Tip A major component of academia is being aware of current research. It’s easier than ever to do this. Set up an RSS feed (Google Reader is a good one, but there are others), which will consolidate the table of contents of multiple journals for you in one place. Here are some journals where you should be at least skimming the titles and abstracts:

- Biostatistics
- Biometrics
- Annals of Applied Statistics
- JASA/JRSSB/Annals of Statistics
- Journals in the scientific area you (want to) work in

http://www.nature.com/nature/journal/v465/n7298/full/nature09040.html
In the previous class we focused on a general approach to regression based on estimating equations (EEs). This approach is sometimes called the generalized method of moments or robust regression. The basic steps were:

▶ Define a parameter as a *functional* of a super-population
▶ Set up an estimating function that defines that parameter
▶ Plug in our data to get an empirical version of the estimating equation
▶ Solve the equation (numerically/analytically) to get our estimator
▶ Use the CLT to calculate asymptotic/approximate behavior of our statistic
▶ Finished with the Sandwich estimator of variance
Today

- An example of the EE in practice
- Adding assumptions
  - Mean models
  - Mean + variance models (quasi-likelihood)
Suppose we are interested in $\theta$ such that

$$E_F[Y] = \exp(\theta)$$

where $Y > 0$ (Covariates appear in later examples)

With independent sample $Y_i$, this motivates the estimating equation:

$$\sum_{i=1}^{n} G(\hat{\theta}, Y_i) = \sum_{i=1}^{n} Y_i - \exp(\hat{\theta}) = 0$$

which can be solved in closed form

$$\hat{\theta} = \log \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \log \bar{Y}$$

... this also qualifies as a “sane” estimate.
Standard error estimates: example math

To get the empirical variance estimates for $\hat{\theta}$, we calculate

$$G(\theta, Y) = Y - e^\theta$$

$$\frac{\partial}{\partial \theta} G(\theta, Y) = -e^\theta$$

and so

$$\hat{A} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} G(\hat{\theta}, Y_i) = -e^{\hat{\theta}} = -\bar{Y}$$

$$\hat{B} = \frac{1}{n} \sum_{i=1}^{n} G(\hat{\theta}, Y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - e^{-\hat{\theta}})^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

therefore we estimate

$$\hat{\text{Var}}(\hat{\theta}) = \frac{1}{n} \hat{A}^{-1} \hat{B} \hat{A}^{-1} = \frac{1}{n^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

“Robust” intervals are $\hat{\theta} \pm 1.96 \times \text{robust std error}$:

$$\{ \log \bar{Y} - 1.96 \times \sqrt{\sum (Y_i - \bar{Y})^2 / n\bar{Y}}, \log \bar{Y} + 1.96 \times \sqrt{\sum (Y_i - \bar{Y})^2 / n\bar{Y}} \}$$
To get (valid large-sample) 95% intervals for $\mu = e^\theta$, we could just exponentiate the previous output. An alternative is to use the estimating equation:

$$\sum_{i=1}^{n} G(\mu, Y_i) = \sum_{i=1}^{n} Y_i - \mu = 0$$

giving the (sane) estimate $\hat{\mu} = \bar{Y}$. Note that $\hat{\mu}$ is “agnostic” to whether $Y > 0$. To calculate the sandwich:

$$\hat{A} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} G(\hat{\theta}, Y_i) = \frac{1}{n} \sum_{i=1}^{n} -1 = -\bar{Y}$$

$$\hat{B} = \frac{1}{n} \sum_{i=1}^{n} G(\hat{\theta}, Y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\mu})^2 \equiv \frac{n - 1}{n} s^2$$

where $s^2 = Var(Y) \frac{n}{n-1}$, the “corrected” sample variance. Hence we estimate $\hat{Var}[\hat{\mu}] = \hat{B} / n$, and obtain intervals:

$$\{ \bar{Y} - 1.96 \times n^{-1} \sqrt{(n-1)s^2}, \bar{Y} + 1.96 \times n^{-1} \sqrt{(n-1)s^2} \}$$
“Corrected sample variances” should jog your memory; what ‘cookbook’ procedure is the “$\mu$” approach (almost) reproducing?

Try it with $G(\mu, Y) = Y/\mu - 1$ [for fun/practice]

Our $e^\theta, \mu$ approaches are not equivalent - although both intervals shrink with $n^{-1/2}$ - but neither is “wrong”
  
  The “$\mu$” approach is (probably) the easiest to explain - but its interval will go negative, for large-enough confidence levels. This can look somewhat insane.
  
  Using $e^\theta$ is (perhaps) harder to explain, but doesn’t give negative intervals
  
  In large, samples, both are fine, and any difference is likely negligible

Which approach is “best”? It depends on our (subjective) bias-variance trade-off, and $F$ - which we don’t know
Here is a regression example, based on Brockmann et al. (1996, Ethology)

**Some background...**

Horseshoe crabs (of the family *Limulidae*) are “living fossils”, having remained unchanged in $\sim 350$ million years. They have blue blood, two spines, five pairs of legs, and nine yes. Their “carapace” (shell) is horseshoe-shaped, and may be greenish-grey to dark brown in color. Some species are endangered.

Dunlap 1999, Carolrhoda Books (Ages 3+ …not a course text)
Horseshoe crabs form “nests”, with one female in the center and orbiting male “satellites”

We have data on $n = 173$ (independent) nests, including

- Number of satellites $(Y)$
- Carapace width, in cm $(X)$
We’re interested in the multiplicative difference in $Y$, comparing nests where females’ widths differ by 1cm. We estimate this by solving

$$\sum_i G(\hat{\beta}, Y_i, X_i) = \sum_{i=1}^n \{1, X_i\}^T (Y_i - e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}) = 0$$

A straightforward way to do this uses Newton’s method

$$\hat{\beta}^{[k+1]} = \hat{\beta}^{[k]} - \left( \frac{\partial}{\partial \beta} \sum_i G(\hat{\beta}^{[k]}, Y_i, X_i) \right)^{-1} \left( \sum_i G(\hat{\beta}^{[k]}, Y_i, X_i) \right)$$

- iterating until convergence. (Derivatives are taken with respect to $\beta$, with $\hat{\beta}^{[k]}$ then “plugged in”)

Note that the derivative term is $n \times \hat{A}$ evaluated at the “working” $\hat{\beta}^{[k]}$, so uses code we’d be interested in anyway.
Here's some R code to do it, using matrices

# Some Preliminaries
crab = read.table("crab.txt",header=F)
names(crab) = c("id","c","s","w","wt","Sa")
y <- crab$Sa
x <- crab$w
n <- 173
Xmat = cbind(rep(1,n),x)
beta.k <- c(0,0)

# Newton Raphson
repeat({
  exp.term <- as.vector(exp(Xmat %*% beta.k)) # just 'shorthand'
  sumG <- t(Xmat) %*% (y - exp.term)
  Amat <- -t(Xmat) %*% diag(exp.term) %*% Xmat/n # crossprod() is faster [Ex]
  beta.kplus1 <- beta.k - solve(n*Amat, sumG) # the updated value
  if( max(abs(beta.kplus1 - beta.k)) < 1E-6 ) break() # have we converged?
  beta.k <- beta.kplus1 # move to the update }

Bmat <- t(Xmat) %*% diag((y-exp.term)^2) %*% Xmat/n
Vhat <- solve(Amat) %*% Bmat %*% solve(Amat)/n
We have computed $\hat{\beta}$, and the “sandwich” estimate of $\text{Var}_F[\beta]$

We are only interested in $\hat{\beta^1}$ component. Here are the point estimate, its standard error, and an asymptotically justified 95% interval

```r
> beta.k[2]  # R indexes start at 1, not at 0
[1] 0.1640451
> beta.k[2] + qnorm(c(0.025, 0.975)) * sqrt(Vhat[2,2])
[1] 0.1042989 0.2237913
```

This is all for the log of the multiplicative difference. Hence it would be more natural to report

```r
> exp( beta.k[2] )
[1] 1.178267
> exp( beta.k[2] + qnorm(c(0.025, 0.975)) * sqrt(Vhat[2,2]) )
[1] 1.109932 1.250810
```
Although rarely-used in estimation, note that the ‘‘full’’ sandwich estimate is of $\text{Var}_F(\hat{\theta}_n)$ i.e. the covariance matrix of vector valued $\hat{\theta}$. In the crab example, what’s estimated here?

```
> Vhat
0.70619867 -0.0255097839
-0.02550978  0.0009292323

> cov2cor(Vhat)
1.0000000 -0.9958208
-0.9958208  1.0000000
```

Using $\hat{\theta}_n - \theta \rightarrow N(0, \text{Var}_F \hat{\theta}_n)$ as before, confidence ellipses around $\hat{\theta}$ are constructed by giving the central e.g. 95% of the $N(\hat{\theta}_n, \text{Var}_F \hat{\theta}_n)$ distribution\(^1\). So, for the ‘‘full’’ crab example...

\(^1\)For sub-vectors, just take the respective sub-matrix of the sandwich. Also note that tests involving vector-valued parameters are very common.
library(ellipse)
plot(ellipse(x=Vhat,centre=as.vector(beta.k),level=0.95),type="l",lwd=5,col="grey")
points(beta.k[1],beta.k[2],col="red",pch=19,cex=2)
Digression: on R and coding†

- Using matrix calculations saves headaches. Coding summations of individual terms in e.g. \( \hat{\mathbf{A}} \), \( \hat{\mathbf{B}} \) is extremely error-prone, as well as being tedious.

- There are no prizes for doing everything in one line. I could have put \( \text{sumG} \), \( \text{Amat} \), and \( \text{beta.kplus1} \) in one step- but this just makes debugging harder.

- You are expected to be able to code at this “low” level; solving equations, calculating derivatives, etc. If implementing a new method (perhaps your new method\(^2\) then expect to have to code it.

- However, “off the shelf” regression can be implemented with “off the shelf” functions - as we’ll see, soon. When you can use them. Don’t re-invent the wheel.

\(^2\)Becoming an R wizard is a great idea. It will come in very handy when you are developing your own methods and will quickly make you invaluable to collaborators.
Standard error estimates: is that it?†

The form of the EE/sandwich approach here should give sane answers for most situations seen in 754. You can view the methods given as “the general case” - for independent data, they are incredibly versatile.

In the next few lectures, we will discuss how (helpful) modifications can be (and are) made to default EE/sandwich. Where would “help” be required? Consider the conditions we gave for using the EE/sandwich

- Parameter $\theta$ of interest is given by $\mathbb{E}_F[G(\theta, Y, X)] = 0$
- Large-enough $n$ to make the asymptotic approximations usefully accurate
- Our $\{Y, X\}$ are a random sample from $F$ (typically a simple random sample)

To motivate studying other methods (and to avoid thinking of the EE/sandwich as the be one true regression strategy) we consider each of these in turn...
Parameter $\theta$ is given by $\mathbb{E}_F[G(\theta, Y, X)] = 0$

- Describing parameters is hard. Using lines/curves will help us - soon - but this requires more careful interpretation than our statements that $\mathbb{E}_F[G(\theta, Y, X)] = 0$.

- Even then, these tools are not enough when figuring out which regression(s) allow us to make causal statements, rather than just associative ones.

- Sometimes we really don’t know what we want to know. Formal “inference” (à la 570) is not then available, but regression “tools” can be used to explore datasets, and to generate some targets for subsequent inference. If you interpret the output as “only exploratory”, this is Not Evil. (Fussing over precise frequentist validity seems particularly futile, in exploratory work)
Large-enough $n$ to make the asymptotic approximations used accurate

- EE approaches will work eventually, with minimal assumptions - but not until $n$ is vast, in some cases. What might we do in such situations?
- Making some (sane) assumptions can allow us to get more efficient estimates, and better-behaved standard errors.
- We should expect to pay a price, if these assumptions are wrong - a big price, if they are badly wrong
Standard error estimates: is that it?†

Our \{Y,X\} are a (simple) random sample from F

- Sampling the wrong super-population will generally lead to bias problems
- Stating that the target super-population is “the set of all experiments done like the one I have” sidesteps this problem - but its basically a tautology, and Not a Good Idea
- A separate problem for EE/sandwich approaches is that X may not be random. If we chose exactly 100 people in their 20s, 30s, 40s etc., this is not a random sample from any natural target superpopulation. But it is a reasonably sane design - so it may be more natural to provide standard errors that treat X as fixed, not random.
Our \{Y,X\} are a (simple) random sample from \(F\).

You are probably familiar with the idea of fitting lines to data.

Up to now, we have stressed that we don’t need line-fitting to justify regression. But it is a useful idea, because

- It can provide more intuitive interpretation of “difference” parameters
- “More-parametric” methods than EE/the sandwich often make assumptions which can be expressed as linearity, on some scale. Viewing line fitting as an EE/sandwich operation will (later) help us understand how/where extra assumptions can help.
Line fitting: straight lines†

What’s a sensible “straight line” summarizing this superpopulation?
One sane summary is the straight line with intercept $\beta_0$, slope $\beta_X$ that minimizes

$$\mathbb{E}_F \left[ (Y - \beta_0 - \beta_X X)^2 \right]$$

Empirically, we can estimate this by finding the $\hat{\beta}$ that minimize

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_X X_i)^2$$

... i.e. the Least Squares Estimator. With some algebra, you should find that $\hat{\beta}_X = \text{Cov}(X, Y)/\text{Var}(X)$ [verify this] This formula should look familiar!

Use of sandwich standard errors based on

$$\sum_{i} \{1, X_i\}^T (Y - \beta_0 - \beta_X X_i) = 0$$

would give the same confidence intervals as for our “weighted mean slope” parameter. This is because the two parameters are exactly the same thing - expressed in two different ways.
Line fitting: straight lines

The equivalence means we are free to write about

- Estimating the “weighted mean slope” (as before)
- Estimating the “difference in Y for a 1-unit difference in X” (with implicit up-weighting of distant X’s)
- Estimating the best-fitting straight line $y = \beta_0 + \beta_x x$ (a line that increases by $\beta_x$ over each $[x, x + 1]$) where “best-fitting” implies “ordinary least squares”
- Estimating the “linear trend” or “linear association” of Y with X (This is more informal, but still useful)

Stating “we did linear regression” is also okay, and is concise, but requires the reader to interpret the output in the same way as you. (Our lack of assumptions isn’t conveyed very well by this phrase; many readers would infer that you assumed *something* about $F$ is truly linear).
Line fitting: straight lines?†

If X is binary, you can also talk about the “expected difference in Y for a 1-unit difference in X…”

...because there is only one difference!
Line fitting: straight lines?†

Otherwise, be careful with this phrase - what exactly do you mean by the “Expected difference in Y for a 1-unit difference in X”?

Is it the “Difference in $\mathbb{E}Y$ for a 1-unit difference in X”? Are you averaging over X? Is your reader?
Without further assumptions, you will in general look foolish claiming that all 1-unit differences are the same. So don't do this! Careful writing should make it unambiguous that you’re estimating a linear trend - perhaps defined by least squares.

Pragmatically, it’s particularly important to avoid statements about “true linearity” when the data are very noisy, or when the data are obviously non-linear.

Expect a struggle, if/when your reviewer only believes in “truly linear” universes.

If the linear trend isn’t of scientific interest - say when predicting new Y’s - you’ll get (and will deserve!) a hostile reception.

---

In practice it is just as bad to write something that’s perceived to claim this, even if it’s not what you meant.
However you write about it, the “least squares” argument generalizes straightforwardly to higher dimensions; I assume you know that

\[ \hat{\beta} = (X^T X)^{-1} X Y \]

is the least squares estimate when fitting the (hyper) plane defined as \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \) (You can also interpret this as estimating a multivariate trend, à la Jacobi).

Least-squares also works for curves, consider estimating \( \beta \) minimizing \( \mathbb{E}_F[(Y - e^{X^T \beta})^2] \)

You would do this (I hope) by solving

\[ \sum_{i=1}^{n} X_i e^{X_i^T \beta}(Y_i - e^{X_i^T \beta}) = 0 \]

and then using the sandwich.
But why always “plain vanilla” least squares? We could “fit a line” (and estimate a sane parameter) by instead minimizing

$$\sum_{i=1}^{n} w_i (Y_i - e^{x_i^T \hat{\beta}})^2$$

meaning $\hat{\beta}$ is the solution to

$$\sum_{i=1}^{n} w_i X_i e^{x_i^T \hat{\beta}} (Y_i - e^{x_i^T \hat{\beta}}) = 0$$

Informally, one way to proceed then sets $w_i = e^{-x_i^T \hat{\beta}}$ and solves

$$\sum_{i=1}^{n} X_i (Y_i - e^{x_i \hat{\beta}}) = 0$$

- where did we see this before?
This informal argument aids intuition, but is “cheating”! The EEs we just gave are actually equivalent to minimizing e.g.

$$\sum_{i=1}^{n} Y_i X_i^T \beta - e^{X_i^T \beta}$$

but non-parametrically, it’s not at all obvious this is a sane criterion it would be hard to defend, presented like this.

But the EE we stated gives the same estimates as “minimizing”

$$\sum_{i=1}^{n} w_i (Y_i - e^{X_i^T \beta})^2$$

for $w_i = e^{-X_i^T \beta}$ by a modified Newton-Raphson method

$$\beta^{[k+1]} = \arg\min_{\beta} \sum_i w_i (\beta^{[k+1]})(Y_i - e^{X_i^T \beta})^2$$

This method is Iteratively reWeighted Least Squares. It’d be an insane way to define parameters - but is smart equation-solving.
Line fitting: with crabs!†

Fitting the crab data, with IWLS using $w_i = e^{-x_i^T \beta}$, and unweighted least squares (R code follows shortly)
“Fitting the line” \( y = e^{\beta_0 + \beta_1 x} \) in these two ways, we get:

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Inference (sandwich-based)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta_1 )</td>
</tr>
<tr>
<td>“Plain” Least Squares</td>
<td>0.139 (0.087, 0.191)</td>
</tr>
<tr>
<td>IWLS Approach</td>
<td>0.164 (0.104, 0.224)</td>
</tr>
</tbody>
</table>

- The two estimates/intervals differ slightly
- For general \( F \), the two parameters are different - despite the Greek letters! Which parameter do you want?
- Write about the “log-linear trend” if you don’t like “multiplicative differences”
- The “intercept” \( \beta_0 \) is a “nuisance” parameter; for inference on differences alone, remember we only care about the slope of the best-fitting line (a log-linear line, in this case)
Line fitting: general case

Up to now, we have seen linear regression, and two forms of log-linear regression. But these are only examples, and can be absorbed into a larger framework. Consider the general EE:

$$\sum_{i=1}^{n} \frac{\partial g(X_i^T \beta)}{\partial \beta} w(X_i^T \beta)(Y - g(X_i^T \beta)) = 0$$

$g(\cdot)$ tells you what line you’re fitting. But its inverse $g^{-1}(\cdot)$ (mapping from $Y$-space to $\mathbb{R}$) is called the link function.

$$
\begin{align*}
    g(X_i^T \beta) &= X_i^T \beta & \text{identity link} \\
    g(X_i^T \beta) &= \exp(X_i^T \beta) & \text{log link} \\
    g(X_i^T \beta) &= 1/X_i^T \beta & \text{inverse link}
\end{align*}
$$

Formally, $\beta$ tells you about “slopes” of fitted surfaces, on a scale defined by $g^{-1}(\cdot)$. Other (occasional) choices include square-root link, and the “probit” function - the inverse of the Normal distribution function.
Line fitting: general case†

Having specified a link, the choice of weights is made by specifying a family

\[
\begin{align*}
  w(\mathbf{X}_i^T \beta) &= 1 & \text{Gaussian/Normal family} \\
  w(\mathbf{X}_i^T \beta) &= 1/g(\mathbf{X}_i^T \beta) & \text{Poisson family} \\
  w(\mathbf{X}_i^T \beta) &= \frac{1}{g(\mathbf{X}_i^T \beta)(1-g(\mathbf{X}_i^T \beta))} & \text{Binomial family}
\end{align*}
\]

- As you might guess from the names, these have connections to parametric models - but keep in mind that our interpretation does not need/use them.
- Judicious, *canonical* choice of \(w(\cdot)\) leads to “pretty” cancellation, and (often) an EE of the form

\[
\sum_{i=1}^{n} \mathbf{X}_i(Y_i - g(\mathbf{X}_i^T \beta)) = 0
\]
The general $g(\cdot), w(\cdot)$ case describe the estimating equations you would get maximizing the likelihood of a **Generalized Linear Model** (we’ll talk more about this later)

- If you believe the data come from that GLM, this is sane
- If you don’t? Solve the EEs anyway. Quite distinct from their MLE interpretation, these EEs (as we saw) typically estimate sane, interpretable parameters.

- This isn’t how they were developed! GLMs are the models where the MLE can be calculated by IWLS. While this was a Big Deal in the 70s, it isn’t now. Choosing parameters simply because we can compute their estimates in a (very) few steps looks rather contrived - to modern statisticians.

So why am I telling you about GLMs...?
Line fitting: general case†

... because they enable off-the-shelf use of the \texttt{glm()} function in \texttt{R}. Here's all the \texttt{R} code for the horseshoe crabs:

```r
# Fit the weighted least squares and least squares lines for the crabs
glm1 <- \texttt{glm}(y ~ x, data=crab, family=poisson)
glm2 <- \texttt{glm}(y ~ x, family=gaussian(link="log"), start=c(-3,0.2))

# Get the sandwich estimates of variance
library(sandwich)
vhat1 <- \texttt{vcovHC}(glm1, "HC0")
vhat2 <- \texttt{vcovHC}(glm2, "HC0")
coef(glm1)[2] + \texttt{qnorm}(c(0.5, 0.025, 0.975))*sqrt(vhat1[2,2])
coef(glm2)[2] + \texttt{qnorm}(c(0.5, 0.025, 0.975))*sqrt(vhat2[2,2])
```

▶ Understanding this code is important
▶ I hope \texttt{y \sim x} is familiar to you. Also, you are aware that you can use \texttt{lm()}, for Gaussian family/identity link
▶ Check the help files for “default” links in \texttt{glm} [for fun/practice]
▶ Non-“pretty” (non-canonical) choices may need \texttt{start} values
▶ The \texttt{vcovHC()} function gives (a form of) sandwich estimates.
Line fitting: reparameterizing†

Exactly the same line/plane can be described in more than one way. For example

$$y = g(\beta_0 + \beta_1 x)$$

is the same line as

$$y = g(\gamma_0 + \gamma_1 (x - c))$$
$$y = g(\psi_0 + \psi_1 (ax - b))$$

for fixed and known $a, b, c$

- Fit the same line $\implies$ get the same inference
- $\gamma$ and $\psi$ are re-parameterizations or transformations of $\beta$
  - and of each other
- Rules for transforming random variables don’t apply here - because we’re not transforming random variables!
- Choosing $c = \bar{X}$ is a common (and sane) choice; what does $\gamma_0$ then tell you? In what sense is $\gamma_0$ a “nuisance”?
- Why is this less clever-looking, if $X = 1_{\text{Male}}$?
You can probably deal with transforming $X$ from grams $\leftrightarrow$ in your head. Celsius $\leftrightarrow$ Fahrenheit is also not too tricky.

For more complex reparameterizations, such as

$$y = g(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4)$$
$$y = g(\psi_0 + \psi(x - c) + \psi_2(x - c)^2 + \psi_3(x - c)^3 + \psi_4(x - c)^4)$$

... you may find it easier to **fit precisely what you want**, and not “patch things up” post hoc$^4$

- What does $\psi_0$ tell you, if $c = \bar{X}$ here?
- Note this discussion is limited to transformations of $X$ which are **linear** (and can be inverted)

$^4$Some parameterizations do make fitting numerically more stable/accurate. But your computer is from $\sim$ 2011, not from 1979
Mean models†

We can now increase the level of assumptions, and suppose our data come from a superpopulation where the mean is constrained

$$\mathbb{E}_F[Y|X = x] = g(x^T\beta)$$

where $5 \ x = \{1, x_1, \ldots, x_{p-1}\}$ and $\beta \in \mathbb{R}^p$. This is known as “assuming a mean model”.

At this stage, we make no other assumptions (other than independence and regularity conditions) - in particular nothing is assumed about the form of variances, skewness, etc.

We now have some finite-dimensional constraints (on the mean) and some unconstrained properties (variance, etc.) - making this a semi-parametric approach. (The previous EE/sandwhich development was non-parametric).

---

5 The intercept term is included here, in line with standard terminology. Also, as before, transformations of $X$ will not affect anything meaningful.
If you believe the assumed mean model, interpretation gets a bit easier. Say, for example

\[ \mathbb{E}_F[Y|X = x] = \beta_0 + \beta_1 x \]

then \( \beta_1 \) is the difference in expected \( Y \), comparing subjects whose \( x \)-values differ by 1 unit. (Note that this hold for any 1-unit difference). Writing about the “expected difference in \( Y \) for a 1-unit difference in \( x \)” is common, and justifiable.

If we instead assumed that

\[ \mathbb{E}_F[Y|X = x] = \exp(\beta_0 + \beta_1 x) \]

then \( \beta_1 \) is the log of the ratio of the expected value of \( Y \), comparing subjects whose \( x \)-values differ by 1 unit. Ratios and expectations don’t commute, so avoid writing this up as “expected ratio” or “expected multiplicative difference”
With $p > 2$, the interpretation is noticeably easier than “multidimensional trends” or “fitted hyperplanes”. Assuming

$$E_F[Y|X = x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1}$$

then $\beta_1$ is the difference in expected $Y$, comparing subjects whose $x_1$ values differ by 1 unit, and for whom everything else is the same. 6

Use this interpretation carefully, it may not be sane!

$$E_F[Y|X = x] = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$$

You can’t compare $x_1 = c$ to $x_1 = c + 1$ with everything else the same! A reasonable default here is to give inference on differences in expected $Y$ by interpreting $\{\beta_1, \beta_2, \beta_3\}$ jointly. (But for some jobs, e.g. assessing strict linearity, you may not need all terms).

6For a log-linear mean, it would be the log of the ratio of the expected values of $Y$, comparing subjects whose $x$-values differ by 1 unit and for whom everything else is the same.
Mean models : misinterpretation†

You may find “for whom everything else is the same” too passive. Many authors instead refer to comparisons “keeping everything else fixed” or “holding other variables constant”

However, to many readers these phrases imply you can meaningfully hold everything else constant. This may not be true! - for example, under a mean model for \( Y = \) Soccer ball sales in a city and two covariates

\[
E[Y|X = x] = \beta_0 + \beta_1 \#(\text{Youth Teams}) + \beta_2 1_{(\text{Have MLS Team})}
\]

Soccer is a sport that is steadily growing in popularity in the US and the DC metro area got an MLS team in 1996. Think about the DC’s parks/shops in 1990 and 2000.

- If \( X_j \) and \( X_k \) are correlated, the conditional (or partial) \( \beta_k \) coefficient doesn’t tell you about the marginal \( Y - X_k \) association - and this result generalizes to \( X \in \mathbb{R}^p \)

- Outside linear, log-linear mean models, even independence of \( X_j, X_k \) won’t save you!

- You must write very clearly about the comparisons your coefficients provide.
Mean models: misinterpretation†

For estimation (not intervals), fitting the mean model follows as for line-fitting; with an assumed mean model
\[ \mathbb{E}[Y|X = x] = g(x^T\beta), \]
we can estimate \( \beta \) by solving
\[
\sum_{i=1}^{n} x_i(Y_i - g(x_i^T\hat{\beta})) = 0
\]

Recall you can think of this as a “canonical” form of weighted least squares fitting, of the line/plane \( y = g(x^T\beta) \). More generally we can solve
\[
\sum_{i=1}^{n} \frac{\partial g(x_i^T\beta)}{\partial \beta} \bigg|_{\beta=\hat{\beta}} w(x_i^T\beta)(Y_i - g(x_i^T\beta)) = 0
\]

For general \( \beta \) and \( w(\cdot) \), under the assumed mean model and mild regularity conditions, this \( \hat{\beta} \) consistently estimates \( \beta \)
Mean models : choice of estimate†

Up to now, for each parameter we have usually considered a single “empirical” estimate.

- Now we have more than one consistent estimator, for the same parameter - \( w(\cdot) \) defines our choice
- Any two consistent estimators need not agree for any one dataset (Though for large \( n \), they should both be close to the truth).

This approach raises two new (and broad) issues

- How much does conditioning on the X-values matter (compared to them being random, as in EE/sandwich)
- Which weights \( w(\cdot) \) might be better/worse?

Having assumed a mean model, we may also seek re-assurance that it’s right - or at least not grossly wrong
We have stated the mean model conditional on $X$; i.e.
\[
E[Y|X = x] = g(x^T \beta^c).
\]
But for frequentist properties, should we be considering replicates where the $X_i$ are

- random? - as we saw for EE
- fixed? - as we might have in an experimental setting e.g. assigning $n/8$ to each of $2^3$ possible treatments

We just talked about how $\hat{\beta}$ is consistent for parameter $\beta^c$ in the mean model (for fixed $X=x$)

Recall from last lecture that the same $\hat{\beta}$ is consistent for

\[
\beta^m : \mathbb{E}_F \left[ \frac{\partial g(X^T \beta^m)}{\partial \beta^m} w(X^T \hat{\beta}^m)(Y - g(X^T \hat{\beta}^m)) \right] = 0
\]

where we assumed $F$ took simple random samples of $\{Y, X\}$ i.e. random $X$. 
Mean models: fixed or random $X^{\dagger}$

If the mean model is right, both $\beta[c]$ and $\beta[m]$ seem like reasonable parameters for inference. But are they actually the same thing?

Suppose $\mathbb{E}_F[Y|X=x] = x^T \beta[c]$. Then

$$\mathbb{E}_F[X(Y - X^T \beta[m])] = 0$$

$$\implies \mathbb{E}_F[X(X^T \beta[c] - X^T \beta[m])] = 0$$

$$\implies \mathbb{E}_F[XX^T](\beta[c] - \beta[m]) = 0$$

$$\implies \beta[c] = \beta[m]$$

... if the $p \times p$ matrix $\mathbb{E}_F[XX^T]$ is of full rank.

So in (sane) applications of Ordinary Least Squares, with the right mean model, the answer is “Yes!”, and the distinction doesn't matter.

With independent data, $\beta[c] = \beta[m]$ very generally, the proof will appear (probably) in your stat theory class.
Mean models: fixed or random \( X \uparrow \)

So why make a fuss about this? We have to consider what happens when the assumed mean model is **wrong**. Consider thought experiment \# 2 - the resistors:

True: \( \mathbb{E}[Y|X = x, Z = z] = \gamma_0 + \gamma_1 x + \gamma_2 z \)

Assumed: \( \mathbb{E}[Y|X = x, ] = \beta_0 + \beta_1 x \)

- Using OLS, we’d estimate \( \hat{\beta}_1 = \text{Cov}(Y, X)/\text{Var}(X) \), as usual. This will only be consistent for \( \gamma_1 \) if \( X \) and \( Z \) are uncorrelated or if \( \gamma_2 = 0 \). Perhaps these hold - but either would usually be a strong assumption. (For unmeasured \( Z \) they seem particularly hard to defend)

- Yet with EE, \( \hat{\beta}_1 \) is consistent for the superpopulation’s “linear trend” \( \beta_1 \) parameter.

In general, if the mean model is wrong, expect consistency only for “trends” (Randomized trials give an important exception).
Mean models: fixed or random $X^\dagger$

Other notes on fixed-$X$ versus random-$X$ estimation

- A lot of “classical” work emphasizes mean model parameters - and worries (a lot) about whether the mean model is right.
- Superpopulation “trends” are okay! But they take work to explain, and can look contrived if the corresponding line-fit looks implausible.
- When you claim a procedure is “biased” or “inconsistent”, you must state whether this is with respect to a mean model parameter, or a superpopulation parameter.
- “Speaking precisely” is really important.
- With dependent data (and unobserved parts of $X$) $\beta^c$ and $\beta^m$ can differ even when the mean model is right (we’ll see that later).
In the mean model setting, we can imagine constructing intervals which have 95% coverage among replications of the universe where $X = x$ happened exactly as in our experiment.

Contrast this with the “default” EE/sandwich setting, where we constructed intervals which cover the true parameter value in e.g. 95% of replications where $X$ varies randomly. Cox (1958) showed these don’t have to be the same...
Mean models: intervals

In case you thought this was all getting aimless

Based on a simple coin toss, the red or blue archer shoots at the bullseye. You know red gets 95% of shots within 3cm of it; blue gets 95% to within 7cm.
Mean models: intervals†

In case you thought this was all getting aimless

YOU'RE BEHIND THE TARGET, AND DON'T KNOW THE LOCATION OF THE BULLSEYE. ONE ARROW IS FIRED - YOU DO KNOW WHO FIRED IT!
Mean models: intervals†

In case you thought this was all getting aimless
Suppose the red distribution was as follows:
Mean models: intervals†

Suppose the blue distribution was as follows:

Drawing radius 3cm/7cm circles around $Y$ gives 95% coverage
But we could ignore the red/blue data

Fixed-radius $\sim 6$cm around $Y$ also gives 95% coverage marginally
Mean models: intervals†

In a designed fixed-X setting, if we have a valid confidence interval “recipe” $\text{Cl}(Y,X)$ for parameter $\beta^c$. If $\beta^c = \beta^m$.

- You get accurate asymptotic coverage for the conditional parameter under fixed $X$ and for the marginal parameter for fixed or random $X$.

If $\beta^c \neq \beta^m$ (as is usual when the mean model is wrong) - then the result doesn’t hold

- For the conditional parameter you get the wrong coverage - usually under coverage
- You may still get decent coverage for the marginal parameter. But this only makes sense as a trend-type parameter.
Assuming a mean model with fixed $X$

$$\mathbb{E}_F[Y_i|X_i = x_i] = g(x^T \beta)$$

means the data aren’t i.i.d. so our original sandwich estimate doesn’t apply but if $\hat{\theta}_n$ solves

$$G_n(\theta) = \sum_{i=1}^{n} G_i(Y_i, \theta) = 0$$

- where the summands depend on the fixed $x_i$, then

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow_D N_p(0, A_n^{-1} B_n A_n^{-1})$$

$$A_n = \mathbb{E}_F \left[ \frac{1}{n} \frac{\partial}{\partial \theta} G_n(\theta) \right] = \sum_{i=1}^{n} \mathbb{E}_F \left[ \frac{\partial}{\partial \theta} G_i(Y_i, \theta) \right]$$

$$B_n = \mathbb{E}_F[ G_n(\theta) G_n(\theta)^T ] = \text{Cov}[G_n(\theta)]$$

So we can get a “sandwich” for the covariance.
As before we plug in empirical estimates

\[ \hat{A}_n = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} G_i(\hat{\theta}_n, Y_i) \]

\[ \hat{B}_n = \sum_{i=1}^{n} G_i(\hat{\theta}_n, Y_i) G_i(\hat{\theta}_n, Y_i)^T \]

- For weighted least squares EE $\hat{A}_n$ is consistent for $A_n$.
- $\hat{B}_n$ is generally slightly larger than $B_n$\(^7\)
- Intervals based on $\hat{\theta} \sim N(\theta, \hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1})$ are slightly conservative, asymptotically.

\[^7\text{because it estimates } \mathbb{E}_F G_n G_n^T \text{ not } \text{Cov}_F[G_n] \]
Efficient Inference: EE’s†

Now back to the general form of the EE’s, which we can interpret as fitting some line/surface to the data.

\[
U(\beta) = \sum_{i=1}^{n} \left[ \frac{\partial g(X_i^T \beta)}{\partial \beta} \right]^T w(X_i^T \beta)(Y_i - g(X_i^T \beta)) = 0
\]

If we have specified the mean model correctly so

\[
\mathbb{E}_F[Y|X = x] = g(x^T \beta)
\]

for unknown \( \beta \), then any sane weights \( w(\cdot) \) will give consistency. What choice of weights work well?

- Just as in linear regression, only specifying weights up to a constant won’t change anything.
- “Small \( \text{Var}[U(\beta)] \)” is too unspecific

---

8Formally, the class of linear unbiased estimating equations
Godambe and Heyde’s (1987) “fix” for the scaling problem was to consider behavior of not $U$, but of a “normalized” form:

$$U^* = \mathbb{E}_F [\frac{\partial U(\beta)}{\partial \beta}]^{-1} U(\beta)$$

–note that the normalizing constant is free of $Y$. They showed that in the class of estimators on the previous slide, EEs of the form

$$U(\beta) = \sum_{i=1}^{n} \left[ \frac{\partial g(X_i^T \beta)}{\partial \beta} \right]^T \text{Var}[Y_i|X_i = x_i]^{-1} (Y_i - g(X_i^T \beta)) = 0$$

are optimal \(9\) in the sense that:

$$\mathbb{E} \left[ (U^*(\beta))^2 \right] \leq e \left[ (U'^*(\beta))^2 \right]$$

for all other choices $U'$ of ‘standardized’ unbiased EEs - which are then normalized.

\(9\) Optimal and quasi-score estimating functions refer to the same thing.
This is another very important result; for a wide range of regressions it tells us to weight contributions in inverse proportion to the variance of their respective $Y_i$’s

- ...because this approach minimizes the variability of the (standardized) EEs, which leads to maximally-stable solutions of unbiased EEs, i.e. good $\hat{\beta}$’s
- We already knew this for linear regression (Aitken/Gauss-Markov) here the result is more general
- It’s sane; we’re giving maximum weight to maximally-stable contributions and least weight to the flakiest
- Putting all the weight on one data point is almost never sane
- The result is **semiparametric** we must specify $\mathbb{E}[Y|X = x]$ and $\text{Var}[Y|X = x]$, but not a full likelihood.
Quasi-likelihood†

'Quasimodo'
Victor Hugo

'Quasar'
Jocelyn Bell

'Quasi-likelihood'
Wedderburn, Nelder

Like a new-born
Like a star
Like a likelihood (!)
We now increase the level of assumptions again;

\[
\begin{align*}
\mathbb{E}_F[Y|X=x] &= g(x^T\beta) \\
\text{Var}_F[Y|X=x] &= \alpha V(g(x^T\beta))
\end{align*}
\]

We assume we know

- the "true" structure of the mean and variance
- that the true variance depends only on an unknown scale parameter \(\alpha\), and a particular function of the mean -denoted \(V(\cdot)\).

The second assumption is quite strong; it does not permit

\[
\begin{align*}
\mathbb{E}_F[Y|X=x] &= \beta_0 + \beta_1 x + \beta_2 z \\
\text{Var}_F[Y|X=x] &= \alpha(1 + \gamma z^2)
\end{align*}
\]
With these assumptions, how might we get estimates/inference that are

- Consistent for $\beta$ in the assumed mean
- Efficient, or nearly efficient, given the assumed variance
- Easy to compute?

After devising GLMs (with John Nelder), Robert Wedderburn (1974) suggested “Quasi-likelihood” methods this name reflects that we get some properties of likelihood-based approaches. However, the methods can also be interpreted as “mean and variance”-modeling, where the mean-variance relationship is (very) restricted.
Wedderburn’s key insight was that, for independent data with mean \( \mu = \{ g(x_i^T \beta) \} \) and (diagonal) variance elements \( \{ \alpha V(\mu_i(\beta)) \} \) defining

\[
U_i = \frac{Y_i - \mu_i}{\alpha V(\mu_i)}, \quad i = 1, \ldots, n
\]

we get the following (helpful) properties)

\[
\mathbb{E}_F[U] = 0 \\
\text{Var}_V[U] = \text{diag}\{1/\alpha V(\mu_i)\} \\
-\mathbb{E}_F[\frac{\partial U_i}{\partial \mu}] = 1/\alpha V(\mu_i)
\]

-note that these are similar in structure to earlier \( \mathbb{E}[G], A, B \) formulations.

Wedderburn realized they could be motivated by minimizing a weighted sum of squares with IWLS, not e.g. Newton-Raphson.
However you get there, some clearly sane EEs are

$$D(\hat{\beta})^T V(\hat{\beta})^{-1} \{Y - \mu(\hat{\beta})\}/\alpha = 0$$

where $D$ is the $n \times p$ matrix of derivatives with elements $\partial \mu_i/\partial \beta_j$.

As shorthand we write $U(\beta) = D^T V^{-1} \{Y - \mu\}/\alpha$ - we are back to a linear function of $U$, making evaluation straightforward. Following from $U$ we find

$$E[U(\beta)] = 0$$
$$\text{Cov}\{U(\beta)\} = D^T V^{-1} D/\alpha$$
$$-E \left[ \frac{\partial U}{\partial \beta} \right] = \text{Cov}\{U(\beta)\} = D^T V^{-1} D/\alpha$$
These are just estimating equations, so we get the usual result

\[(D^T V^{-1} D)^{1/2}(\hat{\beta}_n - \beta) \rightarrow_D N_p(0, \alpha I_p)\]

This is almost all we need to use the method

- Solve the EEs, to get \(\hat{\beta}\)
- Plug \(\hat{\beta}\) into formula for \(D, V\) matrices, to get estimates of everything on LHS
- With an estimate of \(\alpha\), get an asymptotically accurate standard error estimate, and construct intervals in the usual \(\hat{\beta}_j \pm 1.96 \times \text{Est. Std. Error} \) format
Quasi-likelihood: unknown $\alpha^\dagger$

We get estimates of $\alpha$ by essentially ad hoc arguments (not a real problem in practice). For example, since

$$\mathbb{E}_F[(Y - \mu)V^{-1}(\mu)(Y - \mu)] = n\alpha$$

an unbiased estimator of $\alpha$ is given by

$$\hat{\alpha} = (Y - \mu)^T V^{-1}(\mu)(Y - \mu)/n$$

A “degrees of freedom-corrected” (but not in general, unbiased) estimate is given by the Pearson statistic, divided by its degrees of freedom

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

The asymptotic distribution that is used in practice is therefore given by

$$(\hat{D}^T \hat{V}^{-1} \hat{D}/\hat{\alpha})^{1/2}(\hat{\beta}_n - \beta) \rightarrow_D N_p(0, I_p)$$