BIO5312 Biostatistics
Lecture 10: Regression and Correlation Methods

Dr. Junchao Xia
Center of Biophysics and Computational Biology
Fall 2016
In this lecture, we will discuss topics in Chapter 11:

- Methods of regression and correlation analysis in which two different variables in the same sample are related.

- Multiple-regression analysis, where the relationship between more than two variables at a time is considered.

- Linear regression methods where we learn how to relate a normally distributed outcome variable $y$ to one or more predictor variables $x_1, \ldots, x_k$ where the $x$’s may be either continuous or categorical variables.

Linear relationship between $y$ and $x$: $E(y|x) = \alpha + \beta x$

The line $y = \alpha + \beta x$ is the regression line, $\alpha$ is the intercept and $\beta$ is the slope of the line. $y = \alpha + \beta x$ is not expected to be true for every point. Need add an error term $e$, which assumes a normal distribution with mean 0 and variance $\sigma^2$. Then, we have $y = \alpha + \beta x + e$

For any linear-regression equation of the form $y = \alpha + \beta x + e$, $y$ is called the dependent variable and $x$ is called the independent variable because we are trying to predict $y$ as a function of $x$.

**Figure 11.1** Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

Regression line $y = 21.52 + 0.608x$

If $\beta$ is equal to 0, then there is no linear relationship between $x$ and $y$.

If $\sigma^2$ equal to 0, then every point would fall exactly on the regression line. The larger $\sigma^2$ is the more scatter occurs about the regression line.
Fitting Regression Lines—The Method of Least Squares

The least-squares line, or estimated regression line, is the line \( y = a + bx \) that minimizes the sum of the squared distances of the sample points from the line given by \( S = \sum_{i=1}^{n} d_i^2 \).

This method of estimating the parameters of a regression line is known as the method of least squares.

\[
S = \sum_{i=1}^{n} (y_i - a - bx_i)^2
\]

\[
\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = 0
\]

\[
\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0
\]
Sum of Squares and Estimations of the Least-Squares Line

The following notation is needed to define the slope and intercept of a regression line.

- **Raw sum of squares for** $x$ **is defined by** $\sum_{i=1}^{n} x_i^2$
- **Corrected sum of squares for** $x$ **is denoted by** $L_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 / n$

It represents the sum of squares of the deviations of the $x_i$ from the mean.

- **Raw sum of squares for** $y$ **is defined by** $\sum_{i=1}^{n} y_i^2$
- **Corrected sum of squares for** $y$ **is denoted by** $L_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2 / n$

- **Raw sum of cross products** is defined by $\sum_{i=1}^{n} x_i y_i$
- **Corrected sum of cross products** is defined by $L_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

With a short form $\sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)(\sum_{i=1}^{n} y_i) / n$

- The coefficients of the least-squares line $y = a + bx$ are given by $b = L_{xy}/L_{xx}$ and $a = \bar{y} - b\bar{x} = \left(\frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}\right)$

- The **predicted or average value of** $y$ **for a given value of** $x$, as estimated from the fitted regression line, is denoted by $\hat{y} = a + bx$. Thus, the point $(x, a+bx)$ is always on the regression line.
The point \((\bar{x}, \bar{y})\) falls on the regression line. This is common to all estimated regression lines because a regression line can be represented as \(y = a + bx = \bar{y} - b\bar{x} + bx = \bar{y} + b(x - \bar{x})\) or \(y - \bar{y} = b(x - \bar{x})\).

### Figure 11.5 Goodness of fit of a regression line

- For any sample point \((x_i, y_i)\), the **residual or residual component**, of that point about the regression line is defined by \(y_i - \hat{y}_i\).
- For any sample point \((x_i, y_i)\), the **regression component** of that point about the regression line is defined by \(\hat{y}_i - \bar{y}\).
Regression Lines with Varying Residual Components

Figure 11.6  Regression lines with varying goodness-of-fit relationships

(a) Large regression, small residual components
(b) Large regression, large residual components
(c) Small regression, small residual components
(d) Small regression, large residual components
Decomposition of the Total Sum of Squares

- **Total sum of squares** or **Total SS** is the sum of squares of the deviations of the individual sample points from the sample mean:
  \[ \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

- **Regression sum of squares** or **Reg SS** is the sum of squares of the regression components:
  \[ \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]

- **Residual sum of squares** or **Res SS** is the sum of squares of the residual components:
  \[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

- Decomposition of the total sum of squares into regression and residual components:
  \[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

or **Total SS = Reg SS + Res SS**
Goodness-of-fit is considered as the ratio of the regression sum of squares to the residual sum of squares. A large ratio indicates a good fit, whereas a small ratio indicates a poor fit.

The regression mean square, or \( \text{Reg MS} \), is the \( \text{Reg SS} \) divided by the number of predictor variables (\( k \)) in the model (not including the constant). Thus, \( \text{Reg MS} = \text{Reg SS}/k \). For simple linear regression, \( k = 1 \) and thus \( \text{Reg MS} = \text{Reg SS} \). For multiple regression, \( k \) is > 1. \( k \) is referred to as the degrees of freedom for the regression sum of squares or \( \text{Reg df} \).

The residual mean square, or \( \text{Res MS} \), is the ratio of the \( \text{Res SS} \) divided by \( (n - k - 1) \), or \( \text{Res MS} = \text{Res SS}/(n - k - 1) \). For simple linear regression, \( k = 1 \) and \( \text{Res MS} = \text{Res SS}/(n-2) \). We refer to \( n - k - 1 \) as the degrees of freedom for the residual sum of squares, or \( \text{Res df} \). \( \text{Res MS} \) is also sometimes denoted by \( s^2_{y,x} \).

\[
\begin{align*}
  s^2_x &= L_{xx}/(n-1) \\
  s^2_y &= L_{yy}/(n-1) \\
  s^2_{y,x} \neq s^2_{x,y} &= L_{xy}/(n-1)
\end{align*}
\]
Short Computational Form for Regression and Residual SS
Regression SS = \( bL_{xy} = b^2L_{xx} = \frac{L_{2,xy}}{L_{xx}} \)
Residual SS = Total SS – Regression SS = \( L_{yy} - \frac{L_{2,xy}}{L_{xx}} \)

\( F \) test for simple linear regression
To test \( H_0: \beta=0 \) vs. \( H_1: \beta \neq 0 \), use the following procedure:
1) Compute the test statistic
\( F = \frac{\text{Reg MS}}{\text{Res MS}} = \frac{(L_{2,xy}/L_{xx})}{[(L_{yy} - L_{2,xy}/L_{xx})/(n-2)]} \)
That follows an \( F_{1,n-2} \) distribution under \( H_0 \).
2) For a two-sided test with significance level \( \alpha \), if
\( F > F_{1,n-2,1-\alpha} \) then reject \( H_0 \); if \( F \leq F_{1,n-2,1-\alpha} \) then accept \( H_0 \).
3) The exact \( p \)-value is given by \( Pr(F_{1,n-2} > F) \).
Acceptance and Rejection Regions, ANOVA

Acceptance and rejection regions for the simple linear-regression $F$ test

$$F = \frac{\text{Reg MS/Res MS}}{L_{xx}/L_{xx}} = \frac{(L_{xy}^2/L_{xx})}{(L_{yy}^2/L_{xx})/(n - 2)}$$

$F_{1, n-2}$ distribution

$F \leq F_{1, n-2, 1-\alpha}$ Acceptance region

$F > F_{1, n-2, 1-\alpha}$ Rejection region

Computation of the $p$-value for the simple linear-regression $F$ test

$p = \Pr(F_{1, n-2} > F)$

ANOVA table for displaying regression results

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>$df$</th>
<th>MS</th>
<th>$F$ statistic</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$(a)^a$</td>
<td>1</td>
<td>$(a)/1$</td>
<td>$F = [(a)/1]/[(b)/(n-2)]$</td>
<td>$Pr(F_{1,n-2} &gt; F)$</td>
</tr>
<tr>
<td>Residual</td>
<td>$(b)^b$</td>
<td>$n-2$</td>
<td>$(b)/(n-2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$(a) + (b)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$a(a) = \text{Regression SS.}$

$b(b) = \text{Residual SS.}$
The computation of the p-value for regression F test are summarized in an analysis of variance (ANOVA) table. Results displayed in the ANOVA table have been obtained using the MINITAB REGRESSION program.

### Table 11.3

ANOVA results for the birthweight–estriol data in Example 11.12

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>250.57</td>
<td>250.57</td>
<td>17.16</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>29</td>
<td>423.43</td>
<td>14.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>674.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
$R^2$ for Measuring Goodness of Fit

A summary measure of goodness of fit frequently referred to in the literature is $R^2$, which is defined as $\text{Reg SS}/\text{Total SS}$. $R^2$ can be thought of as the proportion of the variance $y$ that is explained by $x$.

i. if $R^2 = 1$, then all variation in $y$ can be explained by variation in $x$, and all data points fall on the regression line.

ii. If $R^2 = 0$, then $x$ gives no information about $y$ and the variance of $y$ is the same with or without knowing $x$.

iii. If $0 < R^2 < 1$, then for a given value of $x$, the variance of $y$ is lower than it would be if $x$ were unknown but is still greater than 0.

$R^2 = \frac{\text{Reg SS}}{\text{Total SS}}$, \thickapprox \text{Adjusted } R^2 = 1 - \frac{s^2_{y,x}}{s^2_y}$ when $n$ is large.

Adjusted $R^2$ is more sensitive when $n$ is small.
To test the hypothesis \( H_0: \beta = 0 \) vs. \( H_1: \beta \neq 0 \), use the following procedure:

1) Compute the test statistic \( t = b / (s_{yx} / L_{xx})^{1/2} \)

2) For a two-sided test with significance level \( \alpha \),
   - If \( t > t_{n-2,1-\alpha/2} \) or \( t < t_{n-2,\alpha/2} = -t_{n-2,1-\alpha/2} \) then reject \( H_0 \).
   - If \( -t_{n-2,1-\alpha/2} \leq t \leq t_{n-2,1-\alpha/2} \) then accept \( H_0 \).

3) The \( p \)-value is given by
   - \( p = 2 \times (\text{area to the left of } t \text{ under a } t_{n-2} \text{ distribution}) \) if \( t < 0 \)
   - \( p = 2 \times (\text{area to the right of } t \text{ under a } t_{n-2} \text{ distribution}) \) if \( t \geq 0 \)

**Figure 11.10** Computation of the \( p \)-value for the \( t \) test for simple linear regression
Interval Estimation for Linear Regression

Interval estimates for the parameters of a regression line:
Standard errors and are often computed to determine the precision of estimates.

\[
\text{se}(b) = \sqrt{\frac{s^2_{y|x}}{L_{xx}}}
\]
\[
\text{se}(a) = \sqrt{s^2_{y|x} \left( \frac{1}{n} + \frac{x^2}{L_{xx}} \right)}
\]

Two-sided 100% × (1-\(\alpha\)) confidence intervals for the parameters of a regression line

If \(b\) and \(a\) are the estimated slope and intercept of a regression line, resp., and \(se(b)\), \(se(a)\) are the estimated standard errors, then the two-sided 100% × (1-\(\alpha\)) confidence intervals for \(\beta\) and \(\alpha\) are given by

\[b \pm t_{n-2,1-\alpha/2}se(b)\] and \[a \pm t_{n-2,1-\alpha/2}se(a)\], resp.
Interval Estimation for Linear Regression

Predictions made from regression lines for Individual Observations

The distribution of observed $y$ values for the subset of individuals with independent variable $x$ is normal with mean $= \hat{y} = a + bx$ and standard deviation given by

$$se_1(\hat{y}) = \sqrt{s_{y,x}^2 \left[ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}} \right]}$$

Furthermore, $100% \times (1-\alpha)$ of the observed values will fall within the interval $\hat{y} \pm t_{n-2,1-\alpha/2}se_1(\hat{y})$

This interval is sometimes called a $100% \times (1-\alpha)$ prediction interval for $y$.

Standard error and confidence interval for predictions made from regression lines for the average value of $y$ for a given $x$

The best estimate of the average value of $y$ for a given $x$ is $\hat{y} = a + bx$. Its standard error, denoted by $se_2(\hat{y})$, is given by

$$se_2(\hat{y}) = \sqrt{s_{y,x}^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}} \right]}$$

Furthermore, a two-sided $100% \times (1-\alpha)$ confidence interval for the average value of $y$ is $\hat{y} \pm t_{n-2,1-\alpha/2}se_2(\hat{y})$
Assessing the Goodness of Fit of Regression Lines

Assumptions made in linear-regression models

1. For any given value of x, the corresponding value of y has an average value $\alpha + \beta x$, which is a linear function of x.

2. For any given value of x, the corresponding value of y is normally distributed about $\alpha + \beta x$ with the same variance $\sigma^2$ for any x.

3. For any two data points $(x_1,y_1), (x_2,y_2)$ the error terms $e_1, e_2$ are independent of each other.

These assumptions may be tested by using several different kinds of plots. The simplest being the x-y scatter plot. Here, we plot the dependent variable y vs. the independent variable x and superimpose the regression line $y = a + bx$ on the same plot.

Standard deviation of residuals about the fitted regression line

Let $(x_i, y_i)$ be a sample point used in estimating the regression line, $y = \alpha + \beta x$.

If $y = a + bx$ is the estimated regression line and $\hat{e}_i = \text{residual for the point (}x_i,y_i\text{)}$ about the estimated regression line, then $\hat{e}_i = y_i - (a + bx_i)$ and

$$sd(\hat{e}_i) = \sqrt{\sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}}\right]}$$

The Studentized residual corresponding to the point $(x, y)$ is

$$\frac{\hat{e}_i}{sd(\hat{e}_i)}$$
Plot of Studentized Residuals (Example)

Figure 11.11  Plot of Studentized residuals vs. the predicted value of birthweight for the birthweight–estriol data in Table 11.1
One commonly used strategy that can be used if unequal residual variances are present is to transform the dependent variable (y) to a different scale. This type of transformation is called a variance-stabilizing transformation.

The most common transformations when the residual variance is an increasing function of x are either the ln or square-root transformations.

The square-root transformation is useful when the residual variance is proportional to the average value of y. The log transformation is useful when the residual variance is proportional to the square of the average values.

Sometimes, data may be retained in the original scale but a weighted regression employed in which the weight is approximately inversely proportional to the residual variance.

Goodness-of-fit of a regression line may also be judged based on outliers and influential points.

Influential points are defined heuristically as points that have an important influence on the coefficients of the fitted regression lines.

An outlier \((x_i, y_i)\) may or may not be influential depending on its location relative to the remaining sample points.

If \(|x_i - \bar{x}|\) is small, then even a gross outlier will have a relatively small influence on the slope estimate but will have an important influence on the intercept estimate.
Correlation Coefficient

The **sample (Pearson) correlation coefficient** \((r)\) is defined by \(L_{xy}/\sqrt{L_{xx}L_{yy}}\). The correlation is not affected by changes in location or scale in either variable and must lie between -1 and +1. It is a useful tool for quantifying the relationship between variables.

**Interpretation of the sample correlation coefficient**

- If the correlation is greater than 0, then the variables are said to be **positively correlated**. Two variables \((x,y)\) are positively correlated if as \(x\) increases, \(y\) tends to increase, whereas as \(x\) decreases, \(y\) tends to decrease.
- If the correlation is less than 0, then the variables are said to be **negatively correlated**. Two variables \((x,y)\) are **negatively correlated** if as \(x\) increases, \(y\) tends to decrease, whereas as \(x\) decreases, \(y\) tends to increase.
- If the correlation is exactly 0, then the variables are said to be **uncorrelated**. Two variables \((x,y)\) are **uncorrelated** if there is no linear relationship between \(x\) and \(y\).
- The correlation coefficient provides a **quantitative measure** of the dependence between two variables: the closer \(|r|\) is to 1, the more closely related the variables are; if \(|r| = 1\), then one variable can be predicted exactly from the other.
Interpreting the sample correlation coefficient \( r \) in terms of degree of dependence is only correct if the variables \( x \) and \( y \) are normally distributed and in certain other special cases. If the variables are not normally distributed, then the interpretation may not be correct.

The sample correlation coefficient \( r \) can be written as

\[
 r = \frac{L_{xy}/(n-1)}{\sqrt{L_{xx}/(n-1) L_{yy}/(n-1)}}
\]

where \( s^2_x = L_{xx}/(n-1) \) and \( s^2_y = L_{yy}/(n-1) \) are sample variances. If we defined sample covariance \( s^2_{xy} = L_{xy}/(n-1) \), we can now re-express the relation as

\[
 r = \frac{s_{xy}}{s_x s_y} = \frac{\text{sample covariance between } x \text{ and } y}{(\text{sample standard deviation of } x)(\text{sample standard deviation of } y)}
\]

which is completely analogous to the definition of the population correlation coefficient \( \rho \),

\[
 \rho = Corr(X,Y) = Cov(X,Y)/(\sigma_x \sigma_y) = \sigma_{xy} / (\sigma_x \sigma_y)
\]
Where the regression coefficient \( b \) can be interpreted as a rescaled version of the correlation coefficient \( r \) where the scale factor is the ratio of the standard deviation of \( y \) to that of \( x \). \( r \) will be unchanged by a change in the units of \( x \) or \( y \), whereas \( b \) is in the units of \( y/x \).

The regression coefficient is used when we specifically want to predict one variable from another.

The correlation coefficient is used when we simply want to describe the linear relationship between two variables but do not want to make predictions.
One sample \( t \) test for a correlation coefficient

To test the hypothesis \( H_0: \rho = 0 \) vs. \( H_1: \rho \neq 0 \) use the following procedure

1) Compute the sample correlation coefficient \( r \)

2) Compute the test statistic \( t = \frac{r(n-2)^{1/2}/(1-r^2)^{1/2}}{\sqrt{1-r^2}} \)

Which under \( H_0 \) follows a \( t \) distribution with \( n-2 \) df

For a two-sided level \( \alpha \) test,

- if \( t > t_{n-2,1-\alpha/2} \) or \( t < -t_{n-2,1-\alpha/2} \) then reject \( H_0 \).
- If \( -t_{n-2,1-\alpha/2} \leq t \leq t_{n-2,1-\alpha/2} \), then accept \( H_0 \).

3) The \( p \)-value is given by

- \( p = 2 \times (\text{area to the left of } t \text{ under a } t_{n-2} \text{ distribution}) \text{ if } t < 0 \)
- \( p = 2 \times (\text{area to the right of } t \text{ under a } t_{n-2} \text{ distribution}) \text{ if } t \geq 0 \)

We assume an underlying normal distribution for each of the random variables used to compute \( r \).
Fisher’s z transformation of the sample correlation coefficient $r$

The z transformation of $r$ is $z = \frac{1}{2} \ln\left[\frac{1+r}{1-r}\right]$ is approximately normally distributed under $H_0$ with mean $z_0 = \frac{1}{2} \ln\left[\frac{1+\rho_0}{1-\rho_0}\right]$ and variance $1/(n-3)$.

The z transformation is very close to $r$ for small values of $r$ but tends to deviate substantially from $r$ for larger values of $r$.

One sample z test for a correlation coefficient

To test the hypothesis $H_0: \rho = \rho_0$ vs. $H_1: \rho \neq \rho_0$, use the following procedure

1) Compute the sample correlation coefficient $r$ and the z transformation of $r$

2) Compute the test statistic $\lambda = (z - z_0)\sqrt{n-3}$

   If $\lambda > z_{1-\alpha/2}$ or $\lambda < -z_{1-\alpha/2}$ reject $H_0$.

   If $-z_{1-\alpha/2} \leq \lambda \leq z_{1-\alpha/2}$ accept $H_0$.

3) The exact $p$-value is given by

   $p = 2 \times \Phi(\lambda)$ if $\lambda \leq 0$

   $p = 2 \times [1 - \Phi(\lambda)]$ if $\lambda > 0$

Assume and underlying normal distribution for each of the random variables used to compute $r$ and $z$. 

Figure 11.15: Acceptance and rejection regions for the one-sample z test for a correlation coefficient
The z test is used to test hypotheses about nonzero null correlations, whereas the t test is used to test hypotheses about null correlations of zero.

The z test can also be used to test correlations of zero under the null hypothesis, but the t test is slightly more powerful and is preferred.

However, if $\rho_0 \neq 0$, then the one-sample z test is very sensitive to non-normality of either $x$ or $y$. 
Confidence limits for $\rho$ can be derived based on the approximate normality of Fisher’s $z$ transformation of $r$. Suppose we have a sample correlation coefficient $r$ based on a sample of $n$ pairs of observations. To obtain a two-sided $100\% \times (1-\alpha)$ confidence interval for the population correlation coefficient ($\rho$).

1) Compute Fisher’s $z$ transformation of $r = z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$.

2) Let $z_\rho = \text{Fisher’s } z \text{ transformation of } \rho = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right)$.

A two-sided $100\% \times (1-\alpha)$ confidence interval is given for $z_\rho = (z_1, z_2)$ where

$$z_1 = z - z_{1-\alpha/2} / \sqrt{n-3}$$
$$z_2 = z + z_{1-\alpha/2} / \sqrt{n-3}$$

And $z_{1-\alpha/2} = 100\% \times (1-\alpha/2)$ percentile of an $N(0,1)$ distribution

3) A two-sided $100\% \times (1-\alpha)$ confidence interval for $\rho$ is then given by $(\rho_1, \rho_2)$

$$\rho_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$$
$$\rho_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$$

The interval $(z_1, z_2) = z \pm z_{1-\alpha/2} / \sqrt{n-3}$

_Solving for $r$ in terms of $z$, we get $r = (e^{2z} - 1) / (e^{2z} + 1)$._
Fisher’s z transformation can also be extended to two-sample (independent) problems for comparing two correlation coefficients.

To test the hypothesis $H_0: \rho_1 = \rho_2$ vs. $H_1: \rho_1 \neq \rho_2$, use the following procedure:

1) Compute the sample correlation coefficients $(r_1, r_2)$ and Fisher’s z transformation $(z_1, z_2)$ for each of the two samples.

2) Compute the test statistic

$$\lambda = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0,1) \text{ under } H_0$$

- If $\lambda > z_{1-\alpha/2}$ or $\lambda > -z_{1-\alpha/2}$ reject $H_0$.
- If $-z_{1-\alpha/2} \leq \lambda \leq z_{1-\alpha/2}$ accept $H_0$.

3) The exact $p$-value is given by

$$P = 2\Phi(\lambda) \text{ if } \lambda \leq 0$$

$$P = 2\times[1 - \Phi(\lambda)] \text{ if } \lambda > 0$$

Assume an underlying normal distribution for each of the random variables used to compute $r_1, r_2$ and $z_1, z_2$. 

Two-Sample Test for Correlations
Wolfe’s Test for Comparing Dependent Correlation Coefficients

Previous slide pertains to the comparison of correlation coefficients obtained from two independent samples. In some cases, we are interested in comparing two correlation coefficients obtained from the same subjects. Suppose we want to test the hypothesis $H_0: \rho_{XZ} = \rho_{YZ}$ vs. $H_1: \rho_{XZ} \neq \rho_{YZ}$ where $X, Y, \text{ and } Z$ are obtained from the same subjects. We assume $\sigma_X = \sigma_Y$.

1) Under this assumption, these hypotheses are equivalent to the hypothesis:
   
   $H_0: \rho_{X-Y,Z} = 0$ vs. $H_1: \rho_{X-Y,Z} \neq 0$. Hence,

2) We use the one-sample t test for correlation based on the following test statistic $t = r\sqrt{n-2}/\sqrt{1-r^2} \sim t_{n-2}$ under $H_0$, where $r = \text{Corr}(X_i - Y_i, Z_i)$. We reject $H_0$ if $t > t_{n-2,1-\alpha/2}$ or if $t < t_{n-2,\alpha/2}$.

   We accept $H_0$ if $t_{n-2,\alpha/2} \leq t \leq t_{n-2,1-\alpha/2}$.

3) The $p$-value is given by

   $2 \times Pr(t_{n-2} > t)$ if $t \geq 0$,
   
   $2 \times Pr(t_{n-2} < t)$ if $t < 0$. 

Multiple regression analysis involves determining the relationship between each of the more than one independent variables \((x_1, ..., x_k)\) and the dependent variable \((y)\) after taking into account the remaining independent variables.

Estimation of the regression equation \(y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e\) where \(e\) is an error term that is normally distributed with mean 0 and variance \(\sigma^2\)

If we have \(k\) independent variables \(x_1, ..., x_k\) then a linear-regression model relating \(y\) to \(x_1, ..., x_k\) is of the form \(y = \alpha + \sum_{j=1}^{k} \beta_j x_j + e\)

We estimate \(\alpha, \beta_1, ..., \beta_k\) by \(a, b_1, ..., b_k\) using the same method of least squares, where we minimize the sum of

\[
[y - \hat{y}]^2 = \left[ y - (a + \sum_{j=1}^{k} b_j x_j) \right]^2
\]
An Example of Multiple Regression

Use the SAS PROC REG program to obtain the least squares estimates.
### An Example of Multiple Regression

#### Table 11.10  Least-squares estimates of the regression parameters for the newborn blood-pressure data in Table 11.9 using the SAS PROC REG program

**The REG Procedure**

Model:  MODELL1  
Dependent Variable: sysbp  
Number of Observations Read  16  
Number of Observations Used  16  

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>591.03564</td>
<td>295.51782</td>
<td>48.08</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>13</td>
<td>79.90186</td>
<td>6.14630</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>15</td>
<td>670.93750</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE  2.47917  
R-Square  0.8809  
Dependent Mean  88.06250  
Adj R-Sq  0.8626  
Coeff Var  2.81524  

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| | Standardized Estimate | Squared Partial |
|----------|----|--------------------|----------------|---------|------|------|-----------------------|-----------------|
| Intercept| 1  | 53.45019           | 4.53189        | 11.79   | <.0001 | 0 |                      |                 |
| brthwgts| 1  | 0.12558            | 0.03434        | 3.66    | 0.0029 | 0.35208 | 0.50715               |
| agedys  | 1  | 5.88772            | 0.68021        | 8.66    | <.0001 | 0.83323 | 0.85214               |
Partial Regression Coefficients

Suppose we consider the multiple-regression model

\[ y = \alpha + \sum_{j=1}^{k} \beta_j x_j + e \]

where \( e \) follows a normal distribution with mean 0 and variance \( \sigma^2 \). The \( \beta_j, j = 1, 2..., k \) are referred to as partial-regression coefficients. \( \beta_j \) represents the average increase in \( y \) per unit increase in \( x_j \), with all other variables held constant (or stated another way, after adjusting for all other variables in the model), and is estimated by the parameter \( b_j \).

- Partial regression coefficients differ from simple linear-regression coefficients. The latter represent the average increase in \( y \) per unit increase in \( x \), without considering any other independent variables.

- If there are strong relationships among the independent variables in a multiple-regression model, then the partial-regression coefficients may differ considerably from the simple linear-regression coefficients obtained from considering each independent variable separately.

- The **standardized regression coefficient** \( (b_s) \) is given by \( b \times (s_x/s_y) \). It represents the estimated average increase in \( y \) (expressed in standard deviation units of \( y \)) per standard deviation increase in \( x \), after adjusting for all other variables in the model.

- It is a useful measure for comparing the predictive value of several independent variables because it tells us the predicted increase in standard-deviation units of \( y \) per standard-deviation increase in \( x \).

- By expressing change in standard-deviation units of \( x \), we can control for differences in the units of measurement for different independent variables.
Hypothesis Testing for Multiple Linear Regression: F Test

F test for testing the hypothesis

\( H_0: \beta_1 = \beta_2 = ... = \beta_k = 0 \) vs. \( H_1: \) at least one of the \( \beta_j \neq 0 \) in multiple linear regression

1) Estimate the regression parameters using the method of least squares, and compute Reg SS and Res SS, where

\[
\text{Reg SS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
\text{Res SS} = \text{Total SS} - \text{Reg SS}
\]

\[
\hat{y}_i = a + \sum_{j=1}^{k} b_j x_{ij}
\]

\( x_{ij} = j \text{th independent variable for the } i \text{th subject}, j = 1,...,k; i = 1,...,n \)

2) Compute Reg MS = Red SS/k, Res MS = Res SS/(n-k-1)

3) Compute the test statistic \( F = \text{Reg MS}/\text{Res MS} \) which follows an \( F_{k,n-k-1} \) distribution under \( H_0 \).

4) For a level \( \alpha \) test, if \( F > F_{k,n-k-1,1-\alpha} \) then reject \( H_0 \). if \( F \leq F_{k,n-k-1,1-\alpha} \) then accept \( H_0 \).

5) The exact p-value is given by the area to the right of \( F \) under an \( F_{k,n-k-1} \) distribution = \( Pr(F_{k,n-k-1} > F) \)
Rejection Regions and p-Value for F Test

Acceptance and rejection regions for testing the hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. $H_1$: at least one of the $\beta_j \neq 0$ in multiple linear regression

$p = Pr(F_{k, n-k-1} > F)$, where $F = \text{Reg MS/Res MS}$

Computation of the p-value for testing the hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. $H_1$: at least one of the $\beta_j \neq 0$ in multiple linear regression
Hypothesis Testing for Independent Contribution: t Test

The significant p-values for the previous F test can be attributed to either variable. We would like to perform significance tests to identify the independent contributions of each variable.

t test for testing the hypothesis $H_0: \beta_l = 0$, All other $\beta_j \neq 0$ vs. $H_1: \beta_l \neq 0$, all other $\beta_j \neq 0$ in multiple linear regression

1) Compute $t = b_l / se(b_l)$ which should follow a $t$ distribution with $n - k - 1 \text{ df}$ under $H_0$.

   If $t < t_{n-k-1, \alpha/2}$ or $t > t_{n-k-1, \alpha/2}$ then reject $H_0$.

   If $t_{n-k-1, \alpha/2} \leq t \leq t_{n-k-1, 1-\alpha/2}$ then accept $H_0$.

2) The exact $p$-value is given by

   $2 \times Pr(t_{n-k-1} > t)$ if $t \geq 0$

   $2 \times Pr(t_{n-k-1} \leq t)$ if $t < 0$
Rejection Regions and p-Value for t Test

Figure 11.21  Acceptance and rejection regions for the t test for multiple linear regression

Figure 11.22  Computation of the exact p-value for the t test for multiple linear regression
Partial $F$ test for Partial-Regression Coefficients in Multiple Linear Regression

To test the hypothesis $H_0: \beta_1 = 0$, all other $\beta_j \neq 0$ vs. $H_1: \beta_1 \neq 0$, all other $\beta_j \neq 0$ in multiple linear regression, we

1) Compute $F$ as

$$F = \frac{(\text{Reg SS}_{\text{full model}} - \text{Reg SS}_{\text{all variables except } \beta_1 \text{ in the model}})}{\text{Res MS}_{\text{full model}}}$$

which should follow an $F_{1,n-k-1}$ distribution under $H_0$.

2) The exact $p$-value is given by $Pr(F_{1,n-k-1} > F)$

3) It can be shown that the $p$-value from using the partial $F$ test given in 2) is the same as the $p$-value obtained from using the $t$ test as previous slide.
Residual analysis can be performed as the simple linear regression case. Outliers (with a Studentized residual > 3.0) can be removed for another fitting model.

In a multiple regression model, \( y \) is normally distributed with expected value \( \alpha_l + \beta_l x_l \) and variance \( \sigma^2 \) where \( \alpha_l = \alpha + \beta_1 x_1 + \ldots + \beta_{l-1} x_{l-1} + \beta_{l+1} x_{l+1} + \ldots + \beta_k x_k \).

Given the values of all other independent variables \( (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k) \)

1) The average value of \( y \) is linearly related to \( x_l \)
2) The variance of \( y \) is constant (ie., \( \sigma^2 \))
3) \( y \) is normally distributed.

These assumptions can be validated by a partial-residual plot.

A **partial-residual plot** characterizing the relationship between the dependent variable \( y \) and a specific independent variable \( x_l \) in a multiple-regression setting is constructed as follows:

1) A multiple regression is performed of \( y \) on all predictors other than \( x_l \) (ie., \( x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k \)) and the residuals are saved.
2) A multiple regression is performed of \( x_l \) on all other predictors (ie., \( x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k \)) and the residuals are saved.
3) The partial-residual plot is a scatter plot of the residuals in step 1 on the \( y \) axis against the residuals in step 2 on the \( x \) axis.
Partial-residual Plot Example

If the multiple-regression model holds, then the residuals in step 1 should be linearly related to the residuals in step 2 with slope $= \beta_i$ and variance of $\sigma^2$.

**Partial-residual plot of SBP vs. age in days for the model in Table 11.10**

**Partial-residual plot of SBP vs. birthweight for the model in Table 11.10**
Partial Correlation and Multiple Correlation

Partial Correlation: assess the degree of association between two variables after controlling for other covariates.

Suppose we are interested in the association between two variables x and y but want to control for other covariates $z_1, ..., z_k$. The partial correlation is defined as the Pearson correlation between two derived variables $e_x$ and $e_y$, where

$e_x = \text{the residual from the linear regression of } x \text{ on } z_1, ..., z_k$

$e_y = \text{the residual from the linear regression of } y \text{ on } z_1, ..., z_k$

Multiple Correlation: assess the degree of association between one outcome variable and a linear combination of multiple variables.

Suppose we have an outcome variable $y$ and a set of predictors $x_1, ..., x_k$.

The maximum possible correlation between $y$ and a linear combination of the predictors $c_1 x_1 + ... + c_k x_k$ is given by the correlation between $y$ and the regression function $\beta_1 x_1 + ... + \beta_k x_k$ and is called the multiple correlation between $y$ and $[x_1, ..., x_k]$.

It is estimated by the Pearson correlation between $y$ and $b_1 x_1 + ... + b_k x_k$, where $b_1, ..., b_k$ are the least-squares estimates of $\beta_1, ..., \beta_k$.

The multiple correlation can also be shown to be equivalent to $\sqrt{\text{Reg SS/Total SS}} = \sqrt{R^2}$.
**Rank Correlation and t Test**

**Spearman rank-correlation coefficient** \( (r_s) \) is an ordinary correlation coefficient based on ranks. Thus, \( r_s = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} \) where the L’s are computed from the ranks rather than from the actual scores.

**T test for Spearman Rank Correlation**

1) Compute the test statistic \( t_s = \frac{r_s(\sqrt{n} - 2)}{\sqrt{1 - r_s^2}} \) which under the null hypothesis of no correlation follows a t distribution with \( n-2 \) degrees of freedom.

2) For a two-sided level \( \alpha \) test, if \( t_s > t_{n-2,1-\alpha/2} \) or \( t_s < t_{n-2,\alpha/2} = -t_{n-2,1-\alpha/2} \) then reject \( H_0 \); otherwise, accept \( H_0 \).

3) The exact p-value is given by
   
   \[ p = 2 \times (\text{area to the left of } t_s \text{ under a } t_{n-2} \text{ distribution}) \text{ if } t_s < 0 \]
   \[ p = 2 \times (\text{area to the right of } t_s \text{ under a } t_{n-2} \text{ distribution}) \text{ if } t_s \geq 0 \]

4) This test is valid only if \( n \geq 10 \)
Rejection Regions and p-Value

Acceptance and rejection regions for the t test for a Spearman rank-correlation coefficient

\[ t_s = \frac{r_s \sqrt{n - 2}}{\sqrt{1 - r_s^2}} \]

Acceptance region
\[-t_{n-2, 1 - \alpha/2} \leq t_s \leq t_{n-2, 1 - \alpha/2} \]

Rejection region
\[ t_s < -t_{n-2, 1 - \alpha/2} \]
\[ t_s > t_{n-2, 1 - \alpha/2} \]

Computation of the exact p-value for the t test for a Spearman rank-correlation coefficient

If \( t_s < 0 \), then \( p = 2 \times \) (area to the left of \( t_s \) under a \( t_{n-2} \) distribution).

If \( t_s \geq 0 \), then \( p = 2 \times \) (area to the right of \( t_s \) under a \( t_{n-2} \) distribution).
Interval Estimation for Spearman Rank-Correlation Coefficients

Suppose we have an estimated Spearman rank-correlation \( r_s \) based on a sample of size \( n \). To obtain an approximate two-sided 100\% \( \times(1-\alpha) \) confidence interval for \( \rho_s \) (the underlying rank correlation) we proceed as follows:

1) Compute the sample probit \( H_i \) and \( H_i^* \) corresponding to \( X_i, Y_i \), where \( H_i = \Phi^{-1}(P_i) \), \( H_i^* = \Phi^{-1}(P_i^*) \) and \( P_i = \text{rank}(X_i)/(n+1) \) and \( P_i^* = \text{rank}(X_i)/(n+1) \). The probit has been referred to as the inverse normal distribution. Thus, probit (0.5) = \( z_{.5} = 0 \), probit (0.975) = \( z_{.975} = 1.96 \), etc.

2) Compute the Pearson correlation \( r \) between sample probits given by \( r_h = \text{corr}(H_i, H_i^*) \), which is a sample estimate of the probit correlation \( \rho_h = \text{corr}(H_i, H_i^*) \) where \( H_i = \Phi^{-1}(P_i) \), \( H_i^* = \Phi^{-1}(P_i^*) \).

3) Because \( r_h \) is a slightly negatively biased estimate of \( \rho_h \), we compute the bias-corrected estimator of \( \rho_h \) given by \( r_{\text{cor},h} = r_h\{1+(1-r_h^2)/(2(n-4))\} \).

4) Let \( z_h = \text{Fisher’s } z \text{ transform of } \rho_h \equiv 0.5 \ln[(1+ \rho_h)/(1- \rho_h)] \).
5) Compute a 100% \times (1-\alpha) confidence interval for \( z_h \) given by \((z_{1h}, z_{2h}) = z_h \pm \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \) where \( z_h = \text{Fisher's z-transform of } r_{cor,h} = 0.5 \ln[(1-r_{cor,h})/(1-r_{cor,h})] \).

6) The corresponding 100% \times (1-\alpha) confidence interval for \( \rho_h \) is \((r_{1h}, r_{2h})\), where \( r_{1h} = \frac{\exp(2z_{1h})-1}{\exp(2z_{1h})+1} \), \( r_{2h} = \frac{\exp(2z_{2h})-1}{\exp(2z_{2h})+1} \).

7) Furthermore, a 100% \times (1-\alpha) confidence interval for \( \rho_s \) is \((r_{s1}, r_{s2})\), where \((r_{s1}, r_{s2}) = \left[\left(\frac{6}{\pi}\right)\sin^{-1}(r_{1h}/2),\left(\frac{6}{\pi}\right)\sin^{-1}(r_{2h}/2)\right]\).

8) This procedure is valid for \( n \geq 10 \). The rationale for this procedure is that for normally distributed scales such as \( H \) and \( H^* \), there is a relationship between the underlying rank correlation and Pearson correlation given by \( \rho_{s,h} = (6/\pi)\sin^{-1}(\rho_h/2) \) where \( \rho_h = \text{corr}(H_i, H_i^*) \) and \( \rho_{s,h} = \text{corr}(P_i, P_i^*) \). However, because the probit transformation is rank-preserving, \( P_i \) and \( P_i^* \) are the same in the probit scale and the original scale. Thus, \( \rho_{s,h} = \rho_s = \text{corr}(P_i, P_i^*) \)
In this lecture for Chapter 11, we discussed

- Statistical inference methods for investigating the relationship between two or more variables.

- If only two variables, both of which are continuous, are being studied, and we wish to predict one variable (the dependent variable) as a function of the other variable (the independent variable) then simple linear regression analysis is used.

- Pearson correlation methods are used to determine the association between two normally distributed variables without distinguishing between dependent and independent variables.

- Rank correlation may be used if both variables are continuous but not normally distributed or are ordinal variables.

- Multiple regression methods may be used to predict the value of one variable (the dependent variable which is normally distributed) as a function of several independent variables.
The End