# Class 5: Additional Regression Considerations

MFE 402

Dan Yavorsky

# Last Class

- Discussed asymptotics
  - Introduced convergence in probability – major use was LLN
  - Introduced convergence in distribution – major use was CLT
  - Put these together via Slutsky's Theorem to get the asymptotic distribution of $\hat{\beta}$
  - Effectively, $\hat{\beta} \overset{a}{\sim} N(\beta, \mathbf{V}_\beta)$ with $V_\beta = \mathbf{Q}_{XX}^{-1}\mathbb{E}[XX'e^2]\mathbf{Q}_{XX}^{-1}$
- Enables testing hypotheses and constructing confidence intervals for single parameters:
  - Use the test statistic $T(\beta_j^0) = (\hat{\beta}_j - \beta_j^0) \Big/ \sqrt{\left[\hat{\mathbf{V}}_{\hat{\beta}}\right]_{jj}}$
- If we assume errors are normally distributed
  - All homoskedastic results hold
  - $T(\beta_j^0)$ has an exact $t$-distribution instead of asymptotic Normal distribution
- Also discussed hypothesis testing for multiple coefficients
  - Single linear hypothesis $r'\beta = w$
  - Sets of linear hypotheses $R'\beta = w$ including the omnibus $F$-test

# Perspective on Model Specification

Once upon a time, econometricians tended to assume that the model provided by economic theory represented accurately the real-world mechanism generating the data, and viewed their role as one of providing "good" estimates for key parameters of that model.

That view of econometrics is obsolete. All econometric models are "false".

The goal then, is to construct models that (1) **correspond to facts** and (2) **are useful**:

- "All models are wrong, but some are useful" – *George Box*
- "A useful model is... parsimonious, plausible, and informative" – *Martin Feldstein*
- "Models are to be used, but not to be believed." – *Ed Leamer*
- "[Models] are simply rough guides to understanding." – *Danny Quah*

# Topics for Today

There is no single-best regression model. But there are many ways to make a model better or worse.

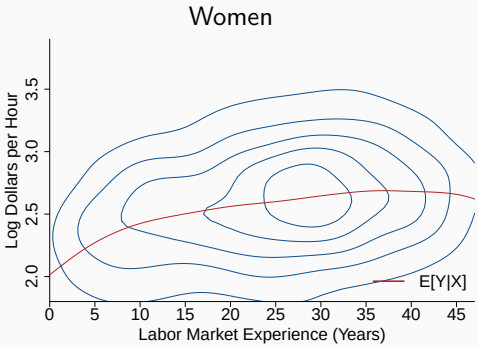
- Categorical $X$ Variables
- Log-Linear an Log-Log models
- Multicollinearity
- Errors in Variables
- Omitted Variables
- Leverage and Outliers
- Forecasting

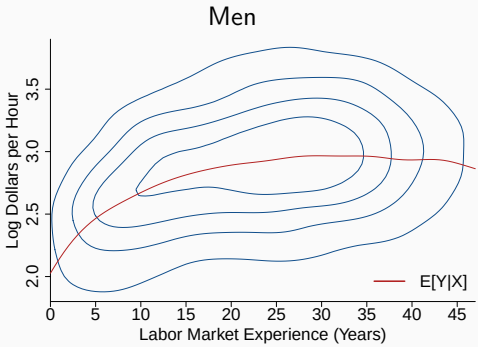
# Categorical Variables

# Categorical Variables

Categorical (or qualitative) variables take on a finite number of values:

- Binary variables: $X_j \in \{\text{yes}, \text{no}\}$ or $X_j \in \{\text{hit}, \text{miss}\}$
- Multi-category variables: $X_j \in \{\text{HS}, \text{BA}, \text{ADV}\}$ or $X_j \in \{\text{red}, \text{green}, \text{blue}\}$



# Dummy coding in R

A random variable is a mapping from an outcome to a value in $\mathbb{R}$.

Should not use $X_j = \{1, 2, 3\}$ for red/green/blue because

- $1 < 2 < 3$ is not meaningful
- The difference between blue-green (3-2) $\neq$ the difference between green-red (2-1)

Use a "sensible" mapping that facilitates interpretation:

- Binary variables: use 0/1
- Multi-category variables: use a *set* of binary ("dummy") indicator variables

In R, the `model.matrix()` function creates these binary variables for us when we run `lm()`:

```
x1 <- c(1.5, 2.5, 3.5)
x2 <- c("red", "green", "blue")
model.matrix( ~ x1 + x2)
```

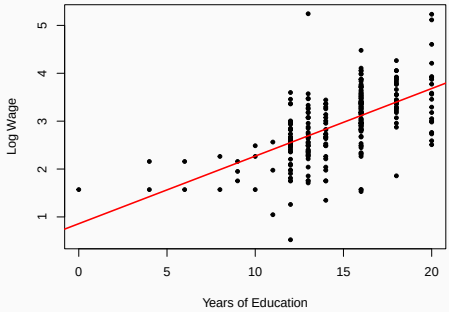
```
  (Intercept)  x1 x2green x2red
1           1 1.5       0     1
2           1 2.5       1     0
3           1 3.5       0     0
```

# Interpreting coefficients on dummy variables

Let's use a regression of log-wage on **years of education** as a running example:

```
out1 <- lm(lwage ~ edu, data=dat)
summary(out1)$coef |> round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8584     0.1763  4.8687        0
edu           0.1411     0.0118 12.0096        0
```

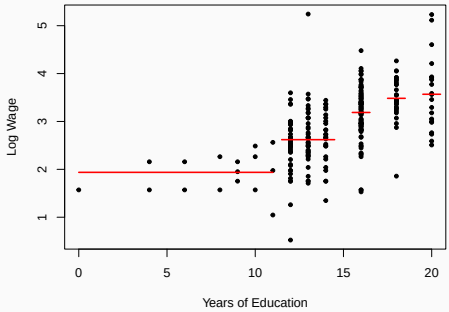


# Interpreting coefficients on dummy variables

Regress log-wage on a **dummy variable for each degree**:

```
out2 <- lm(lwage ~ degree, data=dat)
summary(out2)$coef |> round(4)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.9371 | 0.1469 | 13.1862 | 0 |
| degreeHS | 0.6825 | 0.1563 | 4.3654 | 0 |
| degreeBA | 1.2506 | 0.1588 | 7.8756 | 0 |
| degreeMA | 1.5458 | 0.1799 | 8.5919 | 0 |
| degreePhD | 1.6309 | 0.1943 | 8.3924 | 0 |




Coefficient interpretation depends on the baseline:

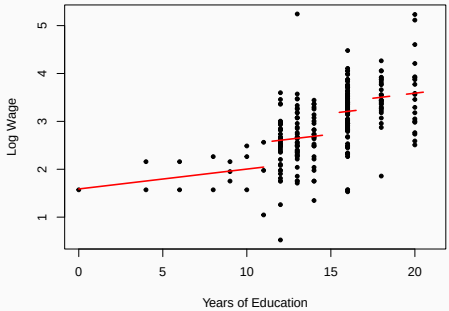
- $\mathbb{E}[\text{lwage} \mid \text{edu} < \text{HS}] = \beta_0$
- $\mathbb{E}[\text{lwage} \mid \text{edu} = \text{HS}] = \beta_0 + \beta_1$
- $\mathbb{E}[\text{lwage} \mid \text{edu} = \text{PhD}] = \beta_0 + \beta_4$

# Interpreting coefficients on dummy variables

Regress log-wage on years of education **and** a dummy variable for each degree:

```
out3 <- lm(lwage ~ edu + degree, data=dat)
summary(out3)$coef |> round(4)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.5883 | 0.4203 | 3.7788 | 0.0002 |
| edu | 0.0415 | 0.0469 | 0.8856 | 0.3767 |
| degreeHS | 0.4985 | 0.2601 | 1.9169 | 0.0563 |
| degreeBA | 0.9351 | 0.3901 | 2.3970 | 0.0172 |
| degreeMA | 1.1473 | 0.4847 | 2.3669 | 0.0187 |
| degreePhD | 1.1493 | 0.5775 | 1.9901 | 0.0476 |




The "slope" on years of education is the same, but the intercepts differ by degree.

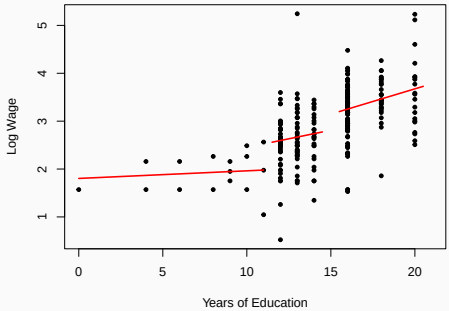
- $\mathbb{E}[\text{lwage} \mid \text{edu} = 0, \text{degree} = \text{MA}] = (\beta_0 + \beta_4) + \beta_1 \times \text{EDU}$

# Interpreting coefficients on dummy variables

Regress log-wage on years of education and degree (only 3 categories now) **and their interaction**:

```
dat$d3 <- cut(dat$edu,
              breaks=c(0,11,15,20),
              labels=c("<HS", "HS", "BA+"))
out4 <- lm(lwage ~ edu + d3 + edu:d3, data=dat)
summary(out4)$coef |> round(4)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.8036 | 0.5560 | 3.2436 | 0.0013 |
| edu | 0.0159 | 0.0638 | 0.2489 | 0.8037 |
| d3HS | -0.1028 | 1.0487 | -0.0981 | 0.9220 |
| d3BA+ | -0.2924 | 0.7908 | -0.3697 | 0.7119 |
| edu:d3HS | 0.0557 | 0.0941 | 0.5919 | 0.5544 |
| edu:d3BA+ | 0.0896 | 0.0718 | 1.2480 | 0.2132 |




Each category now has it's own intercept and slope:

- $\mathbb{E}[\text{lwage} \mid \text{degree} = \text{<HS}] = \beta_0 + \beta_1 \times \text{EDU}$
- $\mathbb{E}[\text{lwage} \mid \text{degree} = \text{HS}] = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) \times \text{EDU}$
- $\mathbb{E}[\text{lwage} \mid \text{degree} = \text{BA+}] = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) \times \text{EDU}$

# Log-Linear and Log-Log Models

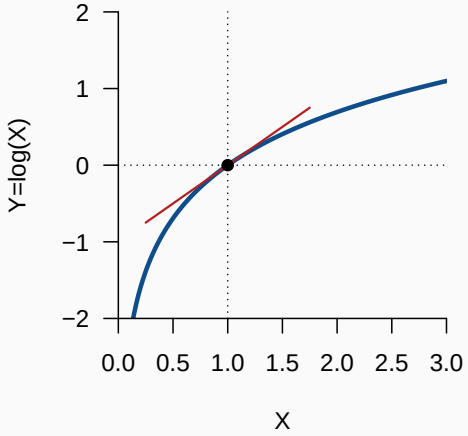
# Logs and Percentage Changes

Small changes in the natural log of a variable are directly interpretable as percentage changes, to a close approximation.

The reason is that the graph of $Y = \log(X)$ passes through the point (1,0) and at that point it has a slope of 1:

$$\frac{d}{dX}\log(X)\Big|_{X=1} = \frac{1}{X}\Big|_{X=1} = 1$$

Thus $\log(1 + r) \approx r$ when $r$ is small
Or equivalently $r \approx e^r$ - 1



# Log-Linear Regression

Suppose our model is $log(Y) = \beta_0 + \beta_1 X + e$ with any set of standard assumptions on $e$.

- **Basic interpretation**: a one-unit increase in $X$ is associated with a $\beta_1$ **unit** increase in $log(Y)$
- **Iintuitive interpretation**: a one-unit increase in $X$ is associated with a $\beta_1$ **percent** increase in $Y$.

First, find the expected values of $Y$ before and after a one-unit increase in $X$:

$$\log(Y_1) = \beta_0 + \beta_1 X \quad \Rightarrow \quad Y_1 = e^{\beta_0 + \beta_1 X}$$
$$\log(Y_2) = \beta_0 + \beta_1 (X+1) \quad \Rightarrow \quad Y_2 = e^{\beta_0 + \beta_1 (X+1)} = e^{\beta_0 + \beta_1 X + \beta_1}$$

The expected change in $Y$ for a one-unit change in X is:

$$\log(Y_2) - \log(Y_1) = \log\left(\frac{Y_2}{Y_1}\right) = \beta_1 \quad \text{because} \quad \frac{Y_2}{Y_1} = \frac{e^{\beta_0 + \beta_1 X + \beta_1}}{e^{\beta_0 + \beta_1 X}} = e^{\beta_1}$$

So $\%\Delta Y = Y_2/Y_1 - 1 = e^{\beta_1} - 1$ and per last slide, $e^{\beta_1} - 1 \approx \beta_1$ for small $\beta_1$.

# Log-Linear Regression Simulation

Let's illustrate with a simulated numerical example

```
# specify parameters
N <- 5000
b0 <- 2
b1 <- .08

# generate some data
set.seed(1234)
x <- runif(N, 5, 10)
y <- exp(b0 + b1*x + rnorm(N))

# fit regression
out <- lm(log(y) ~ x)
```

```
# check approximation
actual <- exp(coef(out)[2]) - 1
approx <- coef(out)[2]
names(approx) <- names(actual) <- NULL
print(c(actual=actual,
        approx=approx), digits=3)
```

```
actual approx
0.1005 0.0958
```

# Linear-Log Regression

What if we log-transform the $X$ variable instead of the $Y$ variable?

Our model is now: $Y = \alpha + \beta \log(X) + e$ with any set of standard assumptions on $e$.

Suppose we change $X$ by $p$ percent (for small $p$) and calculate the expected values of $Y$ before and after that change:

$$Y_1 = \beta_0 + \beta_1 \log(X)$$
$$Y_2 = \beta_0 + \beta_1 \log(X(1+p))$$

Now when we take the difference, we find:

$$Y_2 \text{ - } Y_1 = (\beta_0 \text{ - } \beta_0) + \beta_1 [\log(X(1+p)) \text{ - } \log(X)] = \beta_1 \log\left(\frac{X(1+p)}{X}\right) = \beta_1 \log(1+p)$$

Recall, $\log(1+p) \approx p$ thus $\beta_1 \log(1+p) \approx \beta_1 p$.

**Intuitive interpretation** of $\beta_1$: a $p$ percent increase in $X$ is associated with a $\beta_1 p$ unit change in $Y$.

# Linear-Log Regression Simulation

Let's illustrate with another simulated numerical example

```r
# specify parameters
N <- 5000
b0 <- 2
b1 <- 4

# generate some data
set.seed(1234)
x <- runif(N, 5, 10)
y <- b0 + b1*log(x) + rnorm(N)

# fit regression
out <- lm(y ~ log(x))
```

```r
# check approximation
pred <- predict(out, new=data.frame(x=c(2, 2*1.01)))
actual <- pred[2]-pred[1]
approx <- coef(out)[2]*0.01
names(approx) <- names(actual) <- NULL
print(c(actual=actual, approx=approx), digits=3)
```

```
actual approx
0.0409 0.0411
```

# Log-Log Regression

What if we log transform both $X$ and $Y$?

Consider $\log(Y) = \beta_0 + \beta_1 \log(X) + e$ with any set of standard assumptions on $e$.

Derive the interpretation of the slope coefficient by taking a first derivative:

$$\frac{dY}{dX} = \frac{\beta_1}{X} e^{\beta_0 + \beta_1 \log(X) + \varepsilon} = \frac{\beta_1 Y}{X} \qquad \Rightarrow \qquad \beta_1 = \frac{dY}{dX}\frac{X}{Y} = \frac{\partial Y / Y}{\partial X / X} = \frac{\%\Delta Y}{\%\Delta X}$$

This is the definition of an elasticity

If $X$ is our price, $Y$ is our good's demand (i.e., quantity sales), and $\beta_1 =$ -0.6, then a 1% increase in the price of our good would lead to a 0.6% decrease in demand for it

# Log-Log Regression Simulation

Let's illustrate with a final simulated numerical example

```
# specify parameters
N <- 5000
b0 <- 2
b1 <- 4

# generate some data
set.seed(1234)
x <- runif(N, 5, 10)
y <- exp(b0 + b1*log(x) + rnorm(N))

# fit regression
out <- lm(log(y) ~ log(x))
```

```
# check approximation
pred <- predict(out, new=data.frame(x=c(2, 2*1.01)))
actual <- exp(pred[2]) / exp(pred[1]) - 1
approx <- coef(out)[2] / 100
names(approx) <- names(actual) <- NULL
print(c(actual=actual, approx=approx), digits=3)
```

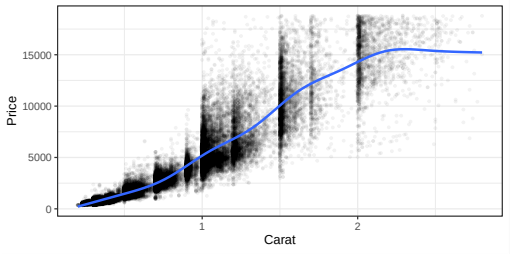
```
actual approx
0.0417 0.0411
```

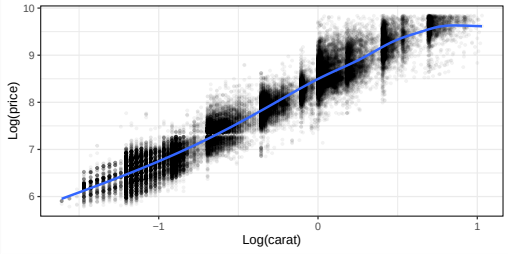
# Additional Benefit of Logging

Two added benefits of log-transformations:

1. they can sometimes transform non-linear relationships into linear ones
2. they can sometimes turn heteroskedasticity into homoskedasticity

The `diamonds` dataset in the `ggplot2` package offers a good example of both simultaneously!



# Multicollinearity

# Perfect/Strict Multicollinearity

**Perfect Multicollinearity**: If $\mathbf{X}'\mathbf{X}$ is singular then $(\mathbf{X}'\mathbf{X})^{-1}$ is not invertible and $\hat{\beta}$ is not defined

This typically arises when sets of regressors are included which are identically linearly related.

For example:

- including the same regressor twice
- including linear combinations of regressors
  - (eg, education, experience, and age when experience = age - education - 6)
- including a dummy variable and its square
- estimating a regression on a subsample where a dummary variable is either all 1's or all 0's
- including a dummy variable interaction which yields all zeros
- including more regressors than observations

Perfect multicollinearity leads to multiple solutions for $\hat{\beta}$ and therefore should be addressed.

# (Near) Multicollinearity

Multicollinearity:

- It is about **lack of variable-specific information**
- It is not a "disease" or a violation of model assumptions

In certain situations, the dependence among the $X$ variables can be so strong that it may be difficult to estimate the regression coefficients because values in $(\mathbf{X}'\mathbf{X})^{-1}$ are too large/small for a computer to accurately compute

Suppose the model and data are such that:

$$Y = \beta_1 X_1 + \beta_2 X_2 + e \quad \text{and} \quad \frac{1}{n}\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Then

$$\text{Var}(\hat{\beta}|\mathbf{X}) = \frac{\sigma^2}{n(1 - \rho^2)} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

Thus high correlation $\rho$ between regressors yields the same issue as small $n$: **imprecision**

# (Near) Multicollinearity (cont.)

How does it present?

- Large standard errors (thus small $t$-stats), but large $F$-stat

How to (possibly) detect it?

- Assess correlations of $X$ variables
- Regress $X_j$ on other $X$'s and look for high $R_j^2$

**Warning**: Do not remove correlated $X$'s *only* because they're correlated (the "sample size effect" may outweigh the "correlation effect")

What to do about it?

- "Re-frame" your interpretation of $t$-stats
- Combine highly correlated $X$'s into an index, if doing so makes senses
- Possibly remove some $X$'s from the regression

# Errors in Variables

# Measurement Error in $Y$ Variable

Suppose there is mean-zero, independent, homoskedastic, additive error $u$ in the measurement of the dependent variable: $Y = Y^* + u$ (where $Y$ is observed and $Y^*$ is measured without error)

Then the feasible linear regression model is equal to the desired model with extra error:

$$Y^* = X'\beta + e$$
$$Y = X'\beta + (e + u)$$

Implications:

- $\beta$ can still be estimated consistently by least squares
- There is less precision because $\sigma_u^2 + \sigma_e^2 > \sigma_e^2$
- Thus, relative to the no-measurement-error case $V_{\hat{\beta}}$ increases and $R^2$ decreases

If the error is not additive, if it is correlated with $X$, or if it is heteroskedastic, then nothing in general can be said about the consequences of measurement error.

# Measurement Error in $X$ Variables

Suppose there is mean-zero, independent, homoskedastic, additive error $u$ in the measurement of the independent variable: $X = X^* + u$ ($X$ is observed and $X^*$ is measured without error)

Then, taking the scalar-$X$ case as a simple example

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

Define the combined error $\nu = e - \beta_1 u$

The error $\nu$ is related to $X^*$ through $u$ and $X$ is related to $X^*$ so

- $\mathbb{E}[X^* \nu] \neq 0$, violating a key assumption of our linear CEF model
- $\hat{\beta}$ from estimating the regression of $Y$ on $X$ is inconsistent; all we know is downward bias

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\mathbb{E}[X\nu]}{\text{Var}(X)} = \beta_1 \left( 1 - \frac{\sigma_u^2}{\sigma_{X^*}^2 + \sigma_u^2} \right) < \beta_1$$

# Omitted Variables

# Omitted Variables

Let the true regression be $Y = X'\beta + e$.

Suppose we partition the $X$ vector into observable variables $X_1$ and unobservable variables $X_2$, so that $Y = X_1'\beta_1 + X_2'\beta_2 + e$.

Consider a projection (ie, regression) on $X_1$ only:

$$Y = X_1'\gamma_1 + u \qquad \text{with} \qquad \mathbb{E}[X_1 u] = 0$$

Then $\gamma_1 \neq \beta_1$ except in special cases ($\beta_2 = 0$ and/or $\text{cor}(X_1, X_2) = 0$):

$$\begin{aligned}
\gamma_1 &= (\mathbb{E}[X_1 X_1'])^{-1}\mathbb{E}[X_1 Y] \\
&= (\mathbb{E}[X_1 X_1'])^{-1}\mathbb{E}[X_1(X_1'\beta_1 + X_2'\beta_2 + e)] \\
&= \beta_1 + (\mathbb{E}[X_1 X_1'])^{-1}\mathbb{E}[X_1 X_2']\beta_2 \\
&= \beta_1 + \Gamma_{12}\beta_2
\end{aligned}$$

where $\Gamma_{12}$ is the coefficient matrix from a projection (ie, regression) of $X_2$ on $X_1$.

# Example: Reasoning about Omitted Variables

Suppose $Y$ is log wages, $X_1$ is education, and $X_2$ is intellectual ability.

We might reason that, on average:

- Highly able individuals attain more education ($\Gamma_{12} > 0$)
- Conditional on education, individuals with higher intelligence earn higher wages ($\beta_2 > 0$)

Then we can conclude that:

$$\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 > \beta_1$$

ie, a regression of log wages on education will overestimate the true return to education

# Endogenous Variables

We say the linear regression model $Y = X'\beta + e$ suffers from endogeneity in $\mathbb{E}[Xe] \neq 0$

- Measurement error in the $X$'s or relevant omitted variables are special cases

With endogeneity, the linear projection coefficient $\beta^*$ does not equal the structural parameter of interest $\beta$:

$$\begin{aligned}\beta^* &= (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY] \\ &= (\mathbb{E}[XX'])^{-1}\mathbb{E}[X(X'\beta + e)] \\ &= \beta + (\mathbb{E}[XX'])^{-1}\mathbb{E}[Xe]\end{aligned}$$

And it's estimator is not consistent for $\beta$:

$$\hat{\beta}_{OLS} \xrightarrow{p} (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY] = \beta^* \neq \beta$$

Possible solutions: Collect $X$ without error, collect omitted $X$, or use an Instrumental Variables technique (not covered)

# Leverage and Outliers

# Intuition about Unusual Observations

**Q:** What do we mean by "outlier" (ie, an outlying observation in our sample)?

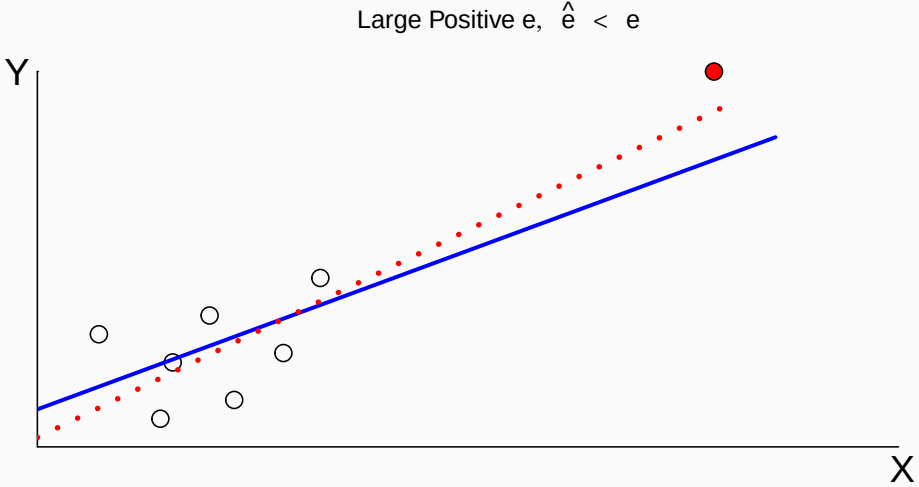
**A:** An observation that deviates markedly from the rest of the sample, due to...

- **Large residual**: the distance between the actual data point $(Y_n|X_n)$ and its fitted value $(\hat{Y}_T)$ is much larger for the outlier than for other observations
- **High leverage**: the data point has an unusual combination of values for the explanatory variable values (ie, it's in a "remote" part of the $X$-space)

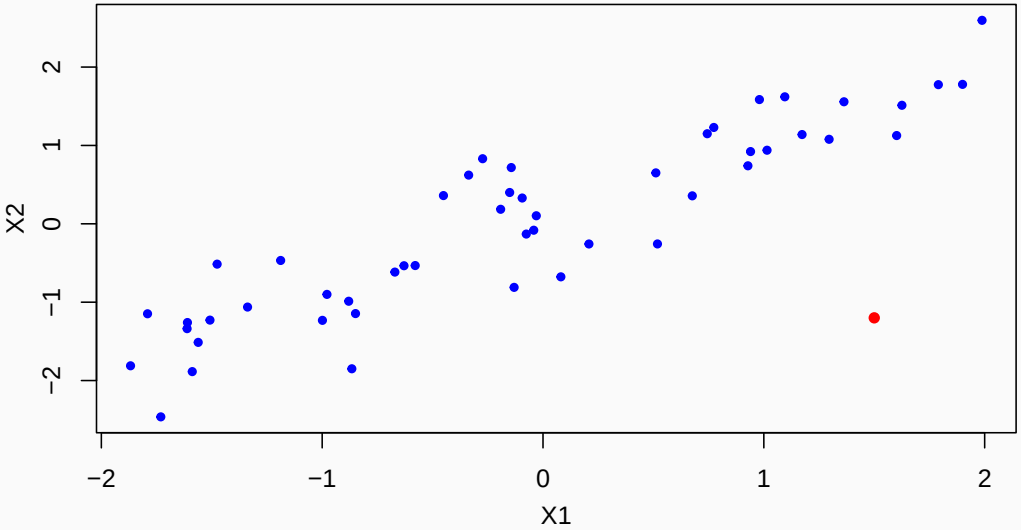
Why do we care?

- Observations with high leverage and large residuals are *influential*: if you dropped that observation from the data, coefficient estimates would change markedly
- May suggest something is wrong with the model specification
- Understand if predicted values are driven more by data or modeling assumptions

# Visual Example: Influential Observation



# Visual Example: “Hidden” Leverage



# Leverage Values

The leverage value of observation $i$ is the $i^{\text{th}}$ diagonal element of the "hat" matrix $P = X(X'X)^{-1}X'$:

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

- $h_{ii}$ is a standardized distance metric in $\mathbb{R}^k$
- $h_{ii}$ is bounded: $0 < h_{ii} < 1$
- $\sum_{i=1}^{n} h_{ii} = k$ and thus $\bar{h}_{ii} = k/n$

Rule of thumb:

- flag the observation if $h_{ii} > 3k/n$
- high leverage values have the *potential* to influence coefficient estimates

# The Effect on the Coefficients and Fitted Values (part 1)

Let's assess the effect on the coefficient estimates when we leave out observation $i$:

$$\begin{aligned}\hat{\beta}_{(-i)} &= (\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)} \\ &= \left(\sum_{j\neq i} X_j X_j'\right)^{-1}\left(\sum_{j\neq i} X_j Y_j\right) \\ &= (\mathbf{X}'\mathbf{X} - X_i X_i')^{-1}(\mathbf{X}'\mathbf{y} - X_i Y_i)\end{aligned}$$

Multiply both sides by $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X} - X_i X_i')$ to get

$$\hat{\beta}_{(-i)} - (\mathbf{X}'\mathbf{X})^{-1}X_i X_i'\hat{\beta}_{(-i)} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - X_i Y_i) = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}X_i Y_i$$

Rearrange to find:

$$\hat{\beta} - \hat{\beta}_{(-i)} = (\mathbf{X}'\mathbf{X})^{-1}X_i(Y_i - X_i\hat{\beta}_{(-i)}) = (\mathbf{X}'\mathbf{X})^{-1}X_i\tilde{e}_i$$

# The Effect on the Coefficients and Fitted Values (part 2)

Pre-multiply by $X_i'$ and subtract from $Y_i$ to find:

$$\hat{e}_i = Y_i - X_i'\hat{\beta} = Y_i - X_i'\hat{\beta}_{(-i)} - X_i'(\mathbf{X}'\mathbf{X})^{-1}X_i\tilde{e}_i = (1 - h_{ii})\tilde{e}_i$$
$$\Longrightarrow \hat{\beta} - \hat{\beta}_{(-i)} = (\mathbf{X}'\mathbf{X})^{-1}X_i(1 - h_{ii})^{-1}\hat{e}_i$$

Alternatively:

$$\hat{Y}_i - \tilde{Y}_i = X_i'\hat{\beta} - X_i'\hat{\beta}_{(-i)} = X_i'(\mathbf{X}'\mathbf{X})^{-1}X_i\tilde{e}_i = h_{ii}(1 - h_{ii})^{-1}\hat{e}_i$$

Thus both differences (in the coefficient estimates and in the fitted values) are functions of the observation's leverage and residual.

# Some Intuition

Because OLS minimizes squared errors, observations with high leverage can "pull" the regression line toward the observation in order to minimize the error. Thus, observations with high leverage ($h_{ii}$) will have smaller residuals ($\hat{e}_i$) as a result of this influence.

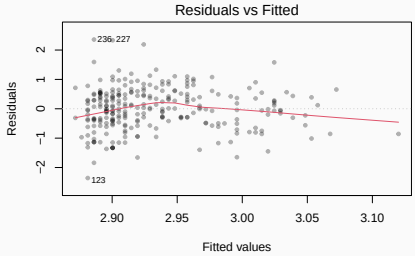
Under homoskedasticity:

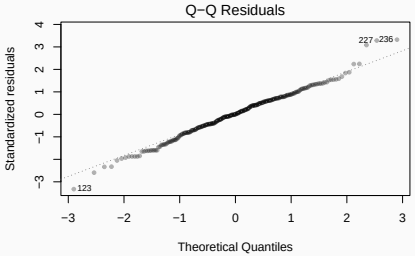
$$\begin{aligned} \text{var}(\hat{e}) = \text{var}(Me) = \mathbb{E}\left[Mee'M'\right] = M\mathbb{E}[ee']M = \sigma^2 M \\ \implies \text{var}(\hat{e}_i) = \sigma^2\left(1 - X_i(\mathbf{X}'\mathbf{X})^{-1}X_i\right) = \sigma^2(1 - h_{ii}) \end{aligned}$$

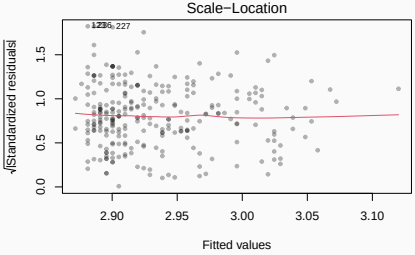
So when looking for unusual observations, you should look for those with high leverage and high *standardized* residuals, defined as:

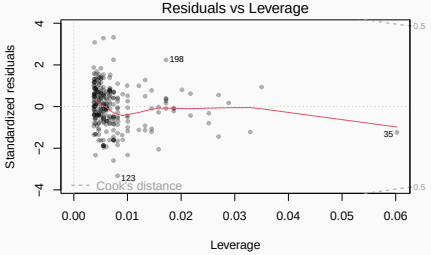
$$\hat{r}_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}} \approx \frac{e_i}{\sigma} \sim N(0, 1)$$

# Regression Diagnostic Plots: BEH Income-Wage Example

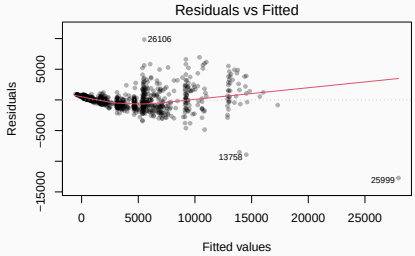


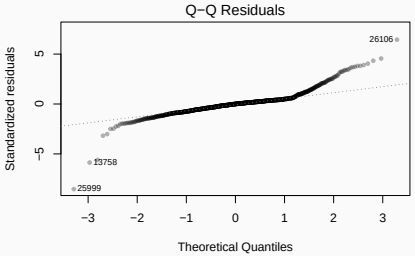



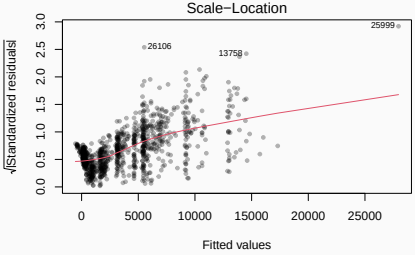



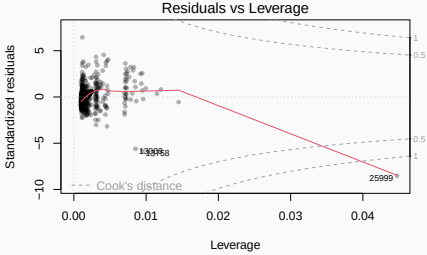



# Regression Diagnostic Plots: Diamonds Data Subset (No Logs)



# Regression Diagnostic Plots

**Residuals vs Fitted**

- Want to see no systematic relationships (ie, want a horizontal red line)
- U or inverted-U pattern suggests nonlinearities
- Systematic changes to vertical spread (eg, fan pattern) suggest heteroskedasticity

**Q-Q Residuals**

- Compares the empirical CDF of the residuals to a Normal CDF
- Marked deviations from the 45-degree line suggest non-normality of residuals

**Scale-Location**

- Want to see a horizontal line with constant spread; if not, suggests heterskedasticity

**Residuals vs Leverage**

- Values outside of Cook's distance lines are influential

# What to do with Influential Observations

If the influence is a result of wrong data:

- correct the data
- omit the observations

If the influence is a result of an unusual circumstance:

- likely modify the model to incorporate the observation

Otherwise, it's not obvious:

- Reconsider your choice of model
- Reconsider the specification (the $X$'s and non-linear transformations of them)
- Reconsider your estimation strategy (use a robust estimator instead of OLS)

# Forecasting

# Prediction and Prediction Error

Consider an out-of-sample realization $(Y_{n+1}, X_{n+1})$ where $X_{n+1}$ is observed; $Y_{n+1}$ is not.

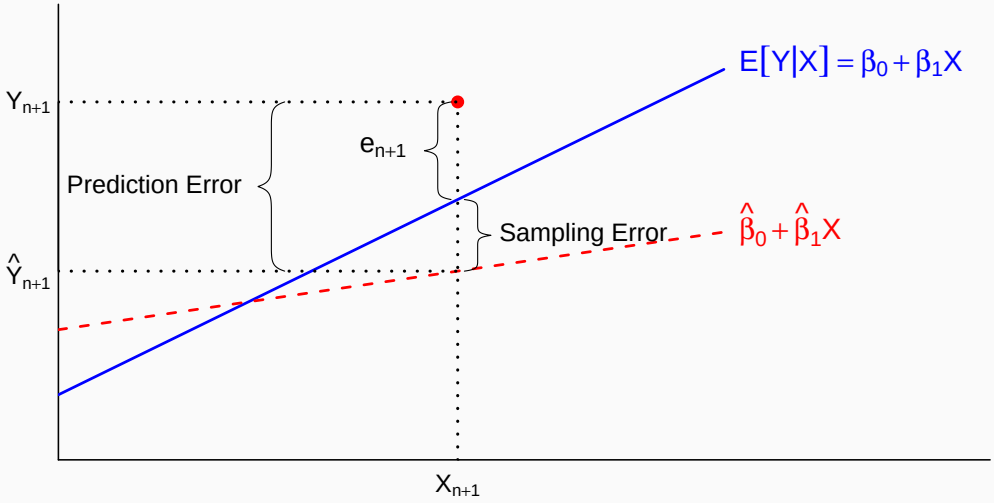
Define the predicted value:

$$\tilde{Y}_{n+1} = X'_{n+1}\hat{\beta}$$

We can decompose the prediction error:

$$\begin{aligned}
\tilde{e}_{n+1} &= Y_{n+1} - \hat{Y}_{n+1} \\
&= (X'_{n+1}\beta + e_{n+1}) - (X'_{n+1}\hat{\beta}) \\
&= e_{n+1} + X'_{n+1}(\beta - \hat{\beta}) \\
&= \underbrace{e_{n+1}}_{\substack{\text{Inherent} \\ \text{Randomness}}} + \underbrace{\left[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_{1,n+1} + \ldots + (\beta_k - \hat{\beta}_k)X_{k,n+1}\right]}_{\text{Sampling Error}}
\end{aligned}$$

# Prediction Error Visualization



# MSFE: Mean Squared Forecast Error

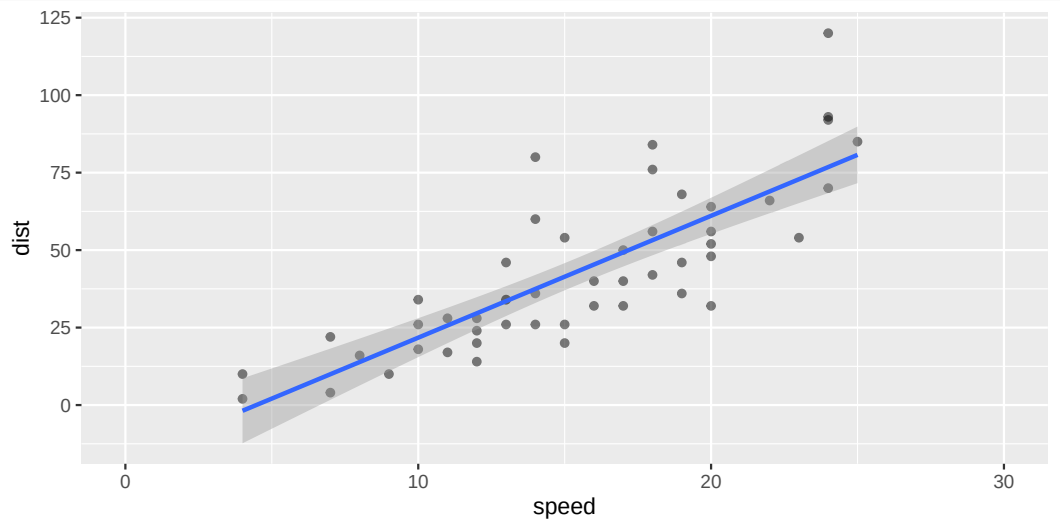
Define the mean-squared forecast error MSFE $= \mathbb{E}[\tilde{e}_{n+1}^2]$. Then:

$$\begin{aligned}\mathbb{E}[\tilde{e}_{n+1}^2] &= \mathbb{E}[e_{n+1} - X_{n+1}'(\hat{\beta} - \beta)] \\ &= \mathbb{E}[e_{n+1}^2] - 2\mathbb{E}[e_{n+1}X_{n+1}'(\hat{\beta} - \beta)] + \mathbb{E}[X_{n+1}'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'X_{n+1}] \\ &= \sigma^2 - 0 + \mathbb{E}\left[X_{n+1}' V_{\hat{\beta}} X_{n+1}\right]\end{aligned}$$

Notice that;

- MSFE $> \sigma^2$
- MSFE depends on $X_{n+1}$
- In particular, MSFE gets larger as $X_{n+1}$ is further from $\bar{X}$

# MSFE: Example from the Cars Dataset



# Estimating MSFE & Prediction Intervals

Under homoskedasticity (i.e., $\text{var}(e_i|X_i) = \sigma^2$ for all $i$), this simplifies to:

$$\text{MSFE} = \sigma^2 \left(1 + \mathbb{E}\left[X'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}X_{n+1}\right]\right)$$

Replace $\sigma^2$ with $\hat{\sigma}^2$ to get an estimator $\hat{\text{MSFE}}$ of the MSFE:

$$\hat{\text{MSFE}} = \hat{\sigma}^2 \left(1 + X'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}X_{n+1}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{e}_i}{1 - h_{ii}}\right)^2$$

Assuming the errors are normally distributed (i.e., $e \sim N(0, \sigma^2)$), we can construct prediction intervals using the estimated MSFE, where $c$ is the critical value from a $t_{n-k}$ distribution:

$$PI_{n+1} = \left(X'_{n+1}\hat{\beta} - c \times \sqrt{\hat{\text{MSFE}}},\ X'_{n+1}\hat{\beta} + c \times \sqrt{\hat{\text{MSFE}}}\right)$$

# Computation

```
# get data
dat <- read.table("support/cps09mar.txt")
exper <- dat[,1] - dat[,4] - 6
lwage <- log( dat[,5]/(dat[,6]*dat[,7]) )
sam <- dat[,11]==4 & dat[,12]==7 & dat[,2]==0
dat <- data.frame(exper=exper[sam],
                  lwage=lwage[sam])

y <- matrix(lwage[sam], ncol=1)
x <- cbind(1, exper[sam])

xxi <- solve(crossprod(x))
xy <- crossprod(x,y)
betahat <- xxi %*% xy

yhat <- x %*% betahat
ehat <- y - yhat
```

```
# calculate MSFE (following BEH pg 112)
P <- x %*% xxi %*% t(x)
etilde <- ehat / (1-diag(P))
msfe <- mean(etilde^2)

# suppose x_{n+1} = 10, the 95% PI is:
xnew <- c(1, 10)
c(low  = xnew %*% betahat - qnorm(0.975) * sqrt(msf
  high = xnew %*% betahat + qnorm(0.975) * sqrt(msf
```

```
     low     high
1.523048 4.325503
```

# Generalized Least Squares

# GLS Setup with OLS Estimator

Take the linear regression model in matrix format and consider a generalized situation where the observed errors are possibly correlated and/or heteroskedastic with $n \times n$ matrix $\Sigma$ and scalar $\sigma^2$:

$$Y = \mathbf{X}\beta + e \qquad \mathbb{E}[e|X] = 0 \qquad \text{Var}(e|X) = \sigma^2\Sigma$$

Then the sampling distribution of $\hat{\beta}_{OLS}$ has the following properties:

$$\mathbb{E}[\hat{\beta}_{OLS}|\mathbf{X}] = \beta$$
$$\text{Var}(\hat{\beta}_{OLS}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Sigma\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

# GLS Estimator

Consider an estimator constructed by pre-multiplying by $L = \Sigma^{-1/2}$:

$$\tilde{\mathbf{y}} = L\mathbf{y} = (L\mathbf{X})\beta + (L\mathbf{e}) = \tilde{\mathbf{X}}\beta + \tilde{\mathbf{e}}$$

Consider OLS estimation for this equation:

$$\begin{aligned}
\tilde{\beta}_{GLS} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\
&= ((L\mathbf{X})'L\mathbf{X})^{-1}(L\mathbf{X})'L\mathbf{y} \\
&= (\mathbf{X}'L'L\mathbf{X})^{-1}\mathbf{X}'L'L\mathbf{y} \\
&= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}X'\Sigma^{-1}\mathbf{y}
\end{aligned}$$

The GLS estimator $\tilde{\beta}_{GLS}$ is BUE:

$$\begin{aligned}
\mathbb{E}[\tilde{\beta}_{GLS}|\mathbf{X}] &= \beta \\
\mathsf{Var}(\tilde{\beta}_{GLS}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \leq \mathsf{Var}(\hat{\beta}_{OLS}|\mathbf{X})
\end{aligned}$$

# Feasible GLS (FGSL)

Because $\Sigma$ is unknown, $\tilde{\beta}_{GLS}$ cannot be calculated.

A feasible estimator is constructed by first estimating $\Sigma$ with some $\hat{\Sigma}$

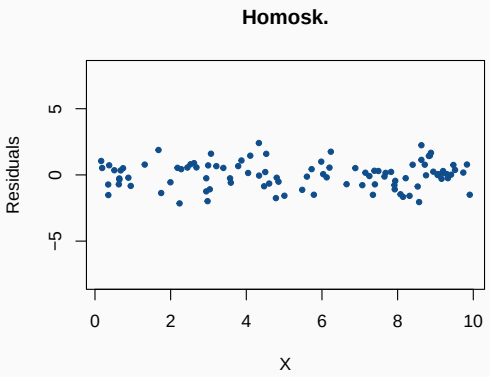
- When there is heterosk. $\Sigma$ is diagonal and we often denote it $D$
- Take the same idea used to construct $\hat{V}_{\hat{\beta}}^{HC0}$ and replace $D$ with $\hat{D}$, which has the squared residuals on the diagonal
- This version of FGLS is often called weighted-least-squares because pre-multiplying by $L$ here is equivalent to scaling the $y_i$ and $x_i$ by $1/\hat{e}_i$

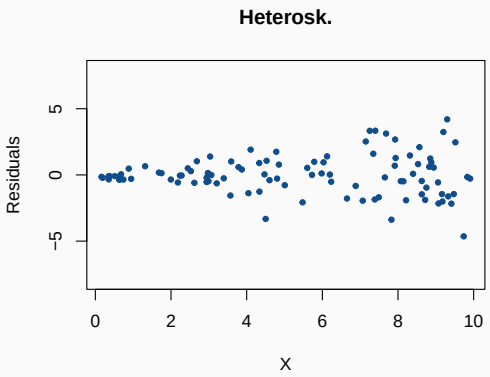
# Testing for Heteroskedasticity

# The Eyeball Test

Plot the residuals (or the squared residuals) against each $X$ variable.

If the absolute magnitudes of the residuals (or their square) are the same regardless of the value of $X$, then homoskedasticity is probably a reasonable assumption.



# Two-Group (Goldfeld-Quandt) Test

Suppose $H_0 : \sigma_i^2 = \sigma^2$

Recipe:

- Break the dataset into two groups: one with low values of $X$ and one with high values of $X$.
  - It's common to leave out some obs "in the middle" between the two groups
- Regress $Y$ on $X$ separately for the two groups.
- Calculate error variance estimates $s_l^2$ and $s_h^2$
- The ratio has an approximate $F$ distribution: $F^{GQ} = s_h^2/s_l^2 \overset{a}{\sim} F_{n_h\text{-}k,\, n_l\text{-}k}$
- Reject the Null Hypothesis if $F$-stat $> c$

# Multiplicative Test

Suppose $H_0 : \alpha = 0$ and $H_a : \alpha \neq 0$ for the function $\sigma_i^2 = \sigma^2 \exp\{z_i'\alpha\}$ where $z$ is a length-$J$ vector that is a subset of $x$.

Recipe:

- Take logs to find: $\log(e_i^2) = \big(\log(\sigma^2) + \log(n)\big) + z_i'\alpha + \nu_i$
- Regress $\log(\hat{e}_i^2)$ on the $z_i$'s to find $\hat{\alpha}$'s
- Calculate the overall $F$-test for the $\alpha$'s
- Reject the Null Hypothesis of homoskedasticity if $F$-stat $> c$

Side note: There is no reason to suspect that $\mathbb{E}[\nu] = 0$ which means that the regression will not provide a good estimate of the intercept, but that's OK because we're focusing on the $\alpha$'s

# Breusch-Pagan Test

Suppose $H_0 : \alpha = 0$ and $H_a : \alpha \neq 0$ for the function $\sigma_i^2 = \sigma^2 h(z_i'\alpha)$ where $z$ is again a length-$J$ subset of $x$ and $h$ is a positive-valued function with $h(0) = 1$

Recipe:

- Miraculously, a test can be derived independently of the function $h()$
- Regress the squared residuals $e_i^2$ on $z_i$ and a constant
- Then $n \times R^2$ from the "auxiliary" regression has a $\chi_J^2$ distribution
- Reject the Null Hypothesis of homoskedasticity if $\chi^2$-stat $> c$

Note: This test is popular because it is easy to compute and you don't need to know $h()$. The generality is also its weakness: more powerful tests can be constructed if the functional form of $h()$ were known.

# White Test

Like prior tests, we use an auxilary regression to assess the possibility that the error variance depends on the values of the original regressors in a linear, quadratic, or interacted way.

Recipe:

- Regress the squared residuals $\hat{e}_i^2$ on $z_i$'s, $z_i^2$'s, cross-products ($z_i \times z_j$), and a constant
- Then $n \times R^2$ from the "auxiliary" regression has a $\chi^2_{2J+J(J\text{-}1)/2}$ distribution
- Reject the Null Hypothesis of homoskedasticity if $\chi^2$-stat $> c$

# Putting it together

1. Funny looking errors should first be interpreted as signaling a specification error
   - eg, omission of an explanatory variable would mean that the error term in the misspecified equation will embody the influence of that omitted variable, which could easily be responsible for any measured heteroskedasticity
2. Choice of an appropriate test for hetersk. is determined by how explicit you want to be about the form of heteroskedasticity
   - In general, the more explicit you are, the more powerful the test will be (ie, correctly reject $H_0$)
   - However, if you get the form of heterosk. wrong, you increase prob. a Type II error (ie, fail to reject $H_0$)
   - Visual inspection of residuals or economic theory can sometimes guide choice of test
3. Then either FGLS or OLS with robust standard errors
   - If the basic structure of hetersk. is unknown, go with OLS and robust standard errors
   - If the magnitude of heterosk. is substantial and you can reasonably estimate its structure, go with FGLS

# Next Time

Midterm!

Computational Topics Potpourri:

- Step-wise Regression
- Cross-Validation
- Bootstrapping