Session 6

OLS / Linear Regression Topics

Anthony Tay

This Version: 22 Aug 2025

Session 6

- Summary of OLS and Linear Regression Theory
- When OLS Assumptions Do Not Hold
  - Consequences
  - How To Check
  - How To Rectify (Partial Discussion)

Quick Recap

Estimates population conditional expectation

- Assume linear-in-parameters:

$$E(Y \mid X_1, \dots, X_{K-1}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1}.$$

E.g.,

$$E(\ln earn \mid age, educ) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ + \beta_4 male + \beta_5 male \cdot educ$$

We can write the regression model as

$$Y = E(Y \mid X_1, \dots, X_{K-1}) + \epsilon = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \epsilon\,.$$

Quick Recap

As long as the conditional expectation is correctly specified, we have

$$E(\epsilon \mid X_1, \ldots, X_{K-1}) = 0\,.$$

which implies

- $E(\epsilon) = 0$, and
- $E(\epsilon X_k) = 0$ for $k = 0, \ldots, K-1$

That is, ϵ is uncorrelated with each of the regressors $X_1, \ldots, X_{K-1}$

We'll add

- $Var(\epsilon \mid X_1, \ldots, X_{K-1}) = \sigma^2$

Quick Recap

Assume

- iid sample $\{Y_i, X_{i1}, \dots, X_{i,K-1}\}_{i=1}^{n}$ from the population

Then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{K-1} X_{i,K-1} + \epsilon_i\,, i = 1, \dots, n$$

where

$$E(\epsilon_i \mid X_{11}, \dots, X_{n1}; \dots; X_{1,K-1}, \dots, X_{n,K-1}) = 0$$

$$E(\epsilon_i^2 \mid X_{11}, \dots, X_{n1}; \dots; X_{1,K-1}, \dots, X_{n,K-1}) = \sigma^2$$

$$E(\epsilon_i \epsilon_j \mid X_{11}, \dots, X_{n1}; \dots; X_{1,K-1}, \dots, X_{n,K-1}) = 0$$

for all $i, j = 1, \dots, n$, $i \neq j$

Quick Recap

Can write in matrix form as

$$y = X\beta + \varepsilon \,,\; E(\varepsilon \mid X) = 0 \,,\;\; E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$$

The OLS estimator for β is

$$\widehat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

This requires $X^{\mathrm{T}}X$ to be invertible, which will be the case if the columns of X are linearly independent, meaning that

$$Xc = 0_n \iff c = 0_K$$

No "perfect collinearity"

Quick Recap

OLS fitted values: $\hat{y}_{ols} = X\hat{\beta}^{ols}$

OLS residuals: $\hat{\varepsilon}_{ols} = y - \hat{y}_{ols}$

Results:

- the OLS estimator $\hat{\beta}$ is unbiased.
- Key assumption: $E(\epsilon_i \mid X_{11}, \dots, X_{n1}; \dots; X_{1,K-1}, \dots, X_{n,K-1}) = 0$
- If the homoskedasticity assumption holds, then

$$Var(\hat{\beta}_{ols} \mid X) = \sigma^2(X^{\mathrm{T}}X)^{-1}$$

Quick Recap

- The noise variance σ^2 can be estimated using: $\widehat{\sigma^2} = \dfrac{\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}}{n-K}$
- Estimate $Var(\hat{\beta}^{ols} \mid X)$ with: $\widehat{Var}(\hat{\beta}^{ols} \mid X) = \widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1}$
- If errors are homoskedastic, then OLS estimators are "Best Linear Unbiased"

That is, given any other linear unbiased estimator $\tilde{\beta}$, we have

$$Var(c^{\mathrm{T}}\hat{\beta}^{ols} \mid X) \leq Var(c^{\mathrm{T}}\tilde{\beta} \mid X)$$

for any $K \times 1$ vector $c \neq 0_K$

Quick Recap

- For goodness-of-fit, the R^2 statistic

$$R^2 = 1 - \frac{\widehat{\varepsilon}_{ols}^{\mathrm{T}}\widehat{\varepsilon}_{ols}}{y^{\mathrm{T}}M_0 y} = 1 - \frac{RSS}{TSS}$$

  measures the proportion of variation in Y_i is accounted for by the explanatory variables

- R^2 never decreases (and usually increases) as more regressors are included in the regression

- Adjusted-R^2

$$\text{adj.-}R^2 = 1 - \frac{RSS/(n-K)}{TSS/(n-1)} = 1 - \frac{n-1}{n-K}\frac{RSS}{TSS}.$$

  is sometimes reported.

Quick Recap

If noise terms are normally distributed

- For testing single hypothesis $H_0 : r^{\mathrm{T}}\beta = r_0$ vs $r^{\mathrm{T}}\beta \neq r_0$

$$t = \frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}\widehat{Var}(\hat{\beta}^{ols} \mid X) r}} \sim \mathrm{t}(n-K)$$

Quick Recap

- For multiple hypotheses $H_0 : \mathcal{R}\beta = r_0$ vs $H_A : \mathcal{R}\beta \neq r_0$

$$F = \frac{(\hat{\varepsilon}_{rls}^{\mathrm{T}}\hat{\varepsilon}_{rls} - \hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols})/J}{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}/(n-K)}$$

$$= \frac{(R_{ur}^2 - R_r^2)/J}{(1-R_{ur}^2)/(n-K)}$$

$$= (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}\,\widehat{Var}(\hat{\beta}^{ols} \mid X)\,\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0)/J \sim \mathsf{F}(J, n-K)$$

Otherwise, we can usually rely on large-sample approximate tests

$$t \overset{a}{\sim} \text{Normal}(0,1) \quad \text{and} \quad JF \overset{a}{\sim} \chi^2(J)\,.$$

Normality of Noise Terms

Non-normality of errors

- does not affect unbiasedness or consistency or efficiency of the OLS estimator
- Primary function to provide the finite sample distribution for the t- and F-statistics
- One way to test for normality, use fact that $X \sim \text{Normal}(\mu, \sigma^2)$ implies

$$S = E((X-\mu)^3)/\sigma^3 = 0 \quad \text{and} \quad Kur = E((X-\mu)^4)/\sigma^4 = 3$$

- Can estimate S and Kur with

$$\widehat{S} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{3/2}} \quad \text{and} \quad \widehat{Kur} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{2}}.$$

Normality of Noise Terms

Jarque-Bera statistic: $JB = \frac{n-K}{6}\left(\widehat{S}^2 + \frac{1}{4}(\widehat{Kur}-3)^2\right) \sim \chi^2(2)$

```
Skew <- function(x){
  return(mean((x-mean(x))^3)/(mean((x-mean(x))^2)^(3/2)))
}
Kurt <- function(x){
  return(mean((x-mean(x))^4)/(mean((x-mean(x))^2)^2))
}
JB <- function(mdl){
  # requires lm object, returns JB Stat, p-val, Skewness and Kurtosis Coef.
  N <- nobs(mdl)
  K <- summary(mdl)$df[1]
  ehat <- residuals(mdl)
  JBSkew <- Skew(ehat)
  JBKurt <- Kurt(ehat)
  JBstat <- ((N-K)/6*(JBSkew^2 + (1/4)*(JBKurt-3)^2))
  JBpval <- 1-pchisq(JBstat,2)
  return(list("JBstat"=JBstat, "JBpval"=JBpval, "Skewness"=JBSkew, "Kurtosis"=JBKurt))
}
```

Normality of Noise Terms

E.g., We test for normality of the residuals in the regression

$$\ln earn_i = \beta_0 + \beta_1 \ln wexp_i + \beta_2 \ln tenure_i + \epsilon_i$$

```
df_earn <-read_csv("data\\earnings2019.csv",show_col_types=FALSE)
mdl <- lm(log(earn)~log(wexp)+log(tenure), data=df_earn)
JBtest <- JB(mdl)
fmt <- function(x){format(round(x,4),nsmall=4)}
cat("JB:", fmt(JBtest$JBstat), "  p-val:", fmt(JBtest$JBpval),
    "  Skewness:", fmt(JBtest$Skewness), "  Kurtosis:", fmt(JBtest$Kurtosis),"\n")
```

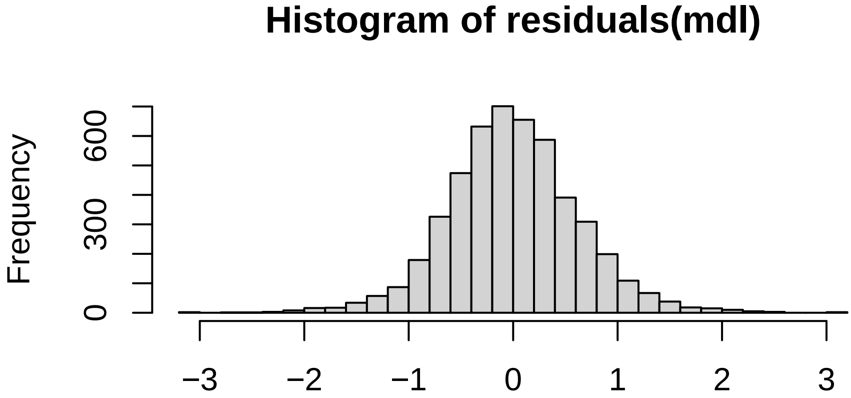
```
JB: 338.6182   p-val: 0.0000   Skewness: 0.0816   Kurtosis: 4.2718
```

The null of normality is rejected

- distribution symmetric, but there is ‘excess kurtosis’ (kurtosis in excess of 3)

Normality of Noise Terms

```
hist(residuals(mdl), 40)
```



Heteroskedasticity

Heteroskedasticity of the noise terms does not affect the unbiasedness or consistency of OLS estimators of the linear regression model

- Have calculate the standard errors properly, use use

$$\widehat{Var}_{HC0}(\hat{\beta}^{ols}) = \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}$$

  or one of its variants, instead of the usual $\widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1}$

- heteroskedasticity-robust variance-covariance matrix remains consistent for $Var(\hat{\beta}^{ols})$ even under homoskedasticity

Heteroskedasticity

- OLS estimators are efficient when the noise terms are homoskedastic, and may not be so otherwise

When there is heteroskedasticity, the question is:

- Can do better than OLS (in terms of getting more precise estimators)

Example

We show inefficiency of OLS under heteroskedasticity with a simple illustrative example

Suppose

$$Y_i = \beta_1 X_{i1} + \epsilon_i \, , \; i = 1, 2, ..., n \, .$$

We assume all of the usual OLS assumptions continue to hold, except that

$$E(\epsilon_i^2 \mid X_{11}, \dots, X_{n1}) = \sigma^2 X_{i1}^2 \quad \text{for all} \;\; i = 1, \dots, n$$

The OLS estimator for β_1 in this example is

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n} X_{i1} Y_i}{\sum_{i=1}^{n} X_{i1}^2}$$

which is unbiased for β_1 (see exercises)

Example

Under our assumptions, the variance of the OLS estimator is

$$Var(\hat{\beta}_1^{ols} \mid X_{11}, \dots, X_{n1})$$

$$= \frac{\sum_{i=1}^n X_{i1}^2 Var(\epsilon_i \mid X_{11}, \dots, X_{n1})}{\left(\sum_{i=1}^n X_{i1}^2\right)^2} = \frac{\sum_{i=1}^n X_{i1}^2 (\sigma^2 X_{i1}^2)}{\left(\sum_{i=1}^n X_{i1}^2\right)^2} = \frac{\sigma^2 \sum_{i=1}^n X_{i1}^4}{\left(\sum_{i=1}^n X_{i1}^2\right)^2}$$

We will show that OLS is inefficient, by presenting a more efficient linear unbiased estimator

Example

- weight each observation by $1/X_{i1}$ and run the regression

$$\frac{Y_i}{X_{i1}} = \beta_1 + \frac{\epsilon_i}{X_{i1}} = \beta_1 + \epsilon_i^* \,.$$

  That is, simply regress Y_i/X_{i1} on a constant

- Modified noise terms will continue to have zero conditional expectation

$$E(\epsilon_i/X_{i1} \mid X_{i1}, \dots, X_{n1}) = (1/X_{i1})E(\epsilon_i \mid X_{i1}, \dots, X_{n1}) = 0$$

- and uncorrelated (exercise)
- Furthermore, its conditional variance is now constant:

$$Var(\epsilon_i/X_{i1} \mid X_{i1}, \dots, X_{n1}) = (1/X_{i1}^2)\,Var(\epsilon_i \mid X_{i1}, \dots, X_{n1}) = \sigma^2 \,.$$

Example

OLS estimation applied to modified regression model gives the estimator

$$\hat{\beta}_1^{wls} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_{i1}}\,.$$

'wls' stands for "weighted least squares"

- $\hat{\beta}_1^{wls}$ is unbiased (exercise)
- $\hat{\beta}_1^{wls}$ is a linear estimator, since

$$\hat{\beta}_1^{wls} = \sum_{i=1}^{n}\left(\frac{1}{nX_{i1}}\right)Y_i\,,$$

Example

We show

$$Var(\widehat{\beta}_1^{wls} \mid X_{11}, \dots, X_{n1}) \leq Var(\widehat{\beta}_1^{ols} \mid X_{11}, \dots, X_{n1})$$

directly

- $\widehat{\beta}_1^{wls}$ is a sample mean of n observations of a random variable with variance σ^2, its variance is
$$Var(\widehat{\beta}_1^{wls} \mid X_{11}, \dots, X_{n1}) = \frac{\sigma^2}{n}\,.$$
- It can be shown (see exercises) that
$$\frac{\sum_{i=1}^{n} X_{i1}^4}{\left(\sum_{i=1}^{n} X_{i1}^2\right)^2} \geq \frac{1}{n}\,,$$

Example

Therefore

$$Var(\hat{\beta}_1^{wls} \mid X_{11}, \ldots, X_{n1}) = \frac{\sigma^2}{n} \leq \frac{\sigma^2 \sum_{i=1}^n X_{i1}^4}{\left(\sum_{i=1}^n X_{i1}^2\right)^2} = Var(\hat{\beta}_1^{ols} \mid X_{11}, \ldots, X_{n1})$$

OLS estimators are inefficient when there is conditional heteroskedasticity because

- OLS makes no use of the fact that observations with large noise variances are less informative about the population regression line than observations with smaller noise variances; Information ignored leads to inefficiency
- WLS approach, on the other hand, uses this information directly by assigning less weight to noisier observations

Weighted Least Squares

Suppose our linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i \,,\ i = 1, \ldots, n,$$

with $E(\epsilon_i \mid X) = 0\,,\ i = 1, \ldots, n$, $E(\epsilon_i \epsilon_j \mid X) = 0$, $i, j = 1, \ldots, n$ and $i \neq j$, and

$$E(\epsilon_i^2 \mid X) = \sigma_i^2 = \sigma^2 \eta_i \,,\ i = 1, \ldots, n$$

where η_i is a **completely known** function of the regressors

Weighted Least Squares

That is, in $\sigma^2\eta_i$, the only unknown parameter is σ^2. For instance,

$$\sigma_i^2 = \sigma^2|X_i| \quad \text{or} \quad \sigma_i^2 = \sigma^2 X_i^2 \quad \text{or } \sigma_i^2 = \sigma^2 \exp(X_i)\,.$$

The idea of weighted least squares: weight each observation so that the weighted noise terms are no longer heteroskedastic

That is, we modify the regression equation to

$$\frac{Y_i}{\sqrt{\eta_i}} = \beta_0 \frac{1}{\sqrt{\eta_i}} + \beta_1 \frac{X_{i1}}{\sqrt{\eta_i}} + \cdots + \beta_{K-1} \frac{X_{i,K-1}}{\sqrt{\eta_i}} + \frac{\epsilon_i}{\sqrt{\eta_i}}$$

which we can write as

$$Y_i^* = \beta_0 X_{0,i}^* + \beta_1 X_{i1}^* + \cdots + \beta_{K-1} X_{i,K-1}^* + \epsilon_i^*\,.$$

Weighted Least Squares

Since η_i is fixed conditional on the regressors, we have

$$E(\epsilon_i^* \mid X) = E\left(\left(\frac{\epsilon_i}{\sqrt{\eta_i}}\right) \middle| X\right) = \frac{1}{\sqrt{\eta_i}} E(\epsilon_i \mid X) = 0\ ;$$

$$E({\epsilon_i^*}^2 \mid X) = E\left(\left(\frac{\epsilon_i}{\sqrt{\eta_i}}\right)^2 \middle| X\right) = \frac{1}{\eta_i} E(\epsilon_i^2 \mid X) = \frac{1}{\eta_i}\sigma^2\eta_i = \sigma^2\ ;$$

$$E(\epsilon_i^*\epsilon_j^* \mid X) = E\left(\left(\frac{\epsilon_i\epsilon_j}{\sqrt{\eta_i}\sqrt{\eta_j}}\right) \middle| X\right) = \frac{1}{\sqrt{\eta_i}\sqrt{\eta_j}} E(\epsilon_i\epsilon_j \mid X) = 0\ .$$

OLS estimation of modified regression will produce BLU estimators of the coefficients since all the necessary conditions for OLS to be BLU are met

Weighted Least Squares

Applying OLS on the transformed regression equation is equivalent to choosing $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_{K-1}$ to minimize

$$
\begin{aligned}
&\sum_{i=1}^{n}(Y_i^* - \widehat{\beta}_0 X_{i0}^* - \widehat{\beta}_1 X_{i1}^* - \cdots - \widehat{\beta}_{K-1} X_{i,K-1}^*)^2 \\
&= \sum_{i=1}^{n}(Y_i/\sqrt{\eta_i} - \widehat{\beta}_0(1/\sqrt{\eta_i}) - \widehat{\beta}_1 X_{i1}/\sqrt{\eta_i} - \cdots - \widehat{\beta}_{K-1} X_{i,K-1}/\sqrt{\eta_i})^2 \\
&= \sum_{i=1}^{n}\frac{1}{\eta_i}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \cdots - \widehat{\beta}_{K-1} X_{i,K-1})^2 .
\end{aligned}
$$

That is, we are minimizing a sum of *weighted* squared residuals $\sum_{i=1}^{n}\omega_i\widehat{\epsilon}_i^2$ where the weights are $\omega_i = 1/\eta_i$

Weighted Least Squares

After obtaining $\hat{\beta}_0^{wls}, \ldots, \hat{\beta}_{K-1}^{wls}$, you should report your results as

$$\hat{Y} = \hat{\beta}_0^{wls} + \hat{\beta}_1^{wls} X_1 + \ldots \hat{\beta}_{K-1}^{wls} X_{K-1} \,,$$

WLS fitted values and residuals are computed as

$$\hat{Y}_i^{wls} = \hat{\beta}_0^{wls} + \hat{\beta}_1^{wls} X_{i1} + \cdots + \hat{\beta}_{K-1}^{wls} X_{i,K-1}$$

$$\hat{\epsilon}_{i,wls} = Y_i - \hat{Y}_i^{wls} = Y_i - \hat{\beta}_0^{wls} - \hat{\beta}_1^{wls} X_{i1} - \cdots - \hat{\beta}_{K-1}^{wls} X_{i,K-1}.$$

Weighted Least Squares

For assessing goodness-of-fit, we should use the WLS residuals to construct the R^2:

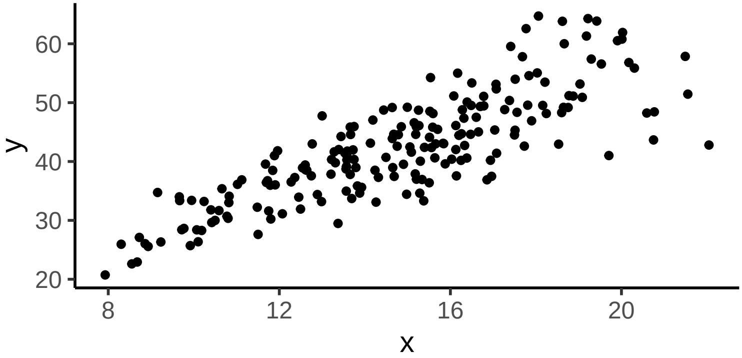
$$R^2_{wls} = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}^2_{i,wls}}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}.$$

- R^2_{wls} will generally be less than the R^2 from OLS estimation (why?)
- may even be negative

Example

Simple linear regression of y on x (with intercept) in the data set *heterosk.csv*.

```
df_het <- read_csv("data\\heterosk.csv",col_types = c("n","n","n"))
ggplot(data=df_het) + geom_point(aes(x=x,y=y), size=1) + theme_classic()
```



Example

OLS estimation of this regression gives the following output.

```
ols <- lm(y~x, data=df_het)
sum_ols <- summary(ols)
coef(sum_ols)
cat("R-squared: ",sum_ols$r.squared,"\n")
```

```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 6.285792  1.7817576  3.52786 5.207066e-04
x           2.430744  0.1177712 20.63955 3.034896e-51
R-squared:  0.6826877
```

Because there is obvious heteroskedasticity in this example, we should not trust the standard errors, t-statistics and p-values presented above

Example

If we wish to stick with OLS, then we have to calculate the heteroskedasticity-robust standard error

```
rbst_V <- sandwich::vcovHC(ols, type="HC0")
rbst_se <- sqrt(diag(rbst_V))
rbst_output <- coef(sum_ols)
colnames(rbst_output) <- c("Estimate", "rbst-se", "rbst-t", "p-val")
rbst_output[,'rbst-se'] <- rbst_se
rbst_output[,'rbst-t'] <- rbst_output[,'Estimate']/rbst_se
rbst_output[,'p-val'] <- 2*(1-pt(abs(rbst_output[,'rbst-t']),sum_ols$df[2]))
round(rbst_output,6)
```

```
            Estimate  rbst-se    rbst-t    p-val
(Intercept) 6.285792 1.863926  3.372339 0.000896
x           2.430744 0.136242 17.841435 0.000000
```

Example

Now we assume that $Var(\epsilon_i \mid X) = \sigma^2 X_i^2$, which seems reasonable assumption

We run WLS using the `weights` option in `lm()` function

`weights` option refers to weights on the squared residuals, i.e., $1/\eta_i$ if $\sigma_i^2 = \sigma^2 \eta_i$

```
df_het$wt <- 1/df_het$x^2
wls2 <- lm(y~x,data=df_het, weights=wt)
sum_wls2 <- summary(wls2)
coef(sum_wls2)
cat("R-squared: ", sum_wls2$r.squared,"\n")
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   4.82571  1.4065451  3.430896 7.321779e-04
x             2.53166  0.1023702 24.730430 1.839271e-62
R-squared:   0.755433
```

Example

- The standard errors are lower than in the OLS regression, which is not unexpected
- Note: R^2 reported above is not the R^2_{wls} that was recommended earlier
- We calculate R^2_{wls} below

```
ehat <- df_het$y - coef(wls2)[1] - coef(wls2)[2]*df_het$x
rss <- sum(ehat^2)
tss <- sum((df_het$y - mean(df_het$y))^2)
R2 <- 1 - rss/tss
cat("R-square: ", R2,"\n")
```

```
R-square:  0.6814966
```

R^2_{wls} is lower than in the OLS regression, as expected, but only slightly so

Example

The R^2 provided by `lm()` when using `weights` is a "weighted R-squared"

$$\text{weighted-}R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{Y}_i^{wls} - \overline{Y}_{wls})^2}{\sum_{i=1}^{n} w_i(Y_i^{wls} - \overline{Y}_{wls})^2}\,.$$

where $\overline{Y}_{wls}$ is the weighted mean of $\{Y_i\}_{i=1}^{n}$

In other words, it is the weighted fitted sum of squares divided by the weighted total sum of squares, centered on the weighted mean of $\{Y_i\}_{i=1}^{n}$

Example

We replicate the `lm()` weighted R-squared below:

```
wls0 <- lm(y~1, data=df_het, weights=wt)  # Regression on intercept only
tss_wtd <- sum(df_het$wt * (df_het$y - coef(wls0))^2)
fss_wtd <- sum(df_het$wt*(wls2$fitted.values - coef(wls0))^2)
WeightedR2 <- fss_wtd/tss_wtd
WeightedR2
```

```
[1] 0.755433
```

Example

One difficulty with WLS is that we generally do not know the form of the heteroskedasticity, especially in multiple regression case

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_{K-1,i} X_{K-1,i} + \epsilon_i \, , i = 1, \ldots, n \, .$$

We might have something like

$$\eta_i = \exp(\alpha_1 X_{i1} + \cdots + \alpha_{K-1} X_{i,K-1})$$

- the exponentiation is to ensure the variance is positive
- to implement WLS, we first have to estimate the parameters in the variance equation

Testing for Heteroskedasticity

The following are some possible tests for heteroskedasticity

- estimate the main regression by OLS and obtaining the OLS residuals $\hat{\epsilon}_{i,ols}$
- then run the regression

$$\hat{\epsilon}^2_{i,ols} = \alpha_0 + \alpha_1 X_{i1} + \ldots \alpha_{K-1} X_{i,K-1} + u_i$$

  and test $H_0 : \alpha_1 = \cdots = \alpha_{K-1} = 0$ using an F test
- An alternative is to use an "LM" test after running the regression above: under the null hypothesis, we have

$$nR^2_\epsilon \overset{a}{\sim} \chi^2(K).$$

  "Breusch-Pagan Test"

Testing for Heteroskedasticity

To allow for possible non-linear forms:

$$\hat{\epsilon}^2_{i,ols} = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_{K-1} X_{i,K-1} \\ + \delta_1 X^2_{i1} + \cdots + \delta_{K-1} X^2_{i,K-1} + \gamma_{12} X_{i1} X_{i2} + \cdots + u_i$$

Test if all of the coefficients (not including the intercept) are zero

Obviously you lose degrees of freedom quickly as the number of regressors grow

One way around this problem is to run the regression

$$\hat{\epsilon}^2_{i,ols} = \alpha_0 + \alpha_1 \hat{Y}_{i,ols} + \alpha_2 \hat{Y}^2_{i,ols} + u_i$$

Test the hypothesis $H_0 : \alpha_1 = \alpha_2 = 0$ using an F test or an LM test **"White Test"**

Testing for Heteroskedasticity

We apply the Breusch-Pagan test for heteroskedasticity to the regression

$$\ln earn = \beta_0 + \beta_1 \ln wexp_i + \beta_2 \ln tenure_i + \epsilon_i.$$

```
mdl <- lm(log(earn)~log(wexp)+log(tenure), data=df_earn)  #--Main Equation
cat("Main Regression\n")                                 #--Main Regr Output Title
round(summary(mdl)$coefficients,4)                       #--Main Regr Output
```

```
Main Regression
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   2.9167     0.0229 127.3595        0
log(wexp)    -0.0411     0.0095  -4.3294        0
log(tenure)   0.1746     0.0091  19.1215        0
```

Testing for Heteroskedasticity

```
df_earn$ehat <- residuals(mdl)                                 #--Get OLS Residuals
heteq <- lm((ehat^2)~log(wexp)+log(tenure), data=df_earn)     #--BP-Test Regression
cat("Heteroskedasticity Test Regression\n")                    #--Test Regr Output Title
round(summary(heteq)$coefficients,4)                           #--Test Regr Output
```

```
Heteroskedasticity Test Regression
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4668     0.0261 17.9070   0.0000
log(wexp)    -0.0224     0.0108 -2.0728   0.0382
log(tenure)  -0.0167     0.0104 -1.6028   0.1090
```

```
## BP, F-version
f_het <- summary(heteq)$fstatistic  #--Retrieve F-Stat (stat, df1, df2)
cat("BP-F Stat: ", f_het[1], "    p-val: ", 1-pf(f_het[1], f_het[2], f_het[3]), "\n")
```

```
BP-F Stat:  4.291278     p-val:  0.01373845
```

Testing for Heteroskedasticity

```
## BP, LM-version
lm_het <- nobs(heteq)*summary(heteq)$r.squared        # Calc. LM Stat.
lm_pval <- 1 - pchisq(lm_het, 2)  # 2 restrictions
cat("BP-LM Stat: ", lm_het, "     p-val: ", lm_pval, "\n", sep="")
```

```
BP-LM Stat: 8.57288     p-val: 0.0137538
```

- There is some evidence of heteroskedasticity
- individual t-tests suggest that the noise variance decreases with work experience

Misspecification

If you specify $E(Y \mid X_1) = \beta_0 + \beta_1 X_1$ but the conditional expectation does not have this form, then parameters become hard to interpret, if not meaningless, and predictions from the model will usually be biased.

Rough check for functional form adequacy

- plot residuals against individual regressors. Does residuals should look like random noise?

More formal tests:

- RESET
- Tests of non-nested alternatives

RESET test for functional form misspecification

Given a regression specification

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i \, ,$$

- Check if adding powers (X_{i1}^2, X_{i2}^2,...) and interaction terms ($X_{i1}X_{i2}$, $X_{i1}X_{i3}$, etc.) of the regressors can significantly improve the fit
- Problem: Severe loss of degrees-of-freedom in test regression

RESET test for functional form misspecification

Regression Equation Specification Error Test (RESET) test add $\hat{Y}^2_{i,ols}$, $\hat{Y}^2_{i,ols}$, ...into the regression specification and tests if significant

Test equation is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1} X_{i,K-1} + \alpha_2 \hat{Y}^2_{i,ols} + \cdots + \alpha_p \hat{Y}^p_{i,ols} + \epsilon_i \,,$$

and the hypothesis of adequacy of the functional form specification is

$$H_0 : \alpha_2 = \cdots = \alpha_p = 0.$$

RESET test for functional form misspecification

Using data in `earnings2019.csv`, we apply the RESET test to the regression

$$\ln earn_i = \beta_0 + \beta_1 wexp_i + \beta_2 tenure_i + \epsilon_i.$$

```
mdl_base <- lm(log(earn)~wexp+tenure, data=df_earn)
cat("Base Regression:\n")
round(summary(mdl_base)$coefficients, 4)
```

```
Base Regression:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   3.0249     0.0156 194.1339        0
wexp         -0.0049     0.0011  -4.3724        0
tenure        0.0187     0.0011  17.1674        0
```

RESET test for functional form misspecification

```
df_earn$yhat <- fitted(mdl_base)
mdl_test <- lm(log(earn)~wexp+tenure+I(yhat^2), data=df_earn)
cat("\nTest Regression:\n")
round(summary(mdl_test)$coefficients, 4)
```

```
Test Regression:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3285     2.3951  7.2350        0
wexp         -0.0531     0.0082 -6.5134        0
tenure        0.2092     0.0319  6.5528        0
I(yhat^2)    -1.5680     0.2625 -5.9722        0
```

The hypothesis of adequacy of the functional form in the base regression is rejected.

Testing Nonnested Alternatives

Regression specifications such as

$$[\text{A}] \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

$$\text{and} \quad [\text{B}] \quad Y_i = \beta_0 + \beta_1 \ln X_{i1} + \beta_2 \ln X_{i2} + \epsilon_i$$

are "non-nested alternatives", i.e., one is not a special case of the other.

Which is better?

Testing Nonnested Alternatives

One way: construct "super-model" that includes both [A] and [B] as restricted cases, i.e.,

$$\text{[A]} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 \ln X_{i1} + \beta_4 \ln X_{i2} + \epsilon_i$$

and to test for coefficient significance

- multicollinearity problems

Testing Nonnested Alternatives

Alternative approach:

- fit both models separately, collect their fitted values,
- include each fitted value series as a regressor in the other specification, i.e., regress

$$[\text{A'}] \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \delta_1 \widehat{Y}_i^B + \epsilon_i$$

$$\text{and} \quad [\text{B'}] \quad Y_i = \beta_0 + \beta_1 \ln X_{i1} + \beta_2 \ln X_{i2} + \delta_2 \widehat{Y}_i^A + \epsilon_i$$

and test (separately) if the coefficients on the fitted values are statistically significant

Testing Nonnested Alternatives

We compare the specifications

$$\begin{aligned} &[A] \quad \ln earn_i = \beta_0 + \beta_1 wexp_i + \beta_2 tenure_i + \epsilon_i \\ \text{and} \quad &[B] \quad \ln earn_i = \beta_0 + \beta_1 \ln wexp_i + \beta_2 \ln tenure_i + \epsilon_i \end{aligned}$$

```r
mdlA <- lm(log(earn)~wexp+tenure, data=df_earn)
mdlB <- lm(log(earn)~log(wexp)+log(tenure), data=df_earn)
df_earn$yhatA <- fitted(mdlA)
df_earn$yhatB <- fitted(mdlB)
cat("Model A plus yhatB:\n")
mdlAplusB <- lm(log(earn)~wexp+tenure+yhatB, data=df_earn)
round(summary(mdlAplusB)$coefficients,4)
```

```
Model A plus yhatB:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2825     0.3331  0.8482   0.3963
wexp         -0.0012     0.0012 -0.9975   0.3186
tenure        0.0022     0.0023  0.9532   0.3406
yhatB         0.9075     0.1101  8.2430   0.0000
```

Testing Nonnested Alternatives

```
cat("\nModel B plus yhatA:\n")
mdlBplusA <- lm(log(earn)~log(wexp)+log(tenure)+yhatA, data=df_earn)
round(summary(mdlBplusA)$coefficients,4)
```

```
Model B plus yhatA:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5557     0.3606  7.0876   0.0000
log(wexp)    -0.0373     0.0103 -3.6330   0.0003
log(tenure)   0.1577     0.0192  8.2204   0.0000
yhatA         0.1219     0.1215  1.0033   0.3158
```

It appears that the specification [B] is preferred over specification [A]

Omitted Variables

Basic problem: you estimate

$$E(Y \mid X_1, \dots, X_{K-1}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1}$$

when you really want to be estimating

$$\begin{aligned} &E(Y \mid X_1, \dots, X_{K-1}, X_K, \dots, X_M) \\ &\quad = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \beta_K X_K + \cdots + \beta_{K+M} X_{K+M} \,. \end{aligned}$$

Omitted Variables

The solution to the omitted variable problem is, of course, to include the missing variables

- this assumes that the missing variables are available
- What if not available?

$$\ln earn = \beta_0 + \beta_1 educ + (\text{controls}) + \epsilon$$

Controls may include race and sex variables, tenure, age, etc.

Some measure of innate ability? Hard to find

One solution is to use "instrumental variables" and "instrumental variable estimation" (Next chapter)

Sampling Issues: Truncated Sampling

Consider the simple linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

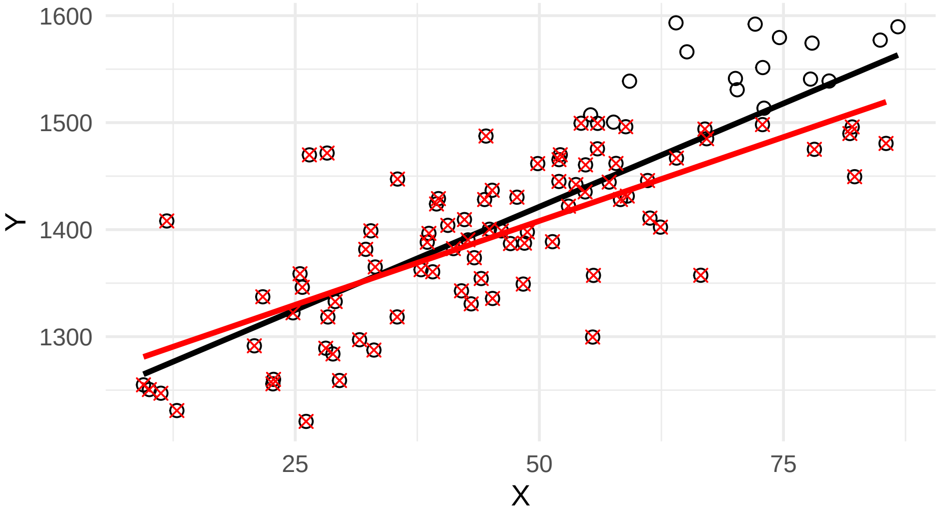
and suppose β_1 is positive

Suppose you have a "truncated sample" where you cannot observe any observation where $Y_i > c$

This causes correlation between X_{i1} and ϵ_i

- Missing observations tend to be those with large X_{i1} and large positive ϵ_i
- Non-missing observations with large X_{i1} associated with negative ϵ_i
- negative correlation between X_{i1} and ϵ_i

Sampling Issues: Truncated Sampling



Measurement Error

Suppose $Y = \beta_0 + \beta_1 X + \epsilon$

X is only observed with error, i.e., you observe $X^* = X + u$

Assume that the measurement error u is independent of X. Then

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon \\ &= \beta_0 + \beta_1 (X^* - u) + \epsilon \\ &= \beta_0 + \beta_1 X^* + (\epsilon - \beta_1 u) \\ &= \beta_0 + \beta_1 X^* + v \end{aligned}$$

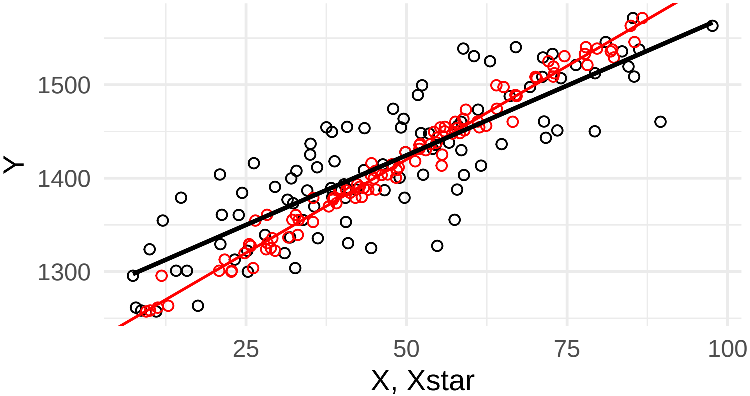
You proceed with only feasible option to you, which is to run the regression

$$Y = \beta_0 + \beta_1 X^* + v$$

Since u is correlated with X^*, the assumption $E(v \mid X^*) = 0$ does not hold

Measurement Error

Simulated example (red circles: no measurement error; black circles: measurement error in the X_i's)



Simultaneity Bias

Consider a demand-and-supply example

$$
\begin{aligned}
Q_t^d &= \delta_0 + \delta_1 P_t + \epsilon_t^d && (\text{Demand Eq } \delta_1 < 0) \\
Q_t^s &= \alpha_0 + \alpha_1 P_t + \epsilon_t^s && (\text{Supply Eq } \alpha_1 > 0) \\
Q_t^s &= Q_t^d && (\text{Market Clearing})
\end{aligned}
$$

Suppose you want to estimate market demand

Simultaneity Bias

Market clearing means that observed quantity and prices occur at the intersection of the demand and supply equations, i.e., observed prices are such that

$$\delta_0 + \delta_1 P_t + \epsilon_t^d = \alpha_0 + \alpha_1 P_t + \epsilon_t^s$$

which we can solve to get

$$P_t = \frac{\alpha_0 - \delta_0}{\delta_1 - \alpha_1} + \frac{\epsilon_t^s - \epsilon_t^d}{\delta_1 - \alpha_1}.$$

Substituting this expression for prices into either the demand or supply equation gives

$$Q_t = \left(\delta_0 + \delta_1 \frac{\alpha_0 - \delta_0}{\delta_1 - \alpha_1}\right) + \frac{\delta_1 \epsilon_t^s - \alpha_1 \epsilon_t^d}{\delta_1 - \alpha_1}.$$

Simultaneity Bias

This implies

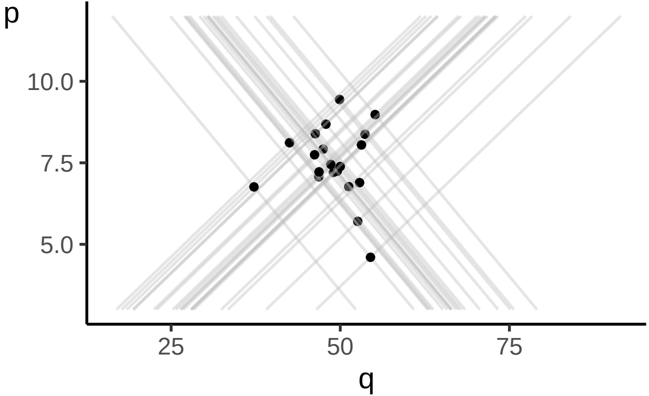
$$Var(P_t) = \frac{\sigma_s^2 + \sigma_d^2}{(\delta_1 - \alpha_1)^2} \quad \text{and} \quad Cov(P_t, Q_t) = \frac{\delta_1\sigma_s^2 + \alpha_1\sigma_d^2}{(\delta_1 - \alpha_1)^2}.$$

This means that in a regression of $Q_t = \beta_0 + \beta_1 P_t + \epsilon_t$, we will get

$$\hat{\beta}_1 \xrightarrow{p} \frac{Cov(Q_t, P_t)}{Var(P_t)} = \frac{\delta_1\sigma_s^2 + \alpha_1\sigma_d^2}{\sigma_s^2 + \sigma_d^2} \tag{1}$$

which is neither the price elasticity of demand nor the price elasticity of supply, but a linear combination of the two.

Simultaneity Bias



Data looks neither like demand or supply curve

Simultaneity Bias

The problem here:

- prices and quantities are simultaneously determined by the intersection of the demand and supply functions
- both prices and quantities are "endogenous" variables

(Next week) Instrumental variables may be able to help with this situation