Session: September 20, 2016

- Linear models and ANOVA
  - Factors in R
  - ANOVA table
  - Sums of Squares
- Automatic model selection
- Model quality measures
- Model extensions

Linear regression and ANOVA

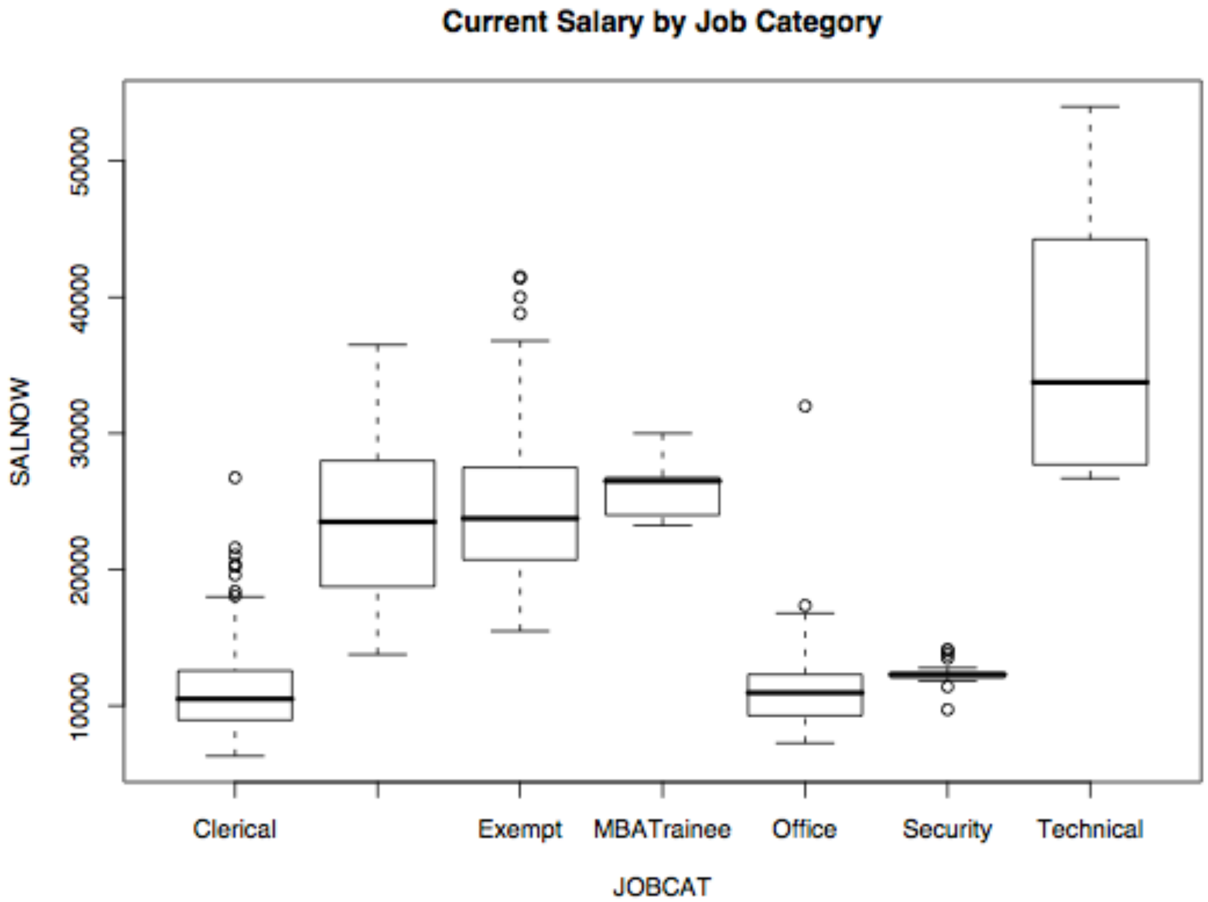
- Traditionally there was a clear differentiation between linear regression and ANOVA (analysis of variance)
- Linear regression = continuous predictors
- ANOVA = categorical predictors (experimental set-up)

- Technically, they are the same
- In praxis, most of the times you have mixed predictors
- Software accepts both kinds of predictors
- Manually, via dummy coding

- However: the devil is in the details

Regression with categorical predictors

- Ex: Bank data
  - Salary depending on job category



Linear regression and ANOVA

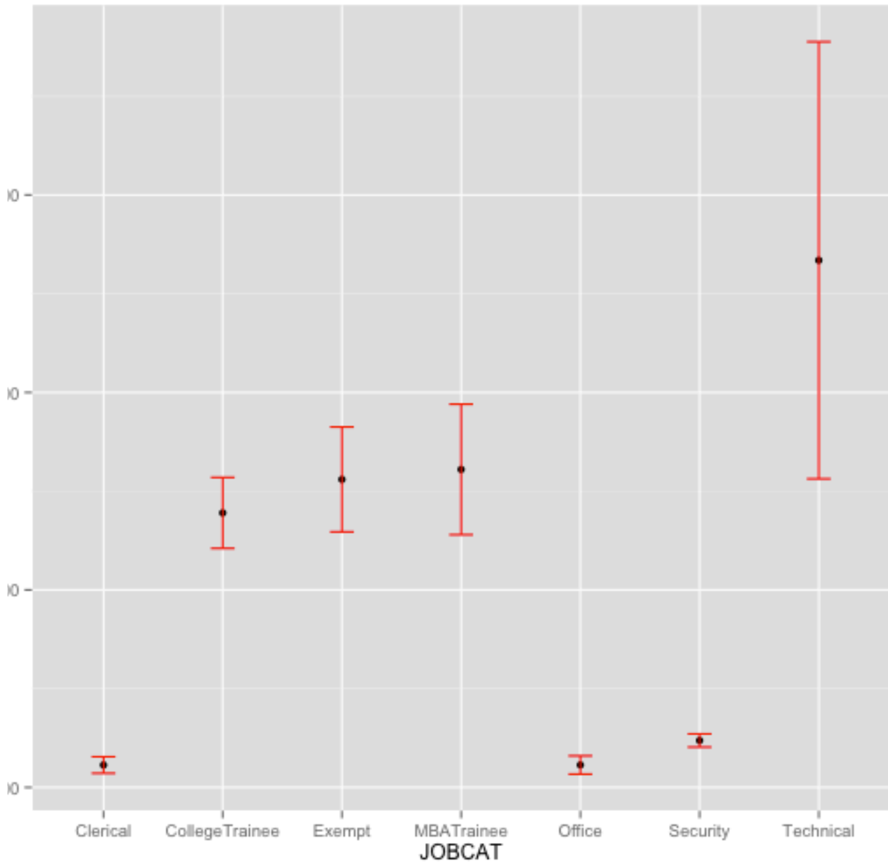
- Categorical predictor with more than two categories
- Let's use job category as a predictor

```
Call:
lm(formula = SALNOW ~ JOBCAT, data = bank)

Residuals:
     Min       1Q   Median       3Q      Max 
-10137.1  -2136.4   -454.8   1405.0  20863.6 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       11134.819    255.102  43.649   <2e-16 ***
JOBCATCollegeTrainee 12766.254    652.213  19.574   <2e-16 ***
JOBCATExempt      14460.806    725.753  19.925   <2e-16 ***
JOBCATMBATrainee  14965.181   1737.692   8.612   <2e-16 ***
JOBCATOffice          1.592    416.771   0.004    0.997    
JOBCATSecurity     1240.736    782.436   1.586    0.113    
JOBCATTechnical   25556.847   1589.703  16.076   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3843 on 467 degrees of freedom
Multiple R-squared: 0.6874, Adjusted R-squared: 0.6834 
F-statistic: 171.1 on 6 and 467 DF,  p-value: < 2.2e-16
```



Factors in R

- If we treat a variable as a factor, R includes an intercept and omits the alphabetically first level of the factor.
- The intercept is the estimated mean for the reference level.
- The intercept t-test tests for whether or not the mean for the reference level is 0.
- All other t-tests are for comparisons of the other levels versus the reference level.
- Other group means are obtained the intercept plus their coefficient.
- If we omit an intercept, then it includes terms for all levels of the factor.
- Group means are now the coefficients.
- Tests are tests of whether the groups are different than zero.
- If we want comparisons between two levels, neither of which is the reference level, we could refit the model with one of them as the reference level.

Linear regression and ANOVA

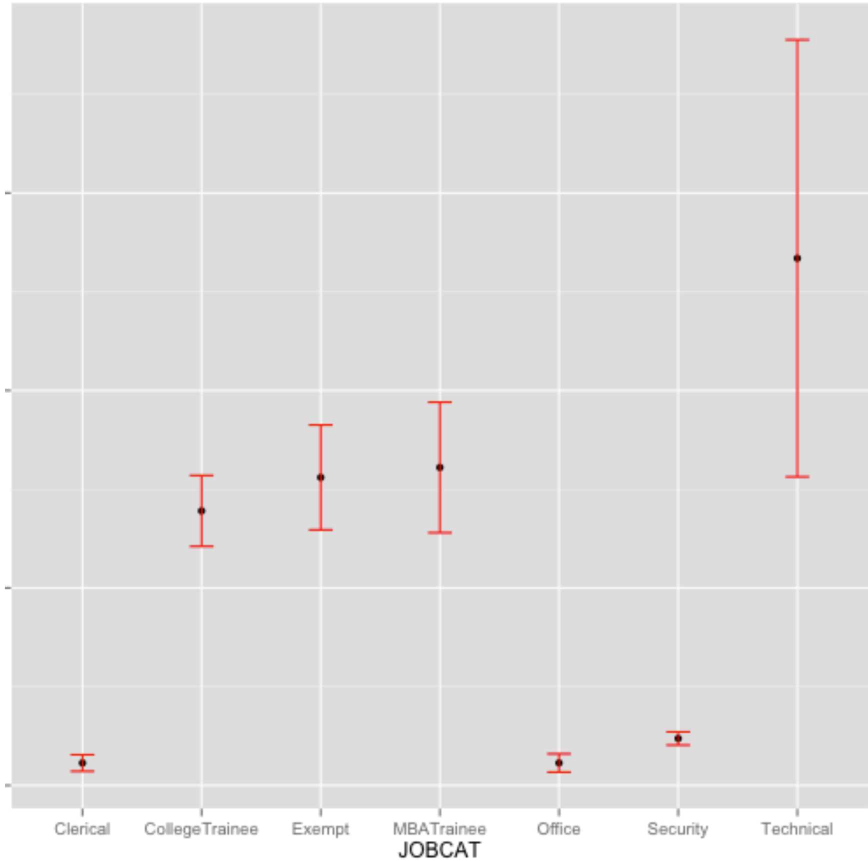
- Let’s use job category as a predictor without intercept

```
Call:
lm(formula = SALNOW ~ JOBCAT - 1, data = bank)

Residuals:
     Min       1Q   Median       3Q      Max
-10137.1  -2136.4   -454.8   1405.0  20863.6

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
JOBCATClerical        11134.8      255.1   43.65   <2e-16 ***
JOBCATCollegeTrainee  23901.1      600.3   39.82   <2e-16 ***
JOBCATExempt          25595.6      679.4   37.67   <2e-16 ***
JOBCATMBATrainee      26100.0     1718.9   15.18   <2e-16 ***
JOBCATOffice          11136.4      329.6   33.79   <2e-16 ***
JOBCATSecurity        12375.6      739.7   16.73   <2e-16 ***
JOBCATTechnical       36691.7     1569.1   23.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3843 on 467 degrees of freedom
Multiple R-squared:  0.9384,    Adjusted R-squared:  0.9374
F-statistic:  1016 on 7 and 467 DF,  p-value: < 2.2e-16
```



Linear regression and ANOVA

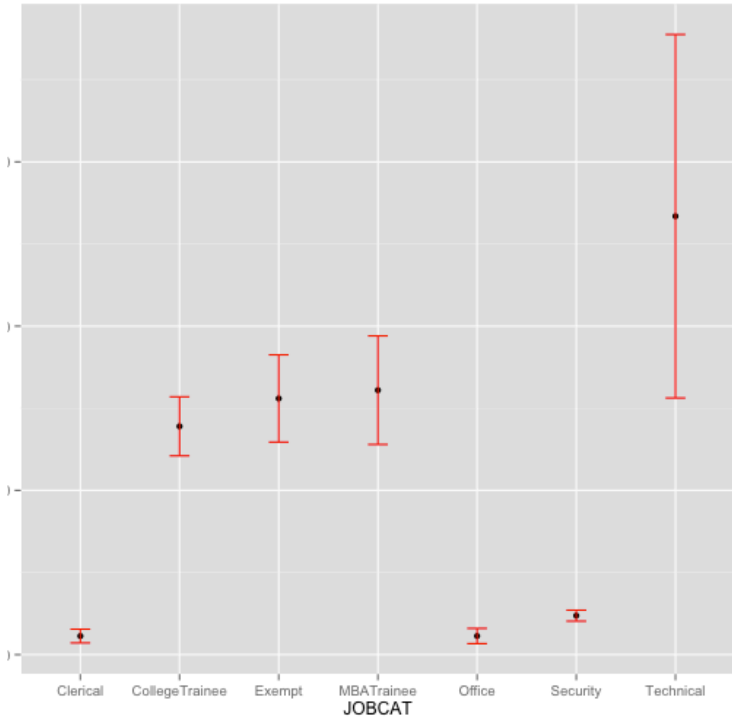
- Changing the reference category
- Let's use job category security as reference category
- bank$JOBCAT <- relevel(bank$JOBCAT, ref="Security")

```
Call:
lm(formula = SALNOW ~ JOBCAT, data = bank)

Residuals:
     Min       1Q   Median       3Q      Max 
-10137.1  -2136.4   -454.8   1405.0  20863.6 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           12375.6      739.7  16.731  < 2e-16 ***
JOBCATClerical        -1240.7      782.4  -1.586    0.113    
JOBCATCollegeTrainee  11525.5      952.6  12.099  < 2e-16 ***
JOBCATExempt          13220.1     1004.4  13.162  < 2e-16 ***
JOBCATMBATrainee      13724.4     1871.3   7.334 9.92e-13 ***
JOBCATOffice          -1239.1      809.8  -1.530    0.127    
JOBCATTechnical       24316.1     1734.7  14.017  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3843 on 467 degrees of freedom
Multiple R-squared:  0.6874,	Adjusted R-squared:  0.6834 
F-statistic: 171.1 on 6 and 467 DF,  p-value: < 2.2e-16
```



Linear regression and ANOVA

- For categorical predictor: slopes are identical to difference in group means from reference category
- Intercept corresponds to average salary for reference category, here: clerical
- Based on “alphabetical order”

```
> coef(bank.lm10)
        (Intercept) JOBCATCollegeTrainee         JOBCATExempt     JOBCATMBATrainee         JOBCATOffice
       11134.819383         12766.253787         14460.805617         14965.180617             1.592381
     JOBCATSecurity      JOBCATTechnical
        1240.736172         25556.847283
> bank.jobcat-bank.jobcat[1]
     Clerical  CollegeTrainee          Exempt      MBATrainee          Office        Security       Technical
     0.000000    12766.253787    14460.805617    14965.180617        1.592381     1240.736172    25556.847283
```

Linear regression and ANOVA

- To assess whether categorical predictor is statistically significant we prefer to have a summary assessment instead of significance of individual coefficients
- Hence we look at ANOVA table

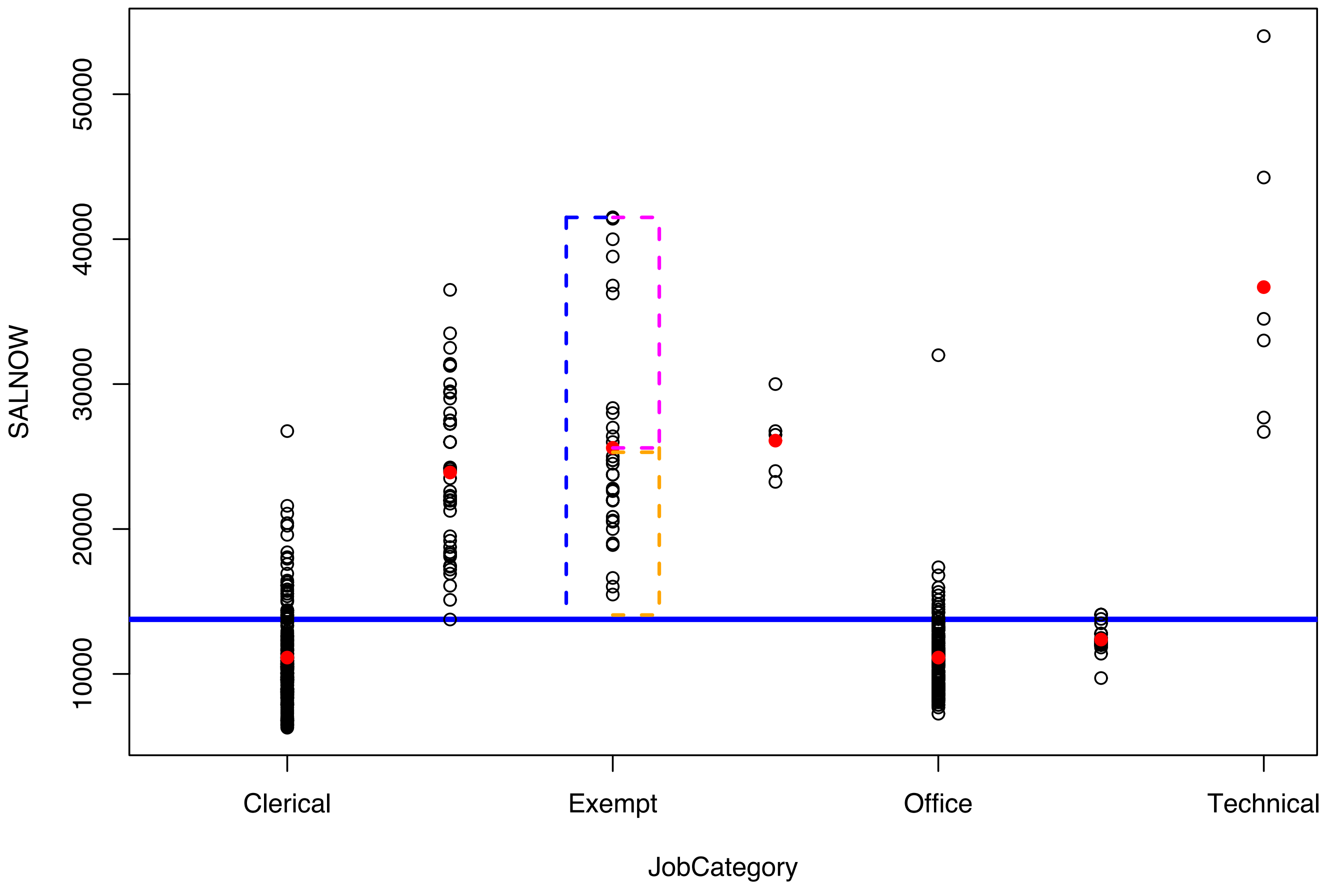
```
> anova(bank.lm10)
Analysis of Variance Table

Response: SALNOW
           Df     Sum Sq    Mean Sq F value    Pr(>F)
JOBCAT      6 1.5168e+10 2527982082  171.13 < 2.2e-16 ***
Residuals 467 6.8987e+09   14772477
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear regression and ANOVA

- How is ANOVA table derived?
- Split of total variation into between-groups and within-groups variation

Do you remember?



Linear regression and ANOVA

- Split of total variation into between-groups and within-groups variation

$$x_{ij} - \bar{x}_{..} = (x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}_{..})$$

$$\begin{aligned} SS_T &= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 \\ &= \sum_{i=1}^{g} n_i (\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2. \end{aligned}$$

Do you remember?

$$SS_T = SS_B + SS_W$$

Linear regression and ANOVA

- Notation used:

Do you remember?

n_i	size of $i-$th group
$n = \sum_{i=1}^{g} n_i$	total sample size
$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$	mean of $i-$th group
$\bar{x}_{..} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} x_{ij}$	(grand) mean (overall or total mean)
SS_B	sums of squares between groups
SS_W	sums of squares within groups
SS_T	total sums of squares

Linear regression and ANOVA

- ANOVA just tests one hypothesis per predictor
- Nothing new for continuous predictors (i.e. one slope per predictor)
- For categorical predictor ANOVA only tells us that there are some group differences
- From ANOVA table alone, we do not know which groups differ
- To get individual differences either look at coefficients from regression output or use post-hoc test

Linear regression and ANOVA

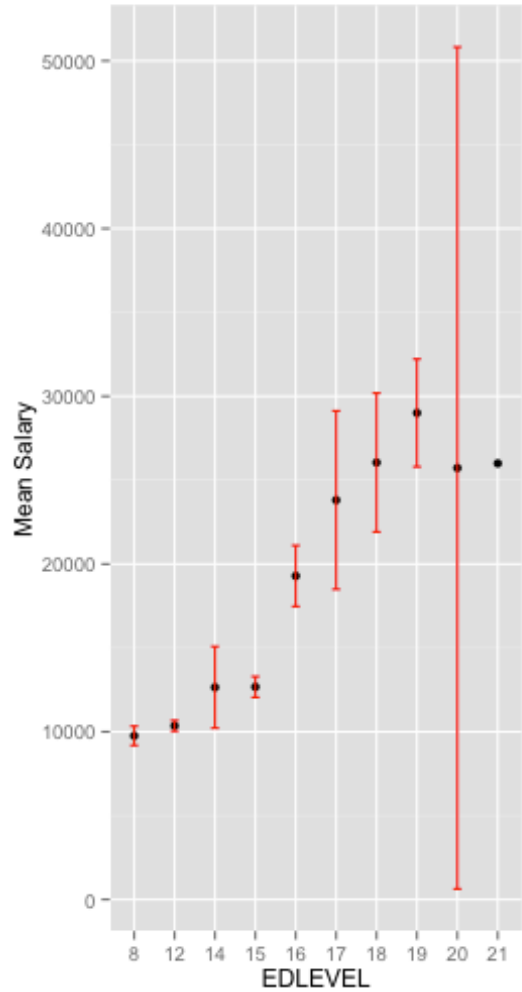
- Using categorical predictor as factor is different from using it as numeric variable

```
Call:
lm(formula = SALNOW ~ factor(EDLEVEL), data = bank)

Residuals:
     Min       1Q   Median       3Q      Max 
-14608.1  -1925.4   -484.9   1626.6  24991.9 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         9759.6      566.1  17.239  < 2e-16 ***
factor(EDLEVEL)12    595.2      640.3   0.930 0.353010    
factor(EDLEVEL)14   2890.4     1775.3   1.628 0.104183    
factor(EDLEVEL)15   2914.4      683.3   4.265 2.42e-05 ***
factor(EDLEVEL)16   9530.8      780.0  12.219  < 2e-16 ***
factor(EDLEVEL)17  14051.3     1365.6  10.290  < 2e-16 ***
factor(EDLEVEL)18  16291.5     1485.9  10.964  < 2e-16 ***
factor(EDLEVEL)19  19248.5      974.5  19.752  < 2e-16 ***
factor(EDLEVEL)20  15965.4     2968.9   5.378 1.20e-07 ***
factor(EDLEVEL)21  16240.4     4160.3   3.904 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4122 on 464 degrees of freedom
Multiple R-squared: 0.6428, Adjusted R-squared: 0.6359 
F-statistic: 92.78 on 9 and 464 DF,  p-value: < 2.2e-16
```



Linear regression and ANOVA

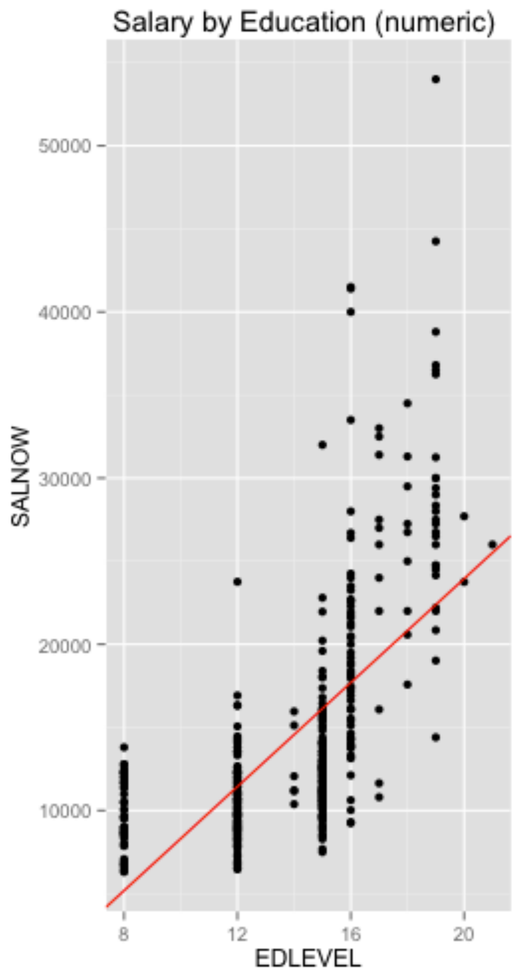
- Using categorical predictor as factor is different from using it as numeric variable

```
Call:
lm(formula = SALNOW ~ EDLEVEL, data = bank)

Residuals:
   Min     1Q Median     3Q    Max 
 -8627  -3284  -1001   2351  31617 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7332.47    1128.76  -6.496  2.1e-10 ***
EDLEVEL      1563.96      81.82  19.115  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5133 on 472 degrees of freedom
Multiple R-squared: 0.4363, Adjusted R-squared: 0.4351
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```



ANOVA-Table

- Analysis of Variance Table
- Partition of total variability as measured by sum of squares
- For continuous or binary predictors: same p-values as in regression coefficient table
- For categorical predictors (> 2 categories): summarize impact of factor in one score
- F-values are just the squares of corresponding t-values

- $SS_{Total} = SS_{Regression} + RSS$ (residual sum of squares)
- $R^2 = SS_{Regression} / SS_{Total}$

  $= 1 – RSS / SS_{Total}$

Recap: Linear regression and ANOVA

- ANOVA just tests one hypothesis per predictor
- Nothing new for continuous predictors (i.e. one slope per predictor)
- For categorical predictor ANOVA only tells us that there are some group differences
- From ANOVA table alone, we do not know which groups differ
- To get individual differences either look at coefficients from regression output or use post-hoc test
- For categorical predictors, linear model tests effect of difference from reference category
  - This is necessary due to overparametrization
  - There exist different standard parametrizations of the same "overall model"
- Mathematically, the overall model is uniquely defined, but not the individual contributions of each predictor
- Different ways of splitting impact between predictors

ANOVA table

- Remember: Main source of information is variability
- General idea of statistics: split variability into systematic (=explainable) part and random fluctuation
- Variance, standard deviation and other measures of variability depend on sum of squared differences from mean
- Model quality measures such as R-squared also depend on sum of squared differences from mean
- In a linear model, different ways of assigning overall variability of response to the individual predictors
- -> different partitions of sum of squares

Sum of squares partitions

Let us look at the two-way full factorial ANOVA model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

- tests for interaction and man effects can be constructed by the incremental sum of squares approach
- $SS(\alpha, \beta, (\alpha\beta))$ denotes sum of squares for the full model
- $SS(\alpha, \beta)$ denotes sum of squares for the no-interaction model
- $SS(\alpha)$ denotes sum of squares for the one-way ANOVA model

Sum of squares

incremental sum of squares are given by differences between sums of squares for alternative models

$$
\begin{aligned}
SS((\alpha\beta)|\alpha,\beta) &= SS(\alpha,\beta,(\alpha\beta)) - SS(\alpha,\beta) \\
SS(\alpha|\beta,(\alpha\beta)) &= SS(\alpha,\beta,(\alpha\beta)) - SS(\beta,(\alpha\beta)) \\
SS(\beta|\alpha,(\alpha\beta)) &= SS(\alpha,\beta,(\alpha\beta)) - SS(\alpha,(\alpha\beta)) \\
SS(\alpha|\beta) &= SS(\alpha,\beta) - SS(\beta) \\
SS(\beta|\alpha) &= SS(\alpha,\beta) - SS(\alpha)
\end{aligned}
$$

We read $SS((\alpha\beta)|\alpha,\beta)$ as the sum of squares for interaction after the main effects
and $SS(\alpha,\beta)$ as the sum of squares for the row main effect after the column main effect ignoring the interaction

Sum of squares types

Type I "sequential": $SS(\alpha), SS(\beta|\alpha)$ and $SS((\alpha\beta)|\alpha, \beta$ do not provide an appropriate test for the row main effect (one-way ANOVA)

Type II $SS(\alpha|\beta)$ and $SS(\beta|\alpha)$ for main-effects (more powerful if interactions are absent)

Type III "orthogonal" $SS(\alpha|\beta, (\alpha\beta))$ and $SS(\beta|\alpha, (\alpha\beta))$ straight-forward, if interaction is present (default in SPSS)

Type IV same as Type III as long as there are no empty cells

anova() uses Type I,
Anova() in package **car** offers Type II and III

Example: Birthweights

- Source: Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression. New York: Wiley
- Data: The data were collected at Baystate Medical Center, Springfield, Mass during 1986.
- Description of the variables.
  - low: indicator of birth weight less than 2.5 kg.
  - age: mother’s age in years
  - lwt: mother’s weight in pounds at last menstrual period race mother’s
  - race (1 = white, 2 = black, 3 = other)
  - smoke: smoking status during pregnancy
  - ptl: number of previous premature labours
  - ht: history of hypertension
  - ui: presence of uterine irritability
  - ftv: number of physician visits during the first trimester
  - bwt: birthweight in grams
- data(birthwt, package="MASS")

Example: Birthweights

- Running a linear regression
  - birthwt.ols <- lm(b
- summary(birthwt.ols)

```
Call:
lm(formula = bwt ~ . - low, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max 
-1816.51  -426.79    16.29   492.06  1654.01 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3129.4594   344.2424   9.091  < 2e-16 ***
age           -0.2658     9.5947  -0.028  0.97793    
lwt            3.4351     1.6999   2.021  0.04478 *  
race        -188.4895    57.7339  -3.265  0.00131 ** 
smoke       -358.4552   107.5172  -3.334  0.00104 ** 
ptl          -51.1526   103.0003  -0.497  0.62006    
ht          -600.6465   204.3454  -2.939  0.00372 ** 
ui          -511.2513   140.2792  -3.645  0.00035 ***
ftv          -15.5358    46.9354  -0.331  0.74103    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 656.9 on 180 degrees of freedom
Multiple R-squared:  0.223,	Adjusted R-squared:  0.1884 
F-statistic: 6.456 on 8 and 180 DF,  p-value: 2.232e-07
```

p-values based on regression t-test coincide with Type II sum of squares F-tests

Example: Birthweights

- ```
  > Anova(birthwt.ols)
  Anova Table (Type II tests)

  Response: bwt
             Sum Sq  Df F value    Pr(>F)
  age           331   1  0.0008 0.9779291
  lwt       1762311   1  4.0836 0.0447838 *
  race      4599967   1 10.6589 0.0013112 **
  smoke     4796844   1 11.1151 0.0010396 **
  ptl        106439   1  0.2466 0.6200592
  ht        3728637   1  8.6399 0.0037201 **
  ui        5732239   1 13.2826 0.0003503 ***
  ftv         47283   1  0.1096 0.7410265
  Residuals 77680946 180
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  ```

Provides partition of total variation (sum of squares) due to (accounted) contribution of each predictor
  ```

Example: Birthweights

```
> anova(birthwt.ols)
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value    Pr(>F)
age         1   815483  815483  1.8896 0.1709544
lwt         1  2967339 2967339  6.8758 0.0094853 **
race        1  2545071 2545071  5.8974 0.0161473 *
smoke       1  6513374 6513374 15.0926 0.0001437 ***
ptl         1   754368  754368  1.7480 0.1878060
ht          1  2937814 2937814  6.8074 0.0098415 **
ui          1  5707978 5707978 13.2264 0.0003603 ***
ftv         1    47283   47283  0.1096 0.7410265
Residuals 180 77680946  431561
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Birthweights

- Real strength of anova() command comes when comparing nested models
- Assume there are two competitive models: set of predictors for first model is a subset of predictors for the second model

```
> anova(birthwt.ols2,birthwt.ols)
Analysis of Variance Table

Model 1: bwt ~ age + lwt + race + smoke + ptl + ht
Model 2: bwt ~ (low + age + lwt + race + smoke + ptl + ht + ui + ftv) -
    low
  Res.Df      RSS Df Sum of Sq     F   Pr(>F)
1    182 83436207
2    180 77680946  2   5755261 6.668 0.001608 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we want to have sequential sum of squares and a test between the models

Assessment of model quality

- Multiple Correlation Coefficient: R
  - Pearson correlation coefficient between observed and predicted response values
- Coefficient of Multiple Determination: R^2
  - Percent variability explaine[illegible] model
  - $R^2 = 1 – RSS/SS_{Total}$
  - RSS = residual sum of squares = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Adjusted $R^2 = R^2 - p/(n-p+1)[1- R^2] = 1- MSE/Var(Y)$
  - Includes relative complexity of the model
  - Corrects bias towards sample prediction equation
  - Can decrease when we add explanatory variable

```
Multiple R-squared:  0.4879,    Adjusted R-squared:  0.3903
```

Other measures of model quality

- *Akaike Information Criterion AIC*
- *Bayesian Information Criterion BIC* (p = number of parameters in model, n = number of cases)
- penalize complexity of model
- The smaller, the better.

$$
\begin{aligned}
AIC &= -2\log \text{likelihood} + 2 \cdot p \\
&= 2\log(\frac{1}{n}RSS) + 2 \cdot p \qquad \text{for OLS}
\end{aligned}
$$

$$
\begin{aligned}
BIC &= -2\log \text{likelihood} + \log n \cdot p \\
&= 2\log(\frac{1}{n}RSS) + \log n \cdot p \qquad \text{for OLS}
\end{aligned}
$$

Causality

- Regression does not prove causality!
- Choice of DV and IV already implies the causal direction!
- For causality in observational data analysis you need:
  - Statistical correlation
  - Temporal order
  - All alternative explanations are ruled out
- Post hoc, ergo propter hoc (logical fallcy)
  - *The drunk scientist conducts an experiment to see why he gets hangovers. He decides to keep a diary.*
    - *Monday night, scotch and soda; Tuesday morning, hangover.*
    - *Tuesday night, gin and soda; Wednesday morning, hangover.*
    - *Wednesday night: vodka and soda; Thursday morning, hangover.*
    - *Thursday night, rum and soda; Friday morning, hangover.*
    - *On Friday night before going out for a drink, the drunk scientist has an epiphany. "Aha!" he says to himself, "I've got it! **Soda** causes hangovers!"*
- *All models are wrong. But some models are useful!*

George E.P. Box

Guidelines for Model and variable selection

Include enough explanatory variables

- model should be useful for theoretical and predictive purposes
- model building process should allow to exclude alternative explanations for causality
  - Spurious relationship
  - Conditional relationship
  - Intervening variables

KISS principle

Model building

- Theoretical approaches vs. exploratory approaches
- Theoretical approach: aims at testing a specific model to decide about impact of some predictor(s) while controlling for others
- Exploratory approach: given a set of potential predictors find the best model

Exploratory model building: Automatic Selection Procedures

- Backward Elimination: Start with all predictors, remove non-significant predictors (one at each step) until model contains only significant predictors
- Variable deleted at each stage is the one that yields smallest decrease in R^2 , *AIC or BIC*.
- A variable once removed remains out

Exploratory model building: Automatic Selection Procedures

- Forward Selection: Starts with no predictor, adds one variable at a time until no further significant partial contribution can be found
- Variable included at each stage is the one that yields largest boost in R^2 , *AIC or BIC*.
- A variable once entered remains in the model

Exploratory model building: Automatic Selection Procedures

- Stepwise Regression: Starts as forward selection, but after each addition, it checks whether some variable no longer makes a significant partial contribution
- A variable once entered may be removed later

Exploratory model building: Automatic Selection Procedures

- Require some exploratory aspect of research
- Multiple comparisons
- Collinearity yields arbitrary results

Example: Birthweights

- Making variable race a factor
  - birthwt$race <- factor(birthwt$race, labels=c('white','black','other'))
- Running a linear regression
  - birthwt.ols <- lm(bwt~ . -low, data=birthwt)
- Running a stepwise model selection
  - birthwt.ols.best <- stepwise(birthwt.ols, direction='backward/forward', criterion='BIC')
  - summary(birthwt.ols.best)
  - anova(birthwt.ols, birthwt.ols.best, test="F")

Example: Birthweights

```
Call:
lm(formula = bwt ~ lwt + race + smoke + ht + ui, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max 
-1842.14  -433.19    67.09   459.21  1631.03 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2837.264    243.676  11.644  < 2e-16 ***
lwt            4.242      1.675   2.532 0.012198 *  
raceblack   -475.058    145.603  -3.263 0.001318 ** 
raceother   -348.150    112.361  -3.099 0.002254 ** 
smoke       -356.321    103.444  -3.445 0.000710 ***
ht          -585.193    199.644  -2.931 0.003810 ** 
ui          -525.524    134.675  -3.902 0.000134 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 645.9 on 182 degrees of freedom
Multiple R-squared:  0.2404,	Adjusted R-squared:  0.2154 
F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

```
> anova(birthwt.ols, birthwt.ols.best, test="F")
Analysis of Variance Table

Model 1: bwt ~ (low + age + lwt + race + smoke + ptl + ht + ui + ftv) - 
    low
Model 2: bwt ~ lwt + race + smoke + ht + ui
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    179 75702317                           
2    182 75937505 -3   -235188 0.1854 0.9062
```

Checking Model Assumptions

- Main assumptions for linear models (and ANOVA, t-test)
  - Normality of residuals
  - Linearity of relationship
  - Homoscedasticity
  - Independence of cases
  - No Multi-collinearity (i.e. predictors need not be completely linearly dependent)

- Some other general data quality assumptions should hold as well
  - No outliers
  - Accurate measurements
  - Sufficient sample size

Checking Model Assumptions

- Checking for Normality
  - Q-Q plots
  - Kolmogorv-Smirnov test
  - Shapiro-Wilks test
  - ….
- Checking for Homoscedasticity
  - Plotting response against predictor
  - Computing variance ratio (for categorical predictors), rule of thumb: max variance ratio smaller than three
  - Bartlett test
  - Levene’s test
  - Variance test
  - Plotting residuals against predictor, fitted, …

Checking Model Assumptions

- There are many more regression diagnostics, see overview by John Fox or here
  - Checking also for linearity
  - Leverage effects
  - Outliers or other unusual observations

- Check for Multi-collinearity
  - Regress each explanatory variable on others, if any R^2 value is close to 1, then multi-collinearity exists
  - Huge changes in regression coefficient if new variable is included signals multi-collinearity
  - Variance Inflation Factor (VIF) (various rules of thumb: > 4, 5, 10)

Model extensions

- In practice, linear regression assumptions are often violated
  - Dependent variable not normally distributed
    - Solution: GLM (later)
  - Relationship not linear
    - Transformation
    - Inclusion of quadratic (polynomial) effects
    - Curve fitting
  - Heteroscedasticity
    - Transformations
    - Econometrics
  - Correlation of residuals
    - Auto-correlation, Time series
    - Spatial dependencies

Polynomial effects

- Are just handled as additional effects
- Same assessment as for “regular” coefficients for statistical significance
- When looking at (practical) effect of predictor, combine linear and other (e.g. quadratic) effects

Example: UN Demography

```
Call:
lm(formula = tfr ~ l2gdp + illiteracyFemale + +contraception +
    region + I(illiteracyFemale^2), data = UN.all)

 Analysis of Variance Table

 Response: tfr
                        Df  Sum Sq Mean Sq  F value    Pr(>F)
 l2gdp                   1 108.918 108.918 186.0705 < 2.2e-16 ***
 illiteracyFemale        1 103.610 103.610 177.0024 < 2.2e-16 ***
 contraception           1  23.380  23.380  39.9414 5.866e-09 ***
 region                  4  25.472   6.368  10.8787 1.854e-07 ***
 I(illiteracyFemale^2)   1   2.418   2.418   4.1302   0.04456 *
 Residuals             109  63.804   0.585
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: UN Demography

```
Call:
lm(formula = tfr ~ l2gdp + illiteracyFemale + +contraception +
    region + I(illiteracyFemale^2), data = UN.all)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7421 -0.4780  0.0124  0.4268  2.1731 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           5.5173536  0.5100543  10.817  < 2e-16 ***
l2gdp                -0.1124340  0.0466507  -2.410 0.017621 *  
illiteracyFemale      0.0419655  0.0110936   3.783 0.000254 ***
contraception        -0.0273566  0.0046271  -5.912 3.93e-08 ***
regionAmerica        -0.3259714  0.2485698  -1.311 0.192482    
regionAsia           -0.4927969  0.2087354  -2.361 0.020009 *  
regionEurope         -1.6343947  0.3345177  -4.886 3.56e-06 ***
regionOceania        -0.0018046  0.3609227  -0.005 0.996020    
I(illiteracyFemale^2) -0.0002594  0.0001277  -2.032 0.044557 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7651 on 109 degrees of freedom
  (89 observations deleted due to missingness)
Multiple R-squared:  0.8052,	Adjusted R-squared:  0.7909 
F-statistic: 56.33 on 8 and 109 DF,  p-value: < 2.2e-16
```

Summary

- ANOVA and Linear regression
- ANOVA table and coefficient table
  - Incremental sums of squares
- R^2 , AIC and BIC
- Automatic variable selection procedures
- Checking assumptions of linear models
- Model extensions

- Thanks for your attention!