Week 2: Linear Models - a Practical Recap

MATH-516 Applied Statistics

Tomas Masak

Feb 20th 2023

Section 1

Data and Intepretation

Data

- data: $(Y_1, Z_1^\top)^\top, \ldots, (Y_N, Z_N^\top)^\top$ where
  - $Z_n \in \mathbb{R}^q$ are explanatory variables
  - Y_n are responses
- model: $\mathbb{E}[Y_n \mid X_n] = \beta_0 f_0(Z_n) + \beta_1 f_1(Z_n) + \ldots + \beta_{p-1} f_{p-1}(Z_n)$ where
  - f_j are known functions
- model matrix:

$$\mathbf{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_N^\top \end{pmatrix}$$

where

$$X_n = \begin{pmatrix} X_{n,0} \\ \vdots \\ X_{n,p-1} \end{pmatrix} = \begin{pmatrix} f_0(Z_n) \\ \vdots \\ f_{p-1}(Z_n) \end{pmatrix}$$

  - X_n is a parametrization of Z_n

Data

- Let $Z_n \in \mathbb{R}$, i.e. there is just a single explanatory variable
  - can be parametrized to many columns of $\mathbf{X}$
- Example 1: polynomial regression of order p in variable Z
  - $\mathbb{E}[Y_N \mid Z_N] = \beta_0 + \beta_1 Z_n + \ldots + \beta_p Z_n$, i.e. the expected value of the response is a polynomial of order p in the explanatory variable
  - $$\mathbf{X} = \begin{pmatrix} 1 & Z_1 & Z_1^2 & \ldots & Z_1^p \\ 1 & Z_2 & Z_2^2 & \ldots & Z_2^p \\ \vdots & & & & \\ 1 & Z_N & Z_N^2 & \ldots & Z_N^p \end{pmatrix}$$

Data

- Example 2: Z_n is a factor, e.g. Z_n is 0 for a child, 1 for a man and 2 for a woman
  - Z has no numerical interpretation $\Rightarrow$ it should be considered as a factor, i.e. every group is allowed to have its own mean
  - the means have to be parametrized somehow, for example:
  - say the model is $\mathbb{E}[Y_n \mid Z_n] = \beta_0 + \beta_1 \mathbb{I}_{[n\text{-th obs is a man}]} + \beta_2 \mathbb{I}_{[n\text{-th obs is a woman}]}$
  - say we have 2 children followed by 2 men and then 2 women in the data
  - $$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

A Single Factor

- Let $Z_n = 1, \ldots, G$ denote a group membership, i.e. it is a factor (and the only variable).
- The largest possible model with only this information allows for different means $\mu_1, \ldots, \mu_G$ for every group.
  - have to be related to variables $\beta_0, \ldots, \beta_{G-1}$
- The naive parametrization:

$$\beta_0 \equiv \mu_1, \ldots, \beta_{G-1} \equiv \mu_G$$

  - the model matrix has rows of the identity matrix (each row replicated by number of observations in that group)
  - does not contain the intercept (an all-one column vector)
  - does not generalize naturally to multiple factors
- other parametrizations possible
  - we choose depending on the interpretation we seek

"contr.treatment" parametrization (the default in R)

- A better parametrization:

$$
\begin{aligned}
\mu_1 &= \beta_0 & \beta_0 &= \mu_1 \\
\mu_2 &= \beta_0 + \beta_1 & \beta_1 &= \mu_2 - \mu_1 \\
&\vdots & &\vdots \\
\mu_G &= \beta_0 + \beta_{G-1} & \beta_G &= \mu_g - \mu_1
\end{aligned}
$$

  - the model matrix has rows

$$
\begin{pmatrix}
1 & 0 & \dots & 0 \\
1 & 1 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 0 & \dots & 1
\end{pmatrix}
$$

  - β_0 is the mean of the first (reference) group
  - for $j = 1, \dots, G-1$, β_j is the difference in means between the j-th group and the reference group

"contr.sum" parametrization

- Another parametrization:

$$
\begin{aligned}
\mu_1 &= \beta_0 + \beta_1 & \beta_0 &= \frac{1}{G}\sum_{g=1}^{G} \mu_g =: \bar{\mu} \\
&\vdots & &\vdots \\
\mu_{G-1} &= \beta_0 + \beta_{G-1} & \beta_1 &= \mu_1 - \bar{\mu} \\
\mu_G &= \beta_0 - \sum_{g=1}^{G-1} \beta_g & \beta_G &= \mu_{G-1} - \bar{\mu}
\end{aligned}
$$

  - the model matrix has rows

$$
\begin{pmatrix}
1 & 1 & \ldots & 0 \\
1 & \vdots & \ddots & \vdots \\
1 & 0 & \ldots & 1 \\
1 & -1 & \ldots & -1
\end{pmatrix}
$$

  - has the advantage of β_0 being the mean of group means

Linear Model

Definition. The data $Y \in \mathbb{R}^N$, $\mathbf{X} \in \mathbb{R}^{N \times p}$ follow a linear model if $Y \mid X \sim (\mathbf{X}\beta, \sigma^2 I)$, that is when $\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}\beta$ and $\mathrm{var}(Y \mid \mathbf{X}) = \sigma^2 \mathbf{I}$

- the model is linear because the dependency on the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$ is linear
- we will assume that the the model is full-rank: $P(\mathrm{rank}(X) = p) = 1$ implying $p \leq n$
- we do not assume Gaussianity here
    - many results require it, but it can be mostly bypassed with asymptotics
- we assume homoscedasticity $(\mathrm{var}(Y \mid \mathbf{X}) = \sigma^2 \mathbf{I})$
    - under heteroscedasticity $(\mathrm{var}(Y_n \mid X_n) = \sigma^2(X_n))$, use *sandwich*
- we do not assume independence here
    - if we have Gaussianity, it follows from uncorrelatedness
    - without Gaussianity, it is crucial to have it for anything else than basic least-squares results such as Gauss-Markov
- the most important assumption is having a correct form for the expectation!

Interpretation

- let $x = (x_1, \ldots, x_p)^\top$ and $\tilde{x}^{(j)} = (x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p)^\top$
- then $\beta_j = \mathbb{E}[Y_n \mid X_n = \tilde{x}^{(j)}] - \mathbb{E}[Y_n \mid X_n = x]$
  - β_j is the expected change in the response when the j-th regressor increases by one
  - the change is multiplicative: a change of x_j by δ suggests a change of Y_n by $\delta\beta_j$
- when there is intercept, then β_0 is the expected value of Y under all other regressors being zero
  - it makes sense to work with centered regressors
- when the j-th regressor is on the log-scale: when $\log(x_j) \mapsto \log(x_j) + 1$, the expected response increases by β_j
  - $\log(x_j) \mapsto \log(x_j) + 1 \Leftrightarrow x_j \mapsto e x_j$
  - it is better to work with base 2 or 10 for the log

Interpretation

If linear model holds for log-transformed response:

- $\log(Y_n) = X_n^\top \beta + \epsilon \Leftrightarrow Y_n = e^{X_n^\top \beta} e^{\epsilon_n}$
- since $\mathbb{E}[Y_n \mid X_n] = e^{X_n^\top \beta} \mathbb{E} e^{\epsilon_n} = e^{X_n^\top \beta + \log(\mathbb{E} e^{\epsilon_n})}$
    - we cannot interpret the intercept, but
    - $\log(x_j) \mapsto \log(x_j) + 1$ can be interpreted as the e^{β_j}-multiplicative increase of the response, because

$$\frac{\mathbb{E}[Y_n \mid X_n = \tilde{x}^{(j)}]}{\mathbb{E}[Y_n \mid X_n = x]} = e^{\beta_j}$$

- for other transformations of the response (e.g. Box-Cox), we do not have such a nice interpretation
    - this is partly why we love logarithmic transformations

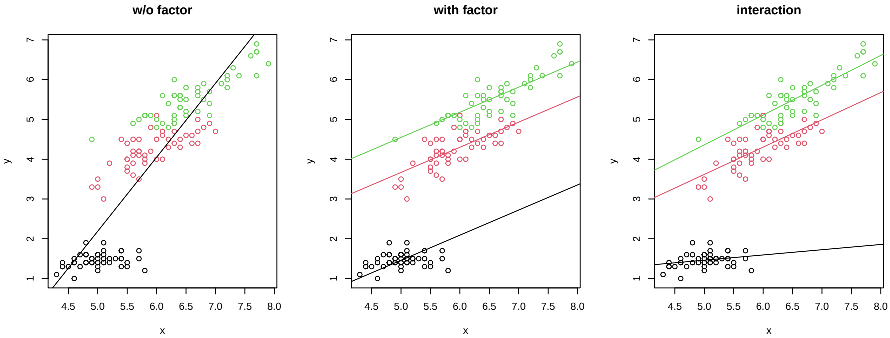
Interactions

- a specific model is given by the model matrix $\mathbf{X}$
- each variable can have a single or numerous corresponding columns of $\mathbf{X}$
  - true both for a numerical variables Z, W and factors A, B
- adding an interaction for two variables means simply to add to the model matrix entry-wise products between all columns of one variable and all columns of the other variable. Adding an interaction...
  - between two numerical variables Z, W has no particular interpretation, for example

$$\mathbb{E}Y_n = \beta_0 + \beta_1 Z_n + \beta_2 W_n \quad \Rightarrow \quad \mathbb{E}Y_n = \beta_0 + \beta_1 Z_n + \beta_2 W_n + \beta_3 Z_n W_n$$

  - between two factors A, B with G_1 and G_2 groups, respectively, creates a partition into $G_1 G_2$ groups
  - between Z and A allows for any form of dependence on Z to be treated separately in the groups given by A (example on next slide)

Example: interaction between a numeric and a factor



- no two lines are exactly parallel on the right-hand plot

Section 2

Least Squares

Projections

- let $\mathcal{M}(\mathbf{X})$ denote the linear space spanned by the columns of $\mathbf{X} \in \mathbb{R}^{N \times p}$
- let $\mathbf{Q}$ be a basis of $\mathbf{X}$ and $\mathbf{P} = (\mathbf{Q} \mid \mathbf{N})$ be the basis of $\mathbb{R}^p$
  - basis $\equiv$ orthonormal basis (for us)

$$\mathbf{I} = \mathbf{P}^\top \mathbf{P} = \mathbf{Q}\mathbf{Q}^\top + \mathbf{N}\mathbf{Q}^\top \mathbf{Q}\mathbf{N}^\top + \mathbf{N}\mathbf{N}^\top = \mathbf{Q}\mathbf{Q}^\top + \mathbf{N}\mathbf{N}^\top =: \mathbf{H} + \mathbf{M}$$

- as projection matrices, $\mathbf{H}$ and $\mathbf{M}$ are
  - unique
  - with eigenvalues 0 or 1
  - symmetric
  - idempotent ($\mathbf{A}\mathbf{A} = \mathbf{A}$)
  - $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ [$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X} = \mathbf{X}$ and properties above]
- hence $Y = \mathbf{I}Y = \mathbf{H}Y + \mathbf{M}Y = \widehat{Y} + E$
  - $\widehat{Y}$ are *fitted values*
  - E are *residuals*

Least Squares

Also follows from the projection properties above (i.e. linear algebra):

$$\widehat{Y} = \operatorname*{arg\,min}_{\widetilde{Y} \in \mathcal{M}(\mathbf{X})} \left\| Y - \widetilde{Y} \right\|_2^2 \qquad \text{or} \qquad \widehat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \| Y - \mathbf{X}\beta \|_2^2$$

Theorem. (Gauss-Markov) Let $Y \mid \mathbf{X} \sim (\mathbf{X}\beta, \sigma^2 \mathbf{I})$, then $\widehat{Y} = \mathbf{H}Y$ is the BLUE (best linear unbiased estimator) of Y and $\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y$ is the BLUE of β.

- $\widehat{Y} \mid \mathbf{X} \sim (0, \sigma^2 \mathbf{H})$ and $E \mid \mathbf{X} \sim (0, \sigma^2 \mathbf{M})$
    - e.g. $\mathbb{E}[E \mid \mathbf{X}] = \mathbf{M}\mathbb{E}Y = \mathbf{M}\mathbf{X}\beta = 0$ and $\operatorname{var}(E \mid \mathbf{X}) = \mathbf{M}\operatorname{var}(Y)\mathbf{M}^\top = \sigma^2 \mathbf{M}\mathbf{M}^\top = \sigma^2 \mathbf{M}$
- hence $s^2 := \|E\|_2^2/(N-p)$ is an unbiased estimator of σ^2
    - since $\mathbb{E}\|E\|_2^2 = \operatorname{tr}(\sigma^2 \mathbf{M}) = \sigma^2 \operatorname{tr}(\mathbf{M}) = \sigma^2 (n-p)$
    - $\|E\|_2^2$ is the residual sum of squares

FWL Theorem

Theorem. Let $\mathbf{X} = (\mathbf{X}_1 \mid \mathbf{X}_2)$ be a partitioned matrix and consider two regressions:

1. $\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$, and
2. $\mathbb{E}[(\mathbf{I} - \mathbf{H}_1)Y|\mathbf{X}_2] = (\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2\gamma_2$, where $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top$.

Then the least squares estimates of β_2 and γ_2 coincide.

- almost no assumptions (the models do not even need to hold), just a property of least squares when working with linear models
- when we add new regressors, we are just trying to explain whatever we failed to explain with the original regressors
  - $(\mathbf{I} - \mathbf{H}_1)Y$ are the residuals from the regression $E[Y \mid \mathbf{X}_1] = \mathbf{X}_1\beta_1$
  - $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2$ is the part of $\mathbf{X}_2$ orthogonal to $\mathbf{X}_1$

Model-Submodel Testing

Definition. Consider two models $M^0 : Y \mid \mathbf{X} \sim (\mathbf{X}^0 \beta^0, \sigma^2 \mathbf{I})$ and $M : Y \mid \mathbf{X} \sim (\mathbf{X}\beta, \sigma^2 \mathbf{I})$. M^0 is a submodel of M if $\mathcal{M}(\mathbf{X}^0) \subset \mathcal{M}(\mathbf{X})$.

- choose a basis $(\mathbf{Q}_0 \mid \mathbf{Q}_1 \mid \mathbf{N})$ in $\mathbb{R}^N$ such that $\mathcal{M}(Q_0) = \mathcal{M}(\mathbf{X}^0)$ and $\mathcal{M}(\mathbf{Q}_0 \mid \mathbf{Q}_1) = \mathcal{M}(\mathbf{X})$
- then $Y = \mathbf{Q}_0 \mathbf{Q}_0^\top Y + \mathbf{Q}_1 \mathbf{Q}_1^\top Y + \mathbf{N}\mathbf{N}^\top Y = \widehat{Y}^0 + \underbrace{D + E}_{E^0} = \widehat{Y} + E$

Theorem. Consider models M and M^0 above and let M^0 hold with the assumption of Gaussianity, i.e. $Y \mid \mathbf{X} \sim \mathcal{N}_n(\mathbf{X}^0 \beta^0, \sigma^2 \mathbf{I})$. Then $\|D\|_2^2 = \|E^0\|_2^2 - \|E\|_2^2$ and

$$F = \frac{\frac{\|E^0\|_2^2 - \|E\|_2^2}{p - p_0}}{\frac{\|E\|_2^2}{N - p}} \sim F_{p-p_0, N-p}$$

Uncertainty Quantification

Theorem. Let $Y \mid \mathbf{X} \sim \mathcal{N}_N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$ and $c \in \mathbb{R}^p$, $c \neq 0$. Then

$$T = \frac{c^\top \widehat{\beta} - c^\top \beta}{\sqrt{s^2 c^\top (\mathbf{X}^\top \mathbf{X})^{-1} c}} \sim t_{N-p}$$

- we can take e.g. $c = (1, 0, \ldots, 0)$ to obtain a CI for the first component of β, etc.
- we can take $c = x_\star$, where $x_\star$ are values of the regressors for a new datum, to obtain a CI for the regression function at a new data point

Uncertainty Quantification (cntd.)

- if we want a CI for $y_\star$ itself, we have to through in the additional uncertainty:
  - under the model: $y_\star = x_\star^\top \beta + \epsilon_\star$ where $\epsilon_\star \sim N(0, \sigma^2)$ is the error, i.e. $y_\star - x_\star^\top \beta \sim N(0, \sigma^2)$
  - from the (proof of) theorem above: $x_\star^\top \widehat{\beta} - x_\star^\top \beta \sim N(0, \sigma^2 x_\star^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_\star)$
  - and the two distributions above are independent (since the new error is independent of everything) hence:

$$y_\star - x_\star^\top \widehat{\beta} \sim N(0, \sigma^2 [1 + x_\star^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_\star])$$

  - and from plugging in the estimator and Cramer-Slutzsky we obtain:

$$\frac{y_\star - x_\star^\top \widehat{\beta}}{\sqrt{s^2 [1 + x_\star^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_\star]}} \sim t_{N-p}$$

    from which we can construct a prediction interval

Asymptotics

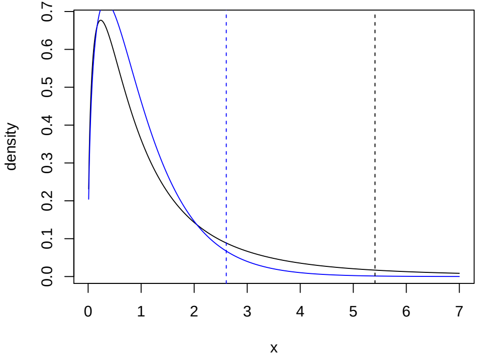
If we do not have Gaussianity (but independence), we can replace:

- the t_{N-p} distribution by the $N(0,1)$ distribution and
- the $F_{p-p_0,N-p}$ distribution by the $\chi^2_{p-p_0}/(p-p_0)$ distribution
  - i.e. doing a likelihood ratio test instead of an F-test

In both cases, relevant quantiles of the asymptotically valid distributions are smaller in magnitude, so using the exact distributions for inference is

- not really a problem for confidence intervals, we are simply being conservative and have wider intervals
  - only use t-tests for CIs, nothing else
- a problem for model-submodel tests, since maybe we should have rejected the submodel, but instead we have accepted

Exact vs. Asymptotic distributions



- black: $F_{3,5}$ distribution and its 95 % quantile (dashed)
- blue: $F_{3,\infty} = \chi^2_3/3$ distribution and its 95 % quantile (dashed)

If the F statistics is between the dashed lines and Gaussianity does not hold, the model-submodel test wrongly arrives to the submodel.

Section 3

Diagnostics

Measures of Model Quality

The first measure of model quality is the `Multiple R-squared`

$$R^2 := 1 - \frac{\|E\|_2^2/N}{\sum_n (Y_n - \bar{Y}_N)^2/N},$$

measuring the proportion of variance explained by the regression.

`Multiple R-squared` always increases with a new predictor added, partly because the two variance estimators in the fraction are biased. `Adjusted R-squared` uses unbiased estimators instead:

$$R^2_{adj} := 1 - \frac{\|E\|_2^2/(N-p)}{\sum_n (Y_n - \bar{Y}_N)^2/(N-1)} = 1 - \frac{\widehat{\sigma}^2}{\sum_n (Y_n - \bar{Y}_N)^2/(N-1)}$$

Still, this tends to favor larger models, so we have

- $AIC = 2N \log(\widehat{\sigma}) + 2p$, which still tends to favor larger models, so
- $AIC_c = AIC + 2p(p+1)/(N-p-1)$
  - note that smaller AIC is better because of a smaller residual variance
  - trade-off between smaller residual variance and the number of predictors

Assumptions to be Checked

1. validity
    - have we "included all relevant predictors"?
    - can we even answer the questions of interest?
2. independence
    - errors have to be independent (or uncorrelated under Gaussianity)
3. linearity
    - correct form for the expectation?
4. homoscedasticity
    - errors have the same variance
5. Gaussianity
    - are the errors

Also, we should check potentially problematic observations (outliers and leverage points).

How to Check the Assumptions

1. validity
   - we cannot really do much about this once data are given to us, but we should always think critically
2. independence
   - this can only be checked in a rather specific cases (whether some subgroups of observations are correlated or whether there is serial dependence in time, provided time matters)
3. linearity
   - plot residuals against regressors, there should be no patterns
   - FWL theorem!
4. homoscedasticity
   - plot (standardized) residuals against fitted values, there should be no pattern
5. Gaussianity
   - QQ-plot and/or histogram of the residuals

One can also perform statistical tests.

Problematic Observations

- $\mathrm{var}(Y - \widehat{Y}) = \sigma^2(\mathbf{I} - \mathbf{H}) \Rightarrow \mathrm{var}(Y_n - \widehat{Y}_n) = \sigma^2(1 - h_{nn})$
- $\mathrm{tr}(\mathbf{H}) = \sum_n h_{nn} = p$ is the no. of model degrees of freedom (no. of parameters)
- h_{nn} is called the leverage of n-th observation
- if h_{nn} is large, it means that a single obs. is usurping too much of the model fit freedom to itself $\Rightarrow$ potential problems (the n-th obs.is called a leverage point)
- if E_n is large in magnitude, the model does not fit the n-th obs. well $\Rightarrow$ potential problems (the n-th obs.is called an outlier)
- when the n-th obs. is outlier AND leverage point $\Rightarrow$ problems!
- Cook's statistic combines the two notions:

$$C_n = \frac{E_n^2 h_{nn}}{p(1 - h_{nn})}$$

- plot (standardized) residuals against leverages and draw some Cook's contours (ROT: $8/(N-2p)$ or $4/N$) to see what's what

Section 4

Linear Models in R

Important Functions

- `model <- lm(formula, data)` estimates a linear model given by `formula` (next slide) specifying a parametrization for a data frame `data`, returns a fitted model object
- `plot(model, which=1:6)` shows 6 default residual plots for a fitted `model`
- `summary(model)` produces summary information for a fitted `model`
- `resid(model)` extracts the residuals
- `fitted(model)` extracts the fitted values
- `predict(model, newdata)` obtain predicted values from a fitted `model` for the values of the regressors specified in `newdata`
- `anova(model)` or `anova(m1,m2)` provides F-tests either between two models `m1` and `m2` or sequentially adding variables to an intercept-only model until `model`

Note: apart from `lm` itself, all function names should end with `.lm`, e.g. `plot.lm()`, but this can be omitted when called on a `lm` object (such as `model` above).

`lm()` model formula

Consider an example call: `lm(y ~ x + I(x^2) + a*b + w:z -1)`

- `y`,`x`, `a`, `b`, `w` and `z` are names of variables in the data frame
- `~` separates the response variable on the LHS from the regressors on the RHS
- `+` is really an "and", specifies that the model depends on whatever is on the left and on the right of `+`
- `:` adds an interaction between `w` and `z`, i.e. adds to the model matrix all element-wise products between all columns corresponding to `x` and all columns corresponsing to `y`
- `*` is an interaction with the main terms, i.e. `a*b` $\equiv$ `a + b + a:b`
  - since we basically never want an interaction without main terms, `*` is much more useful than `:`
- `I()` without this `x^2` would be added as `x` (stupid), so `I()` is mostly used to allow for polynomial dependencies
- `-1` specifies that there should be no intercept (otherwise there is by default)
- `.` includes on the RHS all but the response variable (specified on the LHS)

Model summary

Stupid example: (fitting a quadratic polynomial to intercept plus noise)

```
y <- 1+rnorm(100)
x <- 1:100
model <- lm(y ~ x + I(x^2))
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3408 -0.6419 -0.1258  0.7880  2.7498
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.3966068  0.3215390   4.344 3.45e-05 ***
## x           -0.0183678  0.0146952  -1.250    0.214
## I(x^2)       0.0001935  0.0001410   1.373    0.173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.05 on 97 degrees of freedom
## Multiple R-squared:  0.02009,    Adjusted R-squared:  -0.0001153
## F-statistic: 0.9943 on 2 and 97 DF,  p-value: 0.3737
```

Model summary

- `Call` repeats the model formula
- `Residuals` provides a `summary` of the residuals
- `Coefficients` provides a table with the estimates, their standard errors, values of the T-statistic and p-values of the student t-tests, and also significance codes for a visual appeal
  - one can have the first clues about which variables are significant from the t-tests, but it should always be decided by `anova` whether a variable should be dropped, e.g. here one would fit `submodel <- lm(y~1)` and call `anova(model,submodel)` to see whether the quadratic dependence on `x` can be removed
- `Residual standard error` gives $\widehat{\sigma}$ where $\widehat{\sigma}^2 = \|E\|_2^2/(n-p)$ and $n-p$ are the `degrees of freedom`
- `Multiple R-squared` and `Adjusted R-squared` are self-explanatory
- `F-statistic` for the model-submodel test between the model and intercept-only model
  - i.e. exactly `anova(model,submodel)` from the few lines above
  - informally tests whether the model is of any use

Overview

- linear models are fitted by least squares
- CIs and model-submodel tests are exact given Gaussianity
- prediction intervals are easy to compute analytically
- residuals allow us to check different model assumptions

Section 5

Practical Modeling

Model Building

- either manual or automated (forward/backward elimination, criterion must be chosen)
- possible criteria:
  - model-submodel testing
  - R^2, R^2_{adj}, AIC, AIC_c (and many others)
  - prediction error

Depending on what we want...

Inference	Prediction
Statistics	Machine Learning
Manual	Automated
Simple Models	(Mixures of) Complicated Models
Model-submodel Tests	Prediction Error
AIC_c	R^2

Manual Meta-algorithm (modified from Prof. Davison)

- explore data
  - standardization?
  - can suggest transformations for response and/or regressors
- consider what models are coherent with the problems/questions
  - variables of a particular interest?
- iterate:
  - fit models, compare their quality (comes next)
  - interpret model parameters
  - check fit (comes next)
- provide conclusions
  - careful interpretation of the best model(s) in terms of the original problem
  - consider deficiencies

Model Checking/Comparison

1. residual diagnostics
   - are our assumptions satisfied?
2. sensitivity/stability inspection
   - how much inference/conclusions change when model changes to another plausible one?
   - what if some special observations are omitted or different transformations used?
3. predictive checking
   - does our model provide good/reasonable predictions?

Section 6

Example: CEO Salaries

Data & Objective

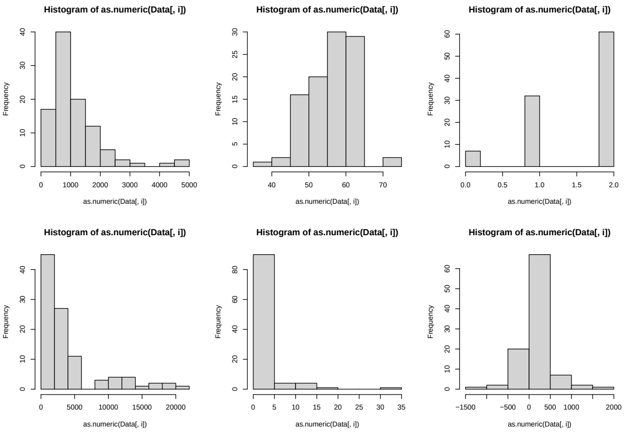
Data: from Forbes (1992) on 100 of the largest firms in the US:

- `comp` - CEO salary
- `age` - CEO age
- `educatn` - CEO education
- `pcntown` - percentage of firm owned by the CEO
- `sales` - firm's sales
- `prof` - firm's profits
  - some other variables also available, but we will not consider them here

Goal: assess the effect of education

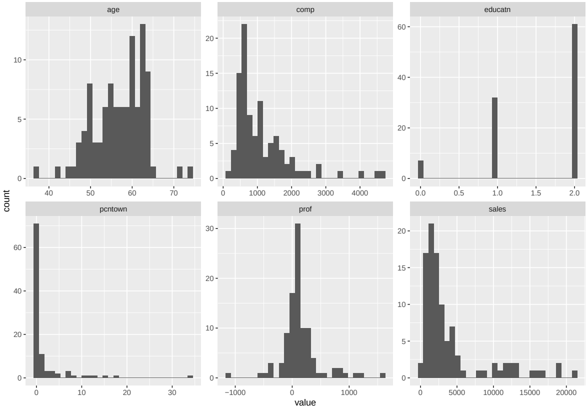
Data Exploration - Histograms with Base R

```
Data <- read.csv("../Project-0/CEO_compensations.csv")
names(Data) <- tolower(names(Data)) # variable names to lower-case
Data <- Data[,c(1,2,3,7,9,10)]
par(mfrow=c(2,3))
for(i in 1:6) hist(as.numeric(Data[,i]))
```



Data Exploration - Histograms with tidyverse

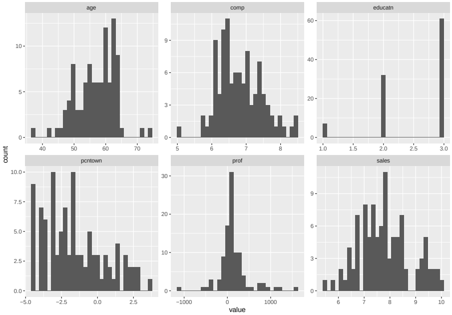
```
library(tidyverse)
Data <- read.csv("../Project-0/CEO_compensations.csv")
names(Data) <- tolower(names(Data))
Data <- Data %>% select(comp, age, educatn, pcntown, sales, prof)
Data %>% pivot_longer(everything()) %>% ggplot(aes(value)) +
  facet_wrap(~ name, scales = "free") + geom_histogram()
```



Transformations

- log-tranformation for `comp` and `sales` seems an obvious choice
- `pcntown` unclear since these are percentages (=0? let's try…)
  - if some were indeed 0, should we create an additional factor?
- education should be considered a factor

```
Data <- Data %>% mutate(comp=log(comp),
                        sales=log(sales),
                        pcntown=log(pcntown),
                        educatn=as.factor(educatn))
```



Anova Table - Type I

- fit a model with all the variables and test for their significance using model-submodel tests
- however, `anova()` does this sequentially (not entirely useful, since it depends on variable ordering)

```
m1 <- lm(comp~., data=Data)
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: comp
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## age        1  2.2198  2.2198  7.9000  0.006028 **
## educatn    2  2.2332  1.1166  3.9738  0.022079 *
## pcntown    1  0.1192  0.1192  0.4240  0.516531
## sales      1  8.0855  8.0855 28.7750 5.917e-07 ***
## prof       1  1.2846  1.2846  4.5718  0.035123 *
## Residuals 93 26.1321  0.2810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova Table - Type II

- instead, we would like to see what happens when we drop a single variable out of the model

```
library(car)
m1 <- lm(comp~., data=Data)
Anova(m1,type=2)
```

```
## Anova Table (Type II tests)
##
## Response: comp
##            Sum Sq Df F value    Pr(>F)
## age        0.5903  1  2.1008   0.15058
## educatn    0.6027  2  1.0725   0.34635
## pcntown    0.3310  1  1.1779   0.28060
## sales      5.7370  1 20.4170 1.828e-05 ***
## prof       1.2846  1  4.5718   0.03512 *
## Residuals 26.1321 93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction

- add interactions between `educatn` (variable of interest) and other variables
- `Anova(m1,type=2)` suggests only `educatn*age` is significant

Test this manually:

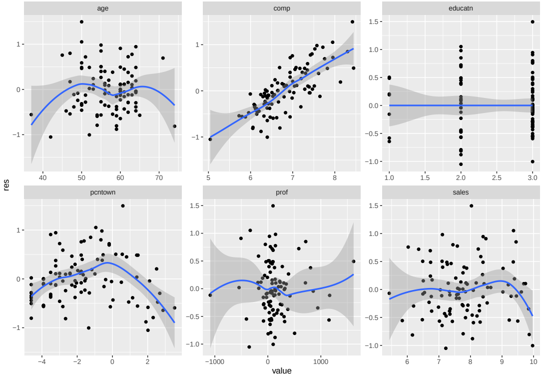
```
m3 <- lm(comp~.*educatn, data=Data)
m2 <- lm(comp~.+educatn:age, data=Data)
anova(m1,m2,m3)
```

```
## Analysis of Variance Table
##
## Model 1: comp ~ age + educatn + pcntown + sales + prof
## Model 2: comp ~ age + educatn + pcntown + sales + prof + educatn:age
## Model 3: comp ~ (age + educatn + pcntown + sales + prof) * educatn
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     93 26.132
## 2     91 24.079  2    2.0535 3.8127 0.02596 *
## 3     85 22.890  6    1.1886 0.7357 0.62228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics

Model 2 seems to be good, let's check the residual plots

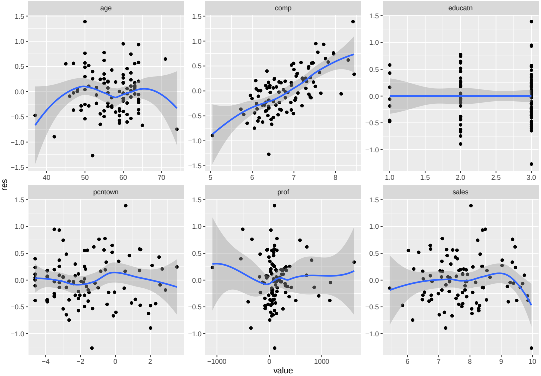
```
Data %>% mutate(res=resid(m2), educatn=as.numeric(educatn)) %>%
  pivot_longer(-res) %>% ggplot(aes(y=res,x=value)) +
  facet_wrap(~ name, scales = "free") + geom_point() + geom_smooth()
```



Diagnostic

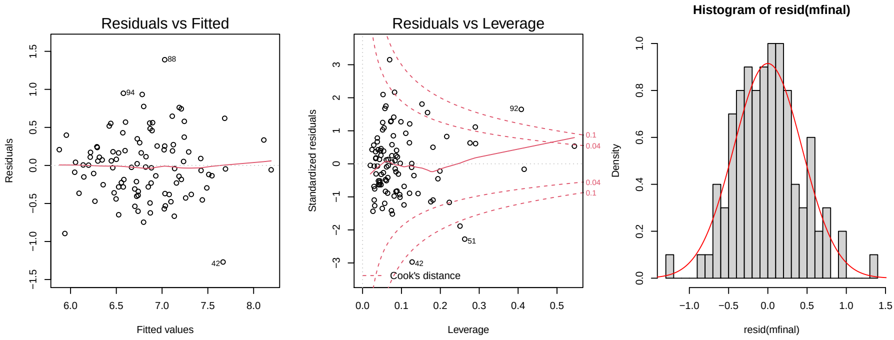
- allow for a quadratic dependence on pcntown

```
mfinal <- lm(comp~.+educatn:age+I(pcntown^2), data=Data)
Data %>% mutate(res=resid(mfinal), educatn=as.numeric(educatn)) %>%
  pivot_longer(-res) %>% ggplot(aes(y=res,x=value)) +
  facet_wrap(~ name, scales = "free") + geom_point() + geom_smooth()
```



Diagnostics

```
par(mfrow=c(1,3))
N <- dim(Data)[1]
p <- length(coef(mfinal))
plot(mfinal,c(1,5),cook.levels=c(8/(N-2*p), 4/N))
hist(resid(mfinal),freq=F, breaks=20)
points(-300:300/100,dnorm(-300:300/100,0,sd(resid(mfinal))),
       type="l",col="red")
```



Interpretation of the model w.r.t. education

```
summary(mfinal)
```

```
##
## Call:
## lm(formula = comp ~ . + educatn:age + I(pcntown^2), data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26944 -0.29867 -0.02302  0.24793  1.38984
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0366250  3.3675641  -0.011   0.9913
## age           0.0839109  0.0562090   1.493   0.1390
## educatn1      3.0167698  3.4603900   0.872   0.3856
## educatn2      4.9056981  3.4191980   1.435   0.1548
## pcntown      -0.0659205  0.0340684  -1.935   0.0561 .
## sales         0.3069079  0.0536386   5.722 1.37e-07 ***
## prof          0.0003272  0.0001475   2.218   0.0291 *
## I(pcntown^2) -0.0509142  0.0101490  -5.017 2.63e-06 ***
## age:educatn1 -0.0554089  0.0575977  -0.962   0.3386
## age:educatn2 -0.0895200  0.0568906  -1.574   0.1191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4572 on 90 degrees of freedom
## Multiple R-squared:  0.5305, Adjusted R-squared:  0.4835
## F-statistic:  11.3 on 9 and 90 DF,  p-value: 1.219e-11
```

Interpretation of the model w.r.t. education

```
Data$age <- Data$age - mean(Data$age) # mean(age)~57
mfinal <- lm(comp~.+educatn:age+I(pcntown^2), data=Data)
summary(mfinal)
```

```
##
## Call:
## lm(formula = comp ~ . + educatn:age + I(pcntown^2), data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26944 -0.29867 -0.02302  0.24793  1.38984
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7404199  0.4722648  10.038 2.39e-16 ***
## age           0.0839109  0.0562090   1.493   0.1390
## educatn1     -0.1376569  0.2695196  -0.511   0.6108
## educatn2     -0.1906744  0.2699710  -0.706   0.4818
## pcntown      -0.0659205  0.0340684  -1.935   0.0561 .
## sales         0.3069079  0.0536386   5.722 1.37e-07 ***
## prof          0.0003272  0.0001475   2.218   0.0291 *
## I(pcntown^2) -0.0509142  0.0101490  -5.017 2.63e-06 ***
## age:educatn1 -0.0554089  0.0575977  -0.962   0.3386
## age:educatn2 -0.0895200  0.0568906  -1.574   0.1191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4572 on 90 degrees of freedom
## Multiple R-squared:  0.5305, Adjusted R-squared:  0.4835
## F-statistic:  11.3 on 9 and 90 DF,  p-value: 1.219e-11
```

Rest

For sensitivity inspection, predictive checking, and more careful interpretation, check out `./Project-0/Project-0.html`

- there is also a `rough_work` script and a separate `cv-script`
- you can use this as a guidance for your project reports

References (for the 1st half of this course)

- Venables & Ripley (2002) Modern Applied Statistics with S (4th ed.)
  - while S is the predecessor of R, it has basically the same syntax (though some packages went some way since 2002)
  - an amazing reference (though a bit hard to swallow with little previous exposition to the material)
- Wood (2017) Generalized Additive Models: an Introduction with R (2nd ed.)
  - even though mainly about GAMs, this book has a short and practical exposition to linear models and GLMs that has a value of its own
  - computational flavor
- Davison (2003) Statistical Models
  - nice reference due to the breadth, more self-contained than Venables & Ripley, but no `R` code
- Gelman & Hill (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models
  - focuses very much on interpretation
  - somehow an opposite of Venables & Ripley in that it is eloquent/lengthy and not always to the point (or precise)
- Wickham & Grolemund (2017) R for Data Science
  - useful guide to `tidyverse`, i.e. data exploration and manipulation