Notes

STA302H1 - Fall 2020

Ziyue Yang

December 14, 2020

Contents

1 Module 1 - Introduction to Data Analysis

1.1 What is SLR?

1.1.1 Basic Concepts of a Regression Model

A regression model is a tendency of the response variable Y to vary with the predictor X in a systematic fashion. It's usually a scattering of points around the curve of statistical relationship.

There is a probability distribution of Y for each level of X. The means of these probability distributions vary in some systematic fashion with X.

2 Module 2 - The SLR Model

2.1 What is a Linear Model?

Definition 2.1 (General Form of Models)**.** General form of a model for Y in terms of three predictors:

$$Y = f(X_1, X_2, X_3) + e \tag{2.1}$$

- f is some unknown function
- e is the error not accounted for f

> Notice there issue here: if f is a smooth, continuous function, then there are many possibilities for f. Also, we would need infinite data to estimate f directly.
>
> A solution is to restrict f to a *linear form.*

Definition 2.2 (Linear Model)**.** In a linear model for Y, the *parameters enter linearly* or *Y is linear in terms of the parameters.*

Example 2.1. Here are some examples of linear models

$$Y = \beta_0 + \beta_1 x + e \tag{2.2}$$
$$Y = \beta_0 + \beta_1 \log(x) + e \tag{2.3}$$
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e \tag{2.4}$$
$$Y = \beta_0 + \beta_1 \log(x_2) + \beta_2 x_2 + \beta_3 x_1 x_2 + e \tag{2.5}$$
$$Y = \beta_0 x^{\beta_1} e \tag{2.6}$$
$$Y = \exp(\beta_0 + \beta_1 x + e) \tag{2.7}$$

> Tip: apply suitable transformations to Y to see that the model is linear.
>
> Take (2.6) as an example, we can perform a log-transformation such that
>
> $$\log(Y) = \log(\beta_0) + \beta_1 \log(x) + \log(e) \tag{2.8}$$

Example 2.2. Here are some examples of non-linear models

$$Y = \beta_0 + \exp(\beta_1 x) + e \tag{2.9}$$
$$Y = \exp(\beta_0 + \exp(\beta_1 x)) + e \tag{2.10}$$
$$Y = \beta_0 + \beta_1 x \beta_2 + e \tag{2.11}$$
$$Y = \beta_0 + \beta_1 x - \exp(\beta_2 + \beta_3 x) + e \tag{2.12}$$

2.1.1 Linear and Nonlinear Models

True nonlinear models are rare. Linear models can handle complex datasets. Since predictors can be transformed and combined in many ways, lienar models are *very flexible.*

Note that all straight lines are linear models, but all linear models are not just straight lines.

2.2 Simple Linear Regression Models

$$Y = \beta_0 + \beta_1 X + e \tag{2.13}$$

- Y - dependent or response or output variable
- X - independent or explanatory or predictor or input variable
- β_0 - intercept parameter
- β_1 - slope parameter
- e - random error/noise, variation in measures that we cannot account for

2.3 Estimating the Regression Parameters

Fitting an SLR Model

We wish to fit the SLR model

$$Y = \beta_0 + \beta_1 X + e \tag{2.14}$$

Our aim is, given a specific value of X, i.e. $X = x$, find the expected value of Y,

$$\mathbb{E}(Y \mid X = x) \tag{2.15}$$

We need estimates of the regression parameters β_0, β_1, and we need to assess the fit.

Get data (observational or experimental):

n pairs of bivariate data:

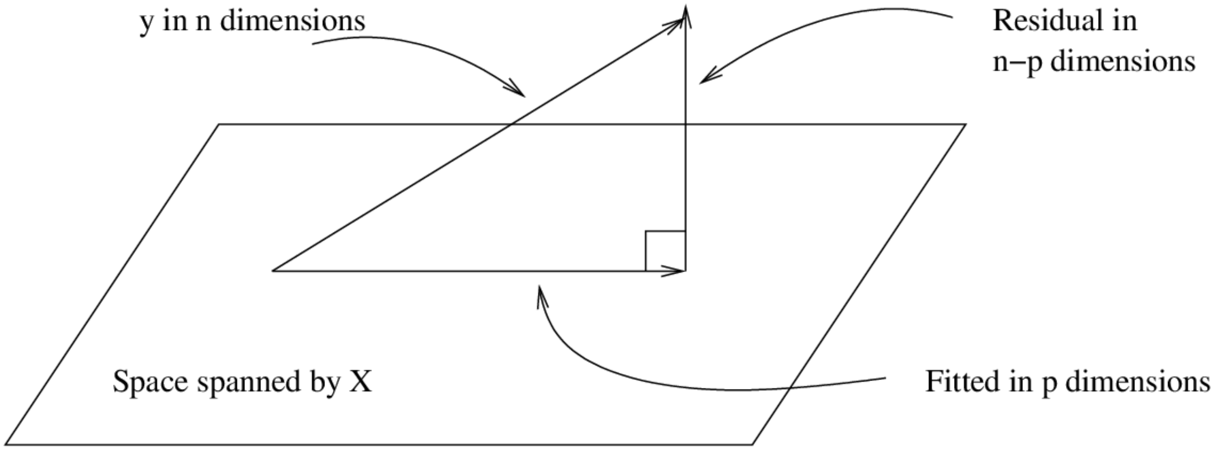
$$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n) \tag{2.16}$$

Notation:

Estimators: $\hat{\beta}_0, \hat{\beta}_1$

Estimates: b_0, b_1

Geometrical Representation of Estimating β



- The response Y is an n-dimensional space, $\boldsymbol{Y} \in \mathbb{R}^n$
- The regression parameters are in a $p+1$-dimensional space, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, where p is the number of predictors, so $p+1$ is the number of regression parameters: $p < n$.

2.3.1 The Least Squares Method

Definition 2.3 (**Least Squares Method**)**.** Consider

$$RSS = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \tag{2.17}$$

This is called the 'least squares criterion'.

The **least squares method** aims to find the esitmators b_0, b_1 such that RSS is minimized.

For each x_i, the predicted/fitted value is

$$\hat{y_i} = b_0 + b_1 x_i \tag{2.18}$$

the residuals are

$$\hat{e_i} = y_i - \hat{y_i} \tag{2.19}$$

Question 2.1. Why are we using vertical distances instead of horizontal distances?

We want to predict Y from X and so we want $\hat{y_i}$ to be as close as possible to y_i.

Note that *regression is not symmetric*. If we minimize the horizontal distances, we will get a different answer for b_0 and $b - 1$. It matters which variable is dependent and which is independent.

Why squared deviations?

1. Square deviations make NO satistical assumptions

2. MSE is the most common way to measure error in statistics

3. LS estimators have *good* properties

Example 2.3. Often, raw deviation $\sum_{i=1}^{n}(y_i - \hat{y}_i) \equiv 0$.

Analytical Derivations

Consider the least square criterion

$$RSS = \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2 \tag{2.20}$$

Using calculus to minimize RSS, we get the normal equations:

wrt b_0:

$$\begin{align}
\frac{\partial RSS}{\partial b_0} &= -\sum(y_i - b_0 - b_1 x_i) \tag{2.21}\\
&\overset{\text{set}}{\equiv} 0 \tag{2.22}\\
\implies \sum y_i &= \sum b_0 + \sum b_1 x_i \tag{2.23}\\
\implies n\bar{y} &= nb_0 + b_1 n\bar{x} \tag{2.24}\\
\implies b_0 &= \bar{y} - b_1\bar{x} \tag{2.25}
\end{align}$$

wrt b_1:

$$\begin{align}
\frac{\partial RSS}{\partial b_1} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i \tag{2.26}\\
&\overset{\text{set}}{\equiv} 0 \tag{2.27}\\
\implies \sum x_i y_i &= \sum b_0 x_i + \sum b_1 x_i^2 \tag{2.28}\\
\implies \sum x_i y_i &= (\bar{y} - b_1\bar{x})\sum x_i + b_1 \sum x_i^2 \qquad \text{substitute (2.25)} \tag{2.29}\\
\implies b_1 &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \tag{2.30}
\end{align}$$

Therefore the **normal equations** are

$$\boxed{b_0 = \bar{y} - b_1\bar{x}} \tag{2.31}$$

$$\boxed{b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}} \tag{2.32}$$

2.3.2 The Maximum Likelihood Estimator (MLE)

Parameter $\boldsymbol{\theta}$, Estimator $\hat{\boldsymbol{\theta}}_{MLE}$

Steps for MLE:

1. Define the **likelihood function** as a function of the parameter(s) $\boldsymbol{\theta}$:

$$\mathcal{L}(\boldsymbol{\theta}) = Distribution(Y \mid \boldsymbol{\theta}) \tag{2.33}$$

   conseidered a working model of the parameter given the specific data

2. Find the **value of the parameter that maximizes the likelihood function**; i.e. the estimator that gives the highest probability density to the observed data

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\theta} \mathcal{L}(\boldsymbol{\theta}) \tag{2.34}$$

MLE Properties

- Regularity conditions are required to derive the asymptotic distribution of the MLE
- Inference follows the frequentist paradigm
- MLE's have nice properties:
  - Asymptotically Unbiased
  - Consistent
  - Sufficient
  - Have minimum variance
  - Invariant principle holds

Example 2.4. Consider a normal likelihood for Y interms of the parameters $\boldsymbol{\beta}, \sigma^2$

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}_n(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

Using calculus we get

$$\hat{\beta}_{0,MLE}, \hat{\beta}_{1,MLE} \text{ same as in LS method}$$
$$\hat{\sigma}^2_{MLE} = \frac{\sum (y_i - \hat{y}_i^2)}{n}$$

2.3.3 Bayesian Approach

In the Bayesian method, the parameters are considered random - instead of fixed constants

$$p(\boldsymbol{\beta}), p(\sigma^2)$$

1. Therefore, the parameters have a **prior distribution**. Priors could be either *proper* or *improper* (check out STA261 notes for further details).

2. Assume a **likelihood** for Y, as a function of the parameters

$$\mathcal{L}(\boldsymbol{beta}, \sigma^2) = Distribution(Y \mid \boldsymbol{\beta, \sigma^2})$$

3. Derive the **posterior distribution** of the parameters given the data

$$p(\boldsymbol{\beta}, \sigma^2 \mid y) \propto \mathcal{L}(\boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta}, \sigma^2) \tag{2.35}$$

> We obtain **credible** (rather than confidence) intervals for $\boldsymbol{\beta}$. Note that the interpretation differs.
>
> With a credible interval, we speak about the probability that the unknown parameter falls into the interval.
>
> Computationally, the Bayesian approach is more challenging than the LS/ML approach.

Example 2.5. An example of Bayesian approach

1. Choose improper prior
$$\pi(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}) \times p(\sigma^2) \propto \sigma^2$$

2. Assume a likelihood for Y
$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}_n(\boldsymbol{x\beta}, \sigma^2 \mathbf{I}_n)$$

3. The posterior distribution of $\boldsymbol{\beta}$ given the data is the kernel of a $(k+1)$-dimensional t distribution

> Results
>
> Posterior mean results are identical to the LS approach.
>
> Posterior mean results are identical to the ML approach *under normality.*
>
> $100(1-\alpha)\%$ credible intervals yield the same results as the $100(1-\alpha)\%$ confidence intervals, though the interpretations are different.

2.4 Properties of Least-Square Estimators

2.4.1 Properties of the Fitted Line

Intercept parameter estimate (2.25)

$$b_0 = \bar{y} - b_1 \bar{x} \tag{2.36}$$

Slope parameter estimate

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SXY}{SXX} \tag{2.37}$$

> *Derivation of* (2.37).
>
> Numerator
>
> $$\begin{aligned} \sum(x_i - \bar{x})(y_i - \bar{y}) &= \sum(x_i - \bar{x})y_i - \sum(x_i - \bar{x})\bar{y} \\ &= \sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x}) \end{aligned}$$
>
> Note that $\sum(x_i - \bar{x}) \equiv 0$, since
>
> $$\begin{aligned} \sum x_i - \sum \bar{x} &= \sum x_i - n\bar{x} \\ &= \sum x_i - n\left(\frac{\sum x_i}{n}\right) = 0 \end{aligned}$$
>
> Same applies, $\sum(y_i - \bar{y}) \equiv 0$.
>
> Denominator
>
> $$\begin{aligned} \sum(x_i - \bar{x})^2 &= \sum(x_i - \bar{x})x_i - \sum(x_i - \bar{x})\bar{x} \\ &= \sum x_i^2 - \bar{x}\sum x_i - \bar{x}\sum(x_i - \bar{x}) \\ &= \sum x_i^2 - n\bar{x}^2 \qquad \text{since } \sum x_i = n\bar{x} \end{aligned}$$

2.4.2 Interpreting Regression Parameter Estimates

Slope, b_1

When x changes by 1 unit, the corresponding average change in y is the slope.

Intercept, b_0

The average value of y when $x = 0$. (No parctical interpretation unless 0 is within the range of the predictor (x) values.)

2.4.3 Properties of Fitted LS Regression Line

Consider a fitted line

$$\hat{y} = b_0 + b_1 x$$

1. The average of the residuals is always 0, i.e.

$$\sum_{i=1}^{n} \hat{e}_i \equiv 0 \tag{2.38}$$

   Proof.

$$\begin{aligned}\sum(y_i - \hat{y}_i) &= \sum(y_i - b_0 - b_1 x_i)\\ &= \sum y_i - nb_0 - b_1 \sum x_i\\ &= n\bar{y} - n(\bar{y} - b_1\bar{x}) - b_1 n\bar{x}\\ &= 0\end{aligned}$$

□

2. The Sum of Squares of Residuals is NOT 0, *unless the fit to the data is perfect.*

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 \neq 0 \tag{2.39}$$

   This criteria is a minimum as reqiuired by the LS method.

3. $\sum \hat{e}_i x_i = 0$
4. $\sum \hat{e}_i \hat{y}_i = 0$
5. $\sum \hat{y}_i = \sum y_i$ (sum of fitted values = sum of observed)

2.4.4 Gauss-Markov Theorem

Theorem 2.1 (Gauss-Markov Theorem)**.** Under the conditions of the SLR model, the least-squares parameter estimators are BLUE ("Best Linear Unbiased Estimators").

- Parameter θ; Estimator $\hat{\theta}$
- Linear - linear in parameters
- Unbiased, $\mathbb{E}(\theta)$, i.e. does NOT overestimate or underestimate systematically
- "Best" - Obtain **minimum variance** among all unbiased linear estimators

2.4.5 Properties of Slope Estimator: Expectation

Recall from (2.37) that

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{SXY}{SXX} \tag{2.40}$$

Since $\sum (x_i - \bar{x}) = 0$,

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i \bar{x}) y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x}) y_i$$

Let $c_i = \frac{x_i - \bar{x}}{SXX}$, then rewrite b_1 as

$$\boxed{b_1 = \sum_{9=1}^n c_i y_i} \tag{2.41}$$

Treat X's as fixed, the the mean of slope estimate, b_1, is

$$\begin{aligned}
\mathbb{E}(b_1 \mid X) &= \mathbb{E}\left[\sum_{i=1}^n c_i y_i \mid X = x_i\right] \\
&= \sum c_i \mathbb{E}(y_i \mid X = x_i) \\
&= \sum c_i (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \sum c_i + \beta_1 \sum c_i x_i,
\end{aligned}$$

Note that

1. $\sum c_i = \sum (x_i - \bar{x})/SXX = 0$
2. $\sum c_i x_i = \sum (x_i - \bar{x}) x_i / SXX = SXX/SXX = 1$

$$\implies \boxed{\mathbb{E}(b_1 \mid X) = \beta_1} \qquad \text{i.e. } b_1 \text{ is unbiased.} \tag{2.42}$$

2.4.6 Properties of Intercept Estimator: Expectation

Recall that $b_0 = \bar{y} - b_1 \bar{x}$

Mean of intercept estimate, b_0

$$
\begin{aligned}
\mathbb{E}(b_0 \mid X) &= \mathbb{E}[(\bar{y} - b_1\bar{x})\; X = x_i] \\
&= \mathbb{E}(\bar{y} \mid X = x_i) - \bar{x}\mathbb{E}(b_1 \mid X = x_i) \\
&= \mathbb{E}\left(\frac{\sum y_i}{n} \mid X = x_i\right) - k\bar{x}\beta_1 \\
&= \frac{1}{n}\sum \mathbb{E}(y_i \mid X = x_i) - \bar{x}\beta_1 \\
&= \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 \\
&= \frac{1}{n}(n\beta_0 + \beta_1 \sum x_i) - \bar{x}\beta_1 \\
&= \beta_0 + \beta_1\frac{n\bar{x}}{n} - \bar{x}\beta_1 \\
&= \beta_0
\end{aligned}
$$

Hence b_0 is unbiased for β_0.

2.4.7 Properties of Slope Estimator: Variance

Variance of slope estimate, b_1

$$
\begin{aligned}
Var(b_1 \mid X) &= Var\left[\sum_{i=1}^{n} c_i y_i \mid X = x_i\right] \qquad (2.43) \\
&= \sum c_i^2 Var(y_i \mid X = x_i) \qquad (2.44) \\
&= \sum c_i^2 \sigma^2 \qquad (2.45) \\
&= \sigma^2 \sum c_i^2 \qquad (2.46) \\
&= \sigma^2 \sum \frac{(x_i - \bar{x})^2}{(SXX)^2} \qquad (2.47) \\
&= \sigma^2 \frac{SXX}{(SXX)^2} \qquad (2.48) \\
&= \frac{\sigma^2}{SXX} \qquad (2.49)
\end{aligned}
$$

2.4.8 Properties of Intercept Estimator: Variance

Vairance of intercept estimate, b_0

$$
\begin{aligned}
Var(b_0 \mid X) &= Var[(\bar{y} - b_1\bar{x}) \mid X = x_i] && (2.50)\\
&= Var(\bar{y} \mid X = x_i) + Var(b_1\bar{x} \mid X = x_i) - 2Cov[(\bar{y}, b_1\bar{x}) \mid X = x_i] && (2.51)\\
&= \frac{\sigma^2}{n} + \bar{x}^2\left(\frac{\sigma^2}{SXX}\right) - 2\bar{x}Cov\left[\left(\frac{1}{n}\sum y_i, \sum c_i y_i\right) \mid X = x_i\right] && (2.52)\\
&= \frac{\sigma^2}{n} + \bar{x}^2\left(\frac{\sigma^2}{SXX}\right) - 2\frac{\bar{x}}{n}Cov\left[(y_i, y_i) \mid X = x_i\right] && (2.53)\\
&= \frac{\sigma^2}{n} + \frac{\bar{x}\sigma^2}{SXX} - \frac{2\bar{x}}{n}\sum c_i Var(y_i \mid X = x_i) && (2.54)\\
&= \frac{\sigma^2}{n} + \frac{\bar{x}\sigma^2}{SXX} - 2\frac{\bar{x}}{n}\sigma^2\sum c_i && (2.55)\\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right) && (2.56)
\end{aligned}
$$

2.5 Statistical Assumptions of SLR

2.5.1 SLR Assumptions

1. We assumed that Y is related to x by the SLR model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1 \ldots n$$

OR

$$\mathbb{E}(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$$

In other words, **the linear model is appropriate**.

And the following three Gauss-Markov Conditions:

2. **The errors have mean of 0**, i.e.,

$$\mathbb{E}(e_i) = 0 \quad i = 1 \ldots n$$

3. **The errors have a common variance** σ^2, i.e.

$$Var(e_i) = \sigma^2 \quad i = 1 \ldots n$$

Meaning the variation is the same for all observations, i.e. **homoscedastic**.

4. **The errors are uncorrelated**, i.e.

$$Cov(e_i, e_j) = 0 \quad \forall i \neq j$$

2.5.2 Estimating Variance of Random Error Term

The random error e_i has mean 0 and variance σ^2. The parameter σ^2 is another parameter of the SLR model.

Our goal is to estimate σ^2, in order to measure the variability of our estimates of Y and carry out inference on our model.

Notice that

$$e_i = Y_i - (\beta_0 + \beta_1 x_i) = Y_i - \text{unknown regression line at } x_i \tag{2.57}$$

Replacing β_0, β_1 by their LS estimates, we estimate the errors by

$$\hat{e}_i = Y - (b_0 + b_1 x_i) = Y_i - \text{unknown regression line at } x_i \tag{2.58}$$

Using the estimated errors, we can show that an *unbiased* estimate of σ^2 is

$$\boxed{S^2 = \frac{\sum \hat{e}_i^2}{n-2} = \frac{RSS}{n-2}} \tag{2.59}$$

2.5.3 Statistical Assumption for Inference

In order to make inferences, we need one more assumption about the errors, e_i's.

Assume that **the errors are Normally distributed**,

$$e_i \sim \mathcal{N}(0, \sigma^2)$$
$$\text{OR}$$
$$\boldsymbol{e} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_n)$$

$\star$**Implications**

1. The Normality assumption implies that the errors are independent (since they are uncorrelated).
2. Since $y_i = \beta_0 + \beta_1 x_i + e_i$, $Y_i \mid x_i$ is normally distributed.
3. The LS estimates of β_0 and β_1 are equivalent to their MLE's.

2.5.4 Sampling Distributions of Slope and Intercept Estimators

Slope: Since $b_1 = \sum c_i y_i$ is a lincomb of the y_i's, $b_1 \mid x$ is also normally distributed, i.e.

$$\hat{\beta}_1 = \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SXX}\right) \tag{2.60}$$

Intercept: Since $b_1 \,|\, X$ is normally distributed, $\bar{y}$ is normally distributed and b_0/x is a lincomb of $b_1 \mid X$ and $\bar{y}$, we have that

$$\hat{\beta}_0 \sim \mathcal{N}\left[\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)\right] \tag{2.61}$$

3 Module 3 - Theory and Inference for SLR I

3.1 Mean and Variance of a Sample Mean

Definition 3.1 (The Sample Mean)**.** Let $Y_1 \dots Y_n$ be a random sample of size n from a population. Then the sample mean is defined as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{3.1}$$

Mean of $\bar{Y}$:

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) &= \frac{1}{n}\mathbb{E}\left(\sum Y_i\right) && (3.2)\\
&= \frac{1}{n}\sum\left(\mathbb{E}(Y_i)\right) && (3.3)\\
&= \frac{1}{n}\sum \mu && (3.4)\\
&= \frac{n\mu}{n} = \mu && (3.5)
\end{aligned}$$

Variance of $\bar{Y}$:

$$\begin{aligned}
Var\left(\frac{1}{n}\sum Y_i\right) &= \left(\frac{1}{n}\right)^2 Var\left(\sum Y_i\right) && (3.6)\\
&= \left(\frac{1}{n}\right)^2 \sum Var(Y_i) \qquad \text{since } Y_i\text{'s iid} && (3.7)\\
&= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} && (3.8)
\end{aligned}$$

Theorem 3.1 (Chi-Square Random Variable)**.** Let $Y_1 \dots Y_n$ be a random sample of size n from a normal distribution with mean $\mathbb{E}(Y_i) = \mu$ and variance $Var(Y_i) = \sigma^2$, and for $i = 1 \dots n$, define Z_i by

$$Z_i = \frac{Y_i - \mu}{\sigma} \tag{3.9}$$

Then $\sum Z_i^2$ has a χ^2 distribution with n degrees of freedom (df).

> In other words, the square of a standard normal random variable is a chi-square random variable with 1 df.
>
> A lincomb of independent chi-square random variables leads to another chi-square random variable. We can use MGFs to show these properties.

Question 3.1. Suppose that $Y_1 \dots Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$. Then for $s^2 = frac1n - 1\sum(Y_i - \bar{Y})^2$ (the sample variance), which distribution does the following term follow?

$$\frac{(n-1)s^2}{\sigma^2}$$

- $\mathbb{E}(s^2) = \sigma^2$, hence …

4 Module 4 - Theory and Inference for SLR, Part II

4.1 Regression Analysis of Variance (ANOVA) Approach to Regression

Q. How well does the regression line summarize the data?

(Figure here)

Notice that $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$. Squaring both sides yields

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \tag{4.1}$$

Now we define and derive $SST, SSReg, RSS$

$$\begin{align} SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}\hat{Y}_i(Y_i - \hat{Y}_i) + 2\sum_{i=1}^{n}(Y_i - \bar{Y}_i)(-\bar{Y}) \tag{4.2} \\ = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 + 2\sum\hat{Y}_i\hat{e}_i - 2\bar{Y} \tag{4.3} \end{align}$$

> Total Sum of Squares

$$\begin{align} SSReg &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{4.4} \\ &= b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4.5} \end{align}$$

> Model SS or Regression SS
>
> Amount of variation in y's explained by regression line

$$\begin{align} RSS &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.6} \\ &= \sum_{i=1}^{n}\hat{e}_i^2 \tag{4.7} \end{align}$$

> Residual SS or Error SS
>
> Least square criterion
>
> Unexplained variation in y's

Decomposition of Sums of Squares

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = SSReg + RSS \tag{4.8}$$

$$= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.9}$$

$$= b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n} \hat{e}_i \tag{4.10}$$

$$= b_1^2 SXX + RSS \tag{4.11}$$

> $\sum_{i=1}^{n} \hat{e}_i(x_i - \bar{x}) = 0$ since
>
> - $\sum \hat{e}_i = 0$
> - $\sum \hat{e}_i x_i = 0$

Source	SS	df	MS = SS/df
Regression Line	$SSReg = b_1^2 SXX$	1	$\frac{b_1^2 SXX}{1}$
Error	$RSS = \sum \hat{e}_i^2$	$n-2$	$\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2} = S^2$
Total	SST	$n-1$	

4.1.1 The F-Distributions

Definition 4.1. Suppose V and W are independent random variables such that $V \sim \chi^2_\nu$ and $W \sim \chi^2_\omega$. Then

$$\frac{V/\nu}{W/\omega} \sim F(\nu, \omega) \tag{4.12}$$

Theorem 4.1 (Equivalent Test of $H_0 : \beta_1 = 0$)**.** Under $H_0, \beta_1 = 0$,

$$F_{obs} = \frac{MSReg}{MSE} = \frac{SSReg/1}{SSE/(n-2)} \sim F(1, n-2) \tag{4.13}$$

Theorem 4.2. Suppose $Q \sim T_\nu$. Then

$$Q^2 \sim F(1, \nu) \tag{4.14}$$

4.2 Connection Between Regression and Correlation Analysis

4.2.1 The Bivariate Normal Distribution

Definition 4.2 (Jointly Distributed)**.** X, Y are **jointly distributed** if their join density function is

$$f(x, y) = \frac{e^{-Q/2}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}, \qquad -\infty < x < \infty, -\infty < y < \infty \tag{4.15}$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right] \tag{4.16}$$

and the **correlation between** X **and** Y is

$$\rho = \frac{\mathbb{E}[(X-\mu_x)(Y-\mu_y)]}{\sqrt{\sigma_x^2\sigma_y^2}} \tag{4.17}$$

Properties of the Bivariate Normal Distribution

- The two **marginal** distributions are

$$X \sim \mathbb{N}(\mu_x, \sigma_x^2) \text{ and } Y \sim \mathcal{N}(\mu_y, \sigma_y^2) \tag{4.18}$$

- The **conditional** distribution of Y given $X = x$ is

$$Y \mid X = x \sim \mathcal{N}\left(\mu_y + \rho\sigma^y\left(\frac{x-\mu_x}{\sigma_x}\right), (1-\rho^2)\sigma_y^2\right) \tag{4.19}$$

- **Linear combinations** of X and Y are normally distributed.
- A zero covariance between X and Y implies that they are **statistically independent**. *Note that this is not true in general for any two uncorrelated random variables.*

4.2.2 Pearson's Sample Correlation Coefficient

Definition 4.3. If X and Y are random variables, *a symmetric measure of the direction and strength of the linear dependence between them* is their **correlation**, ρ.

Based on a sample of n observed pairs (x_i, y_i), the estimate of the population correlation, ρ is **Pearson's Product-Moment Correlation Coefficient** r, formulated as

$$r = \frac{\sum_i^n(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i^n(x_i-\bar{x})^2\sum_i^n(y_i-\bar{y})^2}} = \frac{SXY}{\sqrt{SXX \cdot SYY}} \tag{4.20}$$

▌ It is the MLE of ρ, i.e. $\hat{\rho}_{MLE} = r$.

Facts about the Sample Correlation Coefficient

- It measures the **strength and direction of the linear relationship** between X and Y
- It's **unit-free**
- Is always a member $\in [-1, 1]$
  - $r = 0$: no linear association
  - $r = -1$: perfect negative linear relationship
  - $r = 1$: perfect positive linear relationship
  - The strength of the linear relationship increases as r moves away from 0

Figure of scatterplots in Slides

4.2.3 Relationship between Regression and Correlation

Question 4.1. How is correlation linked to SLR?

> Symmetric relationship $Cov(X, Y) = Cov(Y, X)$
>
> In SLR, we
>
> - Have a response
> - Specify a predictor
> - Estimate $\mathbb{E}(Y \mid X)$/predict Y

Question 4.2. What if we test $H_0 : \rho = 0$?

> Involve r
>
> But how is this related to $H_0 : \beta_1 = 0$?

4.2.4 Coefficient of Determination / R-Squared

Definition 4.4 (Properties and Interpretation of R-Squared)**.**

$$R^2 = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{\sum_i^n (Y_i - \bar{Y})^2} = \frac{SSReg}{SST} = \frac{SST - RSS}{SST} = 1 - \frac{RSS}{SST} \tag{4.21}$$

Figure for Extreme Cases

- $R^2 \in [0, 1]$
- R^2 is NOT resistant to outliers, affected by the spacing of X.
- The higher the variation in X, the higher R^2 will be.

Use and Limitations of R^2

R^2 gives the **percentage** of variation in y's explained by regression line

A high R^2 does NOT indicate that the estimated regression line is a good fit.

- no absolute rules about how 'big' it should be
- can get very high by *overfitting*

R^2 is not meaningful for models without intercept.

To compare two models, R^2 is only useful if

1. same observations, y's in original units (not transformed)
2. one set of predictor variables is a subset of the other

Theorem 4.3. In SLR,

$$R^2 = r^2 \tag{4.22}$$

Proof. Recall from (4.20), (4.21) that

$$r = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}} = \frac{SXY}{\sqrt{SXX \cdot SYY}} = \frac{SXY}{\sqrt{SXX \cdot SYY}}$$

$$R^2 = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{\sum_i^n (Y_i - \bar{Y})^2} = \frac{SSReg}{SST} = 1 - \frac{RSS}{SST} = \frac{b_1^2 SXX}{SYY}$$

Hence

$$\begin{aligned} b_1 = \frac{SXY}{SXX}, b_1^2 = \frac{(SXY)^2}{(SXX)^2} \implies R^2 &= \frac{SXY^2}{SXX^2} \cdot \frac{SXX}{SYY} \\ &= \frac{SXY^2}{SXX \cdot SYY} \\ &= r^2 \end{aligned}$$

□

4.2.5 Testing $H_0 : \rho = 0$

When the population is **bivariate normal** and we test

$$H_0 : \rho_{XY} = 0,$$
$$H_A : \rho_{XY} \neq 0.$$

It can be shown that the test statistic

$$t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1}{\sqrt{\frac{S^2}{SXX}}} \sim T_{n-2} \tag{4.23}$$

where b_1 corresponds to the slope estimate for the normal error SLR model.

> Equivalent to $H_0 : \beta_1 = 0$
>
> Further, $b_1 = \frac{SXY}{SXX} \implies r = \frac{SXY}{\sqrt{SXX \cdot SYY}}$

4.3 Dummy Variable Regression

Consider SLR where the predictor variable X is categorical rather than quantitative.

- Simplest case is when X has only two levels
- Use a dummy or indicator variable for the two levels, i.e. takes on two values: 0 or 1 only
- The resulting regression model is used to test for the difference in two means

Regression model:

$$\mu_{Y/x} = \beta_0 + \beta_1 x \tag{4.24}$$

Fitted Regression Line:

$$\hat{y} = b_0 + b_1 x \tag{4.25}$$

CASE I	Category Level A, $x = 0$	Category Level B, $x = 1$
$\mu_{Y/x}$	$\beta_0 = \mu_A$	$\beta_0 + \beta_1 = \mu_B$
$\hat{y}$	$b_0 = \bar{y}_A$	$b_0 + b_1 = \bar{y}_B$
Diff	$\mu_A - \mu_B = \beta_0 - \beta_0 - \beta_1 = -\beta_1$,	$\bar{y}_A - \bar{y}_B = -b_1$
CASE II	Category Level A, $x = 1$	Category Level B, $x = 0$
$\mu_{Y/x}$	$\beta_0 + \beta_1 = \mu_A$	$\beta_0 = \mu_B$
$\hat{y}$	$b_0 + b_1 = \bar{y}_A$	$b_0 = \bar{y}_B$

β_1 = rep. the diff in means.

4.3.1 The Pooled Two-Sample T-Procedure

Aim: Compare two independent populations means

$$H_0 : \mu_A = \mu_B$$

Conditions:

- Sampel 1: $A_1, A_2, \ldots, A_n \overset{iid}{\sim} \mathcal{N}(\mu_A, \sigma_A^2)$
- Sample 2: $B_1, B_2, \ldots, B_n \overset{iid}{\sim} \mathcal{N}(\mu_A, \sigma_A^2)$
- The two samples are independent
- The two populations have the same variance, i.e. $\sigma_A^2 = \sigma_B^2 = \sigma^2$

Test statistic:

$$t_{obs} = \frac{(\bar{A} - \bar{B}) - (\mu_A - \mu_B)}{s_p^2 \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \overset{H_0}{\sim} T_{n_A + n_B - 2}$$

where $s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)}$.

- Population parameter: $\mu_a - \mu_B$
- Sample estimator: $\bar{A} - \bar{B}$
- Variance of sample estimator:

$$\begin{aligned} Var(\bar{A} - \bar{B}) &= Var(\bar{A}) + Var(\bar{B}) \qquad \text{since samples are independent} \\ &= \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} \\ &= \sigma^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \qquad \text{since } \sigma_A^2 = \sigma_B^2 = \sigma^2. \end{aligned}$$

- Pivotal Quantity:

$$\frac{(\bar{A}-\bar{B})-(\mu_A-\mu_B)}{s_p^2\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} = \frac{\frac{(\bar{A}-\bar{B})-(\mu_A-\mu_B)}{\sigma^2\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}}{\sqrt{\frac{\sum(n_i-1)\left[\frac{\sum(n_i-1)s_i^2}{\sum(n_i-1)}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\right]}{\sigma^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sum(n_i-1)}}}$$

  Note: $Z/\chi^2_{df/df}$.

Example 4.1 (Pooled Two-Sample T Examples).

1. The difference between the average recovery times of an old and new drug.
2. The difference in average condo price between downtown condos and uptown condos.
3. The difference in mean length between Trout and Bass adult
4. . . .

4.4 Summary A: Data Analysis Tips

1. Is inference for:
   - estimating a mean response,
   - predicting a new observation or
   - estimating the regression parameters
2. Do not extrapolate outside the data range of X
3. The conditional bivariate Normal correlation model is equivalent to the Normal error SLR model.
4. Correlation does not imply causation. Consider the context of the data, whether experimental or non-experimental.
5. Statistical significance does NOT imply practical significance and vice versa.

5 Module 5 - Diagnostics in SLR

5.1 Probability Plots

Probability Plots are extremely useful tools to graphically assess goodness-of-fit. To see if the observed data follow a particular distribution, we plot the (observed) order statistics (from sorted data) against the (expected) theoretical quantiles (from a prob dist).

If the data follows the particular distribution, the plot should look roughly linear.

The most common probability plot is the **Normal Q-Q plot**.

The Shapiro-Wilk Test

1. Null Hypotheses: The data follows a Normal Distribution
2. Test Statistic:
$$W = \frac{(\sum_i^n a_i x(i))^2}{\sum_i^n (x_i - \bar{x})^2}$$
   where
   - $a_i = \frac{m'V^{-1}}{\sqrt{m'V^{-1}V^{-1}m}}$
   - $m' = (m_1, \ldots, m_n)$ are the expected values of standard normal order statistics
   - V is the covariance matrix of those normal order statistics
3. Decision Rule: If the p-value is less than the significance level, there is evidence that the data is **not Normal**.

5.2 Properties of Residuals

Definition 5.1 (True Random Error)**.** Where $y_i = \mathbb{E}(y_i \mid X = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i$, e_i is the random fluctuation (or error) in y_i, i.e.

$$e_i = y_i - \mathbb{E}(y_i \mid X = x_i) = y_i - (\beta_0 + \beta_1 x_i) \tag{5.1}$$

Under normal error SLR assumptions, we have that

$$e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{5.2}$$

Definition 5.2 (Estimated Errors/Residuals)**.**

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i \tag{5.3}$$

Some Properties *From the LS method*

1. $\sum \hat{e}_i = 0$
2. $\sum \hat{e}_i x_i = 0$
3. $\sum \hat{e}_i \hat{y}_i = 0$

5.2.1 Mean and Variance of Residuals

Expectation:

$$\mathbb{E}(\hat{e}_i) = \mathbb{E}(y_i - \hat{y}_i) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i) = 0 \tag{5.4}$$

Variance:

$$Var(\hat{e}_i) = (1 - h_{ii})\sigma^2 \tag{5.5}$$

where h is the **leverage**, defined as

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}, \quad \text{fix } i, j = 1 \dots n \tag{5.6}$$

$$\implies i = j, \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} \tag{5.7}$$

$$\tag{5.8}$$

$$\hat{y}_i = b_0 + b_1 x_i = \bar{y} - h\bar{x} + b_1 x_i = \bar{y} + b_1(x_i - \bar{x}) \tag{5.9}$$

Recall $b_1 = \sum_j (x_j - \bar{x})/SXX, \bar{y} = \sum y_j/n$, hence

$$\hat{y}_i = \frac{\sum y_j}{n} + \frac{\sum (x_i - \bar{x})(x_j - \bar{x})}{SXX} y_j \tag{5.10}$$

$$= \sum_{j=1}^{n} h_{ij} y_j \tag{5.11}$$

$$= h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \tag{5.12}$$

5.2.2 Covariance and Normality of Residuals

Covariance:

$$Cov(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2, \quad i \neq j \tag{5.13}$$

Note: Residuals are NOT UNCORRELATED.

Normality:

6 Module 9 - Multiple Linear Regression Analysis

6.1 Review

Multiple Linear Regression (MLR)

- MLR Model (to obtain using the least-squares estimation):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{6.1}$$

  where

  $\boldsymbol{Y} \in M_{n\times 1}$,

  $\boldsymbol{X} \in M_{n\times(p+1)}$,

  $\boldsymbol{\beta} \in M_{p+1}$,

  $\boldsymbol{e} \in M_{n\times 1}$,

- p predictors: $p+1$ $\boldsymbol{\beta}$'s
- Gauss-Markov Conditions: $E(\boldsymbol{e}) = 0, Var(\boldsymbol{e}) = \sigma^2 I$
- Normal Error assumption (for inference)

6.2 R-Squared and Adjusted R-Squared

Definition 6.1 (R^2: Coefficient of Multiple Determination)**.**

$$R^2 = \frac{SSReg}{SST} = 1 - \frac{RSS}{SST} = \frac{\boldsymbol{Y'(H - \frac{1}{n}J)Y}}{\boldsymbol{Y'(I - \frac{1}{n}J)Y}} \tag{6.2}$$

called the coefficient of multiple determination (in the MLR setting).

R^2 gives the percentage of variation in Y explained by the model with all the p predictors.

> Note that it's NOT the square of a sample correlation coefficient (r^2) anymore.

Remark 6.1. For the same Y, as p increases,

SST remains the same,

SSReg stays the same or increases,

RSS stays the same or decreases,

hence R^2 either stays the same or increases.

Question 6.1. Is R^2 helpful in telling whether additional predictors are *useful* for explaining the response?

No.

Definition 6.2 (Adjusted R^2)**.**

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)} = 1 - (n-1)\frac{MSE}{SST} \tag{6.3}$$

where $S^2 = \frac{RSS}{n-p-1}$ is an unbiased estimate of $\sigma^2 = Var(e_i) = Var(Y_i)$.

- adjusted for the number of predictors in the model
- better to use instead of R^2
- always: $R^2_{adj} < R^2$ (note that the inequality is strict)

> Case. $p > 1$ (MLR)
>
> Then $(n-1)/(n-p-1) > 1$
>
> General case: as p increases (adding more predictors to the model), $(n-1)/(n-p-1)$ increases.

6.3 Global and Partial F-tests

6.3.1 Motivational Examples: Salary vs. Experience, Wine Quality

Definition 6.3 (Global F-test)**.** Testing Hypotheses: for $j = 1 \dots p$,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{6.4}$$

$$H_a : \text{at least one } \beta_j \text{ is not } 0 \tag{6.5}$$

To test whether there is a linear association between Y and all predictors.

Test statistic:

$$F_{obs} = \frac{MSReg}{MSE} = \frac{SSReg/p}{RSS/(n-p-1)} \tag{6.6}$$

Under H_0, F_{obs} is an observation from the F distribution with $df = (p, n-p-1)$.

Hence we can conclude that

Small p-value: Rhe model contains at least one significant predictor among the set of p predictors

Large p-value: None of the p predictors are relevant for estimating/predicting Y.

7 Module 10 - Diagnostics in MLR

7.1 Inference for a Single Regression Coefficient

7.1.1 Hypothesis Testing

As in SLR, we are interested in testing

$$H_0 : \beta_j = 0 \text{ vs. } H_a : \beta_j \neq 0. \tag{7.1}$$

Test statistic:

$$t_{obs} = \frac{b_j}{SE(b_j)} \tag{7.2}$$

Under H_0, t_{obs} is an observation from the T distribution with df $= n - p - 1$. This test gives an indication of whether or not the jth predictor, $X_j, j = 1 \dots p$, contributes to the prediction of the response variable **over and above** all the other predictors.

This is the special case of the Partial F-test with $k = 1$

7.1.2 Confidence Interval

Confidence interval for $\beta_j, j = 1 \dots p$, assuming all the other predictors are in the model, is

$$\beta_j \pm t_{\alpha/2,n-p-1} SE(b_j), \tag{7.3}$$

(i.e. unbiased estimate ± Margin of Error, where MOE is the critical value × std error).

where

- b_j: unbiased estimator of β_j
- $SE(b_j)$: standard error of the estimator
- $t_{\alpha/2,n-p-1}$: critical value of $100(1-\alpha/2)$th quantile from the T distribution with df $= n - p - 1$.

7.1.3 Global F-test vs. Individual t-tests

- In SLR, these tests are equivalent
- In MLR, the global F-test is designed to test the *overall model*, while the *t*-tests are designed to test *individual coefficients*.

Case A. If the Global F-test is significant and:

- A.1: All or some or the t-tests are significant, $\implies$ there are some useful explanatory variables for predicting Y.

- A.2: All the t-tests are not significant, $\Longrightarrow$ this is an indication of "multicollinearity", i.e. strongly correlated X's.

  This implies that individual X's do not contribute to the prediction of Y over and above other X's.

Case B. If the Global F-test is NOT significant and:

- B.1: All the t-tests are not significant, $\Longrightarrow$ none of the X's contributes to the prediction of Y.
- B.2: Some of the t-tests are significant, $\Longrightarrow$
  - The model has no predictive ability. Likely, if there are many predictors, there are type I errors in the t-tests.
  - The predictors are poorly chosen. The contribution of one useful predictor among many poor ones may not be enough for the model (Global F-test) to be significant.

7.2 Multicollinearity

Definition 7.1. Multicollinearity occurs when explanatory variables are highly correlated.

> In this case, it is difficult to measure the individual influence of one of the predictors on the response.
>
> - The fitted equation is unstable
> - The estimated regression coefficients vary widely from data set to data set (even if the data sets are similar) and depending on which predictor is included in the model.
> - The estimated regression coefficients may even have opposite sign than what is expected (*e.g. Simpson's Paradox*).

Remark 7.1. When some X's are perfectly correlated, we cannot estimate β because $X'X$ is sigular. Even if $X'X$ is close to singular, its determinant will be close to zero and the standard errors of estimated coefficients will be large.

Remark 7.2. For the general multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e \tag{7.4}$$

$$Var(\hat{\beta_j}) = \frac{1}{1-R_j^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2}, j = 1 \ldots p \tag{7.5}$$

where R_j^2 is the value of R^2 from the regression of x_j on the other x's.

> The jth Variance Inflation Factor (VIF) $= \frac{1}{1-R_j^2}$
>
> A commonly used cut-off is 5.

7.3 Diagnostics and Remedies

7.3.1 Leverage and Influential Points

Definition 7.2 (Identifying Leverage Points)**.** Classify the ith point as a point of high leverage (i.e. a lvg point) in MLR model with p predictors if

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{p+1}{n} \tag{7.6}$$

> In SLR, $p = 1, h_{ii} > 2\left(\frac{2}{n}\right) = \frac{4}{n}$.

Remark 7.3. When a **valid** model has been fit, a plot of standardized residuals against any predictor or any linear combination of the predictors (e.g. the fitted values) will have the following features:

1. A random scatter of points around the horizontal axis, since the mean function of e_i is zero when a *correct model has been fit* (linearity)
2. Constant variability as we look along the horizontal axis, i.e.

$$Var(e) = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} \tag{7.7}$$

Hence, **any pattern** in plot of standardized residuals is indicative that an **invalid** model has been fit to the data. Any nonrandom pattern itself does not provide direct information on how the model is misspecified.

7.3.2 The Box-Cox Transformation

Definition 7.3. The Box-Cox transformation is a general method for transforming a strictly positive (response or predictor) variable.

> It aims to find transformation that makes the transformed variable close to normally distributed.
>
> It considers a family of *power transformations.*
>
> Suppose the power to be λ:
>
> - $\lambda = 0$: Natural log
> - $\lambda = 1$: No transformation
> - $\lambda = 0.5$: Square root transformation
> - $\lambda = -1$: Inverse transformation
>
> It is based on maximizing a likelihood function.

7.4 Added Variable Plot (Need to Review Lecture Part 3)

Definition 7.4. Suppose our current model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \qquad \text{(model } YX\text{)} \tag{7.8}$$

and we are considering the introduction of an additional predictor variable $\boldsymbol{Z}$, that is, our new model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{e} \qquad \text{(model } YXZ\text{)} \tag{7.9}$$

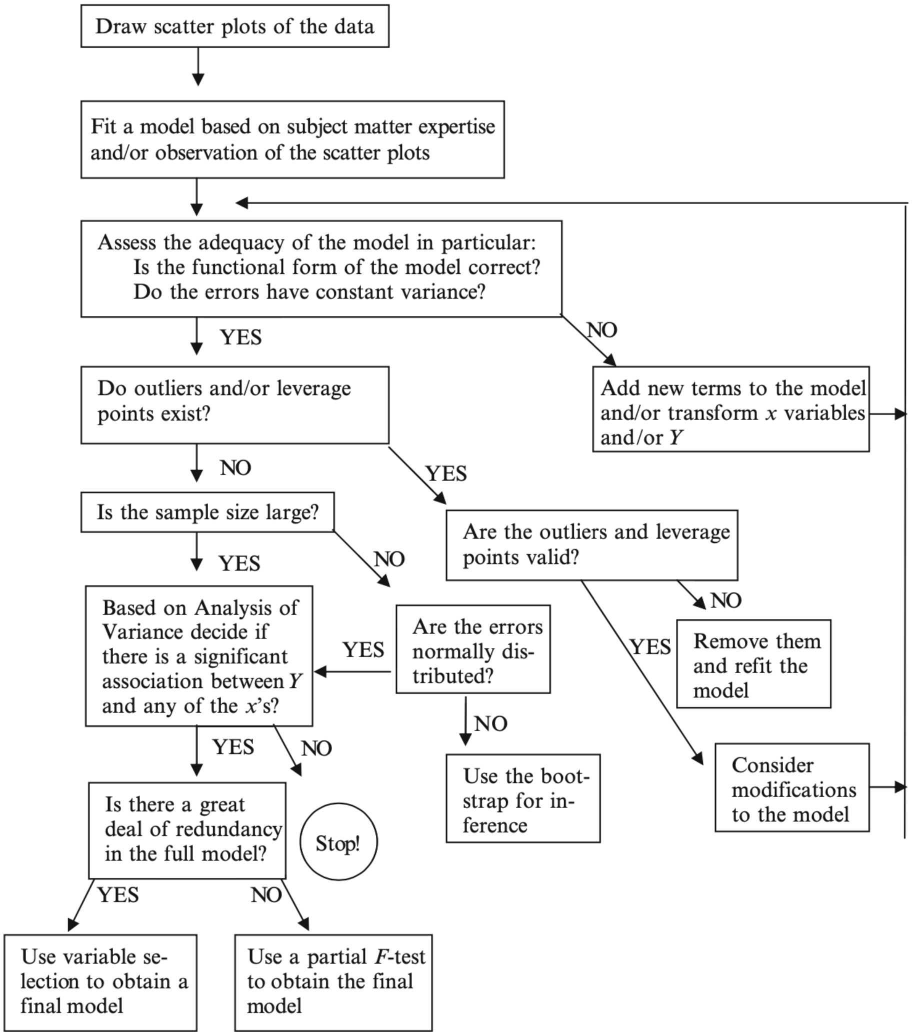
The added-variable plot is obtained by plotting the residuals from model YX against the residuals from the model

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{e} \qquad \text{(model } ZX\text{)} \tag{7.10}$$

Remark 7.4 (Why Added Variable Plot)**.**

- To visually assess the effect of each predictor, having adjusted for the effects of the other predictors
- To visually estimate α
- Can be used to identify points which have undue influence on the least squares estimate of α

7.5 Data Analysis Flow


Draw scatter plots of the data
Fit a model based on subject matter expertise and/or observation of the scatter plots
Assess the adequacy of the model in particular: Is the functional form of the model correct? Do the errors have constant variance?
NO
Add new terms to the model and/or transform x variables and/or Y
YES
Do outliers and/or leverage points exist?
YES
Are the outliers and leverage points valid?
NO
Remove them and refit the model
YES
Consider modifications to the model
NO
Is the sample size large?
NO
Are the errors normally distributed?
NO
Use the bootstrap for inference
YES
YES
Based on Analysis of Variance decide if there is a significant association between Y and any of the x's?
NO
Stop!
YES
Is there a great deal of redundancy in the full model?
YES
Use variable selection to obtain a final model
NO
Use a partial F-test to obtain the final model

8 Module 11 - Variable Selection

8.1 Variable Selection

What is the Goal? *Variable selection* or *Prediction accuracy* **What can occur?**

- Overfitting: Too many predictors are in the "final" regression model.
- Underfitting: Not enough predictors are in the "final" regression model.

The *bias-variance trade-off*: when we add more predictors to a valid model,

- the bias of the predictions gets smaller (↓)
- the variance of the estimated coefficients gets larger (↑)

8.1.1 Variable Selection: MLR and Likelihood-based Criteria

Suppose that $yi, x_{1i}, x_{2i}, \ldots, x_{pi}, i = 1 \ldots n$ are observed values of Normal random variables, and

$$y_i \mid x_{1i}, \ldots, x_{pi} \sim \mathcal{N}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \sigma^2), \tag{8.1}$$

thus the **conditional density** of y_i given $x_{1i}, \ldots, x_{pi}$ is given by

$$f(y_i \mid x_{1i}, x_{2i}, \ldots, x_{pi}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\})^2}{2\sigma^2}\right) \tag{8.2}$$

Definition 8.1. Assuming that the n observations are independent, then the **likelihood function** of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ given Y is given by

$$\begin{aligned} L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 \mid Y) &= \prod_{i=1}^{n} f(y_i \mid x_i) && (8.3) \\ &= \prod_{n=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\})^2}{2\sigma^2}\right) && (8.4) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\})^2\right) && (8.5) \end{aligned}$$

Definition 8.2. The **log-likelihood function** is given by

$$\log L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 \mid Y) \tag{8.6}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\})^2 \tag{8.7}$$

> Note that only the third term contains the regression parameters $\beta_0, \beta_1, \ldots, \beta_p$, hence the MLEs of $\beta_0, \ldots, \beta_p$ can be obtained by minimizing the third term only, which is equivalent to minimizing the residual sum of squares, RSS.
>
> Therefore, MLEs of $\beta_0, \ldots, \beta_p$ are equal to the least square estimates.

Substituting the LSEs of $\beta_0, \beta_1, \ldots, \beta_p$, the **log-likelibood function** is re-written as

$$\log L(\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_p}, \hat{\sigma}^2 \mid Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}RSS, \tag{8.8}$$

where $RSS = \sum_{i=1}^{n}(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\})^2$

Solving for the MLE of σ^2, we obtain

$$\sigma^2_{MLE} = \frac{RSS}{n} \tag{8.9}$$

> Note that it differs slightly from the unbiased estimate of σ^2, namely, $S^2 = RSS/(n-p-1)$.

Substituting the MLE of σ^2 into (8.8), we find that the likelihood associated with the maximum likelihood estimates is given by

$$\begin{aligned}\log L(\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_p}, \hat{\sigma}^2 \mid Y) &= \left(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{RSS}{n} - \frac{1}{2}\left(\frac{RSS}{n}\right)(RSS)\right) &(8.10)\\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{RSS}{n}\right) - \frac{n}{2} &(8.11)\end{aligned}$$

8.1.2 Akaike's Information Criterion (AIC)

Akaike's Information Criterion is a likelihood-based criterion for assessing models, founded by Akaike (1973). It balances *goodness* of fit and a penalty for model complexity.

Definition 8.3 (AIC)**.**

$$\begin{aligned}AIC &= 2\left[-\log L(\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_p}, \hat{\sigma}^2 \mid Y) + K\right] &(8.12)\\ &= 2\left(\frac{n}{2}\log(2\pi) + \frac{n}{2}\log\left(\frac{RSS}{n}\right) + \frac{n}{2} + K\right) &(8.13)\\ &= n\log(2\pi) + n\log\left(\frac{RSS}{n}\right) + \frac{n}{2} + p + 4 &(8.14)\end{aligned}$$

where $K = p + 2$ is the number of parameters in the fitted model.

> In R,
>
> $$AIC = n\log\left(\frac{RSS}{n}\right) + 2p. \tag{8.15}$$
>
> Note that we dropped some terms by *constant shift*.

Remark 8.1. The smaller the value of AIC, the better the model.

When the sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size, AIC tends to *overfit* since the penalty for model complexity is NOT strong enough.

> For this reason, a *bias corrected* verstion of AIC (Hurvich and Tsai, 1989) was developed.

8.1.3 Bayesian Information Criterion (BIC)

Definition 8.4 (BIC)**.**

$$\begin{aligned} BIC &= -2\log L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2 \mid Y) + K\log(n) \qquad (8.16)\\ &= -2\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{RSS}{n}\right) - \frac{n}{2} + K\log n\right) \qquad (8.17)\\ &= n\log(2\pi) + n\log\left(\frac{RSS}{n}\right) + n + p\log n + 2\log n \qquad (8.18)\end{aligned}$$

where $K = p + 2$ is the number of parameters in the fitted model.

> In R:
> $$BIC = n\log\left(\frac{RSS}{n}\right) + p\log(n). \tag{8.19}$$

Remark 8.2. The smaller the value of BIC, the better the model.

When $n \geq 8, \log(n) > 2$, the penalty term in BIC is greater than that in AIC.

Remark 8.3. BIC penalizes complex model more heavily than AIC, thus favours simpler models than AIC.

Remark 8.4. Getting AIC, BIC in R:

```
AIC(fittedmodel)
BIC(fittedmodel)
extractAIC(fittedmodel)
```

8.1.4 A Model Selection Strategy

1. Calculate R^2_{adj}, AIC, corrected AIC, and BIC
2. Compare the models which **minimize** AIC, corrected AIC, and BIC with the model that **maximizes** R^2_{adj}.

Some notes...

- The model with the lowest $MSE = S^2 = RSS/(n - p - 1)$ is equivalent to the model with the highest R^2_{adj} among the same subset of predictors.
- Using R^2_{adj} tends to over-fitting
- Consider $n : p$ ratio

8.2 Stepwise Regression

8.2.1 Stepwise Methods

Stepwise methods are automatic variable selection methods. If there are p terms that can be added to the mean fucntion apart from the intercept, then there are w^p possible regression equations

There are 3 stepwise algorithms: Backward, Forward, Stepwise.

Algorithm 8.1 (Backward Elimination)**. Backward Elimination** starts with all the potential terms in the model, then removes the term with the largest p-value each time to give a smaller information criterion.

Algorithm 8.2 (Forward Selection)**.** The **Forward Selection** method is the *reverse* of the backward method. It begins with no item in the model, then adds one term at a time (with the smallest p-value) until no further terms can be added to produce a smaller information criterion.

Algorithm 8.3 (Stepwise Regression)**. Stepwise Regression** alternates forward steps with backward steps. At each stage, terms can be added, dropeed, or swapped.

> Backward elimination and forward selection consider AT MOST $p+(p-1)+\cdots+1 = \frac{p(p+1)}{2}$ of the 2^p possible predictor subsets, while stepwise regression can consider more subsets.
>
> The idea is to end up with a model where NO variables are redundant, given the other variables in the model. Often, backward elimination and forward selection will produce the same *"final"* model.
>
> Backward elimination is often preferred to forward selection, since its initial estimate of σ^2 will usually be smaller.

8.2.2 Stepwise Regression: Cautions

Selection **overstates** significance.

Estimates of regression coefficients are *biased*

p-values from F- and t- tests are generally smaller than their true values.

Remark 8.5 (Forward and Backward in R)**.** The following illustrates R code for forward selection and backward elimination:

```
forwardAC <- step(interceptmodel, scope=list(lower=~1, upper=~X1+X2+X3))
backwardAC <- step(fullmodel, direction="backward", k=log(n))
```

8.3 Penalized Linear Regression

Penalized linear regression performs variable selection and regression coefficient estimation *simutaneously.* It is a constrained least squares optimization problem

$$\min_{\beta_0} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{x}_i)^2 + \sum_{j=1}^{p} p_\lambda(\cdot), \tag{8.20}$$

where p is the penalty function, and $\lambda \geq 0$ is the penalty parameter.

When $\lambda = 0$, the solution is the least square estimates.

Example 8.1. Some examples:

- Ridge Regression: $p_\lambda = \lambda \beta_j^2$
- LASSO Penalty: $p_\lambda = \lambda |\beta_j|$
- Adaptive LASSO
- SEAD
- Bridge Regression
- ...

8.4 Cross Validation

k-fold Cross-validation is a standard approach to assess the predictive ability of models (y^*) by evaluating their performance on a new data set.

Algorithm 8.4. Cross Validation steps

1. Divide the data (randomly) into k (roughly) equal sets.
2. Stage A: Establish the model by using all but one of the k folds; this set is called the **training data** set.
3. Stage B: Use the remaining data set (the fold that was left out), called the **test data** to evaluate the model.
4. Repeat Stages A & B for k times by changing the k^{th} fold.

8.4.1 Cross Validation Prediction Error

9 Appendix

9.1 Rules of Expectation

- $\mathbb{E}(a) = a, a \in \mathbb{R}$
- $\mathbb{E}(aY) = a\mathbb{E}(Y)$
- $\mathbb{E}(X \pm Y) = \mathbb{E}(X) \pm \mathbb{E}(Y)$
- $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$
- Tower Rule: $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid X)]$