose \( X \) and \( Y \) are random variables, \( a \) and \( b \) are constants:

1. \( E(aX + bY) = aE(X) + bE(Y) \);
2. \( \text{Var}(aX + b) = a^2 \text{Var}(X) \);
3. \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \);
4. If \( X \) and \( Y \) are independent, then \( \text{Cov}(X, Y) = 0 \).
Estimation.

Means, variances and covariances can be measured with certainty only if we know all there is to know about all possible outcomes. In practice, we may obtain a sample of the relevant information needed. Given $x_1, \ldots, x_n$ are random observations of $X$, we want to estimate a population parameter (like the mean or the variance).

The sample mean $\bar{X}$ is an unbiased estimator of the population mean $\mu_X$.

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) = \frac{1}{n} \sum_{i=1}^{n} E(x_i) = \frac{1}{n} \sum_{i=1}^{n} \mu_X = \mu_X.$$ 

The sample variance $s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$ is an unbiased estimator of the population variance $\text{Var}(X)$.

The sample covariance $s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})$ is an unbiased estimator of the population covariance $\text{Cov}(X, Y)$.

The sample correlation coefficient is defined as

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{X})^2 \sum_{i=1}^{n} (y_i - \bar{Y})^2}}.$$ 

Desired properties of estimators:

1. **Lack of bias.** The bias associated with an estimated parameter is defined to be: $\text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta$. $\hat{\beta}$ is an unbiased estimator if the mean or the expected value of $\hat{\beta}$ is equal to the true value, that is, $E(\hat{\beta}) = \beta$.

2. **Consistency.** $\hat{\beta}$ is an consistent estimator of $\beta$ if for any $\delta > 0$, $\lim_{n \to \infty} P(|\beta - \hat{\beta}| < \delta) = 1$. As sample size $n$ approaches infinity, the probability that $\hat{\beta}$ will differ from $\beta$ will get very small.

3. **Efficiency.** We say that $\hat{\beta}$ is an efficient unbiased estimator if for a given sample size, the variance of $\hat{\beta}$ is smaller than the variance of any other unbiased estimators.
Tradeoff between bias and variance of estimators. When the goal is to maximize the precision of the predictions, an estimator with low variance and some bias may be more desirable than an unbiased estimator with high variance. We may want to minimize the mean square error, defined as

$$\text{Mean square error}(\hat{\beta}) = E(\hat{\beta} - \beta)^2.$$  

It can be shown that Mean square error($\hat{\beta}$) = $[\text{Bias}(\hat{\beta})]^2 + \text{Var}(\hat{\beta})$. The criterion of minimizing mean square error take into account of both the variance and the bias of the estimator.

An alternative criterion to consistency is that the mean square error of the estimator approaches zero as the sample size increases. This implies asymptotically, or when the sample size is very large, the estimator is unbiased and its variance goes to zero. An estimator with a mean square error that approaches zero will be consistent estimator but that the reverse need not be true.
The Normal distribution is a continuous bell-shaped probability distribution. It can be fully described by its mean and its variance. If $X$ is normally distributed, we write $X \sim N(\mu_X, \sigma_X^2)$, which is read $X$ is normally distributed with mean $\mu_X$ and variance $\sigma_X^2$. The probability that a single observation of a normally distributed variable will lie within 1.96 standard deviations of its mean is approximately 0.95. The probability that a single observation of a normally distributed variable will lie within 2.57 standard deviations of its mean is approximately 0.99.

\[
P(\mu_X - 1.96\sigma_X < X < \mu_X + 1.96\sigma_X) \approx 0.95,
\]

\[
P(\mu_X - 2.57\sigma_X < X < \mu_X + 2.57\sigma_X) \approx 0.99.
\]

The weighted sum of normal random variables is still normal.
The chi square distribution with N degrees of freedom is the sum of the squares of N independently distributed standard normal random variables (with mean 0 and variance 1).

The chi square distribution starts at the origin, is skewed to the right, and has a tail which extends infinitely to the right. The distribution becomes more and more symmetric as the number of degrees of freedom gets larger. It becomes close to normal distribution when degrees of freedom is very large.

As an example, when we calculate the sample variances $s^2$ of n observations from a normal distribution with variances $\sigma^2$, $(n-1)s^2/\sigma^2$ is chi square with $n-1$ degrees of freedom.
The t distribution. Assume $X$ is normal with mean 0 and variance 1, $Z$ is chi square with $N$ degrees of freedom, $X$ and $Z$ are independent. Then $X/\sqrt{Z/N}$ has a t distribution with $N$ degrees of freedom.

The t distribution is symmetric, has fatter tails than the normal distribution, and approximates the normal distribution when the degrees of freedom is large.

As an example, $X \sim N(\mu_X, \sigma_X^2)$, and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean based on a sample $x_1, \ldots, x_n$ of size $n$. Then $(\bar{X} - \mu_X)/(\sigma_X/\sqrt{n})$ is normal with mean 0 and variance 1. For unknown $\sigma_X^2$, we replace $\sigma_X^2$ by sample variance $s_X^2$, then

$$\frac{\sqrt{n}(\bar{X} - \mu_X)}{s_X} = \frac{(\bar{X} - \mu_X)/(\sigma_X/\sqrt{n})}{\sqrt{((n-1)s_X^2/\sigma_X^2)/(n-1)}}$$

which is a t distribution with $n - 1$ degrees of freedom.
The **F distribution**. If \( X \) and \( Z \) are independent and distributed as chi square with \( N_1 \) and \( N_2 \) degrees of freedom, respectively, then \( (X/N_1)/(Z/N_2) \) is distributed as an F distribution with \( N_1 \) and \( N_2 \) degrees of freedom. These two parameters are called the numerator and the denominator degrees of freedom.
**Forecasting**: to calculate or predict some future event or condition, usually as a result of rational study or analysis of pertinent data.

We all make forecasts: a person waiting for a bus; parents expecting a telephone call from their children; bank manager predicts cash flow for the next quarter; company manager predicts sales or estimates number of man-hours required to meet a given production schedule.

Future events involve uncertainty. Forecasts are not perfect. The objective of forecasting is to reduce the forecast error. Each forecast has its own specifications, and solutions to one are not solutions in another situation.

**General principles for any forecast system:**
Model specification; Model estimation; Diagnostic checking; Forecast generation; Stability checking; Forecast updating.

**Choice of forecast model.** (1) Degrees of accuracy required; (2) forecast horizon; (3) budget; (4) what data is available. One may not construct accurate empirical forecast models from limited and incomplete data base.

**Forecast criteria.** Actual observation at time $t$ is $z_t$. Its forecast, which uses the information up to and including time $t-1$, is $z_{t-1}(1)$. Objective is to make the future forecast error $z_t - z_{t-1}(1)$ as small as possible. However, $z_t$ is unknown, we can only talk about its expected value, conditional on the observed data up to and including time $t-1$. We minimize the mean absolute error $E|z_t - z_{t-1}(1)|$ or the mean square error $E[z_t - z_{t-1}(1)]^2$. We use the mean square error criterion for simpler mathematical calculations.

**What we will discuss.** In single-variable forecasting, we use past history of the series, $z_t$, where $t$ is the time index, to extrapolate into the future. In regression forecasting, we use the relationships between the variable to be forecast and the other variables.

In the most general form, the regression model can be written as

$$y_t = f(x_{t1}, \ldots, x_{tp}; \beta_1, \ldots, \beta_m) + \varepsilon_t$$

Describes relationship between one dependent variable $Y$ and $p$ independent variables $X_1, \ldots, X_p$.

- Index $t$ means at time $t$ or for subject $t$.
- At index $t$, we observe $y_t$ and $x_{t1}, \ldots, x_{tp}$.
- Unknown parameters $\beta_1, \ldots, \beta_m$.
- Known mathematical function form $f$.
- Randomness for this model rises from the error term $\varepsilon_t$. 
In the most general form, the regression model can be written as

$$y_t = f(x_{t1}, \ldots, x_{tp}; \beta_1, \ldots, \beta_m) + \varepsilon_t$$

Describes relationship between one dependent variable $Y$ and $p$ independent variables $X_1, \ldots, X_p$.

- Index $t$ means at time $t$ or for subject $t$.
- At index $t$, we observe $y_t$ and $x_{t1}, \ldots, x_{tp}$.
- Unknown parameters $\beta_1, \ldots, \beta_m$.
- Known mathematical function form $f$.
- Randomness for this model rises from the error term $\varepsilon_t$.

Models **linear in the parameter**:

1. $y = \beta_0 + \beta_1 x_1 + \varepsilon$
2. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
3. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$

Models **linear in the independent variable**:

1. $y = \beta_0 + \beta_1 x_1 + \varepsilon$
2. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

**Regression through origin.**

- $y$: salary of a sales person; $x$: number of products sold. No base salary!

- $y$: merit I get for increase in salary; $x$: number of papers published.

Model: $y = \beta x + \varepsilon$. 
Given observations \((x_1, y_1), \ldots, (x_n, y_n)\), we want to minimize
\[
S(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2.
\]
Take derivatives and set to 0, then we have
\[
S'(\beta) = \frac{dS(\beta)}{d\beta} = -2 \sum_{i=1}^{n} (y_i - \beta x_i)x_i = 0
\]
The least squares estimator of \(\beta\) is
\[
\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.
\]
Take second order derivative to make sure this is a minimizer. Check
\[
S''(\beta) = \frac{d^2S(\beta)}{d\beta^2} = \frac{dS'(\beta)}{d\beta} = 2 \sum_{i=1}^{n} x_i^2 \geq 0.
\]

**Simple linear regression.**
- \(y\): resale price of a preowned car; \(x\): milage!
- \(y\): running time for 10-km road race; \(x\): maximal aerobic capacity (oxygen uptake, milliliter per kilogram per minute, \(ml/(kg \cdot min)\))

**Show MINITAB example! Scatter plot!**
Model: \(y = \beta_0 + \beta_1 x + \varepsilon\).
Given observations \((x_1, y_1), \ldots, (x_n, y_n)\), we want to minimize
\[
S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.
\]
Take derivatives and set to 0, then we have
\[
\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0
\]
\[
\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0
\]

We have two equations and two unknowns, solve for \(\beta_0\) and \(\beta_1\). Then we have
\[
\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},
\]
where
\[
s_{xy} = \frac{1}{n-1} \left( \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \right) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}
\]
\[
s_x^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right] = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}
\]

We still need to take second order derivatives to make sure \((\hat{\beta}_0, \hat{\beta}_1)\) are actual minimizers.

Model assumptions:

1. The relationship between \(x\) and \(y\) is linear, \(y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\);

2. The \(x\) are non-stochastic variable whose values \(x_1, \ldots, x_n\) are fixed;

3. **Normality** The error term \(\varepsilon\) is normally distributed;

   **Homoscedasticity** The error term \(\varepsilon\) has mean zero and constant variance for all observations, \(E(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) = \sigma^2\);

   **Independency** The random error \(\varepsilon_i\) and \(\varepsilon_j\) are independent of each other for \(i \neq j\), or errors corresponding to different observations are independent.

How do we check those assumptions? **DO AN EXAMPLE BY HAND. Show MINITAB RESULTS!**

\(\hat{\beta}_1, \hat{\beta}_0\) as weighted sum

Let \(w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\), then \(\hat{\beta}_1 = \sum_{i=1}^{n} w_i y_i\). We notice that
\[
\sum_{i=1}^{n} w_i = 0, \quad \text{and} \quad \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i (x_i - \bar{x}) = 1.
\]
Thus,

\[ E(\hat{\beta}_1) = E(\sum_{i=1}^{n} w_i y_i) = \sum_{i=1}^{n} w_i E(y_i) = \sum_{i=1}^{n} w_i (\beta_0 + \beta_1 x_i) \]

\[ = \beta_0 \sum_{i=1}^{n} w_i + \beta_1 \sum_{i=1}^{n} w_i x_i = \beta_1, \]

which means \( \hat{\beta}_1 \) is unbiased estimator of \( \beta_1 \). Furthermore

\[ \sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \]

and

\[ \text{Var}(\hat{\beta}_1) = \text{Var}(\sum_{i=1}^{n} w_i y_i) = \sum_{i=1}^{n} w_i^2 \text{Var}(y_i) = \sigma^2 \sum_{i=1}^{n} w_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \]

Similarly, \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{x} \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} (\frac{1}{n} - \bar{x} w_i) y_i. \) Thus

\[ E(\hat{\beta}_0) = \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} w_i \right) E(y_i) = \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} w_i \right) (\beta_0 + \beta_1 x_i) \]

\[ = \beta_0 - \bar{x} \beta_0 \sum_{i=1}^{n} w_i + \beta_1 \bar{x} - \beta_1 \bar{x} \sum_{i=1}^{n} w_i x_i = \beta_0, \]

which means \( \hat{\beta}_0 \) is unbiased estimator of \( \beta_0 \). Furthermore

\[ \text{Var}(\hat{\beta}_0) = \text{Var}(\sum_{i=1}^{n} (\frac{1}{n} - \bar{x} w_i) y_i) = \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} w_i \right)^2 \text{Var}(y_i) = \sigma^2 \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} w_i \right)^2 \]

\[ = \sigma^2 \sum_{i=1}^{n} \left( \frac{1}{n^2} + \bar{x}^2 w_i^2 - \frac{2}{n} \bar{x} w_i \right) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \]

We can also show that

\[ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x} \sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]
Gauss-Markov Theorem: If the previous model assumptions are satisfied, then among all the linear unbiased estimators of $\beta_0$ and $\beta_1$, the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have the smallest variance.

The Theorem implies that for any unbiased estimator of $\beta_1$ with form $\sum w_i y_i$, its variance should be $\geq \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

The fitted least squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is used to get the fitted value $y$.

The i-th residual is the difference between the i-th observation of dependent variable and its fitted value: $r_i = y_i - \hat{y}_i$.

In practice, $\text{Var}(\varepsilon) = \sigma^2$ is unknown and we need to estimate it. The sum of squares of error (SSE; also known as sum of squares of residual) is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$ 

It has $n - 2$ degrees of freedom. We have n random observations, that’s n degrees of freedom. Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$ uses 2 degrees of freedom. The mean squares error (MSE) is defined to be

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}.$$ 

MSE is also denoted by $\hat{\sigma}^2$. It is an unbiased estimator of $\sigma^2$. It can be shown that $E(MSE) = E(\hat{\sigma}^2) = \sigma^2$.

The i-th standardized residual $sr_i = r_i / \hat{\sigma}$.

In summarization, we have

$$\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

$$\hat{\beta}_0 \sim N \left( \beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

$$\frac{SSE}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2.$$
Simple linear regression.

Model: \( y = \beta_0 + \beta_1 x + \varepsilon \).

Model assumptions:

1. The relationship between \( x \) and \( y \) is linear, \( y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \);
2. The \( x \) are non-stochastic variable whose values \( x_1, \ldots, x_n \) are fixed;
3. Normality The error term \( \varepsilon \) is normally distributed;
   Homoscedasticity The error term \( \varepsilon \) has mean zero and constant variance for all observations, \( E(\varepsilon_i) = 0 \), \( \text{Var}(\varepsilon_i) = \sigma^2 \);
   Independency The random error \( \varepsilon_i \) and \( \varepsilon_j \) are independent of each other for \( i \neq j \), or errors corresponding to different observations are independent.

Given observations \((x_1, y_1), \ldots, (x_n, y_n)\), we want to minimize

\[
S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.
\]

Take derivatives and set to 0, then we have

\[
\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0
\]

\[
\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)x_i = 0
\]

We have two equations and two unknowns, solve for \( \beta_0 \) and \( \beta_1 \). Then we have

\[
\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \text{where}
\]
\[ s_{xy} = \frac{1}{n-1} \left( \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \right) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \]

\[ s_x^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 \right] = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \]

We still need to take second order derivatives to make sure \((\hat{\beta}_0, \hat{\beta}_1)\) are actual minimizers.

**The fitted least squares line:** \(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\) is used to get the fitted value \(y\).

**The i-th residual** is the difference between the i-th observation of dependent variable and its fitted value: \(r_i = y_i - \hat{y}_i\).

In practice, \(\text{Var}(\varepsilon) = \sigma^2\) is unknown and we need to estimate it. The sum of squares of error (SSE; also known as sum of squares of residual) is

\[ SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \]

It has \(n-2\) degrees of freedom. We have \(n\) random observations, that’s \(n\) degrees of freedom. Estimating \(\hat{\beta}_0\) and \(\hat{\beta}_1\) uses 2 degrees of freedom. The mean squares error (MSE) is defined to be

\[ MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}. \]

MSE is also denoted by \(\hat{\sigma}^2\). It is an unbiased estimator of \(\sigma^2\). It can be shown that \(E(MSE) = E(\hat{\sigma}^2) = \sigma^2\).

**The i-th standardized residual** \(sr_i = r_i/\hat{\sigma}\).

In summarization, we have

\[
\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).
\]

\[
\hat{\beta}_0 \sim N \left( \beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right).
\]

\[
\frac{SSE}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.
\]
Confidence interval

Denote $\sigma^2_{\hat{\beta}_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $s^2_{\hat{\beta}_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$, then

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0, 1).$$

For known $\sigma^2$, the $100(1 - \alpha)\%$ confidence interval for $\beta_1$ is $\hat{\beta}_1 \pm \frac{z_{\alpha/2}\sigma_{\hat{\beta}_1}}{s_{\hat{\beta}_1}}$.

For unknown $\sigma^2$, plug in $s_{\hat{\beta}_1}$ as an estimate of unknown $\sigma_{\hat{\beta}_1}$ and we have

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{(\hat{\beta}_1 - \beta_1)/\sigma_{\hat{\beta}_1}}{\sqrt{(n-2)\sigma^2/(n-2)}} \sim \frac{N(0, 1)}{\sqrt{\chi^2_{n-2}/(n-2)}} = t_{n-2}.$$

Denote the upper percentile $t_{n-2,\alpha/2}$, which satisfies $P(t > t_{n-2,\alpha/2}) = \alpha/2$ for random variable $t \sim t_{n-2}$. Thus

$$\text{Prob}\left(-t_{n-2,\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} < t_{n-2,\alpha/2}\right) = 100(1 - \alpha)\%$$

$$\text{Prob}(\hat{\beta}_1 - t_{n-2,\alpha/2}s_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2}s_{\hat{\beta}_1}) = 100(1 - \alpha)\%$$

The $100(1 - \alpha)\%$ confidence interval for $\beta_1$ is $\hat{\beta}_1 \pm t_{n-2,\alpha/2}s_{\hat{\beta}_1}$.

Hypothesis test

We are interested whether there is significant predictor effect.

Null hypothesis: $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$.

Test statistics: $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$. Under $H_0$, $t \sim t_{n-2}$.

At $1 - \alpha$ significance level, reject $H_0$ if $|t| > t_{n-2,\alpha/2}$.

We are interested whether there is significant positive predictor effect.

Null hypothesis: $H_0 : \beta_1 = 0$ v.s. $H_0 : \beta_1 > 0$.

Test statistics: $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$. Under $H_0$, $t \sim t_{n-2}$.

At $1 - \alpha$ significance level, reject $H_0$ if $t > t_{n-2,\alpha}$.

We are interested whether the slope is at a certain specified level $c$.

Null hypothesis: $H_0 : \beta_1 = c$ v.s. $H_0 : \beta_1 \neq c$.

Test statistics: $t = \frac{\hat{\beta}_1 - c}{s_{\hat{\beta}_1}}$. Under $H_0$, $t \sim t_{n-2}$.

At $1 - \alpha$ significance level, reject $H_0$ if $|t| > t_{n-2,\alpha/2}$. 
Lecture 5, Feb 19

Please check with your classmates for missed lecture notes. The materials listed here are the key points for ANOVA. Make sure you understand EVERYTHING listed here. We are going to get back to ANOVA when we deal with multiple linear regression.

This version does not have any examples or demonstration in MINITAB. TUCAPTURE should be available soon.

ANOVA

The variation of $y$ has the following decomposition

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

Total Sum of Squares (SST) = Sum of Squares of Regression (SSR) + Sum of Squares of Error (SSE)

$R^2$, or the $R$ squared, coefficient of determination, is the proportion of variation in $y$ explained by the regression equation.

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

The sample correlation coefficient $R = \pm \sqrt{R^2}$. $R > 0$ when $\hat{\beta}_1 > 0$; and $R < 0$ when $\hat{\beta}_1 < 0$. $|R|$ measures how closely the data fit a straight line.

ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$</td>
<td>MSR=SSR/1</td>
<td>F=MSR/MSE</td>
</tr>
<tr>
<td>Error</td>
<td>n-2</td>
<td>$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$</td>
<td>MSE=SSE/(n-2)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To test whether there is **significant predictor effect**, we don’t use $R^2$. We use $F$, which takes into consideration of the degrees of freedom.

Null hypothesis: $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$.

If test statistic $F$ is large, which means a large proportion of variation in $y$ is explained by the regression, the intuition is that we reject $H_0$.

It can be shown that under $H_0$, $F = \frac{MSR}{MSE} \sim F_{1,n-2}$.

Denote the upper percentile $F_{1,n-2,\alpha}$, which satisfies $P(F > F_{1,n-2,\alpha}) = \alpha$ for random variable $F \sim F_{1,n-2}$.

At $1 - \alpha$ significance level, reject $H_0$ if $F = \frac{MSR}{MSE} > F_{1,n-2,\alpha}$.

Notice that

$$
F = \frac{MSR}{MSE} = \frac{SSR}{SSE/(n-2)} = \frac{(n-2)SSR}{SST - SSR} = \frac{(n-2)SSR/SST}{SST/SST - SSR/SST} = \frac{(n-2)R^2}{1 - R^2}
$$

Large $R^2$ with small $n$ may not be significant; but moderate $R^2$ with large sample size $n$ can be highly significant.

It can be shown that the $t$ test and the $F$ test agree with each other.
Lecture 6, Feb 23

Model assumptions:

1. The relationship between $x$ and $y$ is linear, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$;

2. The $x$ are non-stochastic variable whose values $x_1, \ldots, x_n$ are fixed;

3. Normality The error term $\varepsilon$ is normally distributed;
   - Homoscedasticity The error term $\varepsilon$ has mean zero and constant variance for all observations, $E(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) = \sigma^2$;
   - Independency The random error $\varepsilon_i$ and $\varepsilon_j$ are independent of each other for $i \neq j$, or errors corresponding to different observations are independent.

Gauss-Markov Theorem: If the previous model assumptions are satisfied, then among all the linear unbiased estimators of $\beta_0$ and $\beta_1$, the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have the smallest variance.

The Theorem implies that for any unbiased estimator of $\beta_1$ with form $\sum w_i y_i$, its variance should be $\geq \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

Prediction Interval

Consider a college trying to predict first year GPA of a new student based on the student’s high school GPA. Based on present students, the regression equation is $\hat{y} = 0.4 + 0.8x$, where $x$ = high school GPA and $y$ = first year college GPA.

For a student with high school GPA $x_{new} = 3.5$, we estimate its first year college GPA by $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new} = 0.4 + 0.8 \times 3.5 = 3.2$.

The mean of the forecast error is

$$E(y_{new} - \hat{y}_{new}) = E[(\beta_0 + \beta_1 x_{new} + \varepsilon_{new}) - (\hat{\beta}_0 + \hat{\beta}_1 x_{new})] = 0.$$
By the Gauss-Markov theorem, \( \hat{y}_{\text{new}} \) is the minimum mean square error forecast among all linear unbiased forecasts.

The variance of the forecast error is

\[
\text{Var}(y_{\text{new}} - \hat{y}_{\text{new}}) = \text{Var}[(\beta_0 + \beta_1 x_{\text{new}} + \varepsilon_{\text{new}}) - (\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}})]
= \sigma^2 + \text{Var}(\hat{\beta}_0) + x_{\text{new}}^2 \text{Var}(\hat{\beta}_1) + 2x_{\text{new}} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1)
= \sigma^2 + \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n} + \frac{x_{\text{new}}^2 \sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} - \frac{2x_{\text{new}} \bar{x} \sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
= \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)
\]

For known \( \sigma^2 \), the prediction interval for \( y_{\text{new}} \) is

\[
\hat{y}_{\text{new}} \pm z_{\alpha/2} \sqrt{\text{Var}(y_{\text{new}} - \hat{y}_{\text{new}})} = \hat{y}_{\text{new}} \pm z_{\alpha/2} \sqrt{\sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}.
\]

For unknown \( \sigma^2 \), plug in \( MSE = \hat{\sigma}^2 \) and the prediction interval for \( y_{\text{new}} \) is

\[
\hat{y}_{\text{new}} \pm t_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}.
\]

Which prediction interval is wider? Why?

**Confidence Interval for expectation of \( y \)**

If we are not interested in just one particular student with high school GPA=3.5, but we are interested in the expectation of college GPA for all students with high school GPA=3.5. With known \( \sigma^2 \), the Confidence Interval for \( y_0 = E(y|x_0) \) is

\[
\hat{y}_0 \pm z_{\alpha/2} \sqrt{\sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}.
\]

For unknown \( \sigma^2 \), plug in \( MSE = \hat{\sigma}^2 \) and the prediction interval for \( y_0 \) is

\[
\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}.
\]
Here $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Is the prediction interval wider or the confidence interval wider? Why?

**Multiple linear regression**

Instead of studying the relationship between response $y$ and one independent variable (predictor) $x$, we may also use linear model to study the relationship between response $y$ and multiple independent variables $x_1, \ldots, x_p$.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon.$$  

We have $n$ set of observations $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, \ldots, n$. We obtain least square estimators of the unknown parameters by minimizing the SSE:

$$S(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2.$$  

Set the first order derivatives equal to 0:

$$\frac{\partial S}{\partial \beta_0} = 0, \frac{\partial S}{\partial \beta_1} = 0, \cdots, \frac{\partial S}{\partial \beta_p} = 0.$$  

The solutions are denoted by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

The fitted least squares line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

The $i$-th residual is $r_i = y_i - \hat{y}_i$.

**ANOVA**

The variation of $y$ has the same decomposition as in the simple linear regression

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

$$\text{SST} = \text{SSR} + \text{SSE}$$
ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>p</td>
<td>SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2</td>
<td>MSR=SSR/p</td>
<td>\text{F}=\text{MSR}/\text{MSE}</td>
</tr>
<tr>
<td>Error</td>
<td>n-p-1</td>
<td>SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2</td>
<td>MSE=SSE/(n-p-1)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SST = \sum_{i=1}^{n}(y_i - \bar{y})^2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To test whether there is **significant predictor effect**, simultaneous test

Null hypothesis \( H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \) v.s.

Alternative hypothesis \( H_1: \) At least one coefficient not equal to zero.

If test statistic \( F \) is large, which means a large proportion of variation in \( y \) is explained by the regression, the intuition is that we reject \( H_0 \).

It can be shown that under \( H_0, \ F = \text{MSR}/\text{MSE} \sim F_{p,n-p-1}. \)

Denote the upper percentile \( F_{p,n-p-1,\alpha} \), which satisfies \( P(F > F_{p,n-p-1,\alpha}) = \alpha \) for random variable \( F \sim F_{p,n-p-1}. \)

At \( 1 - \alpha \) significance level, reject \( H_0 \) if \( F = \text{MSR}/\text{MSE} > F_{p,n-p-1,\alpha} \), and this means the regression is significant.

Fail to reject \( H_0 \) if \( F = \text{MSR}/\text{MSE} < F_{p,n-p-1,\alpha} \), and this means none of the regression coefficients significantly differs from 0.

Notice that

\[
F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/(p)}{\text{SSE}/(n-p-1)} = \frac{n-p-1}{p} \frac{\text{SSR}}{\text{SST} - \text{SSR}} \\
= \frac{n-p-1}{p} \frac{\text{SSR}/\text{SST}}{\text{SST}/\text{SST} - \text{SSR}/\text{SST}} = \frac{n-p-1}{p} \frac{\text{R}^2}{1 - \text{R}^2}.
\]

The \( \text{R}^2 \), or coefficient of determination, is \( \text{R}^2 = \frac{\text{SSR}/\text{SST}}{\text{SST}/\text{SST} - \text{SSR}/\text{SST}} = \frac{n-p-1}{p} \frac{\text{R}^2}{1 - \text{R}^2} \).

\( \text{R}^2 \) always increase as the number of predictors \( p \) increase, which may result in model overfitting. We need a new criterion.

The corrected \( \text{R}^2 \), or \( \bar{\text{R}}^2 \), is \( \bar{\text{R}}^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)} \). It takes degrees of freedom into consideration and penalizes overfitting.
Lecture 7, March 2

Please read Chapter 2 of the textbook, from page 8 to page 41. You may skip Section 2.5.4 and Section 2.5.5.

For the basics of linear algebra about matrix calculation, you may find the entry from Wikipedia useful: http://en.wikipedia.org/wiki/Matrix_(mathematics)

Matrix representation of multiple linear regression

We have $n$ set of independent observations $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$ from model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon.$$

Let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then we have $Y = X\beta + \varepsilon$, where $\varepsilon$ is multivariate normal with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2 I_n$.

The SSE can be written as

$$S(\beta) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2 = (Y - X\beta)'(Y - X\beta).$$
Take derivatives with respect to $\beta$ and we have the following normal equation:

$$X'X\beta = X'Y,$$

which leads to least squares estimator $\hat{\beta} = (X'X)^{-1}X'Y$.

Because

$$E(\hat{\beta}) = E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'E(X\beta + \varepsilon)$$

$$= (X'X)^{-1}X'[X\beta + E(\varepsilon)] = (X'X)^{-1}X'X\beta = I_{p+1}\beta = \beta,$$

$\hat{\beta}$ is unbiased estimator of $\beta$.

Because $\text{Var}(AX) = A\text{Var}(X)A'$ for random vector $X$ and constant matrix $A$, we have

$$\text{Var}(\hat{\beta}) = \text{Var}[(X'X)^{-1}X'Y] = [(X'X)^{-1}X']\text{Var}(Y)[(X'X)^{-1}X']'$$

$$= \text{Var}[(X'X)^{-1}X'Y] = [(X'X)^{-1}X']\text{Var}(X\beta + \varepsilon)[(X'X)^{-1}X']'$$

$$= \sigma^2(X'X)^{-1}X'1_n(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

Thus we know $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X'X)^{-1})$.

Thus $\hat{\beta}_i \sim N(\beta_i, \sigma^2 d_i)$, where $d_i$ is the $i$-th element on the diagonal of matrix $(X'X)^{-1}$. 

2