Regression review I

Sample $x_i, Y_i, i = 1, \ldots, n$.
Underlying model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, ϵ_i i.i.d $N(0, \sigma)$.
LS estimates b_0 for β_0 (unbiased and linear in Y_i) and b_1 for β_1:

$$b_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad\qquad b_0 = \bar{Y} - b_1 \bar{x}\ .$$

$$\mathrm{SE}(b_1) = S \frac{1}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \mathrm{SE}(b_0) = S \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)}.$$

Where S^2 is an unbiased estimate for σ^2:

$$\widehat{\sigma}^2 \;=\; S^2 \;=\; \frac{\mathrm{SSE}}{\mathrm{df}} \;=\; \frac{\mathrm{SSE}}{n-2} = \frac{\sum_{i=1}^n r_i^2}{n-2}.$$

Distribution to use: t_{n-2}.

Regression review II

Prediction: For a particular predictor value x^*:

The estimate for $\hat{\mu}(x^*)$:

$$\hat{\mu}(x^*) = b_0 + b_1 x^*, \quad \mathrm{SE}(\hat{\mu}(x^*)) = S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

The estimate for $\hat{Y}$ for a new observation from population x^*:

$$\hat{Y} = b_0 + b_1 x^*, \quad \mathrm{SE}(\hat{Y}) = S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Notation:
$S_{xx} = SSX = \sum_{i=1}^{n}(x_i - \bar{x})^2$,
$S_{yy} = SSY = SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Analysis of variance I

Analysis of variance is the term for statistical analyses that break down the variation in data into separate pieces that correspond to different sources of variation. In the regression setting, the observed variation in the responses comes from two sources.

- As the explanatory variable x changes, it "pulls" the response with it along the regression line. This is the **variation along the line** or **regression sum of squares**:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- When x is held fixed, y still varies because not all individuals who share a common x have the same response y. This is the **variation about the line** or **error (residual) sum of squares**:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Analysis of variance II

The ANOVA Equation

It turns out that SSE and SSR together account for *all* the variation in y (i.e. S_{yy}):

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}$$

The degrees of freedom break down in a similar manner:

$$\underbrace{n-1}_{\text{dfT}} = \underbrace{1}_{\text{dfR}} + \underbrace{n-2}_{\text{dfE}}$$

Dividing a sum of squares by its degrees of freedom gives a **mean square (MS)**.

Analysis of variance III

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{\mathrm{dfE}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = S^2$$

Remember:

$$b_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum_i(\hat{Y}_i - \bar{Y})^2}{\mathrm{SST}} = \frac{b_1^2\mathrm{SSX}}{\mathrm{SST}} = r^2$$

The fraction of the total variation in Y explained by the line

Analysis of variance IV

The ANOVA F Statistic

As an alternative test of the hypothesis: $H_0 : \beta_1 = 0$, we use the F statistic:

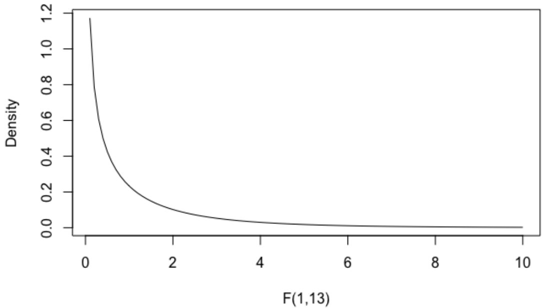
$$
\begin{aligned}
F &= \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/\text{dfR}}{\text{SSE}/\text{dfE}} \\
&= \frac{b_1^2 \text{SSX}}{S^2} \\
&= \left(\frac{b_1}{S/\sqrt{\text{SSX}}} \right)^2 \\
&= \left(\frac{b_1}{SE(b_1)} \right)^2 \\
&= t^2
\end{aligned}
$$

Analysis of variance V

Under H_0,

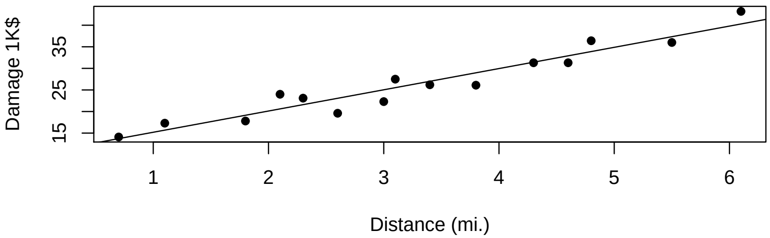
$$F \sim F_{1,n-2}$$

where $F_{1,n-2}$ is an F distribution with 1 and $n-2$ degrees of freedom.



Analysis of variance VI

```r
# Plot the data and the regression line
plot(fire$dist,fire$damage,pch=16,xlab="Distance (mi.)",
     ylab="Damage 1K$")
abline(fire.lm)
```



Analysis of variance VII

```
##
## Call:
## lm(formula = damage ~ dist, data = fire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4682 -1.4705 -0.1311  1.7915  3.3915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
## dist          4.9193     0.3927  12.525 1.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 13 degrees of freedom
## Multiple R-squared:  0.9235,Adjusted R-squared:  0.9176
## F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

Residual standard error is the estimate S of σ.

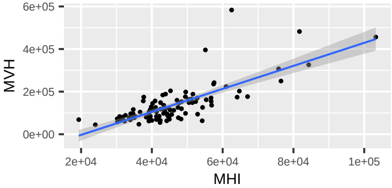
$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad R^2_{adj} = 1 - \frac{SSE/dfE}{SST/dfT} = 1 - \frac{S^2}{SST/dfT}$

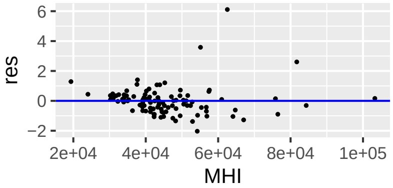
F-statistic is the square of the t-statistic for the slope.

Transformations I

Sometimes a transformation of the response variable yields data that better fits the assumptions.

```
load("countyComplete.Rdata")
CCS=sample(countyComplete,100)
mod=lm(median_val_owner_occupied~median_household_income,data=CCS)
res=residuals(mod)
CCS$res=(res-mean(res))/sd(res)
p1=qplot(median_household_income,median_val_owner_occupied,data=CCS,
         geom=c("point","smooth"),method="lm",ylab="MVH",
         xlab="MHI",size=I(0.5))+theme(text=element_text(size=8))
p2=qplot(median_household_income,res,data=CCS,xlab="MHI",ylab="res",size=I(0.5))+
  geom_hline(yintercept=0,col="blue")
grid.arrange(p1, p2,  ncol=2)
```

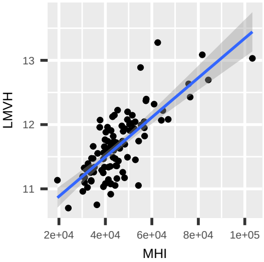




Transformations II

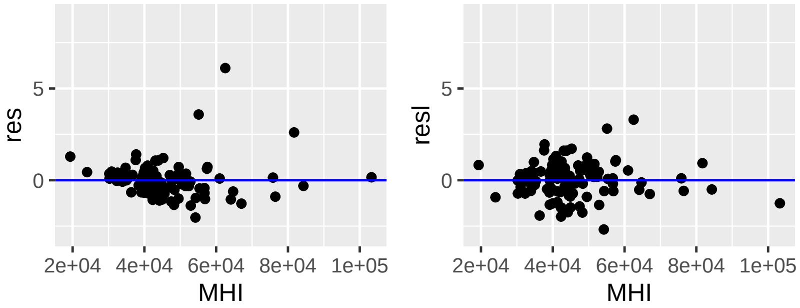
The residual plot seems to have some trends in it. Also data is highly clustered at the lower values. Log transformation can sometimes help.

```
modl=lm(log(median_val_owner_occupied)~median_household_income,data=CCS)
qplot(mapping=aes(x=median_household_income,
                  y=log(median_val_owner_occupied)),data=CCS,geom=c("point","smooth"),
      method="lm",ylab="LMVH",xlab="MHI",size=I(.5))+theme(text=element_text(size=5))
```



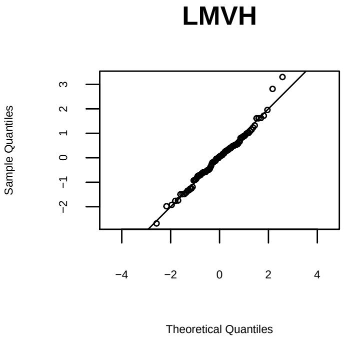
Transformations III

Compare the two residual plots:



Transformations IV

Compare the two qqplots.

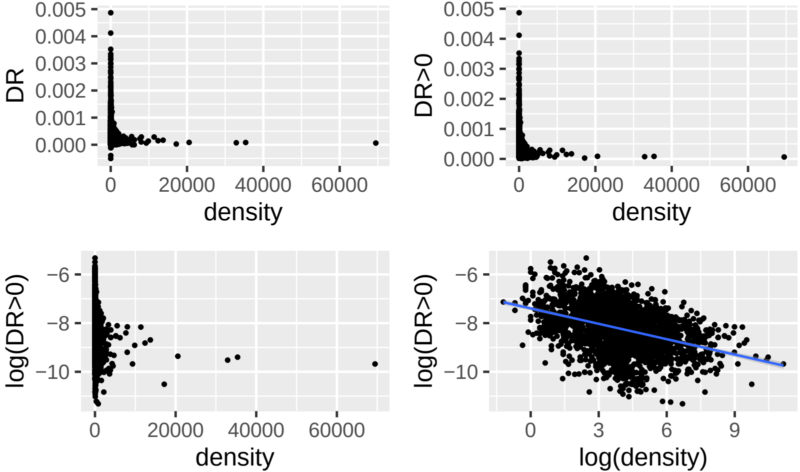


Transformations V

Let's look at the relationship between county population density and COVID death rates in December 2020. Here we want to transform both response and predictor.

```
load("~/Box Sync/tex/courses/234/234_Winter_2021/County_data/County_Data_Demo_Covid_Elec.Rdata")
p1=qplot(density,Dec_Death_rate,data=County_data,ylab="DR",size=I(.5))
p2=qplot(density,Dec_Death_rate,data=subset(County_data,Dec_Death_rate>0),ylab="DR>0",size=I(.5))
p3=qplot(density,log(Dec_Death_rate),data=subset(County_data,Dec_Death_rate>0),
        ylab="log(DR>0)",size=I(.5))
p4=qplot(log(density),log(Dec_Death_rate),
        data=subset(County_data,Dec_Death_rate>0),
        geom=c("point","smooth"),method="lm",ylab="log(DR>0)",size=I(.5))
grid.arrange(p1, p2, p3, p4, ncol=2)
```

Transformations VI



Transformations VII

```
##
## Call:
## lm(formula = log(Dec_Death_rate) ~ log(density), data = subset(County_data,
##     Dec_Death_rate > 0))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.88846 -0.51637  0.04678  0.59286  2.58685
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.39561    0.04138 -178.74   <2e-16 ***
## log(density) -0.21074    0.00961  -21.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8285 on 2786 degrees of freedom
## Multiple R-squared:  0.1472,Adjusted R-squared:  0.1469
## F-statistic: 480.9 on 1 and 2786 DF,  p-value: < 2.2e-16
```

Transformations VIII

Log transform of response:
$log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \ \rightarrow \ Y_i = \exp(\beta_0) \cdot \exp(\beta_1 X_i) \cdot \exp(\epsilon_i)$.

Log transform of both response and predictor:
$\log(Y_i) = \beta_0 + \beta_1 \log X_i + \epsilon_i \ \rightarrow \ Y_i = \exp(\beta_0) X_i^{\beta_1} \exp^{\epsilon_i}$.

Interpretations of the data are then mainly done on the log scales.
But...

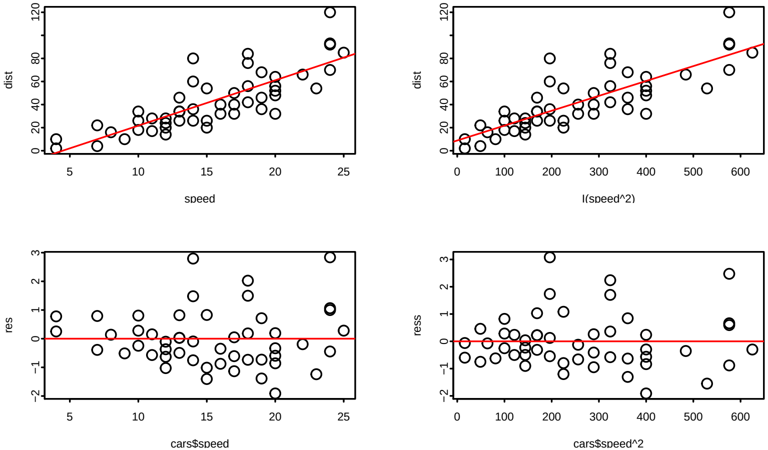
If you get a prediction interval $[l, u]$for $\log(Y)$ at some value x, you can take $[e^l, e^u]$ as prediction interval for Y.

Transformations IX

Sometimes you want a polynomial transformation of the predictor $Y = \beta_0 + \beta_1 X^2 + \epsilon$.

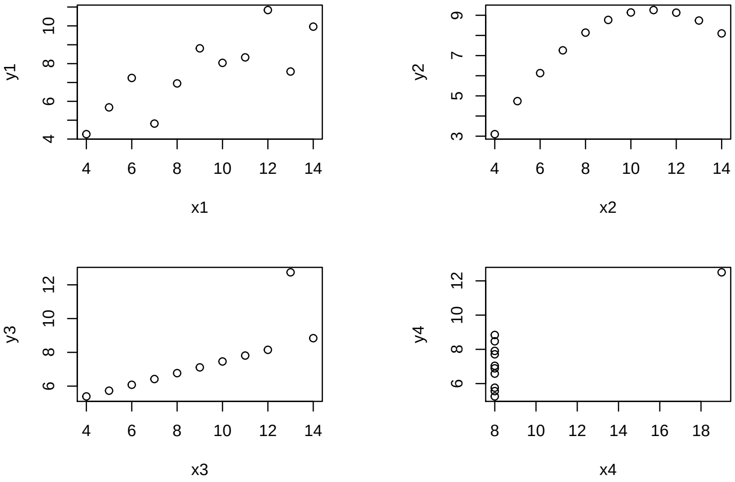
```
data(cars,package="datasets")
mod=lm(dist~speed,data=cars)
res=residuals(mod)
res=(res-mean(res))/sd(res)
mods=lm(dist~I(speed^2),data=cars)
ress=residuals(mods)
ress=(ress-mean(ress))/sd(ress)

par(mfrow=c(2,2),mai = c(.3, 0.3, 0.3, 0.3),cex.lab=.5,cex.main=.5,cex.axis=.5)
plot(dist~speed,data=cars,mgp=c(1,0,0),tck=-.02)
abline(mod,col=2)
plot(dist~I(speed^2),data=cars,mgp=c(1,0,0),tck=-.02)
abline(mods,col=2)
plot(cars$speed,res,mgp=c(1,0,0),tck=-.02)
abline(h=0,col=2)
plot(cars$speed^2,ress,mgp=c(1,0,0),tck=-.02)
abline(h=0,col=2)
```

Transformations X



Simple regression issues I

4 data-sets producing exactly the same regression results - only one of them seems to satisfy the assumptions:



Simple regression issues II

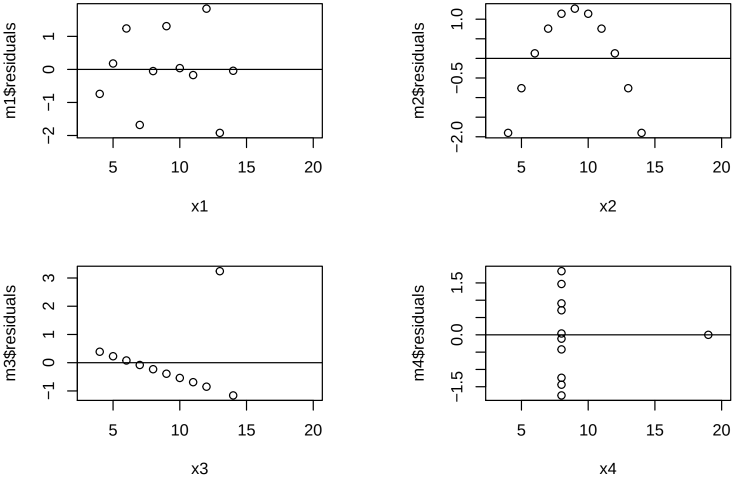
```
s1$coefficients

##             Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0000909  1.1247468 2.667348 0.025734051
## x1          0.5000909  0.1179055 4.241455 0.002169629

s2$coefficients

##             Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.000909  1.1253024 2.666758 0.025758941
## x2          0.500000  0.1179637 4.238590 0.002178816

s3$coefficients

##             Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0024545  1.1244812 2.670080 0.025619109
## x3          0.4997273  0.1178777 4.239372 0.002176305

s4$coefficients

##             Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0017273  1.1239211 2.670763 0.025590425
## x4          0.4999091  0.1178189 4.243028 0.002164602
```

Simple regression issues III

The statistics for linear regression from all 4 datasets are identical!

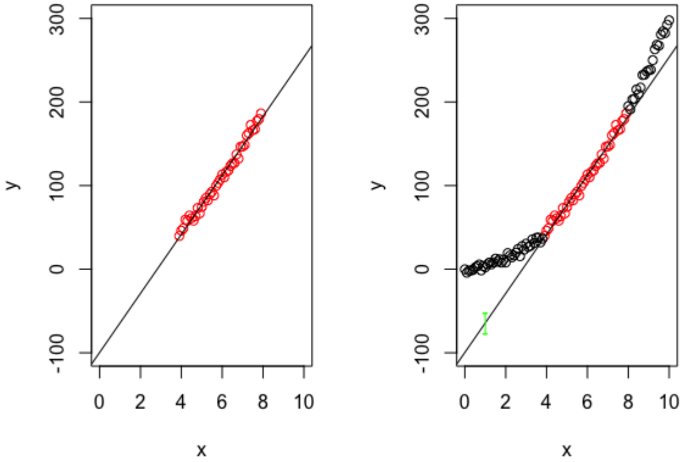
```
c(s1$r.squared,s2$r.squared,s3$r.squared,s4$r.squared)

## [1] 0.6665425 0.6662420 0.6663240 0.6667073
```

Simple regression issues IV



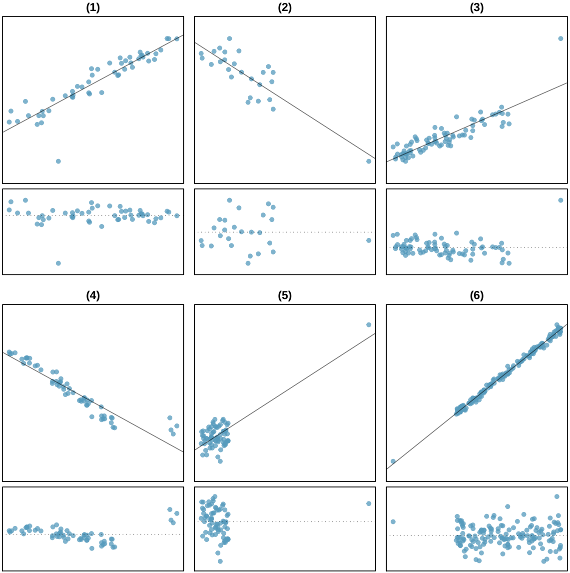
Simple regression issues V

Beware of extrapolation



Simple regression issues VI

Outliers



Simple regression issues VII

- Non-linear relationship between variables.
- Extreme outlier in terms of response
- Extreme outlier in terms of predictor
- Beware of extrapolating beyond range of predictor