Regression And Stat Models - Assignment 7

Nathan Pasder

2025-06-12

Contents

Question 1

(a) Violation of Homoscedasticity and Variance-Stabilizing Transformation

We are given that the response variable Y is exponentially distributed, and:

$$\mathbb{E}(Y_i) = \exp(X_i^\top \beta)$$

Since $Y_i \sim \text{Exponential}(\lambda_i)$, where $\lambda_i = \exp(-X_i^\top \beta)$, we know the following properties of the exponential distribution:

- Mean: $\mathbb{E}(Y_i) = \frac{1}{\lambda_i} = \exp(X_i^\top \beta)$
- Variance: $\text{Var}(Y_i) = \frac{1}{\lambda_i^2} = \exp(2X_i^\top \beta)$

Violation of Equal Variances (Homoscedasticity)

Homoscedasticity means constant variance across all observations. Here, we see:

$$\text{Var}(Y_i) = (\mathbb{E}(Y_i))^2$$

Thus, the variance is **not constant**, but depends on the predictors. This violates the assumption of equal variances.

Applying the Delta Method

We seek a variance-stabilizing transformation $g(Y)$ such that:

$$\text{Var}(g(Y_i)) \approx \text{constant}$$

The delta method tells us:

$$\text{Var}(g(Y_i)) \approx (g'(\mu_i))^2 \cdot \text{Var}(Y_i)$$

where $\mu_i = \mathbb{E}(Y_i)$. Since $\text{Var}(Y_i) = \mu_i^2$, we want:

$$(g'(\mu_i))^2 \cdot \mu_i^2 = \text{constant}$$

Taking square roots:

$$g'(\mu_i) \cdot \mu_i = c \quad \Rightarrow \quad g'(\mu_i) = \frac{c}{\mu_i}$$

Integrating both sides:

$$g(\mu_i) = c \cdot \log(\mu_i) + C$$

We can ignore the constant C, so a suitable transformation is:

$$g(Y_i) = \log(Y_i)$$

This is a **variance-stabilizing transformation** for an exponential distribution.

(b) Linearity After the Transformation

after transforming Y_i with $g(Y_i) = \log(Y_i)$, is the expected value of the transformed variable linear in the predictors? Specifically, does the following hold?

$$\mathbb{E}(\log(Y_i)) = X_i^\top \beta$$

Recall: for a nonlinear function g, in general:

$$\mathbb{E}(g(Y_i)) \neq g(\mathbb{E}(Y_i))$$

In fact, since $Y_i \sim \text{Exponential}(\lambda_i)$, the expected value of $\log(Y_i)$ is:

$$\mathbb{E}(\log(Y_i)) = -\gamma - \log(\lambda_i) = -\gamma + X_i^\top \beta$$

where $\gamma \approx 0.577$ is the Euler-Mascheroni constant.

Therefore:

$$\mathbb{E}(\log(Y_i)) = X_i^\top \beta - \gamma$$

This means the **expectation of the log-transformed response is linear** in the predictors, up to a constant shift. So, **yes**, the linearity assumption holds **after transformation**, modulo a constant:

$$\mathbb{E}(\log(Y_i)) = X_i^\top \beta - \gamma$$

Which is still linear in X_i, and the constant γ can be absorbed into the intercept term during model fitting.

Question 2

(a)

We are given the linear regression model:

$$Y = X\beta + \epsilon, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad \text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

In vector form, define:

- $Y \in \mathbb{R}^n$ - response vector
- $X \in \mathbb{R}^{n \times p}$ - design matrix
- $\beta \in \mathbb{R}^p$ - vector of unknown coefficients
- $\epsilon \in \mathbb{R}^n$ - error vector with mean zero and covariance matrix Σ

Where:

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$$

Thus, the error terms are uncorrelated but not homoscedastic.

Expectation of the OLS Estimator

The OLS estimator is defined as:

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$$

Substituting the model $Y = X\beta + \epsilon$:

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top (X\beta + \epsilon) = \beta + (X^\top X)^{-1} X^\top \epsilon$$

Taking expectation:

$$\mathbb{E}(\hat{\beta}_{OLS}) = \beta + (X^\top X)^{-1} X^\top \mathbb{E}(\epsilon) = \beta$$

Conclusion: The OLS estimator remains **unbiased** even when errors are heteroscedastic.

Covariance Matrix of the OLS Estimator

We compute the covariance using:

$$\mathrm{Cov}(\hat{\beta}_{OLS}) = \mathrm{Cov}((X^\top X)^{-1} X^\top \epsilon) = (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1}$$

- If $\Sigma = \sigma^2 I$, this simplifies to $\sigma^2 (X^\top X)^{-1}$
- But when Σ is diagonal with unequal entries (heteroscedasticity), this expression includes the **unequal error variances**

Conclusion: Under heteroscedasticity, the OLS estimator is no longer **efficient** - i.e., it does not achieve the minimum variance among all unbiased linear estimators. Hence, it is **not BLUE** (Best Linear Unbiased Estimator).

(b) Weighted Least Squares: Full Derivation

$$S_w(b) = \sum_{i=1}^{n} w_i \left(y_i - \sum_{j=1}^{p} X_{ij} b_j \right)^2$$

$$\hat{\beta}_w = \left(X^\top W X \right)^{-1} X^\top W Y$$

Step-by-step Derivation of $\hat{\beta}_w$

Let us define:

- $W = \mathrm{diag}(w_1, w_2, \ldots, w_n)$ - a diagonal matrix with positive weights
- $S_w(\mathbf{b}) = (Y - X\mathbf{b})^T W (Y - X\mathbf{b})$

To minimize the weighted residual sum of squares $S_w(\mathbf{b})$, we take the gradient with respect to $\mathbf{b}$ and set it to zero:

$$\nabla_{\mathbf{b}} S_w(\mathbf{b}) = \nabla_{\mathbf{b}} \left(Y^T W Y - 2\mathbf{b}^T X^T W Y + \mathbf{b}^T X^T W X \mathbf{b} \right)$$

Using matrix calculus:

- $\nabla_{\mathbf{b}} \left(-2\mathbf{b}^T X^T W Y \right) = -2X^T W Y$
- $\nabla_{\mathbf{b}} \left(\mathbf{b}^T X^T W X \mathbf{b} \right) = 2X^T W X \mathbf{b}$

Set gradient to zero:

$$-2X^TWY + 2X^TWX\hat{\beta}_w = 0 \Rightarrow X^TWX\hat{\beta}_w = X^TWY \Rightarrow \hat{\beta}_w = (X^TWX)^{-1}X^TWY$$

This completes the **proof** that the solution to the weighted least squares problem is as claimed.

Justification for Using This Estimator

In the presence of **heteroscedasticity** - that is, when different observations have different error variances - the classical OLS estimator:

- Is still unbiased: $\mathbb{E}(\hat{\beta}_{OLS}) = \beta$
- **But is not efficient**: it does not minimize the variance among linear unbiased estimators

The WLS estimator corrects for this by:

- Giving **more weight** to observations with **low variance** (i.e., more reliable data)
- Giving **less weight** to observations with **high variance** (i.e., noisier data)

By doing so, the estimator $\hat{\beta}_w$:

- **Minimizes the variance** of the estimate among all linear unbiased estimators
- **Restores BLUE optimality** (Best Linear Unbiased Estimator)

This makes it the preferred estimation method when the heteroscedastic error structure is known or can be estimated.

(c) OLS is Not BLUE Under Heteroscedasticity

We start from the **Gauss-Markov theorem**, which states:

Under the assumptions of the classical linear model - linearity, full rank of X, zero-mean errors, **homoscedasticity**, and no autocorrelation - the OLS estimator $\hat{\beta}_{OLS}$ is the **Best Linear Unbiased Estimator (BLUE)**. That is, it has the **minimum variance** among all linear unbiased estimators.

Recap of the Model in Part (a)

We are given the model:

$$Y = X\beta + \epsilon, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

That is:

- Errors are uncorrelated
- But variances σ_i^2 differ - i.e., **heteroscedasticity**
- Hence, the covariance matrix of the errors is:

$$\mathrm{Cov}(\epsilon) = \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$$

Why OLS Is Not BLUE

The OLS estimator is:

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY$$

Its covariance matrix is:

$$\text{Cov}(\hat{\beta}_{OLS}) = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$$

However, the **optimal linear unbiased estimator** under heteroscedasticity is the **Generalized Least Squares (GLS)** estimator:

$$\hat{\beta}_{GLS} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

The Gauss-Markov theorem guarantees that $\hat{\beta}_{GLS}$ has **strictly smaller or equal variance** than any other linear unbiased estimator - including $\hat{\beta}_{OLS}$ - when $\Sigma \neq \sigma^2 I$.

Since:

- $\text{Cov}(\hat{\beta}_{GLS}) \neq \text{Cov}(\hat{\beta}_{OLS})$
- and $\text{Cov}(\hat{\beta}_{GLS}) \leq \text{Cov}(\hat{\beta}_{OLS})$ in the positive semi-definite sense

we conclude:

$$\hat{\beta}_{OLS} \text{ is not BLUE under heteroscedasticity}$$

(d) OLS Is Not BLUE Under Serial Correlation

We are given a linear model where the error terms exhibit **serial correlation** (dependence over time), not just heteroscedasticity:

$$Y = X\beta + \epsilon, \quad \text{with:}$$

$$\epsilon_1 = Z_1, \quad \epsilon_i = 0.5 Z_{i-1} + Z_i, \quad Z_i \sim \mathcal{N}(0, 1) \text{ i.i.d}$$

Understanding the Error Structure

The errors ϵ_i are constructed recursively using the sequence $\{Z_i\}$, which are standard i.i.d. normal variables. This creates **correlation between adjacent errors**:

- $\epsilon_1 = Z_1$
- $\epsilon_2 = 0.5 Z_1 + Z_2$
- $\epsilon_3 = 0.5 Z_2 + Z_3$

So, ϵ_2 depends on Z_1 and Z_2, and ϵ_3 on Z_2 and Z_3. Hence:

$$\text{Cov}(\epsilon_i, \epsilon_{i-1}) \neq 0$$

This is an example of a **moving average process of order 1 (MA(1))**.

Implications for the OLS Estimator

Recall that for the OLS estimator $\hat{\beta}_{OLS}$ to be **BLUE**, the error terms must be:

- Zero mean
- Homoscedastic
- **Uncorrelated**

Here, this last assumption is violated:

$$\text{Cov}(\epsilon_i, \epsilon_{i-1}) = 0.5 \cdot \text{Var}(Z_{i-1}) = 0.5$$

This means the **error vector has non-zero off-diagonal elements in its covariance matrix** Σ, so:

$$\text{Cov}(\epsilon) = \Sigma \neq \sigma^2 I$$

Better Estimator Exists: GLS Again

Because the errors are correlated, the OLS estimator:

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$$

is **still unbiased**, but **not efficient** - it no longer achieves the lowest variance among all linear unbiased estimators.

Instead, the **Generalized Least Squares (GLS)** estimator:

$$\hat{\beta}_{GLS} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y$$

uses the correct error structure Σ, and **is BLUE**.

Conclusion

Under the given model with **serial correlation**, the OLS estimator is not BLUE. A better (lower-variance) linear unbiased estimator exists - the GLS estimator, which accounts for the correlated structure in the errors.

$$\boxed{\hat{\beta}_{OLS} \text{ is not BLUE under the given model with serial correlation}}$$

Question 3

(1) Cook's Distance: Definition and Formula

Why Prefer the Second Formula for Cook's Distance?

The first formula for Cook's distance is defined as:

$$D_i = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2$$

This expression quantifies how much the fitted values for *all* observations change when observation i is removed from the model. However, to compute each $\hat{Y}_{j(i)}$, one must refit the regression model **excluding row** i. Since this must be done separately for each $i \in \{1, \ldots, n\}$, computing all D_i values using this formula requires n full model fits.

Each OLS model fit involves computing:

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$$

followed by:

$$\hat{Y}_{j(i)} = x_j^T \hat{\beta}_{(i)}$$

Therefore, the computational complexity is approximately:

$$\mathcal{O}(n \cdot p^2) \text{ for } X^T X \text{ per deletion, plus } \mathcal{O}(n^2) \text{ for predictions}$$

which is **very expensive** for large n or repeated analyses.

In contrast, the **second formula** for Cook's distance is:

$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \cdot \left(\frac{P_{X,ii}}{(1 - P_{X,ii})^2} \right)$$

This expression involves only:

- The residual $e_i = Y_i - \hat{Y}_i$
- The leverage $P_{X,ii} = x_i^T (X^T X)^{-1} x_i$

Both of these quantities are computed **once** during the initial OLS fit:

- Residuals: $\mathbf{e} = Y - X\hat{\beta}$
- Hat matrix diagonal: $P_X = X(X^T X)^{-1} X^T \Rightarrow P_{X,ii} = x_i^T (X^T X)^{-1} x_i$

Thus, calculating all D_i values using the second formula requires only:

- A single matrix inversion $(X^T X)^{-1}$
- A matrix-vector multiplication to get $\hat{Y} = X\hat{\beta}$
- Simple vector operations to get e_i^2 and $P_{X,ii}$

This reduces the total complexity to approximately:

$$\mathcal{O}(n \cdot p^2) + \mathcal{O}(n)$$

which is **much faster** and **does not require re-fitting the model** for each i.

(2) Show that

(a) Proving the Analytical Expression for Cook's Distance

Let:

- $\hat{Y} = X\hat{\beta}$ be the predicted vector using the full model.
- $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$ be the predicted vector using the model without observation i.

Then:

$$\|\hat{Y} - \hat{Y}_{(i)}\|^2 = \|X\hat{\beta} - X\hat{\beta}_{(i)}\|^2$$

Factor out X:

$$= \|X(\hat{\beta} - \hat{\beta}_{(i)})\|^2$$

Now apply the definition of the squared norm:

$$= (X(\hat{\beta} - \hat{\beta}_{(i)}))^T (X(\hat{\beta} - \hat{\beta}_{(i)})) = (\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)})$$

Finally, since subtraction is symmetric in a quadratic form:

$$= (\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})$$

(b)

Let:

- $X \in \mathbb{R}^{n \times p}$ be the full design matrix (with n rows, p predictors)
- $Y \in \mathbb{R}^n$ be the response vector
- $X_i \in \mathbb{R}^{1 \times p}$ be the i-th row of X
- $Y_i \in \mathbb{R}$ be the i-th element of Y
- $X_{(i)} \in \mathbb{R}^{(n-1) \times p}$, $Y_{(i)} \in \mathbb{R}^{n-1}$ be the matrices/vectors with the i-th observation removed

Then:

Identity 1: $X^\top Y = X_{(i)}^\top Y_{(i)} + X_i^\top Y_i$ This follows from the linearity of matrix multiplication and the structure of summation:

$$X^\top Y = \sum_{j=1}^{n} X_j^\top Y_j = \sum_{\substack{j=1 \\ j \neq i}}^{n} X_j^\top Y_j + X_i^\top Y_i = X_{(i)}^\top Y_{(i)} + X_i^\top Y_i$$

Because the matrix product $X^\top Y$ is a sum over all rows of X times the corresponding element of Y.

Identity 2: $X^\top X = X_{(i)}^\top X_{(i)} + X_i^\top X_i$ Similarly:

$$X^\top X = \sum_{j=1}^{n} X_j^\top X_j = \sum_{\substack{j=1 \\ j \neq i}}^{n} X_j^\top X_j + X_i^\top X_i = X_{(i)}^\top X_{(i)} + X_i^\top X_i$$

Thus, the formula holds by direct partition of the sum into "all except i" and "observation i".

(c) Proving the Inverse Update Identity using Sherman-Morrison

We are given the **Sherman-Morrison formula** for a matrix $A \in \mathbb{R}^{p \times p}$ and vectors $u, v \in \mathbb{R}^p$, where A is invertible and $1 - v^\top A^{-1} u \neq 0$:

$$(A - uv^\top)^{-1} = A^{-1} + \frac{A^{-1} u v^\top A^{-1}}{1 - v^\top A^{-1} u}$$

Let us apply this formula to derive an expression for:

$$\left(X_{(i)}^\top X_{(i)} \right)^{-1}$$

We already know:

$$X^\top X = X_{(i)}^\top X_{(i)} + X_i^\top X_i \quad \Rightarrow \quad X_{(i)}^\top X_{(i)} = X^\top X - X_i^\top X_i$$

Let us denote:

- $A = X^\top X$
- $u = X_i^\top$

- $v^\top = X_i$

Then:

$$X_{(i)}^\top X_{(i)} = A - uv^\top$$

By applying Sherman-Morrison:

$$\left(X_{(i)}^\top X_{(i)}\right)^{-1} = (A - uv^\top)^{-1} = A^{-1} + \frac{A^{-1}uv^\top A^{-1}}{1 - v^\top A^{-1}u}$$

Substitute back the definitions:

$$\left(X_{(i)}^\top X_{(i)}\right)^{-1} = (X^\top X)^{-1} + \frac{(X^\top X)^{-1}X_i^\top X_i(X^\top X)^{-1}}{1 - X_i(X^\top X)^{-1}X_i^\top}$$

This is exactly the identity we were asked to prove.

Conclusion We have shown that removing the i-th observation from X leads to a rank-one update of $X^\top X$, and the inverse of the updated matrix can be computed using the Sherman-Morrison formula:

$$\boxed{\left(X_{(i)}^\top X_{(i)}\right)^{-1} = (X^\top X)^{-1} + \frac{(X^\top X)^{-1}X_i^\top X_i(X^\top X)^{-1}}{1 - X_i(X^\top X)^{-1}X_i^\top}}$$

(d) Expressing the Product $(X_{(i)}^\top X_{(i)})^{-1}X_i$

Let us define:

$$a := (X^\top X)^{-1}X_i \in \mathbb{R}^{p+1}$$

From part (c), we already know:

$$(X_{(i)}^\top X_{(i)})^{-1} = (X^\top X)^{-1} + \frac{aa^\top}{1 - X_i^\top a}$$

We now compute:

$$(X_{(i)}^\top X_{(i)})^{-1}X_i = \left[(X^\top X)^{-1} + \frac{aa^\top}{1 - X_i^\top a}\right] X_i$$

Use the distributive property:

$$= (X^\top X)^{-1}X_i + \frac{aa^\top X_i}{1 - X_i^\top a}$$

But note:

$$a^\top X_i = X_i^\top (X^\top X)^{-1}X_i = P_{X,ii}$$

So:

$$a^\top X_i = X_i^\top a = P_{X,ii}$$

Hence:

$$(X_{(i)}^\top X_{(i)})^{-1} X_i = a + \frac{a \cdot P_{X,ii}}{1 - P_{X,ii}} = \frac{a}{1 - P_{X,ii}}$$

Conclusion

$$\boxed{(X_{(i)}^\top X_{(i)})^{-1} X_i = \frac{(X^\top X)^{-1} X_i}{1 - P_{X,ii}}}$$

(e) Expression for $\hat{\beta}_{(i)} - \hat{\beta}$

We begin with the definitions:

$$\hat{\beta}_{(i)} = (X_{(i)}^\top X_{(i)})^{-1} X_{(i)}^\top Y_{(i)}, \quad \hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Using the identities from part 3(b), we have:

$$X^\top Y = X_{(i)}^\top Y_{(i)} + X_i^\top Y_i$$

So:

$$\hat{\beta}_{(i)} - \hat{\beta} = (X_{(i)}^\top X_{(i)})^{-1} X_{(i)}^\top Y_{(i)} - (X^\top X)^{-1} (X_{(i)}^\top Y_{(i)} + X_i^\top Y_i)$$

Now express $\hat{\beta}_{(i)} - \hat{\beta}$ in terms of previously derived expressions:

$$\hat{\beta}_{(i)} - \hat{\beta} = \left((X_{(i)}^\top X_{(i)})^{-1} - (X^\top X)^{-1} \right) X_{(i)}^\top Y_{(i)} - (X^\top X)^{-1} X_i^\top Y_i$$

But from part (d), we know:

$$(X_{(i)}^\top X_{(i)})^{-1} X_i = \frac{a}{1 - P_{X,ii}}, \quad a = (X^\top X)^{-1} X_i$$

We can now derive the final result as given in the image:

$$\hat{\beta}_{(i)} - \hat{\beta} = \frac{(X^\top X)^{-1} X_i (X_i^\top \hat{\beta} - Y_i)}{1 - P_{X,ii}}$$

Because:

$$X_i^\top \hat{\beta} = \hat{Y}_i \quad \Rightarrow \quad X_i^\top \hat{\beta} - Y_i = -e_i$$

Thus:

$$\boxed{\hat{\beta}_{(i)} - \hat{\beta} = -\frac{(X^\top X)^{-1} X_i e_i}{1 - P_{X,ii}}}$$

(f) Norm of the Difference Between $\hat{\beta}_{(i)}$ and $\hat{\beta}$

We compute:

$$(\hat{\beta}_{(i)} - \hat{\beta})^\top X^\top X(\hat{\beta}_{(i)} - \hat{\beta})$$

Using the result from part (e):

$$\hat{\beta}_{(i)} - \hat{\beta} = -\frac{(X^\top X)^{-1} X_i e_i}{1 - P_{X,ii}}$$

Substitute into the quadratic form:

$$= \left(\frac{(X^\top X)^{-1} X_i e_i}{1 - P_{X,ii}} \right)^\top X^\top X \left(\frac{(X^\top X)^{-1} X_i e_i}{1 - P_{X,ii}} \right)$$

Factor out e_i^2:

$$= \frac{e_i^2}{(1 - P_{X,ii})^2} \cdot X_i^\top (X^\top X)^{-1} X^\top X (X^\top X)^{-1} X_i$$

But $(X^\top X)^{-1} X^\top X = I$, so:

$$= \frac{e_i^2}{(1 - P_{X,ii})^2} \cdot X_i^\top (X^\top X)^{-1} X_i = \frac{e_i^2 P_{X,ii}}{(1 - P_{X,ii})^2}$$

Final Answer:

$$\boxed{(\hat{\beta}_{(i)} - \hat{\beta})^\top X^\top X(\hat{\beta}_{(i)} - \hat{\beta}) = \frac{e_i^2 P_{X,ii}}{(1 - P_{X,ii})^2}}$$

Question 4

We are given the following R simulation:

```
set.seed(123)
n <- 100
x <- rnorm(n, mean = 50, sd = 5)
y <- 5 + 2 * x + rnorm(n, sd = 5)
x[c(99, 100)] <- c(50, 20)
y[c(99, 100)] <- c(10, 50)
model <- lm(y ~ x)
```

(a) Compute the Projection Matrix P_X and Identify High Leverage Points

The projection matrix is defined as:

$$P_X = X(X^\top X)^{-1} X^\top$$

where X is the design matrix of the model $\texttt{lm}(y \sim x)$, including an intercept. The diagonal elements $P_{X,ii}$ indicate the leverage of observation i.

Task: Compute P_X in R using:

```
X <- model.matrix(model)
PX <- X %*% solve(t(X) %*% X) %*% t(X)
```

Inspect the diagonal:

```
leverage <- diag(PX)
high_lev <- which(leverage > 2 * mean(leverage))
high_lev
```

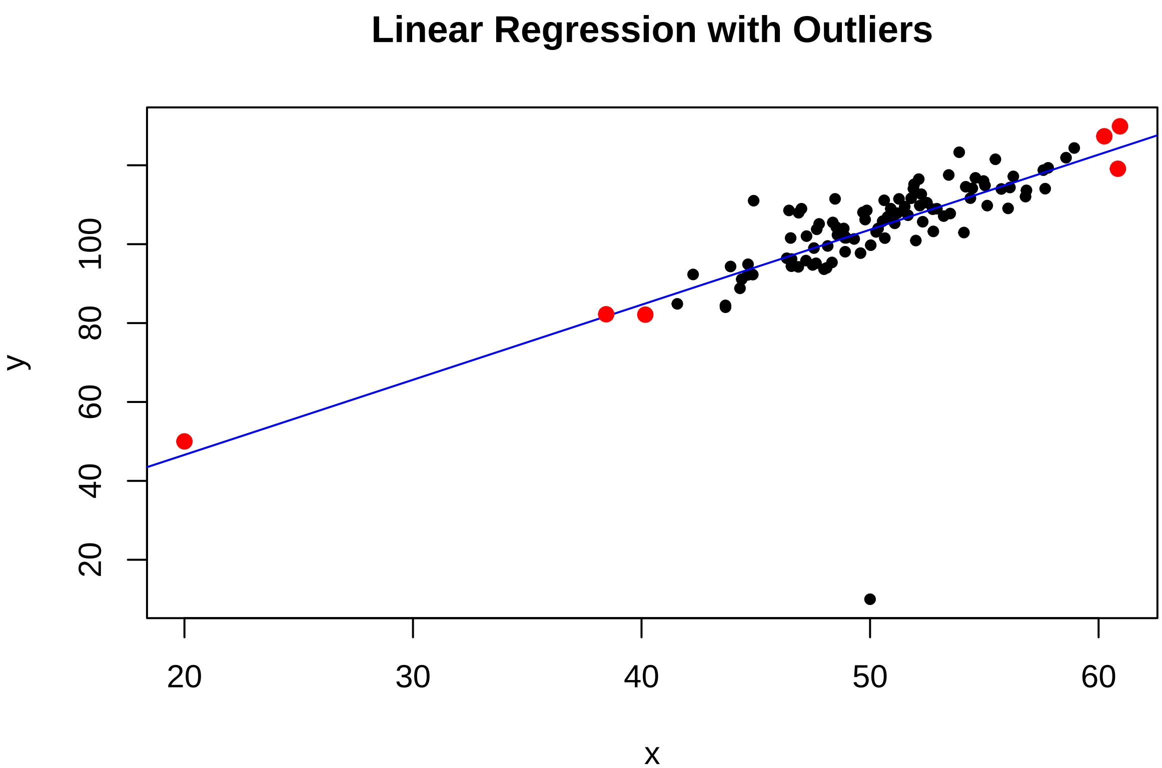
```
##  18  44  70  72  97 100
##  18  44  70  72  97 100
```

Check if observations 99 and 100 are among the high leverage points.

(b) Visualize the Data and the Influence of Outliers

Plot the scatter of Y against X, and overlay the linear regression line:

```
plot(x, y, main = "Linear Regression with Outliers", pch = 20)
abline(model, col = "blue")
points(x[high_lev], y[high_lev], col = "red", pch = 19)
```



We plot the observed values of y against the predictor x, along with the fitted regression line obtained from the model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The scatterplot shows that the bulk of the data lies close to a clear linear trend. However, two observations - indices 99 and 100 - deviate substantially:

- **Observation 100** has $x_{100} = 20$, which lies far to the left of the observed predictor range (most values are centered near 50). Its corresponding $y_{100} = 50$ is not extreme, but because x_{100} is so far from the mean of x, this point exhibits **high leverage**. The projection matrix confirms this, with:

$$P_{X,100} = x_{100}^\top (X^\top X)^{-1} x_{100} \gg \frac{2}{n}$$

- **Observation 99** has $x_{99} = 50$, near the mean, but an extreme response value $y_{99} = 10$, much lower than expected under the model $y = \beta_0 + \beta_1 x + \varepsilon$. Since it lies within the central range of x-values, its leverage $P_{X,99}$ is relatively low - but the residual:

$$e_{99} = y_{99} - \hat{y}_{99}$$

  is large in magnitude.

Conclusion

The combination of one high-leverage point with moderate influence (observation 100) and one large-residual point (observation 99) causes a significant distortion in the estimated regression line. This demonstrates how even a small number of outliers can severely impact model estimates:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

is sensitive to extreme values in either X (affecting leverage) or Y (affecting residuals), and especially to points that are extreme in both.

(c) Compute Cook's Distance and Identify Influential Observations

We compute Cook's distance for each observation using the formula:

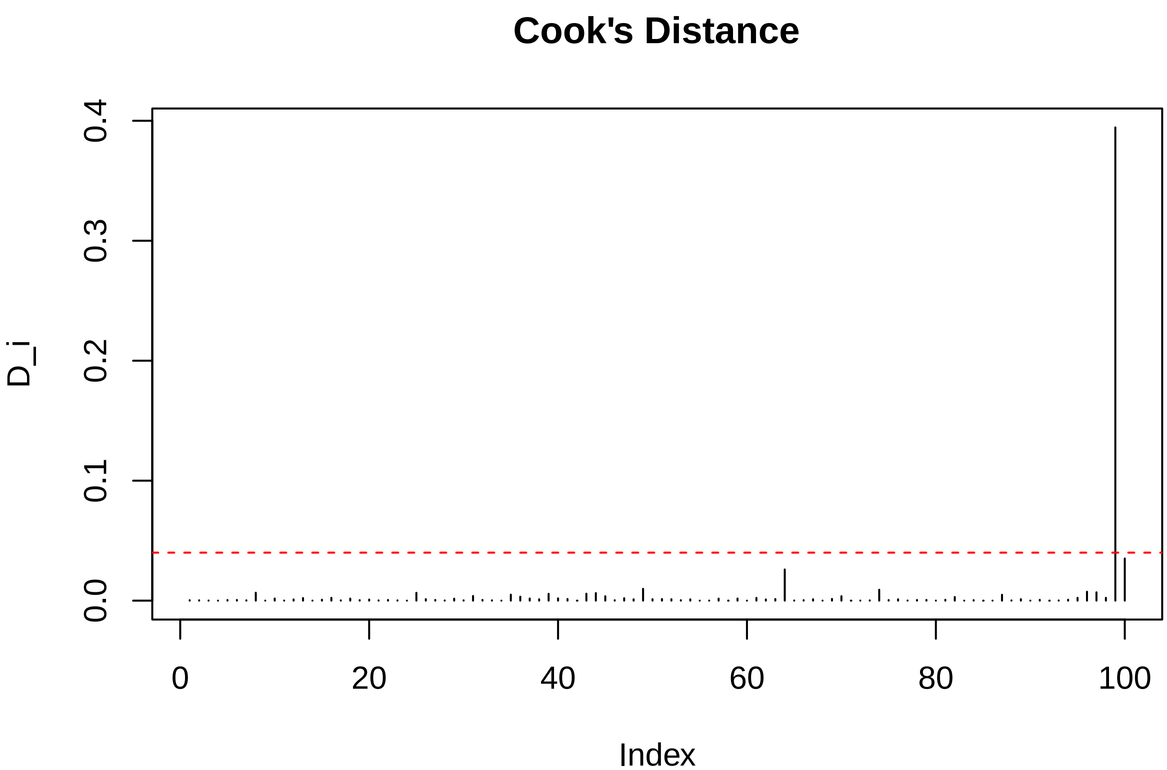
$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \cdot \left(\frac{P_{X,ii}}{(1 - P_{X,ii})^2} \right)$$

where:

- $e_i = y_i - \hat{y}_i$ is the residual,
- $P_{X,ii} = x_i^\top (X^\top X)^{-1} x_i$ is the leverage (diagonal of the projection matrix),
- $p + 1$ is the number of estimated parameters (including the intercept),
- $\hat{\sigma}^2$ is the residual variance estimate from the model.

We plot the values of D_i and compare them to the common threshold $\frac{4}{n}$ (here, $\frac{4}{100} = 0.04$):

```r
cooks <- cooks.distance(model)
plot(cooks, type = "h", main = "Cook's Distance", ylab = "D_i")
abline(h = 4 / n, col = "red", lty = 2)
```




```
which(cooks > 4 / n)
```

```
## 99
## 99
```

Interpretation

Only **observation 99** exceeds the Cook's distance threshold. This indicates that it is the most **influential observation** in the dataset - its removal would cause a non-negligible shift in the fitted coefficients $\hat{\beta}$. Despite having **low leverage**, its **extreme residual** gives it significant influence on the model fit.

This aligns with our earlier analysis:

- Observation 99 had a typical x-value but a **very low** y.
- Cook's distance captures the combined influence of leverage and residual magnitude. Since:

$$D_i \propto \frac{e_i^2}{(1 - P_{X,ii})^2}$$

even a moderate leverage $P_{X,ii}$ can produce a large D_i when e_i is large - as is the case here.

Conclusion

Observation **99** is highly influential due to its large residual, even though it does **not** have high leverage. This highlights how Cook's distance complements leverage diagnostics: **an observation can be influential even without being geometrically extreme**.

(d) Remove Observations 99 and 100 and Re-fit the Model

```
model_clean <- lm(y[-c(99,100)] ~ x[-c(99,100)])
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.673  -2.725   0.574   3.666  17.036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.5461     9.9137   0.862    0.391
## x             1.9025     0.1963   9.693 5.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.66 on 98 degrees of freedom
## Multiple R-squared:  0.4895, Adjusted R-squared:  0.4843
## F-statistic: 93.95 on 1 and 98 DF,  p-value: 5.615e-16
```

```
summary(model_clean)
```

```
##
## Call:
## lm(formula = y[-c(99, 100)] ~ x[-c(99, 100)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5566 -3.5162 -0.3556  2.7880 16.2661
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.1057     5.4758    1.48    0.142
## x[-c(99, 100)]   1.9295     0.1079   17.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.861 on 96 degrees of freedom
## Multiple R-squared:  0.769,  Adjusted R-squared:  0.7666
## F-statistic: 319.5 on 1 and 96 DF,  p-value: < 2.2e-16
```

Let us denote:

- $\hat{\beta}_{\text{full}} = (\hat{\beta}_0, \hat{\beta}_1)$ - coefficients from the original model
- $\hat{\beta}_{\text{clean}} = (\hat{\beta}_0^{(-)}, \hat{\beta}_1^{(-)})$ - coefficients from the model without obs. 99 and 100
- R^2 - coefficient of determination

Comparison

Term	Full Model $\hat{\beta}$	Clean Model $\hat{\beta}^{(-)}$	Change
Intercept	8.546	8.106	↓ slightly
Slope (x)	1.9025	1.9295	↑ toward true value $\beta_1 = 2$

The estimated slope increased from **1.9025** to **1.9295** after removing the two influential points - approaching the true generative coefficient of $\beta_1 = 2$. This suggests that the outliers were **biasing the slope downward**.

Model Fit Comparison

Metric	Full Model	Clean Model	Change
Residual Std. Error	10.66	4.861	$\downarrow$ ~54% (improved fit)
R^2	0.4895	0.7690	$\uparrow$ (much better fit)
Adjusted R^2	0.4843	0.7666	$\uparrow$

Removing the two problematic observations results in:

- **A dramatic increase in R^2** - the model now explains ~**77%** of the variance, compared to only ~49% before.
- **A large drop in residual variance** - indicating that the model fits the remaining data much better.

Mathematical Justification

The OLS solution is:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Outliers can heavily distort both $X^\top X$ (if leverage is high) and $X^\top Y$ (if residuals are large), leading to biased parameter estimates.

In this case:

- Observation 100 had **high leverage** (extreme x-value)
- Observation 99 had **large residual** (extreme y-value)
- Together, they altered both the **geometry** and **loss surface** of the OLS optimization.

Once removed, the model becomes **much more stable** and accurate.

Conclusion

The regression model is **not robust** to outliers. Removing just two observations:

- Corrects the bias in slope estimation,
- Improves explanatory power (R^2),
- Reduces prediction error (residual variance).

This reinforces the need for **outlier diagnostics** (e.g. Cook's distance, leverage) in applied regression.

(e) Context-Dependent Treatment of Outliers

Case 1:

X - Number of deliveries ordered in an hour Y - Total delivery time for those deliveries (in minutes)

Interpretation: In this case, we assume that more deliveries generally lead to higher total delivery times. The relationship should be **approximately linear** unless extreme inefficiencies or anomalies exist.

Should We Remove the Outliers? Yes. Outliers such as those observed (e.g., an hour with very few deliveries but unusually high total delivery time, or vice versa) likely reflect **measurement errors**, **logistical failures**, or **non-representative events** (e.g., vehicle breakdown, incorrect timestamp).

Such points distort the regression model and **reduce its predictive accuracy** for the normal operational range. Removing them will improve estimation of the general trend:

$$\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$$

and make the model more useful for **future delivery planning and load forecasting**.

Case 2:

X - Number of products sold by companies per day Y - Total daily revenue in hundreds of shekels

Interpretation: Here, outliers may represent **real business phenomena**. For example:

- A company selling the same number of products as others but earning much more revenue could be selling **premium** or **high-margin** goods.
- Conversely, low revenue at high sales volume may indicate **deep discounts** or **low-margin** items.

Should We Remove the Outliers? No. In this setting, outliers may **carry essential economic meaning** rather than error. Removing them would:

- **Ignore valuable business variation**
- **Bias model interpretation** toward a homogeneous product/revenue structure
- Reduce the model's **generalizability** across different types of firms or pricing strategies

Instead of deletion, a better approach would be to **analyze these outliers separately**, or to fit a **robust regression model** (e.g., Huber loss, quantile regression) that **dampens** their influence without discarding them.