Lecture 10

Model problems, consequences and fixes

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

In this lecture we looked at the assumptions in our linear model, and considered how they might fail to hold, what the consequences would be, how to detect those problems and how to fix them (when a fix is available). In our strongest assumptions we assume that $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$. We write $y_i = \boldsymbol{x}_i^\mathsf{T}\beta + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The things that could go wrong include

1) $\mathbb{E}(Y) \neq X\beta$,
2) $\mathrm{var}(Y) \neq \sigma^2 I$, and
3) Gaussianity may fail.

The second item includes two problems: unequal variance among the y_i and non-zero correlations among them. We might also face outliers, which in some cases is a special case of the third problem. The assumptions that are hardest to believe are not necessarily the most damaging ones.

10.1 Residuals

Most of the ways that we can detect errors in our model assumptions are based on the residuals $\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i^\mathsf{T}\hat{\beta}$, which we treat as estimates of ε_i.

There are several other residual quantities that we can work with. For example $\hat{\varepsilon}_i/s$ should have approximately mean zero and variance one under our assumptions. With this normalization $|\hat{\varepsilon}_i|/s > 7$ for example looks very unusual and we might then inspect observation i more closely. It is very useful to consider the studentized residuals

$$\frac{\hat{\varepsilon}_i}{s\sqrt{1-H_{ii}}}.$$

Under our model $\mathrm{var}(\hat{\varepsilon}_i) = (1-H_{ii})\sigma^2$ and then $\hat{\varepsilon}_i/\sqrt{1-H_{ii}}$ has variance σ^2 for all i. Further dividing by s makes the common variance roughly 1. (Not exactly 1 because s has variance.)

We could also use a leave-one-out residual such as

$$\frac{y_i - \hat{y}_{-i}}{s_{-i}}\sqrt{1-H_{ii}}$$

where s^2_{-i} is the sample variance estimate after observation i has been held out.

10.2 Non-normality

We can be pretty sure that the ε_i do not have a Gaussian distribution. This is usually not our most serious difficulty because in general the central limit theorem (CLT) will tend to give $\hat{\beta}$ a nearly Gaussian distribution anyway.

Suppose that $\boldsymbol{x}_i$ are IID and independent of ε_i which are also IID with mean 0 and variance σ^2. For data with $y_i = \boldsymbol{x}_i^{\mathsf{T}}\beta + \varepsilon_i$,

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta) &= \sqrt{n}\big[(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(X\beta+\boldsymbol{\varepsilon})-\beta\big]\\ &= \sqrt{n}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{\varepsilon}.\end{aligned}$$

Our assumption makes $\mathbb{E}(\boldsymbol{x}_i\varepsilon_i) = 0$ and now suppose that $\mathbb{E}(\|\boldsymbol{x}_i\varepsilon_i\|^2) < \infty$ and $\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}) = \Sigma \in \mathbb{R}^{p\times p}$ has full rank. Then

$$\sqrt{n}(\hat{\beta}-\beta) = \Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i\varepsilon_i.$$

The first factor converges in probability to Σ^{-1}. The second factor converges in distribution to

$$\mathcal{N}(\boldsymbol{0}, \mathrm{var}(\boldsymbol{x}_i\varepsilon_i)) = \mathcal{N}(\boldsymbol{0}, \Sigma\sigma^2)$$

so

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{\mathrm{d}} \mathcal{N}(\boldsymbol{0}, \Sigma^{-1}\sigma^2).$$

Our t-tests also use s^2 and s^2 converges in probability to σ^2 when $\mathrm{var}(Y) = \sigma^2 I$ regardless of whether Y has a Gaussian distribution. As a result, our t-tests are asymptotically valid along with our p-values, extra sum of squares tests and our confidence intervals for $\boldsymbol{c}\beta$.

We can now reason how to break the CLT. If ε_i have a distribution that is too heavy-tailed for the CLT to hold then we will not get a CLT for $\hat{\beta}$.

We can also make $\boldsymbol{x}_i$ heavy tailed, perhaps with one or more components from a distribution that has infinite variance. Or we can intervene even more directly by putting outlier rows in X. For instance, one violation of the model would have points $\boldsymbol{x}_i = (1, z_i)^\mathsf{T} \in \mathbb{R}^2$ with $z_1, \dots, z_{n-1} \overset{\text{iid}}{\sim} \mathbb{U}[0,1]$ and an outlier at $z_n = n$. Or we could have $\boldsymbol{x}_i = (x_{i1}, x_{i2})^\mathsf{T}$ where $x_{i1} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ while $x_{1,2} = 1$ and $x_{i,2} = 0$ for $2 \leqslant i \leqslant n$. For this model with no intercept we can work out what happens to $X^\mathsf{T}X/n$ and see that it does not converge to a positive definite matrix. Adding an intercept would not fix that problem!

While non-Gaussianity does not ordinarily cause severe trouble for our inferences on β it does have an effect on the prediction intervals we discussed. Those assumed that $y_i - \boldsymbol{x}_i^\mathsf{T}\hat{\beta}$ was Gaussian. Even when the CLT applies to $\boldsymbol{x}_i^\mathsf{T}\hat{\beta}$ it does not apply to y_i.

10.3 $\mathbb{E}(Y) \neq X\beta$

The expectation of Y may not be linear in X. At the individual datapoint level it might hold that

$$\mathbb{E}(y_i) = \boldsymbol{x}_i^\mathsf{T}\beta + \boldsymbol{w}_i^\mathsf{T}\gamma$$

so $\mathbb{E}(Y) = X\beta + W\gamma$ for a collection of missing variables that we call W. It is also possible that the true relationship between $\boldsymbol{x}_i$ and $\mathbb{E}(y_i)$ is nonlinear.

One common explanation of the ε_i in a regression model is that it represents the effect of many unmeasured variables that are very weakly correlated with $\boldsymbol{x}_i$, or in an idealization, uncorrelated with $\boldsymbol{x}_i$. But perhaps there are a small number of unmeasured variables with important correlations with $\boldsymbol{x}_i$. It is clear that there cannot be 1,000,000 such important variables; they would on average contribute only 10^{-6} of their combined total variance explained. This point is studied in Tosh et al. (2025), which seems surprisingly late; how did people not know this in the 1960s or earlier? So we think in terms of low dimensional $\boldsymbol{w}_i$.

Suppose that

$$Y = X\beta + W\gamma + \boldsymbol{\varepsilon} \tag{10.1}$$

where in the above $\boldsymbol{\varepsilon}$ really has mean zero given X and W, which implies that $\mathrm{cor}(\varepsilon_i, \boldsymbol{x}_i) = \mathbf{0}$ and $\mathrm{cor}(\varepsilon_i, \boldsymbol{w}_i) = \mathbf{0}$. If we don't know about W and fit OLS anyway our coefficient estimate is

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(X\beta + W\gamma + \boldsymbol{\varepsilon})$$

and it has expected value

$$\beta + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}W\gamma. \tag{10.2}$$

We generally get a bias unless $X^{\mathsf{T}}W\gamma = \mathbf{0}$. For example if the missing variables W are uncorrelated with X then we don't get a bias. Going with Meehl's 'everything is correlated with everything' viewpoint, the bias will not be zero but it can still be small if W and X are only weakly correlated or if the effect $W\gamma$ of those missing variables is small.

If we are missing an important variable with some association with X then we get an important bias. The consequence of this bias is that we do not estimate β from (10.1) but are instead estimating the quantity in (10.2). If the model in (10.1) is causal in both $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ then our regression will not correctly estimate the causal effect of $\boldsymbol{x}_i$. That is, it will not generalize to settings where we would intervene to change $\boldsymbol{x}$. If the future values $(\boldsymbol{x}_i, \boldsymbol{w}_i, y_i)$ are drawn from the same distribution as our past values then the best linear predictor of past y_i from past $\boldsymbol{x}_i$ can generalize to future linear (non-causal) predictions.

10.4 Added variable plots

Now let's suppose that $W \in \mathbb{R}^{n\times 1}$, so there is just one missing variable, and it is actually equal to one of the variables that we measured but for some reason did not include in our regression. Let's call it X_{p+1}.

We can judge whether X_{p+1} is important by including it in our regression and looking at the reduction in RSS. There is also a graphical way to study the issue. The simplest way is to plot $\hat{\varepsilon}_i$ versus $x_{i,p+1}$ and look for a linear trend.

That simple plot is not the best plot. To understand why, write

$$X_{p+1} = \hat{X}_{p+1} + \hat{\boldsymbol{\varepsilon}}_{p+1}$$

where

$$\begin{aligned}\hat{X}_{p+1} &= X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}X_{p+1} = HX_{p+1} \quad \text{and}\\ \hat{\boldsymbol{\varepsilon}}_{p+1} &= (I-H)X_{p+1}.\end{aligned}$$

We know from our least squares lecture that $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to all of the columns of X. That makes it orthogonal to every column of H. Then $\hat{X}_{p+1}^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}} = X_{p+1}^{\mathsf{T}}H\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$. So X_{p+1} contains a contribution $\hat{\boldsymbol{\varepsilon}}_{p+1}$ that might be correlated with $\hat{\boldsymbol{\varepsilon}}$ plus a contribution $\hat{X}_{p+1}$ that cannot possibly be correlated with $\hat{\boldsymbol{\varepsilon}}$. If X_{p+1} has additional predictive power for Y beyond what we get from X it must be coming from $\hat{\boldsymbol{\varepsilon}}_{p+1}$. Adding $\hat{X}_{p+1}$ can only blur the point cloud. Therefore a better plot for us is the 'added variable plot' of $\hat{\boldsymbol{\varepsilon}}$ versus $\hat{\boldsymbol{\varepsilon}}_{p+1}$.

We might wonder what the plot gives us that an F test for inclusion of X_{p+1} does not give us. The F test is sensitive to a linear relationship between $\hat{\boldsymbol{\varepsilon}}_{p+1}$ and $\hat{\boldsymbol{\varepsilon}}$. A graph could show us a curved relationship or a step function relationship. It might show us a pattern where the n data points appear to follow clusters. Patterns like that could trigger some insights that the F test does not. The graph also has units of y_i versus units of $x_{i,p+1}$ which aid interpretation.

Following those insights and deciding whether or not to include $x_{i,p+1}$ is now a 'researcher degree of freedom'. It takes us out of the pre-specified variable analysis that underlies our hypothesis tests and other inferences. There is no perfect way to account for this. One thing we can do is test several predictions on a holdout data set.

The values of $x_{i,p+1}$ might be functions of some other predictors such as $(x_{i,j} - \bar{x}_{\bullet,j})^2$ to explicitly detect some curvature or $(x_{i,j} - \bar{x}_{\bullet,j})(x_{i,j'} - \bar{x}_{\bullet,j'})$ for $j \neq j'$ to look for an interaction.

10.5 $\mathrm{var}(Y) \neq \sigma^2 I$ and generalized least squares

Our model assumes that y_i have equal variance and that they are uncorrelated. Both of those can easily be unrealistic. Several other models have very plausible alternatives where $\mathrm{var}(Y) = \sigma^2 V$ where $V \in \mathbb{R}^{n\times n}$ is a known matrix, or, depends on a small number of parameters. There are a few examples below where we can explain how one might actually know V.

Suppose that $y_i = (1/m_i)\sum_{j=1}^{m_i} y_{ij}$ so y_i is a group average of some other values. We might not have those other values, just y_i and also m_i. In that case we might take

$$V = \begin{pmatrix} \frac{1}{m_1} & & & \\ & \frac{1}{m_2} & & \\ & & \ddots & \\ & & & \frac{1}{m_n} \end{pmatrix}$$

as a model that averages of more observations have lower variance, where blank entries are zero. This variance model is correct when y_{ij} are uncorrelated with common variance σ^2 and a model can be useful even if it is not quite correct (as Box has told us).

In some circumstances we can reason that observations close together have some kind of correlation. They could be close in time or close in some other ordering like the position of a plant in a row. In this case a variance model like

$$V = \begin{pmatrix} 1 & \rho & & & & \\ \rho & 1 & \rho & & & \\ & \rho & 1 & \rho & & \\ & & \ddots & \ddots & \ddots & \\ & & & \rho & 1 & \rho \\ & & & & \rho & 1 \end{pmatrix} \tag{10.3}$$

might be appropriate. For this model

$$\operatorname{cor}(y_i, y_{i'}) = \begin{cases} 1, & i = i' \\ \rho, & i = i' \pm 1 \\ 0, & \text{else.} \end{cases}$$

Another model for this kind of ordered correlation has

$$\operatorname{cor}(y_i, y_{i'}) = \rho^{|i-i'|} \tag{10.4}$$

for some $\rho \in (-1, 1)$. In practice we would probably have to estimate ρ. One natural estimate is

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}.$$

Note that the numerator with $n-1$ terms is still divided by n. This is the customary estimate in time series analysis. A course on time series is the place to consider the best way to estimate $\operatorname{cor}(Y)$ in problems like this. For our present purposes, we just use equations (10.3) and (10.4) as well-motivated specific examples of $V \neq I$. It is important to recognize when a time series issue is relevant, even if we have not yet studied those solutions.

For the one way random effects model we get a block diagonal matrix

$$V = \begin{pmatrix} B_1(\rho) & & & & \\ & B_2(\rho) & & & \\ & & B_3(\rho) & & \\ & & & \ddots & \\ & & & & B_k(\rho) \end{pmatrix}$$

where block i is

$$B_i(\rho) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n_i \times n_i} \tag{10.5}$$

that is it equals $I(1-\rho) + \rho \mathbf{1}\mathbf{1}^{\mathsf{T}}$. In that setting $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$ and we discussed estimates of those variance components.

A very difficult situation arises when

$$\operatorname{cor}(y_i, y_{i'}) = \begin{cases} 1, & i = i' \\ \rho, & i \neq i'. \end{cases}$$

Then our whole dataset has the equicorrelation structure of just one block like the ones in (10.5). A setting like this arises in political polling. The polls in a collection of k states might all have positively correlated errors. If our data are $y_i = \mu + a + \varepsilon_i$ for random a independent of some IID ε_i then the data do not let us separately estimate μ and a. Any attempt to reason about this setting requires us to make use of other hopefully relevant sources of data.

Now we assume that $Y \sim \mathcal{N}(X\beta, \sigma^2 V)$ for a known $V \in \mathbb{R}^{n\times n}$ of full rank and we develop the appropriate solution. First let

$$V = P^{\mathsf{T}} \Lambda P$$

where $P \in \mathbb{R}^{n\times n}$ is orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ for $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n > 0$. ***Exercise:*** which if any of the examples above do not satisfy this condition?

Now we let $D = \Lambda^{-1/2} P$. We easily find that $D^{\mathsf{T}} D = V^{-1}$, so D is a matrix square root of V^{-1}. With D in hand we now write

$$DY = DX\beta + D\boldsymbol{\varepsilon}$$

which we rewrite as

$$\tilde{Y} = \tilde{X}\beta + \tilde{\boldsymbol{\varepsilon}}. \tag{10.6}$$

This expression has a new X and a new Y and a new $\boldsymbol{\varepsilon}$ but it also has the original β. Next using $\Lambda^{\mathsf{T}} = \Lambda$

$$\begin{aligned} \operatorname{var}(\tilde{\boldsymbol{\varepsilon}}) &= DVD^{\mathsf{T}}\sigma^2 \\ &= \Lambda^{-1/2} P P^{\mathsf{T}} \Lambda P P^{\mathsf{T}} \Lambda^{-1/2} \sigma^2 \\ &= \sigma^2 I. \end{aligned}$$

We now know how to estimate β. We can use ordinary least squares (OLS) on the model in equation (10.6). We call this generalized least squares (GLS) and the GLS parameter estimate is

$$\hat{\beta}_{\mathrm{GLS}} = (\tilde{X}^{\mathsf{T}}\tilde{X})^{-1}\tilde{X}^{\mathsf{T}}\tilde{Y} = (X^{\mathsf{T}}V^{-1}X)^{-1}X^{\mathsf{T}}V^{-1}Y.$$

We know that $\hat{\beta}_{\mathrm{GLS}}$ is the best linear unbiased estimate of β. This follows from the Gauss-Markov theorem applied to equation (10.6).

GLS must obviously have $\mathbb{E}(\hat{\beta}_{\mathrm{GLS}}) = \beta$ because it is OLS for equation (10.6). We can also find with some algebra that

$$\mathrm{var}(\hat{\beta}_{\mathrm{GLS}}) = \cdots = (X^{\mathsf{T}}V^{-1}X)^{-1}\sigma^2.$$

NB: you should be able to fill in the above dots. Also what are $\mathbb{E}(\hat{\beta}_{\mathrm{OLS}})$ and $\mathrm{var}(\hat{\beta}_{\mathrm{OLS}})$?

There are two consequences of using the wrong variance model. The first, from Gauss-Markov is that our OLS estimates are inefficient having higher variance than the GLS estimates we might otherwise have used. We can quantify this inefficiency by the ratio

$$\frac{\mathrm{var}(\boldsymbol{c}\hat{\beta}_{\mathrm{OLS}})}{\mathrm{var}(\boldsymbol{c}\hat{\beta}_{\mathrm{GLS}})}$$

for a vector $\boldsymbol{c} \in \mathbb{R}^{1\times p}$. This inefficiency depends on $\boldsymbol{c}$. So it can affect every different $\hat{\beta}_j$ differently or every different prediction $\boldsymbol{x}_i^{\mathsf{T}}\hat{\beta}$ differently. If the inefficiency equals 2 then we get the same variance from OLS as somebody else would get from GLS with only $n/2$ observations. It is like we are throwing away data.

There is another potentially much more serious problem than inefficiency. It is naivete that comes from underestimating our variance. We can quantify the naivete of OLS via the ratio

$$\frac{\mathrm{var}_{\mathrm{GLS}}(\boldsymbol{c}\hat{\beta}_{\mathrm{OLS}})}{\mathrm{var}_{\mathrm{OLS}}(\boldsymbol{c}\hat{\beta}_{\mathrm{OLS}})}$$

where $\mathrm{var}_{\mathrm{GLS}}(\boldsymbol{c}\hat{\beta}_{\mathrm{OLS}})$ is the variance of the OLS estimate under the GLS model and $\mathrm{var}_{\mathrm{OLS}}(\boldsymbol{c}\hat{\beta}_{\mathrm{OLS}})$ is the expected variance we would get for the OLS estimate using the OLS model.

If we are naive by a factor of 2 then the real variance is twice what we thought. A simple way to be naive by a factor of k is to copy-paste our data set of n rows $k-1$ times getting nk rows and then analyze all nk data points under a model that was appropriate for the original n points.

The next few paragraphs are about looking for an expression for the naivete ratio. It's a personal interest of mine and I don't see it in many other places. For these notes and this course it is enough to know that OLS can give us a severe underestimation of the variance. You may skip them and rejoin the conversation at the case of diagonal V below.

From plain algebra we can find that

$$\begin{aligned}\mathrm{var}_{\mathrm{GLS}}(\boldsymbol{c}\hat{\beta}_{\mathrm{OLS}}) &= \mathrm{var}_{\mathrm{GLS}}(\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(X\beta+\boldsymbol{\varepsilon}))\\ &= \mathrm{var}_{\mathrm{GLS}}(\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{\varepsilon})\\ &= \boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}VX(X^{\mathsf{T}}X)^{-1}\boldsymbol{c}^{\mathsf{T}}\sigma^2.\end{aligned}$$

It is a bit trickier to work out what variance OLS would have given us for the OLS estimate when the data actually followed GLS. The formula would take the form $\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}\boldsymbol{c}^{\mathsf{T}}s^2$, but what s^2 would we have gotten? For $n \gg p$ we can simplify s^2 to $\hat{\sigma}^2 = \mathrm{RSS}/n$. Then

$$\begin{aligned}\mathbb{E}_{\mathrm{GLS}}(\hat{\sigma}^2) &= \frac{1}{n}\mathbb{E}_{\mathrm{GLS}}((Y-X\hat{\beta}_{\mathrm{OLS}})^{\mathsf{T}}(Y-X\hat{\beta}_{\mathrm{OLS}}))\\ &= \frac{1}{n}\mathbb{E}_{\mathrm{GLS}}(Y^{\mathsf{T}}(I-H)Y)\\ &= \frac{1}{n}\mathbb{E}_{\mathrm{GLS}}(\boldsymbol{\varepsilon}^{\mathsf{T}}(I-H)\boldsymbol{\varepsilon})\\ &= \frac{1}{n}\mathrm{tr}((I-H)V)\sigma^2.\end{aligned}$$

The naivete ratio is then

$$\frac{\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}VX(X^{\mathsf{T}}X)^{-1}\boldsymbol{c}^{\mathsf{T}}}{\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}\boldsymbol{c}^{\mathsf{T}}\mathrm{tr}((I-H)V/n)}$$

after canceling σ^2 from numerator and denominator.

This is the place to rejoin the narrative about variances other than $\sigma^2 I$.

Correlations can be so devastating to our inferences because they contribute $n(n-1)$ terms to $\mathrm{var}(\theta^{\mathsf{T}}Y)$. Even small correlations can have a big effect. Now let's look at the case where $\mathrm{var}(Y) = \mathrm{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_n^2)\times\sigma^2$ for unknown σ^2 and known τ_i^2. This setting is called heteroscedastic as distinct from homoscedastic where $\mathrm{var}(y_i) = \sigma^2$ for all $i = 1, \ldots, n$.

For these unequal variances we get

$$\hat{\beta}_{\mathrm{GLS}} = \left(\sum_{i=1}^{n}\tau_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\right)^{-1}\sum_{i=1}^{n}\tau_i^{-2}\boldsymbol{x}_i y_i$$

Note the inverse variance weighting. A somewhat common error is to weight by τ_i^{-1}. A good way to remember the proper weighting is to reconsider the case where $\tau_i^2 = 1/m_i$ above. Then inverse variance weighting weights the data proportionally to the counts m_i.

We can detect some of the problems in this section by plotting residuals. A common phenomenon is that $\mathrm{var}(y_i)$ is an increasing function of $\mathbb{E}(y_i)$. For example with large Poisson counts (that are then approximately normal), $\mathrm{var}(y_i) = \mathbb{E}(y_i)$. We can detect patterns like this by plotting $\hat{\varepsilon}_i$ versus $\hat{y}_i$.

In a simple regression of a scalar y_i on a scalar z_i (so $\boldsymbol{x}_i = (1, z_i)^\mathsf{T}$) we may see a fan-shaped pattern in a plot of $\hat{\varepsilon}_i$ versus z_i. That suggests a model with $\tau_i^2 = z_i^a$ for some power $a > 0$. We can estimate that power by maximum likelihood in a model with

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 z_i, \sigma^2 z_i^a).$$

10.6 Outliers

An outlier is a data point that is far from all of the others. An outlier can have a devastating effect on the parameter estimate $\hat{\beta}$.

In the bodyfat data we saw one observation for a person ostensibly of height 29.5 inches and weight 205 pounds. That data point was far from the others and not believable. When we find an outlier and can explain it as an error then it is straightforward to simply remove it (and disclose why we have done that). In other settings we might be convinced that a point is bad but not have an explanation for why it was bad. Deleting it may still be the wise option but it should be disclosed.

We can detect some outliers by observing large values of $|\hat{\varepsilon}_i|$. This is not always reliable. The regression fits were found by minimizing $\sum_{i=1}^n \hat{\varepsilon}_i^2$ and so they may have led to $\hat{\beta}$ where $|\hat{\varepsilon}_i|$ for the outlier was not especially large.

We can do better by inspecting leave-one-out residual $|y_i - \hat{y}_{-i}|$. Taking $(\boldsymbol{x}_i, y_i)$ out of the fitting process stops that observation from pulling the fit towards itself. That isn't perfectly reliable either. There could be a collection of similar outliers that all pull the regression in the same direction. When we delete one of them the effect of the others is still to give that one only a modestly large leave-one-out residual. This phenomenon is called ***masking***. Each outlier helps to hide the others. If there are k such outliers 'working together' we might have to consider dropping $\binom{n}{k}$ different subsets of k data points to find them. This is hard because we probably don't know which value of k to use and large k will be too expensive.

A related problem is ***swamping***. A collection of outliers can distort the regression model so much that some other points that we think of as good points get large absolute residuals. Then good points might get treated as outliers.

In class I described some computationally intensive ways to get robust linear regression models. Those were the least median of squares algorithm and the least trimmed squares algorithm. Both are robust to outliers but the first one is statistically inefficient.

When we just have to estimate a scalar 'center' of a distribution the usual choice is the sample mean and a robust alternative is the sample median. It is robust because enormous changes to a few of the data values generally do not move the median by much. The mean minimizes $\sum_{i=1}^n (y_i - m)^2$ over m and the median minimizes $\sum_{i=1}^n |y_i - m|$ over m. This suggests an alternative regression criterion namely L_1 regression choosing β to minimize

$$\sum_{i=1}^n |y_i - \boldsymbol{x}_i^\mathsf{T}\beta|.$$

Unfortunately this estimate is not robust even for the simple case of $\boldsymbol{x}_i = (1, z_i)^\mathsf{T}$ for $z_i \in \mathbb{R}$. You can verify this by generating a point cloud of (z_i, y_i) pairs following a regression line and then adding one point with very large $|z_i|$ and y_i far from the regression line.

Bibliography

C. Tosh, P. Greengard, B. Goodrich, A. Gelman, A. Vehtari, and D. Hsu. The piranha problem: Large effects swimming in a small pond. *Notices of the American Mathematical Society*, 72(1):15–25, 2025. doi: 10.1090/noti3044.