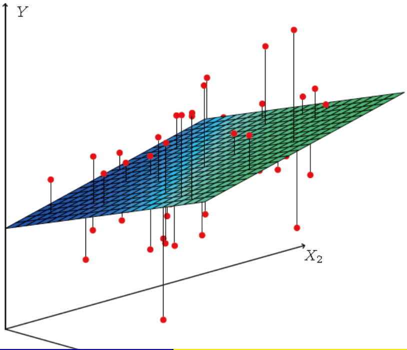
Regression

We have

- list of features $X_1, \ldots, X_p$
- numerical goal variable Y
- training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
- $\mathbf{y} = (y_1, \ldots, y_N)$ denotes the vector of training goal data.

We have or choose

- error measure (loss function) $L(y, \hat{y})$
    - square error loss $L(y, \hat{y}) = (y - \hat{y})^2$



Linear Regression

- assumption about the function $f(X) \approx Y$
  - we assume linear dependence:

$$
\begin{aligned}
f(X) &= \beta_0 + \sum_{i=1}^{p} X_j \beta_j \\
Y &= f(X) + \epsilon \\
\epsilon &\sim N(0, \sigma^2)
\end{aligned}
$$

  σ^2 does not depend on X nor Y
  x_i fixed (not random).

- If $\mathbf{X^T X}$ is not singular, then the unique solution is given by

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

  hat matrix $H = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T}$.

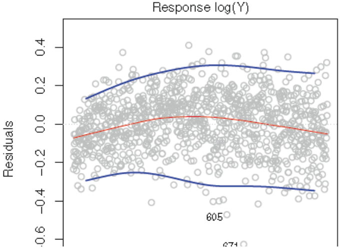
- the estimate $\hat{y}_i$ for a given x_i is $\hat{y}_i = \hat{y}(x_i) = x_i^T \hat{\beta}$.

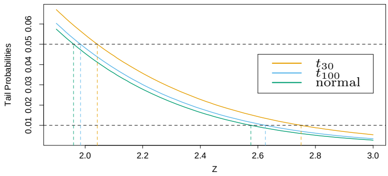
Standard Error, Interval Estimate

- What is the error of the estimate?
- we estimate the variance

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

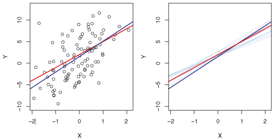
- The $N-p-1$ makes the estimate unbiased, $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$.
- **residual standard error** $\hat{\sigma}$
- and it is with approximately 95% probability in the interval $\hat{y} \in (\hat{y} - 2\sigma, \hat{y} + 2\sigma)$.





Accuracy of Coefficient Estimates

- Different training data lead to different estimates.(red-true, blue-estimated models)



- We assume:

$$
\begin{aligned}
Y &= \mathbb{E}(Y|X_1,\ldots,X_p)+\epsilon \\
&= \beta_0 + \sum_{i=1}^{p} X_j\beta_j + \epsilon
\end{aligned}
$$

- Therefore

$$
\begin{aligned}
\hat{\beta} &\sim N(\beta, (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\sigma^{2}) \\
(N-p-1)\hat{\sigma}^2 &\sim \sigma^2\chi^2_{N-p-1}
\end{aligned}
$$

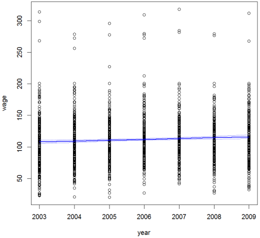
Accuracy of Coefficient Estimates

- For any single β_j, Z-score is (v_j is the j-th diagonal element of $(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$):

$$z_j = \frac{\hat{\beta}}{\hat{\sigma}\sqrt{v_j}}$$

- The entire parameter vector β bounds:

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



Importance of Features

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2595.8616   752.8243  -3.448 0.000572 ***
year            1.3499     0.3753   3.597 0.000328 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.65 on 2998 degrees of freedom
Multiple R-squared:  0.004296,  Adjusted R-squared:  0.003964
F-statistic: 12.94 on 1 and 2998 DF,  p-value: 0.0003277
```

R^2, F Statistics – Comparisons with the Trivial Model

- The proportion of variance explained
- Comparison with the Trivial Model $TSS = \sum_{i=1}^{N}(y_i - \overline{y})^2$
- scale independent, always in [0,1]

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Previous Slide example: wage $R^2 = 0.0043$ is very low.

F measure

- Hypothesis $H_0 \equiv$'coefficients $\beta_{p_0+1}, \ldots, \beta_{p_1}$ are zero, alternative $H_a \equiv$ 'at least one $\beta_i, i = p_0 + 1, \ldots, p_1$ is non–zero'
- $F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}$
- p–value says the probability 'such or further from null–model' data given H_0.

Computational methods

- Cholevsky decomposition; $p^3 + N\frac{p^2}{2}$ operations
  - Decompose $\mathbf{X}^T\mathbf{X}$ to LL^T, where L is a lower diagonal matrix.
- QR decomposition; Np^2 operations

  Regression by Successive Orthogonalization

  1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = 1$.
  2. For $j = 1, 2, \ldots, p$

     Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \frac{\langle \mathbf{z}_\ell, \mathbf{x}_j \rangle}{\langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle}$, $\ell = 0, 1, \ldots, j-1$ and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_{k-1}$.
  3. Regress $\mathbf{y}$ on the residual $\mathbf{z_p}$ to give the estimate $\hat{\beta}_p$.

  - $\mathbf{X} = \mathbf{Z}\Gamma$
  - $\mathbf{Z}$ has $\mathbf{z_j}$ as columns, Γ is the upper triangular matrix with entries $\hat{\gamma}_{kj}$.
  - introducing the diagonal matrix $\mathbf{D}$, $D_{jj} = \|z_j\|$

$$
\begin{aligned}
\mathbf{X} &= \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\Gamma \\
&= \mathbf{Q}\mathbf{R}
\end{aligned}
$$

- We get:

$$
\begin{aligned}
\hat{\beta} &= \mathbf{R}^{-1}\mathbf{Q}^\mathbf{T}\mathbf{y} \\
\hat{y} &= \mathbf{Q}\mathbf{Q}^\mathbf{T}\mathbf{y}
\end{aligned}
$$