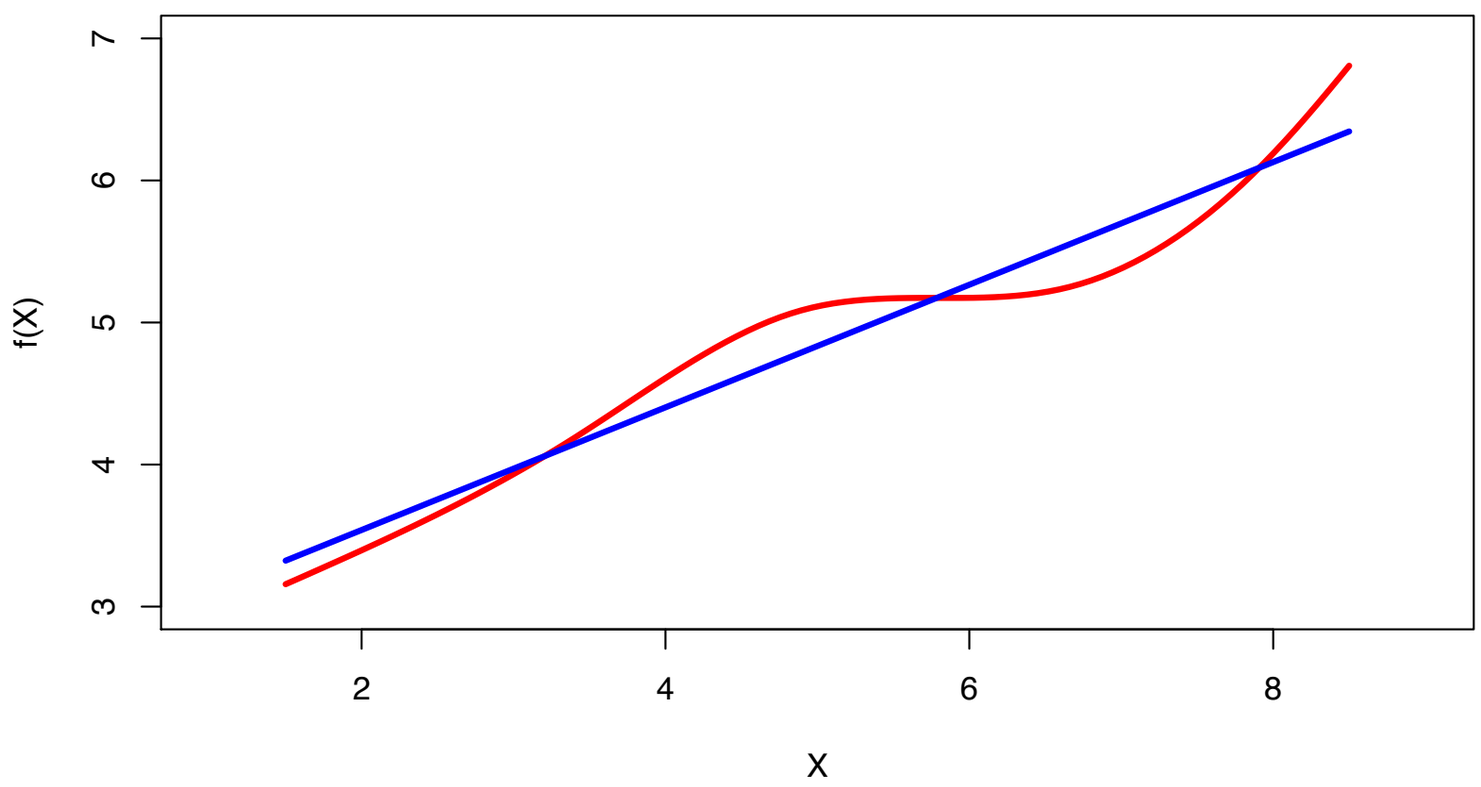
Linear methods for regression

STAT 6312

Department of mathematics, UTA

Linear regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on X_1, X_2, . . . X_p is linear.
- True regression functions are never linear!



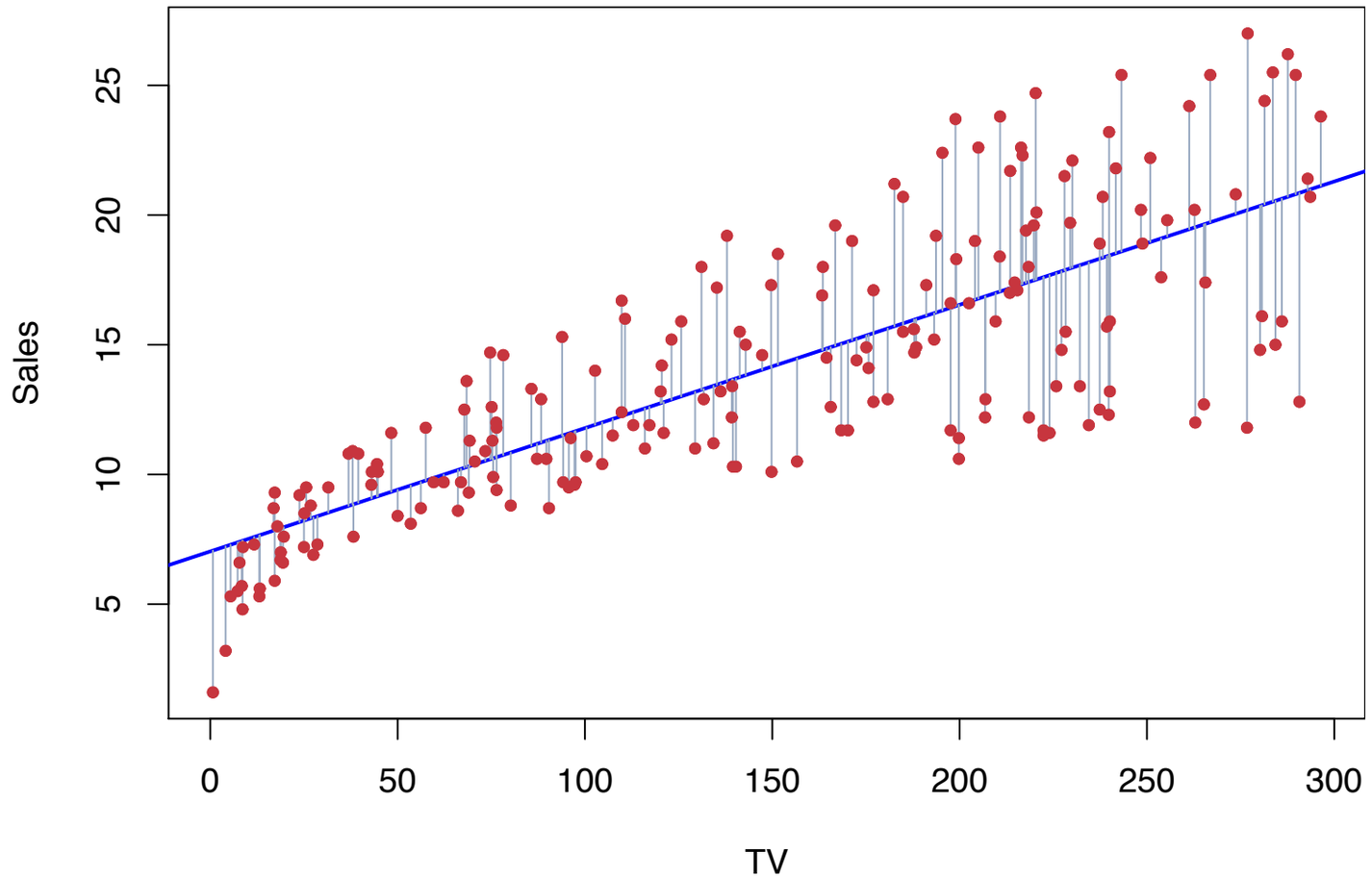
Simple linear regression using a single predictor X

- Model: $Y = \beta_0 + \beta_1 X + \varepsilon$,
  - β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ε is the error term.
- Given estimates for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

  - ``y hat'' indicates a prediction of Y on the basis of X=x.
  - The hat symbol denotes an estimated value.
  - A unit change in X is associated with a β change in Y.

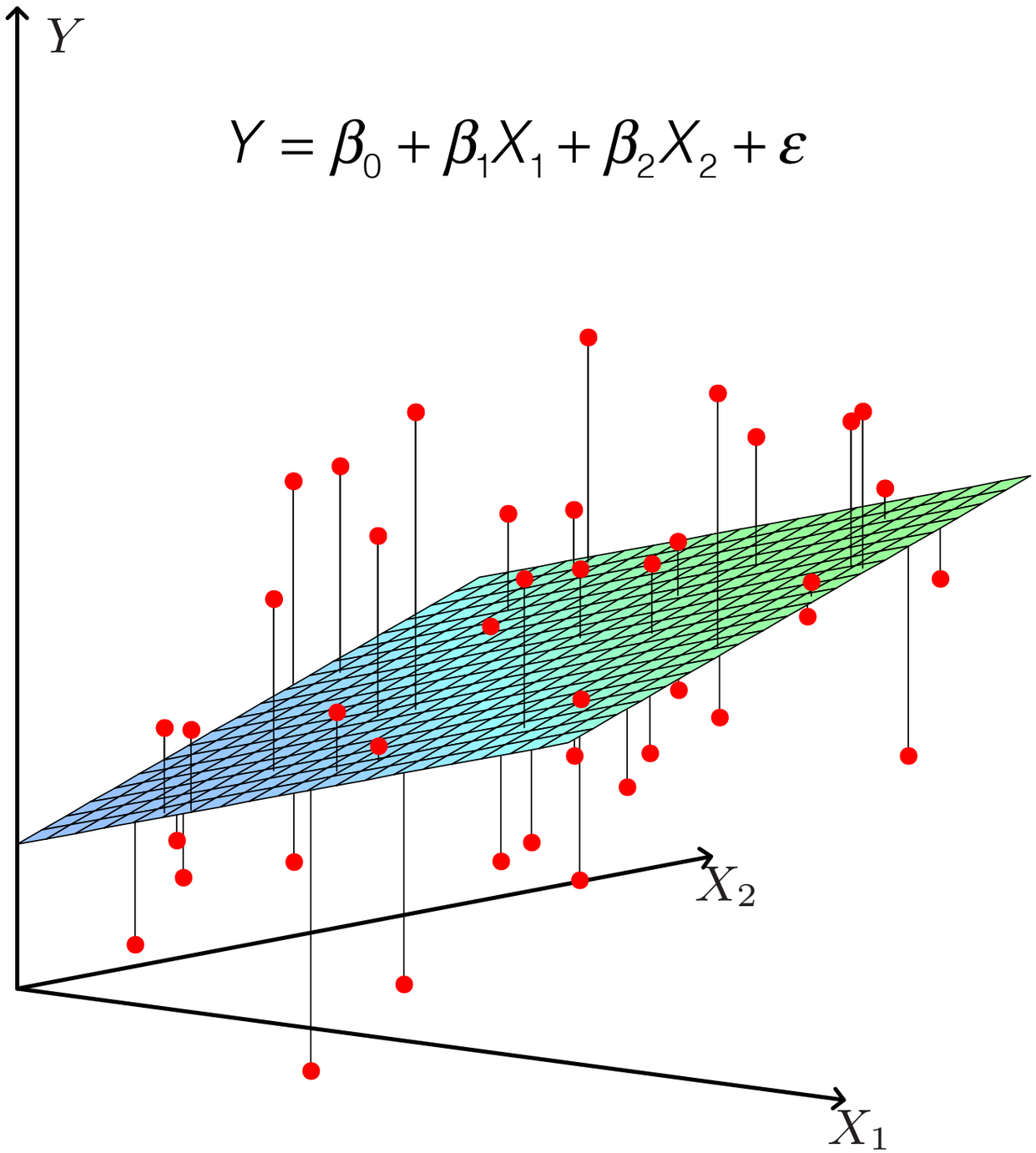
- Residual: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS): $\sum_{i=1}^{N} e_i^2$
- Example: advertising data



- Suppose you can a single predictor X. We suspect that Y and X have non-linear relationship.
  - Q. Can you use the linear regression?
  - We can also consider
    - Smoothing spline, local regression, generalized additive model, etc.
  - There are some benefits assuming the linear regression function although it may not be true.

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$



In general,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as "a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed", are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically (multicollinearity problem).
  - Interpretations become hazardous — when X_j changes, everything else changes.
- Claims of causality should be avoided for observational study.

Matrix representation of linear regression

- Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \; i = 1,\ldots,N$$

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_p \end{pmatrix}'$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{pmatrix}'$$

  - β denotes a vector of regression coefficients, X denotes a design matrix, and ε denotes a vector of random errors

- We want to find β which minimizes (least square method)

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Least square method

- 1st and 2nd derivatives of RSS is given as

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial^2 \mathrm{RSS}}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}.$$

- If X is full-rank, the 2nd derivative is positive definite. (Q: why this matters?)

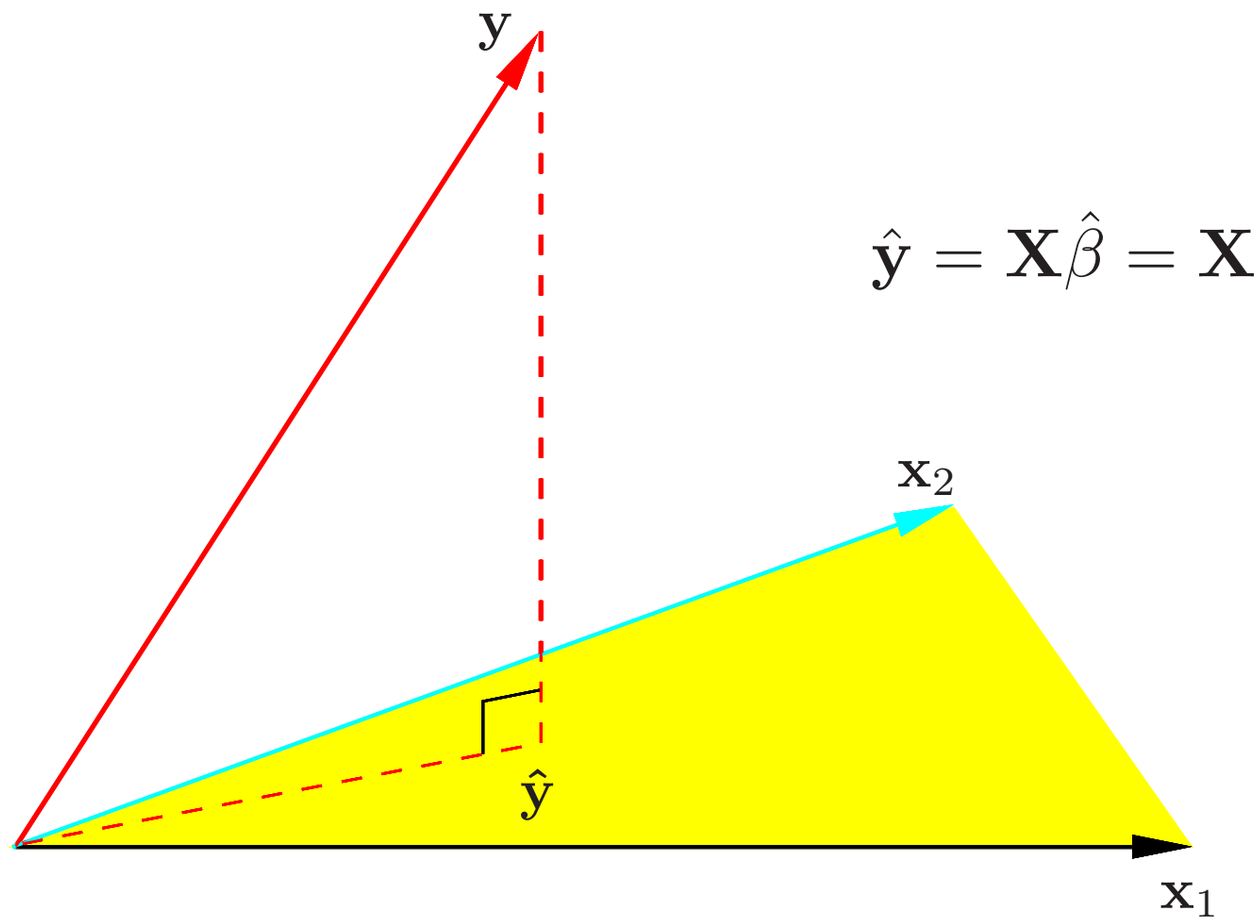
- We set the 1st derivative to zero, and solve it.

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \quad \longrightarrow \quad \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Geometry of the least squares regression

- A matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the "hat" matrix. it is a (orthogonal) projection matrix that projects **y** onto the column space of **X**.



$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Variance-covariance matrix

- For a random vector $y \in \mathbf{R}^N$ with mean $\mu \in \mathbf{R}^N$, the covariance matrix Σ is defined as

$$\Sigma = E[(y - \mu)(y - \mu)']$$

$$\Sigma = (\mathrm{cov}(y_i, y_j))_{ij}$$

- For a matrix $\mathbf{A} \in \mathbf{R}^{mxN}$, the covariance matrix of $\mathbf{A}y$ is

$$\mathrm{var}(\mathbf{A}\mathbf{y}) = \mathbf{A}\Sigma\mathbf{A}'$$

- For a vector $a \in \mathbf{R}^N$, the variance of $a^T y$ is $a^T \Sigma a$.

Uncertainty of estimation

- Suppose a vector of response y has with mean **X**β and variance σ^2**I**

- Then, $\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.

  - You can obtain the variance from $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

- If the error ε is Gaussian,

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2).$$

- The variance estimate is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N - p - 1}$$

- The numerator can be rewritten as

$$(\mathbf{y} - \mathbf{H}\mathbf{y})'(y - \mathbf{H}\mathbf{y}) = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$$

- It can be shown that (Rank(**I**-**H**) = n-p-1)

$$\frac{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}}{\sigma^2} = \frac{\hat{\sigma}^2}{\sigma^2} \sim (N - p - 1)\chi^2_{N-p-1}$$

- Also, $\hat{\sigma}^2$ and $\hat{\beta}$ are independent.
  - Mainly, it is because $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{H}) = 0$.

Hypothesis testing

- The z-score is used to test whether

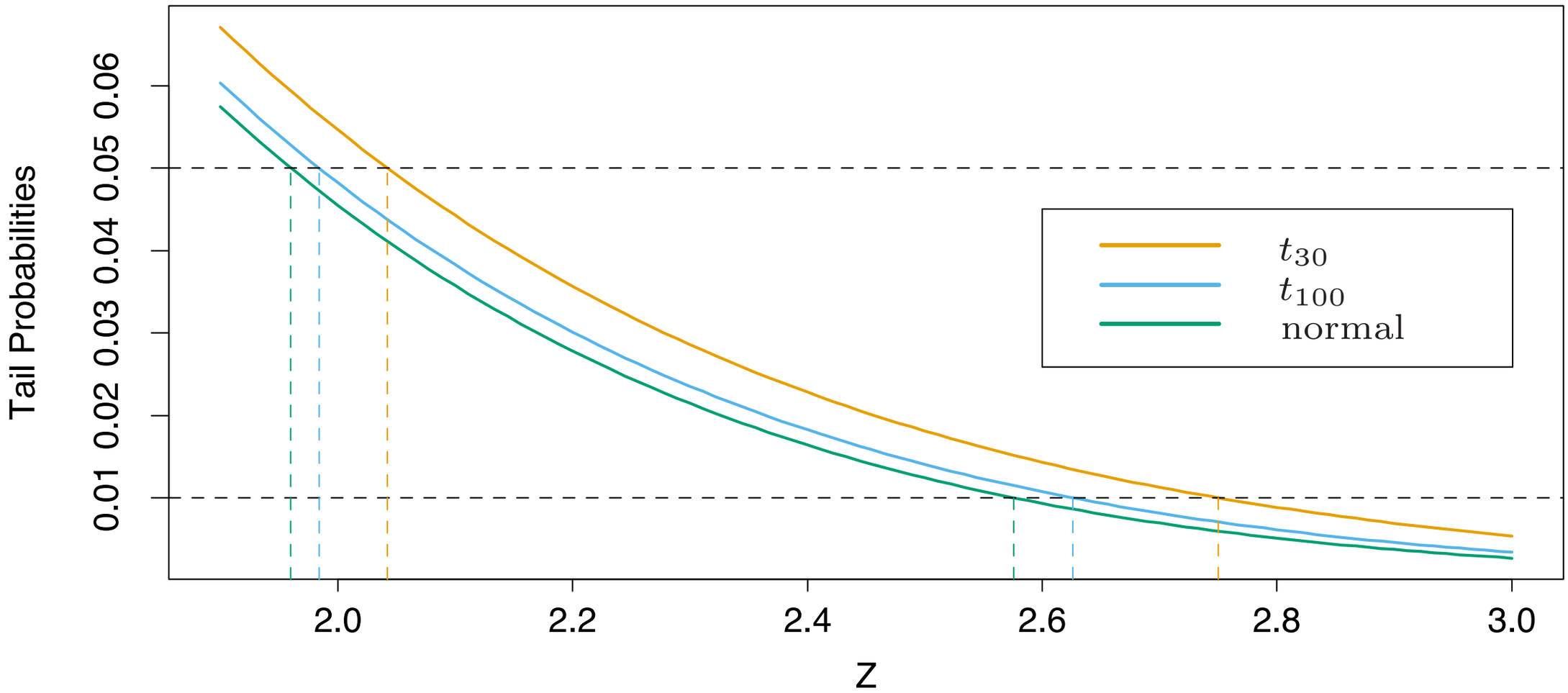
$$H_0: \beta_j = 0 \quad \text{vs} \quad H_1: \beta_j \neq 0$$

$$z_j = \frac{\hat{\beta}_j / (\sigma\sqrt{v_j})}{\sqrt{\hat{\sigma}^2 / \sigma^2}} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t_{N-p-1}$$

v_j : j*th* diagonal of $(X'X)^{-1}$

- If σ is known, the z-score follows the standard normal distribution. For a large N, the difference b/w the normal and t-distribution become negligible.

- Tests based on the z-score can be applied for individual coefficients in a sequential manner.


Tail Probabilities
0.01
0.02
0.03
0.04
0.05
0.06
2.0
2.2
2.4
2.6
2.8
3.0
Z
t_{30}
t_{100}
normal

Confidence interval

- A 1-2xα confidence interval for β_j

$$(\hat{\beta}_j - z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma}, \;\; \hat{\beta}_j + z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma}).$$

  - It is easy to see that the above interval is equivalent to

$$\hat{\beta}_j \pm z^{(1-\alpha)} \text{se}(\hat{\beta}_j)$$

  - One can use the t-distribution for a small sample.

- The confidence set of entire parameter vector is

$$C_\beta = \{\beta | (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \beta) \leq \hat{\sigma}^2 {\chi^2_{p+1}}^{(1-\alpha)}\}$$

  - df = Rank(H) = p+1

F-Tests for nested models

- Suppose we want to drop a few predictors from the full model $\mathbf{y}=\mathbf{X}\beta+\varepsilon$. Let $\mathbf{y}=\mathbf{X}_r\beta_r+\varepsilon$ denote this reduced model.

- The reduced model is nested in the full model. We can also consider two nested models (bigger and smaller).

- Using the F-test, one can test a set of regression coefficients simultaneously.

- Testing for nested models

$$H_0 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \varepsilon$$

$$\updownarrow$$

$$H_0 : \beta_3 = \beta_6 = 0$$
$$H_1 : \beta_3 \neq 0 \text{ or } \beta_6 \neq 0$$

- RSS_1: RSS for the bigger model. RSS_0: RSS for the smaller model. p_1: # of predictors in the bigger model. p_0: # of predictors in the smaller model.

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)}$$

  - F ~ F-distribution with (p_1-p_0, N-p_1-1).

- F-test of overall significance

$$H_0 : Y = \beta_0 + \varepsilon$$
$$H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

- The result of this F-test (null vs full model) is provided in R by default with the summary of the regression fit.

- Failing to reject the null hypothesis means there are no linear association between y and predictors.

Example: Prostate cancer

- The goal is to predict lpsa (log prostate specific antigen).
- Predictors
  - log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

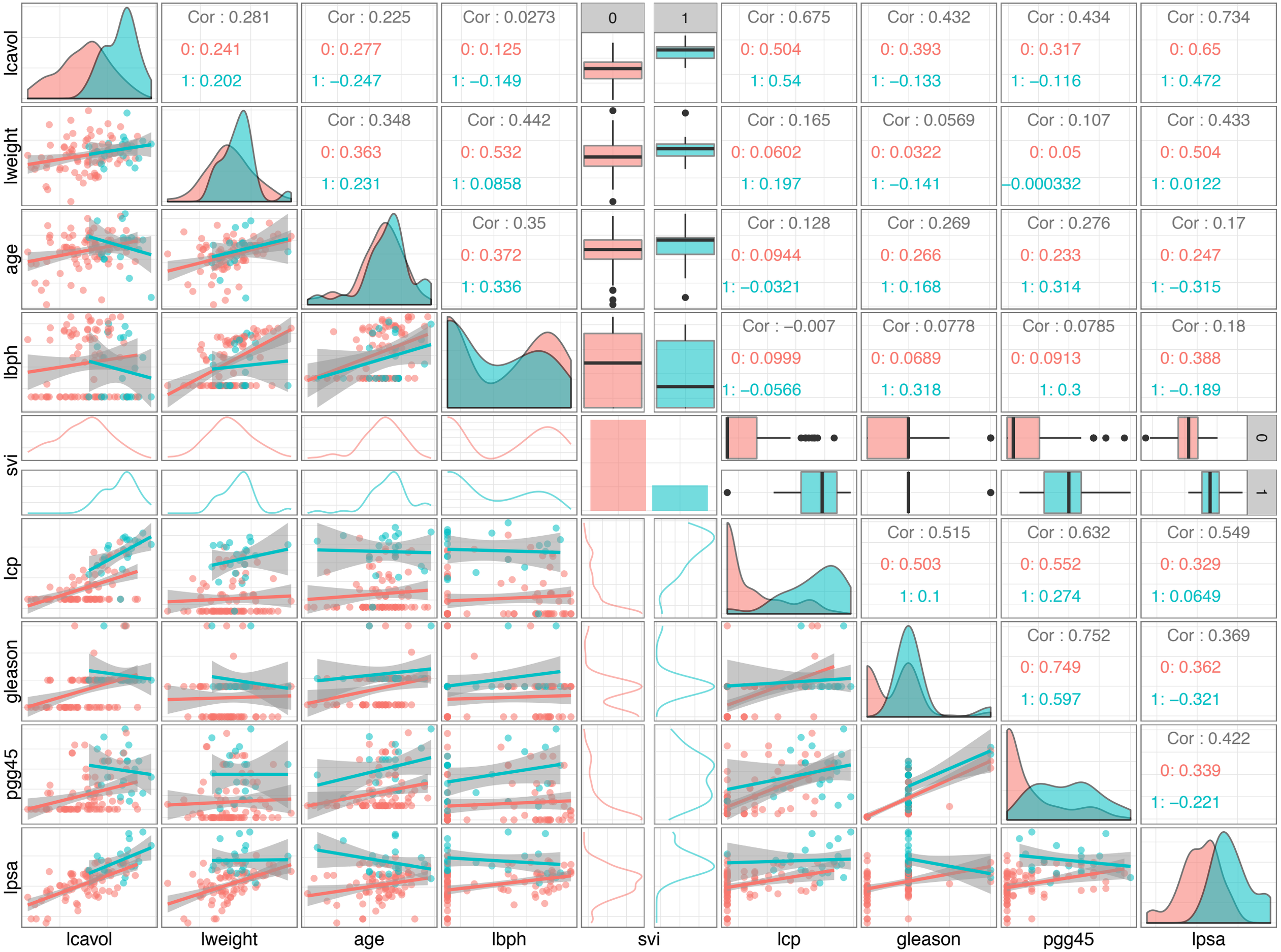

lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
lpsa
Cor : 0.281
0: 0.241
1: 0.202
Cor : 0.225
0: 0.277
1: −0.247
Cor : 0.0273
0: 0.125
1: −0.149
0
1
Cor : 0.675
0: 0.504
1: 0.54
Cor : 0.432
0: 0.393
1: −0.133
Cor : 0.434
0: 0.317
1: −0.116
Cor : 0.734
0: 0.65
1: 0.472
Cor : 0.348
0: 0.363
1: 0.231
Cor : 0.442
0: 0.532
1: 0.0858
Cor : 0.165
0: 0.0602
1: 0.197
Cor : 0.0569
0: 0.0322
1: −0.141
Cor : 0.107
0: 0.05
−0.000332
Cor : 0.433
0: 0.504
1: 0.0122
Cor : 0.35
0: 0.372
1: 0.336
Cor : 0.128
0: 0.0944
1: −0.0321
Cor : 0.269
0: 0.266
1: 0.168
Cor : 0.276
0: 0.233
1: 0.314
Cor : 0.17
0: 0.247
1: −0.315
Cor : −0.007
0: 0.0999
1: −0.0566
Cor : 0.0778
0: 0.0689
1: 0.318
Cor : 0.0785
0: 0.0913
1: 0.3
Cor : 0.18
0: 0.388
1: −0.189
Cor : 0.515
0: 0.503
1: 0.1
Cor : 0.632
0: 0.552
1: 0.274
Cor : 0.549
0: 0.329
1: 0.0649
Cor : 0.752
0: 0.749
1: 0.597
Cor : 0.369
0: 0.362
1: −0.321
Cor : 0.422
0: 0.339
1: −0.221

TABLE 3.1. *Correlations of predictors in the prostate cancer data.*

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

- We consider dropping all the non-significant terms in Table 3.2, namely age, lcp, gleason, and pgg45.

$$F = \frac{(32.81 - 29.43)/(9 - 5)}{29.43/(67 - 9)} = 1.67$$

- P-value is 0.17 ($P(F_{4,58} > 1.67) = 0.17$). Hence, these four predictors are not significant.

Gauss-Markov Theorem

- Bias-variance tradeoff
- The least square estimator is BLUE (best linear unbiased estimator).
  - The estimator with the smallest variance among all linear unbiased estimators.
- Consider the linear combination of parameters $\mathbf{a}^T\beta$.
  - The least square estimate is unbiased
  -

$$\begin{aligned} \mathrm{E}(a^T\hat{\beta}) &= \mathrm{E}(a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\ &= a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= a^T\beta. \end{aligned}$$

- Linear unbiased estimates:

$$\mathrm{E}(\mathbf{c}^T\mathbf{y}) = a^T\beta$$

$$E(\mathbf{c}'\mathbf{y}) = \mathbf{c}'E(\mathbf{y}) = \mathbf{c}'\mathbf{X}\beta = a'\beta$$

$$c'\mathbf{X} = a'$$

- $\mathbf{a}'\hat{\beta}$ is the BLUE iff $\mathrm{Var}(a^T\hat{\beta}) \le \mathrm{Var}(\mathbf{c}^T\mathbf{y})$.

- pf) $\mathrm{var}(\mathbf{a}'\hat{\beta}) = \mathrm{var}(\mathbf{c}'\mathbf{X}\hat{\beta}) = \sigma^2\mathbf{c}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{c}$

  $\mathrm{var}(\mathbf{c}'\mathbf{y}) = \sigma^2\mathbf{c}'\mathbf{c}$

- Q: what should be the next step in the proof?

- Consider the mean squared error of an estimator $\tilde{\theta}$ in estimating θ:

$$\begin{aligned} \mathrm{MSE}(\tilde{\theta}) &= \mathrm{E}(\tilde{\theta}-\theta)^2 \\ &= \mathrm{Var}(\tilde{\theta}) + [\mathrm{E}(\tilde{\theta})-\theta]^2. \end{aligned} \tag{3.20}$$

- MSE = variance + bias^2.

- There may be a biased estimator with smaller variance than the least square estimator (overall, smaller MSE).

  - Ridge regression, LASSO (least absolute shrinkage and selection operator), LARS (least angle regression), etc.

Prediction for a new response

- Prediction of the new response at $x=x_0$.

$$Y_0 = f(x_0) + \varepsilon_0.$$

- The prediction accuracy is measured by the expected prediction error:

true regression function

$$\begin{aligned} \mathrm{E}(Y_0 - \tilde{f}(x_0))^2 &= \sigma^2 + \mathrm{E}(x_0^T\tilde{\beta} - f(x_0))^2 \\ &= \sigma^2 + \mathrm{MSE}(\tilde{f}(x_0)). \end{aligned}$$

- For the prediction, we have more uncertainty. (MSE + σ^2).

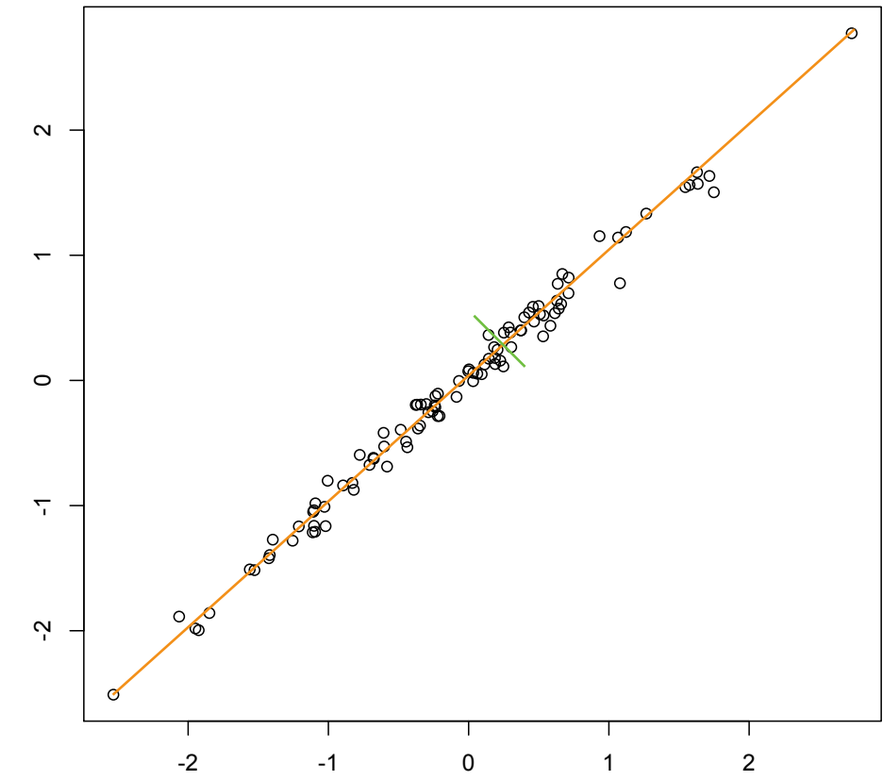
Variance inflation factor (VIF)

- Coefficient of determination ($0 \leq R^2 \leq 1$): squared correlation between y and “y hat”.
  - R^2 indicates lack-of-fit. However, A small R^2 does not imply you haven’t got something interesting!
- Multicollinearity problem: when predictors are highly correlated, the variance for estimate of β increases (unstable estimates).
- $VIF_j = 1/(1-R_j^2)$. $R_j^2 = R^2$ for the linear regression of X_j on other predictors.
- If $VIF_j > 10$, the multicollinearity is severe.

- When you have severe multicollinearity problem:
  - F-test of overall significance says there are significant linear association b/w y and some predictors, but no individual regression coefficients are significant from z-scores.
  - Z-scores may lead to different conclusions depending on which predictors you include in the model.
  - The estimated coefficients is very uncertain (large standard error).

Why these happen?

- Suppose two predictors x_1 and x_2 are highly correlated. Ignoring an intercept, a design matrix is $\mathbf{X}=(x_1,x_2)$.



- Recall that

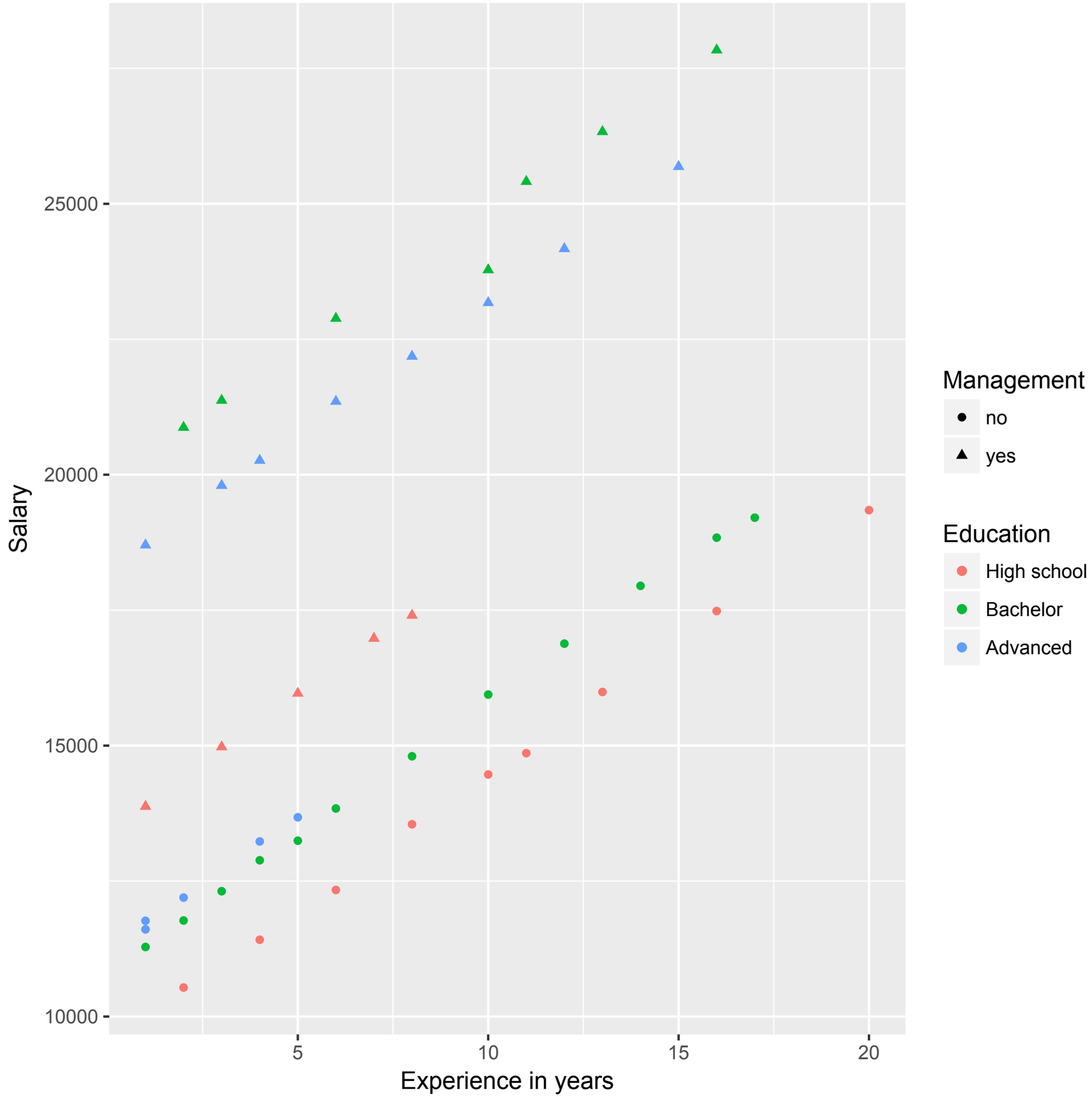
$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.$$

$$X'X = VD^2V' = V\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}V'$$

- Discuss the outcome of the high correlation.

Salary survey data

- The salary survey data was developed from a salary survey of computer professionals in a large corporation. The response variable is salary (S) and the predicators are:
  - X: experience in years
  - E: education coded as
    - 1 for completion of a high school diploma
    - 2 for completion of a bachelor degree
    - 3 for completion of an advanced degree
  - M: management coded as 1 for a person with management responsibility; 0 otherwise.

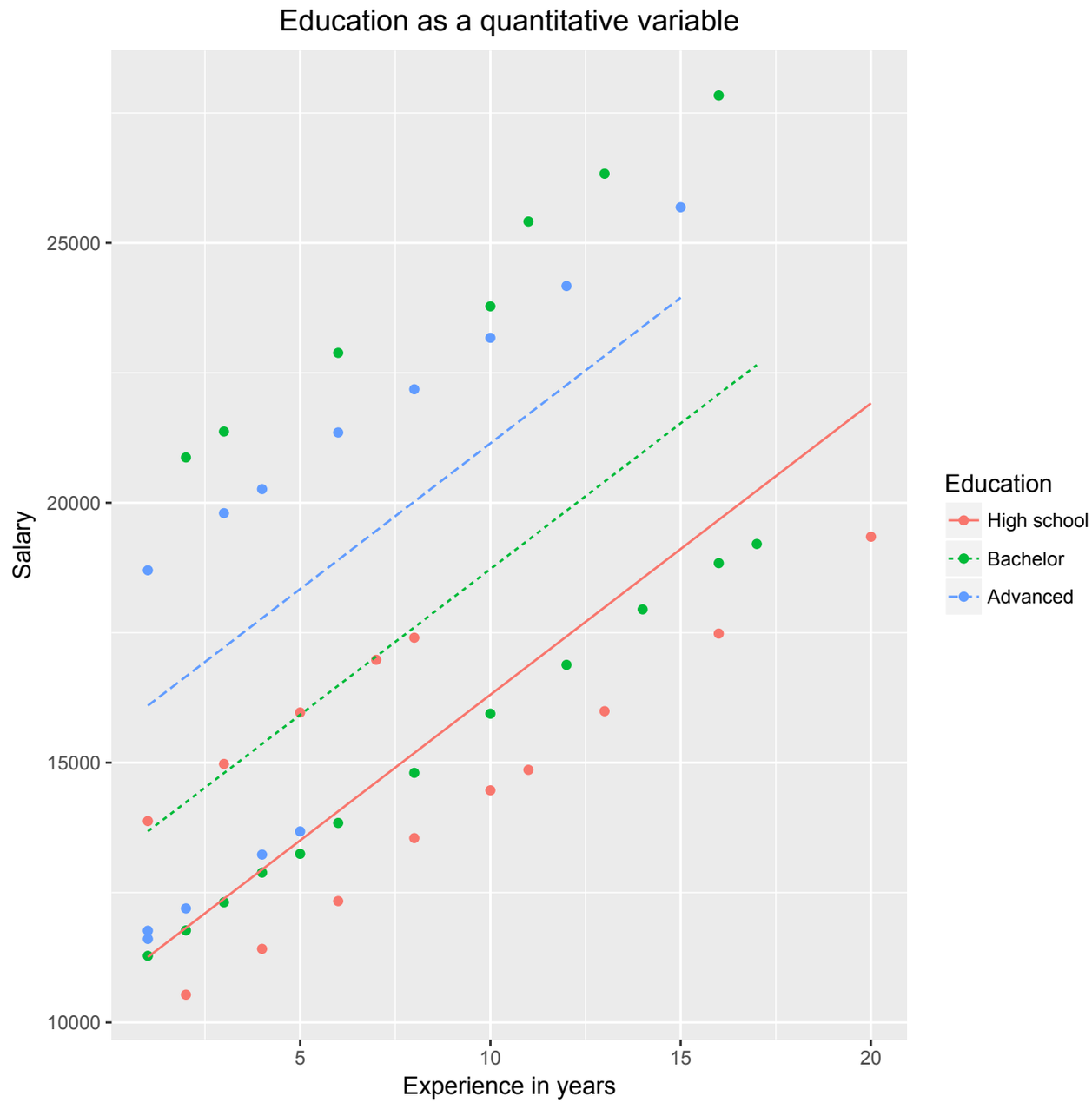

25000
20000
15000
10000
Salary
5
10
15
20
Experience in years
Management
no
yes
Education
High school
Bachelor
Advanced

- Treat E as a quantitative variable (ignoring M).

$$S = \beta_0 + \beta_1 X + \gamma E + \epsilon$$

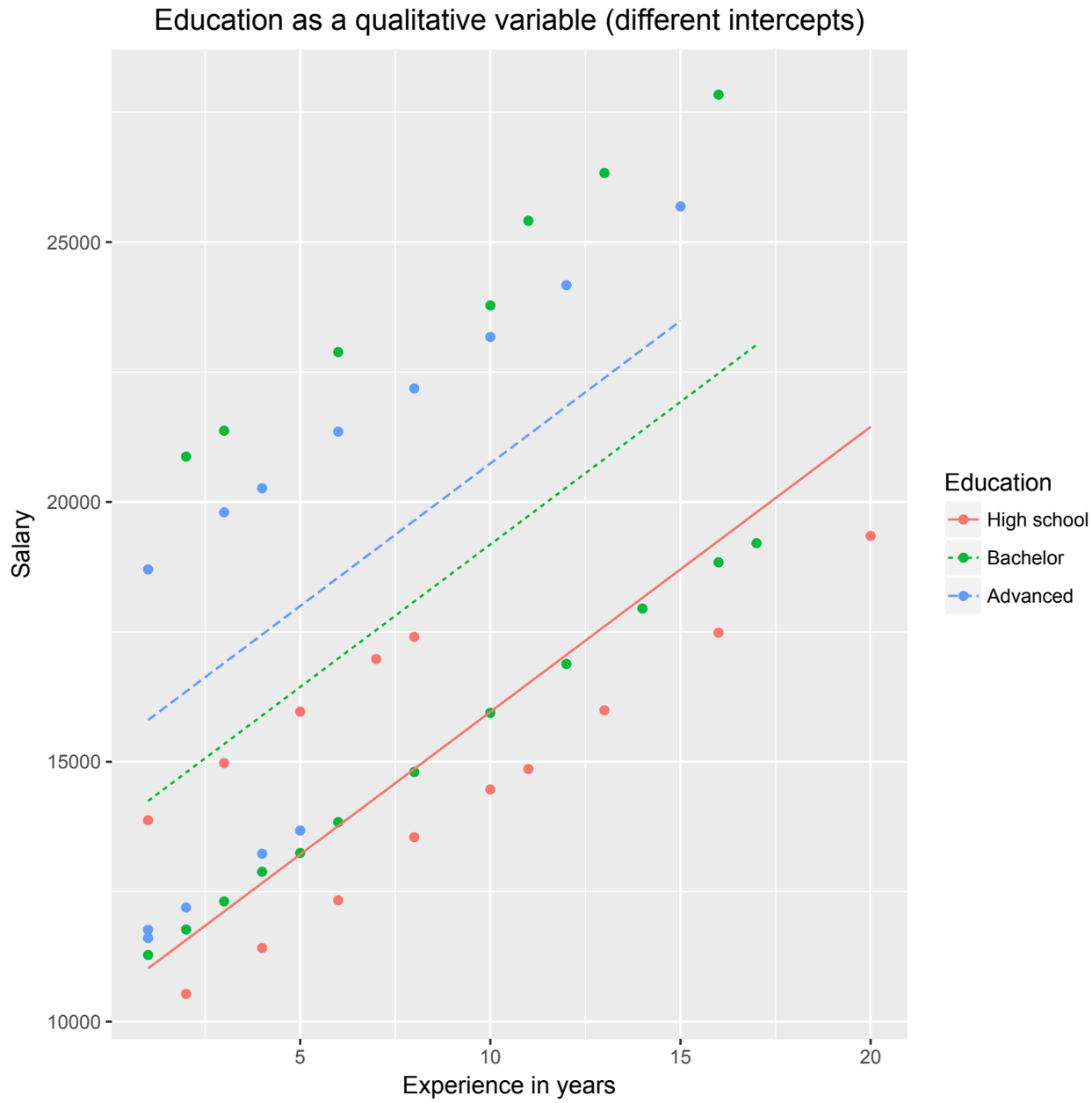
  - Assuming each step up in education is worth a fixed increment to salary.

  - Too restrictive.



- Treat E as a qualitative variable (still ignoring M).
  - Define two indicator variables (dummy variables) to represent the three categories.
    - E2: 1 for high school; 0 otherwise
    - E3: 1 for bachelor degree; 0 otherwise

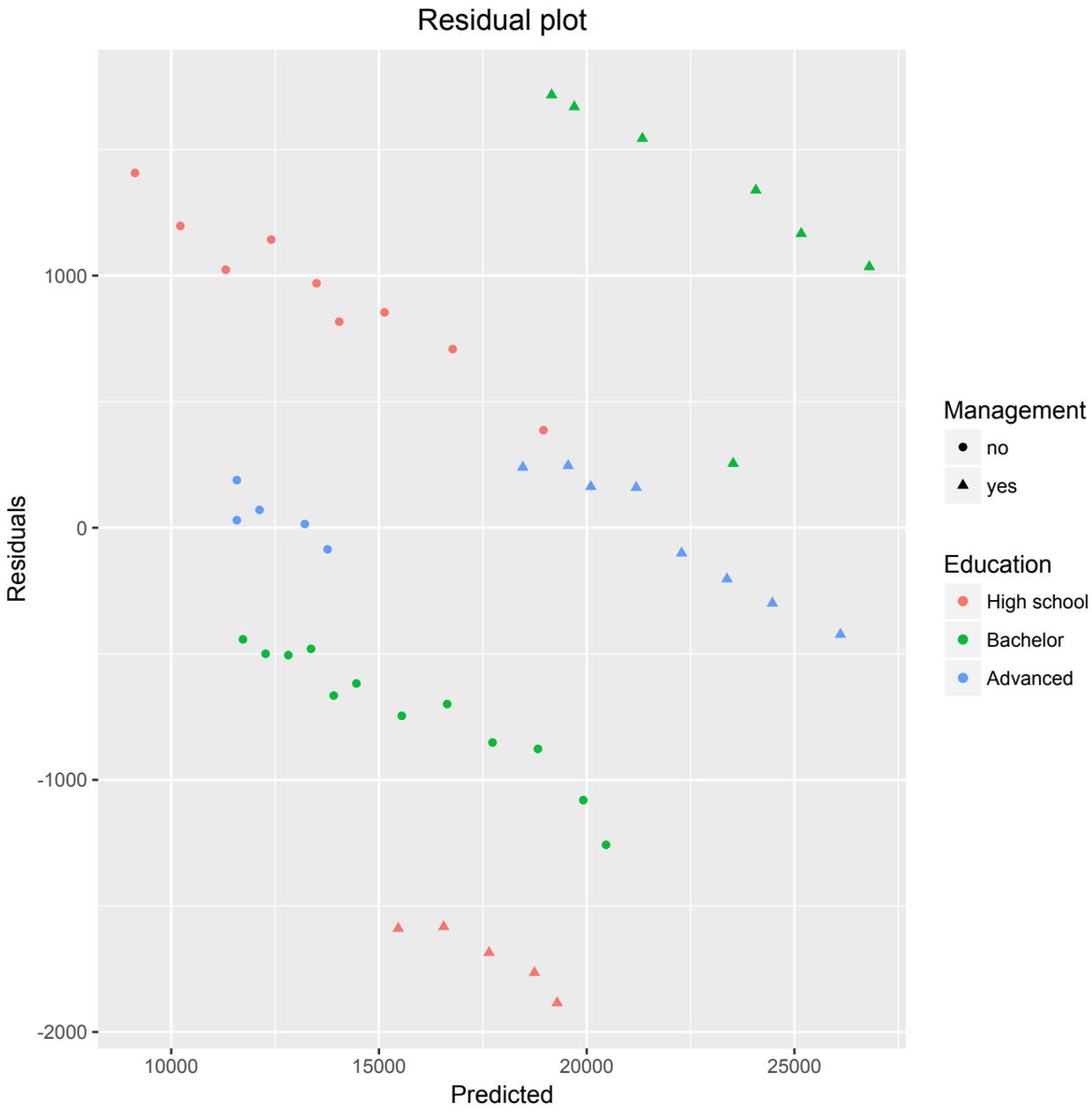
$$S = \beta_0 + \beta_1 X + \gamma_2 E_2 + \gamma_3 E_3 + \epsilon$$



- Full additive model

$$S = \beta_0 + \beta_1 X + \gamma_2 E_2 + \gamma_3 E_3 + \delta_1 M + \epsilon$$

  - 6 categories are present.
  - Describe the regression function for each category.
  - Do you observe any systematic pattern in residuals?

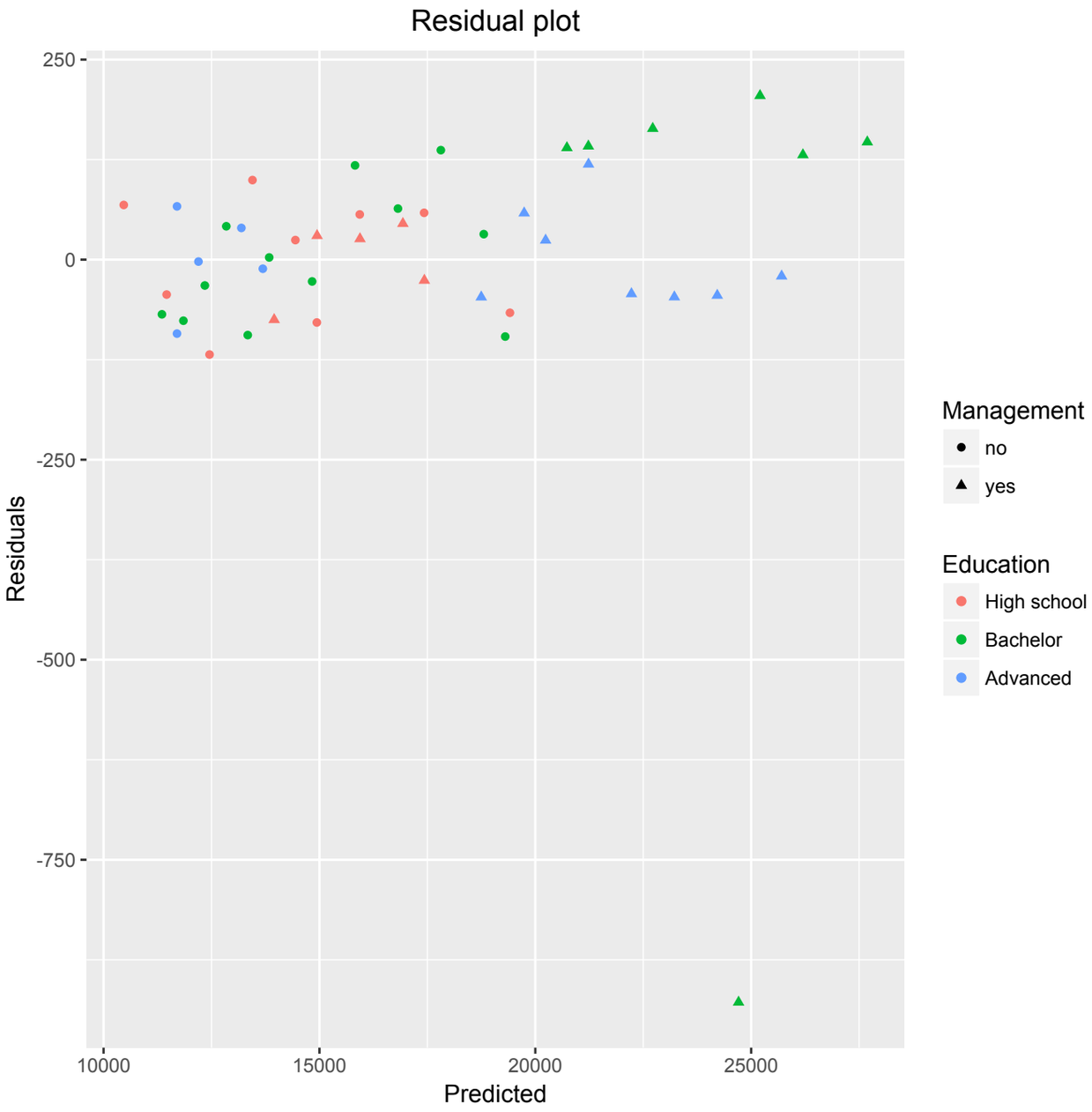


- Interaction variables are defined as products of the existing indicator variables

$$S = \beta_0 + \beta_1 X + \gamma_2 E_2 + \gamma_3 E_3 + \delta_1 M + \alpha_2 E_2 M + \alpha_3 E_3 M + \epsilon$$

- Interaction terms should be included only if an additive model is not good enough.

- Do you see systematic pattern in the residual plot?

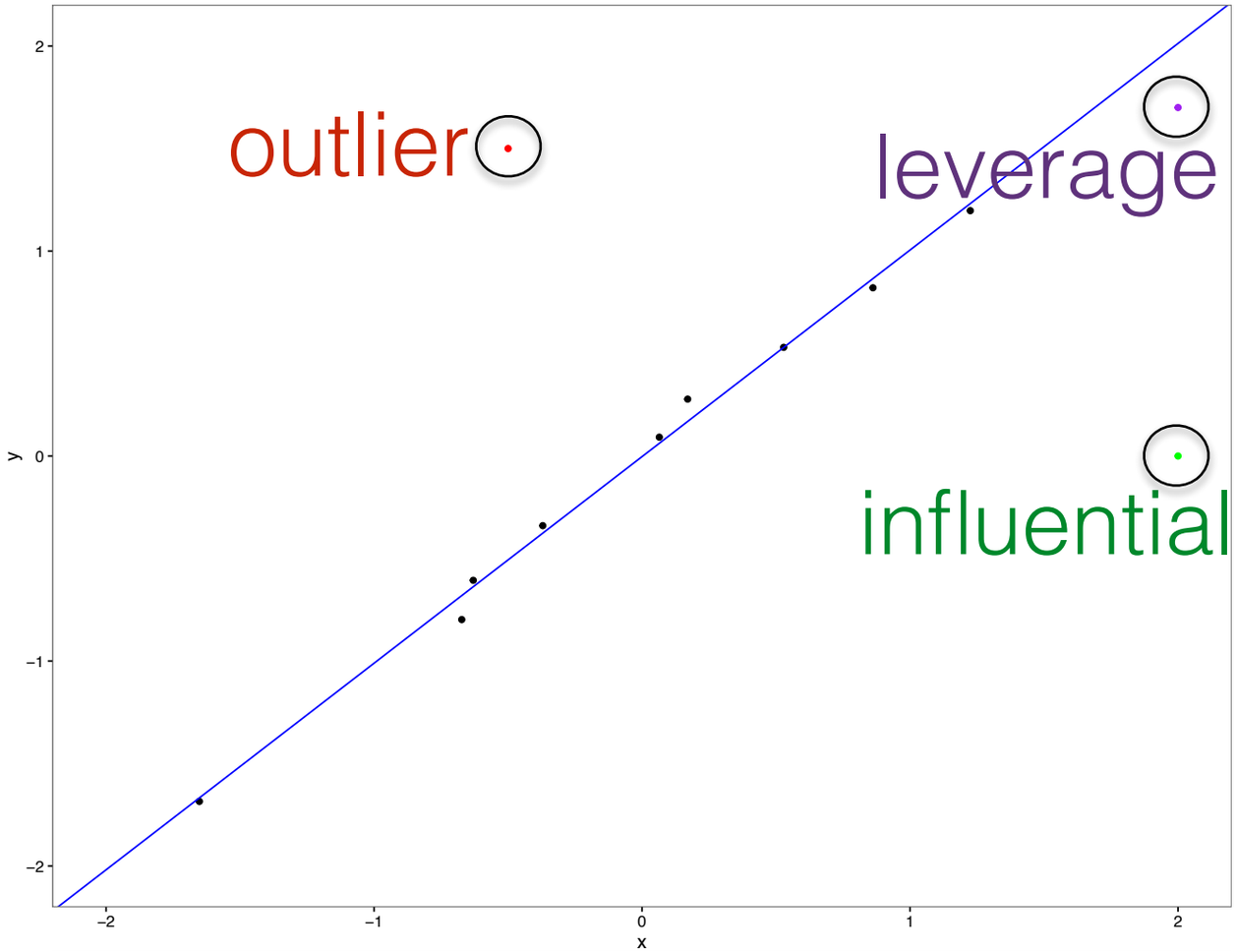


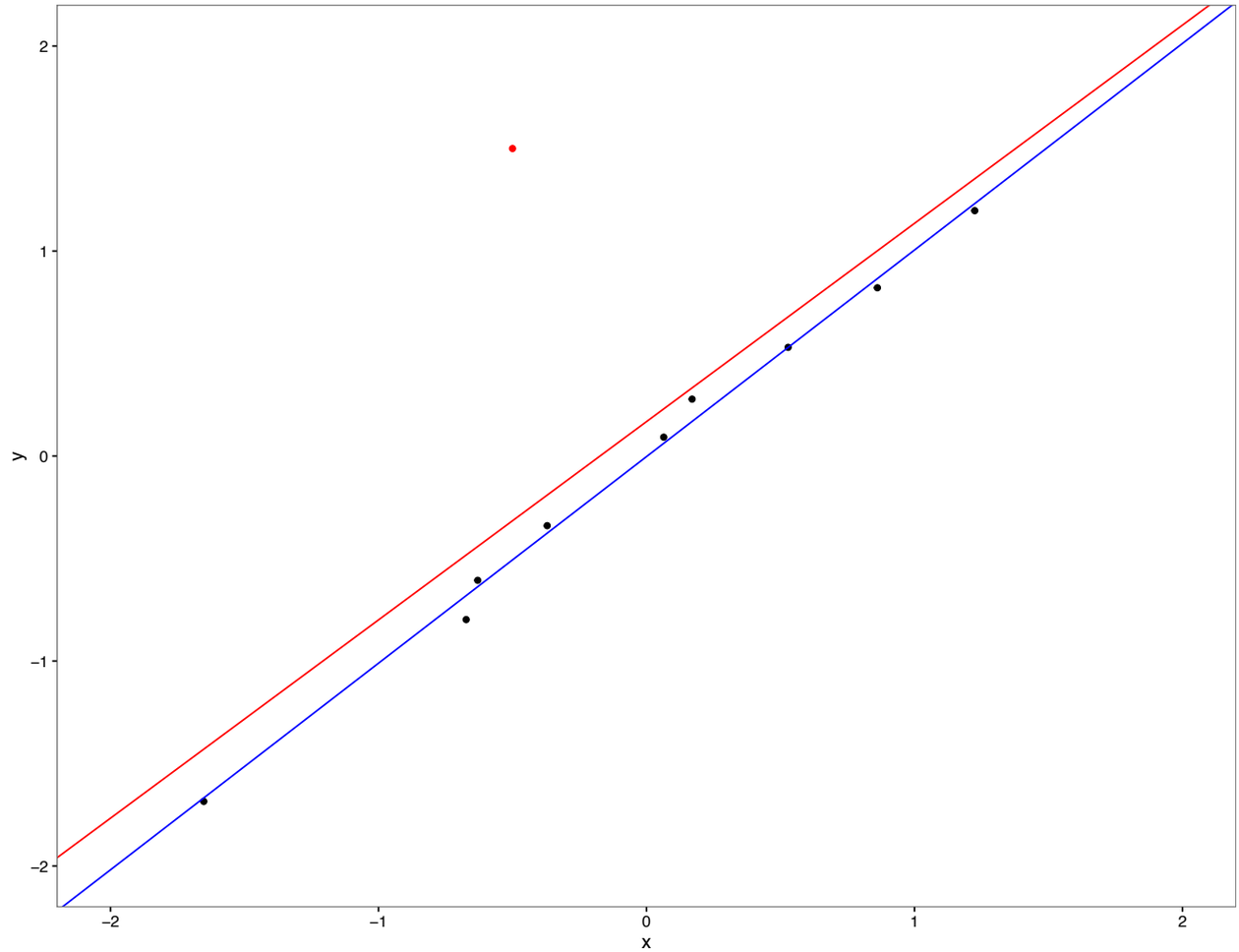
Outlying observations

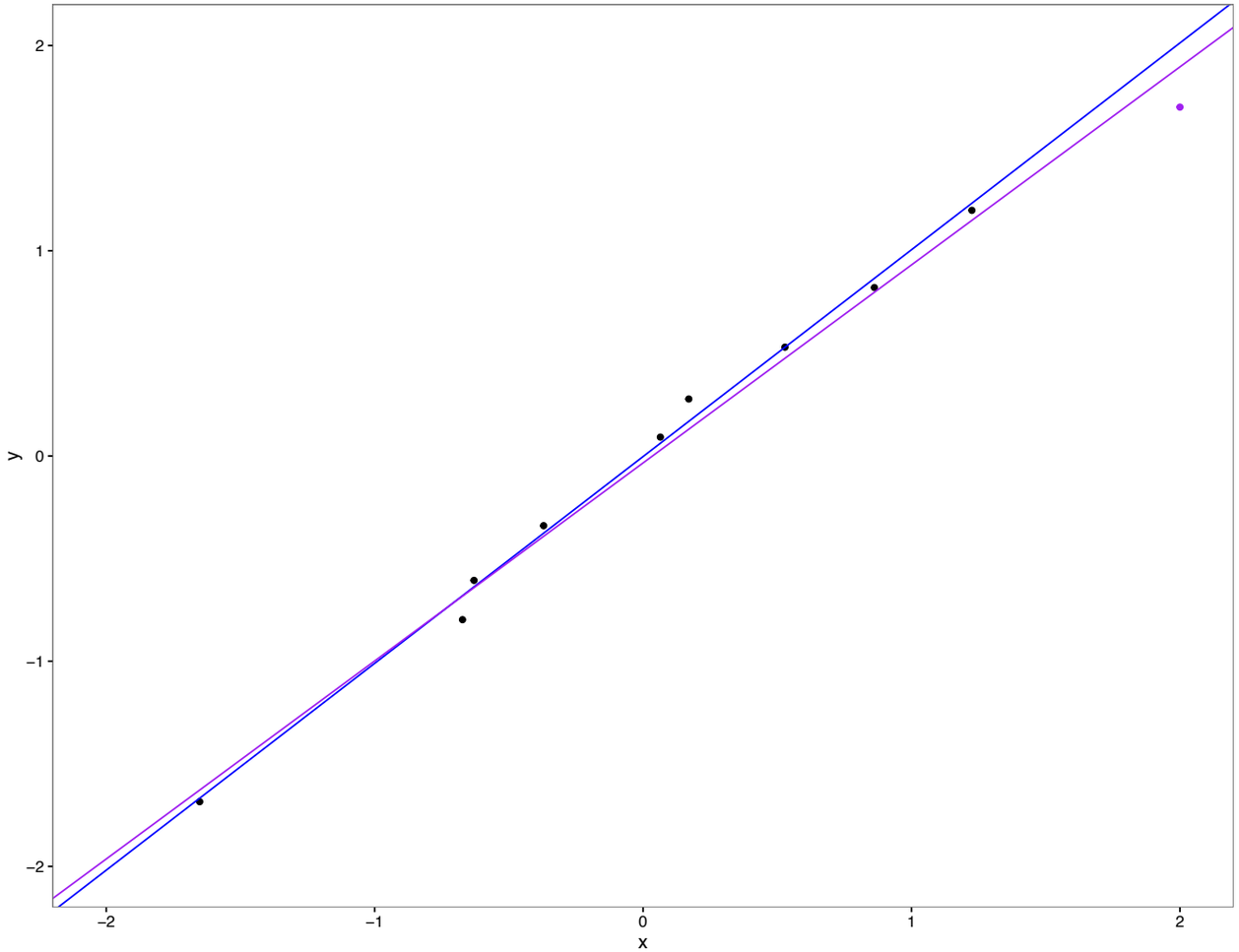
- Outlying observations can cause us to misinterpret patterns.

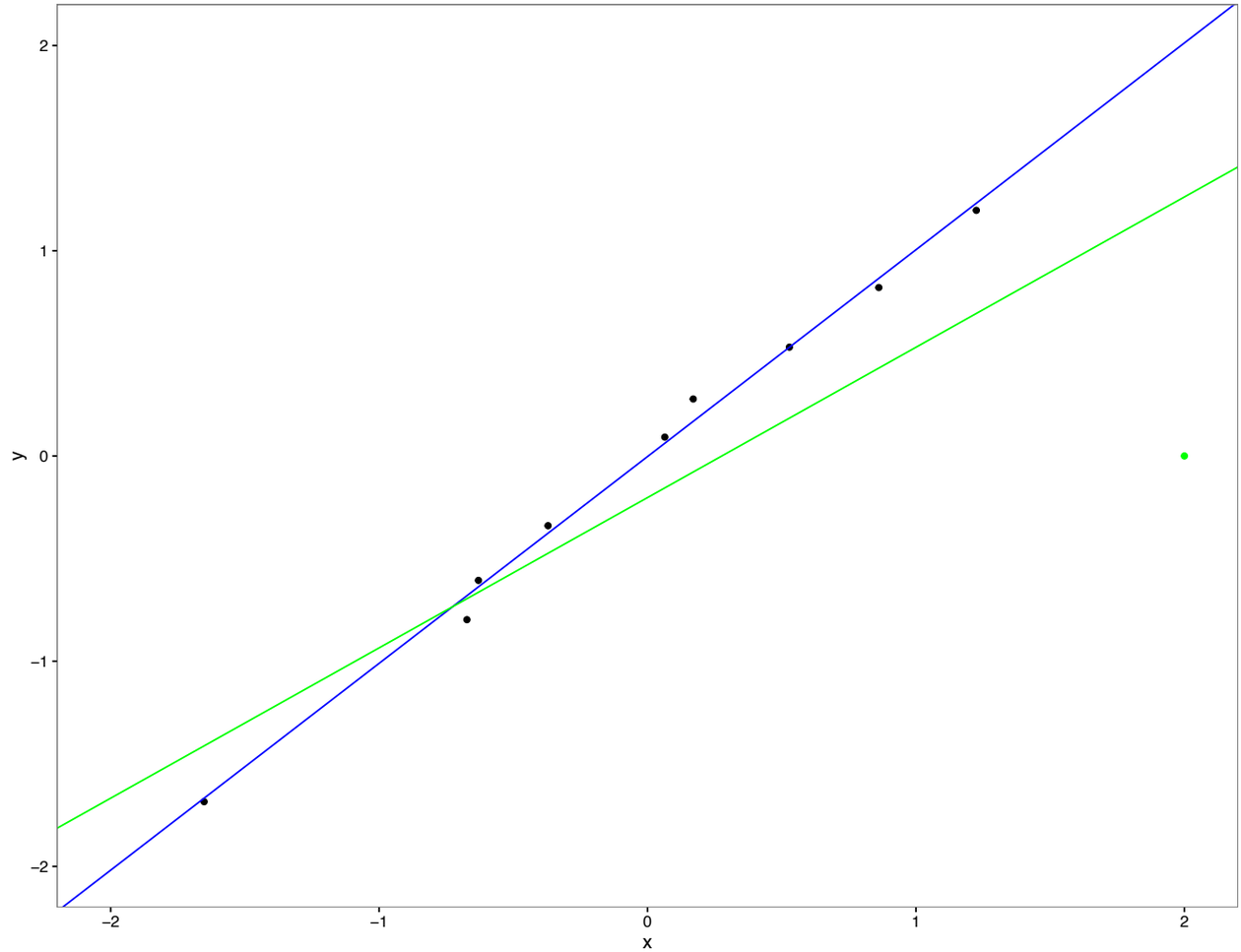
- Outlying observations can have a strong influence on statistical models—deleting outliers from a regression model can sometimes give completely different results.

- Outliers may also indicate that our model fails to capture important characteristics of the data.


outlier
leverage
influential
y
x
2
1
0
−1
−2
−2
−1
0
1
2


y
x
2
1
0
−1
−2
−2
−1
0
1
2


y
x
2
1
0
−1
−2
−2
−1
0
1
2


y
x
2
1
0
−1
−2
−2
−1
0
1
2

Regression outliers

- A regression outlier is an observation that has an unusual value of Y, conditional on X.
  - For a regression outlier, neither the X nor the Y value is necessarily unusual on its own.
- A regression outlier will have a large residual but not necessarily affect the regression slope coefficients.
- Observations with standardized residuals > 3 (in absolute value) is generally deemed outliers.
- $r_i = \frac{e_i}{\text{se}(e_i)} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$ where the h_i is ith diagonal of $\mathbf{H}$.
- $\text{var}(\mathbf{y}-\hat{\mathbf{y}}) = \text{var}((\mathbf{I}-\mathbf{H})\mathbf{y}) = \sigma^2(\mathbf{I}-\mathbf{H})$.

Leverage

- An observation that has an unusual X value (far from the mean of X) has leverage on (potential to influence) the regression line.

- High leverage does not necessarily mean that it influences the regression coefficients. It is possible to have a high leverage and yet follow straight in line with the pattern of the rest of the data.

- A point with high h_i has more leverage on the fitted surface. If $h_i = 1$, the fitted regression line must pass through y_i.
  If $h_i > 2(p+1)/n$, the leverage is large.

- $\mathrm{var}(e_i) = \hat{\sigma}^2(1-h_i)^2$ $\qquad 0 \le h_i \le 1$ and $\sum_{i=1}^{n} h_i = p+1$

  Why this is true?

Influential observations

- When an observation has high leverage and is an outlier in terms of Y-value, it will strongly influence the regression line.

- Cook's distance measures the difference in the regression estimates when the i*th* observation is left out.

- Influence = Leverage x Outlyingness.

$$D_i = \frac{h_i}{1-h_i} \times \frac{r_i^2}{p+1}$$

- Cook's distance is large if $D_i > 4/(n-p-1)$.

Shortcomings of regression

- Predictive ability
  - The linear regression fit often has low bias but high variance.
  - The regression fit often does not predict well, especially when p (the number of predictors) is large (related to multicollinearity).
- Interpretative ability
  - Linear regression “freely” assigns a coefficient to each predictor variable (i.e. no constraints).
  - When p is large, we may sometimes seek, for the sake of interpretation, a smaller set of important variables
  - Hence we want to encourage our fitting procedure to make only a subset of the coefficients large, and others small or even better, zero

What we did not cover

- Section 3.2.3 Multiple regression from univariate regression (optional reading).
- Section 3.2.4 Multiple outputs
- Section 3.4 Subset selection and coefficient shrinkage (we will come back to this later).
- Section 3.5 Methods using derived input directions (we will come back to this later).