Outline

Correlation
- Definition and Estimation
- Interpretation

Linear Regression
- Simple Linear Regression
- Interpretation of Linear Model
- Least Squares Method
Part I

Review and a Bit More about Correlation
Correlation

- $\sigma_{XY}^2$ is the covariance between $X$ and $Y$:

  $$\sigma_{XY}^2 \equiv \text{Cov}(X, Y) \equiv E(X - \mu_X)(Y - \mu_Y).$$

- Pearson’s correlation $\rho$ measures the linear association between two random variables $X$ and $Y$, and equals

  $$\rho = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}. $$
Estimation of Correlation

- $\sigma_{XY}^2$ is estimated by the sample covariance:

$$s_{XY}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- The correlation is in turn estimated by sample correlation:

$$r = \frac{s_{XY}^2}{s_X s_Y} = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_X} \right) \left( \frac{y_i - \bar{y}}{s_Y} \right),$$

where $s_X$ and $s_Y$ are the sample standard deviations. Note the second expression reveals $r$ is the covariance of the standardized observations (the centered and scaled $x_i$ and $y_i$).
Interpretation of Correlation

- If \( \rho > 0 \), then \( X \) and \( Y \) are positively correlated; that is, \( X \) and \( Y \) tend to get bigger or smaller at the same time.

- If \( \rho < 0 \), then \( X \) and \( Y \) are negatively correlated; that is, when \( X \) gets bigger, \( Y \) tends to get smaller, and vice versa.

- If \( \rho = 0 \), then \( X \) and \( Y \) are uncorrelated. We cannot say whether \( Y \) will get bigger or not when \( X \) increases.

- If \( X \) and \( Y \) are independent, then \( \rho = 0 \). The converse is not necessarily true.
Interpretation of Correlation

- $-1 \leq \rho \leq 1$. The closer to 1 or $-1$ $\rho$ is, the stronger the (linear) association.
- $\rho$ does not depend on the scale of the measurements.
- Just like $\bar{X}$ and $s_X$, the sample correlation $r$ is a random variable, and we can estimate its variance.
- As a result, we can test $H_0 : \rho = 0$ based on the data.
- Note that a “strong” correlation is not necessarily statistically significant, and vice versa.
Factors that Influence the Sample Correlation \( r \)

- Non-linear relationship.
- Skewed distribution (non-normality).
- Outliers.
Limitations of Correlation

- \( r \) does not provide an estimate of the effect of \( X \) on \( Y \). As such, it is most often used as quick explanatory tool for summarizing the apparent linear relationship in the data, rather than for formal statistical inference.

- \( r \) can only be used on a pair of variables.

- \( r \) cannot deal with nonlinear relationships.
Part II

Introduction to Linear Regression
Simple Linear Regression

- Often we have a continuous outcome (response) variable, and a single predictor (explanatory) variable.
- The predictor need not be random (e.g., we give rats a drug at several fixed doses, and measure the outcomes).
- Note that unlike correlation, now we are not treating the two variables symmetrically: one is the predictor, and the other is the response.
A regression model is often used to model the distribution of the outcome variable (Y) for given values of the predictor variable (X).

A linear regression model is simplest:

\[ Y = a + bX + \epsilon, \]

where \( a \) is the intercept, \( b \) is the slope, and \( \epsilon \) is the error. Generally we assume \( \epsilon \sim \mathcal{N}(0, \sigma^2) \).

The above model also be written as \( Y|X \sim \mathcal{N}(\mu_{Y|X}, \sigma_{Y|X}^2) \), where

\[ \mu_{Y|X} = a + bX \text{ and } \sigma_{Y|X}^2 = \sigma^2. \]
The Four Key Assumptions of Linear Regression

- The observations \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) are independent.
- For a given value of \(X\), \(Y\) has a normal distribution.
- The variance of \(Y\) does not depend on \(X\); this is called homogeneity of variance, or homoscedasticity (ooh!)
- The mean of \(Y\) is a linear function of \(X\).
Interpretation of Linear Model

- Even though the linear model is unlikely to be exactly true, it is often a good approximation.
- It is a model for the mean of $Y$. For a given $x$, $Y$ is a (normal) random variable with mean $a + bx$ and variance $\sigma^2$.
- The intercept $a$ is the mean of $Y$ when $x = 0$. It is often not of interest, since $X$ may not be able to actually take values near 0.
- The slope $b$ is the increase of the mean of $Y$ when $x$ increases one unit. It is usually the parameter of interest. In particular, we are often interested in testing $H_0 : b = 0$. Note that $b$'s magnitude also depends on the scale of $Y$. 
Misinterpretations of the Regression Line

These statements are all **untrue:**

- If \( b = 0 \), \( X \) and \( Y \) are independent.
- If \( b \neq 0 \), then \( X \) and \( Y \) have a linear relationship.
- Since \( X \) and \( Y \) (almost) never have an exactly linear relationship, linear regression is useless.
- Regression allows us to safely predict \( Y \) for any value of \( x \), even those beyond the range of \( x \)-values spanned by the data (**extrapolation**).
Example: Birth weight vs. Estriol levels in moms

![Scatter plot showing birth weight vs. estriol levels in moms. The plot suggests a positive correlation between the two variables.](image-url)
Least Squares Method

- Recall the error or residual is:

  \[ e_i = y_i - (a + bx_i) \]

- The least-squares (LS) regression line chooses \( \hat{a} \) and \( \hat{b} \) to minimize the residual sum of squares:

  \[ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 \]

- That is, the LS line is the line that makes the sum of the squares of the vertical distances from the data points to the line as small as possible.
LS line for Birth weight vs. Estriol data
SAS Proc REG code for Birth Weight vs. Estriol

```sas
proc reg data = estriol rsquare;
   model bweight = estriol;
   plot bweight * estriol;
run;
```

The REG Procedure  
Model: MODEL1  
Dependent Variable: bweight  

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>1</td>
<td>253.29609</td>
<td>253.29609</td>
<td>17.46</td>
<td>0.0002</td>
</tr>
<tr>
<td>Error</td>
<td>29</td>
<td>420.70391</td>
<td>14.50703</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>674.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE           3.80881  
R-Square           0.3758   
Dependent Mean     32.00000  
Adj R-Sq            0.3543   
Coeff Var          11.90253
### Parameter Estimates

| Variable | DF | Estimate    | Standard Error | t Value | Pr > |t| |
|----------|----|-------------|----------------|---------|-------|-----|
| Intercept| 1  | 21.46651    | 2.61202        | 8.22    | < .0001 |
| estriol  | 1  | 0.61035     | 0.14607        | 4.18    | 0.0002 |
Connecting Correlation and Simple Linear Regression

Recall that the fitted least squares regression line is

\[ \hat{Y}_i = \hat{a} + \hat{b}x_i , \]

where \( \hat{a} \) and \( \hat{b} \) are the least squares estimates of the intercept and slope, and \( \hat{Y}_i \) is the fitted value of the line at \( x_i \).

It turns out that

\[ \hat{b} = r \frac{s_y}{s_x} \quad \text{and} \quad \hat{a} = \bar{y} - \hat{b}\bar{x} . \]

That is, the fitted slope is a rescaled version of the correlation!

Thus a change of one \textit{standard deviation} in \( X \) corresponds to a change of \textit{r standard deviations} in \( Y \).
Connecting Correlation and Simple Linear Regression

- The connection between $r$ and $\hat{b}$ is a special property of least-squares regression that does not hold for other methods of fitting a line to data.

- This (along with its connection to the normal distribution) is a lot of the reason behind the continuing popularity of least-squares regression, despite its sensitivity to outliers (which lead to huge squared differences, and thus exert undue influence on the fitted line).
Connecting Correlation and Simple Linear Regression

- The square of the correlation, $r^2$, is the fraction of the variation in the values of $Y$ that is explained by the least squares regression of $Y$ on $X$.
- More specifically,

$$r^2 = \frac{\text{Variance of predicted values } \hat{Y}_i}{\text{Variance of observed values } Y_i},$$

which of course preserves the property that $0 \leq r^2 \leq 1$.
- So $r^2$ is often thought of as a measure of goodness of fit of the fitted regression.
- See for example the pictures on p.145 of M& M; note that “noisier” data will lead to lower $r^2$ values!
Inference Tasks in Linear Regression Modeling

- **Estimation:** What is an estimate of the effect size $b$ (the regression coefficient) and a corresponding confidence interval?

- **Testing:** Is $b$, the regression coefficient of $X$, significantly different from zero?

- **Prediction:** What can we say about the distribution of $Y$ for a given new value of $x$?

- **Model diagnosis:** Are the four assumptions (independence, normality, linearity, and homoscedasticity) justified for our data? Also, are their outliers or other influential points?