Model Diagnostics

Zhenisbek Assylbekov

Department of Mathematics

Regression Analysis

Diagnostics we have already discussed

- Residuals e_i vs.
  - i (independence)
  - $x_1, \ldots, x_k$ (linearity)
  - $\hat{Y}_i$ (linearity and homogeneity of variance)

Diagnostics we have already discussed

- Residuals e_i vs.
  - i (independence)
  - $x_1, \ldots, x_k$ (linearity)
  - $\hat{Y}_i$ (linearity and homogeneity of variance)
- Q-Q plot of $e_1, \ldots, e_n$.

Diagnostics we have already discussed

- Residuals e_i vs.
  - i (independence)
  - $x_1, \ldots, x_k$ (linearity)
  - $\hat{Y}_i$ (linearity and homogeneity of variance)
- Q-Q plot of $e_1, \ldots, e_n$.
- VIF_j for $j = 1, \ldots, k$.

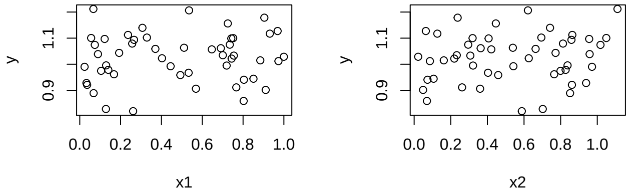
Diagnostics we have already discussed

- Residuals e_i vs.
  - i (independence)
  - $x_1, \ldots, x_k$ (linearity)
  - $\hat{Y}_i$ (linearity and homogeneity of variance)
- Q-Q plot of $e_1, \ldots, e_n$.
- VIF_j for $j = 1, \ldots, k$.
- Significance tests (Runs, Levene's, Shapiro-Wilk)

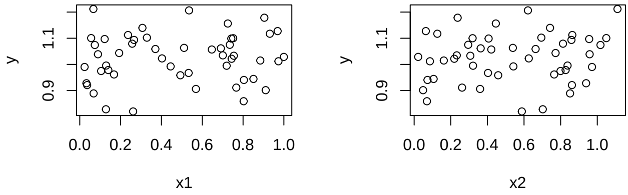
Diagnostics we have already discussed

- Residuals e_i vs.
  - i (independence)
  - $x_1, \ldots, x_k$ (linearity)
  - $\hat{Y}_i$ (linearity and homogeneity of variance)
- Q-Q plot of $e_1, \ldots, e_n$.
- VIF_j for $j = 1, \ldots, k$.
- Significance tests (Runs, Levene's, Shapiro-Wilk)
- Now we'll discuss
  - added variable plots,
  - leverages,
  - DFFITS,
  - Cook's distance.

The problem with marginal plots

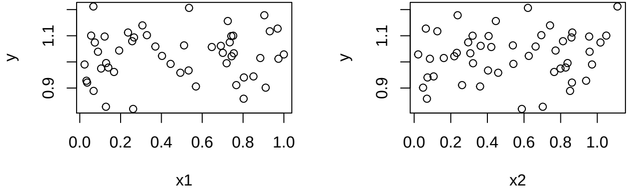


The problem with marginal plots



No dependence b/w Y and x_j marginally, but significant association jointly:

The problem with marginal plots



No dependence b/w Y and x_j marginally, but significant association jointly:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05083    0.07726   0.658    0.514
x1           0.95536    0.07651  12.487   <2e-16 ***
x2           0.96312    0.07658  12.576   <2e-16 ***
```

Added variable plots

Outliers, leverages, DFFITs, Cook's distance

Review of Diagnostics

Example

Problems with marginal plots

- Residuals e_i versus a predictor values x_i, j can show whether x_j may need to be transformed or whether we should add a quadratic term x_j^2.

Problems with marginal plots

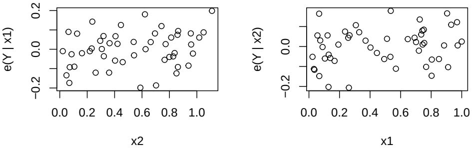
- Residuals e_i versus a predictor values x_i, j can show whether x_j may need to be transformed or whether we should add a quadratic term x_j^2.
- We can omit the predictor from the model and plot the residuals versus the predictor to see if the predictor explains residual variability.

Problems with marginal plots

- Residuals e_i versus a predictor values x_i, j can show whether x_j may need to be transformed or whether we should add a quadratic term x_j^2.
- We can omit the predictor from the model and plot the residuals versus the predictor to see if the predictor explains residual variability.
- However these plots can also be misleading:

Problems with marginal plots

- Residuals e_i versus a predictor values x_i, j can show whether x_j may need to be transformed or whether we should add a quadratic term x_j^2.
- We can omit the predictor from the model and plot the residuals versus the predictor to see if the predictor explains residual variability.
- However these plots can also be misleading: e.g., we can have



where $e(Y \mid x_j)$ are residuals when we regress Y on x_j only.

Added variables plots

- The previous plots suggest that x_2 is not needed when x_1 is in the model,

Added variables plots

- The previous plots suggest that x_2 is not needed when x_1 is in the model, or that x_1 is not needed when x_2 is in the model.

Added variables plots

- The previous plots suggest that x_2 is not needed when x_1 is in the model, or that x_1 is not needed when x_2 is in the model.
- But we still have significance for both x_1 and x_2 when they are *simultaneously* in the model!

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.05083    0.07726   0.658    0.514
x1            0.95536    0.07651  12.487   <2e-16 ***
x2            0.96312    0.07658  12.576   <2e-16 ***
```

Added variables plots

- The previous plots suggest that x_2 is not needed when x_1 is in the model, or that x_1 is not needed when x_2 is in the model.
- But we still have significance for both x_1 and x_2 when they are *simultaneously* in the model!

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.05083    0.07726   0.658    0.514
x1             0.95536    0.07651  12.487   <2e-16 ***
x2             0.96312    0.07658  12.576   <2e-16 ***
```

- An **added variable plot** tries to fix this problem.

Added variables plots

- The previous plots suggest that x_2 is not needed when x_1 is in the model, or that x_1 is not needed when x_2 is in the model.
- But we still have significance for both x_1 and x_2 when they are *simultaneously* in the model!

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05083    0.07726   0.658    0.514
x1           0.95536    0.07651  12.487   <2e-16 ***
x2           0.96312    0.07658  12.576   <2e-16 ***
```

- An **added variable plot** tries to fix this problem.
- It answers the question: Does x_j explain any *residual* variability once the rest of the predictors are in the model?

10.1 Added variable plots

- Consider a pool of predictors $x_1, \ldots, x_k$. Lets consider predictor x_j where $j = 1, \ldots, k$.

10.1 Added variable plots

- Consider a pool of predictors $x_1, \ldots, x_k$. Lets consider predictor x_j where $j = 1, \ldots, k$.
- Regress Y_i vs. all predictors except x_j, call the residuals $e_i(Y \mid \mathbf{x}_{-j})$.

10.1 Added variable plots

- Consider a pool of predictors $x_1, \ldots, x_k$. Lets consider predictor x_j where $j = 1, \ldots, k$.
- Regress Y_i vs. all predictors except x_j, call the residuals $e_i(Y \mid \mathbf{x}_{-j})$.
- Regress x_j vs. all predictors except x_j, call the residuals $e_i(x_j \mid \mathbf{x}_{-j})$.

10.1 Added variable plots

- Consider a pool of predictors $x_1, \ldots, x_k$. Lets consider predictor x_j where $j = 1, \ldots, k$.
- Regress Y_i vs. all predictors except x_j, call the residuals $e_i(Y \mid \mathbf{x}_{-j})$.
- Regress x_j vs. all predictors except x_j, call the residuals $e_i(x_j \mid \mathbf{x}_{-j})$.
- The added variable plot for x_j is $e_i(Y \mid \mathbf{x}_{-j})$ vs. $e_i(x_j \mid \mathbf{x}_{-j})$.

10.1 Added variable plots

- Consider a pool of predictors $x_1, \ldots, x_k$. Lets consider predictor x_j where $j = 1, \ldots, k$.
- Regress Y_i vs. all predictors except x_j, call the residuals $e_i(Y \mid \mathbf{x}_{-j})$.
- Regress x_j vs. all predictors except x_j, call the residuals $e_i(x_j \mid \mathbf{x}_{-j})$.
- The added variable plot for x_j is $e_i(Y \mid \mathbf{x}_{-j})$ vs. $e_i(x_j \mid \mathbf{x}_{-j})$.
- If you fit a simple linear regression

$$e_i(Y \mid \mathbf{x}_{-j}) = \beta_j \cdot e_i(x_j \mid \mathbf{x}_{-j}) + \epsilon_i$$

then the LSE $\hat{\beta}_j$ *is the same* as one would get from fitting the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$.

10.1 Added variable plots

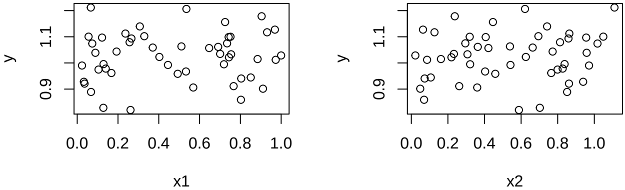
- Consider a pool of predictors $x_1, \ldots, x_k$. Lets consider predictor x_j where $j = 1, \ldots, k$.
- Regress Y_i vs. all predictors except x_j, call the residuals $e_i(Y \mid \mathbf{x}_{-j})$.
- Regress x_j vs. all predictors except x_j, call the residuals $e_i(x_j \mid \mathbf{x}_{-j})$.
- The added variable plot for x_j is $e_i(Y \mid \mathbf{x}_{-j})$ vs. $e_i(x_j \mid \mathbf{x}_{-j})$.
- If you fit a simple linear regression

  $$e_i(Y \mid \mathbf{x}_{-j}) = \beta_j \cdot e_i(x_j \mid \mathbf{x}_{-j}) + \epsilon_i$$

  then the LSE $\hat{\beta}_j$ *is the same* as one would get from fitting the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$.
- Gives an idea of the functional form of x_j: a transformation in x_j should mimic the pattern seen in the plot.

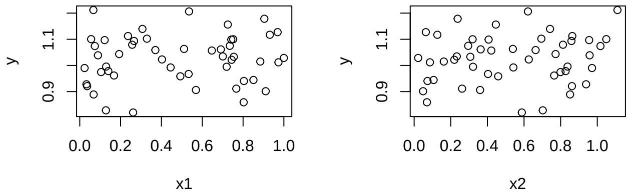
10.1 Added variables plots illustration

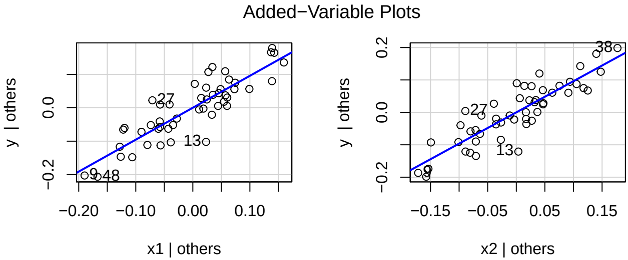
Let's go back to our synthetic example:



10.1 Added variables plots illustration

Let's go back to our synthetic example:

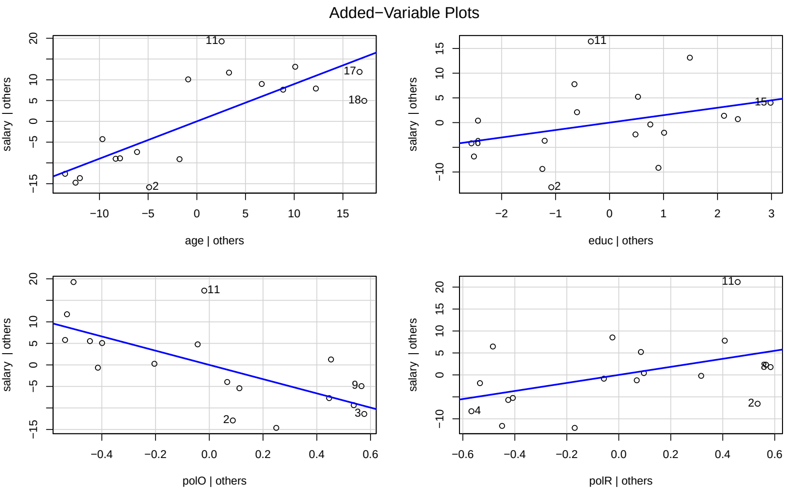




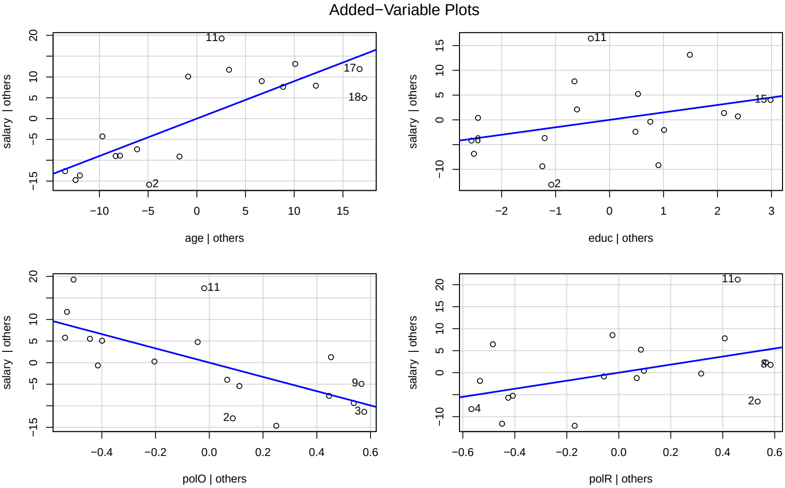
Salary data, first order terms only

```
> head(salary_data)
  salary age educ pol
1     38  25    4   D
2     45  27    4   R
3     28  26    4   O
4     55  39    4   D
5     74  42    4   R
6     43  41    4   O
> m = lm(salary ~ ., data=salary_data)
> summary(m)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.0313     7.3459   2.318  0.03735 *
age           0.8983     0.1968   4.565  0.00053 ***
educ          1.5039     1.1841   1.270  0.22632
polO        -16.5404     4.8807  -3.389  0.00484 **
polR          9.1587     4.8482   1.889  0.08139 .
---

Residual standard error: 8.209 on 13 degrees of freedom
Multiple R-squared:  0.8374,Adjusted R-squared:  0.7873
```

Added variable plots



Added variable plots

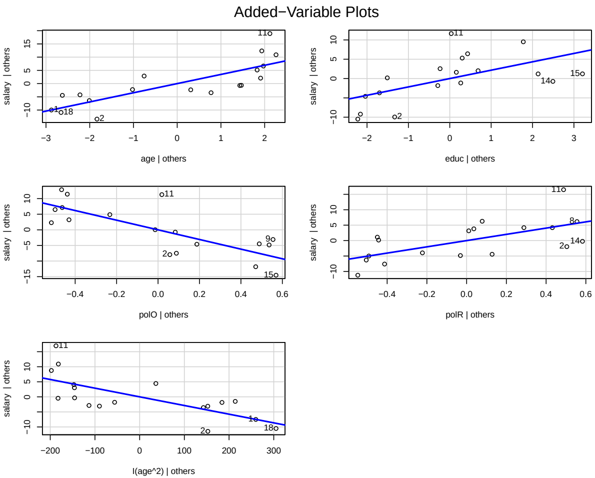


Age effect is nonlinear; let's add a quadratic term.

Salary data, quadratic effect in age

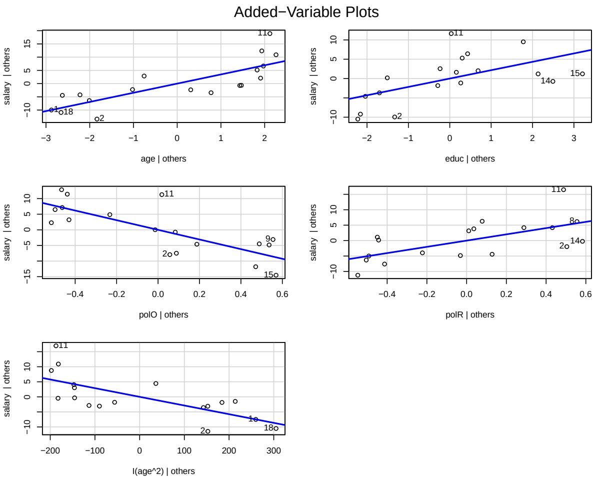
```
> m1 = lm(salary ~ . + I(age^2), data=salary_data)
> summary(m1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.224169  16.810142  -2.333 0.037839 *
age           3.463723   0.740666   4.676 0.000535 ***
educ          2.166475   0.883369   2.453 0.030453 *
polO        -15.455108   3.571147  -4.328 0.000983 ***
polR         10.118144   3.544586   2.855 0.014500 *
I(age^2)     -0.028831   0.008166  -3.530 0.004143 **
---

Residual standard error: 5.984 on 12 degrees of freedom
Multiple R-squared:  0.9202,Adjusted R-squared:  0.887
```

Added variables plots w/ quadratic age



Education is now significant!

Added variables plots w/ quadratic age

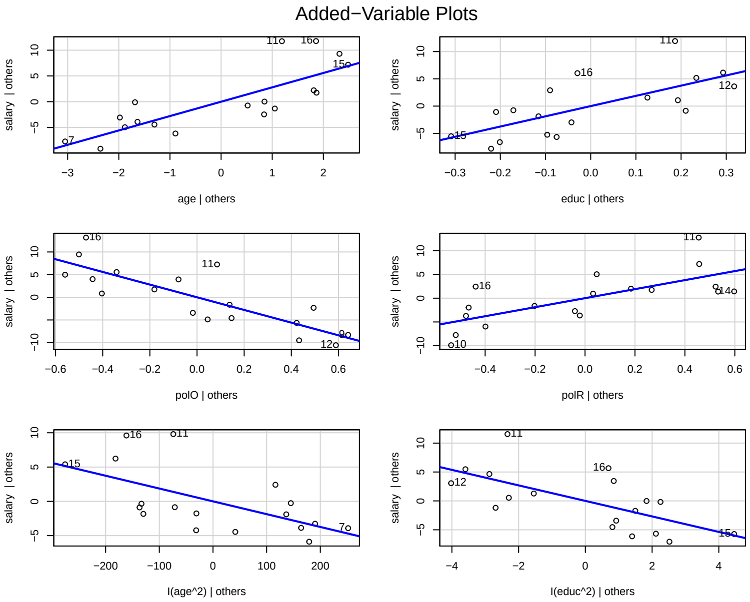


Education is now significant! The incorrect functional form for age was *masking* the importance of education.

Salary data, quadratic effect in education

```
> summary(m2)

Call:
lm(formula = salary ~ . + I(age^2) + I(educ^2), data = salary_data)

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -75.977348  18.262402  -4.160 0.001589 **
age           2.787032   0.626151   4.451 0.000977 ***
educ         18.751324   5.739109   3.267 0.007501 **
polD        -13.976910   2.848879  -4.906 0.000467 ***
polR          9.495127   2.790631   3.403 0.005903 **
I(age^2)     -0.018677   0.007298  -2.559 0.026558 *
I(educ^2)    -1.342341   0.461108  -2.911 0.014161 *
---

Residual standard error: 4.697 on 11 degrees of freedom
Multiple R-squared:  0.9549,Adjusted R-squared:  0.9304
```

Added variable plots w/ quad. age and educ



Outliers

- Outliers are data points which are "far away" from the bulk of data.

Outliers

- Outliers are data points which are “far away” from the bulk of data. Observations may be outlying
  - relative to predictors, i.e. $\mathbf{x}_i$ relative to other $\{\mathbf{x}_j\}_{j \neq i}$

Outliers

- Outliers are data points which are "far away" from the bulk of data. Observations may be outlying
  - relative to predictors, i.e. $\mathbf{x}_i$ relative to other $\{\mathbf{x}_j\}_{j \neq i}$
  - relative to the model, i.e. Y_i relative to $\hat{Y}_i$.

Outliers

- Outliers are data points which are "far away" from the bulk of data. Observations may be outlying
  - relative to predictors, i.e. $\mathbf{x}_i$ relative to other $\{\mathbf{x}_j\}_{j \neq i}$
  - relative to the model, i.e. Y_i relative to $\hat{Y}_i$.
- **Studentized deleted residuals** are designed to detect outlying Y_i observations; **leverages** detect outlying $\mathbf{x}_i$ points.

Outliers

- Outliers are data points which are "far away" from the bulk of data. Observations may be outlying
  - relative to predictors, i.e. $\mathbf{x}_i$ relative to other $\{\mathbf{x}_j\}_{j \neq i}$
  - relative to the model, i.e. Y_i relative to $\hat{Y}_i$.
- **Studentized deleted residuals** are designed to detect outlying Y_i observations; **leverages** detect outlying $\mathbf{x}_i$ points.
- Outliers have the potential to influence the fitted regression function:

Outliers

- Outliers are data points which are "far away" from the bulk of data. Observations may be outlying
  - relative to predictors, i.e. $\mathbf{x}_i$ relative to other $\{\mathbf{x}_j\}_{j \neq i}$
  - relative to the model, i.e. Y_i relative to $\hat{Y}_i$.
- **Studentized deleted residuals** are designed to detect outlying Y_i observations; **leverages** detect outlying $\mathbf{x}_i$ points.
- Outliers have the potential to influence the fitted regression function:
  - if the outlying points follow the modeling assumptions and are representative, they may *strengthen* inference and reduce error in predictions

Outliers

- Outliers are data points which are "far away" from the bulk of data. Observations may be outlying
  - relative to predictors, i.e. $\mathbf{x}_i$ relative to other $\{\mathbf{x}_j\}_{j \neq i}$
  - relative to the model, i.e. Y_i relative to $\hat{Y}_i$.
- **Studentized deleted residuals** are designed to detect outlying Y_i observations; **leverages** detect outlying $\mathbf{x}_i$ points.
- Outliers have the potential to influence the fitted regression function:
  - if the outlying points follow the modeling assumptions and are representative, they may *strengthen* inference and reduce error in predictions
  - if not, outlying values may skew inference a lot and yield models with poor predictive properties.

Outliers & influential points

- Often outliers are flagged and deemed suspect as mistakes or observations not gathered from the same population as the other observations.

Outliers & influential points

- Often outliers are flagged and deemed suspect as mistakes or observations not gathered from the same population as the other observations.
- Outliers are sometimes of interest in their own right and may illustrate aspects of the dataset that require more careful study.

Outliers & influential points

- Often outliers are flagged and deemed suspect as mistakes or observations not gathered from the same population as the other observations.
- Outliers are sometimes of interest in their own right and may illustrate aspects of the dataset that require more careful study.
- Although an observation may be flagged as an outlier, the point *may or may not* affect the fitted regression function more than other points.

Outliers & influential points

- Often outliers are flagged and deemed suspect as mistakes or observations not gathered from the same population as the other observations.
- Outliers are sometimes of interest in their own right and may illustrate aspects of the dataset that require more careful study.
- Although an observation may be flagged as an outlier, the point *may or may not* affect the fitted regression function more than other points.
- A **DFFIT** is a measure of influence that an individual point $(\mathbf{x}_i, Y_i)$ has on the regression surface at $\mathbf{x}_i$.

Outliers & influential points

- Often outliers are flagged and deemed suspect as mistakes or observations not gathered from the same population as the other observations.
- Outliers are sometimes of interest in their own right and may illustrate aspects of the dataset that require more careful study.
- Although an observation may be flagged as an outlier, the point *may or may not* affect the fitted regression function more than other points.
- A **DFFIT** is a measure of influence that an individual point $(\mathbf{x}_i, Y_i)$ has on the regression surface at $\mathbf{x}_i$.
- **Cook's distance** is a consolidated measure of influence the point $(\mathbf{x}_i, Y_i)$ has on the regression surface at all n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} =$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$,

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}}$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$,

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$,
- $\mathbf{Y}$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$,
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$

Thus,

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$,
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$

Thus,

$$\mathrm{Cov}[\hat{\mathbf{Y}}]$$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$,
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$

Thus,

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathrm{Cov}[\mathbf{X}\hat{\boldsymbol{\beta}}]$$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$,
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$

Thus,

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathrm{Cov}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathrm{Cov}[\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}]$$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$,
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$

Thus,

$$\begin{aligned} \mathrm{Cov}[\hat{\mathbf{Y}}] &= \mathrm{Cov}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathrm{Cov}[\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}] \\ &= \mathrm{Cov}[\mathbf{H}\mathbf{Y}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}^\top = \mathbf{H}\sigma^2 \mathbf{I}\mathbf{H}^\top \end{aligned}$$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$,
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$

Thus,

$$
\begin{aligned}
\mathrm{Cov}[\hat{\mathbf{Y}}] &= \mathrm{Cov}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathrm{Cov}[\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}] \\
&= \mathrm{Cov}[\mathbf{H}\mathbf{Y}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}^\top = \mathbf{H}\sigma^2 \mathbf{I}\mathbf{H}^\top \\
&= \sigma^2 \mathbf{H}
\end{aligned}
$$

Variance of $\hat{Y}_i$

Recall that

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$,
- $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$

Thus,

$$\begin{aligned}
\mathrm{Cov}[\hat{\mathbf{Y}}] &= \mathrm{Cov}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathrm{Cov}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}] \\
&= \mathrm{Cov}[\mathbf{H}\mathbf{Y}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}^\top = \mathbf{H}\sigma^2\mathbf{I}\mathbf{H}^\top \\
&= \sigma^2\mathbf{H}
\end{aligned}$$

$\Rightarrow \quad \mathrm{Var}[\hat{Y}_i] = \sigma^2 h_{ii}$, and its unbiased estimator is $\mathrm{MSE} \cdot h_{ii}$.

10.2 Studentized deleted residuals

- The **standardized residuals**

$$r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\text{MSE}(1 - h_{ii})}}$$

have a constant variance of 1.

10.2 Studentized deleted residuals

- The **standardized residuals**

$$r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\mathrm{MSE}(1 - h_{ii})}}$$

  have a constant variance of 1.
- Typically, $|r_i| > 2$ is considered "large."

10.2 Studentized deleted residuals

- The **standardized residuals**

$$r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\text{MSE}(1 - h_{ii})}}$$

  have a constant variance of 1.
- Typically, $|r_i| > 2$ is considered "large."
- $h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$ is called the i^{th} **leverage value**.

10.2 Studentized deleted residuals

- The **standardized residuals**

$$r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\mathrm{MSE}(1 - h_{ii})}}$$

  have a constant variance of 1.
- Typically, $|r_i| > 2$ is considered "large."
- $h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$ is called the i^{th} **leverage value**.
- A refinement of the standardized residual that has a recognizable distribution is the **studentized deleted residual**

$$t_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\mathrm{MSE}_{(i)}(1 - h_{ii})}}$$

  where $\mathrm{MSE}_{i(i)}$ is obtained from the model when i-th example was removed from the data.

Studentized deleted residuals

- In fact, no need to fit n additional regressions, because there is relationship b/w MSE and $\text{MSE}_{(i)}$:

$$(n-p)\text{MSE} = (n-p-1)\text{MSE}_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

(prove it)

Studentized deleted residuals

- In fact, no need to fit n additional regressions, because there is relationship b/w MSE and $\text{MSE}_{(i)}$:

$$(n-p)\text{MSE} = (n-p-1)\text{MSE}_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

  (prove it)
- Studentized deleted residuals are distributed as

$$t_i \sim t_{n-p-1}.$$

Studentized deleted residuals

- In fact, no need to fit n additional regressions, because there is relationship b/w MSE and $\text{MSE}_{(i)}$:

$$(n-p)\text{MSE} = (n-p-1)\text{MSE}_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

  (prove it)
- Studentized deleted residuals are distributed as

$$t_i \sim t_{n-p-1}.$$

- Therefore, outlying Y-values may be flagged by using Bonferroni's adjustment and taking

$$|t_i| > t_{1-\alpha/(2n);n-p-1}$$

  as outlying.

Studentized deleted residuals

- In fact, no need to fit n additional regressions, because there is relationship b/w MSE and $\text{MSE}_{(i)}$:

$$(n-p)\text{MSE} = (n-p-1)\text{MSE}_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

(prove it)
- Studentized deleted residuals are distributed as

$$t_i \sim t_{n-p-1}.$$

- Therefore, outlying Y-values may be flagged by using Bonferroni's adjustment and taking

$$|t_i| > t_{1-\alpha/(2n);n-p-1}$$

as outlying.
- Typically, in practice, one simply flags observations with $|t_i| > t_{1-\alpha/2;n-p-1}$ as *possibly* outlying in consideration with other diagnostics.

10.3 Leverage

- The leverages h_{ii} get larger the further the points $\mathbf{x}_i$ are from the mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, adjusted for "how many" other predictors are in the vicinity of $\mathbf{x}_i$.

10.3 Leverage

- The leverages h_{ii} get larger the further the points $\mathbf{x}_i$ are from the mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, adjusted for "how many" other predictors are in the vicinity of $\mathbf{x}_i$.
- Use the fact that $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}\mathbf{H}$ to show $\sum_{i=1}^{n} h_{ii} = p$ and $0 \leq h_{ii} \leq 1$.

10.3 Leverage

- The leverages h_{ii} get larger the further the points $\mathbf{x}_i$ are from the mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, adjusted for "how many" other predictors are in the vicinity of $\mathbf{x}_i$.
- Use the fact that $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{HH}$ to show $\sum_{i=1}^{n} h_{ii} = p$ and $0 \le h_{ii} \le 1$.
- A large leverage h_{ii} indicates that $\mathbf{x}_i$ is far away from the other predictors $\{\mathbf{x}_j\}_{j\neq i}$

10.3 Leverage

- The leverages h_{ii} get larger the further the points $\mathbf{x}_i$ are from the mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, adjusted for "how many" other predictors are in the vicinity of $\mathbf{x}_i$.
- Use the fact that $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}\mathbf{H}$ to show $\sum_{i=1}^{n} h_{ii} = p$ and $0 \leq h_{ii} \leq 1$.
- A large leverage h_{ii} indicates that $\mathbf{x}_i$ is far away from the other predictors $\{\mathbf{x}_j\}_{j \neq i}$ and that $\mathbf{x}_i$ may influence the fitted value $\hat{Y}_i$ more than other x_j's will influence their respective fitted values.

10.3 Leverage

- The leverages h_{ii} get larger the further the points $\mathbf{x}_i$ are from the mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$, adjusted for "how many" other predictors are in the vicinity of $\mathbf{x}_i$.
- Use the fact that $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{HH}$ to show $\sum_{i=1}^n h_{ii} = p$ and $0 \leq h_{ii} \leq 1$.
- A large leverage h_{ii} indicates that $\mathbf{x}_i$ is far away from the other predictors $\{\mathbf{x}_j\}_{j\neq i}$ and that $\mathbf{x}_i$ may influence the fitted value $\hat{Y}_i$ more than other x_j's will influence their respective fitted values. This is evident in the variance of the residual $\mathrm{Var}[Y_i - \hat{Y}_i] = \sigma^2\sqrt{1-h_{ii}}$. The larger h_{ii} is, the smaller $\mathrm{Var}[Y_i - \hat{Y}_i]$ will be and hence the closer $\hat{Y}_i$ will be to Y_i on average.

10.3 Leverage

- The leverages h_{ii} get larger the further the points $\mathbf{x}_i$ are from the mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, adjusted for "how many" other predictors are in the vicinity of $\mathbf{x}_i$.
- Use the fact that $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}\mathbf{H}$ to show $\sum_{i=1}^{n} h_{ii} = p$ and $0 \leq h_{ii} \leq 1$.
- A large leverage h_{ii} indicates that $\mathbf{x}_i$ is far away from the other predictors $\{\mathbf{x}_j\}_{j\neq i}$ and that $\mathbf{x}_i$ may influence the fitted value $\hat{Y}_i$ more than other x_j's will influence their respective fitted values. This is evident in the variance of the residual $\mathrm{Var}[Y_i - \hat{Y}_i] = \sigma^2\sqrt{1-h_{ii}}$. The larger h_{ii} is, the smaller $\mathrm{Var}[Y_i - \hat{Y}_i]$ will be and hence the closer $\hat{Y}_i$ will be to Y_i on average.
- The rule of thumb is that any leverage h_{ii} that is larger than twice the mean leverage p/n, i.e. $h_{ii} > 2p/n$, is flagged as having "high" leverage.

Leverage

- Note that the leverages h_{ii} depend only on the $\mathbf{x}_i$ and hence indicate which points might *potentially* be influential.

Leverage

- Note that the leverages h_{ii} depend only on the $\mathbf{x}_i$ and hence indicate which points might *potentially* be influential.
- (p. 400) When making predictions $\mathbf{x}_{n+1}$ at a point not in the data set, we consider the measure of distance of this point from the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ given by $h_{n+1} = \mathbf{x}_{n+1}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}_{n+1}$.

Leverage

- Note that the leverages h_{ii} depend only on the $\mathbf{x}_i$ and hence indicate which points might *potentially* be influential.
- (p. 400) When making predictions $\mathbf{x}_{n+1}$ at a point not in the data set, we consider the measure of distance of this point from the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ given by $h_{n+1} = \mathbf{x}_{n+1}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}_{n+1}$.
- If h_{n+1} is much larger than any of the $\{h_{11}, \ldots, h_{nn}\}$ you may be extrapolating far outside the general region of your data.

Leverage

- Note that the leverages h_{ii} depend only on the $\mathbf{x}_i$ and hence indicate which points might *potentially* be influential.
- (p. 400) When making predictions $\mathbf{x}_{n+1}$ at a point not in the data set, we consider the measure of distance of this point from the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ given by $h_{n+1} = \mathbf{x}_{n+1}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}_{n+1}$.
- If h_{n+1} is much larger than any of the $\{h_{11}, \ldots, h_{nn}\}$ you may be extrapolating far outside the general region of your data.
- In `R`, you can get leverages using `hatvalues` command.

10.4 DFFITs

- The i^{th} DFFIT, denoted DFFIT_i, is given by

$$\text{DFFIT}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

10.4 DFFITs

- The i^{th} DFFIT, denoted DFFIT_i, is given by

$$\text{DFFIT}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

where $\hat{Y}_i$ is fitted value of regression surface (calculated using all n observations) at $\mathbf{x}_i$ and $\hat{Y}_{i(i)}$ is fitted value of regression surface *omitting the point* $(\mathbf{x}_i, Y_i)$ at the point $\mathbf{x}_i$.

10.4 DFFITs

- The i^{th} DFFIT, denoted DFFIT_i, is given by

$$\text{DFFIT}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

  where $\hat{Y}_i$ is fitted value of regression surface (calculated using all n observations) at $\mathbf{x}_i$ and $\hat{Y}_{i(i)}$ is fitted value of regression surface *omitting the point* $(\mathbf{x}_i, Y_i)$ at the point $\mathbf{x}_i$.
- DFFIT_i is standardized distance between *fitted* regression surfaces *with* and *without* the point $(\mathbf{x}_i, Y_i)$.

10.4 DFFITs

- The i^{th} DFFIT, denoted DFFIT_i, is given by

$$\text{DFFIT}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

  where $\hat{Y}_i$ is fitted value of regression surface (calculated using all n observations) at $\mathbf{x}_i$ and $\hat{Y}_{i(i)}$ is fitted value of regression surface *omitting the point* $(\mathbf{x}_i, Y_i)$ at the point $\mathbf{x}_i$.
- DFFIT_i is standardized distance between *fitted* regression surfaces *with* and *without* the point $(\mathbf{x}_i, Y_i)$.
- Rule of thumb that DFFIT_i is "large" when $|\text{DFFIT}_i| > 1$ for small to medium-sized data sets and $|\text{DFFIT}_i| > 2\sqrt{p/n}$ for large data sets.

10.4 DFFITs

- The i^{th} DFFIT, denoted DFFIT_i, is given by

$$\text{DFFIT}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

  where $\hat{Y}_i$ is fitted value of regression surface (calculated using all n observations) at $\mathbf{x}_i$ and $\hat{Y}_{i(i)}$ is fitted value of regression surface *omitting the point* $(\mathbf{x}_i, Y_i)$ at the point $\mathbf{x}_i$.
- DFFIT_i is standardized distance between *fitted* regression surfaces *with* and *without* the point $(\mathbf{x}_i, Y_i)$.
- Rule of thumb that DFFIT_i is "large" when $|\text{DFFIT}_i| > 1$ for small to medium-sized data sets and $|\text{DFFIT}_i| > 2\sqrt{p/n}$ for large data sets. We will often just note those DFFIT_i's that are considerable larger than the bulk of the DFFIT_i's.

10.4 Cook's distance

- The i^{th} Cook's distance, denoted D_i, is an aggregate measure of the influence of the i^{th} observation on all n fitted values:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot \text{MSE}}$$

10.4 Cook's distance

- The i^{th} Cook's distance, denoted D_i, is an aggregate measure of the influence of the i^{th} observation on all n fitted values:
$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot \text{MSE}}$$
This is the sum of squared distances, at each $\mathbf{x}_j$, between fitted regression surface calculated with all n points and fitted regression surface calculated with the i^{th} case removed, standardized by $p \cdot \text{MSE}$.

10.4 Cook's distance

- The i^{th} Cook's distance, denoted D_i, is an aggregate measure of the influence of the i^{th} observation on all n fitted values:
$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot \text{MSE}}$$
This is the sum of squared distances, at each $\mathbf{x}_j$, between fitted regression surface calculated with all n points and fitted regression surface calculated with the i^{th} case removed, standardized by $p \cdot \text{MSE}$.
- Look for values of Cook's distance significantly larger than other values; these are cases that have disproportionate influence on the fitted regression surface as a whole.

Review of diagnostics

- Variance inflation factors VIF_j tell you which predictors are highly correlated with other predictors. If you have one or more $\mathrm{VIF}_j > 10$, you may want to eliminate some of the predictors.

Review of diagnostics

- Variance inflation factors VIF_j tell you which predictors are highly correlated with other predictors. If you have one or more $\text{VIF}_j > 10$, you may want to eliminate some of the predictors.

  Multicollinearity affects the interpretability of the model, but does not indicate the model is "bad" in any way.

Review of diagnostics

- Variance inflation factors VIF_j tell you which predictors are highly correlated with other predictors. If you have one or more $\text{VIF}_j > 10$, you may want to eliminate some of the predictors.

  Multicollinearity affects the interpretability of the model, but does not indicate the model is "bad" in any way.

  An alternative approach that allows keeping correlated predictors is ridge regression (Chapter 11).
- Deleted residuals $t_i \sim t_{n-p-1}$, so you can formally define an outlier as being larger than $t_{1-\alpha/(2n),n-p-1}$.

Review of diagnostics

- Residual plots. Plots of e_i or t_i vs. $\hat{Y}_i$ and versus each $x_1, \ldots, x_k$ help assess (a) correct functional form, (b) constant variance, and (c) outlying observations. They may also suggest a transformation for a predictor or two.

Review of diagnostics

- Residual plots. Plots of e_i or t_i vs. $\hat{Y}_i$ and versus each $x_1, \ldots, x_k$ help assess (a) correct functional form, (b) constant variance, and (c) outlying observations. They may also suggest a transformation for a predictor or two.
  - Heteroscedasticy can be corrected by transforming Y, or else modeling the variance directly (Chapter 11).

Review of diagnostics

- Residual plots. Plots of e_i or t_i vs. $\hat{Y}_i$ and versus each $x_1, \ldots, x_k$ help assess (a) correct functional form, (b) constant variance, and (c) outlying observations. They may also suggest a transformation for a predictor or two.
  - Heteroscedasticy can be corrected by transforming Y, or else modeling the variance directly (Chapter 11).
  - Constant variance but nonlinear patterns can be accommodated by introducing quadratic terms.

Review of diagnostics

- Residual plots. Plots of e_i or t_i vs. $\hat{Y}_i$ and versus each $x_1, \ldots, x_k$ help assess (a) correct functional form, (b) constant variance, and (c) outlying observations. They may also suggest a transformation for a predictor or two.
  - Heteroscedasticy can be corrected by transforming Y, or else modeling the variance directly (Chapter 11).
  - Constant variance but nonlinear patterns can be accommodated by introducing quadratic terms.

  Added variable plots help figure out functional form of predictors, and whether significance is being driven by one or two points only.

Review of diagnostics

- **DFFIT**$_i$ and **Cook's distance** D_i tell you which observations *influence* the fitted model the most. Sometimes one or two points can drive the significance of a predictor.

Review of diagnostics

- **DFFIT**$_i$ and **Cook's distance** D_i tell you which observations *influence* the fitted model the most. Sometimes one or two points can drive the significance of a predictor.
- **Leverages** tell you which points *can potentially* influence the fitted model.

Review of diagnostics

- **DFFIT**$_i$ and **Cook's distance** D_i tell you which observations *influence* the fitted model the most. Sometimes one or two points can drive the significance of a predictor.
- **Leverages** tell you which points *can potentially* influence the fitted model.
- A **normal Q-Q plot** of the residuals will indicate departures from normality.

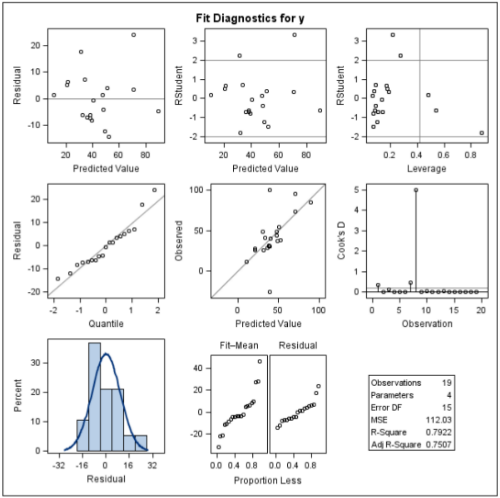
Review of diagnostics

- **DFFIT**$_i$ and **Cook's distance** D_i tell you which observations *influence* the fitted model the most. Sometimes one or two points can drive the significance of a predictor.
- **Leverages** tell you which points *can potentially* influence the fitted model.
- A **normal Q-Q plot** of the residuals will indicate departures from normality.
- A list of the studentized deleted residuals, leverages, and Cook's distances helps to determine outlying values that may be transcription errors or data anomalies and also indicates those observations that affect the fitted regression surface as a whole.

Standard diagnostic plots

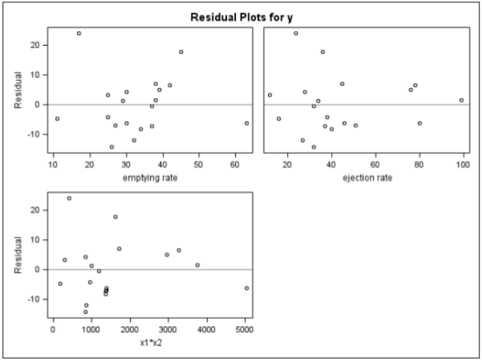
- t_i vs. h_i. Which observations are outlying in **x**-direction, outlying in Y-direction, or both?
- D_i vs. i. Which observations grossly affect fit of regression surface?
- e_i vs. $\hat{Y}_i$ and t_i vs. $\hat{Y}_i$. Constant variance & linearity.
- Y_i vs. $\hat{Y}_i$; how well model predicts its own data. Better models have points close to line $y = x$.
- Normal probability plot of the $e_1, \ldots, e_n$.
- Histogram of $e_1, \ldots, e_n$.
- Plots of e_i vs. each predictor $x_1, \ldots, x_k$.
- One more plot that prof. Hanson never looks at.

An example of diagnostics



Model is $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i$. One highly influential point & one poorly fit.

Residual plots



These look pretty good, except for the one large residual.

Arterial pressure data

```
proc glm;
 model y=x1 x2 x1*x2;
 output out=out cookd=c rstudent=r; run;
proc print; var x1 x2 y c r; run;
-------------------------------------------------------------------
```

Obs	x1	x2	y	c	r
1	45	36	49	0.36904	2.20950
2	30	28	55	0.00383	0.39889
3	11	16	85	0.12052	-0.62921
4	30	46	32	0.00885	-0.60493
5	39	76	26	0.01498	0.51721
6	42	78	28	0.02392	0.66178
7	17	24	95	0.45892	3.31414
8	63	80	26	4.99081	-1.77941
9	25	12	74	0.00724	0.33794
10	32	27	37	0.04100	-1.22324
11	37	37	31	0.01660	-0.71526
12	29	34	49	0.00032	0.12816
13	26	32	38	0.04023	-1.45743
14	38	45	41	0.01271	0.69211
15	38	99	12	0.00817	0.18206
16	25	38	44	0.00422	-0.40213
17	27	51	29	0.02196	-0.70921
18	37	32	40	0.00014	-0.05730
19	34	40	31	0.01371	-0.80210

Obs. 7 has largest arterial pressure. Obs. 8 has relatively small arterial pressure.

Dropping obs. 8 and obs. 7

```
proc glm data=out; model y=x1 x2 x1*x2; run;
--------------------------------------------------------------------
                                    Standard
Parameter         Estimate             Error    t Value    Pr > |t|
Intercept      134.3998664       15.98159869       8.41      <.0001
x1              -2.1330220        0.52215739      -4.09      0.0010
x2              -1.6993299        0.36366865      -4.67      0.0003
x1*x2            0.0333471        0.00928281       3.59      0.0027
--------------------------------------------------------------------
proc glm data=out(where=(c<4)); model y=x1 x2 x1*x2; run;
--------------------------------------------------------------------
                                    Standard
Parameter         Estimate             Error    t Value    Pr > |t|
Intercept      157.5094488       19.79515582       7.96      <.0001
x1              -2.7122125        0.58667658      -4.62      0.0004
x2              -2.7743376        0.69321545      -4.00      0.0013
x1*x2            0.0618590        0.01822201       3.39      0.0044
--------------------------------------------------------------------
proc glm data=out(where=(abs(r)<3)); model y=x1 x2 x1*x2; run;
--------------------------------------------------------------------
                                    Standard
Parameter         Estimate             Error    t Value    Pr > |t|
Intercept      116.3928224       13.52293668       8.61      <.0001
x1              -1.6161083        0.43361763      -3.73      0.0023
x2              -1.4903775        0.28875668      -5.16      0.0001
x1*x2            0.0272510        0.00742428       3.67      0.0025
```

How do 7 and 8 affect the significance and/or magnitude of the effects?