Model Assumptions

- \( Y = X\beta + \epsilon \)
- \( p < n \) and \( X \) has full rank.
- \( \epsilon \perp X \)
- \( \epsilon_1, \ldots, \epsilon_n \) are independent
- \( E(\epsilon) = 0 \)
- \( \text{Var}(\epsilon_i) = \sigma^2 \) for all \( i \)
- \( \epsilon_1, \ldots, \epsilon_n \) are normally distributed
Outline

1. Error Term Assumptions
2. Transformations of $Y$
3. Functional Form
4. Model Building
Normality of Errors Diagnostics

Quantile-Quantile Plot of residuals.

**Normal Q-Q Plot**

- Normally Distributed Residuals

**Normal Q-Q Plot**

- Uniformly Distributed Residuals

**R Command:** `qqnorm(e)`
Normality of Errors Diagnostics

Shapiro-Wilks Test on Residuals

- Null hypothesis is that the $\epsilon_i$’s are normally distributed.
- **Command:** `shapiro.test(e)`, where `e` is the vector of residuals.

![Console output]

- Reject $H_0$ if $p$-value is less than $\alpha$. 
Normality of Errors Remedial Measures

- Transform $Y$
Plot $|e|$ vs. $\hat{Y}$ or $X_j$. 

Constant Error Variance

Nonconstant Error Variance
Brown-Forsythe Test

- The null hypothesis is $\text{Var}(\epsilon_1) = \cdots = \text{Var}(\epsilon_n) = \sigma^2$.
- Divide all observations into two groups based on whether $\hat{Y}$ (or $X_j$) is above or below a certain value.
- Define $e_{i1} =$ $i$th residual in group 1 and $e_{i2} =$ $i$th residual in group 2.
- Let $n_1$ and $n_2$ be the groups sizes, $n = n_1 + n_2$, and $\tilde{e}_1$ and $\tilde{e}_2$ be the medians of the residuals in each group.
- Define $d_{i1} = |e_{i1} - \tilde{e}_1|$ and $d_{i2} = |e_{i2} - \tilde{e}_2|$ for each $i$.
- Perform a two-sample $t$-test using the $d_{i1}$’s and $d_{i2}$’s.
Brown-Forsythe Test (cont)

\[ t = \frac{\bar{d}_1 - \bar{d}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]

\[ s_p^2 = \frac{\sum_{i=1}^{n_1} (d_{i1} - \bar{d}_1)^2 + \sum_{i=1}^{n_2} (d_{i2} - \bar{d}_2)^2}{n - 2} \]

Reject \( H_0 \) if \( |t| > t_{\alpha/2}(n - 2) \).
Transform $Y$

Use GLS
Independence of Errors Diagnostics

- Were the data collected in time order?
- Durbin-Watson Test
- Sequence plot: plot $\epsilon_1, \ldots, \epsilon_n$ vs. $1, \ldots, n$. 

![No Auto Correlation](image1)

![Autocorrelated Residuals](image2)
If data were collected in time order, and the Durbin-Watson test/sequence plot show evidence of autocorrelation, use time series analysis.

If there is a structural reason to believe the $\epsilon_i$'s are dependent, use GLS.
Outline

1. Error Term Assumptions
2. Transformations of Y
3. Functional Form
4. Model Building
MLE for the OLS Model

Theorem

Consider an OLS model \( Y = X\beta + \epsilon \).

Ch4 assumptions hold and \( \epsilon \) is normally distributed.

Then the maximum likelihood estimators for \( \beta \) and \( \sigma^2 \) are

\[
\hat{\beta} = (X'X)^{-1}X'Y \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\|e\|^2.
\]

If \( L \) is the likelihood function, then

\[
-2 \ln(L(\hat{\beta}, \tilde{\sigma}^2)) = n\ln(2\pi) + n\ln(\|e\|^2) - n\ln(n) + n
\]

For linear models with normal disturbance terms, maximizing likelihood is equivalent to minimizing residual sum of squares, \( \|e\|^2 \).
Transformations of $Y$

- Problem: error terms not normal or have nonconstant variance.
- Possible solution: transform $Y$
- Assuming values of $Y$ are **nonnegative**, possible transformations include

\[
\tilde{Y}_i = \sqrt{Y_i}
\]

\[
\tilde{Y}_i = \ln Y_i
\]

\[
\tilde{Y}_i = \frac{1}{Y_i}
\]

- We would then fit the model

\[
\tilde{Y} = X\beta + \epsilon
\]
Box-Cox Transformations

- Assume $Y$ values are nonnegative. If not, add a constant to all $Y$ values.
- Given a power parameter $\lambda \in \mathbb{R}$, the Box-Cox transformation is
  \[
  \tilde{Y} = \begin{cases} 
  Y^\lambda, & \text{if } \lambda \neq 0 \\
  \ln(Y), & \text{if } \lambda = 0 
  \end{cases}
  \]
- The model becomes
  \[
  \tilde{Y} = X\beta + \epsilon
  \]
- $\lambda$ is estimated with maximum likelihood (least squares).
Consider a range of values for $\lambda$, such as $-2, -1.9, -1.8, \ldots, 1.8, 1.9, 2.0$.

For each value of $\lambda$ in this range, perform the following steps.

- Standardize $Y$ as follows:

$$W_i = \begin{cases} K_1(Y_i^\lambda - 1), & \text{if } \lambda \neq 0 \\ K_2(\ln(Y_i)), & \text{if } \lambda = 0, \end{cases}$$

where

$$K_2 = \left( \prod_{i=1}^{n} Y_i \right)^{\frac{1}{n}},$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}.$$

- Fit the model $W = X\beta + \epsilon$ and compute $\|e\|^2$.

The value of $\lambda$ leading to the smallest value of $\|e\|^2$ is the MLE.
Outline

1. Error Term Assumptions
2. Transformations of \( Y \)
3. Functional Form
4. Model Building
Overall Measures of Fit

- \( R^2 = 1 - \frac{\|e\|^2}{\|Y - \bar{Y}\|^2} \)

- Adjusted \( R^2 \)

\[ R^2_a = 1 - \frac{n - 1}{n - p} \frac{\|e\|^2}{\|Y - \bar{Y}\|^2} = 1 - \frac{\hat{\sigma}^2}{\text{Var}(Y)} \]

- Aikake Information Criterion

\[ \text{AIC} = 2p - 2 \ln(L) = 2p + n \ln(2\pi) + n \ln(\|e\|^2) - n \ln(n) + n \]

- How R calculates AIC for linear models

\[ 2p + n \ln(\|e\|^2) - n \ln(n) \]
Measures of Fit Based on Cross-validation

Leave One Out Cross-validation (LOOCV)

- For each $i = 1, \ldots, n$, fit a model based on the other observations $1, \ldots, i - 1, i + 1, \ldots, n$.
- Use this model to predict $Y_i$, and call this prediction $\hat{Y}_i$.
- Find the prediction sum of square errors (PRESS)

$$\text{PRESS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$
Delete-\(d\) Cross-validation

- Choose an integer \(d\) between 1 and \(n\). A value that has been suggested by Shao (1997) is

\[
d = n(1 - (\ln(n) - 1)^{-1}).
\]

- Repeat the following process a large number (say 1000) times:
  - Randomly select \(d\) rows of the data and remove them.
  - Fit a model to the remaining \(n - d\) rows.
  - Use this model to predict the values of \(Y_i\) for the removed rows.
  - Find the prediction sum of square errors

\[
PRESS = \sum_{\text{Removed Rows}} (Y_i - \hat{Y}_i)^2.
\]

- Finally, average all of these PRESS values to find a single overall PRESS value.
Diagnostics and Remedial Measures for Curvature

Diagnostics

- Plot $Y$ vs. $X_j$
- Plot $e$ vs. $\hat{Y}$ or $X_j$
- Compare original model to a model with higher order terms using an $F$-test or using overall measures of fit.

Remedial Measures

- Transform $X_j$ or add higher order terms.
Example Involving Curvature

- Scatterplot of $Y$ vs. $X$

- Do we need higher order terms?
Example Involving Curvature

- Scatterplot of $e$ vs. $X$

- Trend in residual plot indicates functional form is wrong.
Example Involving Curvature

- Fitting quadratic model

\[ Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \epsilon_i \]
Example Involving Curvature

- Scatterplot of $e$ vs. $X$ for quadratic model

- Lack of trend in residual plot indicates functional form is right.
Two Models

- **True Model:**
  \[ Y_i = 50 + 5X_i + 3X_i^2 + \epsilon_i \]

- **Model 1:**
  \[ Y_i = \beta_1 + \beta_2 X_i + \epsilon_i \]

- **Model 2:**
  \[ Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \epsilon_i \]
Two Models

- **True Model:**

\[ Y_i = 50 + 5X_i + 3X_i^2 + \epsilon_i \]

---

**Model 1**

```
Call: lm(formula = y ~ x)

Residuals:
    Min     1Q   Median     3Q    Max
-59.411 -21.309   0.998  20.750  67.615

Coefficients:
            Estimate  Std. Error    t value  Pr(>|t|)
(Intercept)  0.79280   5.2482   0.151     0.88
x           34.8783   0.9022 38.657 <2e-16 ***
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1
```

---

**Model 2**

```
Call: lm(formula = y ~ x + I(x^2))

Residuals:
    Min     1Q   Median     3Q    Max
-37.0814 -11.7763   0.8271  9.0966  36.9520

Coefficients:
            Estimate  Std. Error    t value  Pr(>|t|)
(Intercept) 49.2248   4.6405 10.608 < 2e-16 ***
x           6.3889   2.1208  3.012  0.00331 **
I(x^2)      2.8207   0.2034 13.865 < 2e-16 ***
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1
```
Comparing the Models with an $F$-test

- **Model 1:**
  \[ Y_i = \beta_1 + \beta_2 X_i + \epsilon_i \]

- **Model 2:**
  \[ Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \epsilon_i \]

\[ \text{ftest(model1, model2)} = 9.59 \times 10^{-25} \]

- So, we reject Model 1 in favor of Model 2.
Comparing the Models with Measures of Overall Fit

- $R^2$ (higher is better)
  - Model 1: 0.9385
  - Model 2: 0.9794

- Adjusted $R^2$ (higher is better)
  - Model 1: 0.9378
  - Model 2: 0.9789

- AIC (lower is better)
  - Model 1: 653.94
  - Model 2: 546.67

- Leave One Out PRESS (lower is better)
  - Model 1: 69572.47
  - Model 2: 23633.27

- SSE ($\|e\|^2$)
  - Model 1: 66474.03
  - Model 2: 22293.14
Interaction Terms

- Consider the regression model

\[ Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \]

- An interaction term is a term of the form

\[ \beta_{i_1 i_2} X_{i_1 j_1} X_{i_2 j_2} \]

Example

- Consider the regression model

\[ \text{BloodPressure}_i = \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Cholesterol}_i + \epsilon_i \]

- Here is the same model with an added interaction term for Gender and Cholesterol

\[ \text{BloodPressure}_i = \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Cholesterol}_i + \beta_{12} \text{Gender}_i \text{Cholesterol}_i + \epsilon_i \]
Diagnostics and Remedial Measures for Interactions

Diagnostics

▶ Plot $e$ vs. interaction term.
▶ Compare original model to a model with the interaction term using an $F$-test or using overall measures of fit.

Remedial Measures

▶ Include the interaction term if necessary.
1. Error Term Assumptions
2. Transformations of $Y$
3. Functional Form
4. Model Building
General Guidelines

- Data should be screened for errors.
- Rule of thumb: Sample size should be about 6 to 10 times as large as the number of variables in the pool of potential variables.
- Variables may need to be eliminated if they
  - are not clinically relevant
  - have large measurement errors
  - duplicate other variables
- Clinical considerations should be taken into account. Subject matter experts should be consulted.
Data cleaning/checking

Split Data into a training sample and a validation sample (this step is not necessary if it is possible to generate new data).
  - Univariate Analyses
    - Quantitative variables can be checked for curvature.
    - Appropriate categories can be considered for categorical variables.
  - Variable Selection
  - Diagnostics and Remedial Measures

Model Validation: Can be done by comparing model to
  - New data
  - Data from the validation sample
Variable Selection Methods

- Manually
- Stepwise Method
  ```r
d = data.frame(Y=Y, X=X)
bigmodel = lm(Y ~ ., data = d)
stepmodel = step(bigmodel)
```
- Best Subsets Method
  ```r
library(bestglm)
Xy = as.data.frame(cbind(x1, x2, x3, x4, x5, x6, Y))
bestglm(Xy, family = gaussian, IC = "AIC")
bestglm(Xy, family = gaussian, IC = "CV", t = 10)
```
- Combination: Use stepwise to narrow the list of variables and then apply best subsets to the remaining variables.
Models are validated by assessing their performance on a new data set.

The new data set can actually be newly collected data or can be the validation sample that was set aside at the beginning of model building.

Diagnostics should be used to determine if the fitted model is consistent with the new data.

The Mean Squared Prediction Error (MSPR) should be determined

- Let $Y_i, i = 1, \ldots, n^*$ be the new data set.
- For each $i$, use the model fitted to the training data to predict $Y_i$.
- Call the predicted value $\hat{Y}_i$.
- The MSPR is

$$MSPR = \frac{\sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2}{n^*}.$$
Example with Stepwise/Best Subsets Methods

\[
\begin{align*}
X &= \text{matrix}(\text{runif}(5000), 100, 50) \\
X0 &= \text{cbind}(\text{rep}(1, 100), X[, 1:3]) \\
\beta &= c(50, 5, 10, 30) \\
\epsilon &= \text{rnorm}(100, 0, 1) \\
Y &= X0 \times \beta + \epsilon
\end{align*}
\]

**True Model:**

\[
Y_i = 50 + 5X_{i1} + 10X_{i2} + 30X_{i3} + \epsilon_i, \text{ for } i = 1, \ldots, 100.
\]

Variables \(X_{i4}, \ldots, X_{i50}\) are just noise.
True Model:

\[ Y_i = 50 + 5X_{i1} + 10X_{i2} + 30X_{i3} + \epsilon_i, \text{ for } i = 1, \ldots, 100. \]

\[ x1=X0[,2] \]
\[ x2=X0[,3] \]
\[ x3=X0[,4] \]

model=lm(Y~x1+x2+x3)
summary(model)
d=data.frame(Y=Y, X=X)
bigmodel=lm(Y~., data=d)

stepmodel=step(bigmodel)
summary(stepmodel)

Coefficients:  

|   | Estimate | Std. Error | t value | Pr(>|t|) |
|---|----------|------------|---------|----------|
| (Intercept) | 47.6880 | 0.8455 | 56.404 | < 2e-16 *** |
| X.1 | 3.8129 | 0.3426 | 11.129 | < 2e-16 *** |
| X.2 | 10.1911 | 0.3503 | 29.095 | < 2e-16 *** |
| X.3 | 30.6207 | 0.3688 | 83.021 | < 2e-16 *** |
| X.4 | 0.6246 | 0.3961 | 1.577 | 0.11891 |
| X.6 | 0.7788 | 0.3333 | 2.337 | 0.02203 * |
| X.14 | 0.6239 | 0.3617 | 1.725 | 0.08848 . |
| X.15 | 0.9363 | 0.3304 | 2.834 | 0.00584 ** |
| X.16 | -0.5682 | 0.3289 | -1.727 | 0.08807 . |
| X.21 | -0.6973 | 0.3601 | -1.936 | 0.05644 . |
| X.23 | 0.6019 | 0.3580 | 1.681 | 0.09671 |
| X.27 | 0.4442 | 0.3254 | 1.365 | 0.17614 |
| X.28 | -0.6381 | 0.3559 | -1.793 | 0.07684 . |
| X.30 | 0.9224 | 0.3210 | 2.874 | 0.00522 ** |
| X.35 | -0.8150 | 0.3805 | -2.142 | 0.03532 * |
| X.37 | 0.9491 | 0.3569 | 2.659 | 0.00949 ** |
| X.41 | 0.9769 | 0.3812 | 2.563 | 0.01231 * |
| X.42 | 0.9704 | 0.3344 | 2.902 | 0.00481 ** |
| X.43 | -1.0106 | 0.3639 | -2.777 | 0.00686 ** |
| X.47 | -0.4945 | 0.3286 | -1.505 | 0.13642 |
| X.48 | 0.5735 | 0.3797 | 1.511 | 0.13496 |
| X.50 | 0.6709 | 0.3698 | 1.814 | 0.07352 . |
Xy = as.data.frame(cbind(X[, c(1, 2, 3, 4, 6, 14, 15, 16, 21, 23, 27, 28, 30, 35, 37, 41, 42, 43, 47, 48, 50)], Y))

bestmodel = bestglm(Xy, IC="CV", family=gaussian, t=10)
summary(bestmodel$BestModel)

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.7149   0.3706   134.153 <2e-16  ***
   V1         4.0777   0.3719    10.963 <2e-16  ***
   V2         9.8833   0.3643    27.128 <2e-16  ***
   V3        30.4825   0.3689    82.634 <2e-16  ***
   V15       0.9209   0.3508     2.625  0.0101   *

(Diagnostic and Remedial Measures)
$$\text{truemodel} = \text{lm}(Y \sim x1 + x2 + x3)$$
$$\text{trueYhat} = \text{predict}(\text{truemodel})$$

$$\text{bestYhat} = \text{predict}(\text{bestmodel}\$\text{BestModel})$$

[Graphs of Y vs. trueYhat and Y vs. bestYhat]