Demonstrating Linear Model Assumptions

David Gerard

2018-12-07

Objectives

- Understand assumptions of linear regression.
- Evaluate assumptions of linear regression.
- Solve problems of linear regression.
- Ch 8 in the book.

Assumptions in Decreasing Order of Importance

1. **Linearity** - Does the relationship look like a straight line?
2. **Independence** - knowledge of the value of one observation does not give you any information on the value of another.
3. **Equal Variance** - The spread is the same for every value of x
4. **Normality** - The distribution isn't too skewed and there aren't any too extreme points. (only an issue if you have outliers and a small number of observations because of the CLT).

Problems when Violated

1. **Linearity** - Linear regression line does not pick up actual relationship
2. **Independence** - Linear regression line is unbiased, but standard errors are off.
3. **Equal Variance** - Linear regression line is unbiased, but standard errors are off.
4. **Normality** - Unstable results if outliers are present and sample size is small.

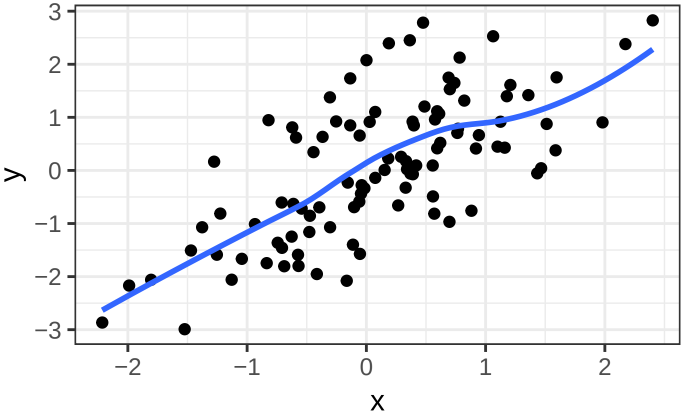
Assessment Tools: Scatterplots and Residual Plots

- Make a scatterplot of the explanatory variable (x-axis) vs the response (y-axis) to check for non-linearity, equal variance, and normality violations.
- Residuals (y-axis) vs fitted values (x-axis) is sometimes more clear because the signal is removed.

Dataset 1: Gold Standard

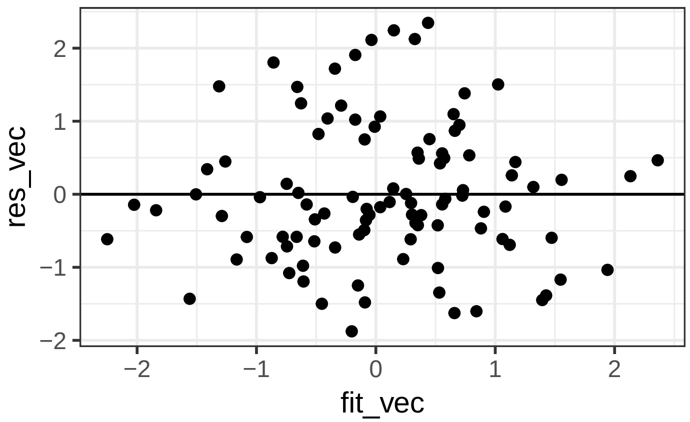
Dataset 1: Scatterplot

```
qplot(x, y) + geom_smooth(se = FALSE)
```



Dataset 1: Residual Plot

```r
lmout <- lm(y ~ x)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



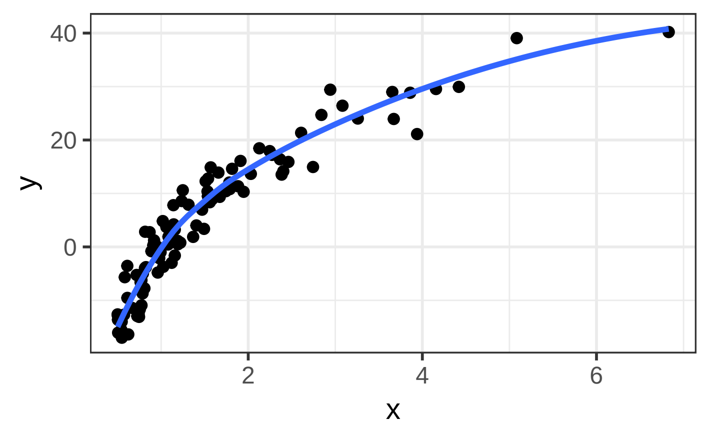
Dataset 1: Summary

- Means are straight lines
- Residuals seem to be centered at 0 for all x
- Variance looks equal for all x
- Everything looks perfect

Dataset 2: Curved Monotone Relationship, Equal Variances

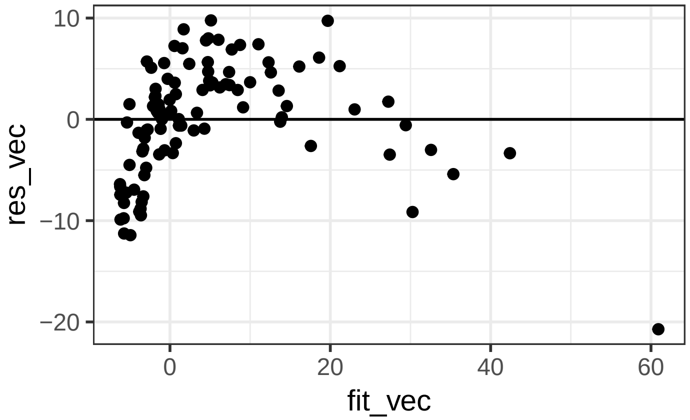
Dataset 2: Scatterplot

```
qplot(x, y) + geom_smooth(se = FALSE)
```



Dataset 2: Residual Plot

```r
lmout <- lm(y ~ x)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```

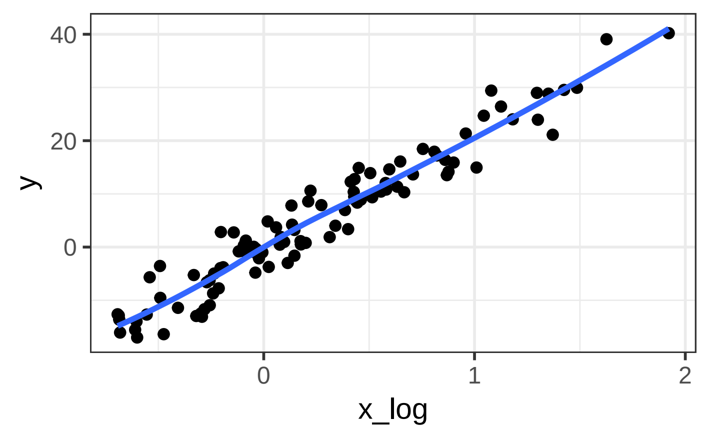


Dataset 2: Summary

- Curved (but always increasing) relationship between x and y.
- Variance looks equal for all x
- Residual plot has a parabolic shape.
- These indicate a log transformation of x could help.

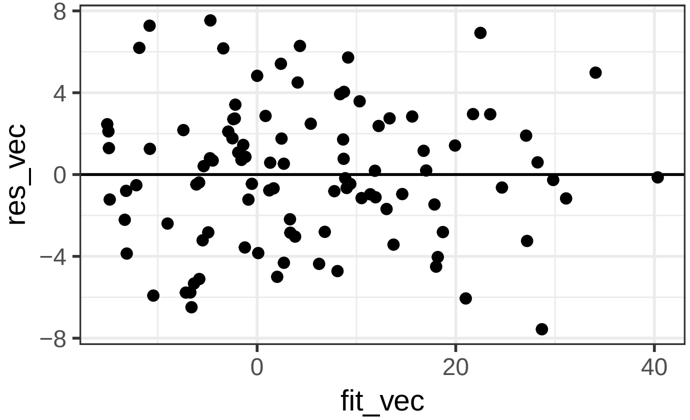
Dataset 2: Transformed x Scatterplot

```
x_log <- log(x)
qplot(x_log, y) + geom_smooth(se = FALSE)
```



Dataset 2: Transformed x Residual Plot

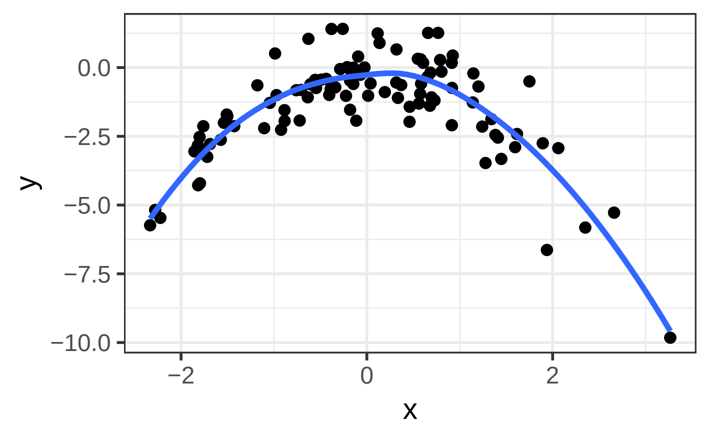
```
lmout <- lm(y ~ x_log)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



Dataset 3: Curved Non-monotone Relationship, Equal Variances

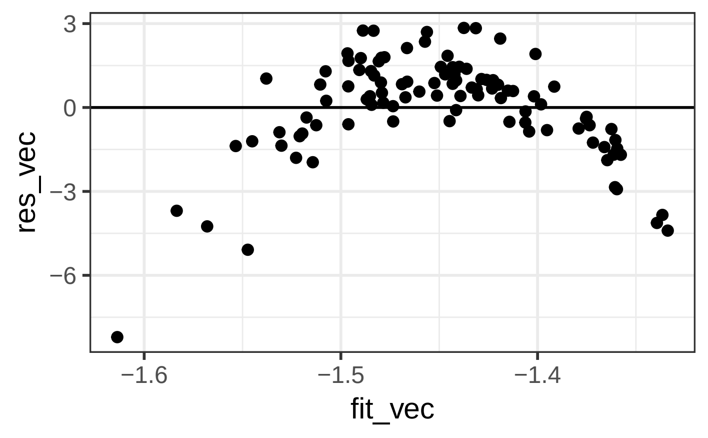
Dataset 3: Scatterplot

```
qplot(x, y) + geom_smooth(se = FALSE)
```



Dataset 3: Residual Plot

```
lmout <- lm(y ~ x)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



Dataset 3: Summary

- Curved relationship between x and y
- Sometimes the relationship is increasing, sometimes it is decreasing.
- Variance looks equal for all x
- Residual plot has a parabolic form.

Dataset 3: Solution

- Two Solutions

1. Fit model:

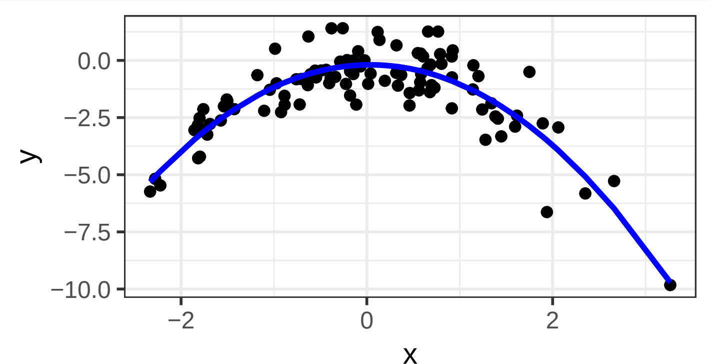
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$

2. Or fit model

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i & \text{if } X_i < C \\ \beta_0^* + \beta_1^* X_i & \text{if } X_i > C \end{cases}$$

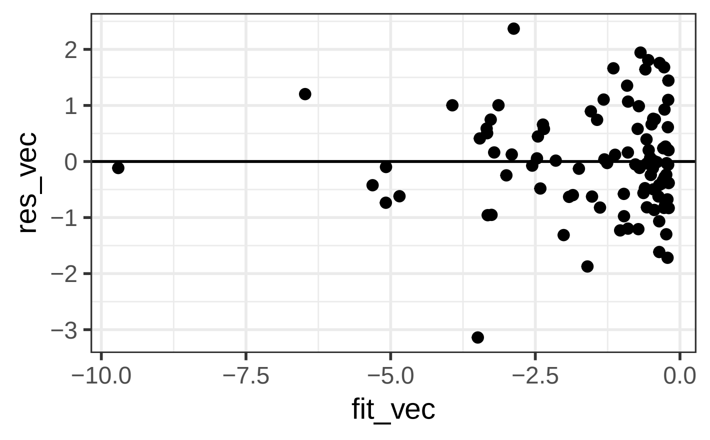
Dataset 3: Solution 1

```
x2 <- x^2
quad_lm <- lm(y ~ x2 + x)
fit_vec <- fitted(quad_lm)
qplot(x, y) +
  geom_line(data = data.frame(x = x, y = fit_vec),
            mapping = aes(x = x, y = y), col = "blue", lwd
```



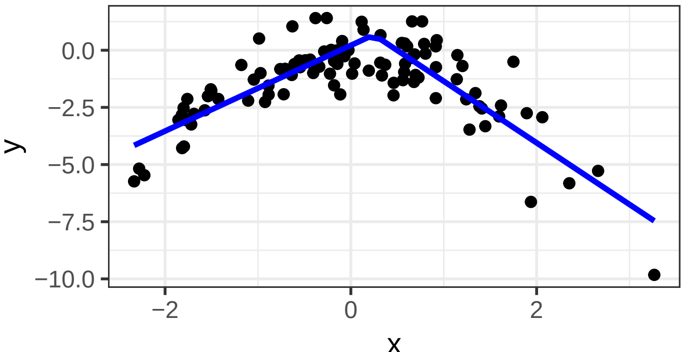
Dataset 3: Solution 1 Residuals

```
res_vec <- resid(quad_lm)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



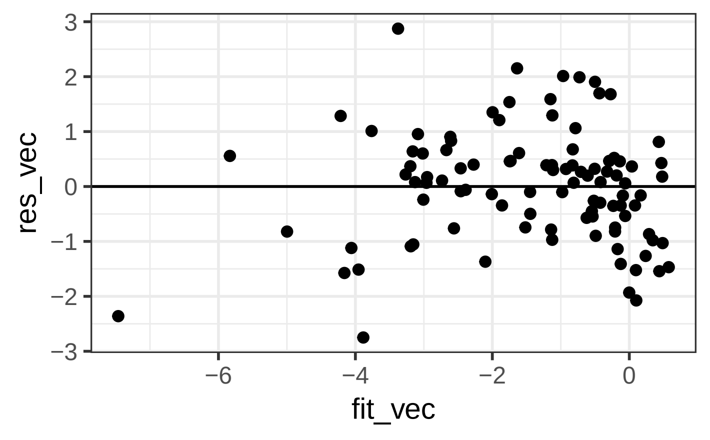
Dataset 3: Solution 2

```r
library(lm.br)
lmbr_out <- lm.br(y ~ x)
fit_vec <- fitted(lmbr_out)
qplot(x, y) +
  geom_line(data = data.frame(x = x, y = fit_vec),
            mapping = aes(x = x, y = y), col = "blue", lwd
```



Dataset 3: Solution 2 Residuals

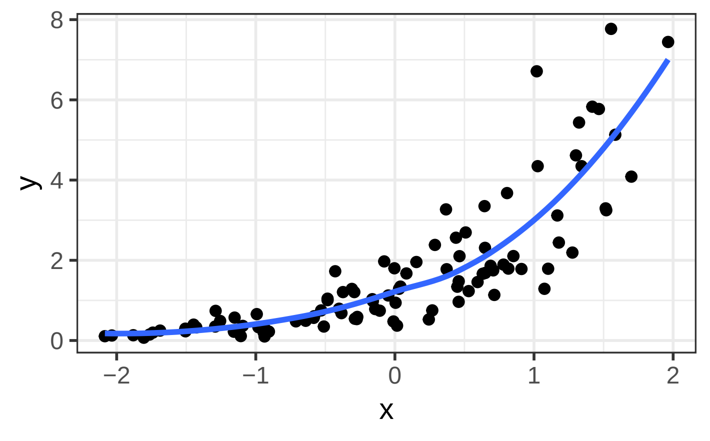
```
res_vec <- resid(lmbr_out)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



Dataset 4: Curved Relationship, Variance Increases with Y

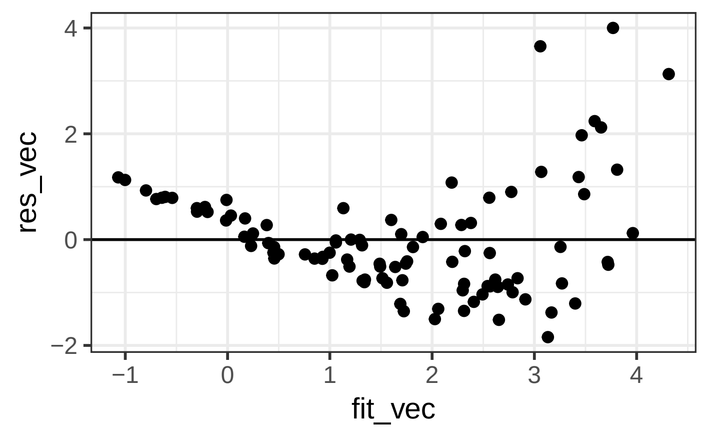
Dataset 4: Scatterplot

```r
qplot(x, y) + geom_smooth(se = FALSE)
```



Dataset 4: Residual Plot

```
lmout <- lm(y ~ x)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



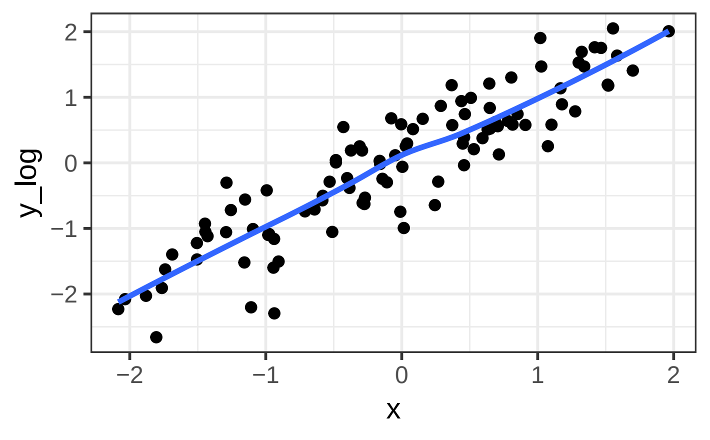
Dataset 4: Summary

- Curved relationship between x and y
- Variance looks like it increases as y increases
- Residual plot has a parabolic form.
- Residual plot variance looks larger to the right and smaller to the left.

Dataset 4: Solution

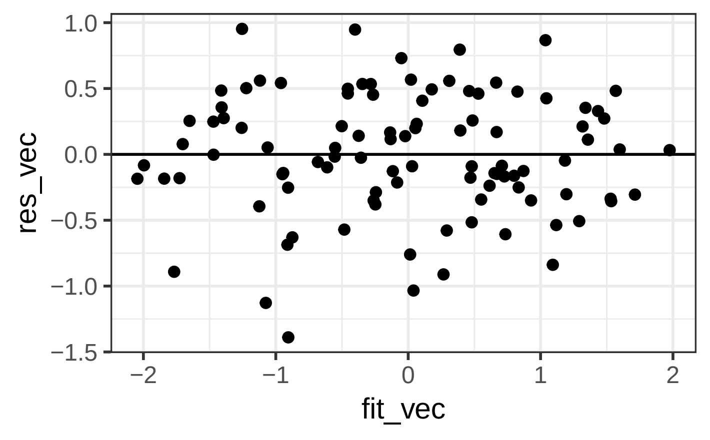
- Take a log-transformation of y.

```
y_log <- log(y)
qplot(x, y_log) + geom_smooth(se = FALSE)
```



Dataset 4: Solution

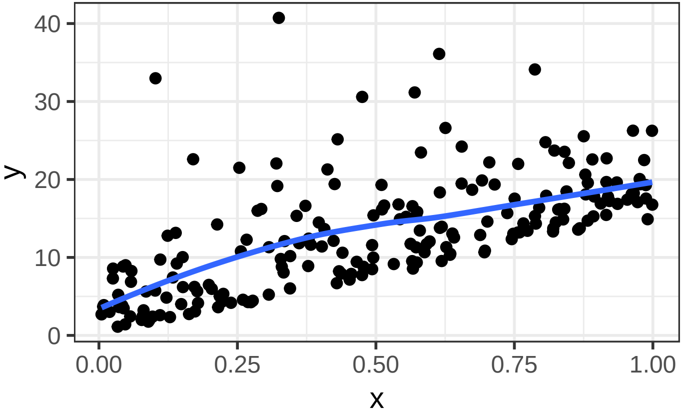
```r
lmout <- lm(y_log ~ x)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



Dataset 5: Linear Relationship, Equal Variances, Skewed Distribution

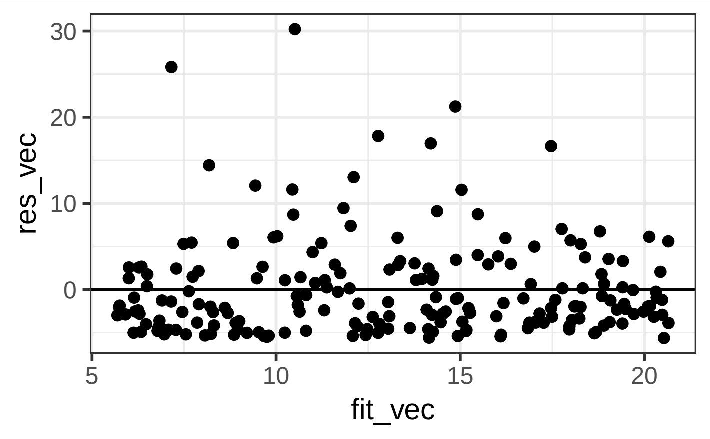
Dataset 5: Scatterplot

```
qplot(x, y) + geom_smooth(se = FALSE)
```



Dataset 5: Residual Plot

```
lmout <- lm(y ~ x)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



Dataset 5: Summary

- Straight line relationship between x and y.
- Variances about equal for all x
- Skew for all x
- Residual plots show skew.

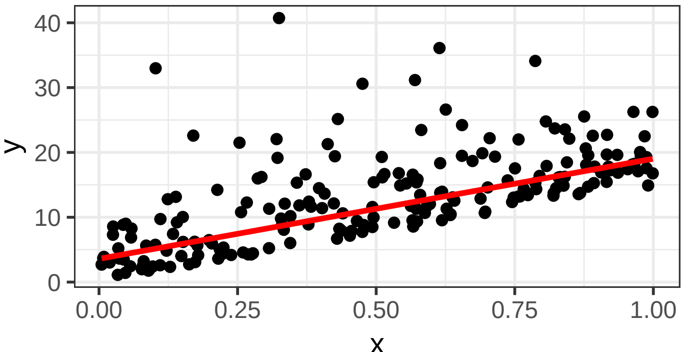
Dataset 5: Solution

- Do nothing, but report skew (usually ok to do)
- Be fancy, fit quantile regression:

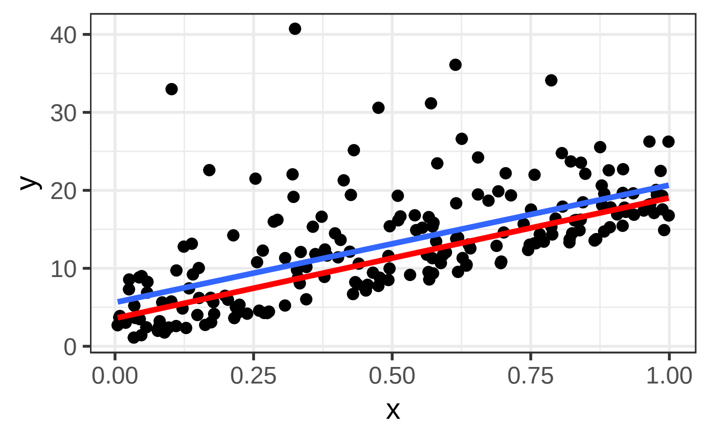
$$Median(Y_i) = \beta_0 + \beta_1 X_i$$

Dataset 5: Quantile Regression

```
library(quantreg)
qr_out <- rq(y ~ x, tau = 0.5)
fit_vec <- fitted(qr_out)
qplot(x, y) +
  geom_line(data = data.frame(x = x, y = fit_vec),
            mapping = aes(x = x, y = y), col = "red", lwd =
```



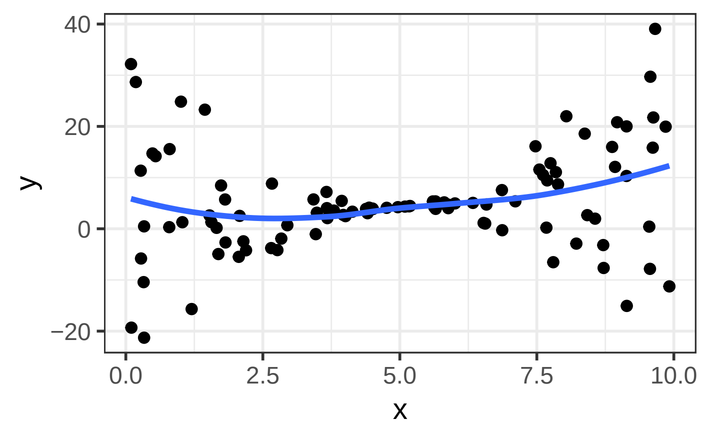
Solution 5: Not too different from regression line



Dataset 6: Linear Relationship, Unequal Variances

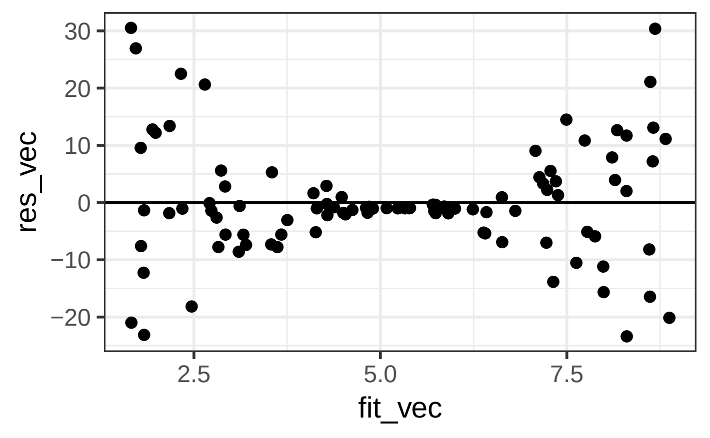
Dataset 6: Scatterplot

```
qplot(x, y) + geom_smooth(se = FALSE)
```



Dataset 6: Residual Plot

```
lmout <- lm(y ~ x)
res_vec <- resid(lmout)
fit_vec <- fitted(lmout)
qplot(fit_vec, res_vec) + geom_hline(yintercept = 0)
```



Dataset 6: Summary

- Linear relationshp between x and y.
- Variance is different for different values of x.
- Residual plots really good at showing this.

Dataset 6: Solution

- The modern solution is to use **sandwich** estimates of the standard errors.

```
library(sandwich)
sandwich(lmout)
```

```
##             (Intercept)       x
## (Intercept)       6.621 -1.1043
## x                -1.104  0.2169
```

- The new standard error of $\hat{\beta}_0$ is the square root of 6.6213
- The new standard error of $\hat{\beta}_1$ is the square root of 0.2169
- The -1.1043 is the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$.

Compare new with old

```r
sqrt(diag(sandwich(lmout)))
```

```
## (Intercept)           x
##      2.5732      0.4657
```

```r
sqrt(diag(vcov(lmout)))
```

```
## (Intercept)           x
##      1.9411      0.3339
```

Using Sandwich in t-tests

```
betahat1_se <- sqrt(sandwich(lmout)[2, 2])
tstat <- coef(lmout)[2] /  betahat1_se
2 * pt(-abs(tstat), df = lmout$df.residual)
```

```
##      x
## 0.1178
```

Compare to Original:

```
coef(summary(lmout))[2, 4]
```

```
## [1] 0.03012
```

Using Sandwich in Confidence Intervals

```r
betahat1_se <- sqrt(sandwich(lmout)[2, 2])
betahat1    <- coef(lmout)[2]
quant975    <- qt(0.975, df = lmout$df.residual)
lower <- betahat1 - quant975 * betahat1_se
upper <- betahat1 + quant975 * betahat1_se
c(lower, upper)
```

```
##       x       x
## -0.1894  1.6589
```

Compare to Original

```r
confint(lmout)
```

```
##                2.5 % 97.5 %
## (Intercept) -2.26518  5.439
## x            0.07213  1.397
```

Intuition of Sandwich Estimator of Variance

- Simplified Model: $Y_i = \beta_1 x_i$ (so zero intercept)
- Using Calculus: $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$

Intuition of Sandwich Estimator of Variance

- Simplified Model: $Y_i = \beta_1 x_i$ (so zero intercept)
- Using Calculus: $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$

So

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) \\ &= \frac{\sum_{i=1}^n x_i^2 Var(y_i|x_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} \end{aligned}$$

Intuition of Sandwich Estimator of Variance

- Simplified Model: $Y_i = \beta_1 x_i$ (so zero intercept)
- Using Calculus: $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$

So

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) \\
&= \frac{\sum_{i=1}^n x_i^2 Var(y_i|x_i)}{\left(\sum_{i=1}^n x_i^2\right)^2}
\end{aligned}
$$

- Usual Method: Estimate $Var(y_i|x_i)$ with s_p^2
  - Assumes variance estimate is same for all i
- Sandwich Method: Estimate $Var(y_i|x_i)$ with $(y_i - \hat{\beta}_1 x_i)^2$
  - Allows variance estimate to differ at each i

Notes on Sandwich

- They result in accurate standard errors of the coefficient estimates as long as
    1. The linearity assumption is satisfied.
    2. You have a large enough sample size.
- You cannot use them for prediction intervals