3.3 Linear Regression

PEAKS2TAILS

Simple Linear Regression

Simple Linear Regression

1 **Simple Linear Regression Model**

$$y = \beta_0 + \beta_1 x + u$$

(y - dependent variable, x - independent variable, u - error term / shock)

2 **Two Fundamental Assumptions**

$\mathbb{E}[u] = 0$ (hardly an assumption as long as intercept β_0 is included)

$\mathbb{E}[u|x] = \mathbb{E}[u]$ (In other words, Cov (x, u) = 0 ... exogeneity)

$\mathbb{E}[u|x] = 0$ (combining the above two, we get Zero Conditional Mean)

3 **Systematic & Unsystematic component of y**

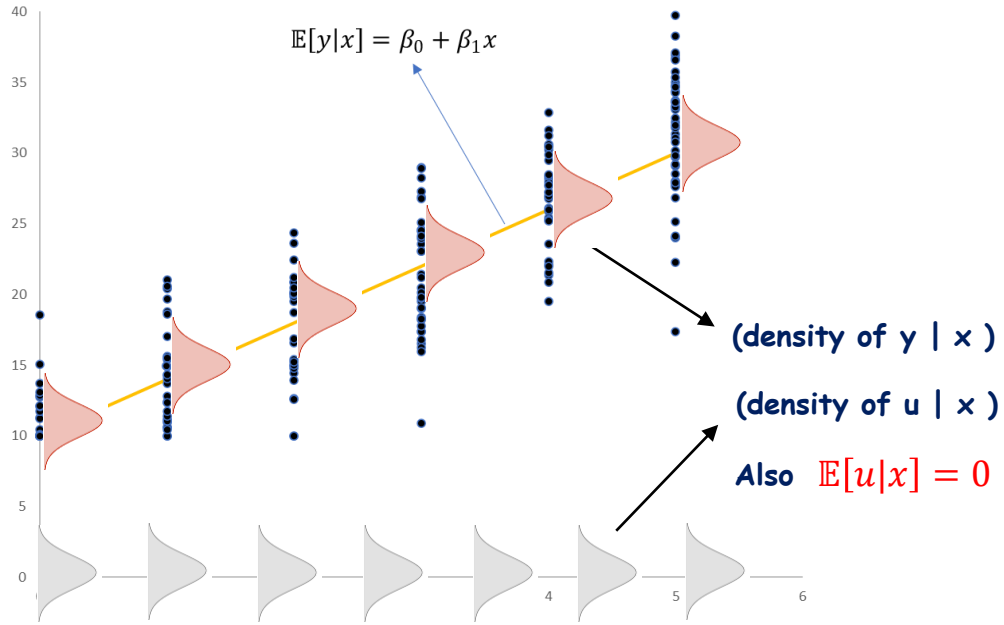
Take conditional Expectation against x and use the above assumption to get ...

$\mathbb{E}[y|x] = \beta_0 + \beta_1 x$ so $y = \mathbb{E}[y|x] + u$ → (unsystematic part)

↓

(systematic part)

4 **Visualization of population dynamics**

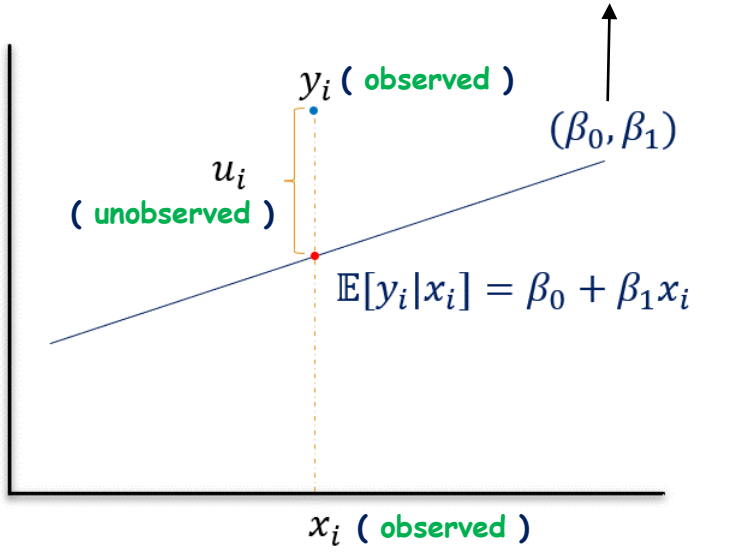


Simple Linear Regression

5 **Population dynamics vs sample dynamics**

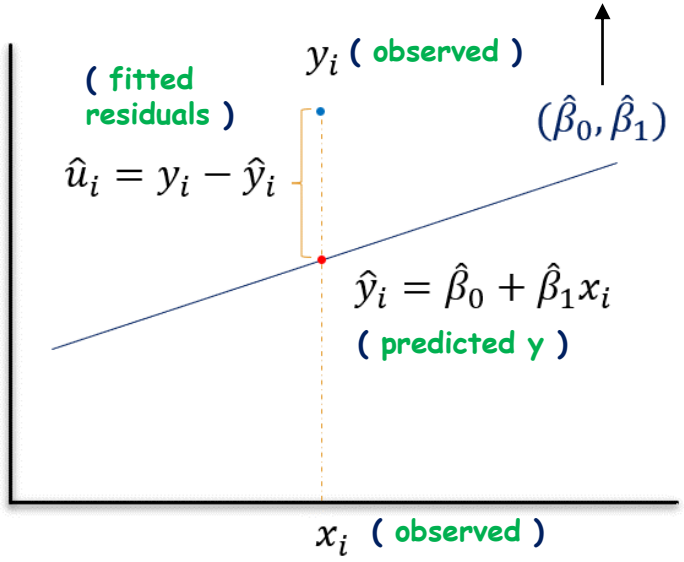
(Population Process) $y_i = \beta_0 + \beta_1 x_i + u_i$

(Population parameters)



(Sample Process) $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$

(estimated from sample)



Sum of Total Squares (SST) : $\sum_{i=1}^{i=n} (y_i - \bar{y})^2$

(total variation of y around its mean)

Explained Sum of Squares (SSE) : $\sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2$

(total variation of y around its mean explained by the model)

Residual Sum of Squares (SSR) : $\sum_{i=1}^{i=n} \hat{u}_i{}^2 = \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2$

(total variation of y not explained by the model)

Simple Linear Regression

6 **Deriving the OLS estimate**

OLS Estimator : Minimize SSR

$$\sum_{i=1}^{i=n} \hat{u}_i{}^2 = \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{i=n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$$\frac{\partial(SSR)}{\partial \beta_0} = \sum_{i=1}^{i=n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\frac{\partial(SSR)}{\partial \beta_1} = \sum_{i=1}^{i=n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

OLS forces the follwing results in the sample

(1) - $\sum_{i=1}^{i=n} \hat{u}_i = 0$

(Sum/mean of residuals is 0)

(2) - $\sum_{i=1}^{i=n} \hat{u}_i x_i = 0$

(residual vector is orthogonal to x or we can say sample cov (x, $\hat{u}$) = 0)

(3) - The point ($\bar{x}, \bar{y}$) always lies on the OLS regression line (because of (1))

(4) - with (1) and (2), SST = SSE + SSR

Solving (1) and (2), we get the estimates of beta coefficients ...

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \quad \text{or} \quad \hat{\beta}_1 = \hat{\rho}_{xy} \left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x}\right)$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ because of (3)

Simple Linear Regression

6 **Goodness of fit**

It's % of total variation of y captured by the model

coefficient of determination (R^2) = SSE / SST = 1 - (SSR / SST)

Also *for a given sample*, $y = \hat{y} + \hat{u}$

$Var(y) = Var(\hat{y}) + Var(\hat{u})$ because cov (x, $\hat{u}$) = 0 , cov ($(\hat{\beta}_0 + \hat{\beta}_1 x)$, $\hat{u}$) = 0

We can also define, $R^2 = \frac{SSE/(n-1)}{SST/(n-1)} = \frac{Var(\hat{y})}{Var(y)}$

$$\frac{Var(\hat{y})}{Var(y)} = \frac{Var(\hat{\beta}_0 + \hat{\beta}_1 x)}{Var(y)} = \hat{\beta}_1{}^2 \frac{Var(x)}{Var(y)} = \left(\hat{\rho}_{xy}\left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x}\right)\right)^2 \left(\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\right)^2 = \hat{\rho}_{xy}{}^2$$

$R^2 = \hat{\rho}_{xy}{}^2$ (square of the correlation (x , y)) $0 \leq R^2 \leq 1$

Higher the R^2 better is the model fit

Simple Linear Regression

7 **Unbiasedness of OLS beta coefficients**

beta coeffcients are unbiased : $\mathbb{E}[\hat{\beta}_0] = \beta_0 \quad \mathbb{E}[\hat{\beta}_1] = \beta_1$

SLR-1 Linear in parameters (population dynamics is linear w.r.t betas)

SLR-2 Random Sampling (sample is drawn randomly from the population)

SLR-3 Sample variation in the regressor (x must not be constant)

SLR-4 Zero Conditional Mean (combines both exogeneity and zero mean of error)

for the same x, sample y can change because of random nature of u. if y changes beta estimates will change

$$\hat{\beta}_1 = \beta_1 + \left(\frac{1}{SST_x}\right)\sum_{i=1}^{i=n}(x_i - \bar{x})u_i$$

$\mathbb{E}[\hat{\beta}_1] = \beta_1$ because of SLR-4

$\mathbb{E}[\hat{\beta}_0] = \beta_0$ follows from $\mathbb{E}[\hat{\beta}_1] = \beta_1$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

Simple Linear Regression

8 **Variance and covariance of OLS beta coefficients**

To simplify variance calculation, we add another assumption

SLR-5 **Homoscedasticity** (constant conditional variance : $Var(u|x) = \sigma^2$)

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \qquad Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{i=n} x_i^2}{n.SST_x} \qquad Cov(\hat{\beta}_1, \hat{\beta}_0) = -\frac{\bar{x}\sigma^2}{SST_x}$$

Assumptions from SLR-1 to SLR-5 are known as Gauss-Markov assumptions. But note that only 1 to 4 are needed for unbiasedness. (More discussion on this in Multiple Regression sheet)

Simple Linear Regression

9 **Error variance**

Error variance can be calculated from the fitted residuals

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{i=n} \hat{u}_i{}^2 = \frac{SSR}{n-2}$$

The denominator loses 2 degree of freedoms, because the fitted residuals are constrained by two equations

(1) - $\sum_{i=1}^{i=n} \hat{u}_i = 0$ (2) - $\sum_{i=1}^{i=n} \hat{u}_i x_i = 0$

resampling y for the same x will generate 'n' new fitted residuals, but only (n-2) are free

The above is an *unbiased estimator* of error variance

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

Simple Linear Regression

10 **Standard Error of Regression**

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ ...is called the SER

This is linked to the standard error of beta coefficients and influences confidence intervals and hypothesis testing of beta coefficients

$$S.E.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}} \qquad S.E.(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{\sum_{i=1}^{i=n} x_i^2}{n.SST_x}}$$

Note that SER $\hat{\sigma}$ is a biased estimator of σ Expectation of a square root is not the square root of an expectation

11 **Interpretation of S.E.**

(1) - Note SSTx is the key denominator for variance of beta coefficients. If x does not vary (SLR - 3), SSTx will be 0 and the variances will be infinity and so the beta estimates will be unreliable.

(2) - The S.E. of beta coefficients increase if the error variance increases. (It makes harder to estimate beta if the error variance is high)

(3) - if the variation in x (SSTx) is high, it actually reduces S.E. of betas. The reason is we will be able to capture major variation of y through the variation of x.

Multiple Linear Regression

Multiple Linear Regression

1 Multiple Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$ 2 regressors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$ k regressors

2 Fundamental Assumption

$$\mathbb{E}[u|x_1, x_2] = 0$$ Zero conditional mean

Similar to SLR, the error term has zero mean conditional on any combination of regressors' values

3 Population and Sample Dynamics in Matrix form

In Multiple regression, it's convenient to express in matrix form

(Population Process) $\boldsymbol{y = X\beta + u}$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

error term $\boldsymbol{u}$ (n x 1)

observed $\boldsymbol{y}$ (n x 1)

beta vector $\boldsymbol{\beta}$ (k x 1)

Design Matrix $\boldsymbol{X}$ (n x (k+1))

(Sample Process) $\boldsymbol{y = X\hat{\beta} + \hat{u}}$

residual vector $\hat{\boldsymbol{u}}$ (n x 1)

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$

observed $\boldsymbol{y}$ (n x 1)

estimated beta vector $\hat{\boldsymbol{\beta}}$ (k x 1)

Design Matrix $\boldsymbol{X}$ (n x (k+1))

Multiple Linear Regression

3 **Deriving OLS estimates**

For a sample, we have $y_i = \hat{y} + \hat{u}_i$

where $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

OLS Estimator : Minimize

$$\sum_{i=1}^{i=n} \hat{u}_i{}^2 = \sum_{i=1}^{i=n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}\right)^2$$

taking partial derivative wr.r.t each of the beta coeffs
we get the following system of equations

$$\sum_{i=1}^{i=n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}\right) = 0$$

$$\sum_{i=1}^{i=n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}\right) x_{1i} = 0$$

$$\sum_{i=1}^{i=n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}\right) x_{ki} = 0$$

(1) - $\sum_{i=1}^{i=n} \hat{u}_i = 0$ (Sum/mean of residuals is 0)

(2) - $\sum_{i=1}^{i=n} \hat{u}_i x_{ji} = 0$ (for each x_j , j=1,2,...k)

(residual vector is orthogonal to each x_j or we can say sample cov $(x_j, \hat{u})$ = 0)

Combining (1) & (2), we can write in compact form

$$X^T \hat{u} = 0$$

Now we start with the sample process $y = X\hat{\beta} + \hat{u}$

Multiplying X^T we get $X^T y = X^T X\hat{\beta} + X^T \hat{u}$

$$X^T y = X^T X\hat{\beta}$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

Multiple Linear Regression

4 **Interpreting OLS coefficients in MLR**

Multiple linear regression is amenable to ceteris paribus analysis

For example, for a regression involving 2 regressors

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x_1}$$ keeping x_2 constant

for example, if we estimated

$$\widehat{marks} = 10 + 4(hours\ of\ study) + 2(class\ attendance)$$

for a sample of students with the same class attendance, average marks will go up by 4 if hours of study increases by 1 hour

Note that we din't have to fix the feature while drawing the sample to come up with the beta estimates. Our sample consisted of both regressors taking different values, but we could still obtain such a powerful equation that allows what if analysis on individual regressors, keeping the effect of others constant

of course, if both variables change, we can deduce the combined effect on $\hat{y}$ i.e. $\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$

Another way to express the slope is

$\hat{r}_1$ is the OLS residual of regressing x_1 on x_2

then regress y on $\hat{r}_1$ to obain $\hat{\beta}_1$

This is called *partialing out* / *netting out* the effect of x_2

Multiple Linear Regression

5 **Comparing OLS beta coeff between SLR and MLR**

Let's say the true model includes two regressors x_1 and x_2 but we under-specified the model (we performed SLR with only x_1)

Dynamics of both the models are below

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ ... true model

$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$ ... underspecified (SLR)

Let's say x_1 and x_2 are correlated such that $\hat{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$

In the underspecified model the error terms will contain x_2 which we are not holding constant. So net change in y in SLR because of change in x_1 will be contaminated through the change in x_2

The true model can be re-written as

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (\tilde{\delta}_0 + \tilde{\delta}_1 x_1)$

if we fit SLR between y and x_1, the OLS estimate for the slope will be

$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \, \tilde{\delta}_1 \qquad \mathbb{E}[\tilde{\beta}_1] = \beta_1 + \beta_2 \, \tilde{\delta}_1$

Clearly underspecification leads to a bias $(\beta_2 \, \tilde{\delta}_1)$

The bias does not exist if

(1) - $\beta_2 = 0$, in other words x_2 is an irrelevant attribute that does not affect y

(2) - $\tilde{\delta}_1 = 0$, in other words x_2 and x_1 are uncorrelated

Multiple Linear Regression

6 Goodness of fit

The goodness of fit R^2 can be calculated as SSE / SST

Alternatively, R^2 is the squared correlation between y and $\hat{y}$

R^2 always increases with inclusion of additional variables. True model performance needs to be adjusted to penalize number of variables taken into the model

We use $Adjusted\ R^2$ that only improves if the new variable has good explanatory power measured against the cost of being added to the model

$$Adjusted\ R^2 = 1-(1-R^2)\frac{n-1}{n-k-1}$$

- $Adjusted\ R^2 \uparrow$ when $R^2 \uparrow$
- $Adjusted\ R^2 \downarrow$ when $k \uparrow$

Also, $Adjusted\ R^2 = 1 - \frac{Var(u)}{Var(y)}$

7 Unbiasedness of OLS estimates

Similar to SLR Assumptions, we need to specify MLR assumptions to derive the unbiasedness of OLS estimates in multiple regression

- MLR-1: Linear in parameters (population dynamics is linear w.r.t betas)
- MLR-2: Random Sampling (sample is drawn randomly from the population)
- MLR-3: No perfect Multicollinearity (regressors are linearly independent)
- MLR-4: Zero Conditional Mean ($\mathbb{E}[u|x_1, x_2, \dots, x_k] = 0$)

Under the above assumptions from (1) to (4), $\mathbb{E}[\hat{\beta}] = \beta$

Multiple Linear Regression

8 Overspecifying the model (addition of irrelevant variables)

Adding an irrelevant regressor which has no effect on y at all will have a beta coefficient of 0 in the population process. Inclusion or exclusion of such a regressor does not impact the unbiasedness of other betas. This follows directly from 5 - (1)

9 Ommited variable bias (excluding a relevant attribute)

Let's say the true population model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

But we specify $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u'$

We dropped x_2 because of either *lack of awarenes* or *data unavailability* or *inability to measure*

We already know $\tilde{\beta}_1$ will be biased in general and the bias is given as $(\beta_2 \tilde{\delta}_1)$. Size of the bias will be difficult to know as we don't know β_2 as we have excluded x_2 . But we would have an idea of the direction of correlation between y and x_2 (in other words sign of β_2). This will give us an idea of direction of the bias at least.

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

In multiple regression model, omitting a variable from the model results in error term capturing that omitted variable. If the omitted variable is correlated with any of the regressors, it violates MLR-4 and results in all beta coefficients being *biased* in general

Multiple Linear Regression

8 **Variance of OLS estimates**

we specify another assumption to derive variance of OLS estimates as we did in case of SLR

MLR-5 **Homoscedasticity** ($Var(u|x_j) = \sigma^2$)

Variance-covariance of beta vector in matrix form $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

Variance of individual beta coeff $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1-R_j{}^2)}$

SST_j is the sum of total squares of jth regressor x_j

$R_j{}^2$ is R^2 of regressing x_j on all other regrssors

9 **Dissecting the OLS variance**

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j{}^2)}$$

Variance of individual beta coefficients depend on 3 factors

(1) - σ^2 : error variance (direct)

(2) - SST_j: $\sum_{i=1}^{i=n}(x_{ji}-\bar{x})^2$ (Inverse)

(3) - $R_j{}^2$: $R_j{}^2$ is R^2 of regressing x_j on all other regressors (direct)

Multiple Linear Regression

9(a) **Effect of Multicollinearity**

If x_j is highly correlated with other regressors, then R_j^2 will be high

This will lead to high variance in $\hat{\beta}_j$

In the extreme case, the estimate blows up in presence of perfect collinearity i.e. $R_j{}^2 = 1$

The problem of multicollinearity can be reduced by increasing more sample size. This will increase SST_j

Dropping a variable that is highly correlated with others is not always the solution because

(1) - it will lead to OBV where other beta coefficients will be biased
(2) - the variable may be an important explanatory variable for economic reasons

Suppose we want to study discrimination while approving loans by measuring the impact of minority % on loan approval rates for various localities

$$loan\ approval\ rates = \beta_0 + \beta_1(minority\%) + \beta_2(avg\ income) + \beta_3(avg\ housing\ price) + u$$

avg income and avg house prices can be correlated but have nothing to do with minority %. So this will not impact the variance of $\hat{\beta}_1$ which is of our primary interest in the study.

Some statisticians have proposed $VIF_j = \frac{1}{(1 - R_j{}^2)}$

High VIF indicates high collinearity for a regressor. We want the VIF to be small, however discarding variables solely on basis of VIF is not a good idea for the reasons discussed before

Multiple Linear Regression

9(b) **Effect of mis-specification on OLS Variance**

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ (1)

$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$ (2)

we know $\tilde{\beta}_1$ will be biased $\tilde{\beta}_1 = \beta_1 + \beta_2 \tilde{\delta}_1$

However, $Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1 - {R_1}^2)}$

wheras $Var(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$

10 **Estimating error variance**

error variance can be estimated from residuals after adjusting for degree of freedom

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{i=n} {\hat{u}_i}^2 = \frac{SSR}{n-k-1}$$

Under assumptions MLR-1 to MLR-5, $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

We can now estimate the S.E. on the beta coefficients

$$S.E.(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - {R_j}^2)}}$$

Multiple Linear Regression

11 **Gauss Markov theorem**

We have already seen that OLS is unbiased

It's also a linear estimator as it depends on y in a linear fashion i.e.

$\hat{\beta} = (X^T X)^{-1} X^T y$ is of the form $\hat{\beta} = Cy$

It's best in terms of efficiency i.e. of the entire class of linear estimators, OLS is the one with the minimum variance

$$
\begin{aligned}
\mathrm{E}[\tilde{\beta}] &= \mathrm{E}[Cy] \\
&= \mathrm{E}\left[\left((X'X)^{-1}X' + D\right)(X\beta + \varepsilon)\right] \\
&= \left((X'X)^{-1}X' + D\right)X\beta + \left((X'X)^{-1}X' + D\right)\mathrm{E}[\varepsilon] \\
&= \left((X'X)^{-1}X' + D\right)X\beta && \mathrm{E}[\varepsilon] = 0 \\
&= (X'X)^{-1}X'X\beta + DX\beta \\
&= (I_K + DX)\beta.
\end{aligned}
$$

Therefore, since β is **un**observable, $\tilde{\beta}$ is unbiased if and only if $DX = 0$. Then:

$$
\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}(Cy) \\
&= C\,\mathrm{Var}(y)C' \\
&= \sigma^2 CC' \\
&= \sigma^2\left((X'X)^{-1}X' + D\right)\left(X(X'X)^{-1} + D'\right) \\
&= \sigma^2\left((X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'\right) \\
&= \sigma^2(X'X)^{-1} + \sigma^2(X'X)^{-1}(DX)' + \sigma^2 DX(X'X)^{-1} + \sigma^2 DD' \\
&= \sigma^2(X'X)^{-1} + \sigma^2 DD' && DX = 0 \\
&= \mathrm{Var}\left(\hat{\beta}\right) + \sigma^2 DD' && \sigma^2(X'X)^{-1} = \mathrm{Var}\left(\hat{\beta}\right)
\end{aligned}
$$

Since DD' is a positive semidefinite matrix, $\mathrm{Var}(\tilde{\beta})$ exceeds $\mathrm{Var}\left(\hat{\beta}\right)$ by a positive semidefinite matrix.

Multiple Linear Regression

9 **Sampling Distribution of OLS estimates**

We derived Mean and Variance of OLS distribution. But we need the full distribution to be able to do hypothesis testing

So now we add another assumption

MLR-6 **Normality** (error terms are normally distributed)

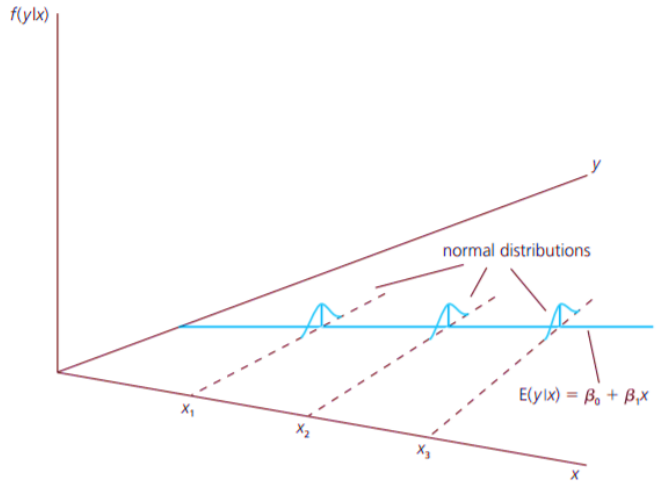
$$u \sim N(0, \sigma^2)$$

This assumption is a much stronger assumption and it encapsulates MLR-4 and MLR-5

Under these 6 assumptions, OLS estimators are the best estimators across all classes of estimators (not just linear)

We can now specify the distribution of **y | x**

$$y|x \sim N\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \ , \sigma^2\right)$$



Now that y is normal and OLS estimator is linear in y, OLS beta estimates will also be normal

$$\hat{\beta}_j \sim N\left(\beta_j, Var(\hat{\beta}_j)\right) \quad \text{or} \quad \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim N(0,1)$$

But, $SE(\hat{\beta}_j)$ is estimated from the sample (unknow sd)

$$S.E.(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j\left(1 - {R_j}^2\right)}}$$

It's well known that if population variance is unknown,

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-k-1} \qquad \text{Also,} \quad (n-k-1)\hat{\sigma}^2 \sim {\chi^2}_{n-k-1}$$

Now we are all set for hypothesis testing.

Multiple Linear Regression

10 Hypthesis testing on beta parameters

It's natural to ask the question if any beta coefficient is 0 i.e. a regressor is irrelevant to explain y

$H_0: \beta_j = 0$ two tailed t-test, reject if $\left|\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right| > c_{\alpha/2}$

$H_0: \beta_j < 0$ one-tailed t test, reject if $\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} > c_{\alpha}$

$H_0: \beta_j > 0$ one-tailed t test, reject if $\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} < c_{1-\alpha}$

Hypothesis can be against different values of beta other than 0. We just need to use the hypthesized mean in the t-stat calculation

11 Economic significance vs statistical significance

In the econometric model, the regressor with the largest beta coefficient influences y the most. In other words, the economic significance is purely the size of the β_j

The statistical significance of β_j is given by $\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

Guidelines :

(1) start with statistical significance, if t-stat is significant discuss the magnitude and impact of β

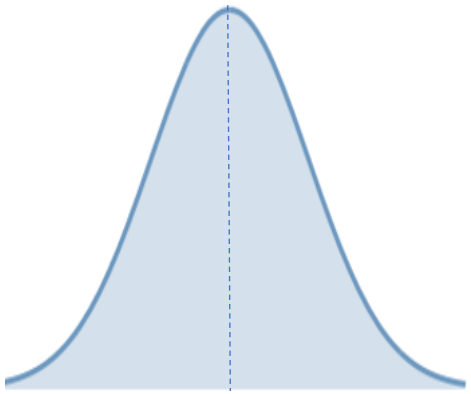
(2) if statistically insignificant, check if the magnitude of β is much larger than expected. The statistical insignificance could be because of random sampling or small sample size. You may still argue to include the variable.

Multiple Linear Regression

12 **Confidence Interval around beta parameters**

Using the t-distribution and desired confidence level, we can construct confidence intervals for the population beta

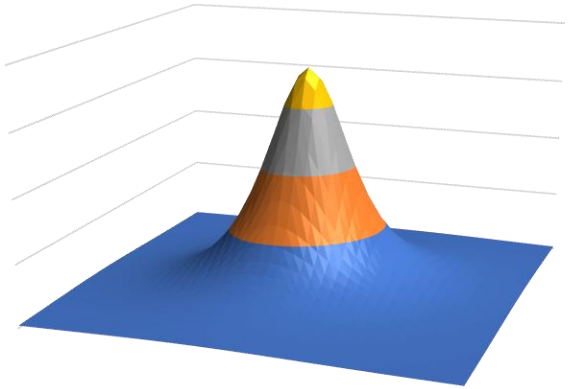
confidence interval for β_j (univariate)



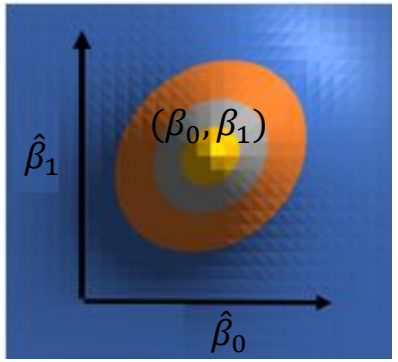
$\hat{\beta}_j - (t_{\alpha/2,n-k-1})SE(\hat{\beta}_j)$ β_j $\hat{\beta}_j - (t_{\alpha/2,n-k-1})SE(\hat{\beta}_j)$

Confidence region can be constructed joint parameters using the multivariate t-distribution of the beta estimates

e.g. we can use the bivariate t-distribution for $< \hat{\beta}_1, \hat{\beta}_2 >$ to construct confidence region for $< \beta_1, \beta_2 >$



bivariate t-distribution for $< \hat{\beta}_0, \hat{\beta}_1 >$



elliptical confidence region for $< \beta_0, \beta_1 >$

Multiple Linear Regression

13 **Hypothesis Testing of a linear combination of parameters**

We may want to test a statement on the linear combination of betas

e.g. $H_0: \beta_1 = \beta_2$

The key is to find the distribution of the test statistic

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

$$Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2cov(\hat{\beta}_1, \hat{\beta}_2)$$

$cov(\hat{\beta}_1, \hat{\beta}_2)$ can be obtained from VarCovar matrix $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)}$$

Multiple Linear Regression

14 **Joint Hypothesis Testing of Multiple conditions**

Suppose our original model contains k+1 parameters and is given as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

We want to jointly test the last q beta coefficients are 0

$$H_0: \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \ldots, H_0: \beta_k = 0$$

We remove the regressors whose beta coefficients we are hypothesizing as 0 and call that model as 'restricted model' and call the original model 'unrestricted model'

The idea is if the hypothesized beta coefficients are zero, and the corresponding regressors are irrelevant, then dropping those should not materially change the error variance. In other words, we will compare the ratio of SSR for unrestricted model vs restricted model adjusted by their degree of freedom. This will be an 'F-test'

$$F = \frac{(SSR^r - SSR^{ur})/q}{SSR^{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

q = df(ur) - df(r)

we will reject the null (restricted = unrestricted), if F > c (threshold)

Note that F-statistic will work if we are testing one beta parameter. It will become square of the t-statistic.

$$t_{n-k-1}{}^2 = F_{1,n-k-1}$$

The above F-statistic can be written in terms of R^2

$$F = \frac{(R_{ur}{}^2 - R_r{}^2)/q}{(1 - R_{ur}{}^2)/(n-k-1)}$$

15 **Overall Significance of Regression**

Whole purpose of doing regression is X explains y
We can jointly test if all slope coefficients are 0 (in other words no regressors explain y)

$$H_0: \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0,$$

Our restricted model will be

$$y = \beta_0 + u$$

$$F = \frac{(R^2)/k}{(1-R^2)/(n-k-1)}$$ (overall significance of regression)

Multiple Linear Regression

16 **Error in Prediction**

The actual y at $x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k$ (new sample) for a given subject (p) would be

$$y_p = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + u^p$$

u^p denotes the unobserved error term for the subject (p)

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

Estimated / Predicted y

$$y_p = \hat{y}_p + \hat{u}_p \qquad \hat{u}_p = y_p - \hat{y}_p$$

source of variance is out of sample error terms — uncorrelated — source of variance is in sample error terms

Variance of the 'prediction error' is

$$Var(\hat{u}_p) = Var(y_p) + Var(\hat{y}_p) = \hat{\sigma}^2 + \boldsymbol{c^T}\hat{\sigma}^2\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{c}$$

$$s.e.(\hat{u}_p) = \hat{\sigma}\sqrt{1 + \boldsymbol{c^T}(\boldsymbol{X^T X})^{-1}\boldsymbol{c}}$$

95% Prediction interval of estimated y is $\hat{y}_p \pm 1.96\ s.e.(\hat{u}_p)$

Multiple Linear Regression

17 **OLS Asymptotics**

Consistency is is an important property of an estimator. It says when the sample size goes to infinity,the distribution of the estimator converges in probability on the population parameter. The distribution will get slimmer and slimmer and collapse onto the population parameter.

Under the assumptions of MLR-1 to MLR-4 , OLS estimate is consistent. Practically, the most important assumption is 'exogeneity' ; the zero conditional mean of error terms.

$$\hat{\beta}_1 = \beta_1 + \left(\frac{1}{SST_x}\right)\sum_{i=1}^{i=n}(x_i - \bar{x})u_i$$

divide by n in the numerator and denominator

$$plim \quad \hat{\beta}_1 = \beta_1 + \frac{Cov(x,u)}{Var(x)}$$

plim - *probability limits*

MLR-4 indicates Cov (x,u) = 0, so the limiting distribution of the estimator collapses on the parameter. This is 'consistency'

We already know that if any of the x is correlated with u, generally all OLS estimates will be biased.

So just the **failure of MLR-4** will make OLS estimates *biased* and *inconsistent*. Bigger sample size won't help to get rid of bias because it's inconsistent.

Omitted variable causes bias and the bias depends on the sample slope coefficient between the regressors

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\,\tilde{\delta}_1$$

$$plim \quad \tilde{\beta}_1 = \beta_1 + \beta_2\delta_1 \quad where\ \delta_1 = \frac{Cov\,(x_1, x_2)}{Var(x_1)}$$

Omitted variable causes bias and inconsistency which are practically the same. A subtle point is, even if the population covariance (x1, x2) = 0, the estimator will be a consistent estimator but can remain biased for the sample of (x1, x2) in the data as sample covariance which determines $\tilde{\delta}_1$ may not be 0.

OLS estimates are asymptotically normal i.e. the estimates follow normal distribution regardless of normality of error terms. This is due to CLT

Multiple Linear Regression

18 **Scaling, standardization of variables**

if y is scaled scaled : beta, residuals, SER unchanged : t-stat, F-stat and R^2

standardization of the variables puts all in equal footing (removes scale or unit dependence) so the standardized beta coefficients can be compared across.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}$$

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \hat{v}$$ (standardized variables)

Here are the effect of standardization

(i) Intercept goes to 0
(ii) beta coefficients get scaled : $\hat{b}_j = \frac{\hat{\sigma}_j}{\hat{\sigma}_y}\hat{\beta}_j$

19 **Interaction terms**

Sometimes the partial effect of a regressor on y depends on the value of another regressor. We achieve this by adding an 'Interaction term'

$$expenditure = \beta_0 + \beta_1(salary) + \beta_2(gender) + \beta_3(gender * salary) + u$$

Interaction term

$$\frac{\partial(expenditure)}{\partial(salary)} = \beta_1 + \beta_3 * (gender)$$

We can see the partial effect of salary on expenditure depends on value of gender

We calculate a term called 'Average Partial Effect (APE)' which in the previous example would be $\hat{\beta}_1 + \hat{\beta}_3 * (\overline{gender})$

Multiple Linear Regression

20

Adding Binary (Dummy) variables

Categorical variables can be dummy coded and added to the linear regression. The only catch is to use one less number of dummy variables than the original number of categories.

e.g. Gender can be added as a variable (1/ 0)

If a variable (Job Type) has 4 values in the data say 'Engineer', 'Doctor', "Teacher', 'Consultant'.. we can use 3 dummy variables

Job Type	Dummy_1	Dummy_2	Dummy_3
Engineer	0	0	0
Doctor	1	0	0
Teacher	0	1	0
Consultant	0	0	1

One variable 'Job Type' can be replaced by 3 dummy variables and combination of binary codes represent a particular level for an observation.

We simply cannot use 4 dummy variables (called dummy variable trap) because then the 4 dummy variables will be perfectly correlated (sum of 4 dummy variables would be always 1)

The beta coefficient against a dummy variable shifts the intercept of the equation for different levels in relative to the base level (base level is the one where all binary codes are 0 for the dummy variables)

The beta coefficient against a dummy variable added as an interaction term affects the slope of the term it is interacting with.

Multiple Linear Regression

21 **Chow Test**

Suppose there are two groups g=1, g=2 and we want to test if the beta coefficients (slopes and intercepts) for the two groups are same or not.

Unrestricted Model (df = n - 2(k+1))

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \alpha_0 g + \alpha_1 g x_1 + \cdots + \alpha_k g x_k + u$$

restricted Model (df = n - (k+1))

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Hypothesis

$$\alpha_0 = 0, \alpha_1 = 0, \ldots, \alpha_k = 0$$

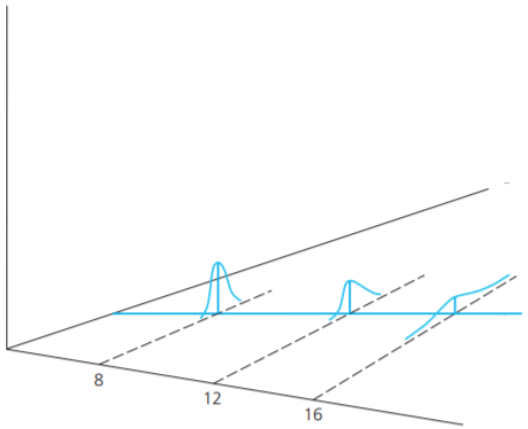
The SSR for the unrestricted model can be calculated by summing SSR1+SSR2 where SSR1 and SSR2 are the residual sum of squares by regressing with group1 and group2 respectively

$$F = \frac{\left(SSR^r - (SSR_1 + SSR_2)\right)/(k+1)}{(SSR_1 + SSR_2)/(n-2k-2)} \sim F_{k+1,n-2k-2}$$ Chow Test

22 **Heteroscedasticity**

Heteroscedasticity violates asumption (MLR - 5) and indcues bias to the OLS variances

Heteroscedasticity plays no role as far as unbiasedness or consistency of the OLS estimates are concerned or population R^2 (depends on unconditional variance)



Heteroscedasticity will make the variance of OLS estimates $Var(\hat{\beta}_j)$ biased and affect inferential statistics (t-tests, F-tests etc.)

If homoscedasticity is violated then OLS is no longer BLUE

Multiple Linear Regression

22(a) **Heteroscedasticity - robust inference after OLS estimators**

robust inference is about estimaing using OLS but adjusting the variance so as to get reliable t-tests, F-tests etc. in the presence of *heteroscedasticity of unknown form*

For example, consider simple linear regression

$y_i = \beta_0 + \beta_1 x_i + u_i$ $\quad Var(u_i \,|x_i) = {\sigma_i}^2$ (error variance depends on value of x)

OLS estimator is

$$\hat{\beta}_1 = \beta_1 + \left(\frac{1}{SST_x}\right)\sum_{i=1}^{i=n}(x_i - \bar{x})u_i \qquad Var(\hat{\beta}_1) = \left(\frac{1}{{SST_x}^2}\right)\sum_{i=1}^{i=n}((x_i - \bar{x})\sigma_i)^2$$

White (1980) showed that a valid estimator in presence of any form of heteroscedasticity is

$$\widehat{Var}(\hat{\beta}_1) = \left(\frac{1}{{SST_x}^2}\right)\sum_{i=1}^{i=n}((x_i - \bar{x})\hat{u}_i)^2$$

where $\hat{u}_i$ are the fitted residuals from OLS regression

Also $\widehat{Var}(\hat{\beta}_j) = \dfrac{\sum_{i=1}^{n} {\hat{r}_{ij}}^2 {\hat{u}_i}^2}{{SSR_j}^2}$

$\hat{r}_{ij}$ is the ith residual of regressing x_j on other regressors

SSR_j is the sum of squared residuals in this regression

$\sqrt{\widehat{Var}(\hat{\beta}_j)}$ is called the heteroscedasticity-robust standard error of OLS.

small sample variance in x_j , multicollinearity of x_j with other variables will bump up the robust standard error

robust t-tests can be performed for hypothesized value of OLS parameters by using the robust s.e.

Multiple Linear Regression

22(b) **Testing for Heteroscedasticity - Breusch-Pagan Test**

If we simply want to use the classical OLS estimates, we may want to check for heteroscedasticity first

We want to test

$H_0: Var(u|x_j) = \sigma^2$ or $H_0: E(u^2|x_j) = \sigma^2$

Assume a simple linear dependence

$$u^2 = \delta_0 + \delta_1 x_1 + \cdots + v$$

We want to test $H_0: \delta_1 = 0, \delta_2 = 0, \ldots$

Since we don't observe the error term u, we can use the fitted OLS residuals $\hat{u}$ as the dependent variable

This can be tested with F-statistic

$$F = \frac{(R^2)/k}{(1-R^2)/(n-k-1)}$$

Another test by the name of Breusch-Pagan test calculates LM statistic as

$$LM = nR^2 \quad \sim \quad {\chi_k}^2$$

Multiple Linear Regression

22(c) **Testing for Heteroscedasticity - White test**

White test for example in 2 variable case is based on estimation of testing correlation of squared error against all x_i , all ${x_i}^2$ and all cross product terms $x_i x_{ij}$

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 {x_1}^2 + \delta_4 {x_2}^2 + \delta_5 x_1 x_2 + \cdots + \nu$$

We want to test $H_0: \delta_1 = 0, \delta_2 = 0, \delta_3 = 0, \delta_4 = 0, \delta_5 = 0$

F-test can be performed as usual with the 5 restrictions

The above functional dependence can be captured simply as follows which includes all linear and quadratic terms

$$u^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \nu$$

22(d) **Weighted Least Square (WLS) method**

Before development of robust-statistics, WLS was used which is superior to OLS in the presence of Heteroscedasticity

Assume, $Var(u_i \,|\boldsymbol{x_i}) = \sigma^2 h(\boldsymbol{x_i})$

$\boldsymbol{x_i}$ is the vector of values of all regressors on ith observation

where $h(\boldsymbol{x})$ is a known function

Consider the variable $u_i / \sqrt{h_i}$ $\quad E\left(\frac{u_i}{\sqrt{h_i}} | x_i\right) = 0$ $\quad Var\left(\frac{u_i}{\sqrt{h_i}} | x_i\right) = \sigma^2$

Divide each observation by $\sqrt{h_i}$ and perform OLS

The OLS estimates thus obtained are a class of GLS estimates and are against the transformed variables

GLS estimator for correcting heteroscedasticity are called WLS estimator as it minimizes the *weighted sum of squared errors*

WLS Estimator : Minimize

$$\sum_{i=1}^{i=n} {\hat{u}_i}^2 = \sum_{i=1}^{i=n} \left(\hat{y}_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i} - \cdots - \hat{b}_k x_{ki}\right)^2 / h_i$$

Multiple Linear Regression

22(e) **Feasible GLS (fGLS)**

In practice, the function $h(x_i)$ is not known and hence need to be estimated

Using $\widehat{h_i}$ instead of h_i in the GLS transformation leads to an estimator called fGLS

We specify a flexible function whose parameters need to be calibrated from the sample

$$Var(u\,|\boldsymbol{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k)$$

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k)\nu \qquad \mathbb{E}[\nu]=1$$

$$\ln(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + e$$

Regress $g = \ln(\hat{u}^2)$ with $x_1, x_2, \ldots, x_k$

Get the fitted values of $\hat{g}$, use $\widehat{h_i} = \exp(\hat{g}_i)$

WLS is BLUE if we know h_i. Having to use $\widehat{h_i}$ makes the estimates biased, However asymptotically consistent

In order to perform joint hypotheses (F-test) , it's important to use the same weights

The OLS and WLS estimates may be different due to random sampling. However there is a formal test (Hausman test) to check if they are statistically different.

If they are different, it could be because of model mis-specification

Multiple Linear Regression

22(f) **prediction error in WLS**

standard error of the 'prediction error' is

$$s.e.(\hat{u}_p) = \hat{\sigma} h(x_p) \sqrt{1 + \boldsymbol{c}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{c}}$$

95% Prediction interval of estimated y is $\hat{y}_p \pm 1.96\, s.e.(\hat{u}_p)$

23 **Model Mis-Specification**

If a regressor is correlated with the error term, we say the variable is endogenous

If an omitted variable is a function of a regressor in the model, then model suffers from functional form mis-specification. In this case proxy variables can be used to mitigate omitted variable bias

Multiple Linear Regression

23(a) **Functional Form Mis-specification - nested and non-nested**

Examples of Model Mis-specification

$y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + u$ True Model

$y = \beta_0 + \beta_1 x_1 + v$ Mis-specified Model

or

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$ True Model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ Mis-specified Model

We can use Ramsey's (1969) regression specification error test (RESET). The idea is to specify additional non linear terms (unrestricted model) and perform an F-test between the unrestricted model vs restricted model (current model)

restricted Model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

unrestriced Model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \delta_1 \hat{y}_1{}^2 + \delta_1 \hat{y}_2{}^3 + v$

F - test follows as usual

The test has a drawback of being very general and not giving any direction on the nature of mis-specification

In case of non-nested models i.e. the set of regressors are disjoint in the current and alternate model, we can perform the following test

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ current Model

$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + u$ Alternate Model

One Approach is to use an unrestricted model

$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \ln(x_1) + \gamma_4 \ln(x_2) + u$ unrestricted Model

Perform $H_0: \gamma_1 = 0, \gamma_2 = 0$ to reject the current Model

Perform $H_0: \gamma_3 = 0, \gamma_4 = 0$ to reject the Alternate Model

This test does not necessarily lead to a clear winner. Both models could be rejectd, or neither model is rejected. In the latter case, we can look at adjusted R^2 to choose between the models

Multiple Linear Regression

23(b) **Functional Form Mis-specification - Proxy Variables**

More challenging is a scenario when the model excludes a key variable which cannot be measured

Consider a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$

where x_3^* is not observed and suppose we only care about unbiased and onsistent estimates for β_1 and β_2 . We find a proxy variable x_3 which is related to x_3^* as

$x_3^* = \delta_0 + \delta_3 x_3 + v_3$

we can regress y on x_1, x_2, x_3 and expect to get consistent estimates at least under the assumption

(1).. u is uncorrelated with x_1, x_2, x_3^*

(2).. v_3 is uncorrelated with x_1, x_2, x_3

combining the two, we get

$\mathbb{E}[x_3^* | x_1, x_2, x_3] = \mathbb{E}[x_3^* | x_3] = \delta_0 + \delta_3 x_3$

So using Proxy variable,

$y = (\beta_0 + \beta_3 \delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \delta_3 x_3 + u + \beta_3 v_3$

$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$

good estimates

Multiple Linear Regression

24 **Measurement Error**

Sometimes we are not able to accurately measure a variable because of practical issues and so we rely on an approximation. The measurement error can be present on the dependent as well as explanatory vvariables.

24(a) **Measurement error in dependent variable**

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

y^* is not accurately measured. Instead we are measuring y

e^0 is the measurement error $y - y^*$

Our model becomes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u + e^0$$

if the measurement error is uncorrelated with each explanatory vvariable, we get unbiased and consistent estimates. However, the error variance will be higher and hence OLS variance.

$$Var(u + e^0) = {\sigma_u}^2 + {\sigma_0}^2$$

24(b) **Measurement error in explanatory variable**

$$y = \beta_0 + \beta_1 {x_1}^* + u$$

e^1 is the measurement error $x_1 - {x_1}^*$

We assume $\mathbb{E}[e^1]=0$

u is uncorrelated with $x_1, {x_1}^*$ $\quad Cov(x_1, e^1) = 0$

We get $y = \beta_0 + \beta_1 x_1 + u - \beta_1 e^1$

The OLS estimates are unbiased and consistent

However, if we instead assume that

$Cov({x_1}^*, e^1) = 0$ Classical errors in variables (CEV)

Then x_1 and e^1 become correlated as $x_1 = {x_1}^* + e^1$

$Cov(x_1, e^1) = {\sigma_{e_1}}^2$ Then

$Cov(x_1, u - \beta_1 e^1) = -\beta_1 {\sigma_{e_1}}^2$

$$\text{plim}\ (\hat{\beta}_1) = \beta_1 + \frac{Cov(x_1, u - \beta_1 e^1)}{Var(x_1)} = \beta_1 \left(\frac{{\sigma_{x_1^*}}^2}{{\sigma_{x_1^*}}^2 + {\sigma_{e_1}}^2} \right)$$

attenuation bias in OLS due to CEV

$$\frac{Var({x_1}^*)}{Var(x_1)} < 1$$