Linear Regression Basics

Roberta De Vito

BROWN
Public Health

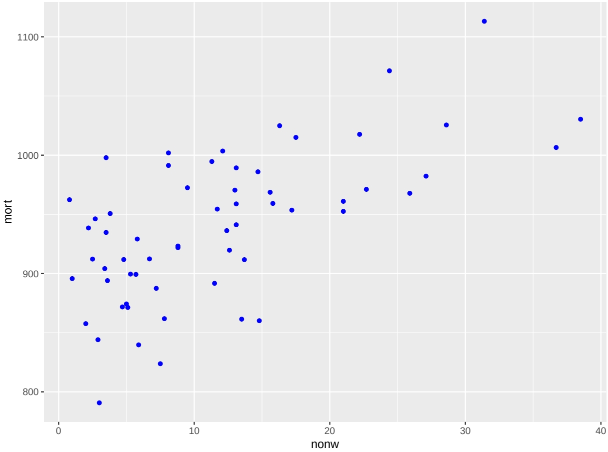
Questions

- Is there a relationship between a predictor X and the outcome Y?
- How strong is the relationship? Is it linear?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How accurately can we predict Y?

Example: Pollution data set

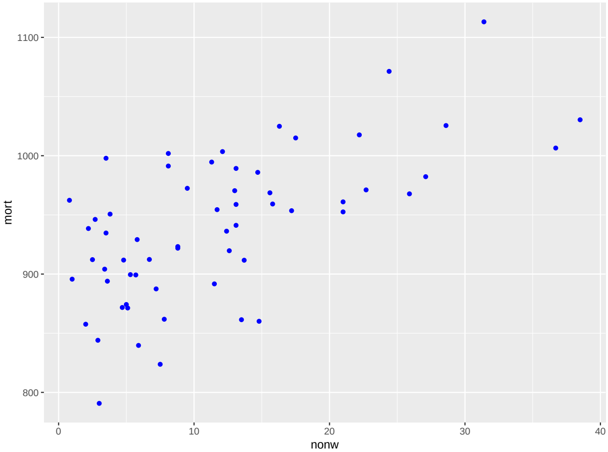
`ID`	code for the identification of the sample
`OVR65`	% of 1960 SMSA population aged 65 or older
`EDUC`	Median school years completed by those over 22
`HOUS`	% of housing units with all facilities
`DENS`	Population per sq. mile in urbanized areas, 1960
`NONW`	% non-white population in urbanized areas, 1960
`WWDRK`	% employed in white collar occupations
`POOR`	% of families with income < 3000
`HC`	Relative hydrocarbon pollution potential
`NOX`	Same for nitric oxides
`SO2`	Same for sulphur dioxide
`HUMID`	Annual average % relative humidity at 1pm
`MORT`	Total age-adjusted mortality rate per 100,000
`PREC`	Average annual precipitation in inches

The correlation in practice




```
   Pearson's product-moment correlation

data:  mort and nonw
t = 6.4067, df = 58, p-value = 2.885e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4659958 0.7715516
sample estimates:
      cor
0.6437473
```

The correlation in practice: Question in prismia




```
   Spearman's rank correlation rho

data:  mort and nonw
S = 14080, p-value = 2.458e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6087923

Warning message:
In cor.test.default(mort, nonw, method = "spearman") :
  Cannot compute exact p-value with ties
```

The linear regression

- $y_i = f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$
- $f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

$$\Downarrow$$

Matrix form: $Y = X\beta + \epsilon$

Question on Prismia

Simple linear regression

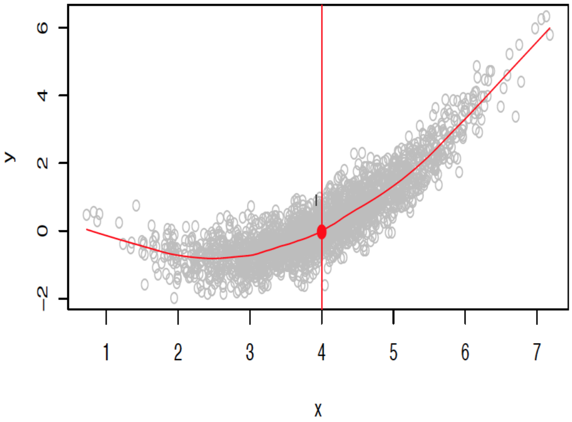
- $y_i = f(x_i) = \beta_0 + \beta_1 x_{i1} + \epsilon_i$
- $f(x_i) = \beta_0 + \beta_1 x_{i1}$

$$\Downarrow$$

$$\text{matrix form: } Y = X\beta + \epsilon$$

Model Assumption 1.

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

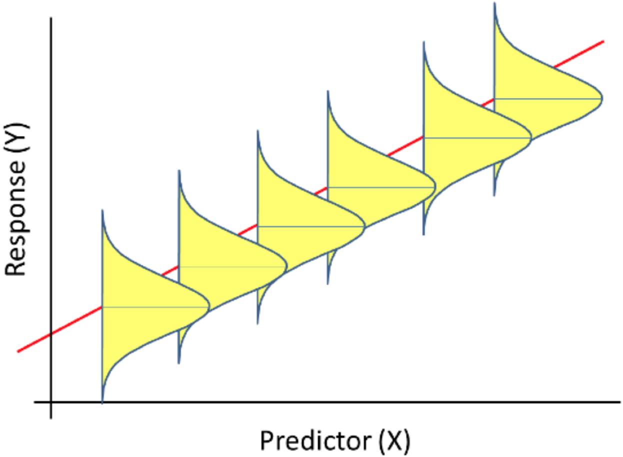


Model Assumption 2.

$$\epsilon \sim N(0, \sigma^2)$$

Model Assumption 2.

$$\epsilon \sim N(0, \sigma^2)$$



Model Assumption 3. and 4.

3. Error term is independent of (uncorrelated with) covariate(s)

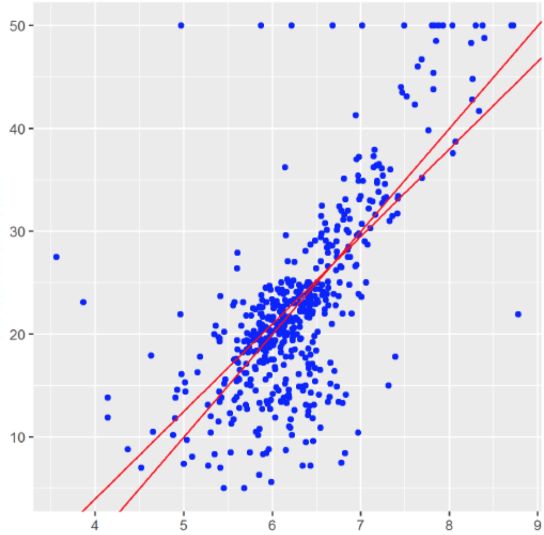
$$Corr(X, \epsilon) = 0$$

4. Variance of error term is same, regardless of value of x (homoscedasticity)

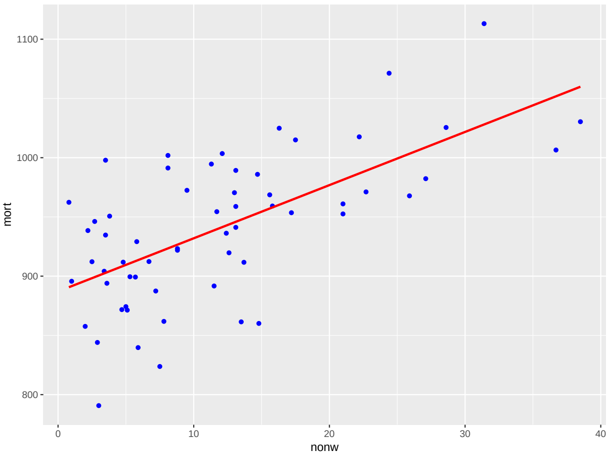
$$Var(\epsilon) = \sigma^2$$

Fitting the best line

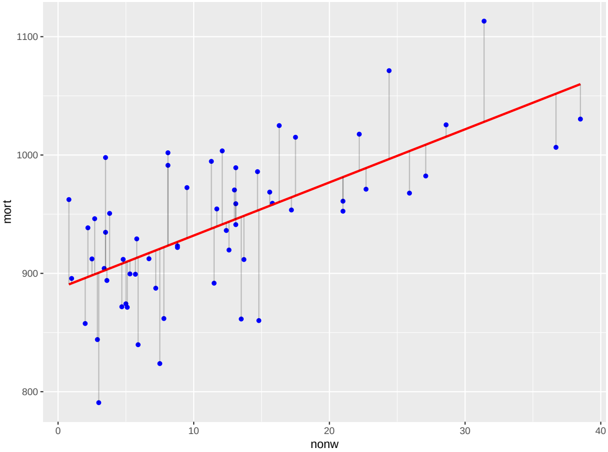
- How do we find regression line that fits best?



The linear regression in practice



The linear regression in practice



The lm function in R: what are we looking?

```
> summary(lm(mort~nonw))

Call:
lm(formula = mort ~ nonw)

Residuals:
     Min       1Q   Median       3Q      Max 
-109.810  -32.757   -4.021   35.053   95.088 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 887.0765    10.3723  85.524  < 2e-16 ***
nonw          4.4888     0.7006   6.407 2.88e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.01 on 58 degrees of freedom
Multiple R-squared:  0.4144,	Adjusted R-squared:  0.4043 
F-statistic: 41.05 on 1 and 58 DF,  p-value: 2.885e-08
```

Interpreting Coefficients

- Intercept term: mean of Y for those having $X = 0$

$$E(Y|X) = \beta_0 + \beta_1 0 = \beta_0 = 887.0765$$

- Frequently, intercept is scientifically meaningless; we can use mean centered covariates (more later)

Interpreting Coefficients

- Slope term

$$
\begin{aligned}
E[Y|X = x + 1] &= \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1(x + 1) \\
E[Y|X = x] &= \beta_0 + \beta_1 x
\end{aligned}
$$

- What happens when taking difference between these means?

Interpreting Coefficients

- Slope term

$$E[Y|X = x + 1] = \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1(x + 1)$$
$$E[Y|X = x] = \beta_0 + \beta_1 x$$

- What happens when taking difference between these means?
- Mean difference in Y for data which differ by one X unit.
- in our case $\beta_1 = 4.4888$

The lm function in R: what are we looking?

```
> summary(lm(mort~nonw))

Call:
lm(formula = mort ~ nonw)

Residuals:
     Min       1Q   Median       3Q      Max
-109.810  -32.757   -4.021   35.053   95.088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 887.0765    10.3723  85.524  < 2e-16 ***
nonw          4.4888     0.7006   6.407 2.88e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.01 on 58 degrees of freedom
Multiple R-squared:  0.4144,	Adjusted R-squared:  0.4043
F-statistic: 41.05 on 1 and 58 DF,  p-value: 2.885e-08
```

Inference

- Variance of $\hat{\beta} = \sigma^2 (X^\top X)^{-1}$, where $\sigma^2 = Var(\epsilon)$
- We can estimate σ using the Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Why the standard error?
Hypothesis test

Inference

- Variance of $\hat{\beta} = \sigma^2(X^\top X)^{-1}$, where $\sigma^2 = Var(\epsilon)$
- We can estimate σ using the Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Why the standard error?

$H_0 : \beta_1 = 0, \quad \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$ has a t-distribution, n-p-1 degrees of freedom

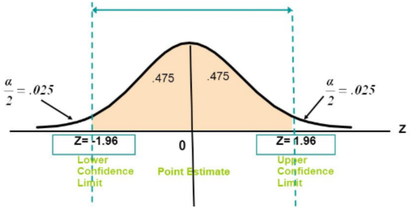
Inference

- Variance of $\hat{\beta} = \sigma^2(X^\top X)^{-1}$, where $\sigma^2 = Var(\epsilon)$
- We can estimate σ using the Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Why the standard error?
95% confidence intervals

$$[\hat{\beta}_j - 1.96SE(\hat{\beta}_j), \hat{\beta}_j + 1.96SE(\hat{\beta}_j)]$$



The lm function in R: what are we looking?

```
> summary(lm(mort~nonw))

Call:
lm(formula = mort ~ nonw)

Residuals:
     Min       1Q   Median       3Q      Max
-109.810  -32.757   -4.021   35.053   95.088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 887.0765    10.3723  85.524  < 2e-16 ***
nonw          4.4888     0.7006   6.407 2.88e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.01 on 58 degrees of freedom
Multiple R-squared:  0.4144,	Adjusted R-squared:  0.4043
F-statistic: 41.05 on 1 and 58 DF,  p-value: 2.885e-08
```

Analysis of fit

- Total Sum of Squares (TSS) or deviance of y

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

- Residual Sum of Squares (RSS)
- R^2

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- $DEV(Y) = DEV(M) + DEV(E)$

$$R^2 = \frac{Dev_m}{Dev_Y} = 1 - \frac{Dev_e}{Dev_Y}$$

Analysis of fit: the F-statistics, another test

Analysis of fit: the F-statistics, another test

- $H_0: \quad \beta_0 = \beta_1 = 0$
- Does the model fit better than a model with only an intercept?

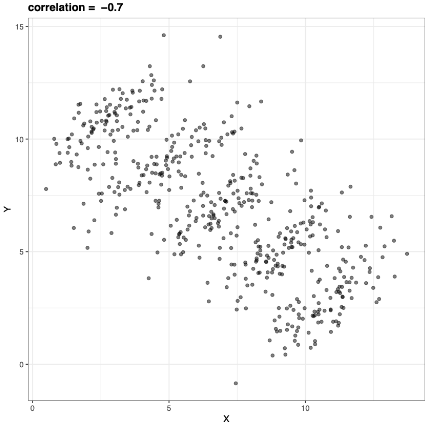
$$F = \frac{SSM/p}{RSS/(n-p-1)}$$

- in our case: $R^2 = 0.62266$, $R^2_{adj} = 0.6066$. Q on Prismia.
- in our case $F = 31.33$, with $p - value \leq 0.05$

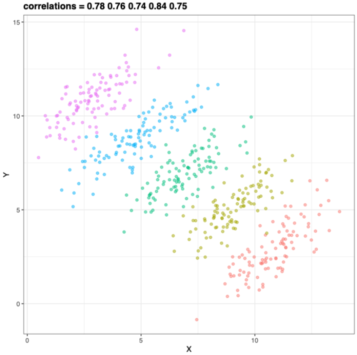
The lm function in R: what are we looking?

```
> summary(lm(mort~nonw))

Call:
lm(formula = mort ~ nonw)

Residuals:
     Min       1Q   Median       3Q      Max
-109.810  -32.757   -4.021   35.053   95.088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 887.0765    10.3723  85.524  < 2e-16 ***
nonw          4.4888     0.7006   6.407 2.88e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.01 on 58 degrees of freedom
Multiple R-squared:  0.4144,	Adjusted R-squared:  0.4043
F-statistic: 41.05 on 1 and 58 DF,  p-value: 2.885e-08
```

Simpson Paradox



Simpson Paradox



Need to account for the effect of the third variable.

Example: Pollution data set

`ID`	code for the identification of the sample
`OVR65`	% of 1960 SMSA population aged 65 or older
`EDUC`	Median school years completed by those over 22
`HOUS`	% of housing units with all facilities
`DENS`	Population per sq. mile in urbanized areas, 1960
`NONW`	% non-white population in urbanized areas, 1960
`WWDRK`	% employed in white collar occupations
`POOR`	% of families with income < 3000
`HC`	Relative hydrocarbon pollution potential
`NOX`	Same for nitric oxides
`SO2`	Same for sulphur dioxide
`HUMID`	Annual average % relative humidity at 1pm
`MORT`	Total age-adjusted mortality rate per 100,000
`PREC`	Average annual precipitation in inches

The lm function in R: what are we looking?

```
Call:
lm(formula = mort ~ nonw + so2 + educ + nonw)

Residuals:
    Min      1Q  Median      3Q     Max
-94.201 -19.410   1.294  16.537  92.986

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1156.06487   71.68018  16.128  < 2e-16 ***
nonw           3.70485    0.58615   6.321 4.55e-08 ***
so2            0.25699    0.08298   3.097 0.003054 **
educ         -24.92413    6.28208  -3.967 0.000209 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.02 on 56 degrees of freedom
Multiple R-squared:  0.6266,	Adjusted R-squared:  0.6066
F-statistic: 31.33 on 3 and 56 DF,  p-value: 5.063e-12
```