MLR: Problems and Solutions

Nate Wells

Math 243: Stat Learning

September 23rd, 2020

Outline

In today's class, we will...

- Look at problematic linear models
- Discuss variable transformations

Section 1

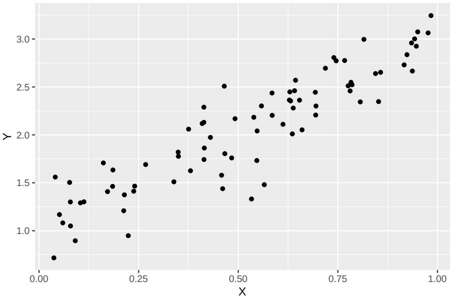
Valid Linear Model

A Valid Model

Previously, we created a valid linear model to use as a baseline:

$$Y = 1 + 2X + \epsilon \qquad \epsilon \sim N(0, 0.25)$$

```
set.seed(700)
X <- runif(80, 0, 1)
e <- rnorm(80, 0, .25)
Y <- 1 + 2*X + e
my_data <- data.frame(X,Y)

ggplot(my_data, aes(x = X , y = Y)) + geom_point()
```

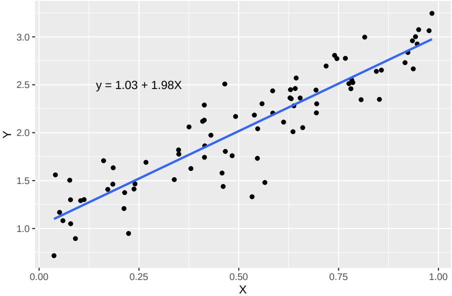


Linear Model

```
my_mod<-lm(Y ~ X, data = my_data)
beta_0 <- summary(my_mod)$coefficients[1]
beta_1 <- summary(my_mod)$coefficients[2]
c(beta_0, beta_1)
```

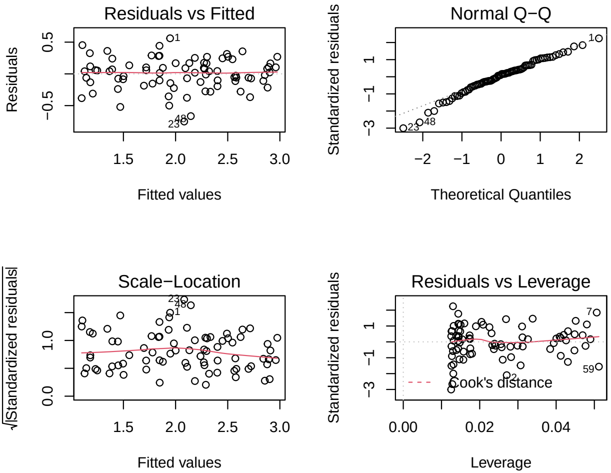
```
## [1] 1.025947 1.981375
```

```
ggplot(my_data, aes(x = X , y = Y)) + geom_point() + geom_smooth(method = "lm", se = F) +
  annotate(geom= "text", x = .25, y = 2.5, label = "y = 1.03 + 1.98X")
```



Plot Quartet

```
par(mfrow = c(2,2))
plot(my_mod)
```



Validation

Assuming the conditions for linear regression are met, then $\hat{\beta}_1$ and RSE^2 are unbiased estimators of β_1 and $\mathrm{Var}(\epsilon)$.

Validation

Assuming the conditions for linear regression are met, then $\hat{\beta}_1$ and RSE^2 are unbiased estimators of β_1 and $\mathrm{Var}(\epsilon)$.

Suppose we randomly generate data and fit models a large number of times.

Validation

Assuming the conditions for linear regression are met, then $\hat{\beta}_1$ and RSE^2 are unbiased estimators of β_1 and $\mathrm{Var}(\epsilon)$.

Suppose we randomly generate data and fit models a large number of times.

- On average, $\hat{\beta}_1 = \beta_1$.

Validation

Assuming the conditions for linear regression are met, then $\hat{\beta}_1$ and RSE^2 are unbiased estimators of β_1 and $\mathrm{Var}(\epsilon)$.

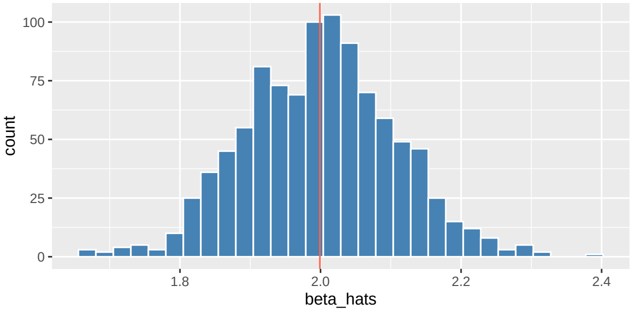
Suppose we randomly generate data and fit models a large number of times.

- On average, $\hat{\beta}_1 = \beta_1$.
- The 95% confidence interval for $\hat{\beta}_1$ should contain the true value of β_1 in approximately 95% of all intervals.

Simulations

```
set.seed(794)

x <- runif(80, 0, 1)

it <- 1000
beta_hats <- rep(NA, it)
capture <- rep(FALSE, it)
for(i in 1:it) {
  e <- rnorm(80, 0, .25)
  y <- 1 + 2*x + e
  m <- lm(y ~ x)
  beta_hats[i] <- m$coef[2]
  ci <- confint(m)[2, ]
  capture[i] <- (ci[1] < 2 & 2 < ci[2])
}
```

Distribution of $\hat{\beta}_1$




```
mean(beta_hats)
```

```
## [1] 1.999127
```

```
mean(capture)
```

```
## [1] 0.947
```

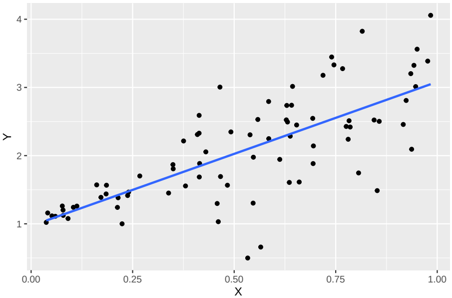
Section 2

Now let's break things

Non-constant variance

$$Y = 1 + 2X + \epsilon \qquad \epsilon \sim N(0, X)$$

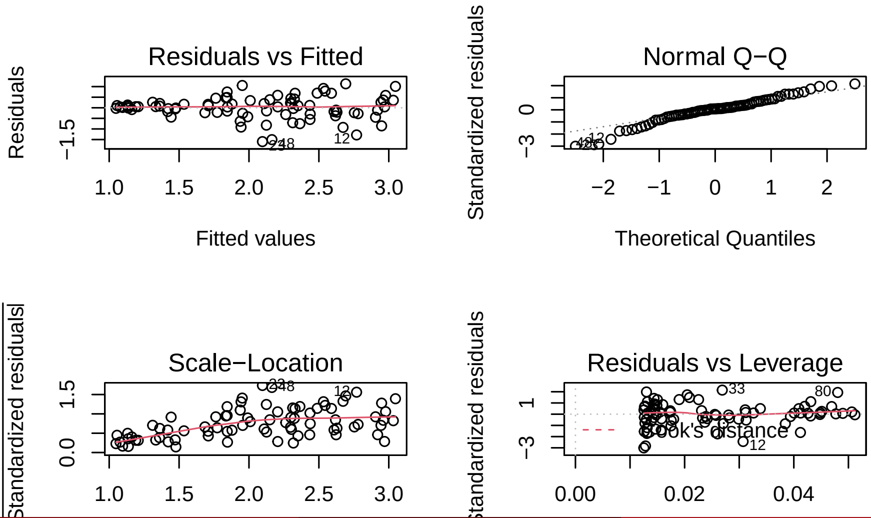
```
set.seed(700)
X <- runif(80, 0, 1)
e <- rnorm(80, 0, sd =  X)
Y <- 1 + 2*X + e
my_data <- data.frame(X,Y)

ggplot(my_data, aes(x = X , y = Y)) + geom_point() +geom_smooth(method = "lm", se = F)
```



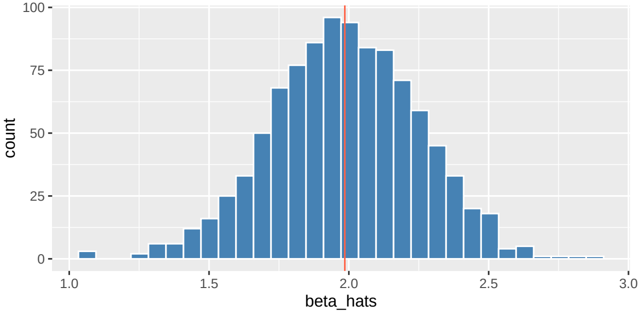
The Linear Model

```
##
## Call:
## lm(formula = Y ~ X, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59764 -0.23174  0.04282  0.27996  1.13194
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9700     0.1280   7.576 6.21e-11 ***
## X             2.1119     0.2175   9.710 4.57e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5345 on 78 degrees of freedom
## Multiple R-squared:  0.5472, Adjusted R-squared:  0.5414
## F-statistic: 94.28 on 1 and 78 DF,  p-value: 4.569e-15
```

Diagnostic plots



Simulations Problematic Model




```
mean(beta_hats)
```

```
## [1] 1.985632
```

```
mean(capture)
```

```
## [1] 0.906
```

A Fix? Weighted least Squares

Each residual contributes to the lm proportional to the reciprocal of its variance.

- This way, all standardized residuals have the same effective variance.
- Downside? We need to estimate the variance of each residual

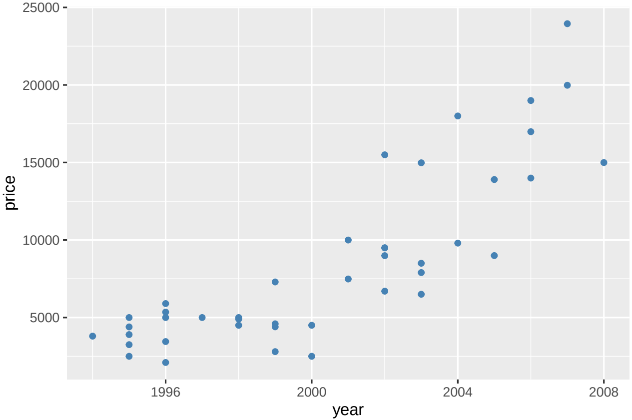
```
##
## Call:
## lm(formula = Y ~ X, data = my_data, weights = 1/X^2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -2.98941 -0.51949  0.08428  0.67098  2.26063
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00807    0.02235    45.1   <2e-16 ***
## X            2.03418    0.14124    14.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.006 on 78 degrees of freedom
## Multiple R-squared:  0.7267, Adjusted R-squared:  0.7232
## F-statistic: 207.4 on 1 and 78 DF,  p-value: < 2.2e-16
```

Section 3

Transformations

Example: Truck Prices

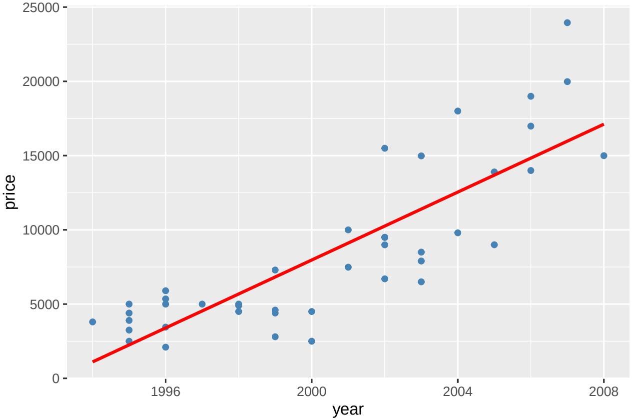
Can we use the age of a truck to predict what its price should be? Consider a random sample of 43 pickup trucks *from the most recent 20 years*.



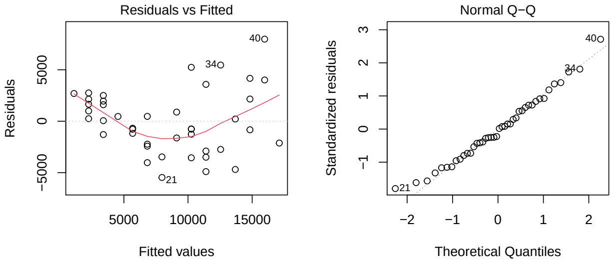
Linear model?

–

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -2278766.230 238325.6991 -9.561563 6.923503e-12
## year            1143.367    119.1371  9.597075 6.237638e-12
```

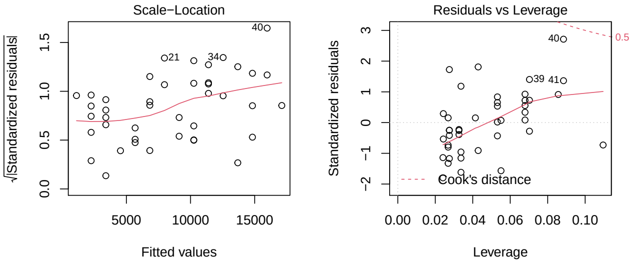


Linearity and normality



- Residuals appear normally distributed.
- But data suggests a non-linear relationship

Variance and leverage

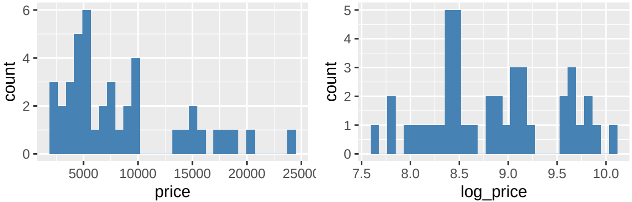


- One observation (44) appears influential.
- There is evidence of increasing variance in the residuals.

Transformations

If the diagnostic plots look bad, try to transform variables by applying functions.

```
pickups <- mutate(pickups, log_price = log(price))
```



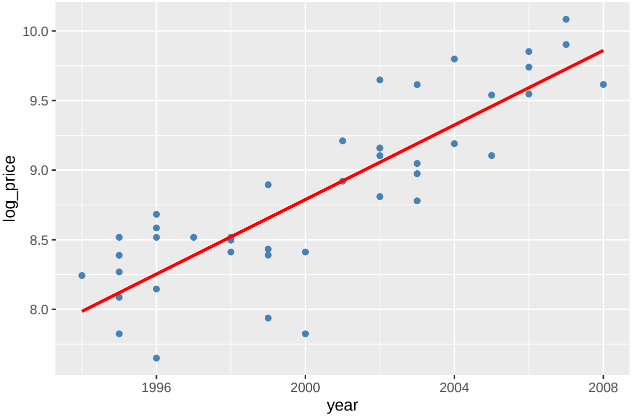
Variables that span multiple orders of magnitude often benefit from a natural log transformation.

$$Y_t = \ln(Y)$$

Log-transformed linear model

```
m2 <- lm(log_price ~ year, data = pickups)
summary(m2)$coef
```

```
##                  Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -258.9980504 26.12294226 -9.914582 2.471946e-12
## year           0.1338934  0.01305865 10.253239 9.342855e-13
```

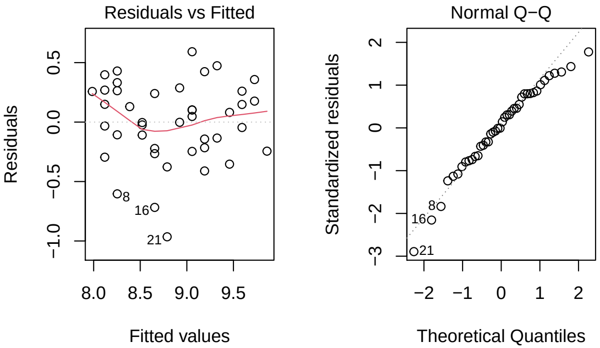


Poll: Interpretation

The slope coefficient in the log-linear model was 0.13. Which of the following interpretations are correct? Select all that apply

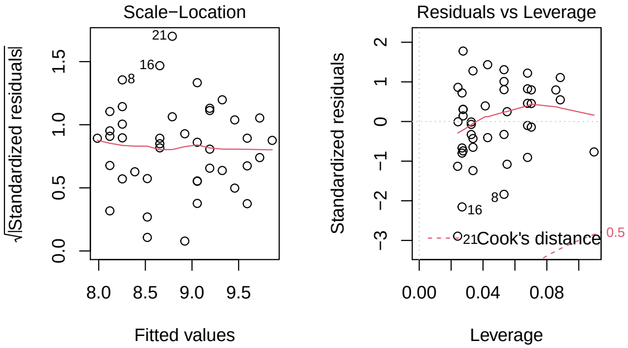
1. Increasing year by 1 increases price by approximately 0.13.
2. Increasing year by 1 produces a relative increase in price of approximately $e^{.13}$.
3. Increasing year by 1 increases the log-price by approximately 0.13.
4. Increasing year by $\ln(1)$ increases price by approximately 0.13.

Linearity and normality



- The residuals from this model appear less normal
- But the quadratic trend is now less apparent.

Constant variance and influence



- There are no points flagged as influential
- The variance has been stabilized

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
    - Count data and prices often benefit from transformations.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
  - Count data and prices often benefit from transformations.
  - The natural log and the square root are the most common, but you can use any transformation you like.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
  - Count data and prices often benefit from transformations.
  - The natural log and the square root are the most common, but you can use any transformation you like.
- Transformations may change model interpretations.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
    - Count data and prices often benefit from transformations.
    - The natural log and the square root are the most common, but you can use any transformation you like.
- Transformations may change model interpretations.
- Non-constant variance is a serious problem but it can sometimes be solved by transforming the response.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
  - Count data and prices often benefit from transformations.
  - The natural log and the square root are the most common, but you can use any transformation you like.
- Transformations may change model interpretations.
- Non-constant variance is a serious problem but it can sometimes be solved by transforming the response.
- Transformations can also fix non-linearity, as can polynomials.