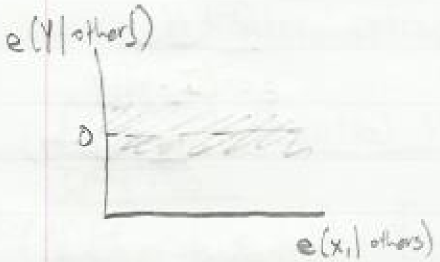
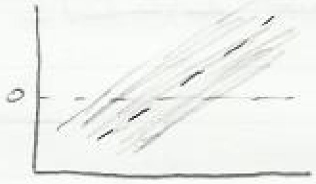
Chapter 10

Added Variable Plots Show the marginal role of a predictor variable given all other predictors are in the model



↑ x_1 contains no new information for predicting Y



↑ x_1 may be a helpful linear addition to the model

↑ x_1 may be a helpful factor to the model (maybe an interaction)

SAS

```
proc reg;
  model y = x1 x2 / partial;
run;
```

Finding Outlying observations

sgscatter

① Outlying w/ respect to its y value, x value or both

(2,3,4) Outlying w/ respect to x as they are larger than the other cases

(1,2) Look like they won't be too influential

(3,4) Will be influential to the regression line

Residuals $e_i = Y_i - \hat{Y}_i$

Semistudentized Residuals $e_i^* = \frac{e_i}{\sqrt{MSE}}$

Hat Matrix $\underline{H}_{n \times n} = \underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}'$ Note $\underline{\hat{Y}} = \underline{H}\underline{Y}$

$0 \le h_{ii} \le 1$ $e = (\underline{I} - H)\underline{Y}$

$\sum_i h_{ii} = p$ $\sigma^2(e) = \underline{\sigma}^2(\underline{I} - H)$

Studentized Residuals $r_i = \frac{e_i}{se\{e_i\}}$ (student residual)

Deleted Residual d_i = actual data - predicted value from model w/o said data point

$$d = Y_i - \hat{Y}_{i(i)}$$
$$= \frac{e_i}{1 - h_{ii}}$$

Studentized Deleted Residuals $t_i = d_i / s\{d_i\}$ (RStudent)

$$= e_i / \sqrt{MSE_{(i)}(1 - h_{ii})}$$
$$= e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2}$$

Outlying Test ①   Bonferroni Critical Value $t(1-\alpha/2; n-p-1) = qt(1-\alpha/2, n-p-1)$ in R

Test: If |Student Deleted Residual| > Bonferroni Critical Value $\exists$ outlier

else we're okay

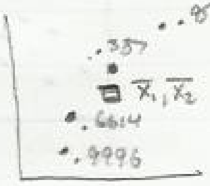
Outlying Test ②   Hat Matrix for Outlying Observation   (Hat Diag H)

- h_{ii} is called the leverage, distance from the center
- h_{ii} is considered large if $h_{ii} > 2\bar{h} = \frac{2\sum h_{ii}}{n} = \frac{2p}{n}$
- Another guideline: $h_{ii} > .5$ high leverage

  $.2 \le h_{ii} \le .5$ moderate leverage

  $.2 < h_{ii}$ low leverage

  } large n



Hat matrix for Extrapolation

$$h_{new,new} = X'_{new}(X'X)^{-1}X_{new}$$

*If $h_{new,new}$ is much larger than the leverage values it indicates extrapolation.

Finding Influential Points

DFFITS

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} = t \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

* Considered influential if $\left\{ \begin{array}{ll} |DFFITS|_i > 1 & \text{for small data} \\ |DFFITS| > 2\sqrt{\frac{p}{n}} & \text{for large data sets} \end{array} \right\}$

Cook's D

$$D_i = \frac{\sum (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE} = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{pMSE} = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

* Considered influential if $pF(D_i, p, n-p) \geq .5$

moderately if $.2 \leq pF(D_i, p, n-p) \leq .5$

DFBetas

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

c_{kk} is the k^{th} diag of $(X'X)^{-1}$

* considered influential if $\left\{ \begin{array}{ll} > 1 & \text{for small data sets} \\ > 2/\sqrt{n} & \text{for large data sets} \end{array} \right\}$

VIF > 10 indicate multicollinearity

Sas

```
proc reg;
  model y= x1 x2 x3 / r influence vif;
run;
```