# Stat 206: Linear Models

## Lecture 17

Nov. 30, 2015

# Remedial Measures in ANOVA

How to make up for unequal variance and/or nonnormality?

- Transformation of the response variable to make its distribution closer to normal and to stabilize the variance.
  - Variance stabilizing transformation.
  - Box-cox procedure.
- If the departures are too extreme such that transformations do not work, then use *nonparametric tests*.
  - Rank F test for equality of means.

# Variance Stabilizing Transformations

When factor level variance is a function of the factor level mean, i.e., $\sigma_i^2 = \phi(\mu_i)$ for $i = 1, \cdots, I$, we can find a transformation $f(\cdot)$ such that:

- Factor level variances of the transformed data $Y_{ij}^* = f(Y_{ij})$ will be approximately equal.
- Often the distribution of the transformed data will also be closer to Normal.

Suppose $\mathrm{E}(Y) = \mu,\ \mathrm{Var}(Y) = \sigma^2 = \phi(\mu)$.

- We want to find a transformation $Y^* = f(Y)$ such that variance of $Y^*$ is a constant, say 1.
- By first order Taylor expansion:

$$Y^* \approx f(\mu) + f'(\mu)(Y - \mu).$$

- Therefore $\mathrm{E}(Y^*) \approx f(\mu)$ and $\mathrm{Var}(Y^*) \approx (f'(\mu))^2\phi(\mu)$
- Choose $f$ such that $(f'(\mu))^2\phi(\mu) = 1$.
- Thus

$$f(\mu) = \int \frac{1}{\sqrt{\phi(\mu)}} d\mu.$$

# Commonly Used Transformations

- If $\sigma_i^2 \propto \mu_i$, then use
- If $\sigma_i \propto \mu_i$, then use
- If $\sigma_i \propto \mu_i^2$, then use
- How to decide on which one to use?
  - Calculate $\frac{s_i^2}{\overline{Y}_{i\cdot}}$, $\frac{s_i}{\overline{Y}_{i\cdot}}$, $\frac{s_i}{(\overline{Y}_{i\cdot})^2}$ for $i = 1, \ldots, I$.
  - The approximate constancy of one of the three statistics across treatment groups suggests the corresponding transformation.

# Commonly Used Transformations

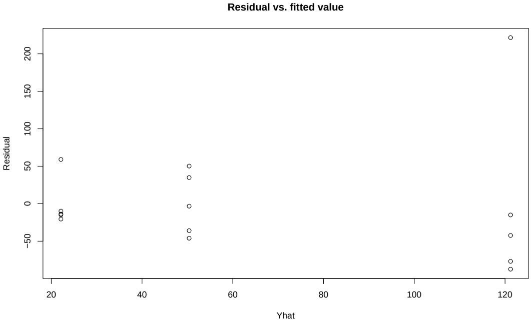
- If $\sigma_i^2 \propto \mu_i$, then use the square root transformation: $Y^* = \sqrt{Y}$.
- If $\sigma_i \propto \mu_i$, then use the log transformation: $Y^* = \log(Y)$.
- If $\sigma_i \propto \mu_i^2$, then use the inverse transformation: $Y^* = 1/Y$.
- How to decide on which one to use?
  - Calculate $\frac{s_i^2}{\overline{Y}_{i\cdot}}$, $\frac{s_i}{\overline{Y}_{i\cdot}}$, $\frac{s_i}{(\overline{Y}_{i\cdot})^2}$ for $i = 1, \ldots, I$.
  - The approximate constancy of one of the three statistics across treatment groups suggests the corresponding transformation.

# Computers

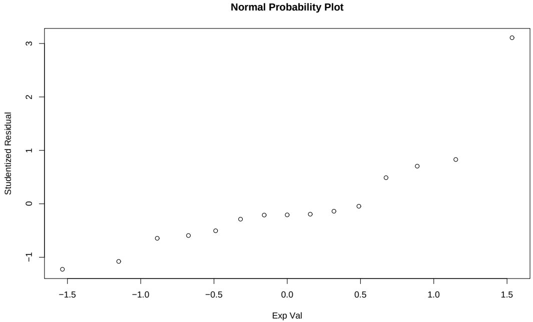
A company operates computers at three different locations. The computers are identical as to make and model, but are subject to different degrees of voltage fluctuation. The length of time between computer failures are recorded for three locations, each for five failure intervals. This is an observational study since no randomization of treatments to experimental units occurred.

| i | $Y_{i1}$ | $Y_{i2}$ | $Y_{i3}$ | $Y_{i4}$ | $Y_{i5}$ | $\overline{Y}_{i\cdot}$ |
|---|---|---|---|---|---|---|
| 1 | 4.41 | 100.65 | 14.45 | 47.13 | 85.21 | 50.37 |
| 2 | 8.24 | 81.16 | 7.35 | 12.29 | 1.61 | 22.13 |
| 3 | 106.19 | 33.83 | 78.88 | 342.81 | 44.33 | 121.2 |
| i | $s_i^2$ | $\frac{s_i^2}{\overline{Y}_{i\cdot}}$ | $\frac{s_i}{\overline{Y}_{i\cdot}}$ | $\frac{s_i}{(\overline{Y}_{i\cdot})^2}$ | | |
| 1 | 1788.7 | 35.5 | 0.84 | 0.017 | | |
| 2 | 1103.454 | 49.9 | 1.5 | 0.068 | | |
| 3 | 16167.4 | 133.4 | 1.05 | 0.009 | | |
| | $\overline{Y}_{\cdot\cdot}$ | = | 64.6; | MSE | = | 6353.2 |

Residual vs. fitted value plot indicates unequal variances.



Normal Q-Q plot indicates non-normality.



Since is nearly constant across *i*, we should use the .

Table: log-transformed data

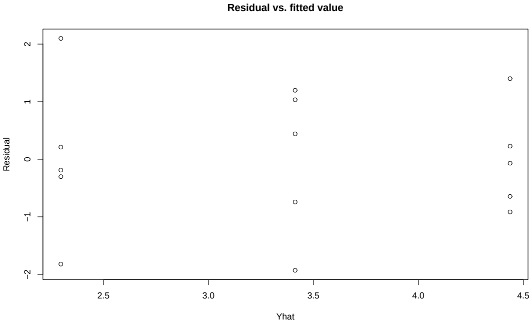
| i | $Y^*_{i1}$ | $Y^*_{i2}$ | $Y^*_{i3}$ | $Y^*_{i4}$ | $Y^*_{i5}$ | $\overline{Y^*}_{i\cdot}$ | $(s^*_i)^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.484 | 4.612 | 2.671 | 3.853 | 4.445 | 3.413 | 1.742 |
| 2 | 2.109 | 4.396 | 1.995 | 2.509 | 0.476 | 2.297 | 1.974 |
| 3 | 4.665 | 3.521 | 4.368 | 5.837 | 3.792 | 4.437 | 0.818 |
|  | $\overline{Y^*}_{\cdot\cdot}$ | = | 3.38 | ; | $MSE^*$ | = | 1.511 |

Since $\frac{s_i}{\overline{Y}_{i\cdot}}$ is nearly constant across $i$, we should use the log-transformation: $Y^* = \log(Y)$.

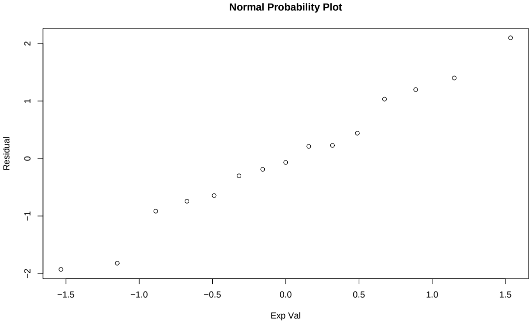
Table: log-transformed data

| i | $Y^*_{i1}$ | $Y^*_{i2}$ | $Y^*_{i3}$ | $Y^*_{i4}$ | $Y^*_{i5}$ | $\overline{Y^*}_{i\cdot}$ | $(s^*_i)^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.484 | 4.612 | 2.671 | 3.853 | 4.445 | 3.413 | 1.742 |
| 2 | 2.109 | 4.396 | 1.995 | 2.509 | 0.476 | 2.297 | 1.974 |
| 3 | 4.665 | 3.521 | 4.368 | 5.837 | 3.792 | 4.437 | 0.818 |
|  | $\overline{Y^*}_{\cdot\cdot}$ | = | 3.38 | ; | $MSE^*$ | = | 1.511 |

Residual vs. fitted value plot of the log-transformed data: Nearly equal variance.



Normal Q-Q plot of the log-transformed data: Nearly linear.



ANOVA on the log-transformed data $Y_{ij}^*$.

- $MSTR^* = 5.726, \ MSE^* = 1.511$.
- F test for equal means: $F_{log}^* = \frac{5.726}{1.511} = 3.789$.
- $I = 3, n_T = 15, pvalue = P(F_{2,12} > 3.789) = 0.053$.
- Can not reject $H_0 : \mu_1^* = \mu_2^* = \mu_3^*$ at 0.05 significance level, but can reject $H_0$ at 0.1 significance level. This is often referred to as “nearly significant”.

# Rank F Test and Rank Transformation

Test $H_0 : \mu_1 = \cdots = \mu_I$ without the normality assumption.

- Nonparametric procedures assume:
  - Model equation $Y_{ij} = \mu_i + \epsilon_{ij}$.
  - $\epsilon_{ij}$ have the **same continuous distribution** which is centered at zero.
  - Consequently the distributions of $Y_{ij}$ are the same up to a location translation (by $\mu_i$).
- **Rank transformation**: Get the rank $R_{ij}$ for each observation $Y_{ij}$.
  - For example, for the data set $3, 4, 2, 5, 6, 4$, the ranks are $2, 3.5, 1, 5, 6, 3.5$.

Rank F-test.

- Apply ANOVA on the ranks $R_{ij}$.
- Derive the F ratio: $F_R^* = MSTR(R)/MSE(R)$, where,

$$MSTR(R) = \frac{\sum_{i=1}^{I} n_i(\overline{R}_{i\cdot} - \overline{R}_{\cdot\cdot})^2}{I-1},$$

$$MSE(R) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(R_{ij} - \overline{R}_{i\cdot})^2}{n_T - I}.$$

- The null distribution of $F^*$ is approximately $F_{I-1,n_T-I}$ provided that $n_i$'s are not too small.
  - If $F_R^* > F(1-\alpha; I-1, n_T-I)$, then reject $H_0 : \mu_1 = \cdots = \mu_I$ at significance level $\alpha$.

# Computer

Table: Ranks of the data

| i | $R_{i1}$ | $R_{i2}$ | $R_{i3}$ | $R_{i4}$ | $R_{i5}$ | $\overline{R}_{i\cdot}$ | $s_i^2(R)$ |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 13 | 6 | 9 | 12 | 8.4 | 20.3 |
| 2 | 4 | 11 | 3 | 5 | 1 | 4.8 | 14.2 |
| 3 | 14 | 7 | 10 | 15 | 8 | 10.8 | 12.7 |
| | $\overline{R}_{\cdot\cdot}$ | = | 8 | ; | $MSE(R)$ | = | 15.7 |

- $MSTR(R) = 45.6,\ MSE(R) = 15.7.$
- $F_R^* = \frac{45.6}{15.7} = 2.90.$
- $pvalue = P(F_{2,12} > 2.90) = 0.094.$
- Reject $H_0$ at significance level 0.1, but can not reject $H_0$ at level 0.05.

- Under the log-transformation, the p-value is 0.053, which is more significant than the p-value under the rank transformation (0.094).
- In practice, for small data sets, simple transformations such as logarithm transformation are often preferred over the rank transformation since the ANOVA tests tend to have greater power under these simple transformations.

# Remedial Measures in Linear Regression

- Multicollinearity remedial measures.
  - Ridge regression.
  - `lm.ridge` function in the `MASS` library.
- Unequal error variance remedial measures.
  - Transformation of observation variable $Y$, e.g., Box-cox procedure.
  - Weighted least squares.
  - `weights` option in the `lm` function.
- Influential cases remedial measures.
  - Robust regression.
  - `rlm` function in the `MASS` library.

# Ridge Regression

Ridge regression introduces bias into the regression estimators while reducing their variances and as such achieves bias-variance trade-off. It is often used to remedy severe multicollinearity among the $X$ variables.

- Ridge regression is often applied to the standardized regression model:

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\epsilon}^*.$$

- The least-squares estimators:

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{*}} = \mathbf{r}_{XX}^{-1} \mathbf{r}_{XY},$$

where $\mathbf{r}_{XX} = \mathbf{X}^{*,T} \mathbf{X}^*, \quad \mathbf{r}_{XY} = \mathbf{X}^{*,T} \mathbf{Y}^*$ are correlation matrices among $X$ and between $X$ and $Y$, respectively.

- Ridge regression estimator is the minimizer of the $\ell_2$ *penalized least-squares criterion*:

$$Q_\lambda(\mathbf{b}^*) = (\mathbf{Y}^* - \mathbf{X}^*\mathbf{b}^*)^T(\mathbf{Y}^* - \mathbf{X}^*\mathbf{b}^*) + \lambda\mathbf{b}^{*,T}\mathbf{b}^*,$$

  where $\lambda \geq 0$ is a *tuning parameter*.

- The norm equation:

$$\frac{\partial Q_\lambda(\mathbf{b}^*)}{\partial \mathbf{b}^*} = 2\mathbf{X}^{*,T}\mathbf{X}^*\mathbf{b}^* + 2\lambda\mathbf{b}^* - 2\mathbf{X}^{*,T}\mathbf{Y}^* = \mathbf{0}.$$

- The Ridge regression estimator:

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{*}}{}_\lambda = (\mathbf{r}_{XX} + \lambda\mathbf{I}_{p-1})^{-1}\,\mathbf{r}_{XY}.$$

- When $\lambda = 0$, $\hat{\boldsymbol{\beta}}^{\boldsymbol{*}}{}_\lambda$ is the LS estimator. When $\lambda > 0$, $\hat{\boldsymbol{\beta}}^{\boldsymbol{*}}{}_\lambda$ is shrunk towards zero and the degree of shrinkage increases as $\lambda$ gets larger.

- The mean of $\hat{\boldsymbol{\beta}}^*{}_\lambda$ (assume the model is correct):

$$\mathbf{E}\{\hat{\boldsymbol{\beta}}^*{}_\lambda\} = \left(\mathbf{r}_{XX} + \lambda \mathbf{I}_{p-1}\right)^{-1} \mathbf{r}_{XX} \boldsymbol{\beta}^*.$$

  When $\lambda > 0$, $\hat{\boldsymbol{\beta}}^*{}_\lambda$ is a biased estimator and the bias increases as $\lambda$ gets larger.

- Variance-covariance matrix of $\hat{\boldsymbol{\beta}}^*{}_\lambda$:

$$\boldsymbol{\sigma}^2\{\hat{\boldsymbol{\beta}}^*{}_\lambda\} = \sigma^{*,2} \left(\mathbf{r}_{XX} + \lambda \mathbf{I}_{p-1}\right)^{-1} \mathbf{r}_{XX} \left(\mathbf{r}_{XX} + \lambda \mathbf{I}_{p-1}\right)^{-1}.$$

  The variance decreases as $\lambda$ gets larger.

- The $\lambda$ that achieves “optimal” bias-variance trade-off can be chosen using a model selection criteria such as AIC, or through cross-validation.

# Weighted Least Squares

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T.$$

- $\epsilon_i \sim N(0, \sigma^2/w_i)$ and $\epsilon_i$s are independent. So $\boldsymbol{\sigma}^2(\mathbf{Y}) = \sigma^2\mathbf{W}^{-1}$.
- $w_i > 0$ ($i = 1, \cdots, n$), $\sum_{i=1}^n w_i = 1$ are prespecified weights (either known or estimated).
- Maximizing the likelihood function is equivalent to minimizing the *weighted least squares criterion*:

$$\begin{aligned} Q_{\mathbf{w}}(\mathbf{b}) &= \sum_{i=1}^{n} w_i \left(Y_i - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1}\right)^2 \\ &= \left(\mathbf{Y} - \mathbf{X}\mathbf{b}\right)' \mathbf{W} \left(\mathbf{Y} - \mathbf{X}\mathbf{b}\right), \end{aligned}$$

  where $\mathbf{W}$ is the $n \times n$ diagonal matrix of the weights.
- The weighted least-squares estimator:

$$\hat{\beta}_{\mathbf{w}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y},$$

- $\hat{\boldsymbol{\beta}}_{\mathbf{w}}$ is an unbiased estimator (assuming the model is correct):

  $$E(\hat{\boldsymbol{\beta}}_{\mathbf{w}}) = \boldsymbol{\beta}.$$

  Note that even when the weight matrix **W** is misspecified, the estimator is still unbiased.

- The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathbf{w}}$:

  $$\boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}}_{\mathbf{w}}) = \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.$$

- An unbiased estimator of $\sigma^2$ is:

  $$MSE_{\mathbf{w}} = \frac{\sum_{i=1}^{n} w_i(Y_i - \widehat{Y}_i)^2}{n-p} = \frac{\sum_{i=1}^{n} w_i e_i^2}{n-p}.$$

# Robust Regression

Robust regression dampens the influence of outlying $Y$ observations, in hope to provide a better fit for the majority of the data and at the same time avoid entirely discarding the outlying cases.

- Robust regression estimators are defined as minimizers of a criterion of the following form:

$$Q_\rho(\mathbf{b}) = \sum_{i=1}^{n} \rho\left(Y_i - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1}\right),$$

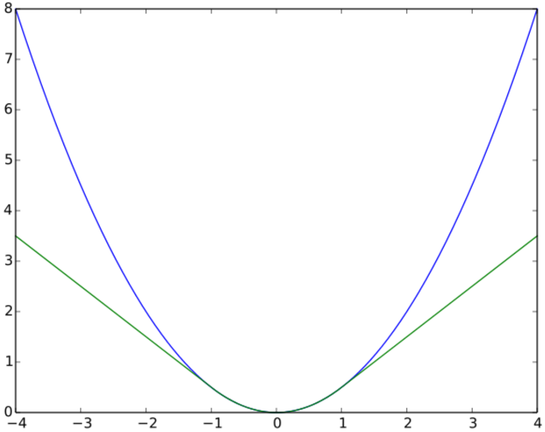
where $\rho(\cdot)$ is a function satisfies certain properties.

- $\rho(x) = x^2$: Ordinary least-squares regression.
- $\rho(x) = |x|$: *Least-absolute deviation (LAD)* regression, a.k.a. $L_1$ *regression*. LAD is implemented by the `rq` function in the `quantreg` library.
- *Huber's loss function*:

$$\rho(x) = \begin{cases} x^2/2 & \textit{if} \quad |x| \le c \\ c|x| - c^2/2 & \textit{if} \quad |x| > c. \end{cases}$$

  where $c$ is a tuning parameter (one choice is $c = 1.345$). Huber's method is implemented by the `rml` function in the `MASS` library.

Figure: Green: Huber's loss ($c = 1$); Blue: squared error loss



# More Topics

- Multi-factor ANOVA models and mixed effects models.
- Generalized linear models (GLM).
    - Logistic regression: The response $Y_i$ are binary.
    - Poisson regression: The response $Y_i$ are counts.
- Principal components regression (PCR) and partial least squares (PLS).
- $\ell_1$ penalized regression (Lasso regression).
- Nonlinear regression.
- Nonparametric regression.