Violation of Error Assumptions: Non-constant variance

Stats 203, Winter 2024

Lecture 10 [Last update: February 24, 2024]

1 Non-constant variance of errors

We will now begin to relax the assumptions of the linear model with Normal errors, starting with the assumption that the errors have equal variance. That is, we now assume that: $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled independently, with

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i | X_i = x \sim N(0, \sigma^2(x)). \tag{1}$$

Notice that ϵ_i is no longer independent of X_i since $\mathrm{Var}[\epsilon_i | X_i = x] = \sigma^2(x)$. Although not clear from the notation, it is also possible that the variance can depend on variables besides the predictors. In statistics, the inscrutable but fun-to-say word *heteroskedastic* is sometimes used to describe non-constant variance. Heteroskedasticity should be checked by making plots of the residuals, as seen in the data analysis example below.

Here are some common reasons why the errors may not have constant variance.

- *Response is an average of independent trials.* Suppose each response is actually an average of n_i independent random variables: $Y_i = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik}$, where Y_{ik} are unobserved. Then if each Y_{ik} follows the usual linear model, the response Y_i that we actually observe is heteroskedastic, with variance $\sigma_i^2 = \frac{\sigma^2}{n_i}$. Pooling observations in this way, and then running a regression on the pooled responses, is sometimes called **ecological inference**.
- *Response is a positive number.* If Y_i must be positive (or at least, non-negative), then the variance of ϵ_i will likely be smaller if $m(x) = \mathbb{E}[Y_i | X_i = x]$ is close to zero.

  Of course, in this example the errors cannot be Normal, since then Y_i might be negative. But one issue at a time.
- *Response is a count.* A special kind of positive number is a count. For many of the distributions we use to model counts, there is an intricate relationship between the mean and variance. For instance, in Binomial regression where $Y_i | X_i = x \sim \mathrm{Bin}(n_i, p(x))$, we have $\mathrm{Var}(Y_i | X_i = x) = n_i p(x)(1 - p(x))$. In Poisson regression, if $Y_i | X_i = x \sim \mathrm{Pois}(m(x))$ then $\mathrm{Var}(Y_i | X_i = x) = m(x)$.

When the errors have unequal variances, then our formulas for the standard error of $\mathrm{Var}[\hat{\beta} | \mathbf{X}]$ are incorrect, and our hypothesis tests and confidence intervals may be off. Even more basically, there will be more accurate estimators than OLS. We start with the second issue first.

2 Best linear unbiased estimator

To understand how to improve on OLS when error variances are unequal, we need to start by showing that, in a certain sense, we cannot improve on OLS when error variance are equal. We have shown previously that if the linear model with all of its usual assumptions is correct, then OLS is unbiased. While it is not the only unbiased estimator, it turns out that it is the *best* linear unbiased estimator, in the sense of having the smallest standard errors. We work through the details.

An estimator $\tilde{\beta}(\mathbf{X}, Y)$ of β is called **linear** if it is a linear transformation of Y given $\mathbf{X}$: $\tilde{\beta} = \tilde{\mathbf{A}}Y$. Here $\tilde{\mathbf{A}} \in \mathbb{R}^{(p+1)\times n}$ that can depend on $\mathbf{X}$. For example, we have shown that OLS is a linear estimator of β since $\hat{\beta}_{\text{OLS}} = \mathbf{A}_{\text{OLS}}Y$ with $\mathbf{A}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

Suppose now a linear estimator $\tilde{\mathbf{A}}Y$ is unbiased for β. This means that, *for all* $\beta \in \mathbb{R}^{p+1}$:

$$\begin{aligned}\beta &= \mathbb{E}[\tilde{\mathbf{A}}Y|\mathbf{X}]\\ &= \tilde{\mathbf{A}}\mathbb{E}[Y|\mathbf{X}]\\ &= \tilde{\mathbf{A}}\mathbf{X}\beta.\end{aligned}$$

Since this holds for all β, it must be the case that $\tilde{\mathbf{A}}\mathbf{X} = \mathbf{I}_{p+1}$. You can check that this is indeed true when $\tilde{\mathbf{A}} = \mathbf{A}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. More broadly, it holds if and only if $\tilde{\mathbf{A}} = \mathbf{A}_{\text{OLS}} + \mathbf{C}$ for some matrix $\mathbf{C} \in \mathbb{R}^{(p+1)\times n}$ such that $\mathbf{C}\mathbf{X} = 0$.

Now we compute the covariance matrix of any such unbiased estimator:

$$\begin{aligned}\text{Cov}[\tilde{\beta}|\mathbf{X}] = \text{Cov}[\tilde{\mathbf{A}}Y|\mathbf{X}] = \tilde{\mathbf{A}}\text{Cov}[Y|\mathbf{X}]\tilde{\mathbf{A}}^\top &= \sigma^2\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top\\ &= \sigma^2(\mathbf{A}_{\text{OLS}} + \mathbf{C})(\mathbf{A}_{\text{OLS}} + \mathbf{C})^\top\\ &= \sigma^2\Big(\mathbf{A}_{\text{OLS}}(\mathbf{A}_{\text{OLS}})^\top + \mathbf{C}\mathbf{C}^\top + \mathbf{A}_{\text{OLS}}\mathbf{C}^\top + \mathbf{C}(\mathbf{A}_{\text{OLS}})^\top\Big)\\ &= \sigma^2\Big(\mathbf{A}_{\text{OLS}}(\mathbf{A}_{\text{OLS}})^\top + \mathbf{C}\mathbf{C}^\top + \Big(\mathbf{C}(\mathbf{A}_{\text{OLS}})^\top\Big)^\top + \mathbf{C}(\mathbf{A}_{\text{OLS}})^\top\Big)\\ &= \sigma^2\Big(\mathbf{A}_{\text{OLS}}(\mathbf{A}_{\text{OLS}})^\top + \mathbf{C}\mathbf{C}^\top + \Big(\mathbf{C}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)^\top + \mathbf{C}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)\\ &= \sigma^2\Big(\mathbf{A}_{\text{OLS}}(\mathbf{A}_{\text{OLS}})^\top + \mathbf{C}\mathbf{C}^\top\Big)\\ &= \sigma^2\Big((\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{C}\mathbf{C}^\top\Big)\end{aligned}$$

Of course, we recognize the first term as the covariance matrix of ordinary least squares. The second term $\mathbf{C}\mathbf{C}^\top$ is a symmetric, positive semi-definite matrix. So in particular it has positive diagonal elements, and thus $\text{Var}[\tilde{\beta}_j|\mathbf{X}] \geq \text{Var}[\hat{\beta}_j|\mathbf{X}]$. For at least one $j = 0, \ldots, p$, the inequality will be strict, unless $\mathbf{C} = \mathbf{0}_{p\times n}$.

In other words, under the usual assumptions of the linear model, ordinary least squares is the best (in the sense of having lowest variance) linear unbiased estimator; sometimes abbreviated to BLUE. This is also known as the **Gauss-Markov** theorem. Notice that it did not rely on the errors being Normal, but it did rely on $\text{Cov}[Y|\mathbf{X}] = \sigma^2\mathbf{I}_n$.

3 Weighted least squares

We return now to the case of errors with non-constant variance. Now, ordinary least squares is still unbiased and linear but it is no longer best. What replaces OLS depends on whether the variance of the errors is known or unknown.

Known error variance. Assume for the moment $\text{Var}[\epsilon_i|X_i] = \sigma^2(X_i)$ were a known quantity. Intuitively, we would like an estimator that gives more weight to observations with less variance, and less weight to observations with more variance. **Weighted least squares** does exactly this, with this specific choice of weights $w_i = 1/\sigma^2(X_i)$:

$$\hat{\beta}_{\text{WLS}} = \underset{b}{\text{argmin}} \sum_{i=1}^{n} w_i\Big(y_i - b_0 + \sum_{j=1}^{p} b_j x_{ij}\Big)^2, \quad \text{where } w_i = \frac{1}{\sigma^2(X_i)}. \tag{2}$$

The solution to (2) is the weighted least squares estimate

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}Y, \quad \mathbf{W} = \text{diag}\Big(\frac{1}{\sigma^2(X_i)}\Big) \tag{3}$$

(The notation diag means $\mathbf{W} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal entries given by $1/\sigma^2(X_i)$.)

This particular choice of weights makes $\hat{\beta}_{\text{WLS}}$ the best linear unbiased estimator for the linear model in (1). This can be shown using similar steps to our derivation for OLS in the constant variance case. Likewise, it is not hard to carry over most of the theory for OLS – e.g. sampling distributions, standard errors, confidence intervals, and hypothesis tests – to WLS, but we will not do that explicitly.

Unknown error variance. We continue to allow for errors with different variances but now suppose that the error variances are *unknown*. We would like to do better than OLS, but WLS is no longer an option as it depends on unknown parameters. In fact, this is quite a hard problem as we have more parameters than observations: n observations but $p+1+n$ parameters. In general the problem is unsolvable. But in certain special cases $\sigma^2(x)$ is a simple enough function of x that we can estimate it from data.

If we are lucky we might have good estimates $\hat{\sigma}_i^2$ for each $\sigma^2(X_i)$. In this case the obvious thing to do is to solve WLS by plugging in our estimates into the weights. This leads to the solution

$$\hat{\beta}_{\text{FLS}} := (\mathbf{X}^\top \widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^\top \widehat{\mathbf{W}}Y, \quad \widehat{\mathbf{W}} = \text{diag}\Big(\frac{1}{\hat{\sigma}_i^2}\Big).$$

This is a special case of **feasible least squares** which is where the subscript "FLS" comes from. Introducing the plug-in approximation $\widehat{W}$ makes feasible least squares biased in general, so that it is no longer BLUE. But the point estimates can be still be very useful in practice.

How to estimate $\sigma^2(x)$? Estimating $\sigma^2(x)$ opens up a second modeling frontier. We will not have time to do the topic justice. Instead we just look at couple of common cases.

- *Counts.* In the Binomial/Poisson examples above, $\sigma^2(x)$ was a simple function of the conditional mean. For the Binomial case, $\sigma^2(x) = n_i p(x)(1-p(x))$. For the Poisson case, $\sigma^2(x) = m(x)$. So if we have a good estimate for the conditional mean – say, the linear regression estimate $\hat{m}(x) = x^\top \hat{\beta}$ – that means we have a good estimate for the conditional variance.
- *Linear in the predictors.* If Y_i is positive, then it is often reasonable to model the error variance as being linear in the predictors: $\sigma^2(x) = \sigma^2(x) = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p$. For example, this will be the case if $\sigma^2(x)$ a linear function of the $m(x)$, since composition of linear functions are linear.

  To estimate $\sigma^2(x)$ in this case, notice that if we observed the errors ϵ_i this would just become a linear modeling problem, since

  $$\mathbb{E}[\epsilon_i^2 | X_i = x] = \text{Var}[\epsilon_i | X_i = x] = \sigma^2(x) = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p.$$

  Of course we do not estimate the errors. Instead, the typical approach is to substitute the OLS residuals e for the errors ϵ, and estimate $\sigma^2(x)$ by minimizing

  $$\sum_{i=1}^{n} \Big(e_i - a_0 - \sum_{j=1}^{p} a_j X_{ij}\Big)^2.$$

In both of these cases, notice that we needed to first compute the OLS solution, and then use its fitted values (in the first case) or residuals (in the second case) to estimate $\sigma^2(x)$. Methods which work in this way are called **reweighted least squares**, and generally speaking proceed according to the following steps:

1. Compute the ordinary least squares solution $\hat{\beta}_{\text{OLS}}$.
2. Estimate $\hat{\sigma}^2(x)$ using the fitted values $\hat{y}_i = \hat{m}(x_i)$ and residuals e_i of the OLS solution.

3. Run feasible least squares with weights $\widehat{\mathbf{W}} = \text{diag}(\frac{1}{\hat{\sigma}^2(X_i)})$, producing $\hat{\beta}_{\text{FLS}}$.

Note that at the end of Step 3 we have a new estimate $\hat{\beta}_{\text{FLS}}$ of β which is (hopefully) better than $\hat{\beta}_{\text{OLS}}$. Using the fitted values and residuals of this estimator in Step 2 could then lead to better estimates of the error variance, which in turn would lead to more accurate estimated weights in feasible least squares, etc. We can take advantage of this by iterating Steps 2 and 3 in the above algorithm for a few steps, if desired. In this case the algorithm is called **iteratively reweighted least squares**.

4 Sampling distribution of OLS and WLS with non-constant error variance

So far we have primarily talked about the implications that heteroskedasticity has for estimation. But, of course, there are also implications for inference.

Suppose the errors have non-constant variance but we still use OLS. The estimator $\hat{\beta}_{\text{OLS}}$ is still unbiased for β. But the covariance matrix of $\hat{\beta}_{\text{OLS}}$ is now

$$\text{Cov}[\hat{\beta}_{\text{OLS}}|\mathbf{X}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Sigma}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}. \tag{4}$$

So our previous expressions for the standard error $\text{SE}[\hat{\beta}_j|\mathbf{X}]$ are wrong, and as a result so are our Wald intervals and T-tests, and F-tests.

If we know the (unequal) error variances, we can plug them in to (4) to figure out the right covariance for $\hat{\beta}_{\text{OLS}}$. If we believe we have good estimates of the error variances, we could plug these instead. Later, we will discuss a better method for computing confidence intervals based on OLS, called the bootstrap, which automatically accounts for unequal error variances.

On the other hand, if we compute weighted least squares $\hat{\beta}_{\text{WLS}}$ with *the correct weights* $\mathbf{W} = \text{diag}(1/\sigma^2(X_i))$, then

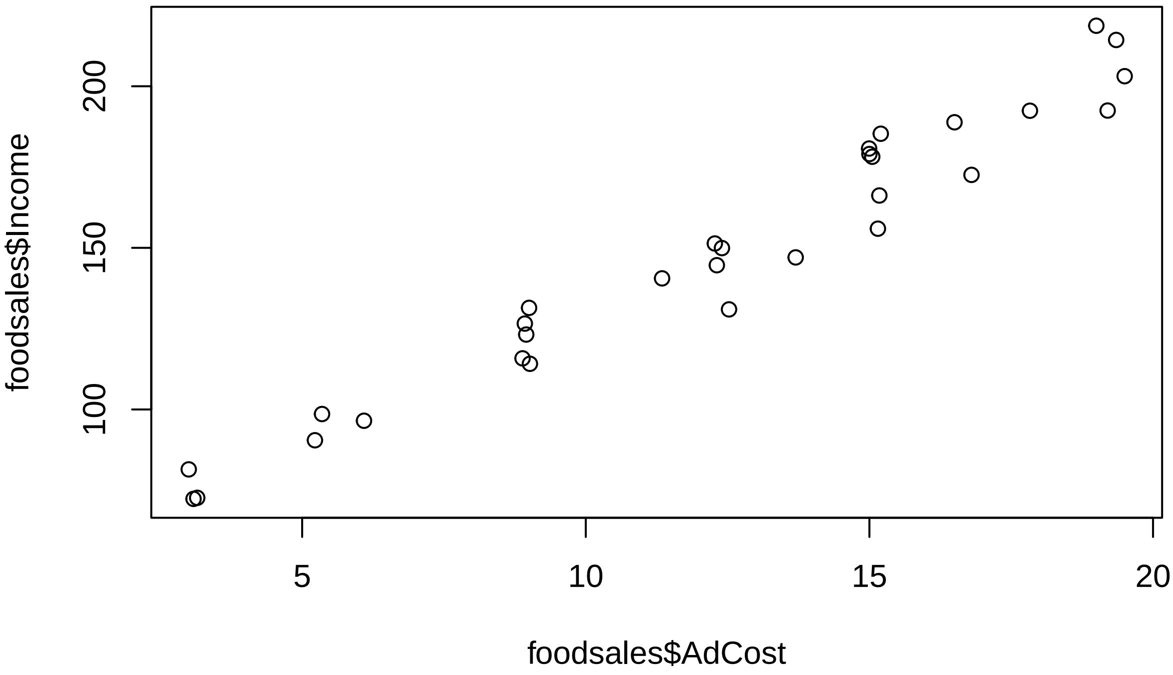
$$\text{Cov}[\hat{\beta}_{\text{WLS}}|\mathbf{X}] = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}.$$

When we run weighted least squares in R, the resulting standard errors, T-statistics, etc. are based on this expression for the covariance of $\hat{\beta}_{\text{WLS}}$.

Reweighted least squares: food cost data

This example is drawn from the Montgomery, Peck, and Vining textbook. We look at $n = 30$ observations of restaurant ad sales and income. We want to know if ad sales are positively associated with restaurant income. We load and plot the data and see that a linear fit is plausible.

```r
foodsales = read.table("foodsales.txt")
foodsales = foodsales/1000 # Put everything in terms of 1000s of dollars.
plot(foodsales$AdCost, foodsales$Income)
```



So we fit the linear model and examine the relationship between residuals and ad sales.

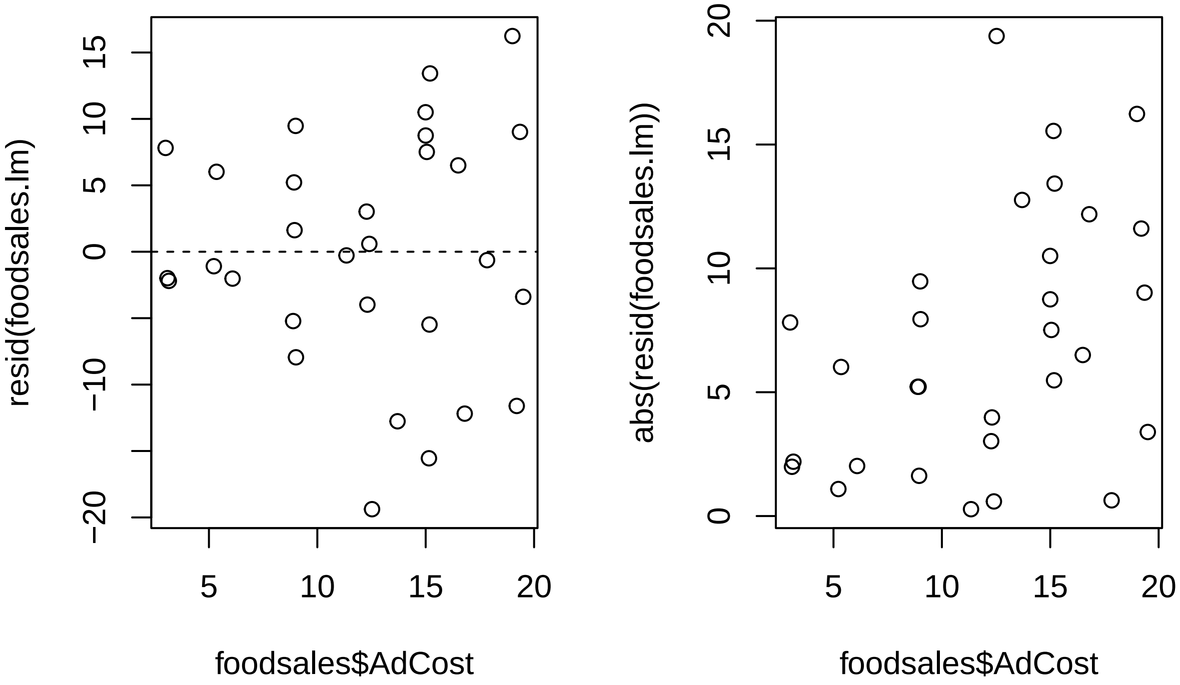
```r
foodsales.lm = lm(Income ~ AdCost, foodsales)
summary(foodsales.lm)
```

```
##
## Call:
## lm(formula = Income ~ AdCost, data = foodsales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.3795  -4.9099  -0.4557   7.2605  16.2361
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.4918     4.2845   11.55 3.64e-12 ***
## AdCost        8.0520     0.3262   24.68  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.99 on 28 degrees of freedom
## Multiple R-squared:  0.9561, Adjusted R-squared:  0.9545
## F-statistic: 609.3 on 1 and 28 DF,  p-value: < 2.2e-16
```

Now we look at residual plots to determine whether the assumption of constant error variance is plausible. Several kinds of residual plots may be helpful. On the y-axis, we could have residuals, absolute residuals, squared residuals, etc. On the x-axis, we could have the fitted values $\hat{y}$, or any of the predictors.

Here I plot the residuals and absolute residuals of the fitted model against the predictor.

```
par(mfrow = c(1,2))
plot(foodsales$AdCost,resid(foodsales.lm))
abline(h = 0,lty = 2)
plot(foodsales$AdCost,abs(resid(foodsales.lm)))
```



The residuals have a classic *fan-like* structure, where the spread in the residuals increases as a function of the predictor variable. This is very typical when the response is positive, as in this example. This suggests that the error variance $\sigma^2(x)$ is increasing as a function of x. We consider modeling the error variance as a linear function of x:

$$\sigma^2(x) = \gamma_0 + \gamma_1 x.$$

In order to fit this linear model, we use the fact that $\sigma^2(x) = \mathbb{E}[\epsilon_i^2 | X_i = x]$. Thus, the conditional mean of the squared errors is linear in x. Since we do not know the squared errors, we use the squared residuals as responses instead instead.

```
var.lm = lm(foodsales.lm$residuals^2 ~ AdCost, foodsales)
sigmahatsq = fitted(var.lm) # Our estimate of error variance
var.lm
```

```
##
## Call:
## lm(formula = foodsales.lm$residuals^2 ~ AdCost, data = foodsales)
##
## Coefficients:
## (Intercept)       AdCost
##      -7.901        6.868
```

Our estimate of error variance is $\hat{\sigma}^2(x) = -7.90 + 6.87x$, where x is AdCost. As expected, we see that our estimate of error variance is increasing with $m(x)$. Now we refit the regression model using weighted least squares, with weights $w_i = 1/\hat{\sigma}^2(X_i)$.

```
foodsales.wlm1 = lm(Income ~ AdCost, foodsales, weights = 1/sigmahatsq)
round(summary.lm(foodsales.wlm1)$coef,4)
```

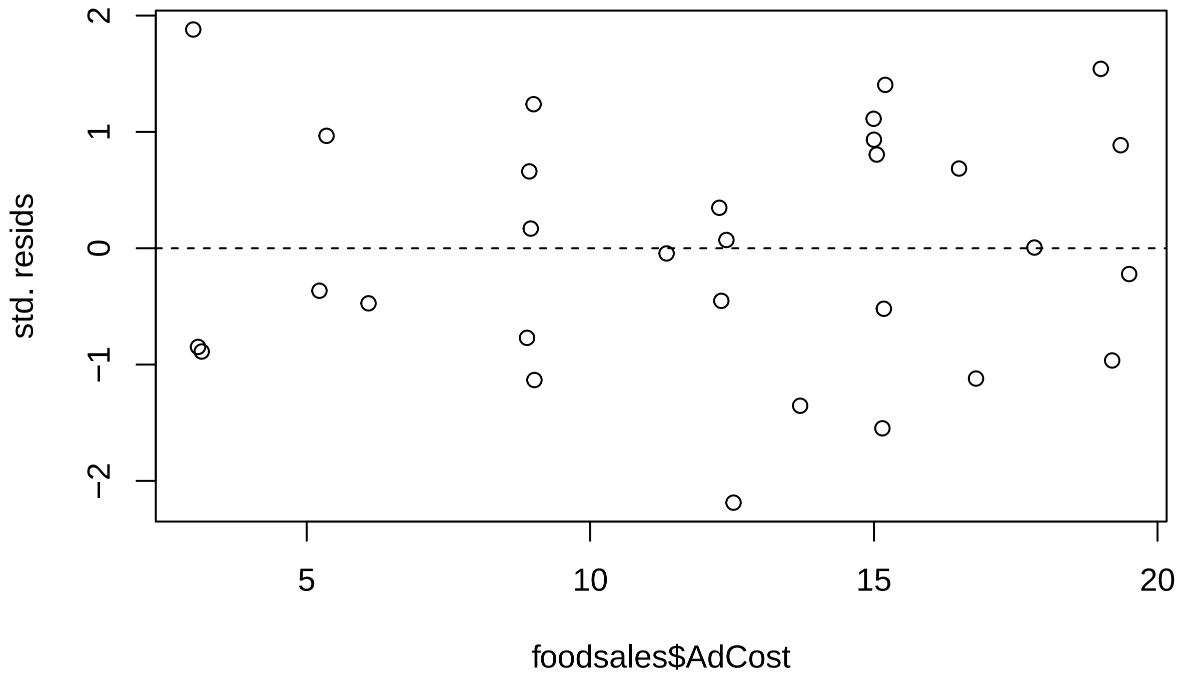
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.9727     2.4486 20.8172        0
## AdCost        7.9299     0.2514 31.5486        0
```

Compared to the unweighted model, our point estimates for intercept and slope are largely unchanged. However, our standard errors are noticeably different. In this case they are smaller, and so we would get narrower Wald intervals, but that will not always be true.

Diagnostics for weighted least squares

If we have used the right weights in WLS, then the residuals in the weighted least squares regression, normalized by our estimates of error standard deviation, should have close to equal variance. This motivates another residual plot, where on the y-axis we have the standardized residuals $e_i/\hat{\sigma}(x_i)$.

```
plot(foodsales$AdCost,resid(foodsales.wlm1)/sqrt(sigmahatsq), ylab = "std. resids")
abline(h = 0,lty = 2)
```



The standardized residuals of the new fitted regression model appear more evenly spread.

Iterated reweighted least squares

We iterate steps 2 and 3 of the reweighted least squares algorithm.

```
var.lm2 = lm(foodsales.wlm1$residuals^2 ~ AdCost, foodsales)
foodsales.wlm2 = lm(Income ~ AdCost, foodsales, weights = 1/fitted(var.lm2))
round(summary.lm(foodsales.wlm2)$coef,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.0607     2.3339 21.8782        0
## AdCost        7.9226     0.2483 31.9090        0
```

We see that both the coefficient estimates and standard errors are quite similar to the previous round of weighted least squares. So we probably don't need any subsequent iterations.