HKU BUSINESS SCHOOL 港大經管學院

Business Statistics

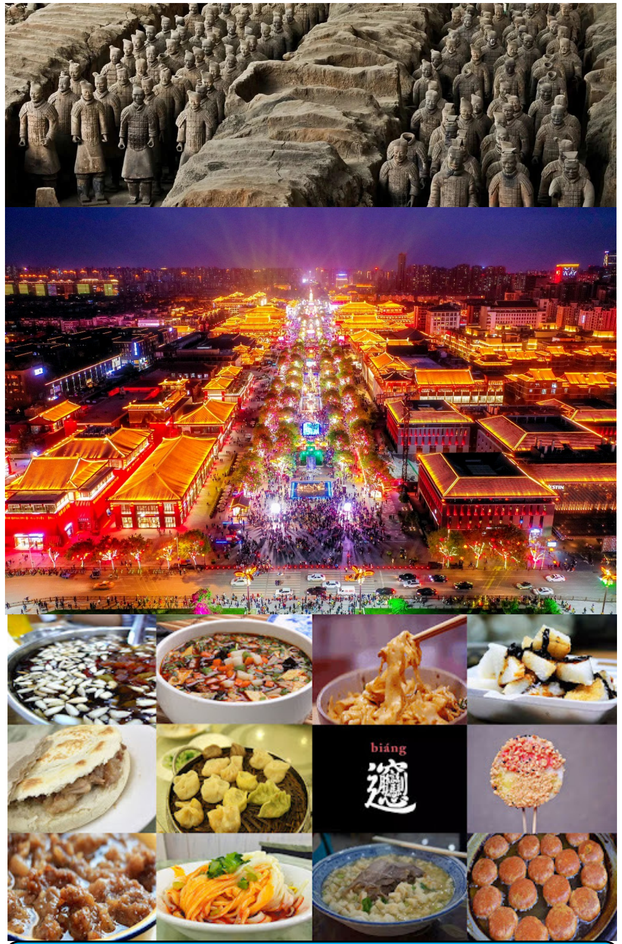
Poisson Regression and Generalized Linear Models

Weichen Wang

Assistant Professor
Innovation and Information Management

ISLR Chapter 4.6

WANG Weichen 王煒辰



Hometown

07-11 Tsinghua University
BS in Math and Physics

11-16 Princeton University
PhD in Financial Engineering

16-21 Two Sigma Investments
Quantitative Researcher

21-… HKU
Assistant Professor

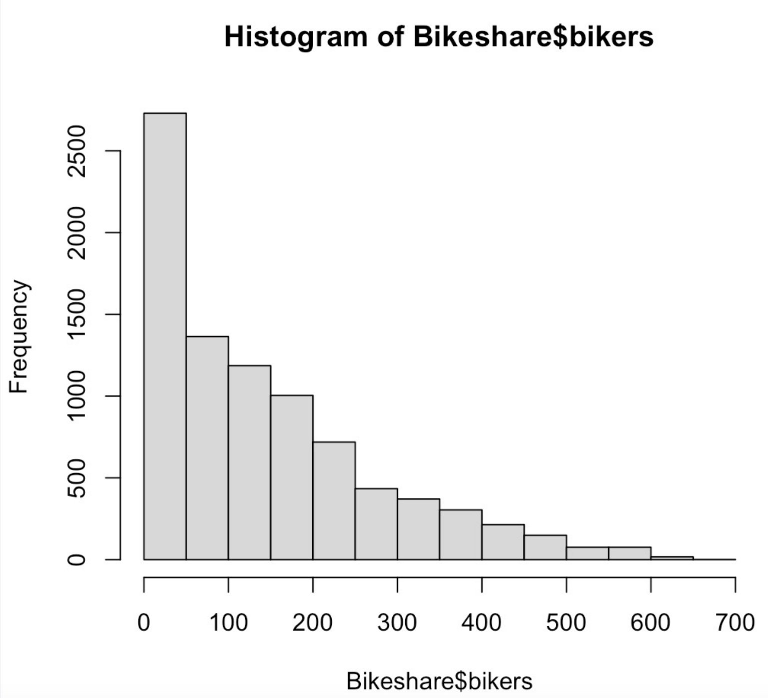
Quick Review up to now

- Linear regression: model assumptions, diagnostics, estimation, inference, transformation, implementation, interpretation…
- Model selection and regularization: forward/backward selection, best subset, Lasso, ridge regression, cross-validation, AIC/BIC…
- Binary/Multinomial logistic regression: MLE, logit, odds ratio, nominal vs ordinal response, implementation, interpretation…

Bikeshare data

- In linear regression, Y is quantitative.
- In logistic regression / classification, Y is qualitative (categorical).
- What if Y is neither of them? E.g., counts.

Here $Y =$ bikers, # hourly users of a bike sharing program in Washington.



Features to predict bikers?

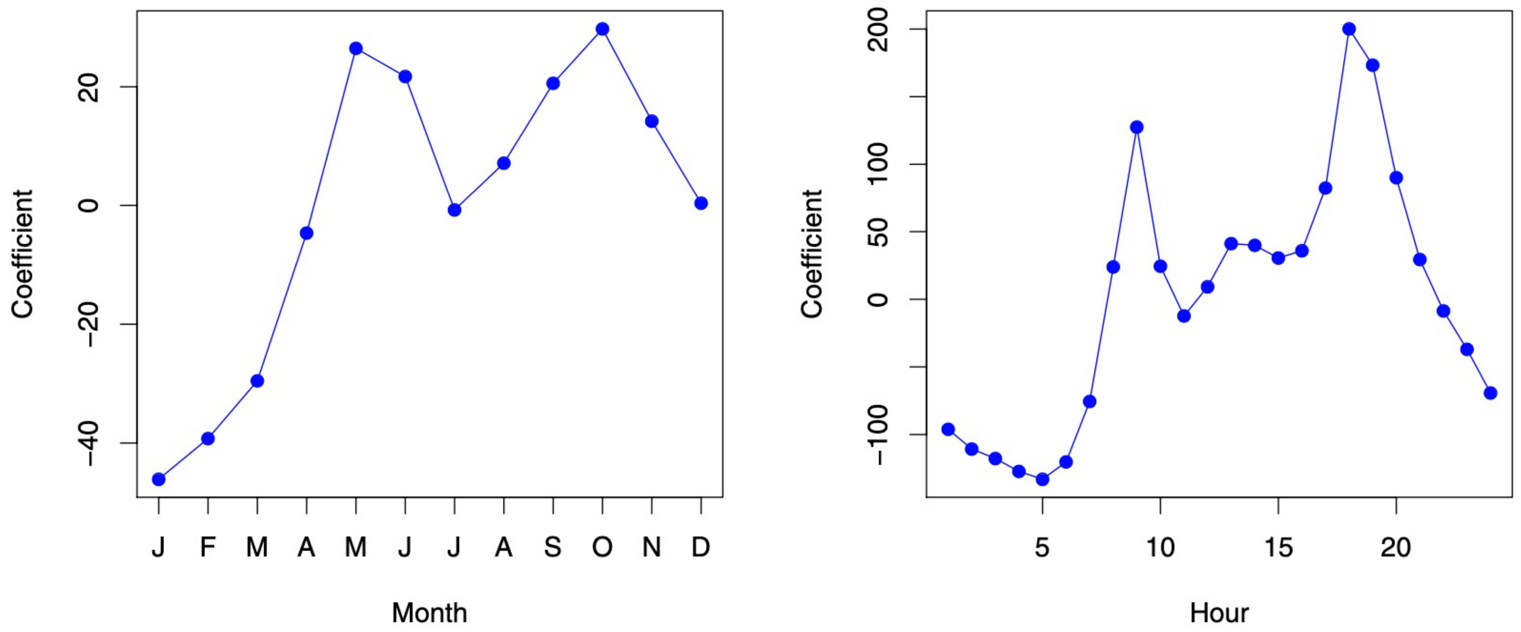
- Temporal dimension?
- Spatial dimension?
- Other consideration?

Bikeshare data

- mnth (month of the year), qualitative
- hr (hour of the day, from 0 to 23), qualitative
- workingday (1 if it is neither a weekend nor a holiday), qualitative
- temp (temperature in Celsius)
- weathersit (clear; misty or cloudy; light rain or light snow; or heavy rain or heavy snow), qualitative

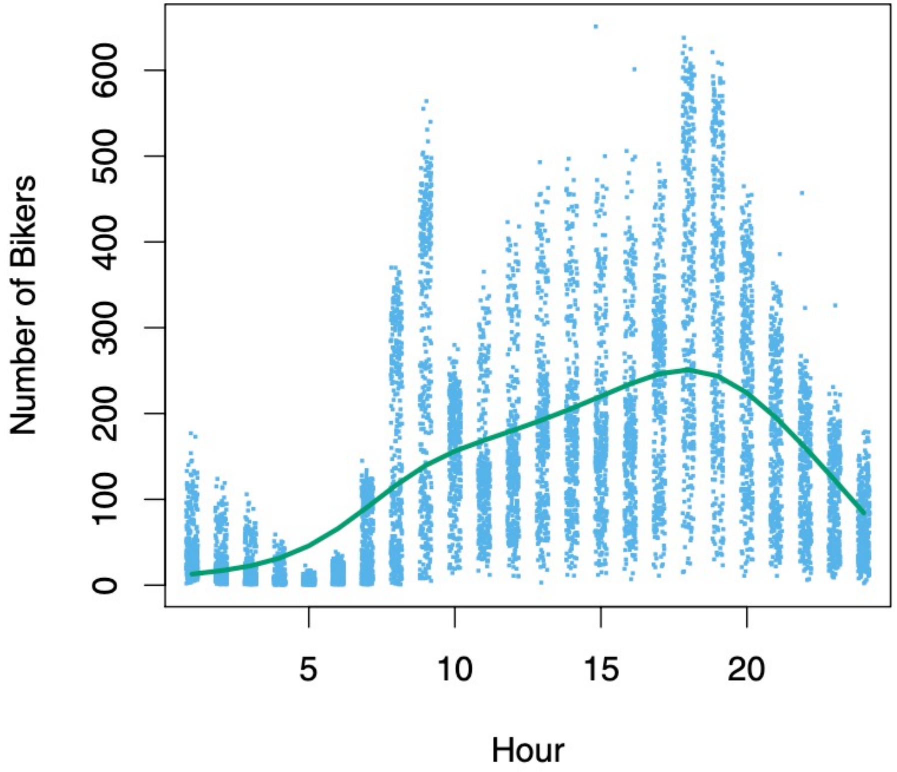
Linear Regression on the Bikeshare Data

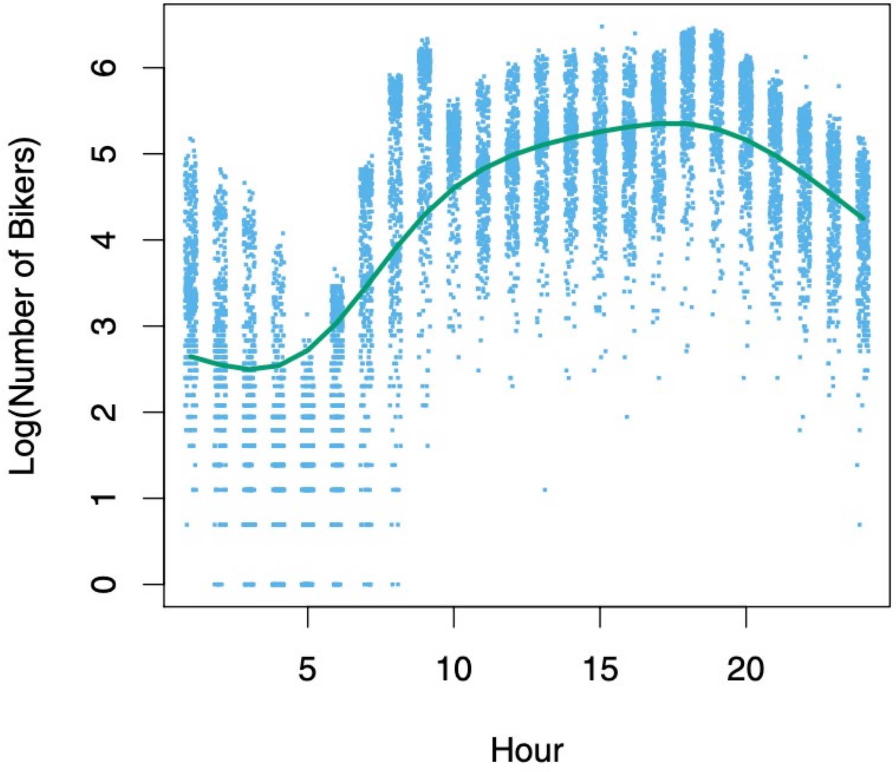
	Coefficient	Std. error	z-statistic	p-value
`Intercept`	73.60	5.13	14.34	0.00
`workingday`	1.27	1.78	0.71	0.48
`temp`	157.21	10.26	15.32	0.00
`weathersit[cloudy/misty]`	-12.89	1.96	-6.56	0.00
`weathersit[light rain/snow]`	-66.49	2.97	-22.43	0.00
`weathersit[heavy rain/snow]`	-109.75	76.67	-1.43	0.15



Issues of Linear Regression

- 9.6% of the fitted values are negative.
- When the expected value of bikers is small, the variance of bikers should also be small.





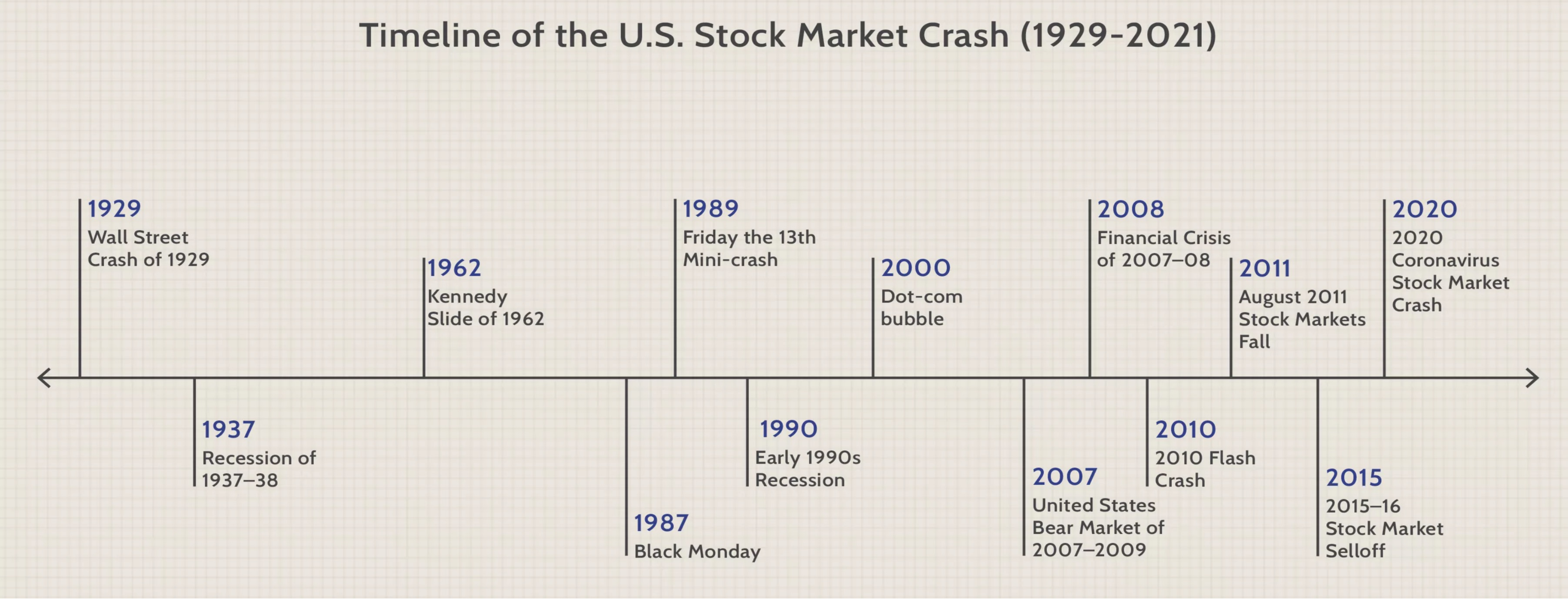
Poisson Distribution

Suppose that a random variable Y takes on nonnegative integer values, i.e. $Y \in \{0, 1, 2, \ldots\}$. If Y follows the Poisson distribution, then

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \ldots. \tag{4.35}$$

Here, $\lambda > 0$ is the expected value of Y, i.e. $\mathrm{E}(Y)$. It turns out that λ also equals the variance of Y, i.e. $\lambda = \mathrm{E}(Y) = \mathrm{Var}(Y)$.

- Poisson distribution is typically used to model counts.



- Assume #crisis in one year follows Possion ($\lambda = 0.13$, in past 100 years, we have 13 crises), what is the probability next year we will see at least one crisis?

$$1 - P(Y = 0) = 1 - \frac{e^{-\lambda}\lambda^0}{0!} = 1 - e^{-0.13} = 12.2\%$$

Poisson Regression

- Three ingredients:
  - Likelihood of response = Poisson density

$$\ell(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!},$$

  - Find expectation of y condition on x

$$\mathrm{E}(Y|X_1, \dots, X_p) = \lambda(X_1, \dots, X_p)$$

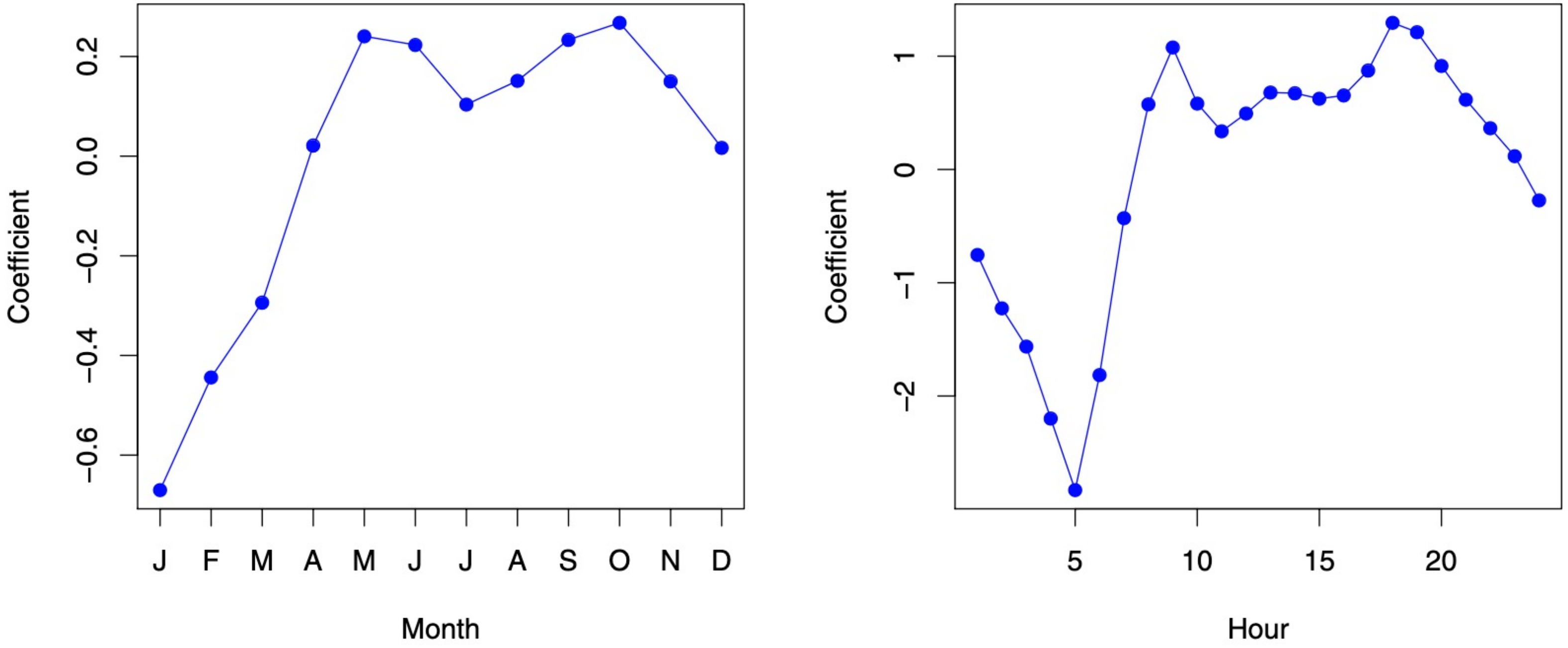
  - Link expectation with the linear function of predictors

$$\log(\lambda(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Estimate β by maximizing the likelihood.

Poisson Regression on the Bikeshare Data

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	4.12	0.01	683.96	0.00
`workingday`	0.01	0.00	7.5	0.00
`temp`	0.79	0.01	68.43	0.00
`weathersit[cloudy/misty]`	-0.08	0.00	-34.53	0.00
`weathersit[light rain/snow]`	-0.58	0.00	-141.91	0.00
`weathersit[heavy rain/snow]`	-0.93	0.17	-5.55	0.00



Poisson vs Linear Regression

- An increase in X_j by one unit is associated with a change in $E(Y) = \lambda = e^{X\beta}$ by a factor of $\exp(\beta_j)$.
  - By contrast, in linear regression, an increase in X_j by one unit is associated with an increase of β_j in $E(Y) = \mu = X\beta$.

- Poisson regression implicitly assumes that mean bike usage equals the variance of bike usage.
  - By contrast, linear regression assumes the variance always takes on a constant value.

- Poisson regression gives non-negative predictions.
  - By contrast, almost 10% of the predictions in Bikeshare data were negative.

R implementation

- Recall logistic R command:

  *fit <- **glm**(Y~X, data, family=**binomial**(logit))*

- "glm" here means generalized linear models.

```
> mod.pois <- glm(
    bikers ~ mnth + hr + workingday + temp + weathersit,
    data = Bikeshare, family = poisson
  )
> summary(mod.pois)
  Call:
  glm(formula = bikers ~ mnth + hr + workingday + temp + weathersit,
      family = poisson, data = Bikeshare)

  Deviance Residuals:
       Min        1Q    Median        3Q       Max
  -20.7574   -3.3441   -0.6549    2.6999   21.9628

  Coefficients:
                        Estimate Std. Error  z value Pr(>|z|)
  (Intercept)           4.118245   0.006021  683.964  < 2e-16 ***
  mnth1                -0.670170   0.005907 -113.445  < 2e-16 ***
  mnth2                -0.444124   0.004860  -91.379  < 2e-16 ***
  mnth3                -0.293733   0.004144  -70.886  < 2e-16 ***
  mnth4                 0.021523   0.003125    6.888 5.66e-12 ***
```

```
...
hr20                       0.914022   0.003700  247.065  < 2e-16 ***
hr21                       0.616201   0.004191  147.045  < 2e-16 ***
hr22                       0.364181   0.004659   78.173  < 2e-16 ***
hr23                       0.117493   0.005225   22.488  < 2e-16 ***
workingday                 0.014665   0.001955    7.502 6.27e-14 ***
temp                       0.785292   0.011475   68.434  < 2e-16 ***
weathersitcloudy/misty    -0.075231   0.002179  -34.528  < 2e-16 ***
weathersitlight rain/snow -0.575800   0.004058 -141.905  < 2e-16 ***
weathersitheavy rain/snow -0.926287   0.166782   -5.554 2.79e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1052921  on 8644  degrees of freedom
Residual deviance:  228041  on 8605  degrees of freedom
AIC: 281159

Number of Fisher Scoring iterations: 5
```

Generalized Linear Models

- Three ingredients:
  - Likelihood of response = some distribution from exponential family

    E.g., Gaussian, Binomial, Poisson, Exponential, Gamma, Negative binomial …

  - Find expectation of y

    E.g., Gaussian $E[y|x] = \mu$, Logistic $E[y|x] = P(y = 1|x)$, Poisson $E[y|x] = \lambda$ …

  - Link the expectation with the linear function of predictors with a link function $\eta(\cdot)$

$$\eta(\mathrm{E}(Y|X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

    E.g., Gaussian $\eta(\mu) = \mu$, Logistic $\eta(\mu) = \log(\mu/(1-\mu))$, Poisson $\eta(\mu) = \log(\mu)$.

- Estimate β by maximizing the likelihood.

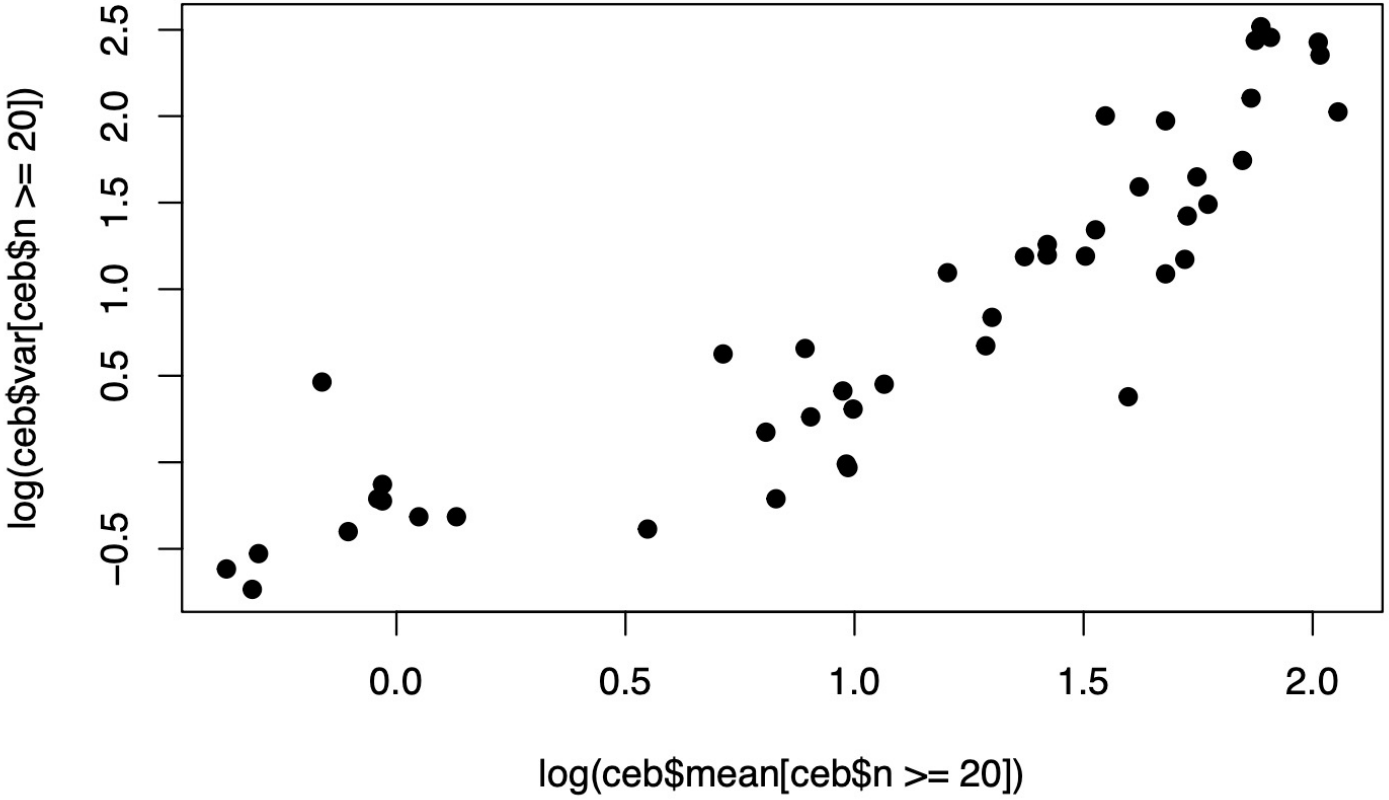
Case Study: Children Ever Born in Fiji

- The dataset has 70 rows representing grouped individual data.
- Each row has entries for:
  - The cell number
  - marriage duration (1=0-4, 2=5-9, 3=10-14, 4=15-19, 5=20-24, 6=25-29)
  - residence (1=Suva, 2=Urban, 3=Rural)
  - education (1=none, 2=lower primary, 3=upper primary, 4=secondary+)
  - mean number of children ever born
  - variance of children ever born
  - number of women in the cell

	dur	res	educ	mean	var	n	y
1	0-4	Suva	none	0.50	1.14	8	4
2	0-4	Suva	lower	1.14	0.73	21	24
3	0-4	Suva	upper	0.90	0.67	42	38
4	0-4	Suva	sec+	0.73	0.48	51	37
5	0-4	urban	none	1.17	1.06	12	14
6	0-4	urban	lower	0.85	1.59	27	23

Case Study: Children Ever Born in Fiji

- Check the mean-variance relationship first.



  – If variance is significantly different from mean, we need additional degree of freedom to fit the data better. We may try the negative binomial regression (not required for this course).

Case Study: Children Ever Born in Fiji

Are we able to still run Poisson regression? Yes, but using the nice "offset" feature of "glm" function. Suppose the l-th woman in a group has Y_l babies. The group total is $Y = \sum_{l=1}^{n} Y_l$ and we have n women in this group. In Poisson regression, if we know each Y_l, we assume

$$\log E[Y_l] = X'\beta.$$

Note that all Y_l's have shared predictors X. Then we have

$$\log E[Y] = \log E[\sum_{l=1}^{n} Y_l] = \log \sum_{l=1}^{n} E[Y_l] = \log nE[Y_l] = \log n + X'\beta.$$

```
fit.ceb <- glm(y ~ dur + res + educ + offset(log(n)),
               family=poisson, data=ceb, contrasts=contrasts)
summary(fit.ceb)
```

Case Study: Children Ever Born in Fiji

```
Call:
glm(formula = y ~ dur + res + educ + offset(log(n)), family = poisson,
    data = ceb, contrasts = contrasts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2960  -0.6641   0.0725   0.6336   3.6782

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11710    0.05491  -2.132 0.032969 *
dur5-9       0.99693    0.05274  18.902  < 2e-16 ***
dur10-14     1.36940    0.05107  26.815  < 2e-16 ***
dur15-19     1.61376    0.05119  31.522  < 2e-16 ***
dur20-24     1.78491    0.05121  34.852  < 2e-16 ***
dur25-29     1.97641    0.05003  39.501  < 2e-16 ***
resrural     0.15166    0.02833   5.353 8.63e-08 ***
resurban     0.11242    0.03250   3.459 0.000541 ***
educlower    0.02297    0.02266   1.014 0.310597
educupper   -0.10127    0.03099  -3.268 0.001082 **
educsec+    -0.31015    0.05521  -5.618 1.94e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3731.852  on 69  degrees of freedom
Residual deviance:   70.665  on 59  degrees of freedom
AIC: 522.14

Number of Fisher Scoring iterations: 4
```

- Expected #children for a Suvanese woman with no education and married 0-4 years? $e^{-0.117} = 0.89$
- By duration 25-29, the expected #children increases by $e^{1.976} = 7.22$ times compared to duration 0-4.
- With secondary or higher education, women are expected to have $1 - e^{-0.310} = 27\%$ fewer kids.

Case Study: Children Ever Born in Fiji

- Is there interaction effect between say education and marital duration?

	dur	res	educ	n	prediction
1	0-4	Suva	none	1	0.8894987
2	0-4	Suva	sec+	1	0.6523026
3	5-9	Suva	none	1	2.4105084
4	5-9	Suva	sec+	1	1.7677157
5	10-14	Suva	none	1	3.4983737
6	10-14	Suva	sec+	1	2.5654879
7	15-19	Suva	none	1	4.4667458
8	15-19	Suva	sec+	1	3.2756312
9	20-24	Suva	none	1	5.3005682
10	20-24	Suva	sec+	1	3.8871043
11	25-29	Suva	none	1	6.4192929
12	25-29	Suva	sec+	1	4.7075068

Summary

- Poisson regression is used for modeling count data.

- Three ingredients of GLM:

  Distribution of Y + Expression of $E[Y]$ + link function $\eta(E[Y|X]) = X'\beta$

- Estimate β by MLE.

- Explain e^{β_k} as the multiplicative change in $E[Y]$ by changing X_k by one unit.