Review of Linear Regression I
Statistics 211 - Statistical Methods II

Presented January 11, 2017

Accounting for Overdispersion

## Plot Pearson residuals vs. fitted values

nhat <- fitted( fit )
plot( nhat, presid^2, xlab="Fitted mean response",
 ylab="Squared Pearson residuals"
)
abline( h=1, col="red", lwd=2 )

sfit <- loess( presid^2 ~ nhat )
lines( sort(sfit$x), sfit$fitted[order(sfit$x)], col="blue", lwd=2 )

abline(h=phihat, lty=2, col="red", lwd=2)

Again, it looks as though the smoother is consistently above the y=1 line, indicating overdispersion.

Assumptions of the LR Model

Estimation of Model Parameters

The OLS estimator

Moments of the OLS estimator

Efficiency of the OLS Estimator

Optimality under classical assumptions

Efficiency under non-normal errors: Gauss-Markov Theorem

Distribution of the OLS estimator

Distribution of the OLS estimator under normality

Asymptotic distribution of the OLS estimator

Summary and Conclusions

Assumptions and properties of the OLS estimator

The role of heteroscedasticity
Review of Linear Regression

Our plan

1. Start with OLS and see where we get
2. What assumptions are not fulfilled with categorical response data (eg. binary or Poisson)
3. ‘Fix up’ OLS to satisfy assumptions and obtain valid inference
Review of Linear Regression

Linear Regression Model

Definition: By a classical (ordinary least squares) linear regression model, we mean a model in which we assume that

1. \( E[Y_i|\bar{X}_i] = \bar{X}_i^T \beta \)

2. \( \epsilon_i = Y_i - \bar{X}_i^T \beta \) and note that the model demands \( E[\epsilon_i] = 0 \)

3. the \( \epsilon_i \)'s are independent

4. \( \text{var}(\epsilon_i) = \sigma^2 \) for all \( i = 1, \ldots, n \)

5. the \( \epsilon_i \)'s are identically distributed

6. \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \)
Review of Linear Regression

Goal

- Conduct a model for the dependence of a response $Y$ on predictors $X_1, X_2, \ldots, X_{p-1}$
  - Two components to the model:
    1. The *systematic* component (mean model)
       \[ \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip-1} \]
    2. The *random* component (error term)
       \[ Y_i = \mu_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2) \]
  - Note: We can write the above model using matrix notation in which the $i$th row of design matrix $X$ is $\vec{X}_i^T$, response vector $\vec{Y}^T = (Y_1, \ldots, Y_n)$, and error vector $\vec{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$ obeys
     \[ \vec{Y} = X\vec{\beta} + \vec{\epsilon} \]
Estimation of Model Parameters

Least Squares

- We consider parameter estimates that minimize the sum of squared errors

\[ \sum_{i=1}^{n} (Y_i - \mu_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{X}_i^T \hat{\beta})^2 = (\hat{Y} - \hat{X} \hat{\beta})^T (\hat{Y} - \hat{X} \hat{\beta}) \]

where \( \hat{X}_i \) is the \( i^{th} \) row of the design matrix (the row vector of covariate values corresponding to the \( i^{th} \) observation) and \( \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})^T \)

- Why focus on the sum of squared errors?
  - Leads to score estimating equation under classical OLS assumptions
  - It is reasonable and mathematically convenient!
Proposition: Assume \( \text{rank}(X^T X) = p \) (ie. the number of observations \( n \) is greater than the number of parameters \( p \) and no predictors are constant or linear combinations of the other predictors). Then the least squares estimate is given by

\[
\hat{\beta} = (X^T X)^{-1} X^T \hat{Y}.
\]
Estimation of Model Parameters

Proof:
Mean and Variance of the OLS Estimator

Mean of the OLS Estimator

- **Proposition:** \( \hat{\beta} \) is unbiased for \( \beta \) (i.e. \( E[\hat{\beta}] = \beta \))

**Proof:**

Assumptions of the LR Model

- Estimation of Model Parameters
  - The OLS estimator
  - Moments of the OLS estimator

- Efficiency of the OLS Estimator
  - Optimality under classical assumptions
  - Efficiency under non-normal errors: Gauss-Markov Theorem

- Distribution of the OLS Estimator
  - Distribution of the OLS estimator under normality
  - Asymptotic distribution of the OLS estimator

Summary and Conclusions

- Assumptions and properties of the OLS estimator
  - The role of heteroscedasticity
### Mean and Variance of the OLS Estimator

#### Variance of the OLS Estimator

> **Proposition:** The variance of the ordinary least squares estimate is

\[
\text{var}(\hat{\beta}) = (X^T X)^{-1} X^T \Sigma X(X^T X)^{-1}
\]

where \( \Sigma = \text{var}(\tilde{Y}) \). When \( \Sigma = \sigma^2 I_n \) (i.e., the \( Y_i \)s are uncorrelated and have equal variance – assumptions (3-4)), this reduces to

\[
\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}
\]

**Proof:** Follows directly from \( \text{var}(A \tilde{Y}) = A \text{var}(\tilde{Y}) A^T \).
Mean and Variance of the OLS Estimator

**Estimation of $\text{Var}[^\hat{\beta}]$**

- **Note:** In practice, we estimate $\sigma^2$ with

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2$$

**It can “easily” be shown that $\hat{\sigma}^2$ is an unbiased and consistent estimator of $\sigma^2$ using the methods of Stat 120B/200B!**
Optimality under classical assumptions

OLS is “optimal” under the normality assumption

- If the $\epsilon_i$ are independent and distributed $\mathcal{N}(0, \sigma^2)$ then the OLS estimator is the MLE

- This means that the OLS estimator is:
  1. Consistent,
  2. Asymptotically normally distributed,
  3. Asymptotically efficient (achieves the Cramer-Rao lower bound)
Gauss-Markov Theorem

- If we do not assume normality, we may appeal to the Gauss-Markov theorem...

- Proposition: (Gauss-Markov Thm) Suppose $\text{Var}(\widetilde{Y}) = \sigma^2 I_n$. Let $\tilde{\beta} = CY$ be an unbiased estimator of $\beta$. Then, the variance of linear functions of $\tilde{\beta}$ is at least as great as the variance of linear functions of $\hat{\beta}$ (that is, the ordinary least squares estimate is the best linear unbiased estimator (BLUE) of $\beta$).
Gauss-Markov Thm

Proof:
Gauss-Markov Theorem

**Note:** Now suppose that \( \text{Var}(\tilde{Y}) = \Sigma \) is arbitrary. For a positive definite symmetric matrix we can find nonsingular symmetric matrix \( A \) such that \( \Sigma = AA \). In that case, then, \( Z = A^{-1} \tilde{Y} \) has expectation \( A^{-1}X\beta \) and variance \( A^{-1}\Sigma A^{-1} = I_n \). Letting \( W = A^{-1}X \) in this transformed model, the ordinary least squares estimate for \( \beta \) would be

\[
\hat{\beta} = (W^T W)^{-1} W^T \tilde{Z}
\]
Efficiency of the OLS estimator under non-normal errors

Gauss-Markov Theorem

- In terms of the original response $\tilde{Y}$ and predictors $X$ this yields generalized least squares estimate

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \tilde{Y}$$

which is unbiased for $\hat{\beta}$ and has variance $(X^T \Sigma^{-1} X)^{-1}$. Note that by the Gauss Markov Thm, this is the best linear unbiased estimate of $\hat{\beta}$ for this general setting.

- Note: Generalized least squares can obviously handle the case of correlated $Y_i$s. In this class, we do not consider such settings. We do however consider the setting in which the $Y_i$s are uncorrelated but do not have equal variance.
Efficiency of the OLS estimator under non-normal errors

Gauss-Markov Theorem

Definition: Consider a linear regression model in which \( \text{Var}(Y_i) = \sigma_i \) and \( \text{Cov}(Y_i, Y_j) = 0 \) for \( i \neq j \). Thus
\[
\Sigma = \text{Var}(\vec{Y}) = \text{diag}(\sigma_1, \ldots, \sigma_n).
\]
The weighted least squares estimate of \( \vec{\beta} \) is given by the generalized least squares estimate using the above definition of \( \Sigma \).

Note: The above optimality (BLUE) of the ordinary, weighted, and generalized least squares estimates is not dependent upon any particular distribution of the \( Y_i \)s beyond their first two moments. However, if we want to make inference after an analysis, we need to know the distribution of the estimates, which in turn requires some assumptions on the regression model.
Distribution of the OLS estimator under normality

Proposition: Suppose the $Y_i$s are jointly normally distributed and are uncorrelated (hence independent) (assumptions 1-6)). Then, the ordinary (weighted, generalized) least squares estimates are multivariately normally distributed. Thus in the case of constant variance,

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

Proof: This follows from linear transformations of multivariate normals.
Distribution of the OLS estimator

Consider testing the null hypothesis $H_0 : \beta_k = \beta_{k,0}$ vs $H_1 : \beta_k \neq \beta_{k,0}$

In Stat 210 you found that for the *Wald test* statistic we have:

$$T = \frac{\hat{\beta}_k - \beta_{k,0}}{\hat{se}(\hat{\beta}_k)} \sim_{H_0} t_{n-p}$$

where $\hat{se}(\hat{\beta}_k)$ is given by the square-root of the $k^{th}$ diagonal element of $\text{Var}[\hat{\beta}] = \hat{\sigma}^2 (X^T X)^{-1}$ with

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

A $100(1-\alpha)\%$ CI for $\beta_k$ is given by computing

$$\hat{\beta}_k \pm t_{n-p,1-\alpha/2} \hat{se}(\hat{\beta}_k)$$
Asymptotic normality of OLS

► **Question:** What happens when the normality assumption is not satisfied?

► **Answer:** Like most (useful) estimators we can approximate the sampling distribution in large samples! To do this, we must appeal to the Lindeberg-Feller Central Limit Theorem...
Lindeberg-Feller Central Limit Theorem

- Proposition: (Lindeberg-Feller Central Limit Theorem) Let $Y_1, Y_2, \ldots$ be independent random variables with $E[Y_i] = 0$ and $\text{var}(Y_i) = \sigma_i^2$. Define $S_n = \sum_{i=1}^{n} Y_i$ and $\sigma^2(n) = \sum_{i=1}^{n} \sigma_i^2$. Then both

1. $S_n/\sigma(n) \xrightarrow{d} N(0, 1)$, and
2. $\lim_{n \to \infty} \max\{\sigma_i^2/\sigma^2(n), 1 \leq i \leq n\} = 0$

if and only if (the Lindeberg condition)

$$\forall \epsilon > 0 \lim_{n \to \infty} \frac{1}{\sigma^2(n)} \sum_{i=1}^{n} E\left[|Y_i|^2 1_{|Y_i| \geq \epsilon \sigma(n)}\right] = 0$$
Distribution of the OLS Estimator

Asymptotic normality of OLS

- Proposition: Consider simple linear regression in which \((Y_i, X_i)\) are pairs of response R.V.'s and known predictors. \(Y_i\)'s are independently distributed \(Y_i \sim (\mu_i, \sigma^2)\) with \(\sigma^2 < \infty\) known. In particular, we will consider regression model of the form \(\mu_i = \beta_0 + \beta_1 (X_i - \bar{X})\) and assume

\[
(Y_i - \mu_i) \sim iid (0, \sigma^2)
\]

Further, let \(X = (1 \quad \bar{X} - \bar{X})\), \(\beta = (\beta_0 \quad \beta_1)^T\) and consider the OLSE \(\hat{\beta} = (X^T X)^{-1} X \hat{Y}\). Then,

\[
\tilde{Z}_n = (X^T X)^{1/2} (\hat{\beta} - \beta)
\]

\[
= \left( \begin{array}{c} \sqrt{n}(\hat{\beta}_0 - \beta_0) \\ \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (\hat{\beta}_1 - \beta_1)} \end{array} \right) \rightarrow_d \mathcal{N}_2(0, \sigma^2 I_2)
\]
Asymptotic normality of OLS

Proof:
Distribution of the OLS Estimator

Asymptotic normality of OLS

- Conclusion: Even if we do not assume normality, but simply have independence between the errors, the ordinary least squares estimate will be asymptotically normally distributed as long as

\[
\max \left\{ \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right\} \to 0 \text{ as } n \to \infty
\]

by the Lindeberg-Feller CLT. In particular, in the case of constant variance we have

\[
\hat{\beta} \sim N(\bar{\beta}, \sigma^2 (X^T X)^{-1})
\]
Consider the regression model $Y_i = \mu_i + \epsilon_i$

### Varying degrees of assumptions

1. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i$

2. $\epsilon$ independent and identically distributed with mean zero

3. $\epsilon$ independent with constant variance and mean zero

4. $\epsilon$ independent with mean zero

5. $\epsilon$ has mean zero
Consider the regression model \( Y_i = \mu_i + \epsilon_i \)

Weaker assumptions lead to weaker properties for the OLS estimator

1. OLS is optimal (consistent, unbiased, most efficient)

2. OLS is consistent and is the ‘best linear unbiased estimator’ (BLUE)

3. OLS is consistent and is the ‘best linear unbiased estimator’ (BLUE)

4. OLS is consistent and asymptotically Normal

5. No guarantees (OLS consistent and asymptotically Normal under additional assumptions)
What is the effect of changing the error distribution?

Thus, changing the error distribution could...

1. Could change \( \text{Var}[\hat{\beta}] \)
   - In repeated experimentation, \( \hat{\beta} \) varies more than it would if \( \epsilon \sim N(0, \sigma^2) \)

2. Could affect the efficiency of \( \hat{\beta} \)
   - In repeated experimentation, \( \hat{\beta} \) varies more than some other estimator of \( \beta \)

3. Could make \( \hat{\sigma}^2(X^T X)^{-1} \) a bad estimate of \( \text{Var}[\hat{\beta}] \)
   - In repeated experimentation, the variability of \( \hat{\beta} \) is greater (or less) than \( \hat{\sigma}^2(X^T X)^{-1} \)
What is the effect of changing the error distribution?

These results are distinct...

- The above results of changing the error distribution are all different phenomena

  - Items (1) and (2) mean that another estimator may be more efficient (smaller variability) than the OLS estimator

  - Item (3) means that if we estimate $\text{Var}([\hat{\beta}])$ by $\sigma^2(X^TX)^{-1}$ then our inference for $\hat{\beta}$ will be wrong:

    - Type I error rate of hypothesis tests will be higher (lower) than the nominal level

    - Confidence intervals will not have the correct coverage probability
The role of heteroscedasticity

(3) occurs when the variance of the error terms is not constant

- Why does this matter to us?

1. Suppose that our response is a binary outcome variable $Y$
   - $Y_i \sim Binom(\mu_i, 1)$
   - Standard linear regression mean model: $E[Y_i] = \mu_i = X_i \beta$
   - Error distribution: $Var[Y_i] = \mu_i (1 - \mu_i)$

2. Suppose that our response $Y$ counts the number of events over a specified interval
   - Might assume $Y_i \sim Poisson(\mu_i)$
   - Standard linear regression mean model: $E[Y_i] = \mu_i = X_i \beta$
   - Error distribution: $Var[Y_i] = \mu_i$

- *Note: Nonconstant variance can also cause (1) and (2)*
The role of heteroscedasticity

Bottom line

► Because of the mean-variance relationship in these (and many other) outcome distributions, we cannot fulfill the constant variance assumption!

► \( \hat{\sigma}^2 (X^T X)^{-1} \) is a bad estimate for \( \hat{\beta} \)

► Invalid inference

► Much of our class will be devoted to deriving a general class of estimators for regression models where a mean-variance assumption exists...